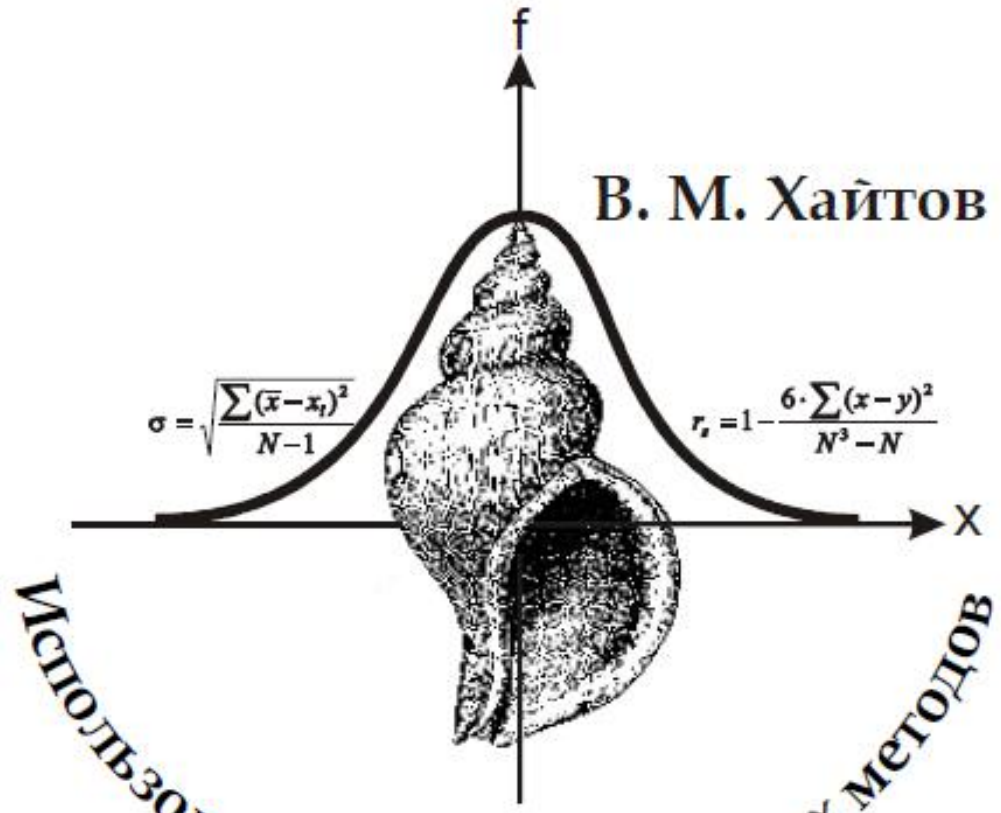

В. М. Хайтов

Использование математических методов в биологических исследованиях школьников

Санкт-Петербург
2005

От автора

В современной литературе имеется довольно много книг и пособий по статистике и биометрии. Я решил написать еще одну. Нужна ли она? На мой взгляд, да. Это продиктовано тем, что в практике работы со школьниками многие книги абсолютно непригодны, поскольку рассчитаны на студентов, освоивших курс высшей математики. Существует и другая серия, более простых книг, например, книги П. Ф. Рокицкого (1967) или Э. В. Ивантера и А. В. Коросова (1992), написанные простым и понятным языком. При некоторых усилиях со стороны преподавателя эти тексты становятся доступными и для школьников. Однако все эти издания - библиографическая редкость. Поэтому я отважился на написание такого труда.

Не скрою, данная брошюра написана по мотивам уже существующих изданий. Однако я решил отказаться от принятой в них строгости в пользу понятности и доступности для школьников 8-9 классов. Я старался построить пособие так, чтобы его мог освоить любой учащийся, знакомый с азами курса алгебры средней школы. Поэтому здесь вы не найдете ни выводов формул, ни доказательств теорем. Однако, имея в руках материал, собранный в экспедициях или экспериментах, учащиеся смогут выбрать метод, нужный для его обработки.

Вместе с тем, слепое применение формул без знания основ статистики, на мой взгляд, – вещь совершенно недопустимая. Поэтому в данном пособии приведена глава «Трудная, но необходимая теория», которая может показаться некоторым школьникам трудной для понимания. Увы, я не знаю, как объяснить этот материал проще. Поэтому данная глава написана предельно кратко. Это позволит прочитать ее несколько раз, что должно помочь понять содержащиеся в ней сведения.

Автор этой работы – не математик и не биометрик, а биолог, практикующий как в области собственно биологии, так и в области работы со школьниками, поэтому многие подводные камни в объяснении тех или иных методов мне известны не понаслышке. Вместе с тем, написание подобных математических пособий дело для автора совершенно новое, что, возможно (и вероятно), сказалось на слоге и несовершенстве дидактики. Поэтому я очень рассчитываю на отклики с вашей стороны и очень хочу услышать критические замечания.

Я искренне признателен д. б. н. Н. С. Ростовой за редакторскую правку, критические замечания и огромное терпение, которое она демонстрировала в то время, когда я отрывал ее от других дел. Большую роль в разработке способов изложения материала, примененных в этом пособии, сыграли лекции и консультации Е. А. Нинбурга. Особая благодарность Ксении Шункиной за помощь в работе над текстом.

Вступительное слово для преподавателей, которое могут прочитать и школьники

В Санкт-Петербурге сложилась уникальная для России традиция проведения школьниками самостоятельных исследовательских работ по биологии и смежным дисциплинам. Первостепенную роль в формировании этой традиции сыграла городская биологическая олимпиада, на которой исследовательская работа оказывается своеобразным пропуском, позволяющим школьнику принять участие в соревновании. Другой источник формирования такой традиции – это регулярно проводимые в Санкт-Петербурге разнообразные конференции, фестивали и конкурсы, на которых начинающие исследователи представляют свои работы и имеют возможность познакомиться с исследованиями своих коллег. Ежегодно в олимпиадах, конференциях и конкурсах принимают участие более тысячи школьников со своими самостоятельными изысканиями.

Большой опыт проводимых мероприятий позволил заметить как положительные, так и отрицательные стороны исследовательской деятельности школьников, как особого

рода педагогической технологии. На плюсах останавливаться нет необходимости, они очевидны, в противном случае педагоги не занимались бы руководством исследовательскими работами. Гораздо интереснее разобраться в отрицательном опыте.

Практически любая научная работа учащегося, оформленная и представленная в окончательном виде, будь то доклад, постер, переплетенная рукопись или публикация – все это результат «симбиоза» школьника и руководителя. Мне не известны случаи, когда бы сколько-нибудь приемлемое по качеству исследование было проведено школьником на пустом месте, без какого-то руководства со стороны взрослых. Поэтому будем рассматривать оконченную работу как результат совместной работы двух людей – школьника и его руководителя. Степень самостоятельности ребенка при проведении работы может быть самой разной. От практически полного непонимания того, что от него хочет руководитель, до практически полной самостоятельности юного исследователя, когда роль руководителя заключается лишь в постановке проблемы и редакции окончательного текста. Подавляющее большинство работ сосредоточено в средней части этого градиента.

Неудачи большинства работ на тех или иных соревнованиях и конкурсах обычно обусловлены неправильно построенными взаимоотношениями в системе «школьник–руководитель». В подавляющем большинстве случаев ответственность за те или иные провалы наших ребят ложится на нас, руководителей. К сожалению, очень часта ситуация, когда в основе неверных представлений учащихся лежат ошибочные взгляды педагогов. Так, например, на вопрос, что такое ошибка среднего, учащиеся (а значит и учителя) отвечают, что это погрешность измерения. Это абсолютно неправильно! Если ученик так считает, то он совершенно не понимает того, что сделал в своей работе. Значит, в этом виноваты мы – педагоги. Поэтому очень важно, сделать с нашей стороны все, чтобы наши воспитанники могли потерять свои баллы только благодаря своим личным особенностям, а не за счет того, что в их представлениях об исследовательской работе зияла прореха, которую не заделали мы. Данная брошюра – это пособие и для учителей, и для школьников. Очень надеюсь, что изложенный в ней материал позволит избежать тех неприятностей, о которых говорилось выше.

Одной из самых распространенных причин, которые приводят к снижению оценок наших воспитанников на конкурсах, оказывается неправильное применение тех или иных методов сбора и обработки материала при проведении биологических исследований. В проверке и оценке работ на олимпиадах и конкурсах принимают участие, главным образом, профессиональные биологи-исследователи. А у таких людей особенно остро выражено неприятие небрежного отношения к методологии исследования. Кроме того, в последнее время, в связи с ростом количества работ, связанных с теми или иными природоохранными проектами, становится заметен и еще один источник провалов – это полное игнорирование общепринятых методов регистрации и фиксации научных фактов. Более того, авторы подобных работ (а это значит, в первую очередь, их руководители) зачастую попросту не задумываются о том, что только безукоризненно собранные и оформленные данные будут приняты как доказательства неблагоприятного экологического состояния объекта. Чиновники, от которых зависит что-либо, – люди грамотные, и они не будут тратить средства на спасение той или иной экосистемы, если в качестве доказательства ее неблагополучия они получат лишь эмоциональные фразы, тем более от детей.

Все изложенные выше соображения делают крайне актуальным создание краткого пособия по методам обработки материала, полученного в биологических исследованиях. Вместе с тем, биология – наука обширная, и надо четко определить, о каких методах пойдет речь. Если несколько упростить ситуацию, то можно выделить всего три типа данных, с которыми работают биологи:

1. Словесное представление данных - записи наблюдений,
2. Графическое представление данных – рисунки,

3. Численные представления данных.

Первый тип представления наблюдений, как правило, оказывается предварительным. Все словесные описания, в конечном итоге, должны быть представлены в виде численных характеристик или в виде каких-то формализованных схем.

Второй тип данных – рисунок – оказывается основным для морфологических работ, доля которых в исследовательском творчестве школьников минимальна.

Данные третьего типа – числовые данные (результаты учетов, измерений, взвешиваний и т. п.) – наиболее часто лежат в основе самостоятельных работ школьников. Методам работы именно с этим типом данных и посвящено это пособие.

Вступительное слово для школьников

В предыдущей главе мы постарались показать, что самостоятельная работа школьника – это результат «симбиоза» преподавателя и учащегося. При этом очень важно, чтобы такое взаимодействие не приобретало черты системы «паразит–хозяин». Опыт показывает, что школьники зачастую используют знания преподавателя без взаимной отдачи. Они выполняют только то, что им учитель скажет – от и до, ни шагу дальше. Такое поведение, как правило, означает, что ученик освоил алгоритм, запомнил, как нажимать на кнопочки калькулятора или компьютера, но не понял, что за этим кроется. Главное же, что может дать школьник своему учителю, – это глубоко понять и освоить то, чему учитель хочет его научить. Единственным критерием перехода к такому взаимовыгодному взаимодействию является самостоятельный, без указки, выбор той или иной методики работы и творческое осмысление полученных результатов. К сожалению, человечество не придумало пока никакого другого метода для достижения этой цели кроме кропотливого повторения пройденного и безбоязненного переспрашивания непонятого.

Эта брошюра для вас, уважаемые юные исследователи. Многим она покажется трудной. А кто сказал, что заниматься наукой (а именно этим вы хотите заниматься, раз читаете данную работу) легко? Ученые не зря едят свой хлеб. И единственный путь попасть в удивительный и непростой мир науки – это, не стесняясь спрашивать своих учителей, внимательно изучать предложенную литературу, а главное, помнить, что в науке можно только научитьСЯ, то есть научить СЕБЯ.

Вступительное слово для любителей компьютерных технологий

Современное научное исследование немыслимо без применений компьютеров. Ученому компьютеры приносят огромную пользу и… не меньший вред. Особенно вредны компьютеры для начинающего исследователя. Почему? Дело в том, что они зашоривают его взгляд. Вместо понимания сути работы того или иного метода анализа, зачастую, авторы демонстрируют знание особенностей работы тех или иных программ. Так, например, часто приходится в ответ на вопрос о том, каким методом был проведен тот или иной анализ, слышать ответ, что так посчитал компьютер. Более того, внедрение вычислительной техники привело к тому, что стали плодиться работы, вся суть которых сводится лишь к массовым обсчетам каких-то числовых данных без серьезного обсуждения того, что же, собственно, получилось с точки зрения биологии.

Бесспорно, компьютерными программами надо уметь пользоваться (очень рекомендую всем юным исследователям освоить Microsoft Excel и Statistica for Windows). Однако все эти умения совершенно лишены смысла, если вы не понимаете в чем суть метода, который заложен в основу той или иной программы. Очень важно, особенно в юности, методы, которые лежат в основе работы тех или компьютерных пакетов, «пощупать руками», пропустить через себя формулы. Поэтому на первых порах (если вы действительно хотите понять, как работают методы исследования) забудьте о компьютерах и приготовьтесь к долгому и муторному счету вручную, в лучшем случае – к расчетам на калькуляторе.

ЧАСТЬ 1. Трудная, но необходимая теория

Материал, изложенный в этой главе, достаточно сложен. В принципе, эту часть пособия можно не читать или прочитать после того, как были изучены другие главы. Однако если вы не освоите всех изложенных в этой главе идей, то ваша работа будет механической, вы не поймете, что же получилось в результате вашего исследования и можно ли полученным результатам доверять. Поэтому я крайне рекомендую рано или поздно эту главу прочитать.

Глава 1.1. Зачем нужны математические методы в биологии

Многие науки прошли путь от чисто описательных к математическим методам. Так было и с физикой, и с химией, и с экономикой. Биология сейчас тоже стремительно движется к тому, чтобы стать теоретической наукой. По сути дела, некоторые биологические дисциплины (генетика, экология и др.) уже перешли эту грань и давно оперируют не словом, но числом. Поэтому начинающий биолог-исследователь должен как можно раньше втягиваться в мир математических методов. Для чего же они используются в биологии?

На мой взгляд, можно выделить три основных области применения математических методов.

Первая область – это **моделирование**. Математическая модель призвана имитировать поведение параметров биологических систем в заданных условиях. Бесспорно, биологи стремятся к построению моделей для как можно большего количества процессов и явлений. Однако сами эти процессы и явления зачастую могут быть обнаружены только после серьезной аналитической части работы. Поэтому, вторая область применения математики в биологии – это **анализ** разнообразных явлений. Многие явления вообще становятся видны только тогда, когда они пропущены через призму математики. Так, например, без математических методов, зачастую, совершенно невозможно увидеть взаимосвязь каких-то процессов. Однако биологические системы изменчивы, нередко исследователь думает, что он увидел нечто, но его коллеги с ним могут не согласиться. В связи с этим, третья область – это **доказательство** с помощью математических методов наличия тех или иных закономерностей.

В данном пособии мы коснемся лишь третьей и отчасти второй задачи. Поскольку именно они наиболее часто появляются в исследовательских работах школьников.

Глава 1.2. О способах выражения признаков

Биолог в своих исследованиях всегда анализирует некоторые признаки объектов. Он может изучать форму или окраску плодов, численность животных на определенной территории, видовой состав биоценозов, частоту встречаемости какого-либо свойства организмов в популяции. Это признаки тех систем, с которыми он работает. В работе с ними очень важно четко представлять себе то, как эти признаки должны быть выражены в материалах, которые далее лягут в основу исследования.

Можно придумать огромное количество способов представления данных. Например, окраску цветка можно охарактеризовать так: розовый, ярко красный, темно красный. Далее с этими описаниями можно работать, даже существуют некоторые математические методы, которые позволяют провести вполне грамотное исследование на базе такого материала. Однако если аналогичные изыскания захотят провести другие люди, то им будет крайне сложно сопоставить свои данные с вашими. Они ведь могут иметь другие представления о том, что является «ярко красным» или «розовым». Поэтому крайне желательно применять такие способы выражения признаков, которые могут быть воспроизведены другими исследователями. Самым лучшим способом будет выражение признаков какими-то численными, измеряемыми величинами. В случае с окраской

цветков это могут быть длины волн отраженного света. Для исследований биоценозов лучше, чтобы это были не выражения типа «много» или «мало», а реальная численность особей того или иного вида на определенной площади. Анализ формы чего-либо тоже можно проводить с помощью численных методов (правда, они в большинстве своем достаточно сложны).

Понятно, что далеко не всегда можно дать строгую численную оценку признака, однако к этому надо стремиться. Степень объективности исследования (а, стало быть, и его качество) во многом определяется тем, насколько в нем используются числовые данные.

Глава 1.3. Типы числовых данных

Несмотря на огромное количество разнообразных возможностей количественного описания материала, можно говорить лишь о трех основных типах числовых данных: *количественные данные*, *балловые данные*, *качественные данные*[1].

Исследователь, планируя тот или иной метод сбора данных об изучаемой системе, должен четко представлять себе, какой тип показателей он получит. В противном случае ему придется уже постфактум искать пути преобразования своих данных, что, в конечном итоге, приводит к огрублению и снижению достоверности результатов. Кроме того, каждый тип величин требует своего подхода к обработке. Поэтому ниже мы дадим краткую характеристику этим типам.

Количественные данные. Как правило, это результат измерений, подсчетов, взвешиваний и т.п. Примерами такого рода данных могут служить измерения длины тела животных, площади листьев, площади проективного покрытия растений на учетной площади, подсчет числа особей в пробе, вес овощей, число семян в плодах и т.д. Это самый распространенный тип величин.

Балловые оценки. Эти величины используются, когда вместо реальных измерений используются балловые оценки. Например, если нет возможности измерить тело организма, но есть возможность выделить визуальные классы среди всех изученных особей, то каждому классу может быть присвоен балл. Так, при изучении размеров деревьев в лесу далеко не всегда можно их измерить, поэтому здесь удобнее применять приблизительную оценку (скажем, размер крупных деревьев оценивается в 3 балла, размер средних – в 2, а мелких – в 1 балл). Однако балловые оценки не следует путать с любыми другими цифровыми обозначениями. Например, поведенческие реакции животного можно обозначить номерами: реакция 1, реакция 2 и т.д., но это не будут баллы. Баллы всегда можно ранжировать от меньшего значения к большему.

Этот тип данных наиболее грубый, но, вместе с тем, наиболее гибкий. Именно поэтому ему следует отдавать предпочтение при работе с материалом, о характере которого мало что известно. К сожалению, в самостоятельных работах школьников этот тип используется крайне редко. Многие авторы совершенно неоправданно привлекают количественные данные там, где их использование некорректно.

Качественные данные. Они получаются в результате учета наблюдений по альтернативной схеме: белый-черный, да-нет, присутствует-отсутствует, что может быть обозначено как «+/-», или «1/0». Примеры такого рода достаточно часто встречаются в практике самостоятельных работ школьников. Особенно активно такие оценки используются в гидробиологических и орнитологических исследованиях, когда производится оценка встречаемости каких-то видов на достаточно обширной территории: вид встречен («да», «+», «1») или не встречен («нет», «-», «0»).

---

[1] Далее курсивом будут обозначаться первые упоминания терминов, а жирным шрифтом будут выделяться смысловые ударения.

Глава 1.4. Типы задач, наиболее часто решающиеся в школьных самостоятельных работах

Опыт работы со школьными научно-исследовательскими работами по биологии говорит о том, что вне зависимости от конкретных целей и задач тех или иных работ можно выделить всего 4 типа элементарных задач, которые решаются в ходе исследования[2]. К числу таких задач относятся:

1. Выявление различий между выборочными показателями. Примерами такого рода задач могут быть следующие: сравнение плотности поселения вида в двух разных местообитаниях; сравнение размеров листьев на освещенных и затененных участках; выявление межгодовых изменений встречаемости вида; выявление различий в успеваемости разных групп школьников.
2. Описание структуры популяций (например, размерно-возрастной анализ популяций; фенетико-генетический частотный анализ структуры популяции; выявление отклонений от ожидаемого частотного распределения).
3. Описание взаимосвязи величин и явлений (например, выявление связи между обилием вида и температурой окружающей среды; ответ на вопрос «существует ли взаимосвязь между размером тела и плодовитостью животного?»).
4. Поиск закономерностей многообразия объектов (например, группировка геоботанических описаний при построении карт распространения растительности; классификация видов по их экологическим характеристикам; выявление закономерности варьирования размерно-возрастной структуры популяции).

В дальнейшем мы рассмотрим каждый тип задач и предложим некоторые **типовые**[3] варианты их решения. Однако предварительно необходимо обсудить еще одну важную проблему, а именно – принципы сбора материала, поскольку в зависимости от того, как собран тот или иной материал, следует (или не следует) применять те или иные методы его обработки.

Глава 1.5. Основные принципы сбора материала

Утверждение, что метод сбора материала должен быть адекватен поставленной цели, очевидно. Вместе с тем, при знакомстве с самостоятельными исследованиями школьников приходится сталкиваться с такими примерами, когда под внешним соответствием целей и методов скрываются просчеты, которые заставляют поставить под сомнение достоверность результатов. В основе большинства этих просчетов лежит неправильное использование *выборочного метода* или неправильное понимание его основ.

Выборочный метод основан на том, что некоторое множество объектов, не доступное для полного изучения, описывается на основании изучения лишь незначительного количества таких объектов. Например, изучение веса тела всех рыб, обитающих в том или ином водоеме, невозможно. Это долго и дорого. Поэтому для характеристики популяции вылавливается лишь небольшое количество особей, у которых и измеряется вес. Это и есть *выборка* из *генеральной совокупности*. Мы, изучая выборку, пытаемся судить о свойствах всей генеральной совокупности. Практически все статистические методы основаны на том, что с помощью анализа тех или иных выборочных показателей мы оцениваем показатели генеральной совокупности.

