МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВПО «КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ
АГРАРНЫЙ УНИВЕРСИТЕТ»

Факультет экологии
Кафедра прикладной экологии

ЭКСПЕРИМЕНТАЛЬНАЯ ЭКОЛОГИЯ

Курс лекций

по направлению подготовки аспирантов 05.06.01 – Науки о Земле

Краснодар
КубГАУ
2015

Составители: Горковенко Н.Е.

Экспериментальная экология: курс лекций / сост. Н.Е. Горковенко – Краснодар : КубГАУ, 2015. – 53 с.

Курс лекций предназначен для аспирантов по направлению подготовки 05.06.01 – Науки о Земле

Рассмотрено и одобрено методической комиссией факультета экологии 29.06.2015г., протокол № 10.

Лекция № 1. Биометрия как основа интерпретации результатов эксперимента

Предметом биометрии служит любой биологический объект, изучаемый с применением счета или меры, т. е. с количественной стороны в целях более или менее точной оценки его качественного состояния. При этом имеются в виду не единичные, а групповые объекты, т. е. явления массовые, в сфере которых проявляют свое действие статистические законы. Например, врач принял больного и назначил необходимое ему лекарство – это единичное явление, отдельный акт. Если же врач принял несколько больных или подверг неоднократному осмотру одного и того же больного, – это массовое явление независимо от того, каким был объект наблюдения – единичным или групповым.

Обычно наблюдения проводят на групповых объектах, например на особях одного и того же вида, пола и возраста, которые рассматривают как составные элементы, или члены группового объекта, и называют единицами наблюдения. Множество относительно однородных, но индивидуально различимых единиц, объединенных для совместного (группового) изучения, называют статистической совокупностью.

Понятие статистической совокупности – одно из фундаментальных биометрических понятий. Оно базируется на принципе качественной однородности ее состава. Нельзя объединять в одну совокупность особей разного пола и возраста, когда речь идет о нормах питания, стандартизации обуви и одежды, поскольку заведомо известно, что с возрастом и в зависимости от пола индивидов меняются их потребности в питании и закономерно изменяются размеры и пропорции тела. Недопустимо изучать закономерность модификационной изменчивости на генетически неоднородном материале, объединяя в одну совокупность чистопородных и гибридных особей и т. д.

Наряду с понятием статистической совокупности существует понятие статистического комплекса. Так, если статистическая совокупность состоит из относительно однородных единиц, то статистический комплекс слагается из разнородных групп, объединяемых для совместного (комплексного) изучения. При этом каждая группа, входящая в состав комплекса, должна состоять из однородных элементов. Например, в массе подопытных животных наряду с контролем может быть образовано несколько групп, отличающихся друг от друга по возрасту, породной или видовой принадлежности и т. п., на которых испытывают действие изучаемого агента. При испытании различных доз удобрений каждую опытную делянку рассматривают как отдельную группу, входящую в состав статистического комплекса.

Вопрос о форме объединения биометрических данных экспериментатор решает сам в зависимости от объекта и цели исследования. Объединяемые в статистическую совокупность или статистический комплекс результаты наблюдений представляют некую систему, не сводимую к сумме составляющих ее единиц или компонентов. В статистических совокупностях и в статистических комплексах существует внутренняя связь между частью и целым, единичным и общим, которая находит свое выражение в статистических закономерностях, обнаруживаемых в сфере массовых явлений. Эти закономерности являются той теоретической платформой, на которой базируется биометрия.

Выборка.

Биометрическое исследование в центр внимания всегда ставит *выборку – множество значений случайной величины, совокупность вариант, набор чисел*; отдельная варианта – это объект, несущий качественный или числовой признак. Термин «выборка» указывает на процесс выбора части из чего-то большего, в данном случае – на процесс получения ограниченного количества значений из генеральной совокупности. *Генеральная совокупность – это множество всех вариант определенного типа* (выборка бесконечного размера). Чаще всего получить все возможные значения в принципе невозможно.

Поэтому судить о генеральной совокупности приходится, исследуя выборки, – по части составлять представление о целом.

Признак

Варианта качественно или количественно выражает признак данного объекта исследования (полученного при данном уровне фактора внешней среды вполне определенным методом). Признак (свойство, показатель, величина, характеристика, переменная) – любая информация о наблюдаемом объекте, выраженная качественно или определенная количественно.

В общем смысле под словом «признак» подразумевают свойство, проявлением которого один предмет отличается от другого. В области биологии признаками, по которым проводят наблюдения над объектами, служат такие характерные особенности в строении и функциях живого, которые позволяют отличать одну единицу наблюдения от другой, сравнивать их между собой. Например, исследователя интересует содержание зерен в колосьях пшеницы или ржи, возделываемой на специально подготовленном участке. Массив данной культуры и будет объектом наблюдения, а признаком – количество зерен в колосьях отдельных растений, которые являются единицами наблюдения, составляя в общей массе, подвергаемой изучению, статистическую совокупность.

Характерным свойством биологических признаков является варьирование величины признаков в определенных пределах при переходе от одной единицы наблюдения к другой. Например, подсчитывая наличие зерен или колосков в колосьях, взвешивая детенышей животных одного и того же помета, определяя жирность молока у животных однородной группы и в других подобных случаях, нетрудно заметить, что величина каждого признака колеблется, образуя совокупность числовых значений признака, по которому проводят наблюдение. Эти колебания величины одного и того же признака, наблюдаемые в массе однородных членов статистической совокупности, называют вариациями (от лат. *variatio* – изменение колебания), а отдельные числовые значения варьирующего признака принятого называть вариантами (от лат. *varians, variantis* – различимый, изменяющийся).

Классификация признаков. Все биологические признаки варьируют, но все они поддаются непосредственному измерению. Отсюда возникает деление признаков на *качественные*, или *атрибутивные*, и *количественные*.

Качественные признаки не поддаются непосредственному измерению и учитываются по наличию их свойств у отдельных членов изучаемой группы. Например, среди растений можно подсчитать количество экземпляров с разной окраской цветков – белой, розовой, красной, фиолетовой и т. д. В массе животных также нетрудно отличить и учесть особей разного пола и масти – серых, вороных, гнедых, пестрых и др.

Количественные признаки поддаются непосредственному измерению или счету. Их делят на *мерные*, или *метрические*, и *счетные*, или *меристические*. Длина колосьев, урожайность той или иной культуры, мясная и молочная продуктивность животных – все это мерные признаки, варьирующие непрерывно: их величина может принимать в определенных пределах (от-до) любые числовые значения. Счетные признаки, такие, например, как число зерен или колосков в колосьях, яйценоскость и другие подобные признаки, варьируют прерывисто или дискретно: их числовые значения выражаются только целыми числами.

Если результаты наблюдений группируются в противопоставляемые друг другу группы, их варьирование в отличие от рядовой изменчивости называют альтернативным и признаки, по которым проводят наблюдение, – альтернативными. Примером могут служить случаи, когда противопоставляют особи женские мужским, больные – здоровым, высокорослые - низкорослым, успевающие – неуспевающим и т. д.

Деление признаков на качественные и количественные весьма условно. Например, в массе однородных индивидов, доступных измерению, можно выделить группы высоких, средних и низких, а также успевающих и неуспе-

вающих и т. д. