В ПОМОЩЬ ИССЛЕДОВАТЕЛЮ

СОВРЕМЕННЫЕ ПРАВИЛА ИСПОЛЬЗОВАНИЯ МЕТОДОВ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ В МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

УДК: 614.1:311
03.01.02 — биофизика
Поступила 26.04.2020 г.

А. П. Баврина
ФГБОУ ВО «Приволжский исследовательский медицинский университет» Минздрава России, Нижний Новгород

В статье рассматриваются основные приемы описательной статистики. Особое внимание уделяется предварительному определению принадлежности совокупности данных нормальному распределению. Дается краткий обзор общепринятых описательных статистик, подробно рассмотрены новейшие наглядные методы описания данных. Разбираются правила описания всех видов медико-биологических данных: количественных (симметрично и несимметрично распределенных дискретных и непрерывных) и качественных (номинальных и порядковых) данных. Дан простой и удобный алгоритм, который поможет правильно подойти к описанию результатов исследования.

Ключевые слова: описательная статистика; виды данных; распределение.

MODERN RULES FOR THE USE OF DESCRIPTIVE STATISTICS METHODS IN BIOMEDICAL RESEARCH

A. P. Bavrina
Privolzhsky Research Medical University, Nizhny Novgorod

The article discusses the basic techniques of descriptive statistics. Particular attention is paid to the preliminary determination of whether a population of data belongs to a normal distribution. A brief overview of generally accepted descriptive statistics is given, the latest visual methods for describing data are discussed in detail. The rules for describing all types of biomedical data are analyzed: quantitative (symmetrically and asymmetrically distributed discrete and continuous data) and qualitative (nominal and ordinal) data. A simple and convenient algorithm is given that can help to provide a correct approach to the description of the research results.

Key words: descriptive statistics; types of data; distribution.

ВВЕДЕНИЕ

Настоящей статьей журнал «Медицинский альманах» открывает серию публикаций по применению медико-биологической статистики в научных исследованиях. В рамках этой серии будут рассмотрены принципы и правила представления медико-биологических данных в научных работах и сформировано целостное представление о статистической обработке результатов, базирующееся на принципе от простого к сложному.

Ни для кого не секрет, что в настоящее время к статистической обработке медико-биологических данных предъявляются все более серьезные требования при представлении их в научные издания. В прошлом анализ статистического наполнения научно-исследовательских работ полностью отдавался на откуп научным рецензентам. Однако сейчас в большинстве научных изданий существует специальная служба рецензентов, специализирующихся только на статистике. К статистической обработке данных диссертационных исследований также должны соответствовать более строгим требованиям. Кроме того, статистическая наука не стоит на месте, и правила представления данных, актуальные 20 лет назад, могут оказаться совершенно устаревшими.

Но прежде чем заниматься аналитической статистикой, которой авторы научных работ уделяют особо

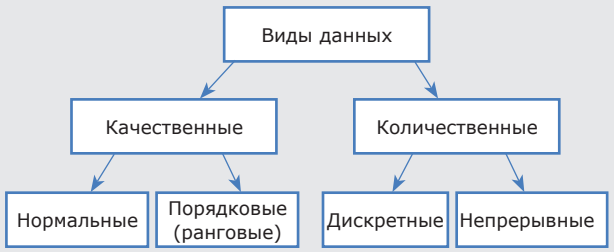


Рис. 1. Виды данных

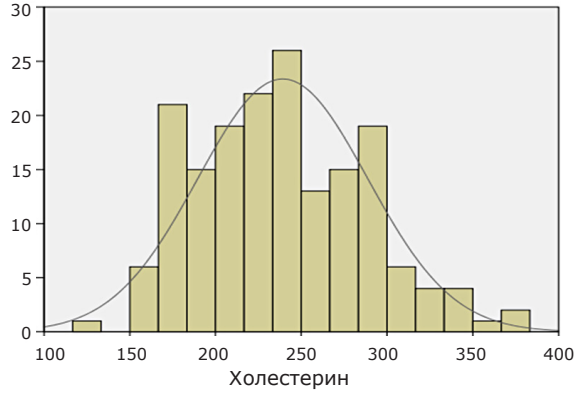


Рис. 2. Графическое представление нормально распределенной совокупности данных

пристальное внимание, следует обратиться к описательной статистике. Первая статья будет посвящена именно *описательной статистике* — важному разделу статистической науки, который занимается описанием и представлением анализируемых данных. От корректного описания и правильного представления полученных результатов по большому счету зависит успех научной публикации.

Чтобы корректно описать данные, необходимо определиться, к какому типу они относятся, поскольку для каждого типа данных существуют свои правила описания [1]. Данные подразделяются на два больших типа: количественные и качественные, которые в свою очередь делятся на подтипы (рис. 1).

I. ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ КОЛИЧЕСТВЕННЫХ ДАННЫХ

Количественные данные — это данные, которые определяются путем измерения. Они подразделяются на две группы [2].

1. К количественным данным относятся результаты измерения артериального давления, температуры, общего белка. Такие данные обычно имеют единицы измерения: в нашем примере — мм рт.ст., градусы Цельсия и граммы в 1 литре соответственно. Вышеперечисленные данные являются *непрерывными*, т.е. принимающими любое значение на непрерывной шкале. Особенность непрерывных количественных данных: они могут иметь бесконечное количество знаков после запятой.

2. Также к количественным данным относятся величины, принимающие лишь определенное значение из диапазона измерения. Они называются *дискретными*. Например, длительность заболевания или госпитализации, количество выкуриваемых сигарет, количество клеток крови, число беременностей у женщины. Дискретные данные также имеют единицы измерения — дни, годы, штуки, разы и т.д. Они обычно являются целочисленными.

Количественные данные (и дискретные, и непрерывные) можно описать с помощью арифметических действий; они поддаются упорядочиванию: их можно расположить в порядке возрастания либо убывания.

