В.П. Ильин

МЕТОДИЧЕСКИЕ ОСОБЕННОСТИ ПРИМЕНЕНИЯ ДЕСКРИПТИВНОЙ СТАТИСТИКИ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

ФГБУ «Научный центр проблем здоровья семьи и репродукции человека» СО РАМН (Иркутск)

В статье обсуждаются методические особенности применения дескриптивных статистик для переменных в различных шкалах с учетом закона распределения и приводится анализ условий переноса выявленных выборочных закономерностей на генеральную совокупность.

Целью данной работы явилось описание методов и особенностей применения дескриптивной (описательной) внутригрупповой статистики.

Отмечено, что выбор параметров дескриптивной статистики основан на двух принципах: 1) приведение набора описательных характеристик центральной тенденции, меры рассеяния и формы распределения каждой наблюдаемой группы; 2) если установлен нормальный (или связанный с нормальным) закон распределения переменной – приведение доверительных интервалов параметров выборочного распределения.

Сделан вывод о том, что при описании первичных данных необходимо приводить: гистограмму первичных данных, наложенную на график выбранного закона распределения; полное описание наблюдаемых данных в группах с описанием выборочного метода их получения; параметры и критерии, обосновывающие отказ или выбор закона распределения; оценки параметров генеральной совокупности (в случае нормального распределения).

Ключевые слова: дескриптивная статистика, шкала представления признака, закон распределения

METHODICAL PECULIARITIES OF USING DESCRIPTIVE STATISTICS IN MEDICAL-BIOLOGICAL RESEARCHES

V.P. Iljin

Scientific Center of Family Health and Human Reproduction Problems SB RAMS, Irkutsk

The article discusses methodical peculiarities of using descriptive statistics for variable in different scales taking into account partition law and presents the analysis of conditions of transfer of revealed sample regularities to the universe.

The aim of the work was to describe methods and peculiarities of using descriptive intragroup statistics.

It was determined that choice of parameters of descriptive statistics was based on two principles: 1) adduction of set of descriptive characteristics of central trend, value of dispersion and form of distribution in studied group; 2) adduction of confidence intervals of parameters of sample distribution in case of normal (or connected with normal) partition law.

It was concluded that at the description of primary data it is necessary to adduct histogram of primary data imposed on the graph of chosen partition law; full description of observed data in groups with description of sample method of their obtaining; parameters and criteria that prove rejection or choice of partition law; values of parameters of universe (in case of normal distribution).

Key words: descriptive statistics, scale of presentation of the variable, partition law

Применение статистических методов анализа в медико-биологических исследованиях — это не самоцель и не дань моде, а объективная необходимость. Дескриптивная статистика предназначена для описания первичных данных таким образом, чтобы за цифрами не потерять суть, и любому читателю, а не только самому исследователю, можно было бы понять изучаемое явление. Дескриптивное описание способствует выявлению степени достоверности полученных результатов наблюдений и установлению закономерностей изучаемых медицинских явлений. Кроме того, особенно объективные, обоснованные знания нужны для принятия верного, доказательного решения в условиях стохастических процессов, к которым, безусловно, относятся медико-биологические исследования. Выбор закона распределения, оценка адекватности первичным данным (модели) служит обязательным этапом дескриптивного анализа, являющегося обоснованием правомерности последующего обобщения, переноса и прогнозирования выявленных

закономерностей с выборочных исследований на генеральную совокупность.

Медицинские исследовательские работы, как правило, состоят из нескольких задач, две из которых являются типовыми с точки зрения статистики. Одна из задач — это внутригрупповая дескриптивная статистика. Другая задача — межгрупповые сравнения. Целью данной работы является описание методов и особенностей применения дескриптивной (описательной) внутригрупповой статистики. Задача сравнительного межгруппового анализа, направленного на проверку их однородности, выявление признаков, по которым группы различаются, с последующей трактовкой выявленных различий и переноса установленных особенностей и закономерностей на генеральную совокупность, изложена в работах [5, 6]. Выбор методов дескриптивной статистики и критериев согласия законам распределения изучаемых переменных в группах осуществляется с учетом типа шкалы.

Тип шкалы первичен в выборе статистического метода. Наиболее известная классификация шкал, предложенная С.С. Стивенсоном в 1959 году, состоит из четырех типов и основана на возможности проведения арифметических операций сложения (вычитания), умножения (деления) и операции сравнения (больше, меньше) с данными, представленными в этих шкалах (Закс Л., 1976; Аффифи А., Эйзен Э., 1982; Боровиков В., 2001).

