

ЧТЕНИЯ ПАМЯТИ ВЛАДИМИРА ЯКОВЛЕВИЧА ЛЕВАНИДОВА

Vladimir Ya. Levanidov's Biennial Memorial Meetings

2025

Вып. 11

<https://doi.org/10.25221/levanidov.11.02>

<https://elibrary.ru/xseuhq>

ВЫБОР МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ ПРИ ОБРАБОТКЕ ГИДРОБИОЛОГИЧЕСКИХ ДАННЫХ (МИНИ-ОБЗОР)

М.В. Астахов

*Федеральный научный центр биоразнообразия наземной биоты Восточной Азии ДВО РАН,
пр. 100-летия Владивостока, 159, г. Владивосток, 690022, Россия. E-mail: mvastakhov@mail.ru*

В идеале мера центральной тенденции, она же параметр положения, рассматривается как типичный показатель выборки (совокупности), обобщающий в себе её отличительные свойства, как единственное число, соответствующее центру симметрии (медиана) или центру тяжести (средние величины) эмпирического распределения. В представленном мини-обзоре обсуждаются ограничения и примеры использования некоторых мер положения при описании количественных гидробиологических материалов.

CHOOSING A MEASURE OF CENTRAL TENDENCY IN THE PROCESSING OF HYDROBIOLOGICAL DATA (MINI-REVIEW)

M.V. Astakhov

*Federal Scientific Center of the East Asia Terrestrial Biodiversity, FEB RAS, 159 Stoletiya Vladivostoka Avenue,
Vladivostok, 690022, Russia. E-mail: mvastakhov@mail.ru*

Ideally, the measure of central tendency, also known as the location parameter, is considered as a typical indicator of the sample (population), generalizing its distinctive properties, as a single number corresponding to the center of symmetry (median) or to the center of gravity (average values) of the empirical distribution. This mini-review discusses the limitations and examples of some measures of location usage in the description of quantitative hydrobiological materials.

Обзор статистических методов (Huber, 1972), содержащий фразу «мы рассмотрели около 65 оценок положения», был опубликован более 50 лет назад, однако за прошедшие годы появилось ещё немало «лучших» мер центральной тенденции, сопровождаемых разнообразными модификациями, причем теоретические изыскания в этом направлении продолжают. Автором настоящего мини-обзора будут упомянуты лишь несколько из самых незамысловатых оценок центра распределения, достойных применения в практике гидробиологических исследований. Для простоты восприятия приведённые примеры содержат выборки минимального ($n \geq 5$) объёма, при котором попытки вычисления меры центральной тенденции могут не стать причиной недовольства знатоков теории математической статистики. Между тем теоретически, согласно Ф.М. Чаттеру (Chutter, 1972), для расчёта с 95 %-ой доверительной вероятности средней плотности распределения зообентоса с точностью ± 5 % от средней генеральной совокупности только на одном перекате надо взять 448 донных проб. А по данным М.М. Сметанина и Н.А. Гордеева (1977),

оценка средней массы 10-летних лещей при допустимой ошибке взвешивания ± 10 г (доверительная вероятность 95 %) может потребовать 1135 особей указанного вида и возраста. Автор не сторонник настолько преувеличенного внимания к выполнению формальных требований, но и к вычислению среднего по 2–3 измерениям относится скептически. Речь, конечно, о среднем арифметическом – самой известной и широко применяемой мере центральной тенденции. Среднее арифметическое понятно всем. Этот показатель центра распределения «проще других и по смыслу, и по свойствам, и по способу получения» (Калинина, Панкин, 2002). В стремлении обосновать «при помощи основных законов теории вероятностей» мнение, что «всегда будет наиболее надежным придерживаться именно» среднего арифметического «из многих найденных значений», великий К.Ф. Гаусс (Гаусс, 1957) сконструировал соответствующую гипотетическую функцию, приложимость которой для описания набора данных теперь рассматривается как признак того, что эти данные имеют нормальное распределение. Т.е. с годами произошла своеобразная инверсия в понимании причинно-следственной связи между средним арифметическим и функцией плотности нормального распределения. В настоящее время подвергается критике вычисление среднего значения выборки без проверки последней на предмет соответствия закону нормального распределения. Считается, что для адекватного осуществления такой проверки выборка должна содержать ≥ 30 наблюдений. Однако в реальной жизни получить выборку «канонического» размера удастся далеко не всегда. Даже тогда, когда это достижимо технически. Возможно ли взять с одного переката 30 или даже 448 донных проб? Если площадь переката позволит, то, наверное, да. Но что в таком случае останется от бентосного сообщества? Поэтому, как правило, с одной станции отбирают не больше 2–3 проб. В теории выборку такого размера проверить на нормальность нельзя, значит, не стоит вычислять по ней среднее; достаточно привести лимиты (min–max). На практике, однако, большинство коллег такое суждение игнорируют. В быту мы настолько привыкли оперировать средним арифметическим, что не каждый догадается соотнести свои навыки в его вычислении с нижеприведённой формулой:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad (1)$$

где $\bar{x}$ – среднее арифметическое, x_i – i -ый член выборки, n – объём выборки.

Широко используя среднее арифметическое, мы далеки от понимания того, над чем ломают головы поколения специалистов в сфере теории статистики. Нам невдомек, что среднее значение должно быть оценкой «состоятельной, несмещённой и эффективной» (Цивинский, 2011), – вопреки заявлению швейцарского статистика П.И. Хубера (Huber, 1972), будто «нет никаких причин, кроме математического удобства, навязывать» подобные требования.

Для гидробиолога-практика важно следующее. При первичной статистической обработке данных надо обращать внимание на показатели варьирования числовых значений выборочного ряда вокруг среднего арифметического, т. е. на меры разброса. Например, на стандартное отклонение или на коэффициент вариации (относительное стандартное отклонение). Последний удобен тем, что позволяет непосредственно оценить адекватность рассчитанного среднего. Если коэффициент вариации > 50 %, то выборка неоднородна, несимметрична, её распределение «заметно отличается от нормального» (Дмитриев, 2009), а полученное среднее арифметическое не является типичной характеристикой совокупности.

