ISSN 2225-6016

ВЕСТНИК

Смоленской государственной медицинской академии

Том 18, №2

2019

УДК 519 253

МЕТОДИКА ОПИСАТЕЛЬНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА НОМИНАЛЬНЫХ ПРИЗНАКОВ В ВЫБОРКАХ МАЛОГО ОБЪЕМА, ПОЛУЧЕННЫХ В РЕЗУЛЬТАТЕ ФАРМАКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

© Лямец Л.Л., Евсеев А.В.

Смоленский государственный медицинский университет, Россия, 214019, Смоленск, ул. Крупской, 28

Резюме

Цель. Цель теоретического исследования заключалась в разработке методики описательного статистического анализа номинальных признаков, т.е. признаков, измеренных в номинальной шкале. Методика разрабатывалась для анализа результатов экспериментальных фармакологических исследований, которые обычно по объективным причинам представлены выборочными совокупностями (выборками) малого объема с числом единиц наблюдений не более 30. Методика представляет собой алгоритм вычислительных действий, который позволит обеспечить проведение статистического анализа номинальных признаков, используемых для описания фармакологических эффектов.

Методика. Проведен обзорный анализ публикаций по фармакологии, в которых для получения новых знаний и обоснования результатов исследований использовались статистические методы анализа экспериментальных данных. В результате обзора выявлены основные, наиболее часто встречающиеся исследовательские задачи, требующие статистического анализа признаков, измеренных в номинальной шкале. Проведена систематизация вычислительных операций, необходимых для проведения статистического анализа номинальных признаков в типичных исследовательских задачах. На основе систематизированных вычислительных операций разработана методика (алгоритм статистического анализа) номинальных признаков, которая позволит обеспечить количественное обоснование индуктивных выводов в научных исследованиях и положений, выносимых на защиту в диссертационных работах по фармакологической тематике.

Результаты. Разработана и обоснована методика для проведения описательного статистического анализа номинальных признаков в фармакологических исследованиях. Методика ориентирована на анализ выборок малого объема в типичных, наиболее часто встречающихся исследовательских задачах. Для реализации методики предложены способы автоматизации вычислений с использованием табличного процессора Excel.

Заключение. В результате обзорного анализа публикаций по фармакологии проведена систематизация вычислительных операций, необходимых для проведения описательного статистического анализа номинальных признаков в типичных исследовательских задачах. На основе систематизированных вычислительных операций разработана методика (алгоритм статистического анализа) номинальных признаков. Показан практический пример автоматизации входящих в методику вычислений с использованием современных информационных технологий. Методика может представлять практический интерес для научных работников, осуществляющих исследования в области фармакологии и использующих в своей работе статистические методы анализа экспериментальных данных.

Ключевые слова: выборочный метод, номинальные признаки, описательный статистический анализ, интервальные оценки, проверка статистических гипотез

METHODOLOGY OF DESCRIPTIVE STATISTICAL ANALYSIS OF THE NOMINAL CHARACTERISTICS IN THE SMALL SAMPLE SIZES OBTAINED A A RESULTS OF PHARMACOLOGICAL STUDIES

Lyamets L.L., Evseev A.V.

Smolensk State Medical University, 28, Krupskoj St., 214019, Smolensk, Russia

Abstract

Objective. The aim of the theoretical study was to develop a method of descriptive statistical analysis of nominal characteristics, i.e. features measured in the nominal scale. The technique was developed to

analyze the results of experimental pharmacological studies, which are usually for objective reasons represented by sample sets (samples) of small volume with the number of units of observations not more than 30. The technique is an algorithm of computational actions, which will provide a statistical analysis of the nominal characteristics used to describe the pharmacological effects.

Method. A review analysis of publications on pharmacology, in which statistical methods of analysis of experimental data were used to obtain new knowledge and substantiate the results of studies, was carried out. The review identified the main, most common research tasks that require statistical analysis of features measured in the nominal scale. The systematization of computational operations necessary for the statistical analysis of nominal characteristics in typical research problems is carried out. On the basis of the systematized computational operations the technique (algorithm of the statistical analysis) of nominal signs which will allow to provide a quantitative justification of inductive conclusions in scientific researches and the positions taken out on protection in dissertations on pharmacological subjects is developed.

Results. The technique for descriptive statistical analysis of nominal characteristics in pharmacological studies is developed and justified. The technique is focused on the analysis of small samples in typical, most common research problems. To implement the methodology, the methods of automation of calculations using the Excel spreadsheet are proposed.

Conclusion. As a result of the review analysis of publications on pharmacology, the systematization of computational operations necessary for the descriptive statistical analysis of nominal characteristics in typical research problems is carried out. On the basis of systematic computational operations the technique (algorithm of statistical analysis) of nominal characteristics is developed. The practical example of automation of the calculations entering into a technique with use of modern information technologies is shown. The technique can be of practical interest for scientists who carry out research in the field of pharmacology and use in their work statistical methods of analysis of experimental data.

Keywords: sampling method, nominal characteristics, descriptive statistical analysis, interval estimates, statistical hypothesis testing

Введение

В настоящее время в доступных литературных источниках по статистическому анализу экспериментальных данных определено большое количество показателей и количественных характеристик (математических конструктов), которые могут быть использованы для описания, объяснения и прогнозирования процессов и явлений в исследуемых статистических совокупностях.

Разнообразие математических конструктов и их семантика дает широкую возможность для разработки программ научных исследований и количественного обоснования их результатов. В связи с этим разнообразием вычислительных действий и соответствующих им математических конструктов возникает необходимость в разработке рациональных методик статистического анализа экспериментальных данных, которые соответствуют целям и задачам конкретного научного исследования.

Первичная статистическая информация может быть получена на основе измерений с использованием номинальных шкал. Эти шкалы также называются шкалами наименований или классификационными шкалами. Номинальный тип шкал соответствует простейшему виду измерений. При этом объектам присваиваются шкальные значения – числа, которые используются лишь как «имена» или символы.