Шкала наименований используется для отнесения единицы наблюдения к одной из групп. Каждая группа имеет свое уникальное обозначение, не совпадающее ни с каким другим. Переменные, представленные в шкале наименований, называются номинальными. Арифметические операции (сложение – вычитание, умножение – деление) и операция сравнения или отношения (больше или меньше) не имеют смысла в применении к данным шкалы наименований. Назначение этой шкалы заключается в проведении качественной классификации (отнесении в группу) без выполнения каких-либо арифметических операций и операций сравнения. Характеристикой является абсолютное число наблюдений или их процент вхождения в группу. Статистика центральной тенденции – мода. Примерами могут служить классификация по полу (мужчины, женщины), состоянию (болен, здоров). В качестве кодов можно использовать цифры: 1 – мужчина, 2 – женщина, либо буквы: М – мужчина, Ж – женщина, либо любой другой набор символов. Например, в группе из 50 человек 20 мужчин и 30 женщин, что составляет: 40 % мужчин и 60 % женщин. Арифметические операции с кодами 1, 2 или их мнемоническими обозначениями «М», «Ж» смысла не имеют. Их нельзя складывать, умножать, делить, вычислять среднее, дисперсию. Нет смысла сравнивать 1 и 2 (кто больше, кто меньше). Номинальные переменные используются только для качественной классификации. Статистической мерой центральной тенденции является мода, но не среднее и не медиана. Такие характеристики, как среднее значение в группе, ошибка среднего, дисперсия, не имеют смысла для такого типа данных. Операции можно выполнять с частотами наблюдаемого признака (кодами), их процентами и сравнивать с аналогичными показателями в других группах. Для шкалы наименований используются только статистические методы неупорядоченных классов, такие, как выводы относительно биномиального распределения (z-критерий), критерий² как мера связности и другие.

Порядковая шкала предназначена для объединения единиц наблюдения в группы. Для данных этой шкалы определены операции сравнения или отношения (больше, меньше), но не имеют смысла арифметические операции. Обозначения классам (группам) для удобства использования присваиваются таким образом, чтобы можно было выделить порядок классов. Если это буквы, то удобно их располагать в порядке алфавита – А, Б, В, Г, ...; если цифры, то по возрастанию – 1, 2, 3, ... либо убыванию – ..., 3, 2, 1; если слова – то по

смыслу слов можно допустить, что тяжелое течение болезни «хуже», чем среднетяжелое, а очень тяжелое – «еще хуже» и т.д. В этом случае можно сказать, у какого пациента течение болезни тяжелее, а у кого легче, но нельзя сказать, насколько (или во сколько раз) хуже или лучше. Статистикой центральной тенденции для группы данных порядковой шкалы являются мода и медиана. В этой шкале данными допустимыми операциями являются операции сравнения или отношения. Переменные, представленные в порядковой шкале, называются порядковыми или ранговыми. Статистические характеристики, такие, как среднее значение в группе, ошибка среднего, дисперсия, не имеют смысла для такого типа данных. Для порядковой шкалы применяются методы, основанные на понятии рангов и относящиеся к непараметрическим статистиками.

Вышеприведенные две шкалы (наименований и порядковые) предназначены для описания качественных переменных.

Описание количественных переменных осуществляется в метрических шкалах, представленных интервальной и шкалой отношений.

Интервальная шкала предназначена не только для количественного сравнения, но и для оценки степени различий между единицами наблюдения, она позволяет и упорядочивать представленные в исследовании данные, и численно выражать различия между ними. С данными, представленными в этой шкале, допустимы операции сложения и вычитания, которые позволяют оценить, «на сколько» данные различаются. Но очень сомнительным является результат сравнения «во сколько раз различаются наблюдаемые данные». Различия между данными, представленными в интервальной шкале, наглядно демонстрируются классическим примером температуры, измеряющихся в шкалах Фаренгейта и Цельсия, различающихся на 273 градуса. Так, два значения в шкале Цельсия 10° и 20° различаются на 10°. Они же будут различаться на 10° и в шкале Фаренгейта. Однако, различие в 2 раза этих данных в шкале Цельсия вызывает большие сомнения в шкале Фаренгейта. Значение 283° (10° + 273°) уже не превосходит в 2 раза значение 293 (20 + 273). Таким образом, интервальные шкалы допускают операцию сложения и не всегда обладают свойством, описываемым операцией умножения.