Выборочный метод имеет свои правила. Первое правило: любая выборка должна быть **случайна**. Это означает, что все объекты, попавшие в выборку, должны попадать

---

[2] В данном пособии не рассмотрены методы обработки материалов экспериментов, основное внимание уделено наблюдениям. Однако во многом изложенные здесь методы могут быть применены и к таким экспериментальным исследованиям.

[3] В каждом конкретном случае могут быть разные формулировки задачи, но принцип решения будет более или менее одинаковым.

туда без выбора, пусть даже подсознательного, со стороны исследователя. В упомянутом примере с рыбой было бы неправильно собирать материал с помощью сачка, «прицельно» вылавливая приглянувшуюся рыбешку. В этой ситуации исследователь может подсознательно вылавливать особей, скажем, более крупных. Понятно, что в результате оценка веса особей, обитающих в водоеме, будет сильно завышена. К сожалению, работы, связанные с природоохранными проектами, буквально переполнены примерами нарушения принципа случайности выборки. Находя большое количество организмов с какими-то аномалиями, многие авторы склонны бить тревогу, утверждая, что это результат «неблагоприятной экологии». Вместе с тем, может оказаться (и очень часто так и бывает), что частота такой аномалии в изучаемом месте ничуть не превосходит фоновое значение, просто исследователь, имея изначальную установку на то, что в данном районе что-то не так, будет подсознательно отбирать факты, которые вписываются в его изначальную установку.

Для того, чтобы избежать подобных искажений, необходимо предпринять *рандомизацию*[4] выборки. Методов рандомизации очень много и зависят они от характера конкретного материала. Например, при описании популяции растений (размер колосков у злаков, площадь листьев у деревьев, количество цветков в соцветии и т.п.) площадки, на которых проводятся измерения, должны располагаться в пространстве случайно. Для этого можно воспользоваться стрельбой из лука с завязанными глазами – где стрела упадет, там и проводить измерения. Или такой вариант – расположить все учетные площадки равномерно по территории, занятой популяцией.

Другой пример - при опросе мнения местного населения (этот тип сбора информации в последнее время чрезвычайно широко используется в природоохранных проектах) необходимо каким-то образом избегать личных пристрастий или антипатий в выборе респондента. В противном случае можно получить картину, отражающую мнения только какой-то одной части населения. Для рандомизации выборки здесь можно использовать, например, такой прием - спрашивать каждого десятого встретившегося на улице человека.

Есть и более сложные (но вместе с тем более корректные) методы рандомизации. В некоторых случаях используют таблицу случайных чисел. Например, исследователь поставил задачу описать суточную активность птиц в период насиживания яиц. Он выбрал для наблюдения 10 гнезд. И решил описывать активность птиц каждые 15 минут. В этой ситуации было бы неправильно навещать все десять гнезд, поскольку птицы могут изменить свое поведение в ответ на посещения. Поэтому корректнее было бы посещать часть из этих гнезд (например, три гнезда), выбранных случайно. Этот случайный выбор номера гнезда и поможет сделать таблица случайных чисел.

Второе правило работы с выборкой гласит, что выборка должна быть *репрезентативна*, или *представительна*. Это означает, что выборка должна отражать структуру генеральной совокупности. Например, исследователь поставил перед собой задачу охарактеризовать частоту заражения рыб каким-нибудь паразитом в пруду. При этом он, зная о правиле случайности выборки, ловил рыб сетью, выбирая из нее всех пойманных особей. Однако размер ячеи этой сети он выбрал такой, что в сеть не попадает молодь. Конечно же, в этой ситуации он получит искаженные представления о распространении паразита в популяции хозяина. В данном случае выборка не отражает структуры всей генеральной совокупности. Сбор материала надо планировать так, чтобы в выборку попадали все представленные в генеральной совокупности разновидности объектов и в тех соотношениях, в которых они представлены в ней.

Вместе с тем, о структуре генеральной совокупности мы зачастую ничего не знаем! Значит, выполнить это правило, казалось бы, нельзя. Однако при выполнении правила случайности выборки и при достаточно большом ее объеме (количестве изученных объектов) мы с высокой вероятностью выполним это правило. В связи с этим, важным

[4] Рандомизация – от английского ”random” – случайный.

требованием при работе с выборочным методом оказывается большой объем выборки. При этом, чем более изменчив изучаемый признак, тем больше должен быть объем выборки. К сожалению, многие объекты, после их попадания в выборку, не могут быть возвращены в генеральную совокупность (зачастую они попросту погибают), что иногда оказывается несовместимым с природоохранными намерениями авторов, поэтому объем выборки в таких ситуациях должен быть оптимальным (ни чрезвычайно большим, ни недопустимо маленьким). К сожалению, формально определить требуемый объем выборки практически невозможно[5], здесь нужен опыт.

Напомним, что все сказанное выше, имело отношение к выборочному методу. Вместе с тем, в некоторых исследованиях выборочный метод не применяется. Это происходит потому, что изучается вся генеральная совокупность целиком. Например, учету подвергаются все крупные животные или проводится тотальный опрос всех людей, населяющих ту или иную местность. При подобных исследованиях мы можем **точно** охарактеризовать состояние системы, а, следовательно, нам **не нужны** будут различные ухищрения, которые придумала статистика для обработки выборочных данных. Для таких тотальных исследований все, изложенное ниже, не имеет никакого значения.

Глава 1.6. Некоторые правила чтения формул

Как бы мы ни стремились к тому, чтобы сделать это пособие наиболее доступным для неспециалистов, нам все-таки не удастся избежать формул. Многие их пугаются и, увидев какой-нибудь знак, вроде такого «$\sum$», в ужасе закрывают книгу. Вместе с тем, страшного здесь ничего нет. Давайте научимся читать формулы. Для этого надо освоить следующую таблицу.

Таблица 1. Некоторые непривычные обозначения, используемые в статистических формулах (приведены лишь те обозначения, которые встречаются в данной брошюре)

Обозначение	Пояснение
σ	Греческая буква «сигма»
ν	Греческая буква «ню»
$\sum$	Знак суммы. Он означает, что все величины, стоящие под этим знаком, надо сложить. Например, $\sum x_i$ означает, что надо просуммировать все величины x_i, то есть, если у нас есть четыре числа, то $\sum x_i = x_1+x_2+x_3+x_4$
Нижние индексы – x_i	Обозначают номер числа. Например, в ряду чисел 1, 2, 45, 15, 24 x_3 =45. Иногда применяются двойные индексы, например $x_{i,j}$ это означает номер числа в двумерном множестве, где i – номер столбца, а j – номер строки. В двумерном массиве 1 2 3 2 8 4 2 5 9 $x_{2,3}$=5. Иногда нижние индексы применяются не для обозначения номера числа в массиве, а для обозначения близких по смыслу величин.

[5] На самом деле, существуют некоторые методы, позволяющие определить требуемый объем выборки, однако они исходят из знаний о масштабах варьирования признака в генеральной совокупности. То есть, для их использования необходимы предварительные исследования.

	Например, m_p означает ошибку p, а m_x – ошибку x.
∞	Знак бесконечности. Мы его будем использовать в смысле «очень большой» или «очень много».
[ ]	Квадратные скобки. Обозначают некоторый интервал значений. Например, выражение [1;10] означает, что мы рассматриваем все числа от 1 до 10 включительно.
∈	Знак принадлежности. Выражение X ∈ [1;10] означает, что величина X принадлежит промежутку от 1 до 10 включительно. То есть X может быть любым числом из интервала [1;10].
→	Знак «стремится». Этот знак означает, что величина стремится при некоторых условиях принять какое-то значение. Например, выражение $\sigma_{N\to} \to S$ надо читать так: «сигма стремится к значению S, когда N стремится к бесконечности».
·	Знак умножения.
| a-b |	Знак модуля. Он означает, что величина рассматривается без учета знака. Например, 10- 15=- 5, но | 10-15 | =5.
±	Плюс-минус. В одном выражении кратко записывается два. Например, выражение $x \pm m_x$ означает, что мы одновременно рассматриваем два случая $x+m_x$ и $x-m_x$.
$\sqrt{}$	Знак квадратного корня. Обозначает действие обратное возведению в степень. Например, если $a^2=b$, то $\sqrt{b}=a$.
$log_a(x)$	Знак логарифма числа *x* по основанию *a*. Обозначает следующее действие: если $a^b=x$, то $log_a(x)=b$. То есть логарифм – это степень, в которую надо возвести число *a*, чтобы получить число *x*.
$lg\ (x)$	Знак десятичного логарифма числа *x*, то есть логарифм числа *x* по основанию *10*.

Особо обратим внимание на следующую особенность формул – в разных книгах одни и те же статистические показатели могут обозначаться разными буквами. Кроме того, некоторые формулы после нехитрых преобразований приобретают несколько другой вид. Всего этого бояться не надо! Надо только внимательно разобраться с тем, что вы видите в той или иной работе.

Глава 1.7. Что такое варьирование биологических признаков?

Школьный курс, в котором нас знакомят с основами науки, имеет один существенный недостаток – в качестве примеров наук, по которым проводятся практические работы, требующие вычислений, нам преподают физику и химию. В этих дисциплинах почти все закономерности можно установить точно. Например, есть формулы, согласно которым можно точно вычислить, на какое расстояние улетит предмет, брошенный под данным углом с данной скоростью. В биологии же все не так. Одним из главнейших свойств биологических систем является изменчивость, или, как говорят в статистике, *варьирование*. Если мы хотим охарактеризовать дальность прыжка какого-то животного, даже имеющего вполне определенный набор физических

параметров, то мы заметим, что при прочих равных условиях это животное будет прыгать каждый раз на разные расстояния. Или другой пример. Если вы посадили семена одного вида растений, обладающие одинаковым весом, и условия произрастания семян будут более или менее однородными, то, все равно, проростки будут иметь разную высоту. Это проявление неотъемлемого свойства живого.

Биологическая изменчивость – это не результат погрешностей измерения, а самостоятельное явление, которое надо учитывать и изучать.

Глава 1.8. Что такое вероятность?

Освоить основы статистики неспециалистам, а тем более школьникам, бывает обычно трудно из-за того, что большинство людей плохо себе представляют основы теории вероятности. Поэтому надо пояснить, что же такое вероятность.

Обычно учащиеся 8-9 классов уже знают, что вероятность выпадения орла или решки у брошенной монеты равна 1/2, или вероятность выпадения определенной стороны игральной кости – 1/6, или вероятность вытащить наугад туз пик из колоды карт – 1/36. Это самые простые примеры работы с вероятностями. Однако можно себе представить более сложную задачу, когда вычислить вероятность события так просто невозможно. Например, мы не можем априори вычислить вероятность встречи медведя в лесу. Для определения вероятности этого события необходимы специальные исследования.

Приведенные примеры свидетельствуют, что применение понятия вероятности может быть очень разным. Общим будет то, что чем выше значение вероятности, тем с большей уверенностью мы будем утверждать, что то или иное событие произойдет.

Вероятности обычно измеряют в процентах или долях от единицы. Если вероятность равна 100% (1), то событие происходит с неизбежностью. Например, вероятность того, что брошенный камень упадет на землю, равна 100%[6]. Если же значение вероятности равно 0, то такое событие абсолютно невозможно. Так, например, совершенно невозможно событие, что Земля в данную секунду начнет вращаться в обратную сторону. Это противоречит всем законам физики.

Важно помнить, что если событие имеет вероятность P, то вероятность того, что событие не произойдет равно 100% – P или, если выражать вероятность в долях от единицы, то 1 – P. Например, если вероятность того, что рост произвольно взятого человека попадает в интервал [1,5 м; 2,0 м] равна 80%, то вероятность того, что рост произвольно взятого человека не попадает в этот интервал, равна 20%.

Глава 1.9. Принципы обработки выборок

Основы статистики – науки, занимающейся, главным образом, анализом выборочных данных, достаточно просты. В основе этой науки лежит представление о *распределении величины*. За этими пугающими словами стоит очень простая вещь – **вероятность встретить в генеральной совокупности то или иное значение величины**. Например, вероятность встречи среди людей карлика или гиганта низка, а вероятность встречи человека со средним ростом – значительно выше. Связь между значением анализируемой величины и вероятностью встречи такого значения в генеральной совокупности и называется распределением. Собственно говоря, исследователя, занимающегося изучением какого-либо признака, и интересует то, как выглядит распределение. Если он знает, как связано значение признака и вероятность встречи данного значения, то он может предсказать, как часто он будет встречать в природе объекты, обладающие теми или иными свойствами.

Связь между величинами обычно отражается формулой (функцией) или графиком. Любая функция имеет свои параметры, от значений которых зависит то, какая именно

[6] Строго говоря, это не совсем так. Существует очень низкая вероятность того, что все молекулы камня начнут двигаться в одну сторону, противоположную направлению падения, тогда камень взлетит. Однако вероятность этого события пренебрежимо мала.

зависимость будет наблюдаться. Так, например, урав нение прямой y=kx+b имеет два параметра: k и b (рис. 1). От значения параметров зависит то, как будет выглядеть связь между X и Y в том или ином конкретном случае.

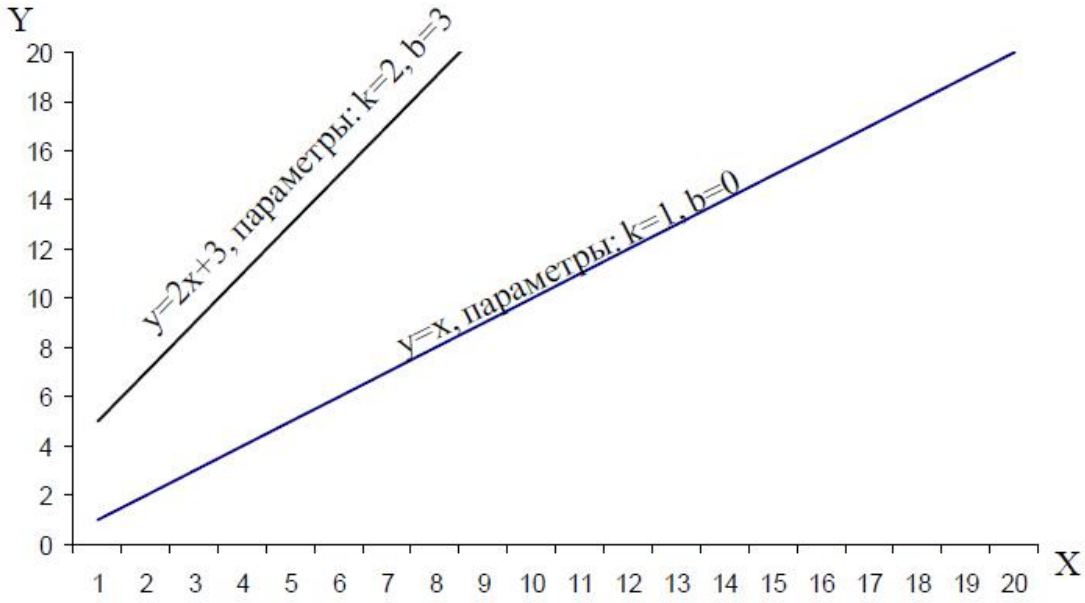

Рисунок 1. Пример простейшей функции, устанавливающей связь между X и Y. Прямая линия - это график функции Y=kX + b. При разных значениях параметров k и b зависимость выглядит по-разному.

Наиболее часто в биологии исходят из того, что связь между величиной и ее вероятностью отражается так называемым ***нормальным распределением***[7]. Нормальное распределение имеет два параметра X и S[8] (рис. 2). Первый параметр равен тому значению величины, вероятность которого наибольшая. Второй параметр описывает размах варьирования величин, степень их разброса в генеральной совокупности. Если все величины в совокупности оказываются одинаковыми, то параметр S=0. Правда, в этой ситуации мы уже не получим нормального распределения, так как параметр S стоит в знаменателе дроби (см. пояснения к рисунку 2).

Для того чтобы описать генеральную совокупность, к чему, собственно, и стремится исследователь, необходимо вычислить параметры распределения изучаемой величины в генеральной совокупности. Если мы знаем эти параметры, то мы можем вычислить, с какой вероятностью мы встретим в данной генеральной совокупности то или иное значение признака, что нас и интересует. Однако эти параметры напрямую измерить

[7] На самом деле, это далеко не всегда так. Существуют такие признаки, которые подчиняются другим типам распределения. Строгий анализ таких величин требует дополнительных ухищрений и особого математического аппарата. К сожалению, в данной брошюре мы не можем остановиться на всех тонкостях работы с такими признаками.

[8] В разных изданиях существует разнобой в обозначениях статистических параметров. В некоторых из них греческими буквами обозначаются генеральные параметры, а римскими – выборочные, в других – принята другая система (все обозначается разными римскими буквами). По большому счету, способ обозначения не важен, важно то, что за этими обозначениями кроется. Обозначения, которые приняты в данной брошюре, всего лишь привычны для автора. Надеюсь, что это не создаст больших трудностей для читателей.

нельзя, так как для этого пришлось бы провести изучение всей генеральной совокупности. Поэтому производят **оценку** генерального показателя на основе выборки.

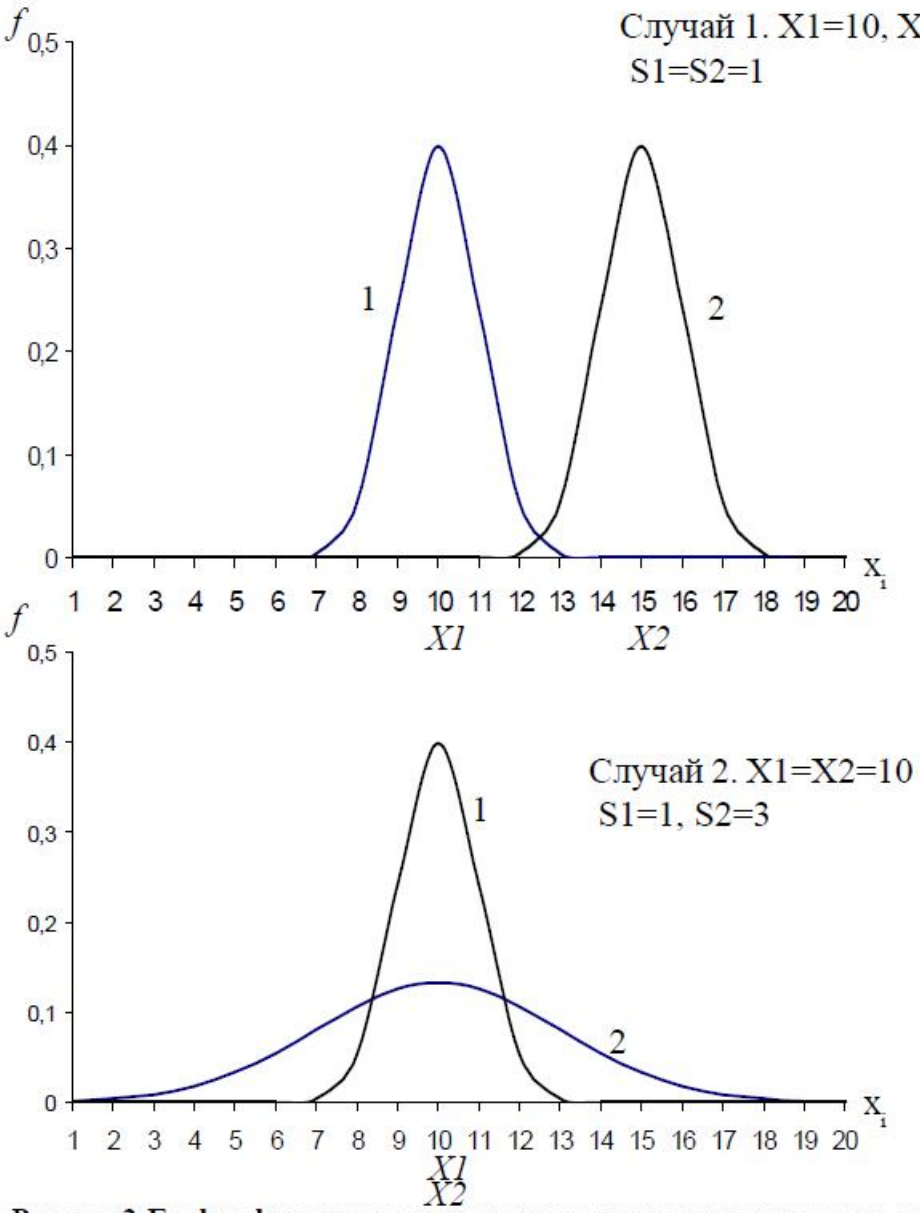

Рисунок 2. График функции нормального распределения при разных парметрах X и S.

Пояснения:

Функция нормального распределения: $f = \frac{1}{S\sqrt{2\pi}} e^{-\frac{(x_i - X)^2}{2S}}$

где f – вероятность встретить в генеральной совокупности величину x_i, X и S – параметры распределения. Числа e (основание натурального логарифма) и π – константы. Число e= 2,7182; π = 3,14159.

Для оценки параметра X используют *среднее арифметическое* значение величины в выборке.

$$\bar{x} = \frac{x_i}{N}$$

В этой формуле $\bar{x}$ – среднеарифметическое значение, x_i- конкретные значения величины, измеренной в выборке, N - объем выборки (количество объектов, попавших в выборку).