При сильно ассиметричных распределениях коэффициент вариации достигает 100 % и выше (Лакин, 1990). Поскольку симметричность – это «одна из основных характеристик нормального распределения» (Реброва, 2002), то, по мнению некоторых исследователей, критерий оценки должен быть жестче, и в качестве пороговой для коэффициента вариации принимают величину 33 % (Закс, 1976).

То, что среднее арифметическое «впитывает» (Рокицкий, 1973) и обобщает информацию обо всех членах выборки, делает этот показатель наилучшей мерой центральной тенденции именно для симметрично распределённых выборок. Однако это же преимущество превращается в «ахиллесову пяту» среднего арифметического при наличии в выборке хотя бы одного экстремального наблюдения. Экстремальное наблюдение, или выброс, это элемент выборочного ряда, числовое значение которого существенно превышает величины других членов выборки. Такое «выпадающее» наблюдение может появиться в результате ошибки, например, при записи первичных данных, а может отражать естественную закономерность, например, как показатель временной динамики. В любом случае выбросы искажают распределение (в плане симметричности) и оказывают непропорциональное завышающее влияние на среднее арифметическое. Оценить наличие выбросов можно посредством специализированных тестов (например, критерий Грабсса), либо графически.

Наглядный пример построен нами по табулированным данным (Chutter, Noble, 1966) учёта личинок ручейника-гидропсихиды *Cheumatopsyche thomasseti* (Ulmer) в бентосных пробах ($n = 10$). Из рисунка (рис. 1А) видно, что величина среднего арифметического (115 экз.) этой выборки превосходит числовые значения восьми из десяти её членов. Значит такое среднее арифметическое не может служить показателем центра распределения. Это же следует из теоретических соображений: слишком высокий коэффициент вариации (137,8 %); распределение наблюдений не согласуется с нормальным (тест Шапиро-Уилка, $p < 0,001$). По данным Н.Н. Рязова (1971), «средние», вычисленные по выборкам с выбросами, В.И. Ленин называл «фиктивными».

Когда даже из-за одного экстремального значения расчёт среднего арифметического может стать причиной ошибочного заключения, то теоретически, лучшей мерой положения считается медиана, которая в отличие от среднего арифметического, всегда остаётся робастной (устойчивой к выбросам). Однако секрет робастности медианы неутешителен. Экстремальные значения при её определении просто не учитываются, так же, как не учитываются и почти все остальные элементы выборки. Строго говоря, при определении медианы подавляющая часть информации, заключённой в выборке, вообще не используется. Если набор данных нечётный, то медиана – это числовое значение срединного элемента упорядоченной по возрастанию выборки, который разделяет последнюю на две части, равные по количеству наблюдений. Если же набор чётный, то при определении медианы появляется возможность учёта информации уже о двух срединных («центральных») наблюдениях. В таком случае, в качестве аналога медианного показателя, из этих двух наблюдений берут среднее арифметическое. Причём то, что среднее арифметическое получают по двум (!) значениям, не смущает виртуозов теоретических изысканий. Они имеют стойкое убеждение будто «медиана удобна». Между тем, если числовой показатель наименьшего из двух «центральных» наблюдений упорядоченной чётной выборки окажется нулём, то медиана будет вычислена как $\frac{1}{2}$ наибольшего. Т.е. тоже на основе информации о единственном элементе выборки. Так, медиана чётной выборки 0, 0, 0, 10, 20, 30 равна 5. Очевидно, что

если нулевые значения составляют более половины набора данных, то нулю равна и медиана (Alexander, 2012).

Следствием того, что медиана не зависит от каждого члена выборки, является невозможность её применения в качестве адекватного обобщающего показателя при исследовании каких-либо изменений во времени (темпы роста, интенсивность дрефта и т. п.). Это ясно из простого числового примера. Рассмотрим упорядоченный набор 8, 9, 10, 11, 12. Здесь и медиана, и среднее арифметическое равны 10. Если, не изменяя объём этой выборки, заменять 8 и 9 на любые два значения ≤ 10 , а 11 и 12 – на любые два значения ≥ 10 , то средние арифметические генерируемых таким способом наборов чисел будут изменяться соответствующим образом, но медиана не изменится совершенно. Например, если 8 и 9 заменить на 1 и 1, а 11 и 12 заменить на 111 и 222, то медиана так и останется равной 10, хотя среднее арифметическое возрастёт до 69. Значит, если принять во внимание медианный показатель, то мы придём к ошибочному заключению, будто с этими выборками не произошло никаких изменений. Т.е. из-за использования медианы в ходе реального исследования временной динамики может остаться незамеченной та или иная тенденция. Числовой пример описан с акцентом на фразе «не изменяя объём» выборки. Это сделано с целью подчеркнуть факт того, что неизменность объёма выборки – залог стабильности медианного показателя. Единственное, что способно сдвинуть медиану с «насиженного» «центра» – это добавление/удаление отдельных наблюдений. Потребность в таких действиях может быть заложена в планах некоторых исследований.

Следует всегда помнить, что медиана, превосходящая среднее арифметическое в аспекте робастности, но существенно уступающая ему в информативности, гораздо больше нуждается в поддержке мерой разброса при практическом применении. В случае выбора медианного показателя мерой разброса служит обычно интерквартильный размах – диапазон значений между 25-м и 75-м процентилями, иначе говоря, половина выборки, центрированная относительно 50-го перцентиля (т. е. самой медианы). Если выборочное распределение соответствует нормальному, то медианный показатель и среднее арифметическое примерно равны. Поэтому, «вычитая медиану из среднего» можно получить приблизительную «оценку асимметричности данных: чем больше разность, тем сильнее» выборка скошена вправо (Oliver, Hall, 1989; Ланг, Сесик, 2011).