В шкале отношений с данными допустимы все логические и арифметические операции, в том числе и преобразование вида $x_1 = ax + b$, не изменяющее структуру шкалы. В дополнение всех свойств интервальной шкалы в шкале отношений добавляется возможность утверждения, во сколько раз одно значение больше или меньше другого. Все арифметические операции с данными этой шкалы имеют смысл. Статистики – среднее, медиана и мода – в метрической шкале являются мерой центральной тенденции.

Шкалы отношений и интервальная отличаются между собой наличием абсолютного начала координат, не оказывая принципиального влияния на

виды статистических характеристик. Отметим, что некоторые исследователи не делают тонкого различия между свойствами интервальной и шкалы отношений и объединяют эти две шкалы в одну, называя её метрической (Боровиков В., 2001). Для шкалы отношений и интервальной применяются практически все статистические методы.

Для полноты классификации заметим, что переменные делятся на непрерывные и дискретные величины. Поименованные и порядковые величины всегда дискретны, а интервальные и отношений, представляющие метрические переменные, могут быть как дискретными, так и непрерывными.

Таким образом, по информативности шкалы можно расположить в порядке возрастания в следующей последовательности: самая «бедная» — шкала наименований, «богаче» — порядковая шкала, самая «богатая» — метрическая шкала (интервальная и отношений). Перевести данные перекодировкой можно только в одном направлении информативности от более «богатой» шкалы к менее «богатой» с потерей некоторой степени информативности. Например, переменная «концентрация пролактина» из метрической шкалы может быть переведена в порядковую с градациями: низкая концентрация (ниже нормы), средняя концентрация (норма), выше средней (выше нормы). Либо можно перевести ее в бинарную шкалу наименований: норма, патология. А обратный перевод уже невозможен.

Итак, различные типы шкал в представлении исходных данных определяют возможность применения различных статистических характеристик и статистических критериев, корректных для описания изучаемого медико-биологического явления.

Дескриптивный анализ начинается с описательного представления первичных наблюдаемых данных и, как правило, с таблиц, графического анализа и представления результатов анализа. Зачастую анализ начинается с того, что изучается вопрос: как часто встречаются те или иные значения изучаемой переменной, являются ли они типичными, располагаются ли в интервалах средних значений или попадают в интервалы минимальных, максимальных значений? Сколько «горбов» в группе, выделяются ли подгруппы и сколько их и т.д. Наиболее наглядно и доступно на эти вопросы отвечают графические методы. Для этого строятся графики, диаграммы, гистограммы и таблицы распределения наблюдаемых частот. Нередко они являются основой для получения ценных содержательных выводов исследования.

Обычно в ходе исследования интересующий исследователя признак измеряется не у одного — двух, а у множества испытуемых (единиц наблюдения). Кроме того, каждая единица наблюдения характеризуется не одним, а целым рядом признаков, измеренных в разных шкалах. Одни признаки представлены в номинативной шкале и указывают на принадлежность испытуемых к той или иной группе (пол, диагноз заболевания, группа наблюдения или контроля и т.д.). Другие признаки могут

быть представлены в порядковой или метрической шкале (артериальное давление — в норме или выше нормы; концентрация пролактина).

Если признак принимает несколько возможных значений, по количеству удобных для просмотра (на практике — до 10–15), тогда есть смысл рассматривать таблицу распределения частот, которая показывает частоту встречаемости каждого значения признака в абсолютных и относительных значениях. При большем числе значений признака предпочтительнее представлять распределение частот в виде гистограммы. Абсолютные значения указывают, сколько раз встречается каждое значение признака, и представляют все абсолютные значения частот распределения. Если указывается доля наблюдений, приходящихся на то или иное значение признака, то говорят об относительных частотах распределения.

Гистограмма распределения частот — это столбиковая диаграмма, каждый столбец которой опирается на конкретное значение признака или значение интервала для сгруппированных частот. Высота столбика пропорциональна частоте встречаемости соответствующего значения. Гистограмма накопленных (кумулятивных) частот отличается от гистограммы распределения тем, что высота каждого столбика пропорциональна частоте, накопленной к данному значению (интервалу).