Итак, как же наилучшим образом описать количественные данные? Ответ на этот вопрос напрямую связан с видом **распределения**, к которому принадлежат исследуемые данные [3]. По сути, целью анализа в данном случае является доказательство того, что исследуемая совокупность данных подчиняется *нормальному закону распределения, или доказательство обратного*.

Существует множество видов распределений случайных величин (равномерное, экспоненциальное, Пуассона и т.д.), но мы остановимся лишь на нормальном распределении, так как при выявлении различия распределения от нормального обычно несущественно, к какому конкретно виду распределения относятся данные.

Нормальное распределение, или **распределение Гаусса,** в природе встречается чаще всего, поэтому оно и было названо нормальным.

Нормальное распределение вероятностей случайной величины — это симметричная относительно среднего значения кривая (рис. 2), часто ее называют колоколообразной. Такое распределение характерно для большинства количественных медико-биологических данных, при условии большого объема исследуемой выборки.

Хорошим примером нормального распределения является рост. Естественно, что экстремально низкий и экстремально высокий рост будут встречаться с наименьшей вероятностью, а на вершине колокола окажется средний рост, наличествующий в популяции чаще всего.

При этом, возвращаясь к графику нормального распределения (см. рис. 2), можно сказать, что вероятность встретить те или иные значения случайной величины в выборке равна площади фигуры под кривой. Логично, что под краями «колокола» находятся наименее вероятные очень высокие и очень низкие уровни холестерина. Максимальную площадь имеют фигуры в центре «колокола», поэтому существует наибольшая вероятность встретить средние значения уровня холестерина в выборке испытуемых.

Проверка распределения данных на нормальность является обязательной на начальном этапе статистической обработки данных. Существует множество

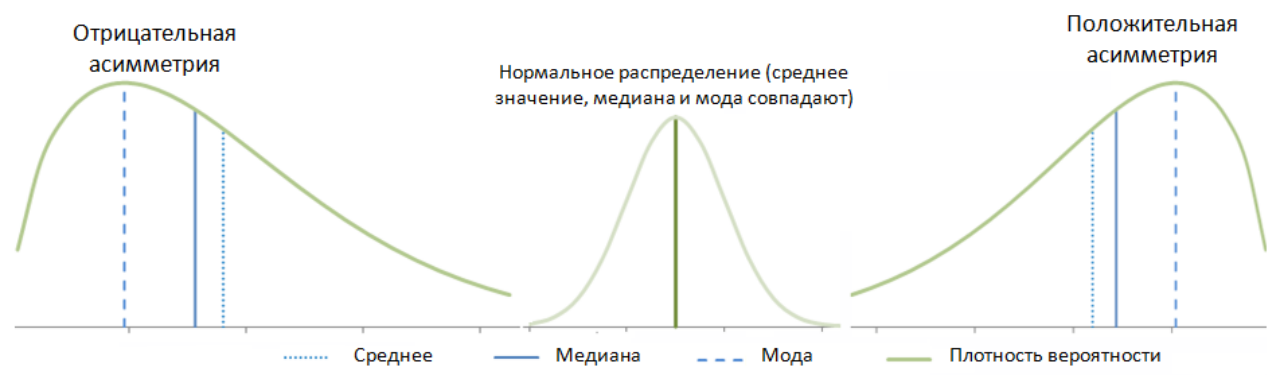


Рис. 3. Внешний вид симметричного и несимметричных распределений данных

вариантов проверки, и автор научного исследования волен выбирать самостоятельно, каким методом воспользоваться. К наиболее распространенным методам проверки принадлежности данных к нормальному распределению относятся:

критерий Колмогорова–Смирнова — рекомендован к использованию при больших объемах выборки (свыше 50 элементов);

критерий Шапиро–Уилка — рекомендован к использованию при малых объемах выборки (менее 50 элементов);

критерий асимметрии и эксцесса;

графики квантилей (Q-Q plot) — удобны при небольших выборках, для которых построение гистограммы не представляется возможным;

простой графический способ (гистограмма); не рекомендуется.

Наиболее широкое применение в научных исследованиях получили так называемые **формальные тесты**, к которым относятся *критерии Колмогорова–Смирнова, Шапиро–Уилка* и многие другие. Они являются разновидностью критериев согласия. Данные критерии используются для проверки нулевой гипотезы H_0, гласящей, что *случайная величина распределена нормально*. Результатом применения формальных тестов является определение уровня *значимости p.* Если полученный $p>0,05$, то H_0 принимается и мы делаем вывод, что распределение исследуемых величин статистически значимо не отличается от нормального распределения. Если же $p\leq0,05$, то H_0 отвергается, следовательно, исследуемое распределение статистически значимо отличается от нормального.

Критерий асимметрии и эксцесса используется гораздо реже, но наряду с критерием Шапиро–Уилка, является очень мощным способом для определения нормальности распределения. *Асимметрия* (As) характеризует степень скошенности графика вправо или влево по сравнению с эталонной кривой нормального распределения (рис. 3).

Очевидно, что колоколообразная кривая нормального распределения имеет асимметрию, равную нулю, так как график симметричен в обе стороны относительно среднего значения. В данном случае среднее значение, мода и медиана совпадают между собой. Коэффициент асимметрии больше нуля свидетельствует о правосторонней (положительной) асимметрии; в этом случае в выборке чаще всего встречаются значения больше среднего. Коэффициент асимметрии меньше нуля указывает на левостороннюю (отрицательную) асимметрию; в этом случае, напротив, в выборке чаще оказываются значения меньше среднего.

Эксцесс, или Kurtosis (E) — это мера остроты пика графика (рис. 4).