Другой параметр распределения (S) оценивается так называемым *среднеквадратичным отклонением*, которое вычисляется по формуле:

$$\sigma = \sqrt{\frac{(\bar{x} - x_i)^2}{N-1}} = \sqrt{\frac{x_i^2 - \frac{(\quad x_i)^2}{N}}{N-1}}.$$

Второй вариант формулы для вычисления сигмы существенно проще для расчетов вручную.

Оба эти показателя имеют очень важное свойство – при очень большом объеме выборки эти показатели равны соответствующим генеральным параметрам. Иными словами:

$$\bar{x} \quad {}_{N\to} \to X$$
$$\sigma \quad {}_{N\to} \to S$$

Это означает, что чем больше объем выборки, тем точнее мы оцениваем с помощью выборочных показателей генеральные параметры.

Вычислить выборочные оценки достаточно просто, но все дело в том, что эти величины оценивают генеральные параметры лишь приблизительно, поскольку объем выборки, который имеется в нашем распоряжении, всегда во много раз меньше объема генеральной совокупности. Однако нам бы хотелось оценить генеральные параметры как можно точнее. Для этого применяют так называемые *интервальные оценки*.

Для начала рассмотрим интервальную оценку параметра *X*. Его приблизительная (как говорят **точечная**) оценка, как мы уже знаем, – это средняя арифметическая ($\bar{x}$). Для получения интервальной оценки необходимо ввести величину, которую называют *ошибкой среднего***.** Она вычисляется по следующей формуле:

$$m_x = \frac{\sigma}{\sqrt{N}}.$$

Математиками было показано, что ошибка среднего обладает двумя важными свойствами.

Первая особенность ошибки заключается в том, что генеральный параметр *X* с вероятностью 95% находится в пределах интервала $\bar{x} \pm 1{,}96 \cdot m_x$. Строго говоря, такая запись не совсем верна. Правильнее было бы написать так: $\bar{x} \pm t \cdot m_x$. Значение t=1,96 появляется тогда, когда объем выборки превосходит 120 объектов. Если выборка меньше, то значение t будет иным. Иными словами, для вычисления параметров интервальной оценки необходимо учитывать объем выборки. Для простоты объяснения далее мы будем пренебрегать этой особенностью. Однако для более строгого анализа, объем выборки надо учитывать обязательно![9]

Указанные выше свойства ошибки позволяют утверждать, что с вероятностью 95% генеральный параметр *X* не может быть выше величины $\bar{x} + 1{,}96 \cdot m_x$ и не может быть меньше $\bar{x} - 1{,}96 \cdot m_x$. Упомянутый интервал называется *доверительным интервалом* (рис. 3), а вероятность попадания генерального параметра *X* в доверительный интервал называется *доверительной вероятностью* ($P_{дов}$). Эта вероятность показывает, **насколько мы можем быть уверены** в том, что искомый параметр находится в пределах доверительного интервала. Степень нашей **неуверенности** в этом будет отражать величина, равная 100%-$P_{дов}$ (если вероятность выражена в долях от единицы, то 1- $P_{дов}$). Такая величина называется *уровнем значимости*. Так, например, мы ошибаемся, утверждая, что *X* принадлежит интервалу [$\bar{x} - 1{,}96 \cdot m_x$; $\bar{x} + 1{,}96 \cdot m_x$] при уровне значимости

[9] О том как это сделать см. специальные учебники, список которых приведен в конце брошюры.

5%, или 0,05. Чем меньше уровень значимости, тем выше наша уверенность, или, как говорят, *достоверность оценки*[10].

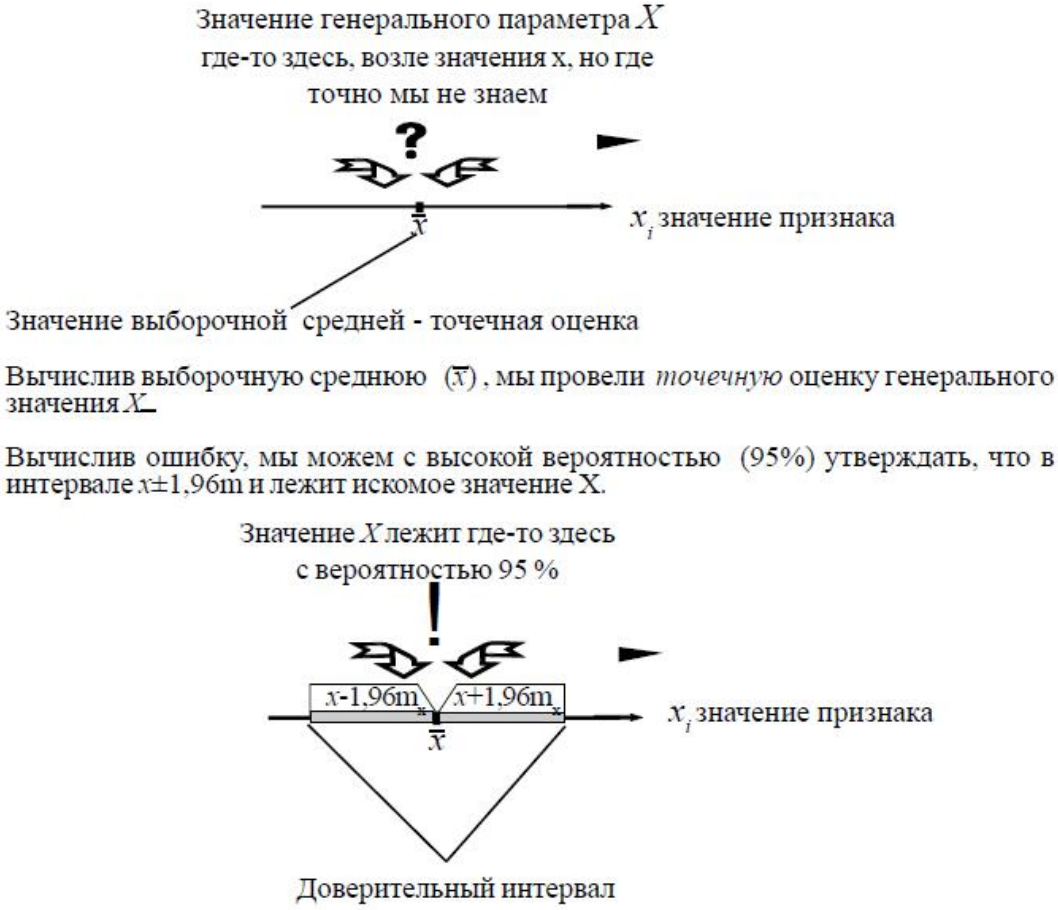

Рисунок 3. Соотношение точечной и интервальной оценки генерального параметра X

Если расширить границы доверительного интервала, например, умножить ошибку (m_x) не на 1,96, а на 2,58 (при том же объеме выборки), то вероятность того, что X попадет в интервал $\bar{x} \pm 2{,}58 \cdot m_x$ увеличится. В этом случае доверительная вероятность будет равна 99%, а, стало быть, вероятность того, что мы ошибаемся (уровень значимости), утверждая, что $X \in [\bar{x} - 2{,}58 \cdot m_x;\ \bar{x} + 2{,}58 \cdot m_x]$, будет равна лишь 1%.

В большинстве биологических исследований принимается, что исследователю достаточно быть уверенным на 95%, чтобы сделать обоснованные выводы или считать, что допустима неверная оценка с вероятностью в 5%.

Вторая особенность статистической ошибки заключается в том, что при увеличении объема выборки значение ошибки уменьшается. То есть $m_x \underset{N\to}{\ } \to 0$. Это означает, что чем больше объем выборки, тем уже становится доверительный интервал, а, стало быть, тем точнее мы оцениваем значение X.

Подобно интервальной оценке генерального параметра X существует и интервальная оценка для параметра S. Она называется *ошибкой среднеквадратичного отклонения*, а ее значение вычисляется по следующей формуле:

[10] О понятии достоверности см. ниже.

$$m_\sigma = \frac{\sigma}{\sqrt{2\ N}}.$$

Все свойства ошибки среднего справедливы и для ошибки среднеквадратичного отклонения. То есть $m_{\sigma\ N\to} \to 0$ и с вероятностью 95% генеральный параметр $S \in [\sigma - 1{,}96 \cdot m_\sigma; \sigma + 1{,}96 \cdot m_\sigma]$.

Все сказанное выше относилось к количественным данным, когда выборка представлена рядом чисел. А как надо рассуждать в случае качественных данных? Напомним, что в этом случае в выборке регистрируется лишь наличие или отсутствие признака у объекта. Как правило, исследователя, работающего с такими данными, интересует встречаемость того или иного признака. Например, его интересует, с какой вероятностью можно встретить больную особь в популяции или с какой частотой встречаются цветки правильного и неправильного строения. И, наконец, этот тип данных используется при социологических исследованиях, например, при опросах местного населения, когда исследователя интересует частота тех или иных ответов на заданные вопросы.

Во всех упомянутых случаях исследователь должен подразумевать, что его целью оказывается оценка *генеральной встречаемости* ($P_{ген.}$), то есть, доля особей с тем или иным признаком в генеральной совокупности. Как уже отмечалось, такая доля практически не может быть вычислена. Поэтому на практике применяется оценочная, приблизительная величина – p. Она вычисляется следующим образом.

$$p = \frac{n}{N}\ 100\%$$

В этой формуле p – выборочная встречаемость, N – объем выборки (число изученных объектов), n – число объектов, имеющих анализируемый признак. Нетрудно заметить, что $p\ _{N\to} \to P_{ген.}$. Особо отметим, что часто это выражение вызывает у школьников недоумение: они замечают, что рост знаменателя должен приводить к уменьшению значения дроби. Однако надо помнить, что при увеличении объема выборки (N) увеличивается и количество особей, несущих анализируемый признак (n), а, стало быть, увеличивается и числитель.

Выборочная оценка встречаемости, как и средняя арифметическая - это точечная, приблизительная оценка генеральной величины. Поэтому для более точной оценки надо вычислить интервальную оценку – *ошибку встречаемости*. Этот показатель вычисляется по следующей формуле:

$$m_p = \frac{\sigma_p}{\sqrt{N}}.$$

То есть он вычисляется аналогично ошибке среднего. Вместе с тем, σ_p вычисляется несколько иначе:

$$\sigma_p = \sqrt{p\ \ (100 - p\ )}.$$

Таким образом, ошибка встречаемости может быть вычислена по формуле:

$$m_p = \sqrt{\frac{p\ \ (100 - p\ )}{N}}.$$

Эта величина обладает всеми теми же свойствами, что и ошибка среднего. То есть $m_{p\ \ N\to} \to 0$ и с вероятностью 95 % $P_{ген.} \in [p - 1{,}96 \cdot m;\ p + 1{,}96 \cdot m]$.

А как быть с балловыми оценками? С ними все несколько сложнее. Дело в том, что баллы придумывают люди, их в генеральной совокупности не существует. Баллы – это «суррогаты», которыми исследователи пользуются для того, чтобы описать некоторую трудную для измерения величину. Поэтому при работе с баллами не применяют точечные и интервальные оценки генеральных параметров распределения (по крайней мере, в том виде, в котором мы обсуждали выше). Однако существуют методы, с помощью которых

можно сделать обоснованное заключение о свойствах генеральной совокупности, изученной с помощью балловых оценок. О таких методах речь пойдет ниже.

Подводя итог всему сказанному выше, надо еще раз подчеркнуть, что главная задача исследователя при работе с выборками – это оценка генеральных параметров, которая всегда делается приблизительно, с некоторой вероятностью. При этом точечные оценки генеральных параметров всегда должны дополняться интервальными (т.е., ошибками). Можно даже сформулировать такой афоризм: тот, кто работает с выборками без статистических ошибок, работает ошибочно.

Глава 1.10. Понятие статистической достоверности

В ходе своей работы исследователь должен выяснять некоторые объективно существующие закономерности. Однако строго судить о том, какая закономерность объективна, а какая нет, зачастую нельзя, так как реальные законы от нас скрыты, мы можем лишь приблизительно их оценить. Поэтому вместо понятий «объективная реальность», «истинные значения» и т. п. в науке чаще используют понятие *достоверности*. Под достоверным утверждением понимается такое высказывание, которое с **высокой вероятностью может оказаться истинным**.

Рассмотрим такой пример. Пусть исследователь поставил своей целью изучить воздействие некоторого лекарственного препарата. Для этого он взял два подопытных животных. Одному ввел препарат, а другому нет. В результате он установил, что животное, которому он ввел препарат, выздоровело, а другое – нет.

Можно ли утверждать, что ученый заметил в своем эксперименте какую-то закономерность? Да, действительно, полученный результат совпал с предположением исследователя о целебных свойствах препарата. Можно ли теперь применять этот препарат для лечения людей. Нет! Может быть, эти два животных были какие-то нетипичные, а мы сразу человека лечить собираемся. В данной ситуации полученная закономерность была недостоверной. Как сделать ее достоверной? Для этого надо, например, увеличить объем выборки, то есть, в каждой группе должно быть не одно животное, а несколько. А самое главное, необходимо применить *методы статистического анализа*, которые всем людям, разбирающимся в науке, докажут, что выявленная закономерность достоверна. О том, как применять такие анализы, мы и поведем разговор в следующих главах.

ЧАСТЬ 2. Статистические методы

Глава 2.1. Методы сравнения двух величин

Одним из самых распространенных недостатков самостоятельных исследовательских работ школьников оказывается обсуждение различий между двумя совокупностями без доказательств достоверности этих отличий.

Рассмотрим достаточно типичный пример[11]. Нередко в работах школьников встречаются приблизительно такого рода утверждения: «В экологически неблагополучном районе частота заболеваний дыхательных путей составляет 50%, в то время как в более благоприятных областях эта величина не превышает 25%». Несмотря на высокую разницу между приведенными величинами, любой грамотный исследователь (или чиновник) не примет данное утверждение за доказательство неблагополучия в первом районе. Прежде всего, остается непонятным, получены ли эти данные путем тотального учета всех без исключения людей или приведенные данные – это результат выборочного учета (например, результаты работы с картотеками некоторых поликлиник).

[11] Почти все примеры, приведенные в данной брошюре, - вымышленные и автор не уверен, что все полученные при разборе примеров закономерности могут быть обнаружены в реальных условиях.

Поскольку в подавляющем большинстве случаев работа идет с выборками, то приведенные величины должны сопровождаться всеми атрибутами выборочных данных, то есть точечными и интервальными оценками. Без приведения всех этих величин определение уровня заболеваний оказывается абсолютно неубедительным. Рассмотрим данный пример с точки зрения правильного описания.

Прежде всего, необходимо представить данные в виде исходной таблицы. Например, такой.

Таблица 2. Частота заболеваемости дыхательных путей в двух районах

	Район	
	Экологически неблагоприятный	Экологически чистый
Общее число обследованных людей (объем выборки, N)	260	800
Число людей, у которых отмечены заболевания дыхательных путей	130	200
Встречаемость заболеваний (%)	50	25

Нетрудно заметить, что две выборки отличаются по своему объему. В первом случае людей обследовано почти в три раза меньше, чем во втором.

Поскольку мы имеем дело с выборкой и при этом полученные данные имеют качественный характер, то выборочные оценки будут иметь следующий вид:

$$p_1 = \frac{n}{N}\ 100\% = \frac{130}{260}\ 100\% = 50\% \quad \text{и} \quad p_2 = \frac{200}{800}\ 100\% = 25\%$$

$$\sigma_{p1} = \sqrt{p_1\ (100 - p_1)} = \sqrt{50\ (100 - 50)} = 50 \text{ и } \sigma_{p2} = \sqrt{25\ (100 - 25)} = 43{,}3$$

$$m_{p1} = \frac{\sigma_{p1}}{\sqrt{N}} = \frac{50}{\sqrt{260}} = 3{,}1 \text{ и } m_{p2} = \frac{43{,}3}{\sqrt{800}} = 1{,}5$$

В окончательном виде результаты исследования можно записать в следующем кратком виде. Встречаемость заболеваний дыхательных путей в экологически неблагоприятном районе составляет p_1=50±3,1% (N=260 человек), в экологически чистом районе частота заболеваний составляет p_2=25±1,5% (N=800 человек)[12]. В таком виде результаты исследования будут приняты любым грамотным исследователем. Напомним, что приведенная запись означает, что в генеральной совокупности точное (генеральное) значение параметра $P_{ген}$ приблизительно равно p, и это точное значение с вероятностью 95% лежит в пределах $\pm 1{,}96 \cdot m_p$. Так, для случая с экологически неблагоприятным районом с вероятностью 95% $P_{ген} \in [50\text{-}1{,}96 \cdot 3{,}1;\ 50+1{,}96 \cdot 3{,}1]$ или $P_{ген} \in [44{,}1;\ 55{,}9]$. После того, как получены правильные выборочные оценки, можно приступить к решению вопроса о том, достоверны отличия между ними или нет. Иными словами, мы должны решить, отличаются ли генеральные параметры в двух генеральных совокупностях. На языке статистики подобная задача называется тестированием *нулевой гипотезы*.

Нулевая гипотеза, или нуль-гипотеза, – это утверждение, которое звучит так: генеральные параметры двух совокупностей равны. Или в символьном виде

$$H_0 = P_{ген1} - P_{ген2} = 0.$$

[12] Обратите внимание на то, что при записи результатов после значка «±» стоит одна ошибка. Это не запись доверительного интервала при доверительной вероятности 95%! Это просто краткая форма представления результатов.

Большинство статистических методов направлено на то, чтобы нулевую гипотезу не доказать, а отвергнуть. С чем это связано? Все дело в том, что данные, которыми мы оперируем, – это выборочные данные, которые оценивают генеральные параметры лишь приблизительно, с определенной вероятностью. Поэтому всегда останется вероятность того, что мы ошиблись при оценке того или иного параметра. Поэтому строго доказать равенство двух генеральных показателей нельзя, однако можно доказать их различия, то есть отвергнуть H_0. В связи с этим, вся логика статистического анализа построена не на поиске сходств, а на выявлении различий. Соответственно и планирование работы надо вести так, чтобы в конечном итоге мы встали **не перед задачей доказательства равенства**, а перед задачей **доказательства различий выборочных величин**. Если мы сможем отвергнуть нулевую гипотезу, то мы сможем с вероятностью 95% утверждать, что различия между сравниваемыми величинами **статистически значимые**, или **достоверны**.

Самым простым и наиболее распространенным методом выявления достоверности различий между двумя выборочными оценками можно считать t-критерий Стьюдента. Если не углубляться в теорию, то использование этого критерия сводится к расчету следующей величины:

$$t = \frac{|p_1 - p_2|}{\sqrt{{m_{p1}}^2 + {m_{p2}}^2}}$$

В приведенном выше примере вычисления будут иметь следующий вид:

$$t = \frac{|50 - 25|}{\sqrt{3{,}1^2 + 1{,}5^2}} = \frac{25}{\sqrt{11{,}86}} = 7{,}26$$

После того как вычислено эмпирическое значение t, его необходимо сравнить с табличным, пороговым значением $t_{табл}$. Это значение находится по специальным таблицам (таблица I) и для правильного его нахождения необходимо иметь в распоряжении еще две величины. Первая величина – это уже знакомая **доверительная вероятность** ($P_{дов}$). Вычислять ее не надо, надо лишь выбрать, какая степень достоверности результата нас устроит. Напомним, что в биологии обычно считается, что в качестве такой вероятности следует выбирать 95%. Это означает, что мы на 95 процентов уверены, что различия между выборочными показателями неслучайны и отражают некоторые реально существующие различия. Лишь 5 % остается на то, что мы ошибаемся в нашем выводе.

Вторая величина требует вычислений. Это так называемое *число степеней свободы***.** Оно определяется по следующей формуле[13]:

$$\nu = N_1 + N_2 - 2$$

В нашем примере ν=260+800−2=1058. Табличное значение при ν=1058 (в таблице такая большая величина обозначается как ∞) и доверительной вероятности 95% равно $t_{табл}$=1,96. Сравнение эмпирического значения t и табличного показывает, что $t > t_{табл.}$. Если эмпирическое значение t превышает табличное, то это означает, что нулевую гипотезу можно с вероятностью 95% (и выше) отвергнуть. То есть, различия между выборочными показателями оказываются **достоверными**, иными словами, мы можем быть практически уверенными (на 95%), что эти **различия неслучайны и отражают реальную закономерность** – в экологически неблагополучном районе частота заболеваемости дыхательных путей выше.

Если бы при наших расчетах получилось, что $t < t_{табл.}$, то это означало бы, что различия **недостоверны** – нуль-гипотеза не может быть отвергнута. Особо отметим, что если мы не смогли отвергнуть нулевую гипотезу, то это не означает, что величины равны, это означает, что различия не доказаны!

[13] Определение числа степеней свободы в разных статистических анализах ведется по-разному.

Приведенный пример был основан на качественных данных. Вместе с тем сравнение выборочных величин можно проводить и на основе количественных показателей.

Рассмотрим пример. Предположим, что в процессе исследования мы изучили длину соцветий лисохвоста в двух разных популяциях, одна из которых подвержена выпасу скота, а другая - нет. При этом мы хотим выяснить, влияет ли выпас коров на рост лисохвоста. Для этого мы случайным образом отобрали несколько растений в двух разных популяциях и измерили у них длину соцветий. Как и в предыдущем случае, мы должны первым делом составить исходную таблицу.