Приведенные два способа представления данных служат основным следующим целям:

1. Обнаружение грубых ошибок.
2. Выявление и локализация «выбросов».
3. Построение эмпирического распределения исходных данных.
4. Представления результатов анализа.

Так, если на гистограмме появился новый код, то это означает ошибку при кодировании. Это тип ошибки наиболее простой, его легко можно увидеть. Немного сложнее ситуация для выявления ошибок заключается в локализации «выбросов». Выбросы — это не только ошибка нового кодирования, но и приписывание известному коду большего числа наблюдений, в то время как таких наблюдений на самом деле нет. Определяется этот тип ошибки нарушением на гистограмме числа относительных частот или преобладанием относительной частоты некоторого значения. Такое ассиметричное распределение может означать ошибочное кодирование. Более сложная ситуация заключается в разрыве значений гистограммы и наличии как минимум двух мод («горбов») в распределении, по сути делящих группу на две подгруппы по числу преобладаний наблюдаемых значений. На гистограмме будет четко видно разделение на две группы наблюдаемых значений переменной. В этом случае решение о логической (выделении новой группы или подгруппы) или технической ошибке (ошибочном представлении данных) целиком лежит на исследователе и зависит от цели и задач проводимого исследования.

Увидеть специфику первичных данных лучше всего с помощью графического представления на

плоскости. Наглядное представление позволяет выявить значения переменных редко, часто, наиболее часто встречающихся в различных интервалах группировки, вокруг каких средних значений данные располагаются, насколько они различаются, какие минимальные и максимальные значения принимают, какой наглядный вид распределения имеют наблюдаемые данные в этих интервалах, являются ли симметричными и сколько «горбов» они имеют.

Гистограмма дает качественное представление о распределении, которое не может быть выражено в полном объеме каким-либо числовым показателем. Важным понятием при построении гистограмм является группировка. В сущности сгруппированные данные представляют собой «дважды классифицируемые данные»: один раз — при измерении, второй раз — при обработке. Объединение нескольких значений по определенному правилу в одну группу с последующим присвоением членом группы одного значения — это и есть группировка значений. Чем шире размер группы, тем «грубее» представление первичных данных или меньше точность измерения изучаемого свойства. Для непрерывных переменных группировка заключается в задании интервалов, а для категориальных переменных — категорий, на которых, как на основаниях, строятся столбцы, описывающие частоты. Группировки могут быть равномерными, неравномерными и «пользовательскими», т.е. на усмотрение исследователя, задаваемыми логическими операторами. Выбор числа групп — это тонкое методическое понятие, не входит в данную работу. Приведем лишь известную формулу Стерджеса (Юнкеров В.И., Григорьев С.Г., 2002), рекомендующую формировать число интервалов m по известному n — объему выборки наблюдений: $m = 1 + 3,32 \times \lg n$.

По данным выборки наблюдений строятся гистограммы, кумулятивные линии распределения, по которым зрительно судят о типе распределения и соответствии наблюдаемого эмпирического распределения определенному теоретическому. Основными законами распределения, используемыми в медико-биологических исследованиях, являются: нормальное, логнормальное, логистическое, бета, экспоненциальное, Лапласа, Пуассона, Релея, Вейбулла. Основными критериями оценки нормальности служат критерии Колмогорова — Смирнова, Лилиефорса, Шапиро — Уилка [1, 2, 4, 8, 9, 10].

ГИСТОГРАММЫ И ДИАГРАММЫ В ДЕСКРИПТИВНОМ АНАЛИЗЕ

Кроме визуального изучения характера и закона распределения, важную роль гистограммы и диаграммы играют при выявлении нарушений линейности и формы зависимости между наблюдаемыми признаками, что служит существенной подсказкой в выборе преобразования при «линеаризации» зависимостей и «нормализации» распределений.

Л. Закс (1972) рекомендует следующий подход при преобразованиях. Асимметричные распреде-

ления достаточно часто преобразуются логарифмированием. Неоднородность размаха выборки несколько уменьшается преобразованием «извлечение квадратного корня», в большей степени в сочетании с логарифмированием. Вычисление обратных значений увеличивает маленькие размахи выборки, преобразование вида $1 / \sqrt{x}$ несколько меньше увеличивает размахи. Возведение в квадрат бывает полезно при увеличении масштаба для значений больше 1, при значениях меньше 1 масштаб уменьшается. Тригонометрическое преобразование нормализует распределение, смещенное вправо. При сравнении двух и более групп необходимо выбирать те преобразования, которые нормализуют форму распределений и приемлемы для всех изучаемых групп.