Островершинные распределения характеризуются положительным эксцессом и имеют пик выше пика нормального распределения. Плосковершинные распределения имеют отрицательный эксцесс, их пик находится ниже пика нормального распределения. Средневершинные распределения обладают нулевым эксцессом, что соответствует пику нормального распределения.

Отдельно остановимся на **графиках квантилей,** или так называемых **Q-Q plots** (Quantile-Quantile plots). Такой вид анализа редко можно встретить

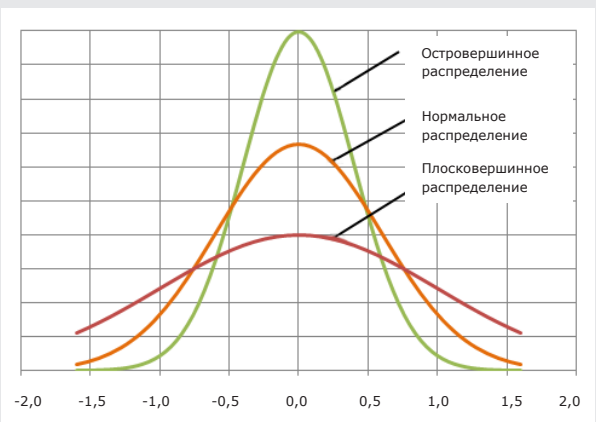
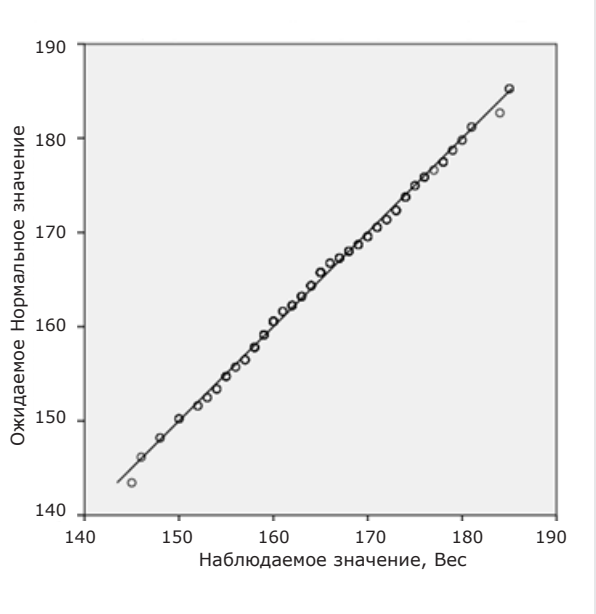
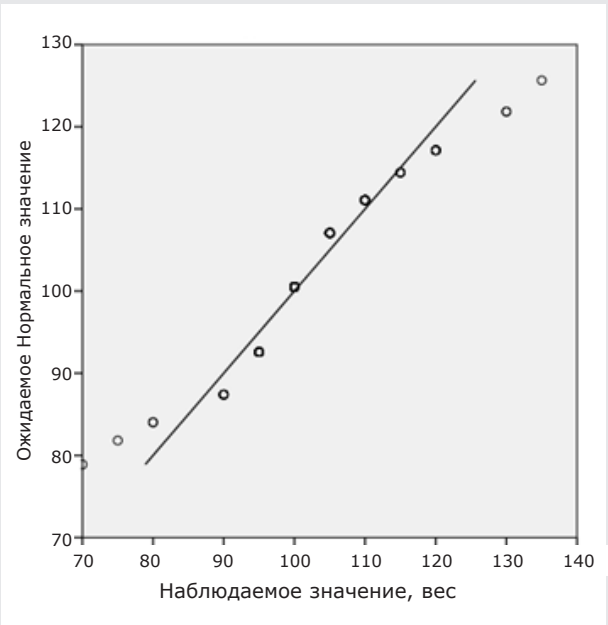


Рис. 4. Внешний вид островершинного, средневершинного (нормального) и плосковершинного распределений данных



а



б

Рис. 5. Q-Q plot для нормального распределения (*а*) и распределения, отличающегося от нормального (*б*)

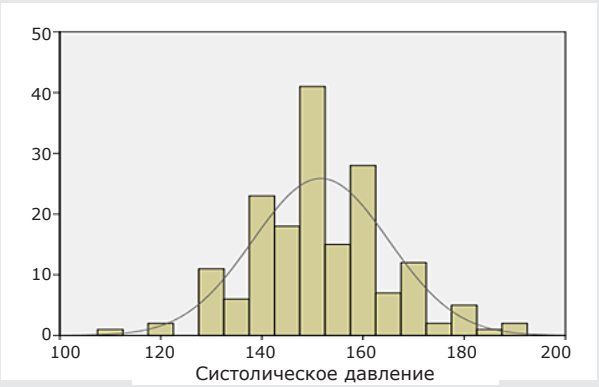


Рис. 6. Графическое представление распределения совокупности данных, отличающегося от нормального

в отечественных публикациях, и напротив, в зарубежных статьях он используется довольно часто.

Как интерпретировать Q-Q plot? На графике представлены квантили двух распределений (рис. 5). Первое распределение в виде прямой линии (восходящая линия под углом 45 градусов из левого нижнего угла графика) — это теоретически ожидаемые квантили нормального распределения. Расположенные вокруг прямой линии круги — это эмпирическое (построенное по реальным данным) распределение. При нормальном распределении квантили исследуемой совокупности должны выстраиваться в прямую линию и практически совпадать с теоретическим нормальным распределением. Соответственно, на рисунке 5а мы видим нормальное распределение, а на рисунке 5б — распределение, отличное от нормального.

Преимущество Q-Q plots по сравнению с формальными тестами является очевидным. Как говорилось выше, результатом формального теста будет лишь значение уровня значимости, по которому автор ориентируется, относится ли совокупность количественных признаков к нормальному распределению или нет. Читатель же научного исследования, рецензент или оппонент не имеют на руках первичного материала. Им остается лишь верить на слово автору, что исследование распределения на нормальность было проведено, даже если к представленным данным имеются вопросы. В свою очередь, Q-Q plot является показательным представлением распределения совокупности данных [4]. Представляя исследование распределения на нормальность в виде Q-Q plot, можно избежать недопонимания и лишних вопросов.