Таблица 3. Размеры соцветий в двух популяциях лисохвоста

Длина соцветия (см)	
Выпас есть	**Выпаса нет**
5,7	4,6
4,1	9,6
13,6	14,1
4,4	5,0
3,3	9,1
9,8	16,6
7,2	9,5
9,3	8,6
12,7	10,8
6,3	10,6
$\bar{x}_1$=7,6	$\bar{x}_2$=9,8
σ_1=3,60	σ_2=3,62
N_1=10	N_2=10
m_{x1}=1,13	m_{x2}=1,13

Здесь $\bar{x}$ – среднее значение, σ – среднеквадратичное отклонение, N - объем выборки (количество измеренных колосков), m_x – ошибка среднего. Как и в предыдущем случае, бросается в глаза, что на поле, где нет выпаса, колоски в среднем длиннее, поскольку $x_2 > x_1$. Но давайте вспомним, что нас интересует разница не выборочных значений, а разница генеральных параметров! То есть, нуль-гипотеза будет иметь следующий вид: $H_0 = X_1 - X_2 = 0$. Для тестирования этой гипотезы воспользуемся t-критерием Стьюдента. Для сравнения средних он будет иметь следующий вид:

$$t = \frac{d}{m_d}$$

$$d = |\bar{x}_1 - \bar{x}_2|$$

$$m_d = \sqrt{m_{x1}{}^2 + m_{x2}{}^2}\ .$$

Для нашего случая вычисления будут выглядеть так:

$$d = |7{,}6 - 9{,}8| = 2.2$$

$$m_d = \sqrt{1{,}13^2 + 1{,}13^2} = 1{,}60$$

$$t = \frac{2{,}2}{1{,}6} = 1{,}34\ .$$

Как и в случае с анализом различий встречаемости, необходимо сравнить эмпирическое значение t и табличное пороговое значение $t_{порог.}$. Для поиска табличного значения надо,

во-первых, выбрать доверительную вероятность. Возьмем ее равной 95%. Во-вторых, надо определить число степеней свободы.

$$\nu=N_1+N_2-2=10+10-2=18.$$

Табличное значение при доверительной вероятности 95% и $\nu=18$ составляет $t_{порог}=2{,}10$. Мы видим, что эмпирическое значение t меньше табличного, порогового. Это означает, что у нас нет оснований отвергнуть нулевую гипотезу. Еще раз напоминаем, что это не означает, что разницы нет, а означает, что разница не доказана.

Таким образом, в нашем примере мы не смогли доказать, что выпас скота влияет на длину соцветий лисохвоста. При этом вполне вероятно, что, увеличив объем выборки (а в нашем случае он чрезвычайно мал), мы выявим достоверные отличия. Вот тогда мы сможем с уверенностью на 95% утверждать, что выпас коров приводит к уменьшению длины соцветия.

В заключение этой части надо отметить, что при сравнении выборочных средних t-критерий имеет некоторые особенности, о которых мы еще не говорили. Главная особенность заключается в том, что если объемы двух выборок сильно отличаются (т.е. $N_1>N_2$ или $N_2>N_1$), то расчеты надо вести немного иначе. Формулы будут такие.

$$d = \left|\bar{x}_1 - \bar{x}_2\right|$$

$$m_d = \sqrt{\frac{\sigma_1^{\,2}\ (N_1-1)+\sigma_2^{\,2}\ (N_2-1)}{N_1+N_2-2}}\ \sqrt{\frac{N_1+N_2}{N_1\ \ N_2}}$$

$$t = \frac{d}{m_d}.$$

То есть, различия будут в методе расчета m_d. В остальном же применение этого метода будет абсолютно таким же, как описано выше.

А как сравнить выборки, если в вашем распоряжении есть только балловые данные? Рассмотрим такой пример. Предположим, вы задались целью сравнить две популяции черных кошек, допустим, в центре города и за городом, по степени выраженности у них белых пятен. При этом вы дали следующие балловые характеристики.

Таблица 4. Балловые характеристики окраски кошек

Балл	**Описание**
0	Нет белых пятен, чисто черное животное
1	одно небольшое белое пятно
2	несколько небольших белых пятен
3	одно крупное белое пятно
4	несколько крупных белых пятен

Обратите внимание, что здесь мы имеем дело именно с балловыми оценками, а не просто с цифровым обозначением разновидностей. В нашем примере чем больше балл, тем меньше черного в окраске кошки. Мы специально остановились на этом, так как иногда, вводя числовые обозначения, люди попадают под магию цифр и пытаются применить математику там, где ее не должно быть (или применяют не ту математику).

Далее вы собрали материал и оформили его в виде таблицы. Предположим, такой.

Таблица 5. Результаты учета кошек в разных районах

В центре города	**За городом**
0	0
0	0
0	1
0	1
0	1
1	1
1	2
1	2
4	2
4	2
4	3
4	3
4	4
4	4
4	4
4	4
	4
	4
	4
	4
Всего кошек: 16	Всего кошек: 20

Наличие цифрового материала в данной таблице сразу вызывает у нас соблазн применить простой метод сравнения средних, который был описан выше. Действительно, чего проще рассчитать средний балл и сравнить эти средние баллы. Давайте попробуем. Для городской популяции средний балл составляет 2,19. Можно округлить до 2. Казалось бы, очень просто и логично. Но давайте задумаемся, а что отражает этот средний балл? Честно говоря, он не отражает ничего! Вспомните, что средняя величина оценивает генеральный параметр распределения *X*. А согласно свойству нормального распределения величины близкие к *X* должны встречаться с максимальной вероятностью (см. рис.2) . В нашем же случае, в выборке кошек из центра города, не было отмечено ни одной особи с баллом 2. Стало быть, вычислив среднее значение балла, мы получили откровенную ложь! К сожалению, очень часто балловые оценки подвергаются некорректному статистическому анализу. С баллами надо быть очень аккуратными, далеко не всегда вычисление средних баллов дает объективную информацию о системе.

Как же нам поступить в данном случае? Самый простой способ – это воспользоваться так называемым критерием χ^2 (читается *хи-квадрат*)[14]. Этот критерий служит для сравнения двух распределений. Помните, мы определили распределение как связь между величиной и вероятностью встречи данной величины в генеральной совокупности. Однако, когда мы работаем с выборкой, то мы тоже можем говорить о распределении, но уже не о генеральном, а о *выборочном распределении*. Выборочное распределение уже нельзя точно описать какой-то функцией, однако его можно представить в виде так называемого *вариационного ряда*. Обычно вариационный ряд представляют в виде таблицы из двух колонок. В первой колонке выписаны значения признака, который мы анализируем, а во второй – число объектов, обладающих данным

[14] На самом деле, существует целый раздел науки, который называется «непараметрическая статистка». Непараметрические методы нужны тогда, когда тип распределения величины в генеральной совокупности неизвестен или заведомо отличается от нормального. В такой ситуации сравнение выборочных параметров x и σ некорректно. Здесь нужны непараметрические критерии достоверности. В данном издании мы коснемся лишь незначительной части из них.

признаком, которые попали в выборку. Для нашего примера вариационные ряды будут иметь следующий вид.

Таблица 6. Вариационные ряды для черных кошек из двух популяций

Балл	**В центре города**	**За городом**
0	5	2
1	3	4
2	0	4
3	0	2
4	8	8
Всего	16	20

Когда вариационные ряды составлены, можно приступить к их сравнению. Формула для вычисления критерия χ^2 в данном случае имеет следующий вид:

$$\chi^2 = \frac{1}{N_1 \ N_2} \quad \frac{(f_{i1} \ N_2 - f_{i2} \ N_1)^2}{f_{i1} + f_{i2}}.$$

В этой формуле N_1 и N_2 – объемы выборок, f_{i1} *и* f_{i2} – количество объектов, обладающих i- тым признаком в первой и во второй совокупностях соответственно. Для вычисления χ^2 проще всего построить такую таблицу.

Таблица 7. Ход вычисления χ^2

Балл	**В центре города (f_1)**	**Загородом (f_2)**	f_{i1}**N2**	f_{i2}**N1**	f_{i1}**N2-** f_{i2}**N1**	$(f_{i1}$**N2-** f_{i2}**N1)²**	$\frac{(f_{i_1} \ N_2 - f_{i_2} \ N_2)^2}{f_{i_1} + f_{i_2}}$
0	5	2	100	32	68	4624	660,6
1	3	4	60	64	-4	16	2,3
2	0	4	0	64	-64	4096	1024,0
3	0	2	0	32	-32	1024	512,0
4	8	8	160	128	32	1024	64,0
	N_1=16	N_2=20					∑= 2262,9

Далее по формуле вычисляем: $\chi^2 = \frac{1}{16 \ 20} \ 2262,6 = 7,07$.

Итак, искомое значение критерия получено. Теперь, как и в предыдущих случаях, надо сравнить это значение с табличным (таблица II). Для поиска табличного, порогового, значения нам опять необходимы две величины – доверительная вероятность и число степеней свободы.

Доверительная вероятность выбирается нами в соответствии с нашими представлениями о степени надежности заключения. Если нас устраивает, что надежность заключения 95% (то есть с вероятностью 5% мы ошибаемся), то мы должны выбрать именно эту доверительную вероятность. В принципе, мы можем согласиться и на менее надежные выводы, взяв, например, доверительную вероятность 90% (но тогда и табличное значение будет другим).

Второй величиной, необходимой нам для поиска порогового значения χ^2 в таблице, как и в предыдущих случаях, оказывается число степеней свободы. В данном случае оно

вычисляется по следующей формуле: $\nu = K-1$[15]. Здесь K – это число классов. В нашем случае число классов определяется количеством баллов. Всего для анализа материала мы использовали 5 баллов, следовательно, число степеней свободы будет: $\nu = 5-1 = 4$. Итак, мы выбрали 95% доверительную вероятность и число степеней свободы равно 4. Этому сочетанию в таблице II соответствует $\chi^2_{порог} = 9{,}49$. Сравнение табличного и эмпирического значений χ^2 говорит о том, что эмпирическое значение ниже табличного. Это, как и в случае с t-критерием Стьюдента, говорит о том, что различия между двумя вариационными рядами недостоверны. Полученные данные позволяют сделать вывод о том, что у нас нет оснований для утверждения, что черные кошки за городом отличаются от черных кошек в центре города. Весьма вероятно, что при большем объеме выборок, если бы мы проанализировали окраску не десятков, а сотен животных, мы выявили бы достоверные отличия. Но пока эти отличия не доказаны, мы не имеем права делать вывод о том, что эти две популяции отличаются друг от друга. Заметим, что мы не имеем права делать и вывод о том, что эти популяции не отличаются!

Между прочим, пример с черными кошками позволяет поставить и другой вопрос. Например, такой: «Где больше вероятность встретить чисто черную кошку в городе или за городом?» Поскольку вопрос задан по-другому, то и метод анализа будет совершенно другим! В данном случае мы должны сравнить встречаемость чисто черных кошек среди черных кошек вообще в городе и за городом. А это уже задача, решаемая с помощью t-критерия Стьюдента. Попробуйте решить эту задачку сами.

Глава 2.2. Методы анализа структуры популяции

Прежде чем мы приступим к разговору о собственно методах анализа структуры популяции, необходимо договориться о некоторых базовых понятиях популяционной биологии.

Под популяцией мы будем понимать группу особей одного вида, населяющих определенную территорию. Заметьте, что для применения статистических методов нам нет нужды вводить представление о том, что эти особи свободно скрещиваются друг с другом.

Структуру популяции мы будем понимать, как некоторую неоднородность особей, входящих в состав популяции. Например, под половой структурой мы будем понимать соотношение числа особей мужского и женского полов.

Если теперь перейти на язык статистики, то вся популяция для нас будет генеральной совокупностью. Она неоднородна по каким-то признакам. Структуру генеральной совокупности мы будем пытаться анализировать с помощью небольшой выборки особей из этой популяции.

Структуру популяции можно изучать с разных точек зрения. Наиболее часто изучаются следующие аспекты.

1. Половая структура популяции – все особи в популяции различаются по своему полу.
2. Фенетическая структура популяции – особи различаются по некоторым дискретным внешним признакам - фенам, например, по окраске тела.
3. Генетическая структура популяции – особи в популяции рассматриваются как носители того или иного аллеля.
4. Размерно-возрастная структура популяции – особи различаются по своим размерам и возрасту.

---

[15] Внимание! В других случаях применения критерия χ^2 формулы для вычисления числа степеней свободы будут другими!

Очень часто в первых трех случаях главной задачей оказывается сравнение эмпирического и теоретического распределения. С рассмотрения решения этой задачи мы и начнем.

Предположим, что мы задались целью проанализировать половую структуру популяции мышей в чулане вашего дома. Для этого анализа необходимо сделать выборку особей. Конечно, для дома было бы полезней произвести анализ всей генеральной совокупности, по крайней мере, всех мышей бы извели, но, предположим, что сделать это невозможно, допустим, грызунов слишком много. Итак, вам предстоит сделать выборку. Как это сделать технически, решайте сами, но, так или иначе, зверьков надо отловить и определить у них пол. Допустим, вы поймали 80 мышей. Из них 15 оказались самцами, а 65 самками. Результаты вашей охоты сразу надо представить в виде исходной таблицы (лучше себя сразу приучить грамотно оформлять результаты исследования).

Таблица 8. Соотношение полов в изученной популяции мышей

Число самцов	**Число самок**
15	65
Всего мышей: 80	

Результаты анализа выборки показывают, что число самцов не равно числу самок. Можем ли мы, исходя из полученных данных, сделать вывод о том, что в нашей популяции самцов меньше, чем самок. Конечно же, нет! Пока нет. Нам необходимо доказать это. Для проверки гипотезы о том, что самцов меньше, чем самок, надо доказать, что наблюдаемое соотношение полов достоверно отличается от соотношения 1:1, при котором количество самцов равно количеству самок. Как это сделать? Для этого используют уже знакомый нам критерий χ^2, только в несколько иной модификации. В данной ситуации нам необходим критерий для сравнения реального (эмпирического) и теоретического распределения.

Напомним, что все статистические критерии призваны тестировать нулевую гипотезу. В нашей ситуации нулевая гипотеза будет сформулирована так: $H_0=f_{\text{ген.}}-f_{(1:1)}=0$, или в словесном выражении она будет звучать так: «в генеральной совокупности соотношение численностей разных полов подчиняется соотношению 1:1».

Первый шаг к применению критерия χ^2 – вычисление *теоретически ожидаемых частот*, или, что то же самое, **построение теоретического распределения**. Выражение, выделенное жирным шрифтом, несмотря на непривычное звучание, имеет очень простой смысл. Что такое распределение, мы уже знаем. Под **частотой** подразумевается количество объектов, обладающих данным признаком. Мы с этим уже тоже встречались, когда говорили о вариационном ряде. Что такое «**теоретически ожидаемые частоты**»? Тоже очень просто! Это означает «сколько **было бы** особей с теми или иными признаками, **если бы** популяция обладала структурой такой, какую мы предполагаем».

Для нашего случая мы предположили, что соотношение самцов и самок должно быть 1:1. Стало быть, теоретически частота самцов должна быть равна 1/2 всех пойманных мышей. Тогда теоретическая частота самцов должна быть 40 и самок – тоже 40. Оформим результаты в виде таблицы.

Таблица 9. Реальные и теоретические частоты самцов и самок мышей

	Самцы	**Самки**
Реальная частота	15	65
Теоретическая частота	40	40

Теперь пришло время воспользоваться критерием χ^2. Для данного типа задач применяется другая формула:

$$\chi^2 = \frac{(E-T)^2}{T}.$$

Здесь E – фактически наблюдаемая (эмпирическая) частота, T – теоретическая частота. Произведем расчеты, занося результаты в таблицу.

Таблица 10. Ход вычисления критерия χ^2

	Самцы	**Самки**
Реальная частота (E)	15	65
Теоретическая частота (T)	40	40
E-T	-25	25
$(E-T)^2$	625	625
$\frac{(E-T)^2}{T}$	15,63	15,63

$$\chi^2 = \frac{(E-T)^2}{T} = 15{,}63+15{,}63 = 31{,}26$$

Итак, вычисленное значение χ^2 =31,26. Теперь, как и в предыдущих случаях, надо сравнить эту величину с табличным значением. Как всегда, надо использовать две величины: доверительную вероятность ($P_{дов}$) и число степеней свободы (ν).

Как обычно, мы выбираем доверительную вероятность $P_{дов}$=95%. Для вычисления числа степеней свободы используется следующая формула:

$$\nu = K - 1,$$

где K – число классов. Таким образом, в нашем случае ν=2-1=1. По таблице значений χ^2 находим пороговое значение для $P_{дов}$=95% и ν =1, оно равно $\chi^2_{порог.}$=3,84. Сравнение табличного и вычисленного значений χ^2 показывает, что $\chi^2 > \chi^2_{порог}$. Это говорит о том, что наблюдаемое соотношение полов с вероятностью 95% отличается от соотношения 1:1. Кстати сказать, если мы возьмем для данного примера доверительную вероятность P=99%, то $\chi^2_{порог}$=6,63. То есть, даже при столь высокой доверительной вероятности, реально наблюдаемое соотношение полов достоверно отличается от теоретического. Это позволяет говорить о том, что даже с вероятностью 99% наблюдаемое распределение не соответствует соотношению 1:1.

Итак, мы доказали, что в популяции мышей самцов меньше, чем самок. На этом статистика свое дело закончила. Теперь надо найти причину, по которой самцов меньше. Может быть, самцы менее осторожны и кошки интенсивнее их вылавливают, а может быть что-то другое. Поиск причин доказанных различий – это самое увлекательное в биологии, но тут уж дать какого-то общего совета или алгоритма действий нельзя. Нужно самому головой поработать!

Разберем теперь другой пример, несколько более сложный, когда в распределении присутствуют не два класса, а много. Давайте попробуем повторить опыт Грегора Менделя и посмотреть на его результаты со статистической точки зрения. Вы помните, что Мендель анализировал наследование двух признаков семян гороха: цвета и формы семени. Во втором гибридном поколении у него получилось следующее соотношение семян по фенотипам: 9 частей желтых гладких семян, 3 части желтых морщинистых, 3 части зеленых гладких и 1 часть зеленых морщинистых.

Предположим, что на своем дачном участке вы решили провести аналогичные исследования. Вы проделали те же самые опыты и, собрав урожай от второго гибридного

поколения, занялись подсчетами различных типов семян. Полученные данные вы тут же записали в исходную таблицу.

Таблица 11. Число семян с разными фенотипами, полученных от растений второго гибридного поколения

Цвет и форма	**Число семян**
Желтые гладкие	311
Желтые морщинистые	113
Зеленые гладкие	102
Зеленые морщинистые	40
ВСЕГО	566

Теперь необходимо провести вычисление теоретически ожидаемых частот. Мы будем предполагать, что теоретически ожидаемое расщепление соответствует менделевскому 9:3:3:1. Как найти теоретические частоты для нашего случая? Для этого надо понять, сколько всего частей должно быть представлено в совокупности, если работает соотношение 9:3:3:1. Таких частей должно быть 16 (9+3+3+1=16). Если частей 16, то можно вычислить, сколько горошин составит одну часть из 566 штук. Для этого делим общее количество горошин на теоретически ожидаемое число частей: 566/16=35,375≈35,4. Получилось дробное число (многие школьники этого очень боятся)! Однако этого не надо смущаться, нужно лишь округлить конечный результат.

Теперь можно вычислять и теоретические частоты. Если работает соотношение 9:3:3:1, то желтых гладких семян мы должны получить 35,4·9≈319, желтых морщинистых 35,4·3≈106, столько же и зеленых гладких, а зеленых морщинистых должно быть 35,4·1≈35 (они, в соответствии с менделевским законом, составляют одну часть). Внимание! При округлении необходимо следить за тем, чтобы суммы эмпирических и теоретических частот были равны. В нашем случае это условие выполнено. Сведем все это в таблицу.

Таблица 12. Эмпирическое и теоретическое распределение по фенотипам семян гороха

	Число семян эмпирическое (E)	**Число семян теоретическое (T)**
Желтые гладкие	311	319
Желтые морщинистые	113	106
Зеленые гладкие	102	106
Зеленые морщинистые	40	35

Теперь все готово для вычисления критерия χ^2, которое ведется по тем же формулам, что и в предыдущем случае.

Таблица 13. Ход вычисления критерия χ^2

	E	**T**	**E-T**	**$(E-T)^2$**	$\frac{(E-T)^2}{T}$
Желтые гладкие	311	319	-8	64	0,2006
Желтые морщинистые	113	106	7	49	0,4623
Зеленые гладкие	102	106	-4	16	0,1509
Зеленые морщинистые	40	35	5	25	0,7143

$$\chi^2 = \frac{(E-T)^2}{T} = 1{,}528$$

Как всегда, далее производится сравнение с табличным значением. Число степеней свободы, как и в предыдущем случае, $\nu = K-1=4-1=3$. При доверительной вероятности $P_{дов}=95\%$ табличное значение $\chi^2_{порог}=7{,}81$. Мы видим, что $\chi^2 < \chi^2_{порог}$. Что это означает? Вы уже можете точно сказать: «Это значит, что **отличия недостоверны**». Стало быть, у нас нет оснований для утверждения, что в нашем эксперименте наблюдаются существенно другие результаты, нежели предсказывает теория. Но очень важно отметить, что полученные данные ни в коей мере не доказывают, что в нашем случае имеет место расщепление 9:3:3:1. Помните, что доказать равенство мы с помощью статистики не можем! То, что мы не получили достоверных отличий от теоретического соотношения, говорит лишь о том, что мы не отвергли нулевую гипотезу и можем ее использовать как **рабочее предположение**. Всегда надо помнить, что при большем объеме выборки могут быть получены другие результаты, только вероятность этого очень низка. Вот такая упрямая штука – логика статистики!