Простейшей номинальной шкалой является дихотомическая шкала. Дихотомическая шкала имеет всего две градации, которые кодируются соответствующими числовыми или (и) буквенными символами. В отношении градаций справедливы следующие утверждения. Все измеряемые единицы наблюдения, отнесенные к одной градации, эквивалентны между собой по определенному регистрируемому свойству. Единицы наблюдения, отнесенные к разным градациям, между собой не эквивалентны. Иные отношения между градациями не определены. Эти особенности номинальной дихотомической шкалы требуют применения соответствующих методов статистического анализа. Примером дихотомии при выполнении фармакологических исследований может служить обнаружение побочного действия после применения лекарственного препарата. При этом побочный эффект либо существует, либо отсутствует. Другой вариант дихотомии наблюдается, например, при проведении опытов по оценке острой токсичности химических соединений, претендующих на включение в перечень лекарственных средств [3]. В

результате измерения в дихотомической шкале отражают либо гибель опытного животного, либо его выживание после введения тестируемой дозы.

Примером дихотомии также является развитие у пациента ощущения эйфории или дисфории после введения морфиноподобных наркотических средств. Из клинических примеров дихотомию можно проиллюстрировать эффектом местных анестетиков при выполнении проводниковой анестезии – чувствительность сохраняется или исчезает, рефлекс регистрируется или отсутствует.

Кроме дихотомических шкал существуют политомические номинальные шкалы, которые имеют три и более градаций измеряемого свойства. Политомия также широко представлена в фармакологических исследованиях. В частности, при постановке экспериментов по оценке влияния веществ на сердечную деятельность могут выявляться (или не выявляться) те или иные эффекты, такие, например, как инотропное действие (изменение силы сокращения), хронотропное действие (изменение частоты сокращений), дромотропное действие (изменение распространения возбуждения по элементам проводящей системы сердца), батмотропное действие (изменение возбудимости миокарда). Следует отметить, что каждый из этих эффектов может быть как положительным, так и отрицательным, что, в свою очередь, является уже проявлением дихотомии [2, 4].

В качестве другого примера политомии может служить эксперимент по оценке спектра антибактериальной активности химиотерапевтических веществ, которые способны оказывать эффект в отношении различных видов бактерий, риккетсий, грибов, простейших. При этом вещество может либо убивать инфекционный агент, либо ограничивать его размножение, что следует рассматривать как дихотомию.

Для выявления и количественного описания индуктивных закономерностей на основе номинальных признаков, используются соответствующие методы статистического анализа и математические конструкты. Обычно статистическое исследование начинается с описательного количественного анализа.

Ниже приводится методика описательного количественного анализа номинальных признаков, которая ориентирована на исследователей, не имеющих специального математического образования, и построена на основе анализа типичных целей и задач, описанных в публикациях и литературных источниках по фармакологическим исследованиям. Предусмотренные методикой вычисления достаточно просто автоматизируются с использованием доступных информационных технологий. Вычисления и полученные на их основе математические конструкты обеспечивают количественное описание закономерностей в спланированных фармакологических исследованиях.

Таким образом, целью исследования явилась разработка методики описательного статистического анализа номинальных признаков, т.е. признаков, измеренных в номинальной шкале.

Методика

В основе любого научного исследования лежат соответствующие целям и задачам исследования научные методы. Для практической реализации выбранного научного метода в конкретном исследовании разрабатывается методика, которая представляет собой определенную процедуру или взаимосвязанную последовательность действий.

Особенность данного исследования заключается в том, что его целью являлась разработка методики статистического анализа номинальных признаков. Очевидно, что решение задачи, направленной на достижение этой цели изначально не имеет строго определенного алгоритма и является поисковой или эвристической. Поэтому для решения исследовательской задачи, использовались эвристические правила морфологического анализа и синтеза, разработанные швейцарским астрономом Ф. Цвики в 1930-х гг. Правила предписывают упорядоченный и систематизированный обзор всех возможных вариантов решений поставленной задачи. Они позволяют реализовать идеи системного подхода для решения данной поисковой задачи и содержат общие рекомендации по организации интеллектуальных действий. Правила морфологического анализа и синтеза формулируются следующим образом: на основании анализа имеющейся информации выбирается группа основных элементов рассматриваемого объекта или системы; для каждого элемента выбирается множество альтернативных вариантов реализации; комбинируя варианты, получают множество решений, из которых синтезируется наиболее рациональное.

Для решения поставленной эвристической задачи были выделены основные морфологические единицы – этапы реализации разрабатываемой методики. На основе анализа публикаций по фармакологии и статистическому анализу данных в доступной литературе для каждого этапа было составлено несколько возможных вариантов его реализации. Затем на основе комбинирования вариантов был осуществлен синтез наиболее рационального эвристического решения, основанного на соответствующих математических конструктах. Обязательным условием для разрабатываемой методики являлась возможность автоматизировать все необходимые вычисления с использованием современных информационных технологий.

Основными этапами разрабатываемой методики (морфологическими единицами) были выбраны следующие действия: 1) описание типичности и вариации номинального признака с использованием выборочных точечных и интервальных оценок; 2) проверка гипотезы о статистической связи (сопряженности) между номинальными признаками и оценка силы статистической связи между ними; 3) вычисление мощности используемых статистических критериев.

Результаты исследования и их обсуждение

В результате решения поисковой задачи на основе эвристических правил была разработана методика анализа номинальных признаков, отражающих основные и побочные фармакологические эффекты в выборочных статистических исследованиях. Корректное применение методики предполагает, что исследователем спланирована и проведена последовательность однотипных независимых испытаний (схема Бернулли), направленных на проверку каких-либо предположений опытным путем. Математический аппарат методики основан на выборочном методе статистического исследования. В основе лежат следующие положения. Объектом статистического исследования является формально определенная через множество признаков включения статистическая совокупность, которая в случае применения выборочного метода называется генеральной. Предметом исследования являются закономерности, присущие номинальным признакам, каждый из которых имеет определенное число градаций. Для количественного описания типичности проявления градаций можно использовать соответствующие этим градациям относительные частоты (вероятности) p_i. Применение выборочного метода статистического исследования основано на том, что подлежащая исследованию генеральная совокупность не может быть исследована сплошным методом, т.е. практически не может быть проведено бесконечно большое количество испытаний. Поэтому все генеральные статистические показатели, в том числе и генеральные вероятности p_i, являются величинами неизвестными. Для их оценки из генеральной совокупности на основе принципа случайного отбора формируется выборочная совокупность (выборка) ограниченного объема N. Выборочные совокупности, объем которых меньше 30 единиц наблюдения условно считаются малыми.