Хотелось бы отметить, что следует с осторожностью использовать **графический способ** (**гистограмму**). Крайне редко гистограмма имеет строго симметричный колоколообразный вид, по форме которой можно сделать вывод о *возможной* принадлежности результатов к нормальному распределению.

Внешний вид гистограммы существенно зависит от шага разбиения данных на классы и от объема выборки. Например, на рисунке 2, к которому мы уже обращались, представлено нормальное распределение, принадлежность к которому доказана несколькими формальными тестами. А на рисунке 6 фигурирует распределение, отличающееся от нормального, кажущаяся симметричность которого может ввести в заблуждение.

Кроме того, нормально распределенные *смешанные*

совокупности могут иметь на гистограмме несимметричный вид (например, если построить гистограмму распределения взрослого населения по росту, график получится не симметричным; разбиение результатов на 2 гистограммы — мужчины и женщины — даст полное соответствие колоколообразной симметричной форме для каждого пола). Поэтому при выборе способа проверки принадлежности к нормальному распределению *следует отдать предпочтение построению графиков квантилей или одному из формальных тестов.*

Приятно отметить, что в настоящее время в научных исследованиях редко можно встретить ссылку на графический метод как способ определения принадлежности распределения к нормальному.

Итак, возвращаемся к ранее заданному вопросу: как наилучшим образом описать количественные данные?

МЕТОДЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ ДЛЯ НОРМАЛЬНО РАСПРЕДЕЛЕННЫХ ДАННЫХ

Задача описательной статистики нормально распределенных величин — наглядно показать, чем отличается наше конкретное распределение от множества других распределений. С этой задачей для нормально распределенных величин успешно справляются основные две характеристики [5, 6].

Среднее арифметическое значение, или просто среднее значение *($\overline{x}$ или M)*:

$$\overline{x}(M) = \frac{\sum x}{n},$$

где x — значение признака; n — число членов совокупности.

Для характеристики нормального распределения также важно количественно показать величину разброса значений относительно среднего (ширину «колокола»). При этом для нормального распределения несущественно, в какую сторону отклоняется значение — в меньшую или в большую. Казалось бы, эту задачу можно решить с помощью **дисперсии** (в разных источниках ее обозначают с помощью σ^2 или D).

$$\sigma^2(D) = \frac{\sum(x - \overline{x})^2}{n}.$$

Однако, анализируя формулу дисперсии, можно увидеть, что результат подсчета дисперсии будет иметь другие единицы измерения, нежели среднее значение. Например, если мы исследовали рост, то среднее значение измеряется в метрах, а полученная дисперсия — в метрах квадратных. Таким образом, две эти характеристики не могут быть использованы вместе, тем более на одном графике.

Несложно сделать вывод, что для получения количественной характеристики разброса необходимо избавиться от квадрата, поместив формулу дисперсии под квадратный корень. Проделав эту процедуру, мы получим формулу **среднего квадратичного отклонения,** или **стандартного отклонения** (σ или S_D в разных источниках) — показателя разброса значений относительно среднего.

$$\sigma^2(S_D) = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x - \overline{x})^2}{n}}.$$

Таким образом, единственно верным представлением нормально распределенных величин является следующий вид: $p_{\overline{x}} \pm \sigma$ или M±S_D (где $\overline{x}(M)$ — среднее значение, а $\sigma(S_D)$ — стандартное отклонение). Исследователи вольны использовать любые общепринятые буквы для обозначения среднего и стандартного отклонения, нужно только указать это в подразделе, посвященном статистической обработке данных, который в настоящий момент является обязательным для любых медико-биологических статей и, тем более, диссертационных исследований.

Хочется обратить внимание на грубую ошибку, которую до сих пор совершают многие авторы намеренно или по незнанию. **Недопустимо** в качестве характеристики разброса относительно среднего значения использовать **стандартную ошибку среднего**. Стандартная ошибка среднего никоим образом не характеризует разброс; ее назначение — отразить точность выборочных оценок. Иными словами, стандартная ошибка среднего показывает, как точно выявленное среднее значение описывает исследуемую генеральную совокупность [7]. Формула стандартной ошибки среднего ($\sigma_{\overline{x}}$ или *m*):

$$\sigma_{\overline{x}}(m) = \frac{\sigma}{\sqrt{n}}.$$

Из формулы видно, что стандартная ошибка среднего в 5–6 раз меньше стандартного отклонения, что может привести к обнаружению статистически значимых различий там, где их в действительности нет. При ошибочном использовании стандартной ошибки среднего разброс будет намного меньше и интервалы двух сравниваемых переменных не перекроются там, где в реальности они должны перекрываться. Поэтому представление данных в виде M±m является неправомерным.

Несомненно, стандартная ошибка среднего может быть использована в научном исследовании, но по назначению — если поставлена задача измерить точность оценки среднего и рассчитать доверительные интервалы, а не для указания разброса относительно среднего.

Кроме среднего значения и стандартного отклонения, для количественных данных очень важной дополнительной описательной статистикой является **доверительный интервал** (ДИ) [8]. Довольно часто в медицинских публикациях встречается загадочное указание на расчет 95% ДИ и далее эта описательная статистика никак не объясняется и не используется. Хотя ДИ может быть крайне полезным для описания медицинских данных.

Что такое ДИ? Это интервал, в пределах которого с заданной вероятностью лежат выборочные оценки статистических характеристик генеральной совокупности. Иными словами, по средним значениям какого-либо исследуемого признака в выборке мы судим о среднем значении того же признака во всей генеральной совокупности.