Итак, из двух наиболее популярных мер центральной тенденции одна (среднее арифметическое) – не считается достаточно робастной, а другая (медиана) – не считается достаточно информативной. Что делать? (Чернышевский, 1975). Вернёмся к примеру с личинками *Ch. thomasseti* (рис. 1А). Если из этого набора (9, 21, 30, 36, 78, 82, 88, 91, 166, 544 экз. – после упорядочивания по возрастанию) исключить максимальное значение, то картина заметно изменится (рис. 1Б). Немного уменьшится медиана (с 80 до 78 экз.), зато среднее арифметическое уменьшится почти вдвое (со 115 до 67 экз.) и переместится к центру распределения. Последнее, если верить результату теста Шапиро-Уилка, станет соответствовать нормальному ($p = 0,278$). Чтобы медиана осталась неизменной, от выборки надо «отсечь» (т. е. исключить из расчёта) не только максимальное (544 экз. – это явный выброс; тест Граббса, $p < 0,001$), но и минимальное значение (9 экз.). Среднее арифметическое (74 экз.), рассчитанное по такой модифицированной выборке станет называться усечённым средним арифметическим. В данном случае (рис. 1В) мы получим 10 % усечённое среднее, поскольку при исходном объёме выборочного ряда в 10 наблюдений (100 %) с каждого его конца будет удалено по 1 наблюдению

(10 %). Что в сумме означает исключение 20 % данных (или 0,2 в долях единицы). Обычно число отсекаемых наблюдений при заданном проценте усечения определяют как результат округления произведения доли усечения на общее количество наблюдений. В нашем случае округление не требуется: $0,1 \times 10 = 1$. В зависимости от первоначального объёма выборки и степени её асимметричности с обоих концов набора данных можно отсечь до 25 % значений (Huber, 1972). При таком усечении проанализирована будет только половина первичного материала, центрированная относительно медианы. Усечённое среднее – это «компромисс» (Hintze, 2007) между информативностью средней арифметической и робастностью медианы. Когда распределение симметрично и согласуется с нормальным, то усечённое среднее равно среднему арифметическому и медиане. В таком случае все три указанные меры положения совпадают с модальным значением (точнее, находятся в пределах

модального класса). Моду, к слову, тоже относят к мерам центральной тенденции, но используют её «ограничено», когда «данные не являются чисто количественными» (Рокицкий, 1973), когда «распределение бимодальное» (Ланг, Сесик, 2011).

Каковы же последствия нашей подготовки к расчёту 10 % усечённого среднего? Медиана осталась такой же, как в исходной выборке (рис. 1); расхождение между нею (80 экз.) и пересчитанным средним (74 экз.) стало ещё менее заметным, нежели при исключении только экстремального значения. Предположение о нормальности снова не отвергается тестом Шапиро-Уилка ($p = 0,233$), хотя коэффициент вариации, который с 72,9 % снизился только до 63 %, по-прежнему свидетельствует об обратном. Добиться однородности выборочного ряда не удалось, но ясно, что при усечении данных с обоих концов исходной выборки была получена более адекватная оценка центра распределения, чем при удалении одного только выброса. Несогласованность оценок теста Шапиро-Уилка и коэффициента вариации, очевидно, связана с недостаточностью количества данных. В таком случае верным представляется выбор в пользу коэффициента вариации. Это мнение обусловлено фактами того, что алгоритм теста Шапиро-Уилка, заложенный в компьютерных статистических программах (например, STATISTICA, SPSS, PAST), способен обнаружить «нормальность» ($p > 0,05$) даже в данных, распределение которых заведомо не соответствует нормальному. Нами такие факты установлены для равномерно распределённых числовых последовательностей от 1, 2, 3, 4 ($n = 4$) и далее, вплоть до 1, 2, ... 50,

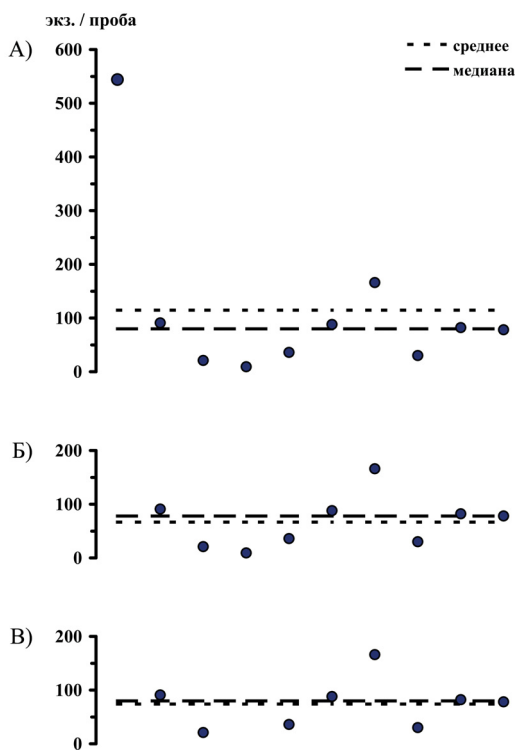


Рис. 1. Визуализация табулированных данных (Chutter, Noble, 1966) подсчёта личинок *Ch. thomasseti* в пробах; каждая проба отобрана с участка дна площадью 1 квадратный фут ($\approx 0,09 \text{ м}^2$). А – полная выборка; Б – выборка после удаления экстремального значения; В – выборка, подготовленная к расчёту 10 % усечённого среднего

51 ($n = 51$). Между тем коэффициент вариации, рассчитанный для таких выборок, всегда оказывался $> 50 \%$.