Методика включает в себя три основных этапа статистического анализа. Первый этап имеет своей целью вычисление выборочных точечных и интервальных оценок для вероятностей p_i, а также вычисление количественных оценок вариации исследуемых номинальных признаков. Также можно сказать, что цель первого этапа состоит в количественном описании статистических закономерностей, присущих типичности и вариации исследуемых номинальных признаков.

Для примера рассмотрим применение вычислительных операций для одного номинального признака A, имеющего k градаций. Пусть первичные экспериментальные данные получены в результате исследования выборочной совокупности объемом N единиц наблюдения. Соответственно проведено N однотипных независимых испытаний. В каждом испытании регистрируется проявившаяся градация номинального признака A. Если номинальный признак имеет k градаций, то для каждой градации A_i, $1 \le i \le k$, вычисляются абсолютные частоты f_i ее проявления в N испытаниях. Для автоматизации вычислений можно использовать статистическую функцию СЧЕТЕСЛИ табличного процессора Microsoft Excel.

Для абсолютных частот f_i вычисляются соответствующие им эмпирические относительные частоты или эмпирические вероятности $\overline{p}_i$. Вычисление эмпирических вероятностей производится по следующей формуле: $\overline{p}_i = \frac{f_i}{N}$, $1 \le i \le k$.

Эмпирические вероятности $\overline{p}_i$ являются приближенными точечными оценками соответствующих неизвестных генеральных вероятностей p_i. Так как выборка, включающая в себя случайным образом отобранные единицы наблюдения для проведения однотипных независимых испытаний, является случайным продуктом, то, следовательно, эмпирические вероятности $\overline{p}_i$, вычисленные на основе результатов проведенных испытаний, являются случайными величинами. Возникает необходимость оценить неизвестную величину через случайную величину. Оценить неизвестные генеральные вероятности p_i через случайные выборочные величины $\overline{p}_i$ можно с использованием интервальных вероятностных оценок – доверительных интервалов. Формальная запись доверительного интервала имеет вид: $P(a \le p_i \le b) = \gamma$, где γ – доверительная вероятность, a_i и b_i – границы доверительного интервала. По сути, доверительный интервал считается определенным, если для заданной вероятности γ вычислены границы доверительного интервала. В приведенной формальной записи символ P означает вероятность события, записанного в скобках в виде двойного неравенства. В данной методике для расчета границ доверительного интервала использован метод, основанный на биномиальном распределении [1]. Для вычисления границ доверительного интервала сначала необходимо задать доверительную вероятность γ и, следовательно, определить уровень значимости $\alpha = 1 - \gamma$. Для медико-биологических исследований вполне приемлемой является доверительная вероятность $\gamma = 0{,}95$ и уровень значимости $\alpha = 1 - 0{,}95 = 0{,}05$. Для заданного объема выборочной совокупности или числа испытаний N и вычисленной для градации A_i абсолютной частоты f_i нижняя граница доверительного интервала вычисляется по следующей формуле: $a_i = \dfrac{f_i}{f_i + (N - f_i + 1) \cdot F_1(d_1; d_2; \nu)}$, где $F_1(d_1; d_2; \nu)$ – квантиль порядка $\nu = 1 - \alpha/2$ статистического *F*-распределения (Фишера) со степенями свободы $d_1 = 2(N - f_i + 1)$ и $d_2 = 2 f_i$. Верхняя граница доверительного интервала вычисляется по формуле: $b_i = \dfrac{(f_i + 1) \cdot F_2(d_3; d_4; \nu)}{N - f_i + (f_i + 1) \cdot F_2(d_3; d_4; \nu)}$, где $F_2(d_3; d_4; \nu)$ – квантиль порядка $\nu = 1 - \alpha/2$ статистического ***F***–распределения (Фишера) со степенями свободы $d_3 = 2(f_i + 1)$ и $d_4 = 2(N - f_i)$. Автоматизировать вычисление квантилей $F_1(d_1; d_2; \nu)$ и $F_2(d_3; d_4; \nu)$ можно в программе Microsoft Excel с использованием статистических функций FРАСПОБР$(\alpha/2; d_1; d_2)$ и FРАСПОБР$(\alpha/2; d_3; d_4)$ соответственно.

Вычисленный доверительный интервал имеет важное практическое значение. Он позволяет дать интервальную вероятностную оценку неизвестной генеральной вероятности p_i. На основании анализа первичных данных можно обоснованно полагать, что с вероятностью γ неизвестный генеральный показатель p_i принадлежит интервалу $[a_i, b_i]$. Формальная запись этого вывода имеет вид: $P(p_i \in [a_i, b_i]) = \gamma$. С практической точки зрения интервальные оценки надежнее точечных оценок $\overline{p}_i$.

Количественная оценка вариации номинального признака позволяет судить об однородности результатов, полученных в однотипных независимых испытаниях. Вариация – это явление, присущее статистической совокупности (множеству единиц наблюдения) и выражающееся в том, что измеряемый признак варьирует, изменяется при переходе от одной единицы наблюдения к другой. Важно отметить, чем меньше вариация, тем больше однородность проведенных измерений и наоборот. Если, например, при проведении измерений с использованием номинального дихотомического признака все единицы наблюдения были отнесены к одной градации, то очевидно, что вариация отсутствует и результаты измерения максимально однородны. Следовательно, показатель, количественно оценивающий вариацию, должен быть равен нулю. В случае если все единицы наблюдения распределились поровну между двумя градациями, то вариация максимальна и однородность таких измерений минимальна. Для оценки вариации номинального признака A можно использовать коэффициент изменчивости категорий (IQV, от англ. index of qualitative variation). Этот коэффициент вычисляется как отношение наблюдаемой вариации к максимально возможной и может принимать значения от 0 до 1.