Но если выборка имеет малый объем, то заявление о том, что среднее значение признака в выборке совпадет со средним значением всей генеральной совокупности, будет некорректным. В данной ситуации правильнее использовать диапазон средних значений генеральной совокупности. Например, 95% ДИ по общему белку составляет 64–83 г/л, т.е. это означает, что истинное среднее значение общего белка в генеральной совокупности с 95-процентной вероятностью лежит в этом интервале, а остальные значения могут выходить за его пределы. Понятно, что во многих ситуациях гораздо корректнее указать данный интервал, чем утверждать, что среднее значение общего белка во всей генеральной совокупности составляет 70 г/л.

Может быть рассчитан и любой другой ДИ, в зависимости от задач исследования. Например, он может быть расширен до 99%, и выводы будут более строгими. Кроме описания данных, ДИ может быть использован при изучении размера эффектов, но это предмет дискуссии в будущих статьях.

Для *дискретных* количественных данных в дополнение к основным описательным статистикам может быть рассчитана **частота**, т.е. количество раз, которое встречается каждая варианта.

Также дополнительной описательной статистикой для *дискретных* нормально распределенных величин является **мода** (см. следующий раздел).

Численные описательные статистики нормально распределенных величин могут быть представлены графическим способом. Классическим графическим методом представления таких данных служит **столбиковая диаграмма,** или **гистограмма [9]**. С ее внешним видом, безусловно, знакомы все, однако следует помнить, что на столбиках, представляющих среднее значение, обязательно должен быть показан разброс относительного среднего (стандартное отклонение). Этой цели служат **планки погрешностей** — отложенное в обе стороны от среднего значения стандартное отклонение ($\overline{X} \pm \sigma$ или $M \pm S_D$).

Часто авторы не добавляют планки погрешностей на рисунки — такое представление данных является некорректным. Например, из рисунка 7 видно, что существует разница между содержанием ГST в сердечной ткани интактной и опытной групп. Однако планки погрешностей показывают, что границы доверительных интервалов сравниваемых групп перекрываются, и дальнейшие выводы о наличии статистически значимых различий требуют дополнительного исследования с помощью статистических критериев. Поэтому отсутствие планок погрешности на диаграммах может ввести читателей в заблуждение.

Методы описательной статистики для несимметрично распределенных данных

Обычно описание нормально распределенных совокупностей не представляет трудностей, в то же время при представлении совокупностей, распределение которых отличается от нормального, наблюдается некоторая путаница. Задачей описательной статистики несимметрично распределенных данных является не просто показать набор каких-то значений, а максимально наглядно охарактеризовать их распределение, используя все доступные способы. Естественно, что полагаться на среднее значение и стандартное отклонение в данном случае нельзя.

Основной характеристикой в случае несимметричного распределения является **медиана** (Me) — такое значение, которое делит их совокупность на две равные по количеству членов части, причем в одной из них все значения меньше медианы, а в другой — больше.

Для характеристики разброса значений относительно медианы используют **процентили**.

Процентили — характеристики совокупности, отсекающие от нее по 0,01 части (они делят совокупность на 100 равных частей, поэтому всего процентилей 99). Таким образом, процентиль какого-либо значения — это процент случаев, которые имеют то же самое или меньшее значение. Естественно, что все 99 процентилей в научных исследованиях не используют, чаще всего применяют 25%, 50% и 75% процентили (рис. 8).

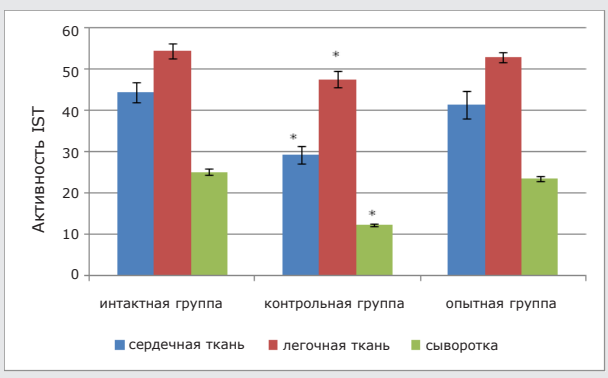


Рис. 7. Гистограмма с планками погрешностей

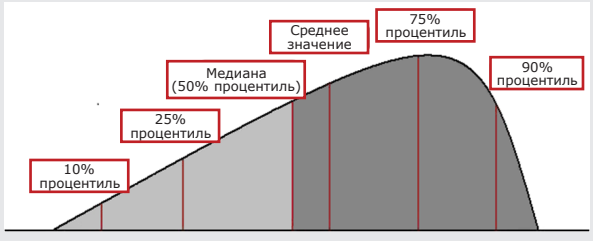


Рис. 8. Основные процентили

50% процентиль равен **медиане** и делит совокупность значений на две равные части.

25% процентиль, или нижний квартиль (Q_1), делит пополам нижнюю часть выборки (значения переменной меньше медианы). Это значит, что 25% значений переменной меньше нижнего квартиля.

75% процентиль, или верхний квартиль (Q_3), делит пополам верхнюю часть выборки (значения переменной больше медианы). Это значит, что 75% значений переменной меньше верхнего квартиля.

Хотя при необходимости можно вычислять любые процентили.

Процентили обычно используют для определения межквартильного размаха и представления межквартильного интервала.

Межквартильный интервал, или интерквартильный интервал (МКИ),— самая распространенная форма представления разброса асимметрично распределенных данных. В отличие от размаха (см. следующий пункт), МКИ — это не конкретное число, а участок между 1-м и 3-м квартилем (25% и 75% процентилем) — [X_{Q1}; X_{Q3}]. Таким образом, основная описательная статистика имеет вид: Me [МКИ]. Например, 34 [2;40] говорит о сильной правосторонней асимметрии распределения (медиана близка к 75% процентилю); можно сделать вывод, что в выборке наблюдается тенденция к высоким значениям исследуемого признака. Таким образом, медианы с межквартильными интервалами не просто ряды чисел, а важные характеристики выборки, по которым можно судить об основных особенностях выборки и наблюдающихся в ней тенденциях.