В примере с личинками *Ch. thomasseti* визуализация получаемых значений (рис. 1) демонстрирует поэтапное «сближение» среднего арифметического с медианой, что свидетельствует о снижении уровня асимметрии распределения. Следовательно, этот простой графический подход можно использовать для определения доли отсечения крайних (в том числе экстремальных) значений с целью максимально возможного приближения эмпирического распределения к симметричному. Разумеется, перед процедурой удаления крайних наблюдений с обоих концов выборки, последнюю надо упорядочить по возрастанию или убыванию. На практике процедура усечения выборочного ряда может быть востребована в ходе подготовки данных для регрессионного, дисперсионного или ковариационного анализа влияния факторов среды на питание, дыхание, экскрецию, рост, размеры гидробионтов, а также на плотность скоплений последних (например, Blumenshine et al., 2012; James, 2022; Pousse et al., 2023).

Желая сохранить полный объём исходной выборки, некоторые коллеги не просто отбрасывают крайние наблюдения, но заменяют их на равные «пороговым». Например, в выборке 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 каждое из двух значений 1 и 2 можно заменить на «пороговое» значение 3, а вместо 9 и 10 дважды поставить 8. Среднее арифметическое, полученное из выборки, модифицированной подобным образом, называется винсоризованным средним (в данном случае 20 %-ым) – в честь американского биостатистика Ч.П. Уинзора (Charles P. Winsor), предложившего такой в подход в первой половине прошлого века (Dixon, 1960). В наши дни винсоризация находит применение, например, при выявлении корреляции между распространённостью паразитов у лососеобразных и видами кормовых объектов этих рыб (Prati et al., 2020) или при изучении проблемы избыточного прилова в процессе промысла (Yan et al., 2022). Тем не менее ясно, что «принимая во внимание затраты и усилия на сбор данных, некоторые учёные-экологи не захотят заменять часть зарегистрированных наблюдений субъективно выбранными пороговыми значениями» (Gregoire, Schabenberger, 1999). Так, в ходе годовичного исследования сезонной динамики бентоса на перекате р. Кедровая (Японское море) минимальная плотность распределения личинок хирономид $61 \text{ экз./0,125 м}^2$ была зарегистрирована в октябре 2006 г., а максимальная $5114 \text{ экз./0,125 м}^2$ – в феврале следующего года (оригинальные данные). В другие месяцы исследования плотность распределения этих беспозвоночных варьировала в пределах $230\text{--}2089 \text{ экз./0,125 м}^2$. Как выброс (и графически, и статистическими тестами) был идентифицирован только февральский показатель. Теоретически, можно исключить этот выброс вместе с минимальным показателем, чтобы для получения «среднегодовой» величины ограничиться расчетом 10 % винсоризованного ($1042 \text{ экз./0,125 м}^2$) или 10 % усечённого ($1018 \text{ экз./0,125 м}^2$) среднего. Однако ... прошло почти 20 лет, но воспоминание о двухнедельном выпутывании 5 тыс. мелких личинок из сплетений нитчатых водорослей не позволяет отказаться от февральского материала. Тем более что речь об уникальных данных сезонной динамики. В подобных случаях, вероятно, следует не исключать крайние (даже экстремальные) значения, а представлять их в виде лимитов (min–max) вместе с усечённым средним арифметическим (или медианой).

Среднее арифметическое как оценка центра распределения наилучшим образом работает с выборками, имеющими приблизительно аддитивную (слагаемую) структуру, т. е. с такими, где при упорядочивании по возрастанию каждое последующее числовое значение больше предыдущего на ту или иную величину. Если эта

величина неизменна, то подобная числовая последовательность представляет собой простую арифметическую прогрессию. При отображении на точечной диаграмме члены такой последовательности располагаются вдоль прямой, поэтому аддитивную связь иногда называют «линейной» (рис. 2А). Однако бывают ситуации, когда на точечной диаграмме расположение упорядоченных по возрастанию (или убыванию) значений явно соответствует не прямой линии, а восходящей (или нисходящей) кривой. Это свидетельствует о наличии между членами выборки отношений, близких к мультипликативным (связанным с умножением). Идеальным примером числовой последовательности с мультипликативной структурой является геометрическая прогрессия, график которой – типичная экспоненциальная кривая (рис. 2Б). Понятно, что при сборе гидробиологического материала получить выборку, характеризующуюся строгой аддитивной или мультипликативной связью едва ли возможно.

В примере с личинками *Ch. thomasseti* (рис. 1А) среднее арифметическое (115 экз.) заметно больше медианы (80 экз.). Этот факт ясно указывает на то, что выборочное распределение имеет положительную (правостороннюю) асимметрию (Ланг, Сесик, 2011). В подобных ситуациях, по мнению К.Н. Несиса (1969), «средняя арифметическая нехарактерна для основной доли изучаемой совокупности ... не имеет биологического смысла и недостаточно устойчива, а потому применение ее для исследования первичных данных нецелесообразно». Чтобы «распределение приблизилось к нормальному», этот известный гидробиолог и малаколог рекомендовал «откладывать на оси абсцисс не сами величины численности, биомассы и т. п., а их логарифмы». Он считал, что результатом такого преобразования станет логнормальное распределение «при котором нормально распределены не сами данные, а их логарифмы». При логнормальном распределении мерой положения, как известно, служит среднее геометрическое:

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^n x_i} \text{ или } \bar{x}_G = a^{\left(\frac{\sum_{i=1}^n \log_a(x_i)}{n} \right)}, \quad (2)$$

где $\bar{x}_G$ – среднее геометрическое, x_i – i -ый член выборки, n – объём выборки, a – основание логарифма.