ормула для расчета IQVимеет вид: $IQV = \frac{k\left(N^2 - \sum_{i=1}^{k} f_i^2\right)}{N^2(k-1)}$, где N - объем выборочной совокупности (число испытаний); k - число градаций номинального признака; f_i - абсолютные частоты в градациях A_i, $1 \le i \le k$.

Пусть, например, признак A дихотомический, т.е. $k = 2$. Абсолютные частоты градаций A_1 и A_2 исследуемого признака соответственно равны $f_1 = N$; $f_2 = 0$. В данном случае вариация признака отсутствует, поскольку все единицы наблюдения в результате измерений отнесены к одной градации A_1. Градация A_2 не встретилась ни разу. Величина IQV, количественно оценивающая вариацию, равна нулю: $IQV = \frac{2\left(N^2 - \sum_{i=1}^{k} f_i^2\right)}{N^2(2-1)} = \frac{2\left(N^2 - N^2\right)}{N^2} = 0$. Если при исследовании дихотомического признака A $(k = 2)$ абсолютные частоты градаций A_1 и A_2 соответственно равны $f_1 = N/2$; $f_2 = N/2$, то вариация будет максимально возможной. Величина IQV, количественно оценивающая вариацию, равна единице: $IQV = \frac{2\left(N^2 - \sum_{i=1}^{k} f_i^2\right)}{N^2(2-1)} = \frac{2\left(N^2 - \left(\frac{N^2}{4} + \frac{N^2}{4}\right)\right)}{N^2} = 2 - 1 = 1$.

На этом вычислительные операции первого этапа методики можно считать законченными. Эмпирические вероятности $\overline{p}_i$ для градаций номинального признака A, интервальные оценки для вероятностей для этих градаций $P(p_i \in [a_i, b_i]) = \gamma$ и количественная оценка вариации через коэффициент IQV, по сути являются важными элементами формального описания статистических закономерностей, присущих типичности и вариации исследуемого номинального признака. Результаты вычислений в текстах научных работ и публикаций удобно представлять табличном виде. Пример табличного представления статистических закономерностей для дихотомического признака приведен в табл. 1.

Таблица 1. Пример табличного представления статистических закономерностей для дихотомического признака

Градации признака A_i	Абсолютные частоты f_i	Вариации признака IQV	Эмпирические вероятности $\overline{p}_i$	Доверительный интервал ($\gamma = 0{,}95$)	
				нижняя граница	верхняя граница
A_1	f_1	IQV	$\overline{p}_1$	a_1	b_1
A_2	f_2		$\overline{p}_2$	a_2	b_2

Данные, приведенные в табл. 1, количественно выражают статистическое распределение исследуемого номинального признака, полученное на основе анализа результатов однотипных независимых испытаний. Предлагаемая формализация представляет собой законченное индуктивное умозаключение, так как выявление статистических закономерностей проводилось от анализа частных случаев к общему выводу.

Второй этап методики имеет своей целью выявление статистической сопряженности (статистической взаимосвязи) между двумя номинальными признаками A и B, которые измеряются у одной и той же единицы наблюдения исследуемой выборочной совокупности при проведении однотипных независимых испытаний. Число градаций признаков A и B обозначим соответственно через k и m.

С целью упрощения будем рассматривать номинальные дихотомические признаки: $k = 2$ и $m = 2$. При необходимости вычислительные операции по аналогии могут быть распространены и для случая, когда один из признаков или оба признака являются политомическими. Предполагается, что для исследуемых признаков уже реализован первый этап методики и полученные эмпирическим путем статистические распределения для обоих признаков представлены в форме табл. 1.

Результаты экспериментальных измерений удобно представить в виде таблицы сопряженности признаков. Для дихотомических признаков A и B их сопряженность в исследуемой выборочной совокупности может быть представлена в табл. 2. Предложенная табличная форма с учетом адаптации под заданное число градаций номинальных признаков может быть использована для формализации статистической информации как при непосредственном проведении научных исследований, так и для наглядного представления конечных результатов в текстах диссертаций и научных публикациях.

Таблица 2. Таблица сопряжённости дихотомических признаков

Градации признаков	B_1	B_2	Всего по признаку A
A_1	f_{11}	f_{12}	f_{A1}
A_2	f_{21}	f_{22}	f_{A2}
Всего по признаку B	f_{B1}	f_{B2}	N

В таблице 2 использованы следующие обозначения: f_{A1} – число единиц наблюдения, у которых зафиксирована градация признака A_1; f_{A2} – число единиц наблюдения, у которых зафиксирована градация признака A_2; f_{B1} – число единиц наблюдения, у которых зафиксирована градация признака B_1; f_{B2} – число единиц наблюдения, у которых зафиксирована градация признака B_2; f_{11} – число единиц наблюдения, у которых зафиксированы градации признаков A_1 и B_1; f_{12} – число единиц наблюдения, у которых зафиксированы градации признаков A_1 и B_2; f_{21} – число единиц наблюдения, у которых зафиксированы градации признаков A_2 и B_1; f_{22} – число единиц наблюдения, у которых зафиксированы градации признаков A_2 и B_2.

Для выявления статистической сопряженности номинальных признаков необходимо один из признаков рассматривать как факторный (группировочный), а другой – как результативный. Если признак A является группировочным, а признак B – результативным, то в этом случае вариация признака B до группировки вычисляется по формуле: $IQV_B = \dfrac{m\left(N^2 - \sum\limits_{j=1}^{m} f_{Bj}^2\right)}{N^2(m-1)}$. Если группировочным является признак B, а признак A – результативным, то в этом случае вариация признака A до группировки вычисляется по формуле: $IQV_A = \dfrac{k\left(N^2 - \sum\limits_{i=1}^{k} f_{Ai}^2\right)}{N^2(k-1)}$.