Одно из основных предназначений дескриптивной статистики заключается в выявлении закона распределения переменных. В зависимости от закона распределения производится «модельное» описание данных набором их параметров. Так, в случае распределения Пуассона — одним параметром, в случае нормального распределения — двумя параметрами (средним и дисперсией), в непараметрическом — набором рангов. Выбор модели первичных данных в дальнейшем позволяет заменить первичные данные на их параметры распределения, изучать, анализировать свойства группы и сравнивать уже не первичные данные, а их параметры с аналогичными параметрами других групп.

Методически важной особенностью дескриптивного анализа является обоснование возможности переноса выявленных закономерностей распределения переменных наблюдаемых выборочных группах на генеральную совокупность, если таковая доказана. Это становится возможным только при доказательстве, что изучаемые переменные имеют распределение нормальное или с ним связанное.

Иными словами методы дескриптивной статистики позволяют наиболее полно описать наблюдаемые данные так, чтобы читатель работы мог составить полное представление об объекте исследования, и далее убедить читателей в корректном выборе одной из множества моделей первичных данных для последующего статистического анализа. Модельная замена первичных данных заключается в наглядном, доступном и сжатом представлении первичных данных, т.е. редукции данных, а именно в оптимальной замене многочисленных наблюдаемых значений переменной (изучаемого свойства) ограниченным набором чисел — статистик (средним значением выборки, медианой, модой, дисперсией, квартилями, минимальным, максимальным значением в группе и др.), позволяющим изучать, анализировать свойства группы по полученным параметрам и в дальнейшем проводить сравнения с аналогичными параметрами других групп. Смысл этих статистик заключается в том, что вместо изучения большого количества реальных индивидуальных значений

наблюдаемого признака в выборке численностью несколько десятков, сотен, а иногда и тысяч единиц наблюдения, изучать их количественные характеристики. В этом случае происходит изучение, например, состояния гормонального статуса не у отдельного пациента, а у группы лиц, т.е. объектом исследования уже является не индивидум, а группа. В ней изучается диапазон изменения показателей, частота встречаемости отдельных значений, минимальных, максимальных, входящих в интервал нормы или выходящих за ее границы и т.д. Например, врач на приеме изучает состояние конкретного пациента. Объектом исследования врача являются показатели разнообразных систем организма пациента. Объектом исследования при статистическом изучении являются группы пациентов при их многообразии состояния различных систем организма. Целью дескриптивного анализа в этом случае является описание всего множества состояний у группы пациентов, как часто встречающихся, так и редких, с указанием их вклада. Описание надо приводить так, чтобы было понятно и доказательно, сколько каких состояний в группе наблюдается. Приведение тех или иных статистик определяется законом распределения и типом шкалы, в которой замерен наблюдаемый признак (изучаемое свойство).

Если закон распределения — нормальный, то замена первичных данных в дальнейшем осуществляется на «колоколообразную кривую» нормального распределения со значениями среднего и дисперсии. Если нормальный тип распределения не принимается, то необходимо обосновать, по какой причине, при нарушениях какой характеристики и по значению какого критерия.

Основными характеристиками распределения являются: меры положения (центральной тенденции), меры рассеяния (изменчивости) и меры формы.

К характеристикам положения или центральной тенденции одномерного распределения относят: среднее арифметическое, среднее геометрическое, среднее гармоническое, медиана, мода, интердециальная ширина.

Среднее арифметическое — это сумма значений переменной в группе наблюдений, деленная на число значений. Показывает центральное положение наблюдаемой переменной и является тем значением, сумма отклонений от которого в группе равна нулю. Это значение — единственное, обладающее этим свойством. Отметим, что сумма квадратов расстояний между всеми наблюдаемыми значениями в группе и их средним арифметическим значением является минимальной. Для любого другого значения (если оно не совпадет со средним значением) в группе этого свойства нет, расстояние будет больше, но никак не меньше. Среднее является характеристикой центральной тенденции для переменных количественного типа.