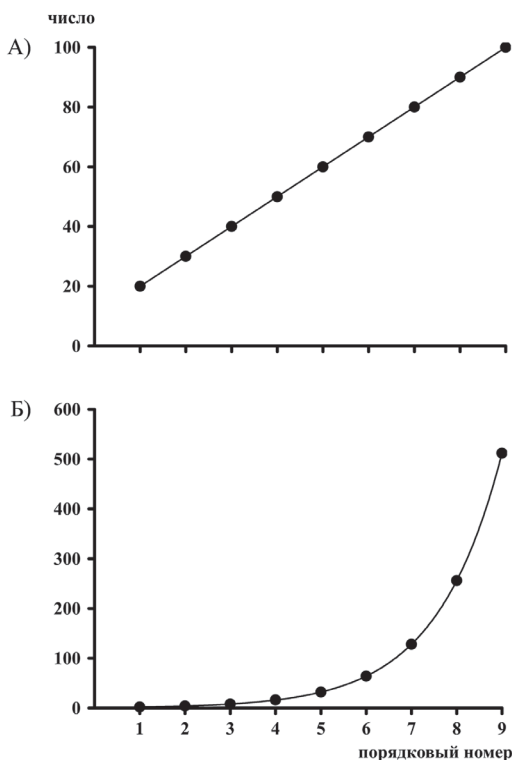


Рис. 2. Примеры набора данных: А – со строго аддитивной структурой (простая арифметическая прогрессия); Б – со строго мультипликативной структурой (геометрическая прогрессия)

В статье об использовании среднего геометрического для оценки распределения гидробионтов К.Н. Несис (1969) подчеркивал, что «Применение средней геометрической необходимо и целесообразно только при логнормальном распределении значений численности, биомассы, частоты встречаемости и других показателей». Однако статья завершалась словами: «Логнормальное распределение численностей, биомасс, частот встречаемости и других показателей, по-видимому, является обычным случаем, если не правилом для совокупностей водных животных. Поэтому практика логарифмического преобразования и расчета средней геометрической должна войти в арсенал математических приемов каждого гидробиолога и ихтиолога». Такое обобщение повлекло за собой критику, причем не всегда однозначную. Так, М.М. Сметанин и А.И. Баканов (1980), предвзяв свою фразу «нам представляется совершенно неверным использовать [среднюю геометрическую] для оценки средней численности или биомассы организмов в случае их ненормального распределения», написали следующее: «распределение данных по биомассе и численности организмов редко бывает нормальным... Поэтому при обработке подобных материалов, если количество проб не позволяет установить точно закон распределения, можно рекомендовать *a priori* принять гипотезу о логнормальном распределении... В этом случае по нормальному закону распределяются логарифмы исходных величин... Это как раз тот случай, когда используется средняя геометрическая».

В действительности же из того, что плотность распределения логарифмов исходных данных бывает близка к нормальной кривой, вовсе не обязательно следует то, что распределение логнормальное. Помимо логнормального, есть другие распределения, которые «при построении на логарифмической («геометрической») основе» «нормализуются и достаточно хорошо описываются кривой Гаусса» (Preston, 1981). Более того, вышеупомянутая правосторонняя асимметрия встречается не только при логнормальном распределении, но также при распределении χ^2 , гамма, экспоненциальном, Вейбулла, геометрическом и отрицательном биномиальном. Собственно, именно последнее считается наиболее полезным при анализе распределения пресноводной фауны (Elliott, 1977). Повод «утверждать, что имеются несомненные показания к применению средней геометрической» (Каминский, 1964) появляется, когда «одни величины во столько же раз больше, во сколько другие меньше какой-либо одной основной величины; тогда при перемножении всех величин противоположные отклонения взаимно уничтожаются и корень соответствующей степени из произведения вариантов, т. е. средняя геометрическая, дает истинное значение основной величины» (Орженцкий, 1921). Более доходчиво такую ситуацию описал К.Б. Уильямс (Williams, 1939), который на протяжении многих лет изучал динамику лёта насекомых: «изменения, происходящие в уловах, имеют геометрическую природу; т. е. улов, в два раза превышающий среднее значение, так же вероятен, как и улов, равный половине среднего значения; улов, в четыре раза превышающий среднее значение, имеет такую же вероятность, как и улов, равный одной четверти среднего значения, и т. д. В этих обстоятельствах арифметические средние дают ошибочные результаты, и следует использовать средние геометрические».

Проиллюстрируем это мнение таким примером. Допустим, что в рамках исследования активности донных беспозвоночных в гипорейной зоне изучали интенсивность их горизонтальных миграций в грунтах разной плотности. Одной из задач было определение средних темпов перемещения животных в промежутках между элементами донных субстратов неодинаковой крупности. В условиях

эксперимента аналогами таких промежутков служили прорези разной ширины и конфигурации, выполненные в непрозрачных перегородках. В каждом случае количество прорезей было разным, они располагались на разном расстоянии друг от друга и от краёв перегородок. Этими перегородками разграничивали каждый из пяти экспериментальных аквариумов на два участка: «впускной» и «аттрактивный». В первом участке не было ничего кроме воды, туда беспозвоночных впускали вначале опыта. Второй участок был затенён, в нём находилась пищевая приманка. Уровень затенения (освещённости) этого участка в каждом аквариуме регулировали. Модельными животными служили одноразмерные взрослые амфиподы *Gammarus* sp. (Gammaridae). Для лучшего контроля степени возможных отклонений во всех опытах использовали выборку ($n = 100$) одних и тех же особей. Началом эксперимента считали момент помещения всех животных во «впускной» участок. При удачном сочетании ширины прорезей и их расположения в перегородке многие гаммарусы проникали в «аттрактивный» участок уже после нескольких минут обследования «впускного». Эксперимент считался окончанным после того, как во «впускном» участке не оставалось ни одной особи. Соответствующий осмотр этого участка проводили каждые 5 минут. Оценим результаты одной из серий экспериментов. В первом аквариуме все 100 особей переместились в «аттрактивный» участок в течение первых 25 мин (средний неокруглённый показатель интенсивности перемещения 4 экз./мин). Во втором аквариуме это произошло за 10 мин (10 экз./мин); в третьем – за 40 мин (2,5 экз./мин); в четвёртом – за 85 мин (1,18 экз./мин); и, наконец, в пятом – за 120 мин (0,83 экз./мин). Если нужно определить среднюю интенсивность перемещения при разных условиях и среднее время перемещения группы, то по формуле среднего арифметического будет получено соответственно 3,7 экз./мин и 56 мин. Естественно полагать, что перемножение этих числовых показателей должно дать величину равную или близкую исходному количеству выпущенных амфипод (100 экз.). Однако результат такого перемножения составит 207 экз., т. е. окажется вдвое большим. Если же применить формулу среднего геометрического, то получится соответственно 2,5 экз./мин и 40 мин, а произведение этих средних будет величиной равной исходной (100 экз.).