Для случая дихотомических признаков $m=2$ и $k=2$ указанные формулы примут вид: $IQV_B = \dfrac{2\left(N^2 - f_{B1}^2 - f_{B2}^2\right)}{N^2}$; $IQV_A = \dfrac{2\left(N^2 - f_{A1}^2 - f_{A2}^2\right)}{N^2}$.

После группировки выборочной совокупности на основе градаций признака A вариацию признака B в образованных группах (внутригрупповую вариацию) можно оценить по формуле: $IQV_{BAi} = \dfrac{m\left(f_{Ai}^2 - \sum\limits_{j=1}^{m} f_{ij}^2\right)}{f_{Ai}^2 \cdot (m-1)}$, $1 \le i \le k$; где индекс BAi указывает на оценку вариации в статистическом распределении признака B, которое соответствует i-ой градации признака A; k – число градаций группировочного признака A.

В результате группировки выборочной совокупности на основе градаций признака B вариацию признака A в образованных группах (внутригрупповую вариацию) можно оценить по формуле: $IQV_{ABj} = \dfrac{k\left(f_{Bj}^2 - \sum\limits_{i=1}^{k} f_{ij}^2\right)}{f_{Bj}^2 \cdot (k-1)}$, $1 \le j \le m$; где индекс ABj указывает на оценку вариации в статистическом распределении признака A, которое соответствует j-й градации признака B; m – число градаций группировочного признака B.

Если признаки A и B дихотомические, как показано в таблице 2, то формулы для оценки внутригрупповых вариаций признака B в группах A_1 и A_2 имеют вид: $IQV_{BA1} = \frac{2\left(f_{A1}^2 - f_{11}^2 - f_{12}^2\right)}{f_{A1}^2}$ – внутригрупповая вариация признака B в группе A_1 с объемом f_{A1}; $IQV_{BA2} = \frac{2\left(f_{A2}^2 - f_{21}^2 - f_{22}^2\right)}{f_{A2}^2}$ – внутригрупповая вариация признака B в группе A_2 с объемом f_{A2}.

Формулы для оценки внутригрупповых вариаций признака A в группах B_1 и B_2 имеют вид: $IQV_{AB1} = \frac{2\left(f_{B1}^2 - f_{11}^2 - f_{21}^2\right)}{f_{B1}^2}$ – внутригрупповая вариация признака A в группе B_1 с объемом f_{B1}; $IQV_{AB2} = \frac{2\left(f_{B2}^2 - f_{12}^2 - f_{22}^2\right)}{f_{B2}^2}$ – внутригрупповая вариация признака A в группе B_2 с объемом f_{B2}.

Средние значения из внутригрупповых вариаций $\overline{IQV}$ вычисляется по следующим формулам: $\overline{IQV_{BA}} = \frac{\sum_{i=1}^{k}\left(f_{Ai} \cdot IQV_{BAi}\right)}{N}$ – среднее значение из внутригрупповых вариаций при группировке признака B по признаку A; $\overline{IQV_{AB}} = \frac{\sum_{j=1}^{m}\left(f_{Bj} \cdot IQV_{ABj}\right)}{N}$ – среднее значение из внутригрупповых вариаций при группировке признака A по признаку B;

Для таблицы 2 формулы для вычисления средних значений из внутригрупповых вариаций имеют вид: $\overline{IQV_{BA}} = \frac{f_{A1} \cdot IQV_{BA1} + f_{A2} \cdot IQV_{BA2}}{N}$; $\overline{IQV_{AB}} = \frac{f_{B1} \cdot IQV_{AB1} + f_{B2} \cdot IQV_{AB2}}{N}$.

Для количественной оценки результатов группировки используется межгрупповая вариация BGV. В случае группировки признака B по признаку A межгрупповая вариация BGV_{BA} вычисляется по следующей формуле: $BGV_{BA} = \frac{\sum_{i=1}^{k}\left(f_{Ai} \cdot \left(IQV_B - IQV_{BAi}\right)\right)}{N}$. При осуществлении группировки признака A по признаку B межгрупповая вариация BGV_{AB} вычисляется по формуле: $BGV_{AB} = \frac{\sum_{j=1}^{m}\left(f_{Bj} \cdot \left(IQV_A - IQV_{ABj}\right)\right)}{N}$.

Для таблицы 2 формулы для вычисления межгрупповой вариации имеют вид: $BGV_{BA} = \frac{f_{A1} \cdot \left(IQV_B - IQV_{BA1}\right) + f_{A2} \cdot \left(IQV_B - IQV_{BA2}\right)}{N}$; $BGV_{AB} = \frac{f_{B1} \cdot \left(IQV_A - IQV_{AB1}\right) + f_{B2} \cdot \left(IQV_A - IQV_{AB2}\right)}{N}$.

Проведение группировки, по своей сути, приводит к расщеплению вариации. Формальная запись этого результата имеет вид: $IQV_B = \overline{IQV_{BA}} + BGV_{BA}$ – при группировке признака B по признаку A; $IQV_A = \overline{IQV_{AB}} + BGV_{AB}$ – при группировке признака A по признаку B.

Вариация признака B до группировки IQV_B равна сумме межгрупповой вариации BGV_{BA} и среднему значению из внутригрупповых дисперсий $\overline{IQV_{BA}}$. Соответственно вариация признака A до группировки IQV_A равна сумме межгрупповой вариации BGV_{AB} и среднему значению из внутригрупповых дисперсий $\overline{IQV_{AB}}$.

Отношение межгрупповой вариации BGV к общей вариации до группировки IQV называется эмпирическим коэффициентом детерминации η^2. Формулы для вычисления имеют следующий вид: $\eta_{BA}^2 = \frac{BGV_{BA}}{IQV_B}$ – при группировке признака B по признаку A; $\eta_{AB}^2 = \frac{BGV_{AB}}{IQV_A}$ – при группировке признака A по признаку B.