Теперь перейдем к разговору об анализе размерно-возрастной структуры популяции. С точки зрения размерно-возрастного анализа все особи в популяции отличаются либо по своему возрасту, либо по размеру. Поскольку этот тип анализа структуры популяции имеет свои тонкости чисто биологического плана, то на них и надо остановиться.

Как правило (хотя не всегда), исследователя интересует, прежде всего, возраст особей. Однако в большинстве случаев возраст измерить крайне трудно. Самой благоприятной оказывается ситуация, когда у организмов имеется какая-то часть тела, на которой этот возраст можно «прочитать». Например, на спиле дерева видны кольца нарастания, на чешуе некоторых рыб и раковинах моллюсков видны кольца зимней остановки роста, у лошадей возраст можно определить по степени изношенности зубов. Однако такого рода признаки встречаются крайне редко и далеко не любой объект можно изучить таким способом. Более того, если вам посчастливится поработать в тропиках, то там вследствие малозаметности сезонных явлений вы таких признаков не найдете. Поэтому гораздо чаще исследователь исходит из анализа не возраста особей, а их размера. При этом предполагается, что чем старше особь, тем больше ее размер. Что, как не трудно заметить, тоже не всегда справедливо, поскольку есть виды, у которых с возрастом рост замедляется или вовсе останавливается. Однако и это препятствие можно обойти, изучая, например, не рост, а вес организмов или и то, и другое. В любом случае, анализируя популяцию того или иного вида, лучше использовать все доступные для измерения признаки (возрастные признаки, размер тела и отдельных его частей, вес организма и т.п.). Однако при этом не стоит увлекаться, так как анализ большого числа признаков на одной особи может занять очень много времени и вы из-за трудоемкости не сможете обработать достаточно большую выборку.

Итак, вы приступили к анализу размерно-возрастной структуры популяции. Какова ваша цель? Главной целью такого рода исследований оказывается выявление размерно-возрастных групп и определение соотношения их численностей в популяции.

Разберем ход анализа на конкретном примере. В таблице приведены результаты измерения длины тела (L) в мм и веса (P) в мг рачков-бокоплавов Pontoporeia femorata, отловленных в мелководном заливе Белого моря.

Таблица 14. Длина тела и вес Pontoporeia femorata

L	P	L	P	L	P	L	P
5,7	4	4,6	4	5,2	4	5	6
6,6	10	6,2	7	5,3	6	6,2	11
7,3	11	5,8	7	6,4	7	5,6	6
4,7	5	4,4	3	6,2	8	11,6	64
5,5	6	5,1	5	6,4	6	13,4	64
5,4	5	6,4	6	5,7	6	5,2	10
6,2	10	4,4	5	6,2	10	6,1	7
6,3	11	10,3	42	5,1	4	6,4	7
10,5	32	12,1	63	5,5	5		
13,5	56	12,3	48	6,7	8		
6,3	10	10,8	39	5,8	10		
5,9	7	5,1	6	6,8	9		
6,2	6	5,2	5	9,5	27		
5,8	6	5,7	7	5,4	6		
4,3	4	5,6	7	11,2	41		
4,7	3	4,2	2	5,2	5		
5,1	4	4,9	4	6	6		
5,2	7	4,3	3	5,9	8		
5,8	6	4,2	3	4,9	3		
4,7	4	5	4	5,2	4		
5,6	6	15,5	98	6	6		
4,3	3	13,4	68	6,5	10		
5,8	7	13,2	79	6,1	6		
4,4	2	13,3	60	5,3	6		
4,3	5	6,1	8	6,3	10		
9,2	22	4,6	3	6,9	10		
4,2	3	6,5	10	6	10		
4,4	3	4,8	5	4,9	6		
6,7	8	6,1	7	5,7	10		

ВСЕГО: 95 особей.

Приведенные данные и являются основой для первого шага изучения размерно-возрастной структуры популяции – *частотного анализа*[16].

Итак, необходимо провести частотный анализ. Суть этого анализа сводится к построению уже известного нам выборочного распределения, т.е. к построению вариационного ряда. Однако, если в предыдущих случаях, построение вариационного ряда сводилось к подсчету количества особей, относящихся к четко отличимым, как говорят, дискретным, классам, то в данной ситуации мы имеем непрерывную величину. Как поступить в этой ситуации? Можно, конечно, поступить просто, например, решить, что в качестве классов мы будем использовать значения конкретных измерений. То есть, в

[16] Частотный анализ – метод очень простой и вместе с тем очень мощный, он позволяет понять структуру ваших данных. Поэтому, когда в вашем распоряжении много числовых данных (не важно для каких целей вы собираетесь их использовать), лучше сначала провести их частотный анализ.

выборке мы будем подсчитывать число объектов, имеющих каждое конкретное значение измеренного признака. В нашем случае это приведет к следующему виду вариационного ряда (для простоты пока остановимся на анализе только длины тела).

Таблица 15. Вариационный ряд при классовых значениях, равных отдельным измерениям

L, мм	Частота (число особей)
4,2	3
4,3	4
4,4	4
4,5	2

и т. д.

Не будем продолжать этот ряд, поскольку в данном случае так его строить не совсем правильно и неудобно. Почему? Во-первых, таких классов получится очень много. Нетрудно подсчитать, что если минимальное значение длины составляет 4,2, а максимальное 15,5, то число классов будет равно 113 (!). При незначительном числе особей большое число классов будет пустым. Во-вторых, и это очень важно, когда производятся измерения с помощью грубых приборов, при таком дробном вариационном ряде начинают играть роль погрешности измерения и округления величин. Наши глаза несовершенны и при измерениях одного и того же объекта разными людьми очень часто получаются немного разные результаты. Поэтому чрезмерная дробность при частотном анализе вредна, поскольку становятся видны посторонние «шумы», не имеющие никакого отношения к реальной структуре популяции.

Вследствие описанных причин частотный анализ лучше проводить несколько иначе. Для этого разделение на классы ведут в соответствии с выбранным *классовым шагом*. Всю шкалу варьирования признака разбивают на несколько *интервалов*. Как выбрать классовый шаг? Можно сделать это чисто формально. В статистике есть специальная формула.

$$i = \frac{L\max - L\min}{1 + 3{,}32\lg(N)}$$

Здесь i – величина классового шага, Lmax – максимальное значение признака, Lmin – минимальное значение признака, lg – знак десятичного логарифма, N – объем выборки.

Вычислим значение классового шага для данных по размерам Pontoporeia femorata.

$$i = \frac{15{,}5 - 4{,}2}{1 + 3{,}32\lg(95)} = 1{,}49 \approx 1{,}5.$$

После того, как получена величина классового шага, можно приступить к разбивке всей шкалы варьирования признака на интервалы. Первая нижняя граница интервала определяется как Lmin, а первая верхняя граница – как Lmin+i, следующая нижняя граница – это верхняя граница предыдущего интервала. В данном случае мы получим следующие интервалы.

4,2 - 5,7
5,7 - 7,2
7,2 - 8,7
8,7 - 10,2
10,2 -11,7
11,7 -13,2
13,2 -14,7
14,7 -16,2
16,2 -17,7

Однако предложенный метод разбивки на интервалы удобен только тогда, когда популяция подвергается разовому анализу и далее она ни с какой другой популяцией сравниваться не будет. Если же вы решили проследить размерную структуру популяции на протяжении нескольких лет, то вполне может сложиться ситуация, при которой объемы выборок в разные годы будут отличаться или изменится размах варьирования размеров. Это приведет к тому, что классовые шаги в разных выборках будут разными, а, стало быть, произойдет и разбивка на разные интервалы. Это сделает сравнение невозможным. Поэтому чаще биологи выбирают для каждого вида некоторый свой единый, стандартный шаг. Так, например, опыт показал, что для Pontoporeia femorata наиболее удобным оказывается классовый шаг, равный 1 мм. Выбор классового шага, таким образом, очень часто происходит в соответствии с интуицией исследователя и знанием особенностей варьирования признака у изучаемого объекта и точности измерений. Пока у вас не появился опыт, лучше все же пользоваться приведенной формулой. Однако для дальнейших построений мы все-таки воспользуемся классовым шагом, равным 1 мм. Тогда мы получим следующие интервалы.

Интервалы для построения вариационного ряда длины тела *Pontoporeia femorata*, построенные при классовом шаге i=1:

4,0 - 5,0
5,0 - 6,0
6,0 - 7,0
7,0 - 8,0
8,0 - 9,0
9,0 - 10,0
10,0 - 11,0
11,0 - 12,0
12,0 - 13,0
12,0 - 13,0
13,0 - 14,0
14,0 - 15,0
15,0 - 16,0
16,0 - 17,0

Теперь необходимо подсчитать количество особей, чьи размеры попадают в соответствующий интервал. Это и будет вариационный ряд. Однако здесь необходимо сделать одно очень важное замечание. Нетрудно заметить, что некоторые значения, размеров могут быть равны одновременно и верхней границе одного интервала и нижней границе следующего. В такой ситуации для распределения конкретных величин по классам используют очень простое правило: **значение, равное «граничному значению», помещается всегда справа от границы**. Так, например значение равное 10,0 попадет в класс 10,0-11,0, а значение 5,0 – в класс 5,0-6,0.

Таблица 16. Вариационный ряд размеров тела Pontoporeia femorata

L, мм	**Частота, экз.**
0,0-1,0	0
1,0-2,0	0
2,0-3,0	0
3,0-4,0	0
4,0-5,0	22
5,0-6,0	33
6,0-7,0	24
7,0-8,0	1
8,0-9,0	0
9,0-10,0	2
10,0-11,0	3
11,0-12,0	2
12,0-13,0	2
13,0-14,0	5
14,0-15,0	0
15,0-16,0	1
16,0-17,0	0

Обратите внимание на то, что в данном вариационном ряду мы начали «отсчет» от нуля. В принципе, не важно, откуда начинать, но для удобства восприятия иногда полезно начинать от начала координат (хотя иногда это может и затруднить восприятие, например, когда значения признака очень сильно отличаются от нуля, а величина классового шага очень мала).

Итак, вариационный ряд построен. Что делать дальше? Следующим ходом в анализе размерно-возрастной структуры популяции будет построение *частотной гистограммы* или *частотного полигона*. Построить гистограмму или полигон очень просто. Необходимо изобразить на листе бумаги координатные оси (для данного случая достаточно лишь положительных осей). По оси OX будем изображать величину признака (в нашей ситуации длину), а по оси OY – частоту.

У многих школьников вызывает недоумение, как на оси OX отложить значение признака, когда все измерения у нас уже разбиты на интервалы. Прежде всего, чтобы избежать подобных вопросов и далее, давайте условимся, что при графическом представлении данных **не надо стремиться к абсолютной точности**. Все равно, на графиках, гистограммах и т.п. видна лишь общая картина. Некоторые исследователи в такой ситуации поступают очень просто. Они разбивают ось OX на несколько равных промежутков (по числу классов), а снизу подписывают значение классов (рис. 4). Другой способ – это откладывать на оси OX лишь граничные точки интервалов. В нашей ситуации на оси абсцисс наиболее удобно отложить следующие точки: 0; 1; 2; 3; ..... 15; 16; 17 (рис. 4). Можно поступить и еще одним способом: в качестве меток на оси OX можно откладывать среднеклассовые значения. В нашем случае такими значениями будут 0,5; 1,5; 2,5; ... 14,5; 15,5; 16,5 (рис. 4). Обычно при построении гистограммы используют первый или второй способы, а при построении полигона – третий.

При построении графиков, гистограмм и других видов графического изображения материала очень важно правильно выбрать масштаб. Масштаб должен быть таким, чтобы данные воспринимались наилучшим способом. Для достижения этой цели можно воспользоваться несколькими простыми правилами.

1. Если значения (абсцисс или ординат) существенно отличаются от нуля, то совершенно необязательно на графике в качестве начала оси выбирать «0». Можно начать отсчет с любого другого значения.
2. На осях OX и OY могут (и должны) быть разные масштабы.
3. При выборе масштаба сначала задайте длину осей для будущего графика, а уж потом выбирайте на ней масштаб. В противном случае (если вы сначала выберете масштаб, а затем начнете рисовать ось) длина оси может быть слишком велика, она может даже не поместиться на одном листе. Это будет некрасиво и неинформативно, так как глаз с близкого расстояния воспринимает обычно небольшие рисунки.
4. Для выбора длины оси надо выявить минимальное и максимальное значение, которое будет отложено на ней. И уже потом вычислить масштаб.

Теперь можно приступить к построению графического изображения. Оформив ось OX (нанеся границы интервалов или среднеклассовые значения), начинаем откладывать точки, соответствующие тем или иным частотам, которые откладываются по оси OY. У этих точек в качестве ординаты будет частота, а в качестве абсциссы любая точка соответствующего интервала, отложенного по оси OX, как правило, в качестве такой точки берется середина интервала. Если вы по оси OX отложили среднеклассовые значения, то они и будут служить абсциссой для этой точки. Далее, если вы соедините последовательно все точки линией, то вы получите *частотный полигон*. Если построите столбики, упирающиеся в эти точки, то получится *частотная гистограмма* (рис. 5).

Следующий этап – самый сложный. Заключается он в осмыслении той информации, которая приведена на гистограмме или полигоне. Для того, чтобы научиться получать эту информацию, необходимо вспомнить, что в генеральной совокупности анализируемая величина распределена в соответствии с некоторым законом. Ранее мы рассматривали в качестве такого закона функцию нормального распределения, но надо постоянно помнить, что величина в генеральной совокупности может иметь и другой закон распределения (в математической статистике описано довольно много таких законов). Однако для простоты давайте остановимся все на том же нормальном распределении. Если в генеральной совокупности действительно оно имеет место, то график, отражающий это распределение, будет иметь уже привычный нам вид колоколообразной кривой (рис. 2).

Однако возможна и другая ситуация, когда вся совокупность объектов включает в себя несколько подсовокупностей, каждая из которых описывается своим собственным нормальным распределением (они отличаются по параметрам распределения: X и S). В такой ситуации в одной генеральной совокупности будут иметь место несколько нормальных распределений (рис. 6).

Например, в популяции некоторых животных представлены самцы и самки, но самки имеют меньшие размеры. В результате этого, распределение размеров в генеральной совокупности будет состоять из двух распределений, отличающихся параметром X. Но если бы мы не умели отличать самцов и самок, а просто анализировали бы размеры тела животных, то распределение было бы не колоколообразным, а двувершинным, или, как говорят, *бимодальным* (рис. 6). Если в генеральной совокупности много подмножеств, то оно называется *полимодальным* (рис. 7).

Поскольку выборочное распределение – это образ генерального распределения, то наличие полимодальности в таком распределении – это верный признак того, что генеральная совокупность состоит из нескольких подмножеств. Что это за подмножества, статистика дать ответа не может. Трактовка этих подмножеств – задача сугубо биологическая. Для ответа на вопрос о том, что это за подмножества, надо очень хорошо знать объект.

Рисунок 4. Разные варианты подписи оси OX при построении частотного распределения.

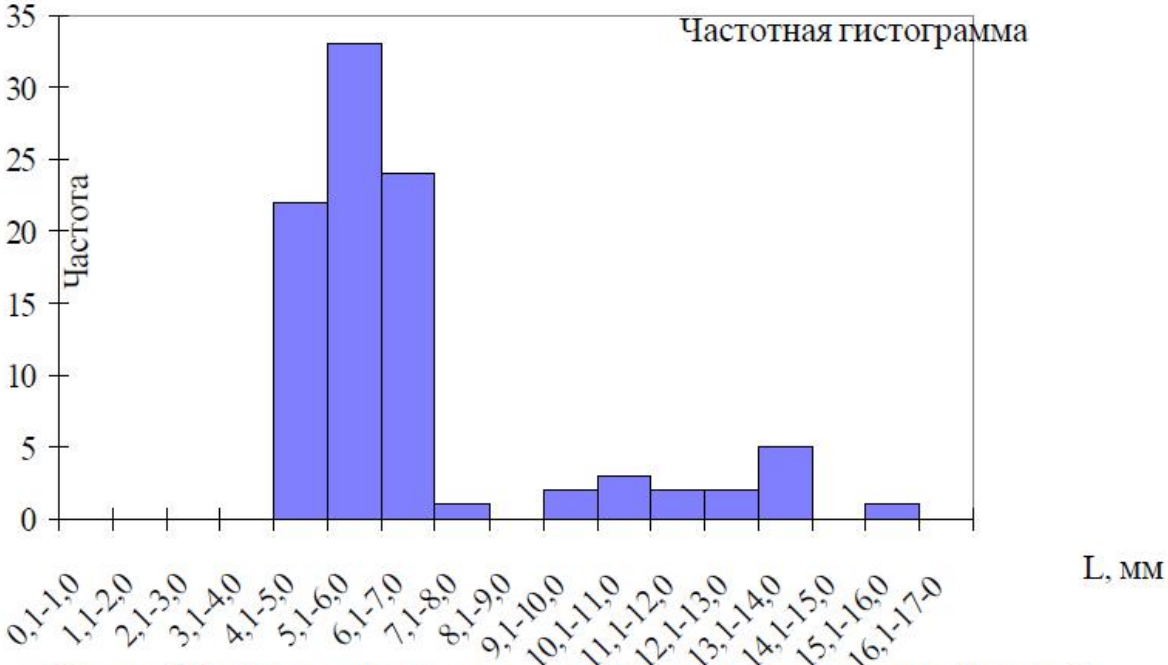

Рисунок 5. Частотное распределение значений длины тела бокоплавов Ponoporeia femorata.

49

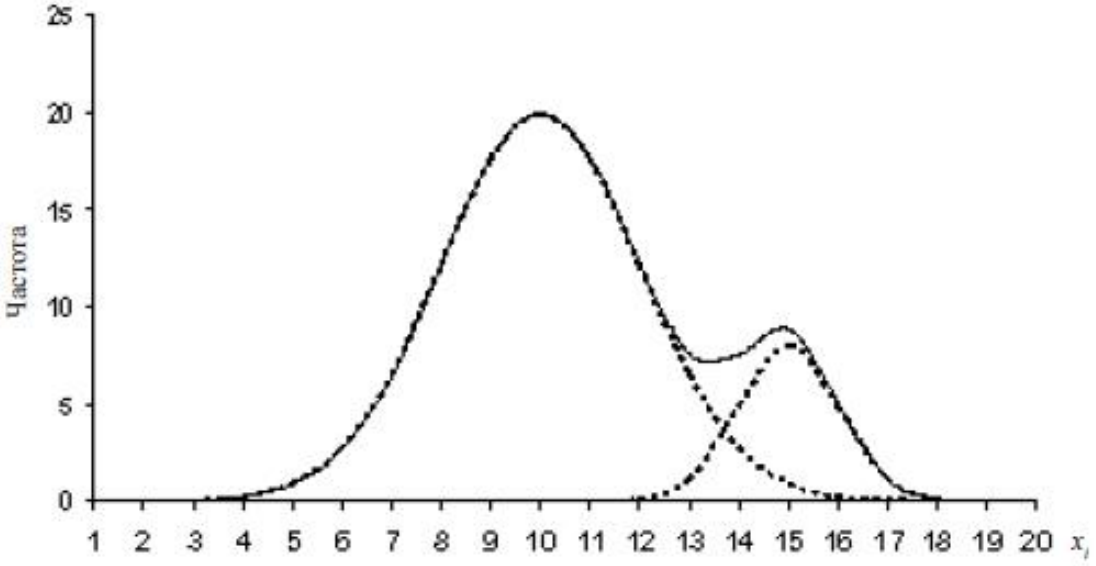

Рисунок 6. Двувершинное (бимодальное) частотное распределине, являющееся результатом наложения двухс нормальных распределений.

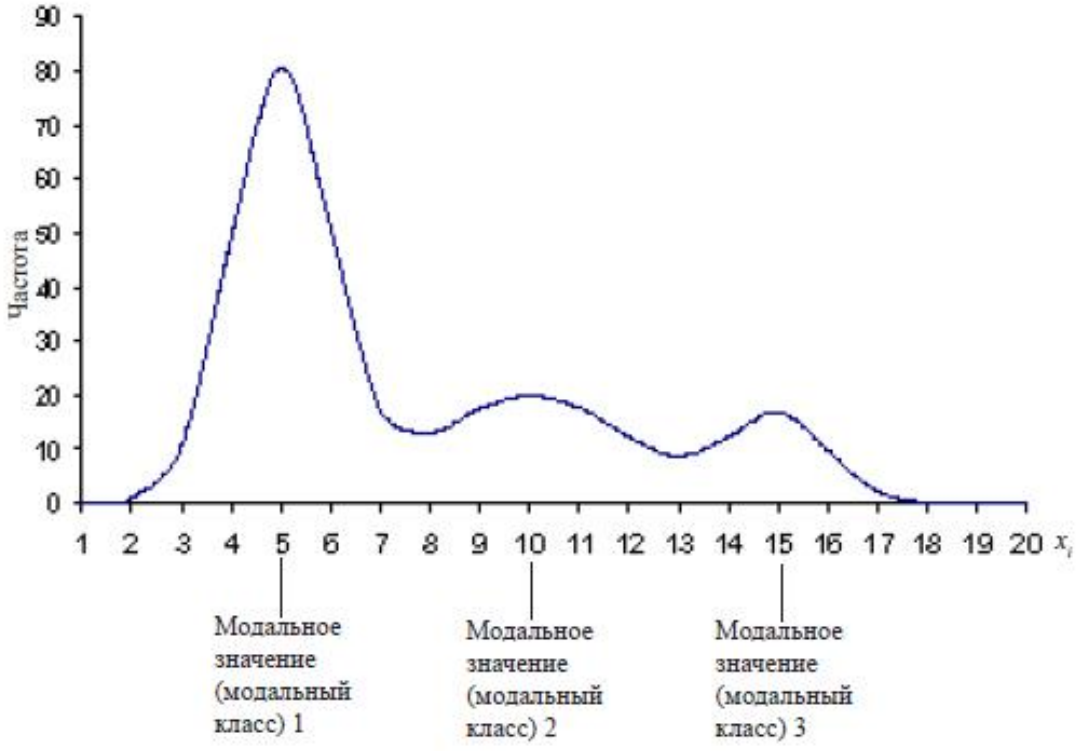

Рисунок 7. Полимодальное частотное распределение.