Например, если систолическое артериальное давление, трижды представленное, составило 120, 130 и 140 единиц, то среднее равно $(120 + 130 +$

$140) / 3 = 130$. Сумма отклонений: $(130 - 120) + (130 - 130) + (130 - 140) = 10 + 0 - 10 = 0$.

Медиана — это значение, которое делит группу на две равные части по числу единиц наблюдения. Если число единиц наблюдений четно, то медиана — среднее значение двух средних значений в группе. Заметим, что для вычисления медианы необходимо предварительно упорядочить наблюдаемые значения по возрастанию или убыванию. Совпадение среднего и медианы является необходимым условием для симметрии распределения переменной, при различающихся значениях среднего и медианы — симметрия отсутствует. Как правило, медиана является характеристикой центральной тенденции для переменных в порядковой шкале.

В примере с трижды представленным систолическим артериальным давлением медиана равна 130. Если бы замеров было четное число, скажем, 4 (120, 130, 140, 150), то медиана была бы равна $(130 + 140) / 2 = 135$.

Мода — это значение переменной в группе наблюдений с наибольшей частотой встречаемости. Значению моды соответствует максимальное значение подъема на графике распределения частот. Если мода единственная, то распределение — унимодальное, иначе — полимодальное. Бимодальное распределение на графике имеет две выраженные вершины. Как правило, мода является характеристикой центральной тенденции для переменных, представленных в номинальной или порядковой шкалах. Нормально распределенные величины — унимодальны, т.е. с одной вершиной.

Итак, для номинативных данных единственной мерой центральной тенденции является мода. Для метрических переменных мерой центральной тенденции являются мода, медиана и среднее, которые в случае нормального распределения, обладающего свойством унимодальности и симметричности, совпадают.

По степени расхождения между мерами центральной тенденции можно судить о степени нарушения симметрии распределения: чем больше нарушения симметрии, тем больше различаются между собой меры центральной тенденции.

Отметим различную роль «выскакивающих значений» в мерах центральной тенденции. Если среди наблюдаемых значений имеется некоторое большое «выскакивающее значение», то величина моды или медианы не изменится. Среднее значение в это время изменится на величину v_3 / n , где: v_3 — «выскакивающее значение», n — число единиц наблюдений в группе. Величина среднего ощутимее изменится при малых объемах выборки, менее ощутимо изменится при больших объемах выборки. Все это по-разному характеризует меры центральной тенденции: существенную чувствительность среднего к «выскакивающим значениям» в малых выборках и нечувствительность моды и медианы к появлению резких изменений в значениях («выскакивающих значений»).

Квантили распределения — это точки на числовой оси, которые делят ось изучаемой переменной

(свойства) на упорядоченные группы с заданным соотношением численности. Медиана — это квантиль, делящая на две равные группы все множество наблюдений переменной. Квартили — это три значения переменной, делящие упорядоченное по возрастанию (убыванию) множество наблюдений на четыре равные по численности группы. Первый квартиль обозначается P_{25} и содержит первые 25 % единиц наблюдений переменной (свойства), второй квартиль обозначается P_{50} и содержит первые 50 % единиц наблюдений и соответствует медиане, третий квартиль обозначается P_{75} и содержит 75 % наблюдений, отделяющий первые 75 % от оставшихся 25 % наблюдений.

Мерами выраженности различий (изменчивости, вариации) изучаемой переменной (свойства) являются следующие количественные показатели.

Дисперсия — это мера изменчивости изучаемой переменной, вычисляемая как сумма квадратов отклонений от среднего значения в группе изменений, деленная на число единиц наблюдения в группе, т.е. усредненное значение величины всех отклонений. Обозначается как σ^2 и равна сумме квадратов отклонений от среднего арифметического значения, деленной на n — число единиц наблюдений в группе. Различают теоретическую дисперсию (в генеральной совокупности) и выборочную дисперсию (реально наблюдаемую в группе). Выборочная дисперсия в вычислении отличается от генеральной дисперсии значением знаменателя, уменьшенного на единицу. При нормальном распределении изучаемой переменной (свойства) в группе выборочная дисперсия является оценкой теоретической дисперсии. Отметим, что значение дисперсии имеет смысл только для метрических величин, всегда величина положительная и по масштабу значений выражается в квадратах исходных единиц значений.

Стандартное отклонение — это мера изменчивости метрических величин (свойств) в масштабе истинных значений исходных данных, равная положительному значению корня квадратного из дисперсии (теоретической и выборочной). Обозначается как σ .