Хочется отметить, что в представлении МКИ автор может использовать не только точку с запятой, но и двоеточие, дефис, запятую, ведь от этого смысл МКИ совершенно не меняется.

В начале статьи мы говорили, что задача описательной статистики — максимально наглядно описать характеристики распределения. Конечно, просматривать ряды медиан с МКИ в таблицах довольно уныло, особенно когда результатов представлено много. В связи с этим все более популярным становится представление Me и МКИ не в численном, а в графическом виде. Этой цели служит так называемая **ящичковая диаграмма с усами,** или просто **«ящик с усами»** (box-and-whiskers diagram) [10, 11]. Такой вид представления основных описательных статистик (рис. 9) является в настоящее время наиболее предпочтительным.

Как интерпретировать данную диаграмму? Верхняя и нижняя границы ящика — это 25% и 75% процентили (межквартильный интервал). Усы — это минимальное и максимальное значения при отсутствии выбросов.

Либо нижняя и верхняя границы усов определяются следующим образом:

$X_1 = Q_1 - к \times (Q_3 - Q_1)$ и $X_2 = Q_3 + к \times (Q_3 - Q_1)$,

где X_1 — нижняя граница уса; X_2 — верхняя граница уса; k — число Тьюки, равное 1,5.

Горизонтальная линия внутри ящика — это медиана (если медиана находится симметрично по центру ящика, то можно говорить, что распределение близко к нормальному). Над ящиком группы 4 расположены два выброса, которые не учитывались в расчетах (число здесь является порядковым номером результата в таблице, положение на диаграмме указывает величину выброса). Несомненно, такой вид представления основных описательных статистик несимметрично распределенных величин намного показательнее, чем привычная Me [МКИ].

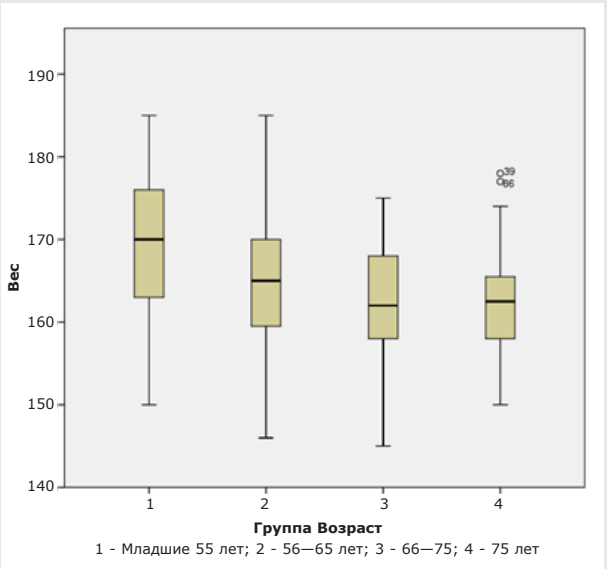


Рис. 9. Диаграмма «ящик с усами»

Размах (Range) — это показатель изменчивости выборки, который незаслуженно игнорируется отечественными исследователями. Размах в исследованиях используется в двух видах: размах вариации и межквартильный размах.

Размах вариации — это разность между самым большим и самым маленьким значением совокупности (между максимумом и минимумом) [9]:

Размах вариации = $X_{max} - X_{min}$

Размах вариации — прекрасный пример наглядной описательной статистики. Чем меньше размах вариации, тем устойчивее исследуемый процесс, т.е. его можно охарактеризовать как более предсказуемый. А это является очень важным при описании выборок. Значение размаха вариации точно отражает изменчивость выборки по возрасту, как, впрочем, и по любому другому признаку. Например, в исследовании участвовала группа пациентов от 18 до 25 лет, в этом случае размах вариации равен 7. Этот результат довольно близок к минимальному значению — нулю, поэтому мы можем сказать, что подверженность выборки колебаниям низкая.

Однако размах вариации имеет серьезный *недостаток:* он не является *робастным*, т.е. размах вариации очень чувствителен к *выбросам* (резко выделяющимся наблюдениям в большую или меньшую

сторону). Дело в том, что наличие в выборке даже небольшого числа выбросов может существенно повлиять на результат исследования — получаемые значения могут перестать нести в себе какой-либо смысл. Например, исследователь хочет показать, кто контрольная и опытная группы имеют схожие колебания по возрасту. При этом возраст в обеих группах варьирует от 18 до 25 лет, но в опытной группе оказался один испытуемый в возрасте 87 лет. Соответственно, ввиду наличия выброса мы получаем размах вариации 7 для контрольной группы, а для опытной — 69. Эти результаты говорят о том, что контрольная группа является устойчивой выборкой по данной характеристике, а опытная — нет. Конечно, этот вывод ошибочен, так как существует правило: размах вариации нельзя использовать при сравнении двух и более выборок, он служит только для характеристики одной выборки.

При необходимости сравнения изменчивости двух и более выборок на помощь приходит **межквартильный размах** — разница между верхним и нижним квартилями. Используя в качестве показателя изменчивости межквартильный размах, мы исключаем экстремальные величины и определяем размах остающихся наблюдений.

Межквартильный размах = $X_{Q3} — X_{Q1}$

Возвращаясь к предыдущему примеру, использование межквартильного размаха вместо размаха вариации обеспечит получение схожих значений изменчивости контрольной и опытной групп по возрасту. Таким образом, размах, в отличие от интервала,— это конкретное число, уменьшение которого путем корректного планирования эксперимента приведет к большей устойчивости исследуемого процесса.