Из этих результатов видно, что для обобщения показателей, изменяющихся во времени, более приемлемым иногда оказывается не среднее арифметическое, а среднее геометрическое. Нередко соответствующие условия складываются в ходе исследований суточной интенсивности дрейфа донных беспозвоночных, ската молоди рыб, а также вылета субимаго и имаго амфибиотических насекомых. Разъясним это на конкретном примере. Необходимо выяснить к какому из двух периодов начала ночи приурочен максимум дрейфа личинок хирономиды *Eukiefferiella* sp. (Orthocladinae): к условному периоду «22 ч» или к условному периоду «23 ч»? На протяжении пяти суток в каждый из этих периодов отбирали пробу дрейфа. Всего было отобрано пять проб в период «22 ч» (8, 6, 7, 6, 6 экз.) и пять проб в период «23 ч» (35, 4, 1, 2, 1 экз.). Видно, что максимальный показатель периода «22 ч» (8 экз.) является величиной одного порядка с другими показателями периода «22 ч», неокруглённое среднее арифметическое которых (без этого максимального показателя) равно 6,3 экз. Наибольший же показатель периода «23 ч» (35 экз.) – величина явно экстремальная по сравнению с другими показателями этого периода, неокруглённое среднее арифметическое которых равно 2,0 экз. Если оба максимальных показателя исключить, то можно уверенно сказать, что максимум дрейфа личинок хирономиды *Eukiefferiella* sp. приурочен к периоду

«22 ч», поскольку среднее арифметическое большей части его показателей (четырёх из пяти) более чем втрое превышает среднее арифметическое четырёх из пяти показателей периода «23 ч». Учёт всех пяти показателей периода «22 ч» даст среднее арифметическое, равное 6,6 экз. и среднее геометрическое, тоже равное 6,6 экз. Обе эти величины близки к «типичным» для этого периода 6,3 экз. (среднее арифметическое большей части показателей периода). При учёте всех пяти показателей периода «23 ч» среднее арифметическое окажется равным 8,6 экз. Такой результат может стать причиной ошибочного вывода, будто максимум дрифта личинок хирономиды *Eukiefferiella* sp. приурочен к периоду «23 ч». Среднее же геометрическое этих пяти показателей равно 3,1 экз. Т.е. несильно отличается от «типичных» для данного периода 2,0 экз. (среднее арифметическое большей части показателей периода «23 ч»). Но важно то, что среднее геометрическое (3,1 экз.) всех пяти показателей периода «23 ч» (т. е. даже с учётом экстремального значения), всё равно меньше среднего геометрического (6,6 экз.) всех пяти показателей периода «22 ч». Это согласуется с фактом, что максимум дрифта личинок хирономиды *Eukiefferiella* sp. приурочен именно к периоду «22 ч».

Таким образом, при изучении суточной активности беспозвоночных с учётом единичных экстремально больших показателей результаты обобщения данных посредством среднего геометрического более точно (и надёжно) передают истинную тенденцию изменений в ряду динамики. Отметим, что если в примере с амфиподами выбор среднего геометрического был обусловлен приблизительно мультипликативными отношениями между результатами эксперимента, то здесь определяющую роль сыграла слабая чувствительность этой меры положения к экстремально большим значениям.

Недостатком среднего геометрического является невозможность расчёта этого показателя для рядов, включающих хотя бы одно нулевое значение. В таких случаях препятствием для вычисления среднего геометрического через корень n -й степени из произведения n значений является то, что умножение любого числа на ноль даёт ноль, а через логарифмы – то, что логарифм нуля не определён. Между тем нули в первичных данных по суточной динамике дрифта, ската или вылета – явление довольно обычное. Решение проблемы нулевых значений нашёл К.Б. Уильямс, который преобразовал формулу среднего геометрического путём прибавления единицы к каждому члену выборки при последующем вычитании её из конечного результата (Williams, 1937):

$$M_w = a^{\left(\frac{\sum_{i=1}^n \log_a(x_i + 1)}{n} \right)} - 1, \quad (3)$$

где M_w – среднее геометрическое Уильямса, x_i – i -ый член выборки, n – объём выборки, a – основание логарифма.

Другой мерой положения, которую нельзя использовать для обобщения наборов данных с нулевыми значениями, является среднее гармоническое:

$$\overline{x_H} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}, \quad (4)$$

где $\overline{x_H}$ – среднее гармоническое, x_i – i -ый член выборки, n – объём выборки.

Из формулы видно, что «среднее гармоническое – это величина, обратная среднему арифметическому обратных величин данных» (Zar, 2010). Обратными

величинами (или мультипликативными обратными), как известно, называют два числа, которые при умножении дают 1. Например, $\frac{1}{3}$ и 3. Невозможность применения формулы среднего гармонического к наборам данных с нулевыми значениями обусловлена тем, что ноль обратной величины не имеет.