Вернемся к нашему примеру с Pontoporeia femorata. Распределение по длине тела этих бокоплавов имеет явно полимодальный характер (рис. 5). Известно, что бокоплавы растут линьками, то есть, у них наблюдается дискретный рост. Это позволяет предположить, что наблюдаемые пики в распределении – это отражение подмножеств, соответствующих возрастным когортам. Внимательный читатель может заметить, что в выборочном распределении размеров бокоплавов пиков можно насчитать несколько (до четырех). Это становится видно, если в качестве «маркеров» пиков рассматривать классы, где находится особей больше, чем в соседних классах. Такие классы называются *модальными*. В нашем выборочном распределении модальными оказываются следующие классы: 5,1-6,0; 10,1-11,0; 13,1-14,0; 15,1-16,0. Означает ли это, что надо каждый из этих пиков рассматривать как проявление самостоятельного подмножества в генеральной совокупности? Формально ответить на этот вопрос нельзя! Может быть, да, а, может быть, и нет. На чем же остановиться? Если вы анализируете только один признак, то окончательно ответить на вопрос о том, сколько же пиков в вашем распределении, может дать только очень большая выборка. Чем больше объем выборки, тем ближе выборочное распределение к генеральному, а, стало быть, менее выраженными становятся разные шумы, связанные с неточностями измерения, малым объемом выборки и т. д. Если увеличение объема выборки не убирает пики, выявленные ранее, то это означает, что они отражают реально существующие внутрипопуляционные подгруппы.

Несколько проще решить вопрос о количестве подмножеств в вашей совокупности, если вы анализируете не один, а два или много признаков. Самый простой метод при наличии такого материала – это построение *скеттер-диаграммы***.**

Если у вас измерено только два признака, то скеттер-диаграмма, или точечная диаграмма, строится очень просто. Для ее построения надо начертить две координатные оси. На одной из них будут откладываться значения одного признака, а на другой – значения другого. При этом заметьте, что здесь уже нет необходимости выделять классы. Каждый отдельный экземпляр вашего объекта будет нанесен на координатной плоскости в виде точки с координатами, соответствующими значениям измеренных признаков.

Вернемся к примеру с бокоплавами. Мы измеряли два параметра: длину тела и вес организма. Если нанести все 95 пар измерений на координатную плоскость, то получим облако точек (рис. 8). На этом облаке четко различимы две группы точек, два «сгущения». Одно располагается в области малых значений длины и веса. Это сгущение достаточно компактно. Второе сгущение – в области более высоких значений, но это скопление более «размазано», оно обладает большей дисперсией, т.е. большим рассеянием. Если внимательно приглядеться, то можно увидеть и еще одну отдельность, состоящую всего из одной особи в области самых высоких значений длины и веса.

Рассмотрение скеттер-диаграммы и частотного распределения позволяет нам придти к выводу, что в популяции бокоплава Pontoporeia femorata присутствуют три подмножества особей, которые мы, учитывая дискретный характер роста, можем трактовать как возрастные группы. На основании полученных статистических данных, мы можем сделать биологический вывод о том, что в изученной популяции бокоплавы живут до трех лет. Кроме того, можно сделать еще ряд любопытных выводов. Например, можно высчитать, какова вероятность особи первого года дожить до второго, а особи второго – до третьего. Для этого достаточно подсчитать отношение численности особей второй размерной группы к численности первой и численности третьей ко второй соответственно. Можно сделать и еще много интересного. Главное, что мы смогли обоснованно разделить всю популяцию на подмножества.

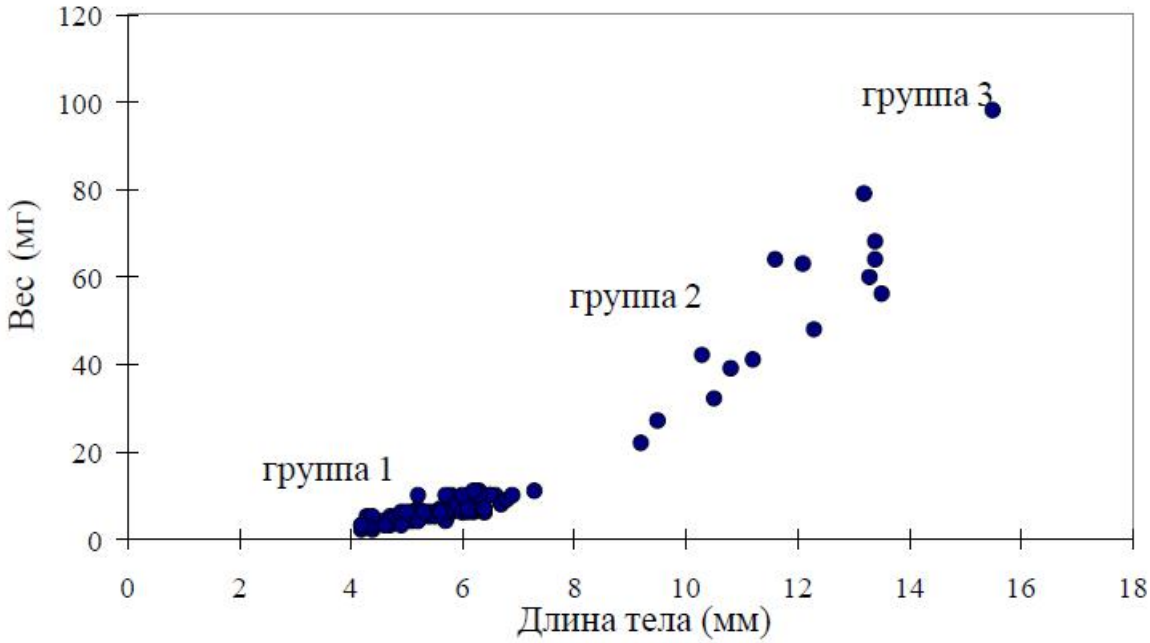

Рисунок 8. Скеттер-диаграмма, отражающая соотношение длины и веса тела у бокоплавов Pontoporeia femorata.

Особый случай анализа размерно-возрастной структуры популяции – это случай, когда отдельные распределения, входящие в состав общего полимодального распределения, сильно перекрываются. В этой ситуации четко провести границу между совокупностями нельзя. Существует много разнообразных ухищрений, которыми пользуются биологи для того, чтобы разделить такие подмножества. Все эти методы очень сложны, и в пределах данного пособия мы просто не можем их описать. Поэтому можно предложить вам менее строгий, но достаточно простой способ разделения популяции на подмножества – в анализ можно включать только особей, имеющих модальные размеры. Например, пусть вам надо выделить в популяции особей разных возрастных групп, и проанализировать их распределение в пространстве. Однако наблюдается существенное перекрывание размерных характеристик у разных возрастов (с полной уверенностью нельзя отнести каждую особь к той или иной совокупности). В такой ситуации вы можете включить в анализ только тех особей, которые имеют размеры, соответствующие модальным классам.

Анализ размерно-возрастной структуры популяции имеет и свои ограничения, суть которых, впрочем, лежит не в статистике, а в биологической природе объектов. Например, не всегда выделенные совокупности можно четко трактовать как возрастные. Они могут иметь и совершенно другую природу. Но в любом случае, если такие совокупности выделяются, их сбрасывать со счетов нельзя. Они что-нибудь да означают!

Глава 2.3. Описание взаимосвязи величин

Речь в этой главе пойдет о *корреляционном анализе*. В практике биологических исследований очень часто стоит вопрос о том, имеется ли какая-нибудь взаимосвязь между явлениями. Здесь нам надо немного отойти от чистой статистики и чуть-чуть поговорить о философии.

Всем хорошо известно, что существует так называемая причинно-следственная связь между явлениями. Например, связь между ветром и деревьями. Ветер дует – деревья качаются. Это взаимосвязь причинно-следственная. Ветер является причиной раскачивания деревьев. Но вспомните себя в глубоком детстве! Описанное соотношение было для вас абсолютно неочевидным – вполне вероятным казалось, что раскачивание деревьев и есть причина возникновения ветра.

Не надо думать, что подобные перевороты в сознании происходят только с детьми, когда те взрослеют, аналогичная ситуация много раз бывала и с взрослыми учеными. Очень часто причинно-следственные связи, которые воспринимались одним образом, позднее начинают восприниматься иначе. Но при всех подобных изменениях в понимании процессов само наличие связи между явлениями не подвергается сомнению. Сколько бы ни было вам лет, связь между ветром и раскачиванием деревьев останется. Вот это и называется *корреляцией*.

Корреляция – это не что иное, как наличие взаимосвязи между явлениями. Причем, заметьте – ничего о причинах и следствиях здесь не говорится. Наличие корреляции не означает, что между явлениями есть причинно-следственная связь, но какая-то связь есть. Она может быть чисто случайной, а может быть и нет. **Поиск причин этой связи не входит в компетенцию статистики!** Статистика занимается лишь ее **выявлением**. Корреляционный же анализ призван количественно измерить силу этой взаимосвязи.

Разберем простейший вариант корреляционного анализа, применяемый для анализа связи между явлениями, которые регистрируются в виде количественных данных. Для этого вернемся к примеру с рачками-бокоплавами Pontoporeia femorata, который мы разобрали ранее (см. главу 2.2). У каждого рачка мы изучали два параметра – длину тела и вес. Оба этих параметра были измерены количественно. Мы можем поставить вопрос о наличии связи между этими величинами. Для ответа на поставленный вопрос надо провести корреляционный анализ. Для простоты возьмем не всю выборку, а лишь первые 29 значений.

Таблица 17. Параметры тела (длина и вес) рачков-бокоплавов Pontoporeia femorata

L	**P**		**L**	**P**
5,7	4		4,7	3
6,6	10		5,1	4
7,3	11		5,2	7
4,7	5		5,8	6
5,5	6		4,7	4
5,4	5		5,6	6
6,2	10		4,3	3
6,3	11		5,8	7
10,5	32		4,4	2
13,5	56		4,3	5
6,3	10		9,2	22
5,9	7		4,2	3
6,2	6		4,4	3
5,8	6		6,7	8
4,3	4			

Для измерения силы взаимосвязи между количественными параметрами применяется *коэффициент корреляции Браве-Пирсона* (r). Этот коэффициент может принимать значения от -1 до 1. Если r =1 или r = -1, то корреляция очень сильная - явления очень сильно взаимосвязаны. Если r =0, то явления не связаны друг с другом. Важное значение имеет и знак коэффициента корреляции. Если коэффициент положительный, то это означает, что чем больше выражено первое явление, тем больше выражено второе. Если коэффициент отрицательный, то связь обратная - чем больше выражено первое явление, тем меньше выражено второе. Вычисление этого коэффициента производится по следующей формуле:

$$r = \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 \quad (y_i - \bar{y})^2}}.$$

В этой формуле x_i, y_i – значения каждого отдельного измерения признака X и Y у i-того объекта, $\bar{x}$ и $\bar{y}$ – средние значения признаков X и Y, соответственно. Эту формулу можно записать несколько иначе – в виде, более удобном для вычисления:

$$r = \frac{x_i y_i - \frac{x_i \quad y_i}{N}}{\sqrt{(\quad x_i^2 - \frac{(\quad x_i)^2}{N})(\quad y_i^2 - \frac{(\quad y_i)^2}{N})}}.$$

Для вычисления по этой формуле надо рассчитать входящие в нее величины.

Таблица 18. Ход вычисления коэффициента корреляции Браве-Пирсона

	L	**P**
Сумма L_i (*или* P_i)	174,6	266,0
Квадрат суммы $(\quad L_i)^2$ (*или* $(\quad P_i)^2$)	30485,2	70756,0
Сумма квадратов $L_i^{\ 2}$ (*или* $P_i^{\ 2}$)	1165,3	5756,0
Сумма произведений $L_i P_i$	2193,10	
Объем выборки N	29	

Подставляем полученные величины в формулу:

$$r = \frac{2193{,}10 - \frac{174{,}6 \;\; 266{,}0}{29}}{\sqrt{(1165{,}3 - \frac{30485{,}2}{29})(5756{,}0 - \frac{70756{,}0}{29})}} = 0{,}96.$$

Итак, мы получили достаточно высокий коэффициент корреляции. Однако мы должны еще проверить, достоверно ли его значение. Ведь мы изучили не всех бокоплавов на Земле, а лишь небольшую выборку. Для ответа на вопрос о достоверности коэффициента корреляции проще всего использовать так называемую таблицу пороговых значений (таблица III). Для работы с ней нам надо знать, как обычно, доверительную вероятность и число степеней свободы. Первую величину мы принимаем равной $P_{дов}$=95%. Число степеней свободы вычисляется по следующей формуле:

$$\nu = N-2,$$

где N – число пар наблюдений. При $\nu = N-2 = 29-2 = 27$ и $P_{дов}$=95% пороговое значение коэффициента корреляции будет равно $r_{порог.}$=0,367. Эмпирическое значение коэффициента

(здесь необходимо брать коэффициент по модулю, т.е. без учета знака) значительно превышает пороговое. Таким образом, вычисленная корреляция в высшей степени достоверна, а значит, мы можем с уверенностью говорить, что между длиной тела бокоплава и его весом связь существует. Более того, мы можем однозначно утверждать, что чем больше длина тела животного, тем больше его вес.

Внимание, контрольный вопрос. А можем ли мы на основании проведенных исследований утверждать, что увеличение длины тела рачка является причиной увеличения его веса? Конечно же, нет! Ведь мы вычислили лишь корреляцию, то есть, только установили наличие связи, но не провели изучение причинно-следственных отношений. Изучение этого аспекта – это уже не задача статистики, а задача биологии.

Итак, мы рассмотрели методы изучения взаимосвязи между явлениями, которые описываются количественными данными. Но у приведенного метода есть одно очень важное ограничение. Этот метод применим не для всех данных, которые выражены численно. Он применим лишь для тех величин, которые имеют **нормальное распределение** (см. главу 1.9). Проверка нормальности распределения величины – задача достаточно сложная, и здесь мы ее рассматривать не будем. Однако многие начинающие биологи по умолчанию считают, что большинство признаков биологических объектов, которые можно численно измерить, подчиняются закону нормального распределения. Вместе с тем, часто это не так. Например, не подчиняются нормальному распределению всевозможные доли (проценты), зачастую, временные промежутки и расстояния на местности. Как же быть в тех случаях, когда вы не уверены в том, что изученная величина подчиняется закону нормального распределения? Для решения задач, в которых используются подобные величины, лучше пользоваться методами непараметрической статистики. В случае корреляционного анализа используются *ранговые коэффициенты корреляции*. Самый распространенный метод анализа в таких случаях был предложен Спирменом. Поэтому коэффициент, предложенный им, обычно называют *коэффициентом Спирмена* (r_s).

Рассмотрим применение такого коэффициента на следующем примере. Предположим, вы задались целью выяснить, существует ли взаимосвязь между следующими явлениями: степенью зараженности прудовиков партенитами печеночного сосальщика и близостью водоема, в котором обитают прудовики, к пастбищам, на которых идет выпас деревенских коров. Первое явление вы оценили с помощью доли зараженных моллюсков в некоторой выборке, взятой в том или ном водоеме. Вторую величину вы оценивали как расстояние от водоема до ближайшего пастбища. Все полученные данные вы свели в следующую таблицу.

Таблица 19. Удаленность водоема от пастбища и доля зараженных прудовиков

№ водоема	**Расстояние до пастбища (м)**	**Доля зараженных прудовиков (%)**
1	0	80
2	12	20
3	11	20
4	0	60
5	1000	5
6	300	2
7	10	40
8	10	20
9	25	50
10	33	20
11	80	20
12	500	10

13	255	10
14	100	20
15	121	22
16	0	65

Несмотря на то, что в нашей таблице приведены количественные данные, взять и вычислить коэффициент корреляции Браве-Пирсона нельзя. Это связано с тем, что ни расстояния, ни доли зараженных моллюсков не имеют нормального распределения. Для выявления корреляции в данном случае необходимо произвести ранжирование данных. Эта процедура очень простая, но требует достаточно внимательной работы (особенно при обработке больших массивов данных).

Итак, давайте осуществим процедуру ранжирования. Для начала, все значения в каждом изученном признаке (параметре) надо упорядочить по мере возрастания. Разберем этот первый шаг на примере данных, оценивающих расстояние от пруда до пастбища.

Таблица 20. Упорядоченные данные по расстоянию (первый шаг ранжирования).

№ водоема	**Расстояние до пастбища**
1	0
4	0
16	0
7	10
8	10
3	11
2	12
9	25
10	33
11	80
14	100
15	121
13	255
6	300
12	500
5	1000

Далее каждому баллу присваивается его порядковый номер в ряду упорядоченном по мере возрастания значения признака.

Таблица 21. Присвоение порядкового номера значениям расстояния (второй шаг ранжирования).

№ водоема	Расстояние до пастбища	Порядковый номер
1	0	1
4	0	2
16	0	3
7	10	4
8	10	5
3	11	6
2	12	7
9	25	8
10	33	9
11	80	10
14	100	11
15	121	12
13	255	13
6	300	14
12	500	15
5	1000	16

Эти порядковые номера и были бы рангами, если бы среди них не было бы повторяющихся значений. Так, например, водоемы 1, 4 и 16 имеют нулевое расстояние от пастбища, но им присвоены разные порядковые номера. Аналогично, водоемы 7 и 8 расположены на одинаковом расстоянии. В такой ситуации всем объектам, имеющим одинаковые значения признака, присваивается ранг, равный среднему значению их порядковых номеров[17].

Таблица 22. Ранги для значений расстояний (заключительный этап ранжирования)

№ водоема	Расстояние до пастбища	Ранги для значений расстояния
1	0	2
4	0	2
16	0	2
7	10	4,5
8	10	4,5
3	11	6
2	12	7
9	25	8
10	33	9
11	80	10
14	100	11
15	121	12
13	255	13
6	300	14
12	500	15
5	1000	16

[17] Крайне желательно спланировать сбор материала так, чтобы повторяющихся значений было бы поменьше, в противном случае придется вычислять некоторые поправки, суть которых мы в данном пособии обсуждать не будем.

Проведя аналогичные операции с величинами, описывающими зараженность моллюсков, переписываем исходную таблицу.

Таблица 23. Ранговые оценки расстояния от водоема до пастбища и зараженности прудовиков

№ водоема	Расстояние до пастбища	Доля зараженных прудовиков	Ранги для значений расстояния	Ранги для значений зараженности
1	0	80	2	16
2	12	20	7	7,5
3	11	20	6	7,5
4	0	60	2	14
5	1000	5	16	2
6	300	2	14	1
7	10	40	4,5	12
8	10	20	4,5	7,5
9	25	50	8	13
10	33	20	9	7,5
11	80	20	10	7,5
12	500	10	15	3,5
13	255	10	13	3,5
14	100	20	11	7,5
15	121	22	12	11
16	0	65	2	15

Внимание! Если вы по каким-то причинам выкинули некоторые пары наблюдений или, наоборот, добавили их, то всю процедуру ранжирования надо повторить сначала!

Теперь можно приступить к вычислениям коэффициента ранговой корреляции Спирмена. Он рассчитывается по следующей формуле:

$$r_s = 1 - \frac{6 \quad (x-y)^2}{N^3 - N}.$$

В этой формуле x - ранг первого признака в паре, y - ранг второго признака в паре, N - число пар.

Для удобства вычислений построим такую такую таблицу.

Таблица 24. Ход вычисления коэффициента Спирмена.