Стандартная ошибка среднего — это мера изменчивости, выраженная как стандартное отклонение, деленное на квадратный корень из n — числа единиц наблюдений в группе. Обозначается: m .

Коэффициент вариации — это мера изменчивости, выраженная как стандартное отклонение, деленное на среднее арифметическое, т.е. суммарная изменчивость, соотношенная к положению среднего значения в группе.

Минимум (minimum) — это минимальное значение из числа значений, реально наблюдаемых в группе. Обозначается как $x_{\min}$.

Максимум (maximum) — это максимальное реально наблюдаемое значение в группе. Обозначается как $x_{\max}$.

Размах (Range) — это наиболее наглядная и простейшая мера изменчивости изучаемой пере-

менной (свойства), которая показывает диапазон изменений, равный разности максимального ($x_{\max}$) и минимального ($x_{\min}$) значений в группе. Обозначается: $R = x_{\max} - x_{\min}$.

МЕРЫ ФОРМЫ

Количественными мерами выраженности отклонения графика распределения частот по горизонтали (нарушения симметричного распределения относительно среднего значения) является показатель асимметрии, а по вертикали (формы вершины) — показатель эксцесса.

Асимметрия (Skewness) — это количественная мера, оценивающая степень отклонения графика от симметричного вида распределения частот относительно среднего значения. Для распределения симметричного вида асимметрия равна нулю. При преобладании значений меньше среднего наблюдается левосторонняя асимметрия, при преобладании значений больше среднего — правосторонняя асимметрия. Для корректного заключения большинство исследователей [1, 2, 3, 5, 8, 9, 10] рекомендуют исходные данные преобразовать в стандартизованные, так называемые z-значения. В этом случае данные с различными масштабами и единицами измерения становятся сравнимыми между собой.

Эксцесс (Kurtosis) — это количественная мера, оценивающая степень отклонения формы графика по виду вершины (плосковершинное или островершинное). Для преобразованных к стандартизованному виду в z-значениях островершинное распределение характеризуется положительным эксцессом ($E_x > 0$), плосковершинное — отрицательным эксцессом ($0 > E_x > -3$) и нормальное (средневершинное) распределение — нулевым ($E_x = 0$).

Напомним, что стандартизация заключается в преобразовании всех наблюдаемых значений в группе по формуле: $z_i = (x_i - M_x) / \sigma$. Смысл этого преобразования заключается в переводе наблюдаемых значений в группе в новую шкалу с нулевым средним значением и единичной дисперсией. В этой шкале все значения будут представлены в единицах стандартного отклонения от среднего значения в группе. Иногда бывает неудобно рассматривать положительные и отрицательные значения, которые будут присутствовать после преобразования данных. Чтобы избежать этого неудобства достаточно перейти в новую шкалу при помощи линейного преобразования так, чтобы все значения находились на положительной оси.

Важным является методический вопрос о том, какие статистики приводить при описании полученных данных в научных работах (статьях, диссертациях и т.д.).

Ответ на поставленный вопрос связан с фундаментальной статистической проблемой описания генеральной совокупности по выборочным исследованиям. Приведем одно обоснование. Выявленные закономерности в выборочном

исследовании могут быть перенесены на генеральную совокупность в случае нормального распределения (или связанного с нормальным) изучаемой величины. При этом необходимо указывать доверительные интервалы, вычисляемые по значениям среднего, ошибки среднего и объема выборки [10].

Описание состояния объекта исследования в научной работе предполагает их различную полноту описания в зависимости от типов данных (метрические, порядковые, номинативные) и их законов распределения. Так, традиционное приведение среднего и дисперсии оправдано только при наличии нормального распределения изучаемой переменной, т.к. это суть параметры теоретического нормального распределения данной величины, и совсем не оправдано для переменных, имеющих иной тип распределения.

Выбор параметров дескриптивной статистики для представления в научной работе основан на следующих принципах:

1. Привести набор описательных характеристик центральной тенденции, меры рассеяния и формы распределения каждой наблюдаемой группы.

2. Если установлен нормальный (или связанный с нормальным) закон распределения переменной — то привести доверительные интервалы параметров выборочного распределения, являющиеся оценками соответствующих параметров генеральной совокупности. Для непараметрических переменных эти оценки приводить не следует.