Корень из эмпирического коэффициента детерминации называется эмпирическим корреляционным отношением $\eta = \sqrt{\eta^2}$. Этот показатель используется для количественной оценки любого вида статистической связи между номинальными признаками. Для рассматриваемых группировок вычисления производятся по следующим формулам: $\eta_{BA} = \sqrt{\frac{BGV_{BA}}{IQV_B}}$ – при группировке признака B по признаку A; $\eta_{AB} = \sqrt{\frac{BGV_{AB}}{IQV_A}}$ – при группировке признака A по признаку B. Эмпирическое корреляционное отношение η есть величина, лежащая в интервале от нуля до единицы включительно. Чем больше значение η, тем сильнее статистическая связь между номинальными признаками. Для качественной оценки статистической связи можно использовать шкалу Чеддока [1], представленную в табл. 3.

Таблица 3. Шкалу Чеддока для качественной оценки эмпирического корреляционного отношения

Корреляционное отношение η	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-1,0
Характеристика силы связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

На этом второй этап методики статистического анализа можно считать законченным.

Целью третьего этапа методики является проверка гипотезы о значимости эмпирического корреляционного отношения, оценка мощности статистического критерия и формулировка выводов о состоятельности эмпирического корреляционного отношения. Для проверки значимости эмпирического корреляционного отношения формулируются следующие статистические гипотезы:

– гипотеза H_0 – эмпирическое корреляционное отношение η значимо не отличается от нуля, т.е. статистическая связь не является значимой;

– гипотеза H_1 – эмпирическое корреляционное отношение значимо отличается от нуля, т.е. статистическая связь является значимой.

Для проверки статистической гипотезы H_0 необходимо зафиксировать ошибку первого рода (уровень значимости) α и выбрать соответствующий статистический критерий. Уровень значимости, например, можно зафиксировать на уровне 0,05. Для проверки гипотезы H_0 следует использовать критерий Фишера. Расчетное значение статистики критерия F_p вычисляется по формулам: $F_p = \frac{BGV_{BA} \cdot (N-k)}{IQV_{BA} \cdot (k-1)}$ – если группировка производится по признаку A; $F_p = \frac{BGV_{AB} \cdot (N-m)}{IQV_{AB} \cdot (m-1)}$ – если группировка производится по признаку B.

Величина p, отражающая вероятность появления статистики F_p при истинной H_0, вычисляется по формуле: $p = \int_{F_p}^{\infty} F du$, где $F = F(u, df_1, df_2)$ – функция плотности распределения вероятности Фишера со степенями свободы df_1 и df_2; u – переменная величина в функции распределения, по которой производится интегрирование. Формулы для вычисления степеней свободы имеют следующий вид: $df_1 = k-1$ и $df_2 = N-k$ – при группировке признака B по признаку A (признак A – группировочный); $df_1 = m-1$ и $df_2 = N-m$ – при группировке признака A по признаку B (признак B – группировочный).

Для автоматизации вычисления вероятности p можно использовать табличный процессор Microsoft Excel, в котором имеется встроенная статистическая функция FРАСП(F_p; df_1; df_2). Если величина $p > \alpha$, то нет оснований отклонить гипотезу H_0, эмпирическое корреляционное отношение η значимо не отличается от нуля, т.е. статистическая связь не является значимой. Если величина $p \le \alpha$, то есть основание отклонить гипотезу H_0, эмпирическое корреляционное

отношение η значимо отличается от нуля, т.е. статистическая связь является значимой и ее можно классифицировать по шкале Чеддока.

Для оценки состоятельности статистических выводов необходимо вычислить мощность F-критерия Фишера. Для этого используется нецентральное распределение Фишера $F_{нц} = F_{нц}(u, \gamma, df_1, df_2)$, где $\gamma = F_p$ – параметр нецентральности. Мощность $1-\beta$ для F-критерия вычисляется на основании следующего выражения: $1-\beta = \int_V^{\infty} F_{нц}\, du$, где нижний предел интегрирования $V = F_{кр}$. Значение $F_{кр}$ вычисляется в результате решения следующего уравнения: $\alpha = \int_V^{\infty} F\, du$. При заданном значении α, например $\alpha = 0{,}05$, и вычисленных степенях свободы df_1 и df_2 для решения этого уравнения можно использовать статистическую функцию FРАСПОБР(α; df_1; df_2). Для вычисления мощности F-критерия $1-\beta$ целесообразно воспользоваться электронный ресурс Keisan online calculator, находящийся в открытом доступе по электронному адресу https://keisan.casio.com.

На практике приемлемой обычно считается мощность статистического критерия $1-\beta_0$, равная или превышающая 0,8, что соответствует вероятности ошибки второго рода β_0 меньшей или равной 0,2. Следовательно, статистические выводы об эмпирическом корреляционном отношении η можно считать состоятельными, если выполняются два условия: $p \le \alpha$ и $(1-\beta) \ge (1-\beta_0)$. Если $p \le \alpha$, но при этом $(1-\beta) < (1-\beta_0)$, то в этом случае гипотезу H_0 на заданном уровне значимости α можно отклонить, но при этом критерий не обладает требуемой чувствительностью (мощностью). Требуется увеличение объема экспериментальных данных. С другой стороны, если для малой выборки условия $p \le \alpha$ и $(1-\beta) \ge (1-\beta_0)$ выполняются, то это означает, что даже имеющегося малого объема экспериментальных данных вполне достаточно для утверждения о состоятельности статистических выводов.

Пример практического применения методики

Покажем применение описанной выше методики на практическом примере. В спланированных фармакологических исследованиях изучалась выборочная совокупность объемом $N = 26$. В ней исследовались два фармакологических эффекта, которые измеряются политомическими признаками A и B. Каждый признак имеет три градации, т.е. $k = 3$ и $m = 3$. Экспериментальные данные представлены в приведенной ниже таблице сопряженности признаков (табл. 4).

Таблица 4. Пример экспериментальных данных

Градации признаков	B_1	B_2	B_3	Всего по признаку A
A_1	15	1	0	16
A_2	1	3	1	5
A_3	0	2	3	5
Всего по признаку B	16	6	4	26

Эмпирические вероятности для градаций признака A: $\overline{p}_1 = \frac{16}{26} = 0{,}616$; $\overline{p}_2 = \frac{5}{26} = 0{,}192$; $\overline{p}_3 = \frac{5}{26} = 0{,}192$.