Ранги для значений расстояния	**Ранги для значений зараженности**	**x-y**	**$(x-y)^2$**
2	16	-14	196
7	7,5	-0,5	0,25
6	7,5	-1,5	2,25
2	14	-12	144
16	2	14	196
14	1	13	169
4,5	12	-7,5	56,25
4,5	7,5	-3	9
8	13	-5	25
9	7,5	1,5	2,25
10	7,5	2,5	6,25
15	3,5	11,5	132,25
13	3,5	9,5	90,25
11	7,5	3,5	12,25
12	11	1	1
2	15	-13	169
Сумма $(x-y)^2$			1211

N=16

N^3=4096

Теперь можно подставить все значения в формулу.

$$r_s = 1 - \frac{6\ \ 1211}{4096 - 16} = -0{,}781$$

Полученное значение коэффициента корреляции далее необходимо сравнить с пороговым значением ($r_{s\text{-порог}}$), найденным по специальной таблице (табл. IV). Эта таблица очень похожа на таблицу пороговых значений коэффициента корреляции, которую мы использовали ранее (табл. III). Однако, поскольку мы работали не истинными значениями признаков, а с их рангами, то появились небольшие отличия. Для ν=16-2=14 при $P_{дов}$=95% оно равно $r_{s\text{-порог}}$=0,501. Поскольку полученное эмпирическое значение (без учета знака) выше табличного, мы можем с высокой степенью уверенности утверждать, что два изученных явления взаимосвязаны. Более того, отрицательное значение коэффициента корреляции говорит том, что чем дальше от пастбища расположен пруд, тем меньше доля зараженных моллюсков. Совершенно очевидно, что здесь имеет место некоторая биологическая закономерность, которая лежит на поверхности (прудовики являются переносчиками сосальщиков). Однако описание этой закономерности – задача не статистическая. С помощью статистики мы лишь выявили наличие связи и доказали, что она значима (неслучайна).

Теперь перейдем к разговору об измерении силы взаимосвязи между явлениями, которые характеризуются качественными данными. Напомним, что это данные, принимающие только два значения (0 или 1, да или нет, + или –), то есть, явление есть или явления нет.

Предположим, что вы решили выяснить, существует ли взаимосвязь между наличием тараканов в квартире и ежедневными уборками в ней. Вооружившись блокнотом, вы обошли всех своих друзей и у каждого из них выясняли следующую информацию: есть

ли у них в квартире тараканы и делает ли мама вашего друга ежедневную уборку. Будучи уже грамотным исследователем, вы оформили свои наблюдения в виде таблицы.

Таблица 25. Наличие тараканов и ежедневная уборка («+» - явление присутствует, «-» - явление не присутствует)

№ опроса	**Наличие тараканов**	**Наличие ежедневной уборки**
1	-	+
2	+	+
3	+	-
4	-	-
5	-	+
6	+	-
7	-	+
8	+	+
9	-	+
10	-	+
11	+	+
12	+	-
13	+	+
14	-	+
15	+	+
16	+	-
17	+	-
18	-	-
19	+	-
20	-	-

После того как первичные результаты наблюдений оформлены, можно приступить к изучению взаимосвязи между этими явлениями. Для этого составляют так называемую *четырехпольную таблицу*. Эта таблица имеет следующий вид.

Таблица 26. Вид четырехпольной таблицы

		Явление I	
		«+»	«-»
Явление II	«+»	a	b
	«-»	c	d

В этой таблице a – число наблюдений, в которых присутствуют оба явления; b – число наблюдений, в которых имеет место второе явление, но отсутствует первое; c – число наблюдений, в которых присутствует первое явление, но отсутствует второе; d – число наблюдений, в которых не проявилось ни первое, ни второе явление.

Для нашего случая с тараканами мы можем построить следующую четырехпольную таблицу.

Таблица 27. Четырехпольная таблица для выяснения взаимосвязи уборки и наличия тараканов.

		Наличие ежедневной уборки	
		«+»	«-»
Наличие тараканов	«+»	5	6
	«-»	6	3

Теперь необходимо вычислить соответствующий показатель согласованности изменений для качественных признаков – так называемый *коэффициент ассоциации* (r_a). Этот коэффициент, как и коэффициент корреляции, отражает силу связи между явлениями. Он так же может принимать значения от -1 до 1. Если r_a =1, то явления сцеплены друг с другом – если происходит одно, то происходит и другое. Если r_a = -1, то при проявлении одного явления не проявляется другое. Если r_a =0, то явления не связаны друг с другом.

В разбираемом нами случае положительная корреляция означала бы, что ежедневная уборка увеличивает возможность встретить тараканов, а наличие отрицательной – напротив, означает, что проведение ежедневной уборки уменьшает шансы тараканов на поселение в квартире. Если бы мы получили нулевую корреляцию, то это означало бы, что никакой связи между наличием тараканов и ежедневной уборкой нет. Теперь можно перейти непосредственно к вычислению коэффициента ассоциации, которое ведется по следующей формуле.

$$r_a = \frac{ad - bc}{\sqrt{(a+b)(b+d)(c+d)(a+c)}}.$$

Подставим в эту формулу результаты наших наблюдений, приведенные в четырехпольной таблице. Тогда получим следующую запись:

$$r_a = \frac{5\ \ 3 - 6\ \ 6}{\sqrt{(5+6)(6+3)(6+3)(5+6)}} = -0{,}21.$$

Получилась величина отрицательная. Можно ли теперь утверждать, что ежедневное мытье полов приводит к уничтожению тараканов? Конечно же, пока нельзя! Мы же опросили не всех людей на земном шаре, а только двадцать своих друзей. Стало быть, нам необходимо определить достоверность вычисленного коэффициента. Это можно сделать с помощью все той же таблицы пороговых значений (табл. III[18]). Для нашего случая ν=20-2=18, $P_{дов}$=95%. Теперь, используя $P_{дов}$ и ν, находим в таблице III пороговую величину. Она равна $r_{порог}$=0,444. После того как найдено значение $r_{порог}$, необходимо его сравнить с полученным эмпирическим коэффициентом, взятым без учета знака. Если эмпирическое значение больше порогового, то корреляция достоверна. Если меньше, то недостоверна. В первом случае мы можем утверждать, что связь между явлениями с вероятностью 95% существует. Во втором – мы ничего утверждать не можем, можем лишь отметить, что достоверной связи не выявлено (что, впрочем, не означает отсутствие связи, она может быть выявится при увеличении объема выборки). Нетрудно заметить, что в нашем случае значение коэффициента ниже чем пороговое. Стало быть, у нас нет оснований для вывода, что между ежедневной уборкой и наличием тараканов есть взаимосвязь.

Завершая разговор о корреляционном анализе, необходимо ответить на следующий вопрос. Для чего нам нужно знать силу и характер связи между явлениями?

Если мы действительно установили, что между явлениями есть взаимосвязь, то нам открывается огромное поле для всевозможных действий. Например, мы, зная, что длина и вес у некоторого животного сильно взаимосвязаны, можем отказаться от трудоемкого

[18] В данной ситуации используется именно таблица III, а не таблица IV.

процесса измерения и заниматься только взвешиваниями (в дальнейшем, пользуясь специальными методами, можно перевести вес в длину). Другой пример: мы показали, что коэффициент ассоциации между двумя видами организмов в некотором сообществе достоверен и имеет знак «-». Это означает, что эти два организма избегают совместного поселения. Значит, между ними имеется некоторая биологическая связь (например, конкуренция или они тяготеют к разным условиям среды). В этом случае корреляционный анализ позволил «нащупать» некоторое новое явление, которое далее можно изучать более внимательно. Наличие корреляции, также позволяет прогнозировать некоторые свойства. Если мы, скажем, установили, что между интенсивностью окраски плода и степенью его сладости существует высокая положительная корреляция, то процесс поиска наиболее сладких плодов заметно упрощается.

Глава 2.4. Методы многомерного анализа

В этой главе мы коснемся задач, которые, в строгом смысле, не являются статистическими, так как в них не будет поставлена цель оценки генеральных параметров распределения. Вот пример такого рода задачи. Пусть мы изучили видовой состав и показатели обилия (например, плотность поселения) видов в десяти водоемах, кроме того, мы оценили степень загрязненности берегов водоема антропогенным мусором. В каждом из водоемов видовой состав и показатели обилия видов разные, различаются и уровни загрязненности. Можно ли каким-либо способом проанализировать степень сходства и различия этих водоемов и выявить зависимость населения водоема от степени его загрязненности?

Можно, конечно, эту задачу решить простым корреляционным анализом, вычислив коэффициенты корреляции обилия каждого из видов с обилием мусора. Однако это даст ответ на вопрос о том, связано ли обилие конкретного вида с количеством мусора. В то же время нас может интересовать реакция всего населения водоема как целого. Для этого и требуются методы многомерного анализа. Для обсуждения этих методов нам сначала необходимо ввести чрезвычайно важное понятие – понятие *гиперпространства признаков*, или *n-мерного пространства признаков*.

Представим себе, что каждый изученный объект обладает всего двумя признаками. С такой ситуацией мы уже сталкивались, когда анализировали размерно-возрастную структуру популяции. Если мы построим скеттер-диаграмму, то точки, соответствующие наиболее сходным объектам, расположатся на диаграмме поблизости. Точки, соответствующие наименее похожим объектам, будут максимально удалены друг от друга. Теперь давайте представим, что мы изучили не два признака, а три. Тогда скеттер-диаграмма будет располагаться не в плоскости, а в объеме. Точки же в ней сформируют подобие облака (обычно так и говорят «облако точек»). Далее надо сделать очень простой ход – необходимо предположить, что мы изучили еще один признак. Таким образом, мы будем уже иметь не трехмерное пространство, а четырехмерное. Вообразить себе это невозможно, так как наш мир трехмерен и мы адаптированы только к нему. Однако такая математическая абстракция вполне допустима. Если уж мы позволили себе работать с четырехмерным пространством, то почему бы не взять любое количество признаков – тогда мы получим n-мерное пространство. Координатами точек (объектов) в данном пространстве будут значения их признаков. Это и есть гиперпространство признаков. В нем наиболее сходные объекты (точнее - точки, им соответствующие) располагаются поблизости друг от друга, а наименее сходные – удалены[19].

[19] Все законы взаиморасположения точек, которые справедливы для трехмерного пространства и плоскости будут справедливы и для n-мерного пространства, поэтому далее некоторые рассуждения лучше иллюстрировать на рисунках в плоскости.

Выше мы говорили о сходстве объектов. Можно ли это сходство как-то выразить количественно? Для этого служат многочисленные коэффициенты сходства или различия[20]. Для того чтобы пояснить, как ими пользоваться, рассмотрим конкретный пример, который мы предложили в начале этой главы. Итак, пусть мы описали обилие видов, населяющих дно в десяти водоемах, и свели все данные в таблицу.

Таблица 28. Плотность поселения (экз./м^2)донных животных в десяти водоемах

	Водоемы									
Виды	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1. Водяной ослик	300	220	2	0	0	12	448	114	228	1
2. Катушка роговая	2	46	1	1	0	173	85	31	34	5
3. Катушка килевая	2	26	2	1	1	31	16	24	26	0
4. Прудовик обыкновенный	1	10	0	0	0	0	0	0	0	0
5. Прудовик ушковый	22	12	0	0	0	10	11	34	88	1
6. Олигохеты	1000	123	1	10	15	188	180	200	568	0
7. Поденки Эфемеры	0	0	100	221	45	0	0	0	0	1451
8. Большая ложноконская пиявка	0	0	0	0	0	0	0	1	0	0
9. Улитковая пиявка	0	0	0	0	0	10	12	15	1	0

В приведенном большом количестве чисел разобраться очень трудно (а ведь это пример учебный, в реальной работе количество видов может измеряться сотнями, а количество сравниваемых объектов – многими десятками), это многообразие необходимо упорядочить. Первый шаг на пути упорядочения – это вычисление коэффициентов сходства или различия.

Сразу надо оговориться, что принципиальной разницы между коэффициентами сходства или различия нет. Высокое сходство – есть низкое различие и наоборот. Поэтому выбор того или иного типа коэффициента – это дело исследователя. Однако чаще применяют коэффициенты, отражающие различия объектов.

Самый простой коэффициент различия между объектами знаком любому человеку, владеющему теоремой Пифагора. Этот коэффициент называется эвклидово расстояние (пояснения см. на рис. 9). Для вывода этого коэффициента предположим, что у двух изученных объектов мы изучили только два признака: X и Y (стало быть, мы будем рассматривать двумерный, плоский рисунок).

Как следует из теоремы Пифагора, расстояние между этими двумя точками будет вычисляться по следующей формуле:

$$R = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} .$$

Это и есть эвклидово расстояние. Поскольку законы плоскости и пространства эквивалентны, то добавление еще одного признака (Z) принципиально ничего не изменит. Формула будет иметь такой вид:

$$R = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} .$$

Если же теперь рассмотреть общий случай, когда изучено n признаков, то формула приобретет следующий вид:

$$R = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} ,$$

[20] Таких коэффициентов огромное множество, можно, не сильно утруждаясь, придумать и свои собственные. Однако суть всех этих показателей одна – они, учитывая все признаки сравниваемых объектов, дают количественную меру сходства или различия между ними.

где A_i и B_i – значения i-тых признаков объектов А и В. Чем меньше расстояние между объектами, тем выше между ними сходство. С этой величиной можно работать точно так же, как с обычным расстоянием в обычном геометрическом пространстве. Только надо помнить, что в нашем гиперпространстве признаков расстояние между точками измеряется не в метрах, а в каких-то абстрактных единицах.

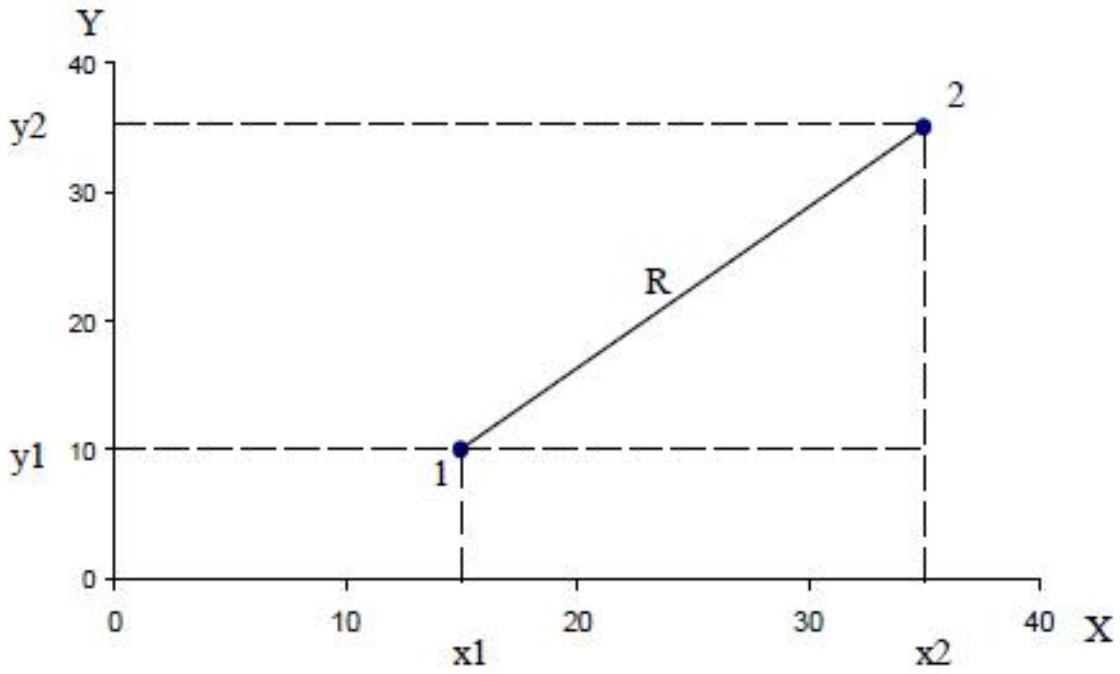

Рисунок 9. Эвклидово расстояние (R) между двумя точками.

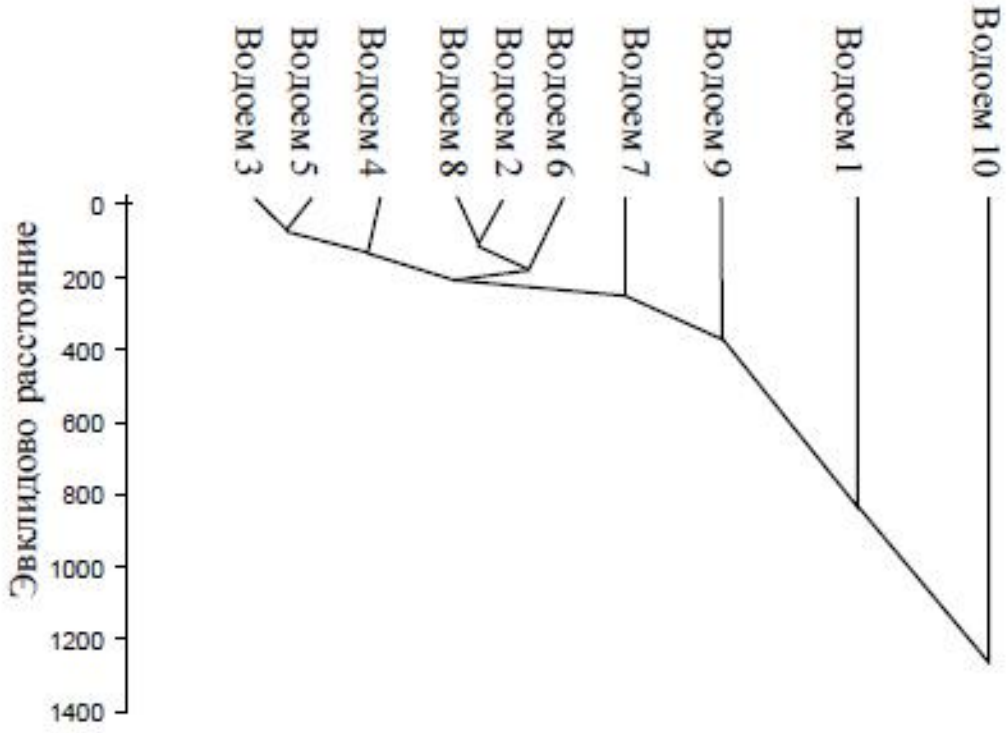

Рисунок 10. Дендрограмма, отражающая степень сходства между водоемами.

Теперь вернемся к рассмотрению нашего примера с водоемами. Понятно, что объектами будут сами водоемы, а признаками - обилие видов в них. Для иллюстрации хода вычислений рассмотрим сравнение водоема №1 и водоема №2.

Таблица 29. Ход вычисления Эвклидова расстояния между Водоемом №1 и Водоемом №2.

№ признака	Водоем 1 (A)	Водоем 2 (B)	A_i-B_i	$(A_i-B_i)^2$
1	300	220	80	6400
2	2	46	-44	1936
3	2	26	-24	576
4	1	10	-9	81
5	22	12	10	100
6	1000	123	877	769129
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
$\sum_{i=1}^{n}(A_i-B_i)^2$				778222
$R=\sqrt{\sum_{i=1}^{n}(A_i-B_i)^2}$				882

Итак, эвклидово расстояние получено. Большое это расстояние или маленькое? Ответить на этот вопрос можно только тогда, когда мы сравним полученное значение расстояния с расстояниями между другими парами объектов. Самый правильный способ для этих целей – это составление так называемой *матрицы расстояний*. Ниже приводится такая матрица, вычисленная для нашего примера.

Таблица 30. Матрица эвклидовых расстояний между изученными водоемами.

	Водоем 1	Водоем 2	Водоем 3	Водоем 4	Водоем 5	Водоем 6	Водоем 7	Водоем 8	Водоем 9	Водоем 10
Водоем 1	0	882	1048	1058	1031	879	838	822	445	1788
Водоем 2	882	0	274	336	255	253	239	135	452	1473
Водоем 3	1048	274	0	121	57	275	498	255	626	1351
Водоем 4	1058	336	121	0	176	334	535	317	649	1230
Водоем 5	1031	255	57	176	0	251	488	228	608	1406
Водоем 6	879	253	275	334	251	0	445	177	465	1473
Водоем 7	838	239	498	535	488	445	0	340	456	1531
Водоем 8	822	135	255	317	228	177	340	0	389	1470
Водоем 9	445	452	626	649	608	465	456	389	0	1578
Водоем 10	1788	1473	1351	1230	1406	1473	1531	1470	1578	0

Нетрудно заметить, что эта матрица отражает расстояния во всех возможных парах сравнения. Кроме того, видно, что она симметрична относительно диагонали, где стоят нули. Последняя особенность позволяет использовать только одну ее половину (например, верхнюю). Если отбросить лишнее, то матрица принимает следующий вид.

Таблица 31. Матрица Эвклидовых расстояний между изученными водоемами, после удаления нулей и дублирующих значений расстояния.

	Водоем 1	**Водоем 2**	**Водоем 3**	**Водоем 4**	**Водоем 5**	**Водоем 6**	**Водоем 7**	**Водоем 8**	**Водоем 9**	**Водоем 10**
Водоем 1		882	1048	1058	1031	879	838	822	445	1788
Водоем 2			274	336	255	253	239	135	452	1473
Водоем 3				121	***57***	275	498	255	626	1351
Водоем 4					176	334	535	317	649	1230
Водоем 5						251	488	228	608	1406
Водоем 6							445	177	465	1473
Водоем 7								340	456	1531
Водоем 8									389	1470
Водоем 9										1578
Водоем 10										

После того как все расстояния вычислены, можно приступать к выяснению того, насколько сходны или различны разные водоемы. Так, например, видно, что наиболее сходными будут водоемы №3 и №5, так как между ними наблюдается минимальное расстояние (R=57). Наименее похож на все остальные водоем № 10. Однако, если мы будем рассматривать результаты таким образом, то нам не хватит никакого места. Для сжатого представления информации об уровне сходства между объектами необходимо провести так называемый *кластерный анализ*.