Конечно же, обязательным параметром является численность выборки, так как, по принципу Фишера, «отклонения от нормального вида, если они не слишком заметны, можно обнаружить лишь для больших выборок».

Представление данных будет различно для переменных с параметрическим законом распределения (нормальным, логнормальным, Пуассона и т.д.) и свободным от параметрического типа распределения (ранговым). В случае с параметрическим распределением рекомендуется в зависимости от типа шкалы приводить статистические параметры, описывающие закон распределения каждой переменной с указанием меры центральной тенденции, изменчивости и формы распределения. В силу того, что нормальное распределение служит теоретическим обоснованием переноса закономерностей с выборки на генеральную совокупность, рекомендуется приводить оценки выборочных параметров распределения с указанием их доверительного интервала, позволяющих с 95% вероятностью судить о соответствующих параметрах генеральной совокупности. Вечный вопрос о том, что приводить в статье: дисперсию или ошибку среднего — решается очень просто. Ошибка среднего — это усредненное стандартное отклонение, получающееся арифметическим пересчетом. Логически дисперсия является характеристикой суммарных различий вариабельности наблюдаемой выборки, а ошибка среднего описывает удельную дисперсию,

приходящуюся на единицу наблюдения, и служит основой для оценки доверительного интервала соответствующего параметра (среднего) генеральной совокупности и используется при переносе на генеральную совокупность, а дисперсия — для характеристики групповых свойств первичных данных.

Особое место в научной работе отводится обоснованию закона распределения. Если это нормальное распределение, то следует приводить доказательство унимодальности (наличие одного горба), симметрии, характеристики формы, а также значение выбранного критерия согласия и уровень принятия гипотезы. Важно привести описание «на хвостах» распределения, т.к. зачастую различия между группами обусловлены этими различиями «на хвостах». Это связано с тем, что теоретическое распределение приводится для бесконечной по численности совокупности в отличие от реальной выборки с конечным числом наблюдений, представленного гистограммой. Если распределение не нормальное, то необходимо приводить обоснование выбора адекватного непараметрического описания и выявлять, нарушение какого условия привело к непринятию нормального распределения (нарушения меры центральной тенденции, изменчивости или формы распределения).

Итак, при описании первичных данных рекомендуется приводить:

- 1) гистограмму первичных данных, наложенную на график выбранного закона распределения;
- 2) полное описание (меры положения, центральной тенденции, меры рассеяния, изменчивости и меры формы распределения) наблюдаемых данных в группах с описанием выборочного метода их получения;
- 3) параметры и критерии, обосновывающие отказ или выбор закона распределения;
- 4) оценки параметров генеральной совокупности (в случае нормального распределения).

ЛИТЕРАТУРА

1. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ / Пер. с англ. — М.: Мир, 1982. — 488 с.
2. Боровиков В. Statistica: искусство анализа данных на компьютере. Для профессионалов. — СПб.: Питер, 2001. — 656 с.
3. Гланц С. Медико-биологическая статистика / Пер. с англ. — М.: Практика, 1998. — 459 с.
4. Закс Л. Статистическое оценивание / Пер. с нем. В.Н. Варыгина; под ред. Ю.П. Адлера, В.Г. Горского. — М.: Статистика, 1976. — 598 с.
5. Ильин В.П. Методические особенности применения статистических непараметрических методов в анализе медико-биологических данных // Бюл. ВСНЦ СО РАМН. — 2011. — № 5 (81). — С. 160 — 164.
6. Ильин В.П. Методические особенности применения Т-критерия Стьюдента в медико-биологических исследованиях // Бюл. ВСНЦ СО РАМН. — 2011. — № 5 (81). — С. 157 — 160.

7. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия / Пер. с англ. Ю.Н. Благовещенского. — М.: Финансы и статистика, 1982. — 317 с.

8. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. — М., Медиасфера, 2002. — 312 с.

9. Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований. — СПб.: ВмедА, 2002. — 266 с.

10. Халафян А.А. Statistica 6. Статистический анализ данных. 3-е изд. — М.: ООО «Бином-Пресс», 2008. — 512 с.

Сведения об авторах

Ильин Владимир Петрович – доктор биологических наук, профессор, ведущий научный сотрудник Научного центра проблем здоровья семьи и репродукции человека СО РАМН (664003, г. Иркутск, ул. Тимирязева, 16; тел.: 8 (3952) 20-74-05)