Вычислим доверительные интервалы $(\gamma = 0{,}95)$ для эмпирических вероятностей градаций признака A. Учитывая, что $\alpha = 1-\gamma$, вычислим величину $v = 1-\alpha/2 = 1-0{,}05/2 = 0{,}975$. Для градации A_1 вычисляются степени свободы $d_1 = 2(N - f_i + 1) = 2(26-16+1) = 22$; $d_2 = 2 f_i = 2 \cdot 16 = 32$; $d_3 = 2(f_i + 1) = 2(16+1) = 34$; $d_4 = 2(N - f_i) = 2(26-16) = 20$. Далее в программе Microsoft Excel с использованием статистических функций $FРАСПОБР(\alpha/2; d_1; d_2)$ и $FРАСПОБР(\alpha/2; d_3; d_4)$ вычисляются квантили $F_1(d_1; d_2; v)$ и $F_2(d_3; d_4; v)$:

$F_1(d_1;d_2;v) = FРАСПОБР(0{,}025;\ 22;\ 32) = 2{,}13$; $F_1(d_1;d_2;v) = FРАСПОБР(0{,}025;\ 34;\ 20) = 2{,}32$.

Границы доверительного интервала для эмпирической вероятности градации A_1 вычисляются по следующим формулам: $a_1 = \dfrac{f_i}{f_i + (N - f_i + 1)\cdot F_1(d_1;d_2;v)} = 0{,}406$; $b_i = \dfrac{(f_i + 1)\cdot F_2(d_3;d_4;v)}{N - f_i + (f_i + 1)\cdot F_2(d_3;d_4;v)} = 0{,}798$.

Аналогичным образом рассчитываются доверительные интервалы для оставшихся эмпирических вероятностей градаций признака A и эмпирических вероятностей градаций признака B.

Количественные оценки вариации признаков A и B до группировки производится по формулам:

$$IQV_A = \frac{k\left(N^2 - \sum_{i=1}^{k} f_{Ai}^2\right)}{N^2(k-1)} = \frac{1110}{1352} = 0{,}821; \qquad IQV_B = \frac{m\left(N^2 - \sum_{j=1}^{m} f_{Bj}^2\right)}{N^2(m-1)} = \frac{1104}{1352} = 0{,}817.$$

Результаты статистического анализа первого этапа методики представлены в таблицах 5 и 6.

Таблица 5. Результаты статистического анализа первого этапа методики для признака А

Градации признака A_i	Абсолютные частоты f_i	Вариации признака IQV	Эмпирические вероятности $\overline{p}_i$	Доверительный интервал ($\gamma = 0{,}95$)	
				нижняя граница	верхняя граница
A_1	16	0,821	0,616	0,406	0,798
A_2	5		0,192	0,065	0,393
A_3	5		0,192	0,065	0,393

Таблица 6. Результаты статистического анализа первого этапа методики для признака В

Градации признака B_i	Абсолютные частоты f_i	Вариации признака IQV	Эмпирические вероятности $\overline{p}_i$	Доверительный интервал ($\gamma = 0{,}95$)	
				нижняя граница	верхняя граница
B_1	16	0,817	0,616	0,406	0,798
B_2	6		0,231	0,089	0,437
B_3	4		0,153	0,043	0,347

В соответствии со вторым этапом методики произведем вычисление эмпирических корреляционных отношений. Вычислим внутригрупповые вариации в группах A_i:

$$IQV_{BA1} = \frac{3\left(f_{A1}^2 - \sum_{j=1}^{3} f_{1j}^2\right)}{f_{A1}^2\cdot(3-1)} = 0{,}176; \qquad IQV_{BA2} = \frac{3\left(f_{A2}^2 - \sum_{j=1}^{3} f_{2j}^2\right)}{f_{A2}^2\cdot(3-1)} = 0{,}84; \qquad IQV_{BA3} = \frac{3\left(f_{A3}^2 - \sum_{j=1}^{3} f_{3j}^2\right)}{f_{A3}^2\cdot(3-1)} = 0{,}72.$$

Вычислим внутригрупповые вариации в группах B_j.

$$IQV_{AB1} = \frac{3\left(f_{B1}^2 - \sum_{i=1}^{3} f_{i1}^2\right)}{f_{B1}^2\cdot(3-1)} = 0{,}176; \qquad IQV_{AB2} = \frac{3\left(f_{B2}^2 - \sum_{j=1}^{3} f_{i2}^2\right)}{f_{B2}^2\cdot(3-1)} = 0{,}917; \qquad IQV_{AB3} = \frac{3\left(f_{B3}^2 - \sum_{j=1}^{3} f_{i3}^2\right)}{f_{B3}^2\cdot(3-1)} = 0{,}563.$$

Средние из внутригрупповых вариаций: $\overline{IQV_{BA}} = \dfrac{\sum_{i=1}^{3}(f_{Ai}\cdot IQV_{BAi})}{26} = 0{,}4082;$

$$\overline{IQV_{AB}} = \frac{\sum_{j=1}^{3}(f_{Bj}\cdot IQV_{ABj})}{26} = 0{,}4063.$$

Межгрупповые вариации: $BGV_{BA} = \frac{\sum_{i=1}^{3}\left(f_{Ai}\cdot\left(IQV_B - IQV_{BAi}\right)\right)}{N} = 0{,}4084$;

$$BGV_{AB} = \frac{\sum_{j=1}^{3}\left(f_{Bj}\cdot\left(IQV_A - IQV_{ABj}\right)\right)}{N} = 0{,}415.$$

Эмпирические корреляционные отношения: $\eta_{BA} = \sqrt{\frac{BGV_{BA}}{IQV_B}}\sqrt{\frac{0{,}4084}{0{,}817}} = 0{,}707$:

$$\eta_{AB} = \sqrt{\frac{BGV_{AB}}{IQV_A}} = \sqrt{\frac{0{,}418}{0{,}821}} = 0{,}711.$$

По шкале Чеддока зависимость между признаками A и B можно классифицировать как высокую. В соответствии с третьим этапом методики проверим гипотезу о незначимости эмпирических корреляционных отношений при $\alpha = 0{,}05$, вычислим мощность статистического критерия и оценим состоятельность статистических выводов при заданной ошибке второго рода $\beta_0 = 0{,}2$.