Так же, как и среднее геометрическое, среднее гармоническое мало (точнее, в ещё меньшей степени) чувствительно к выбросам, но зато ещё больше подвержено влиянию наименьших показателей выборки. Поэтому, если использовать формулу среднего гармонического для обобщения первичных («сырых») данных, то оценка «центра» распределения может оказаться заниженной не только в случае совокупностей с аддитивной структурой (когда желательно использовать среднее арифметическое), но также при доминировании мультипликативных связей (когда наиболее точно положение «центра» отражает среднее геометрическое).

Чем же полезно обращение в дробь, заложенное в формуле среднего гармонического? Среднее гармоническое «необходимо тогда, когда наблюдения» представляют собой «обратную зависимость» (Закс, 1976). Т.е. «при усреднении индексов, выражающих, в частности, число устьиц на 1 см² или величин концентрации, например i мг на 1 л, или других подобных отношений» (Зайцев, 1990). Кроме того, «гармоническое среднее используется при расчете средней продолжительности жизни» (Закс, 1976). Классическим примером применения среднего гармонического служат адаптированные к некоторым областям знания варианты определения средней скорости движения при разном количестве времени, затраченном на преодоление одного и того же расстояния. Так, вариантом из области биологии является расчёт средней скорости перелёта птиц от места ночёвки к месту кормёжки и обратно (Zar, 2010). По условию задачи, «стая птиц летит от места ночлега к месту кормёжки, расположенному на расстоянии 20 км, со скоростью 40 км/ч (что занимает 0,5 ч). Стая возвращается к месту ночлега тем же маршрутом (20 км), летя со скоростью 20 км/ч (на что требуется 1 ч лётного времени)». Если вычислить «среднюю скорость» как среднее арифметическое $((40 \text{ км/ч} + 20 \text{ км/ч}) / 2)$, то будет получен показатель 30 км/ч. Но этот ответ неверен, поскольку в целом на преодоление 40 км птицами было затрачено 1,5 ч, значит, среднюю скорость полёта следует рассчитать как $((40 \text{ км}) / (1,5 \text{ ч}))$. Результат составит 26,7 км/ч. Этот же корректный результат можно получить, если к данным по скорости перелёта сразу применить формулу среднего гармонического (Zar, 2010).

Пример использования среднего гармонического, затрагивающий область гидробиологических исследований, сгенерирован автором на основе примера из руководства Г.Н. Зайцева (Зайцев, 1990, стр. 20). Для лучшего восприятия материала стоит вспомнить, что в качестве меры удельной численности (плотности распределения) животных принято использовать не итоги «сырого» подсчёта особей (экз.), а показатели, уже представляющие собой отношение (экз./м²), т. е. индексы. С этих позиций рассмотрим такую задачу. Перед началом эксперимента в пять водоёмов разной площади интродуцировали по 100 шаровок (*Bivalvia: Sphaeriidae*). Через некоторое время, с целью контроля, в каждом водоёме с учётной площади 1 м² провели разовый количественный отбор этих моллюсков и получили следующие показатели удельной численности: 6, 11, 5, 16, 12 экз./м². Какова средняя для этих водоёмов плотность распределения шаровок? На первый взгляд, верный ответ даст вычисление среднего арифметического: $(6 + 11 + 5 + 16 + 12) / 5 = 10 \text{ экз./м}^2$. Однако проверка с учётом общей численности интродуцированных особей показывает, что это не так. По условию, в каждый из пяти водоёмов было внесено по 100 шаровок. Используя результаты контрольного отбора моллюсков, рассчитаем общую

площадь исследованных водоёмов: $(100 / 6) + (100 / 11) + (100 / 5) + (100 / 16) + (100 / 12) = 60,3 \text{ м}^2$. Если для определения общего количества шаровок, расселённых на этой площади применить рассчитанное выше среднее арифметическое (10 экз./м^2), то результат окажется сильно завышенным: $60,3 \times 10 = 603$ шаровки, хотя из условия очевидно, что расселено было только 500 моллюсков. Зато, если мы воспользуемся средним гармоническим ($8,29 \text{ экз./м}^2$), то результат будет близок к истинной величине: $60,3 \times 8,29 = 499,9 \approx 500$ особей. Предпочтительность среднего гармонического здесь наглядно подтверждается информацией о точном размере исследованной «генеральной» совокупности шаровок (500 экз.). Подтвердить правильность выбора среднего гармонического для обобщения таких относительных показателей как удельная численность или удельная биомасса на основе материала, собранного в естественных условиях, т. е. когда размер генеральной совокупности не определён, способа нет.

Заключение

Расчёты, представленные в настоящем мини-обзоре, ясно свидетельствуют, что привычный безальтернативный выбор среднего арифметического в качестве меры центральной тенденции не всегда безупречен. Иногда вычисление среднего арифметического становится причиной ошибочных заключений. При выборе меры положения следует учитывать как характер данных («сырые» они или уже являются формой некоторых отношений (например, коэффициенты, индексы)), так и особенности изучаемого явления (например, при исследовании временных рядов часто нельзя уловить реальную закономерность без учёта экстремальных «выпадающих» значений). В случаях слишком малых числовых наборов расчёт средней величины порой неуместен; достаточно ограничиться указанием минимального и максимального значения. Нужно подчеркнуть, что в настоящем мини-обзоре были приведены примеры определения только простых (невзвешенных) форм средних, вычисление которых не требует группировки данных. Такая позиция обусловлена, с одной стороны, соображением, что специфика практического применения взвешенных средних заслуживает проведения отдельного исследования, а с другой – необходимостью придерживаться рамок допустимого объёма рукописи.

Благодарности

Работа А.В. Астахова выполнена при поддержке государственного задания Министерства науки и высшего образования Российской Федерации, тема № 124012400285-7.