Мода (Mo) — наиболее часто встречаемое значение, или значение с наибольшей частотой, она находится на вершине несимметричного распределения. Мода может быть очень информативной дополнительной описательной статистикой, когда нужно показать наиболее распространенную величину какого-либо медицинского показателя.

Однако мода в научных исследованиях в качестве описательной статистики встречается довольно редко. Это связано с трудностями ее нахождения для некоторых распределений. Например, отдельные совокупности данных не имеют моды, так как каждое значение встречается один раз. Также нахождение моды затруднено в мультимодальных совокупностях — где 2 или более значений имеют одну частоту встречаемости, т.е. распределение дает 2 или более пиков. Кроме того, мода не подходит для описания непрерывных случайных величин, ее можно использовать только для дискретных количественных данных.

II. МЕТОДЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ КАЧЕСТВЕННЫХ ДАННЫХ

Качественные данные — это данные, которые устанавливаются описательным путем, т.е. их невозможно описать численно. Качественные данные представляют собой условные коды неизмеряемых категорий или условную степень выраженности исследуемого признака. Этот тип данных нельзя описать с помощью арифметических действий. Качественные данные бывают двух видов.

Номинальные качественные данные — это наблюдения, которые классифицируются в группы взаимоисключающих категорий, причем между категориями невозможно определить никаких количественных взаимоотношений. В данной классификации числовые обозначения даются совершенно произвольно, их можно поменять местами либо обозначить другими цифрами. Например, классификация по группе крови, полу, профессии, семейному положению и т.д.

Разберем в качестве примера классификацию по семейному положению: 1 — холост/не замужем; 2 — женат/замужем; 3 — вдовец/вдова; 4 — разведен(а). Если мы обозначим данные категории по-другому: 0 — разведен(а); 1 — вдовец/вдова; 2 — женат/замужем; 3 — холост/не замужем, то для последующей статистической обработки данных это окажется совершенно неважным. В этом состоит особенность номинальных качественных данных.

Порядковые или ранговые качественные данные, в отличие от номинальных, можно расположить в порядке убывания или возрастания, но арифметические действия с ними также проводиться не могут. Для этого типа данных в описываемых категориях наблюдается постепенное изменение эмпирической значимости, иными словами, уменьшение или увеличение интенсивности исследуемого признака.

Наиболее известным примером порядковых данных является интенсивность боли, которая располагается по шкале от 0 до 10, где 0 — отсутствие боли, а 10 — нестерпимая боль. Логично, если мы поменяем местами некоторые категории, то в дальнейшем при проведении статистического анализа возникнет путаница. Таким образом, при работе с порядковыми данными произвольно устанавливать категории нельзя. Другими примерами порядковых данных являются степень тяжести заболевания, стадия болезни, самооценка состояния здоровья, уровень дохода, образование, интенсивность курения.

Описать качественные данные можно только двумя способами [7].

1. Подсчитать, какая **доля** (p) от общего числа объектов приходится на то или иное значение:

$$p = \frac{m}{n} \cdot 100\%$$

В настоящее время при вычислении процентных долей принято указывать их разброс. Для этих целей служит стандартное отклонение процентной доли (σ_p):

$$\sigma_p = \sqrt{\frac{p \cdot (1-p)}{n}}$$

Из формулы стандартной ошибки процентной доли

следует, что она уменьшается при увеличении размера выборки (n). Предположим, что улучшение самочувствия после применения нового препарата отмечали 63% испытуемых опытной группы при объеме выборки n=30. Проведем расчет σ_p.

$$\sigma_{p=0,63} = \sqrt{\frac{0,63 \cdot (1-0,63)}{30}} = 0,09$$

В контрольной группе, принимавшей плацебо, улучшение самочувствия отметили 37% испытуемых. Соответственно, при том же объеме выборки n=30 отклонение будет той же величины. Результаты для обеих групп испытуемых выглядят следующим образом: 63±9% и 37±9%. Мы делаем вывод, что интервалы двух групп не перекрываются, следовательно, применение нового препарата статистически значимо приводит к улучшению самочувствия пациентов.

Теперь допустим, что выборка составила 10 человек (в медицинских исследованиях часто можно видеть необоснованно малые выборки, вплоть до 3 единиц).

$$\sigma_{p=0,63} = \sqrt{\frac{0,63 \cdot (1-0,63)}{10}} = 0,15$$

Результаты для обеих групп 63±15% и 37±15% свидетельствуют о необходимости применения дополнительных статистических методов для окончательного заключения о равенстве процентных долей.

Поэтому использование процентных долей с учетом их стандартных отклонений ($p \pm \sigma_p$%) является такой же необходимостью, как и привычное всем представление количественных данных в виде среднего значения со стандартным отклонением ($M \pm S_d$).

2. Порядковые данные относятся к полуколичественным, поэтому в дополнение к долям для них могут использоваться частота и мода (см. предыдущий раздел).

Особенно хочется отметить, что представление качественных данных в виде абсолютных значений является малоинформативным. Обосновать данное утверждение можно с помощью двух таблиц (в таблице 1 данные представлены в виде абсолютных значений, в таблице 2 — в виде процентных долей).

Очевидно, что первый вариант таблицы «Характеристика пациентов» абсолютно неинформативен. Никому неинтересно (ни читателю, ни рецензенту, ни оппоненту) самостоятельно высчитывать доли качественных признаков в каждой группе пациентов. Да и научное исследование с подобным представлением качественных данных смотрится несерьезно. Поэтому от абсолютных значений при описании любых качественных данных следует отказаться. Это показано в таблице 2: из нее хорошо видно соотношение исследуемых групп пациентов, по указанным отклонениям процентных долей сразу можно понять, какие группы значимо отличаются по качественному признаку.