Слово «*кластер*» происходит от английского «cluster» – группа, пучок, гроздь и т.п. Задача этого анализа – выявление групп сходных объектов и определение степени различия между этими группами. Существует много методов кластерного анализа, однако их процедуры достаточно трудоемки и обычно требуют вычислительной техники. Однако с принципами кластеризации можно познакомиться на самом простом методе, который называется «*методом ближайшего соседа*»[21].

Для кластеризации по методу ближайшего соседа необходимо найти минимальное расстояние. В нашем примере это расстояние между водоемами №3 и №5. Далее заполняем следующую таблицу.

Таблица 32. Ход кластерного анализа.

Группируемые объекты	R
3-5	57

После этого находим следующее по величине значение расстояния – это расстояние между водоемами №4 и №3 (R=121). Однако водоем №3 уже объединен в одну группу с водоемом №5, поэтому на следующем шаге анализа мы записываем уже более длинную цепочку строящегося кластера:

[21] К сожалению, этот метод дает наименее четкие результаты, поэтому большинство исследователей предпочитают им не пользоваться. Однако для объяснения принципа работы кластерного анализа он вполне годится. Ели вы решите профессионально заниматься кластерным анализом собственного материала, то рекомендую использовать процедуру, реализованную в пакете Statistica for Windows. Опыт многих исследователей показал, что из всех методов кластерного анализа наиболее четкие результаты дает не метод ближайшего соседа, а два других – метод парно-групповой взвешенной средней и метод Варда.

Таблица 32. (продолжение)

Группируемые объекты	R
3-5	57
3-5-4	121

Следующее по величине расстояние наблюдается между водоемами №8 и №2 (R=135). Эти два водоема еще не вошли ни в одну группу, поэтому они пока выделяются в отдельный кластер.

Таблица 32. (продолжение)

Группируемые объекты	R
3-5	57
3-5-4	121
8-2	135

Следующая пара – это водоемы № 5 и №4 (R=176). Однако эти два водоема уже попали в один кластер при меньшем расстоянии. Поэтому данное значение пропускаем и берем следующее. Это расстояние между водоемами № 8 и №6 (R=177). Соответственно цепочка удлиняется.

Таблица 32. (продолжение)

Группируемые объекты	R
3-5	57
3-5-4	121
8-2	135
8-2-6	177

Далее мы видим, что водоем №8 и №5 имеют следующее по порядку значение расстояния (R=228). Однако эти водоемы пока относились к разным кластерам. Значит, при R=228 эти два кластера могут быть объединены. Соответственно можно записать следующую цепочку.

Таблица 32. (продолжение)

Группируемые объекты	R
3-5	57
3-5-4	121
8-2	135
8-2-6	177
3-5-4-8-2-6	228

Далее продолжаем аналогичные операции и получаем следующие цепочки.

Таблица 32. (продолжение)

Группируемые объекты	R
3-5	57
3-5-4	121
8-2	135
8-2-6	177
3-5-4-8-2-6	228
3-5-4-8-2-6-7	239
3-5-4-8-2-6-7-9	389
3-5-4-8-2-6-7-9-1	822
3-5-4-8-2-6-7-9-1-10	1230

Когда все объекты попали в цепочку, процесс можно остановить. Теперь можно отразить результаты кластеризации с помощью специальной диаграммы, которая называется *дендрограммой*.

Приведенная на рисунке 10 дендрограмма отражает то, на каком уровне объединяются изученные объекты в группы. Однако, на построении дендрограммы кластерный анализ и заканчивается. Дендрограмма отражает объективный характер группировки объектов. Далее объекты надо разделить на дискретные кластеры. К сожалению, сделать это можно только достаточно субъективно.

Дендрограмма, полученная нами для данного примера, позволяет говорить о том, что отдельно от всех стоят водоемы №10 и № 1, они формируют два отдельных кластера. В одну группу явно объединяются водоемы № 3, 4 и 5, а в другую – водоемы № 8, 2 и 6. Можно даже найти некоторую закономерность в этих группах. Так, водоем №10 обладает наивысшей плотностью поселения поденок рода Эфемера. Водоемы № 8, 2 и 6 имеют большое количество олигохет, моллюсков, пиявок и водяных осликов. Иными словами, после группировки объектов можно искать причины их сходства. Однако это уже задача не математики, а биологии.

Вместе с тем, указанная субъективность в проведении границ между кластерами останавливает многих исследователей перед применением кластерного анализа в чистом виде. Он хорошо работает только в тех случаях, когда группы достаточно четко отличаются друг от друга. Если же все объекты формируют непрерывный ряд (как это наблюдается в нашем случае), то для этих целей более применим другой тип многомерного анализа, который называется *ординацией*, или *шкалированием*.

Суть методов шкалирования достаточно проста – необходимо расположить все объекты вдоль некоторой оси, которая отражала бы сходство между ними. Самый простой способ шкалирования – это упорядочение объектов в соответствии со значениями какого-то одного признака. Например, наши водоемы можно ранжировать согласно обилию олигохет. Однако, признаков у нас много, как быть в такой ситуации?

Для этого служат очень разные методы, самыми мощными из которых являются *методы многомерного шкалирования*[22] и разные формы *факторного анализа*. Однако суть этих методов достаточно сложна, поэтому поясним принципы ординации на самом простом методе, который получил название «*висконсинской полярной ординации*».

Давайте вновь обратимся к матрице сходства (таблица 31), приведенной выше. Можно заметить, что среди всех пар водоемов наименее похожими являются водоемы № 1 и № 10. Между ними максимальное расстояние. Значит, они наиболее далеко отстоят друг от друга в гиперпространстве признаков. Почему они так различаются? Заглянем в исходную таблицу показателей обилий видов (таблица 28). Можно заметить, что в водоеме № 1 наиболее обильны олигохеты и водяные ослики, но полностью отсутствуют поденки.

[22] В компьютерных программах этот метод называется Multi Dimension Scaling (MDS).

В водоеме № 10, напротив, нет олигохет, но многочисленны поденки. Столь высокая разница в населении, очевидно, определяется тем, что эти два водоема имеют какие-то принципиально разные условия. Значит, можно предположить, что эти два объекта находятся на двух противоположных полюсах некоторого градиента условий. Все остальные водоемы располагаются между этими полюсами. Как описать положение всех остальных водоемов в этом градиенте строго и объективно? Это опять чисто математическая задача, сводящаяся к умению использовать теорему косинусов.

Давайте, для примера, определим положение между означенными полюсами водоема № 2. Это можно сделать следующим образом. Представим, что в гиперпространстве признаков полюса (водоем №1 и водоем №10) соединены линией. Все остальные точки распределены где-то в пространстве вокруг этой линии. Расстояния от полюсов до каждой из этих точек нам известны (они приведены в матрице расстояний, таблица 31). Расстояние между полюсами составляет $R_{1\text{-}10}$=1788, расстояние от водоема №1 до водоема №2 равно $R_{1\text{-}2}$=882, а от водоема №10 до водоема №2 – $R_{10\text{-}2}$=1473. Поскольку мы работаем с тремя точками (водоемы №1, №2 и №10), мы можем провести через них плоскость и работать уже не в гиперпространстве признаков, а на плоскости (рисунок 11). Далее мы должны произвести проекцию точки, соответствующей водоему № 2, на линию, соединяющую полюса (водоем № 1 и № 10), и найти расстояние от любого из полюсов до точки проекции. Для этого надо провести очень нехитрые вычисления, суть которых поясняет рисунок 11.

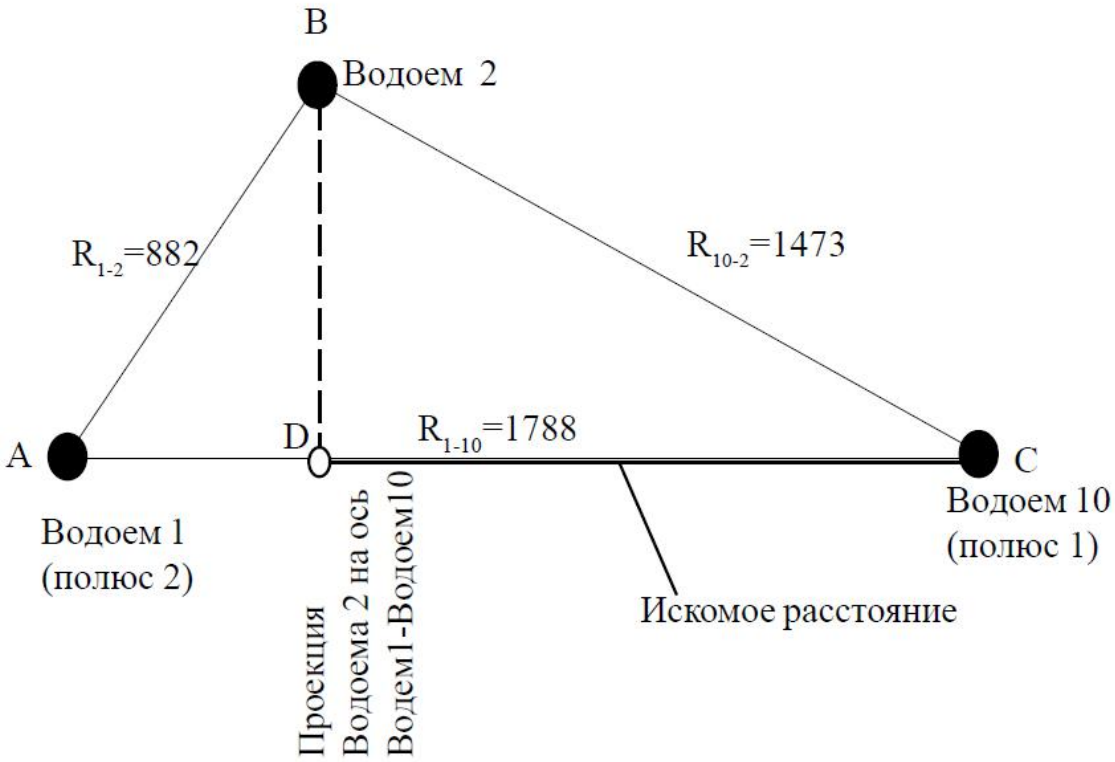

Рисунок 11. Расположение трех точек относительно линии, соединяющей наименее похожие водоемы.

Итак, вычислив все необходимые величины, мы можем сказать, что проекция этого водоема ближе расположена к водоему № 1, чем к водоему № 10. Это означает, что население водоема №2 больше похоже на население водоема №1, чем на население водоема №10 (если вы посмотрите на исходные данные, приведенные в таблице 28, то увидите, что это действительно так). Далее мы должны провести аналогичные вычисления для всех остальных объектов. В результате можно сказать, насколько близок каждый из

водоемов к тому или иному полюсу. В данном случае получились следующие результаты (см. таблицу 33).

Таблица 33. Расстояние от точек проекции водоемов до полюсов

	Расстояние от полюса до точки проекции
Водоем 10 (полюс 1)	0
Водоем 5	1004
Водоем 3	1097
Водоем 2	1150
Водоем 4	1283
Водоем 6	1285
Водоем 7	1309
Водоем 8	1353
Водоем 9	1535
Водоем 1 (полюс 2)	1788

Взаиморасположение полюсов и точек проекций иллюстрирует рисунок 12. Полученные данные позволяют увидеть, что к водоему № 10 ближе всего водоем № 5, 3, 2. Действительно, в этих водоемах водятся эфемеры и мало олигохет и водяных осликов. Значит, эти водоемы имеют сходные с водоемом № 10 условия. К водоему № 1, где много олигохет, близки водоемы № 9,8,7,6 и 4 где население в чем-то похоже, там много олигохет и водяных осликов и отсутствуют эфемеры.

Можно пойти дальше. Можно рассматривать полученные координаты как своего рода комплексные признаки, в которых в свернутом виде заключена информация обо всем населении. Далее можно проанализировать корреляцию этого признака с каким-то фактором. Например, со степенью загрязненности берегов. Пусть степень загрязненности будет следующая.

Таблица 34. Характеристика загрязненности берегов изученных водоемов

Водоем	**Расстояние от полюса до проекции**	**Степень загрязненности (количество пластиковых бутылок на 100 м береговой линии)**
Водоем 1 (полюс 2)	1788	50
Водоем 2	1150	20
Водоем 3	1097	1
Водоем 4	1283	1
Водоем 5	1004	1
Водоем 6	1285	30
Водоем 7	1309	1
Водоем 8	1353	44
Водоем 9	1535	30
Водоем 10 (полюс 1)	0	0

Если вспомнить разговор о корреляционном анализе, то становится ясно, что в данной ситуации наиболее применим ранговый коэффициент Спирмена. Не будем

останавливаться на ходе вычисления этого коэффициента, эта процедура описана в соответствующей главе. В итоге, коэффициент корреляции оказался равен r=0,834 (N=10). Этот коэффициент выше порогового. Значит, между степенью загрязнения и положением водоема в выявленном градиенте есть корреляция. Стало быть, можно сделать вывод о том, что население водоема находится в зависимости от степени антропогенного загрязнения.

Завершая разговор о многомерных методах, замечу, что если объектом вашего исследования являются биоценозы или вы занимаетесь вопросами систематики или какими-то другими исследованиями, требующими учета многих признаков, то освоение методов многомерного анализа для вас совершенно необходимо. Однако для более строгого анализа лучше использовать не описанные здесь простейшие (хотя иногда и весьма эффективные) методы, а более сложные, описание которых дается в специальной литературе, список которой приведен в части 4.

ЧАСТЬ 3. Статистические таблицы

Таблица I. Пороговые значения t-критерия Стьюдента

Число степеней свободы (ν)	**$t_{порог}$ при $P_{дов}$=95%**	**$t_{порог}$ при $P_{дов}$=99%**
1	12,71	63,66
2	4,3	9,93
3	3,18	5,84
4	2,78	4,6
5	2,57	4,03
6	2,45	3,71
7	2,37	3,5
8	2,31	3,36
9	2,26	3,25
10	2,23	3,17
11	2,2	3,11
12	2,18	3,06
13	2,16	3,01
14	2,15	2,98
15	2,13	2,95
16	2,12	2,92
17	2,11	2,9
18	2,1	2,88
19	2,09	2,86
20	2,09	2,85
21	2,08	2,83
22	2,07	2,82
23	2,07	2,81
24	2,06	2,8
25	2,06	2,79
26	2,06	2,78
27	2,05	2,77
28	2,05	2,76
29	2,05	2,76
30	2,04	2,75
более 30	1,96	2,58

Таблица II. Пороговые значения критерия χ^2

Число степеней свободы (ν)	**$\chi^2_{порог}$ при $P_{дов}$=95%**	**$\chi^2_{порог}$ при $P_{дов}$=99%**
1	3,84	6,63
2	5,99	9,21
3	7,81	11,34
4	9,49	13,28
5	11,07	15,09
6	12,59	16,81
7	14,07	18,48
8	15,51	20,09
9	16,92	21,67
10	18,31	23,21
11	19,68	24,72
12	21,03	26,22
13	22,36	27,69
14	23,68	29,14
15	25,00	30,58
16	26,30	32,00
17	27,59	33,41
18	28,87	34,81
19	30,14	36,19
20	31,41	37,57
21	32,67	38,93
22	33,92	40,29
23	35,17	41,64
24	36,42	42,98
25	37,65	44,31
26	38,89	45,64
27	40,11	46,96
28	41,34	48,28
29	42,56	49,59
30	43,77	50,89
40	55,76	63,69
50	67,5	76,15
60	79,08	88,38
70	90,53	100,42
80	101,88	112,33
90	113,14	124,12
100	124,34	135,81

Таблица III. Пороговые значения коэффициентов корреляции ($r_{порог}$)

Число степеней свободы (ν)	**$r_{s\text{-порог}}$ при $P_{дов}$=95%**	**$r_{s\text{-порог}}$ при $P_{дов}$=99%**	**Число степеней свободы (ν)**	**$r_{s\text{-порог}}$ при $P_{дов}$=95%**	**$r_{s\text{-порог}}$ при $P_{дов}$=99%**
1	0,997	1,000	24	0,388	0,496
2	0,950	0,990	25	0,381	0,487
3	0,878	0,959	26	0,374	0,478
4	0,811	0,917	27	0,367	0,470
5	0,754	0,874	28	0,361	0,463
6	0,707	0,834	29	0,355	0,456
7	0,666	0,798	30	0,349	0,449
8	0,632	0,765	35	0,325	0,418
9	0,602	0,735	40	0,304	0,393
10	0,576	0,708	45	0,288	0,372
11	0,553	0,684	50	0,273	0,354
12	0,532	0,661	60	0,250	0,325
13	0,514	0,641	70	0,232	0,302
14	0,497	0,623	80	0,217	0,283
15	0,482	0,606	90	0,205	0,267
16	0,468	0,590	100	0,195	0,254
17	0,456	0,575	125	0,174	0,228
18	0,444	0,561	150	0,159	0,208
19	0,433	0,549	200	0,138	0,181
20	0,423	0,537	300	0,113	0,148
21	0,413	0,526	400	0,098	0,128
22	0,404	0,515	500	0,088	0,115
23	0,396	0,505	1000	0,062	0,081

Таблица IV. Пороговые значения коэффициентов корреляции Спирмена ($r_{s\text{-порог}}$)

Число степеней свободы (ν)	**$r_{s\text{-порог}}$ при $P_{дов}$=95%**	**$r_{s\text{-порог}}$ при $P_{дов}$=99%**	**Число степеней свободы (ν)**	**$r_{s\text{-порог}}$ при$P_{дов}$=95%**	**$r_{s\text{-порог}}$ при $P_{дов}$=99%**
1			24	0,389	0,502
2			25	0,382	0,493
3	0,941		26	0,375	0,484
4	0,848		27	0,368	0,476
5	0,779	0,932	28	0,362	0,468
6	0,724	0,879	29	0,356	0,460
7	0,679	0,833	30	0,350	0,453
8	0,642	0,794	35	0,325	0,422
9	0,610	0,760	40	0,305	0,396
10	0,582	0,729	45	0,288	0,375
11	0,558	0,702	50	0,274	0,356
12	0,537	0,678	60	0,250	0,327
13	0,518	0,656	70	0,232	0,303
14	0,501	0,636	80	0,217	0,284
15	0,485	0,617	90	0,205	0,268
16	0,471	0,600	100	0,195	0,255
17	0,458	0,585	125	0,174	0,229
18	0,446	0,570	150	0,159	0,209
19	0,435	0,557	200	0,138	0,181
20	0,424	0,545	300	0,113	0,148
21	0,415	0,533	400	0,098	0,129
22	0,406	0,522	500	0,088	0,115
23	0,397	0,511	1000	0,062	0,081

ЧАСТЬ 4. Некоторые книги, которые крайне рекомендуется освоить для более глубокого освоения математических методов исследования

В этой главе я решил привести наиболее распространенные и наиболее простые для понимания (к сожалению, это не всегда совпадающие характеристики) издания, сопроводив их краткими замечаниями.

- Урбах В. Ю. Биометрические методы. Статистическая обработка опытных данных в биологии, сельском хозяйстве и медицине. – Издательство «Наука» – Москва, 1964

Самое лучшее русскоязычное пособие по статистике для биологов. В книге приведены и объяснены выводы формул. Все написано простым и понятным языком. Однако обилие формул и строгих выводов может отпугнуть начинающего исследователя. Вместе с тем, эту книгу крайне рекомендуется прочитать.

- Ивантер Э. В., Коросов А. В. Основы биометрии. Введение в статистический анализ биологических явлений и процессов. – Издательство ПГУ.- Петрозаводск, 1992.

Очень хорошее пособие, написанное просто и кратко. Опыт показал, что его школьники осваивают без особых усилий. В книге даются в доступной форме алгоритмы использования тех или иных методов. Однако это пособие, будучи изданным в Петрозаводске, практически неизвестно в других городах.

- Лакин Г. Ф. Биометрия.- Издательство «Высшая школа».- Москва, 1990.
- Рокицкий П. Ф. Биологическая статистика. – Издательство «Вышэйшая школа».-Минск, 1967.

Эти два учебника для студентов ВУЗов содержат все необходимые сведения по курсу биометрии. Они хороши именно как учебники для неспециалистов, для тех, кто не хочет глубоко вникать в тонкости работы тех или иных методов, но стремится к их правильному использованию.

- Коросов А. В. Экологические приложения компонентного анализа. – Издательство ПГУ. – Петрозаводск, 1996.

В книге дается самое простое, из мне известных, описание одного из самых мощных методов многомерного анализа – *метода главных компонент*. Пособие написано очень простым и понятным языком, изобилует примерами.

- Терентьев П. В., Ростова Н. С. Практикум по биометрии. – Издательство Ленинградского гос. университета.- Ленинград, 1977.

Очень полный сборник описаний методов математического анализа биологических данных. Особенно хорошо изложен корреляционный анализ. Разбираются многочисленные примеры.

- Help для пакета Statistica for Windows.

Очень хороший справочник. Его наличие, делает продукт компании Statsoft, пожалуй, самым удобным орудием компьютерного статистического анализа данных. В тексте хелпа даются и ссылки на оригинальную литературу, в которой дается описание соответствующих методов. Единственный недостаток – это то, что все написано на английском языке. Однако компания Statsoft выпустила электронный учебник по статистике на русском языке. На мой взгляд, данный учебник ни в коей мере не заменяет описанных выше пособий – слишком уж он краток, написан «Интернет»-языком (что, впрочем, многим может и понравиться). Поскольку продукт, наверняка, защищен всякими авторскими правами, то по вопросам его приобретения обращайтесь к разработчикам.