Литература

- Гаусс К.Ф. 1957. Избранные геодезические сочинения. Т. 1. Способ наименьших квадратов /Theoria motus corporum coelestium in sectionibus conicis Solem ambientium, Hamburgi, 1809; Göttingische gelehrte Anzeigen, 1821/. М.: Издательство геодезической литературы. 152 с.
- Дмитриев Е.А. 2009. Математическая статистика в почвоведении. М.: Книжный дом «ЛИБРОКОМ». 328 с.
- Зайцев Г.Н. 1990. Математика в экспериментальной ботанике. М.: Наука. 296 с.
- Закс Л. 1976. Статистическое оценивание. М.: Статистика. 598 с.
- Калинина В.Н., Панкин В.Ф. 2002. Математическая статистика. М.: Дрофа. 336 с.
- Каминский Л.С. 1964. О применении геометрической средней в биологии // Применение математических методов в биологии. Сб. 3. Л.: ЛГУ. С. 6–18.
- Лакин Г.Ф. 1990. Биометрия. М.: Высшая школа. 350 с.
- Ланг Т.А., Сесик М. 2011. Как описывать статистику в медицине. Аннотированное руководство для авторов, редакторов и рецензентов. М.: Практическая медицина. 480 с.
- Несис К.Н. 1969. Применение геометрической средней при изучении распределения водных организмов // Труды ВНИРО. Т. 65. С. 304–309.

- Орженцкий Р.М. 1921.** Элементарная теория статистических величин и вычислений. Киевское губернское статистическое бюро. 172 с.
- Реброва О.Ю. 2002.** Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. М.: МедиаСфера. 312 с.
- Рокицкий П.Ф. 1973.** Биологическая статистика. Минск: Вышэйшая школа. 320 с.
- Ряузов Н.Н. 1971.** Общая теория статистики. М.: Статистика. 368 с.
- Сметанин М.М., Гордеев Н.А. 1977.** К оценке объема выборки при изучении роста рыб // Биология внутренних вод. Информационный бюллетень N 34. Л.: Наука. С. 79–83.
- Сметанин М.М., Баканов А.И. 1980.** Об использовании средней арифметической и средней геометрической в гидробиологических и ихтиологических исследованиях // Биология внутренних вод. Информационный бюллетень N 48. Л.: Наука. С. 74–78.
- Цивинский Д.Н. 2011.** Оценка математического ожидания выборки. Самара: СамГТУ. 39 с.
- Чернышевский Н.Г. 1975.** Что делать? Л.: Наука. 872 с.
- Alexander N. 2012.** Review: analysis of parasite and other skewed counts // Tropical Medicine and International Health. V. 17, N 6. P. 684–693.
- Blumenshine S.C., Tsukimura B., Rice A., Rudnick D.A. 2012.** Environmental factors influencing the dynamics of Chinese mitten crab zoeae in the San Francisco Bay-Delta // Aquatic Invasions. V. 7, N 1. P. 111–124.
- Chutter F.M. 1972.** A reappraisal of Needham and Usinger's data on the variability of a stream fauna when sampled with a Surber sampler // Limnology and Oceanography. Vol. 17, N 1. P. 139–141.
- Chutter F.M., Noble R.G. 1966.** Reliability of a method of sampling stream invertebrates // Archiv für Hydrobiologie. Bd. 62, H 1. S. 95–103.
- Dixon W.J. 1960.** Simplified estimation from censored normal samples // The Annals of Mathematical Statistics. V. 31, N 2. P. 385–391.
- Elliott J.M. 1977.** Some methods for the statistical analysis of samples of benthic invertebrates. The Ferry House Ambleside, Westmorland: Freshwater Biological Association, Scientific Publication N 25. 159 p.
- Gregoire T.G., Schabenberger O. 1999.** Sampling-skewed biological populations: behavior of confidence intervals for the population total // Ecology. V. 80, N 3. P. 1056–1065.
- Hintze J.L. 2007.** NCSS user's guide II: descriptive statistics, means, quality control, and design of experiments. Kaysville, Utah. 659 p.
- Huber P.J. 1972.** The 1972 Wald Lecture. Robust statistics: a review // The Annals of Mathematical Statistics. V. 43, N 4. P. 1041–1067.
- James J. 2022.** Ecological responses to climate change: using the common ragworm (*Hediste diversicolor*) as an indicator for benthic ecosystems. University of Hull. 256 p.
- Oliver D., Hall J.C. 1989.** Usage of statistics in the surgical literature and the 'orphan p' phenomenon // Australian and New Zealand Journal of Surgery. V. 59, N 6. P. 449–451.
- Pousse E., Poach M.E., Redman D.H., Sennefelder G., Hubbard W., Osborne K., Munroe D., Hart D., Hennen D., Dixon M.S., Li Y., Milke L.M., Wikfors G.H., Meseck S.L. 2023.** Juvenile Atlantic sea scallop, *Placopecten magellanicus*, energetic response to increased carbon dioxide and temperature changes // PLOS Climate. V. 2, N 2. Article e0000142.
- Prati S., Henriksen E.H., Knudsen R., Amundsen P.-A. 2020.** Seasonal dietary shifts enhance parasite transmission to lake salmonids during ice cover // Ecology and Evolution. V. 10, N 9. P. 4031–4043.
- Preston F.W. 1981.** Pseudo-lognormal distributions // Ecology. V. 62, N 2. P. 355–364.
- Williams C.B. 1937.** The use of logarithms in the interpretation of certain entomological problems // Annals of Applied Biology. V. 24, N 2. P. 404–414.
- Williams C.B. 1939.** An analysis of four years captures of insects in a light trap. Part I. General survey; sex proportion; phenology; and time of flight // Transactions of the Royal Entomological Society of London. V. 89, N 6. P. 79–131.
- Yan Y., Cantoni E., Field C., Treble M., Flemming J.M. 2022.** Spatiotemporal modeling of bycatch data: methods and a practical guide through a case study in a Canadian Arctic fishery // Canadian Journal of Fisheries and Aquatic Sciences. V. 79, N 1. P. 148–158.
- Zar J.H. 2010.** Biostatistical analysis. Upper Saddle River, N.J.: Prentice Hall. 944 p.