Отдельно необходимо уделить внимание описанию такой важной характеристики, как **возраст**, которая фигурирует практически в каждом медицинском научном исследовании. Возраст может быть распределен нормально или нет. Если возраст участвует в последующем статистическом анализе, то это необходимо выяснить и в дальнейшем использовать его согласно правилам статистики.

Однако если возраст не участвует в дальнейшем статистическом анализе, а также при начальном представлении характеристик пациентов его следует представлять особым образом. Например, авторы часто представляют возраст в следующем виде: 46±5,2. Какие выводы неспециалист в статистике сделает, исходя из среднего значения возраста с его разбросом? Специалист может применить правило трех сигм и рассчитать, что возраст 99% пациентов находился в интервале 30,4–61,6 года. Тем более, что такое представление возможно только при нормальном распределении (а возраст редко бывает распределен нормально).

Таблица 1

Характеристика пациентов по морфологической классификации

Показатель	Кол-во пациентов
Мужчины	350
Женщины	312
Катаральный аппендицит	200
Поверхностный аппендицит	138
Флегмонозный аппендицит	103
Флегмонозно-язвенный аппендицит	76
Апостематозный аппендицит	69
Гангренозный аппендицит	47
Перфоративный аппендицит	29

Таблица 2

Характеристика пациентов по морфологической классификации

Показатель	Кол-во пациентов (n=662)
Мужчины (n=350)	53,0±1,9%
Женщины (n=312)	47,0±1,9%
Катаральный аппендицит (n=200)	30,2±1,8%
Поверхностный аппендицит (n=138)	21,0±1,6%
Флегмонозный аппендицит (n=103)	15,5±1,4%
Флезмонозно-язвенный аппендицит (n=76)	11,4±1,2%
Апостематозный аппендицит (n=69)	10,4±1,2%
Гангренозный аппендицит (n=47)	7,1±1%
Перфоративный аппендицит (n=29)	4,4±0,8%

Рис. 10. Алгоритм выбора описательных статистик

Обычно для понимания дизайна исследования необходимо выяснить минимальный и максимальный возраст, т.е. размах. Поэтому правильнее возраст представить в виде Me [Min-Max]. Например, 56 [32–62]. Из такого описания возраста отлично видно, что в исследовании участвовали пациенты от 32 до 62 лет, причем положение медианы говорит о существенном сдвиге количества пациентов в сторону старшего возраста. Если авторам известно о имеющихся выбросах, то предпочтительнее использовать Me [МКИ], но выбор подобной описательной статистики желательно объяснить в тексте.

ЗАКЛЮЧЕНИЕ

Таким образом, описательная статистика является очень важным разделом статистической науки. Для корректного описания результатов исследования хочется предложить простой алгоритм (рис. 10).

Следующая статья будет посвящена основным понятиям аналитической статистики.

Финансирование исследования и конфликт интересов. Исследование не финансировалось каким-либо источником, и конфликты интересов, связанные с данным исследованием, отсутствуют.

ЛИТЕРАТУРА/REFERENCES

1. Pupovac V., Petrovecki M. Summarizing and presenting numerical data. *Biochem Med (Zagreb)* 2011; 21(2): 106–110, https://doi.org/10.11613/bm.2011.018.

2. Реброва О.Ю. *Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA*. М: МедиаСфера; 2000. Rebrova O. Ju. *Statisticheskij analiz medicinskih dannyh. Primenenie paketa prikladnyh programm STATISTICA* [Statistical analysis of medical data. The use of the application package STATISTICA]. Moscow: MediaSfera; 2000.

3. Simundić A.M. Types of variables and distributions. *Acta Medica Croatica* 2006; 60 Suppl 1: 17–35.

4. Pleil J.D. QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics. *J Breath Res* 2016; 10(3): 035001, https://doi.org/10.1088/1752–7155/10/3/035001.

5. Nick T.G. Descriptive statistics. *Methods Mol Biol* 2007; 404: 33–52, https://doi.org/10.1007/978–1–59745–530–5_3.

6. Петри А., Сабин К. *Наглядная медицинская статистика. Учебное пособие.* М: ГЭОТАР-Медиа; 2019. Petri A., Sabin K. *Naglyadnaya meditsinskaya statistika. Uchebnoe posobie* [Clear medical statistics. Textbook]. Moscow: GEOTAR-Media; 2019.

7. Гланц С. *Медико-биологическая статистика*. М: Практика; 1998. Glants S. *Mediko-biologicheskaja statistika* [Biomedical statistics]. Moscow: Praktika; 1998.

8. Michel M.C., Murphy T.J., Motulsky H.J. New author guidelines for displaying data and reporting data analysis and statistical methods in experimental biology. *Mol Pharmacol* 2020; 97(1): 49–60, https://doi.org/10.1124/mol.119.118927.

9. Spriestersbach A., Röhrig B., du Prel J.B., Gerhold-Ay A. Blettner M. Descriptive statistics: the specification of statistical measures and their presentation in tables and graphs. Part 7 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 2009; 106(36): 578–583, https://doi.org/10.3238/arztebl.2009.0578.

10. Martinez E.Z. Description of continuous data using bar graphs: a misleading approach. *Rev Soc Bras Med Trop* 2015; 48(4): 494–497, https://doi.org/10.1590/0037–8682–0013–2015.

11. Buttarazzi D., Pandolfo G., Porzio G.C. A boxplot for circular data. *Biometrics* 2018; 74(4): 1492–1501, https://doi.org/10.1111/biom.12889.

ИНФОРМАЦИЯ ОБ АВТОРЕ:

А. П. Баврина, к. б. н., доцент кафедры медицинской физики и информатики, руководитель Центра биомедицинской статистики, организации исследований и цифровой медицины ФГБОУ ВО «Приволжский исследовательский медицинский университет» Минздрава России

Для контактов: Баврина Анна Петровна e-mail: annabavr@gmail.com