Вычислим статистики F_p: $F_p = \frac{BGV_{BA}\cdot(N-k)}{\overline{IQV_{BA}}\cdot(k-1)} = 11.506$; $F_p = \frac{BGV_{AB}\cdot(N-m)}{\overline{IQV_{AB}}\cdot(m-1)} = 11{,}741$.

Вычислим в программе Microsoft Excel вероятности p для проверки значимости величин η_{BA} и η_{AB} соответственно. Для этого используем статистические функции: $p = FPACП(11{,}506;\ 2;\ 23) = 0{,}00034$; $p = FPACП(11{,}741;\ 2;\ 23) = 0{,}00031$. Очевидно, что эмпирические корреляционные отношения $\eta_{BA} = 0{,}707$ и $\eta_{AB} = 0{,}711$ являются значимыми (значимо отличаются от нуля), так как для них соответственно выполняются условия $p = 0{,}00034 \le \alpha = 0{,}05$ и $p = 0{,}00031 \le \alpha = 0{,}05$.

Для $\alpha = 0{,}05$; $df_1 = 2$ и $df_2 = 23$ в Microsoft Excel вычислим значение статистики $F_{кр}$. Для этого воспользуемся функцией $FРАСПОБР(0{,}05;\ 2;\ 23) = 3{,}422$; т.е. $F_{кр} = 3{,}42$ В завершении с использованием электронного ресурса Keisan online calculator вычислим мощности статистического критерия. В результате вычислений получаем значения мощностей $1-\beta = 0{,}818$ и $1-\beta = 0{,}826$ для параметров нецентральности $F_p = 11.506$ и $F_p = 11{,}741$ соответственно. В результате можно обоснованно утверждать, что вычисленные эмпирические корреляционные отношения $\eta_{BA} = 0{,}707$ и $\eta_{AB} = 0{,}711$ являются значимыми, так как в обоих случаях $p \le 0{,}05$ и мощности критерия удовлетворяет условию $1-\beta > 0{,}8$. Следовательно, статистические выводы о статистической взаимосвязи между исследуемыми признаками являются состоятельными.

Заключение

В результате проведенного теоретического исследования описана и обоснована методика статистического анализа номинальных признаков, используемых для измерения фармакологических эффектов. Вычислительные операции могут быть полностью автоматизированы в программе Microsoft Excel. Для вычисления мощности статистического критерия на основе нецентрального F-распределения можно использовать электронный ресурс Keisan online calculator, находящийся в открытом доступе. Это особенно важно для сокращения временных затрат на проведение вычислений.

Разработанный на основе методики программный модуль существенно упрощает работу для специалистов, не имеющих базового математического образования. По своей сути, предложенный в методике алгоритм статистического анализа можно рассматривать как технологию обработки первичной информации с целью получения формализованной информации более высокого порядка, которая выражает индуктивные закономерности.

Методика позволяет представить выявленные индуктивные закономерности для типичности, вариации и статистической взаимосвязи между номинальными признаками как новое знание, полученное в результате проведения научного исследования.

Литература (references)

1. Медик В.А., Токмачев М.С., Фишман Б.Б. Статистика в медицине и биологии: Руководство. В 2-х томах / Под редакцией Ю.М. Комарова. Т. 1. Теоретическая статистика. – М.: Медицина, 2000. – 412 с. [Medik V.A., Tokmachev M.S., Fishman B.B. *Statistika v medicine i biologii: Rukovodstvo. V 2-h tomah / Pod redakciej Ju.M. Komarova. T. 1. Teoreticheskaja statistika.* Statistics in medicine and biology: a Guide. In 2 volumes / Edited by Yu.M. Komarov. V.1. Theoretical statistics. – Moscow: Medicine, 2000. – 412 p. (in Russian)]
2. Евсеев А.В., Сурменёв Д.В., Евсеева М.А. и др. Сравнительный анализ эффективности металлокомплексных и аминотиоловых антигипоксантов в эксперименте // Обзоры по клинической фармакологии и лекарственной терапии. – 2018. – Т.16, №2. – С. 18-24. [Evseev A.V., Surmenjov D.V., Evseeva M.A. i dr. *Obzory po klinicheskoj farmakologii i lekarstvennoj terapii.* Reviews of clinical pharmacology and drug therapy. – 2018. – V.16, №2. – P. 18-24. in Russian)]
3. Сосин Д.В., Евсеев А.В., Шабанов П.Д. Безопасность новых протекторов острой гипоксии // Обзоры по клинической фармакологии и лекарственной терапии. – 2012. – Т.10, №4. – С. 58-64. [Sosin D.V., Evseev A.V., Shabanov P.D. *Obzory po klinicheskoj farmakologii i lekarstvennoj terapii.* Reviews of clinical pharmacology and drug therapy. – 2012. – V.10, №4. – P. 58-64. in Russian)]
4. Evseev A.V., Surmenev D.V., Evseeva M.A. et al. The impact of the new metal-complex (ZnII) selenium-containing compound πQ2721 on the resistance of rats to acute hypoxic hypoxia // Chronicles of Pharmaceutical Science. – 2018. – V.2, Iss.2. – P. 493-501.

Информация об авторах

Лямец Леонид Леонидович – кандидат технических наук, доцент, заведующий кафедрой физики, математики и медицинской информатики ФГБОУ ВО «Смоленский государственный медицинский университет» Минздрава России. E-mail: lll190965@yandex.ru

Евсеев Андрей Викторович – доктор медицинских наук, профессор, заведующий кафедрой нормальной физиологии ФГБОУ ВО «Смоленский государственный медицинский университет» Минздрава России. E-mail: hypoxia@yandex.ru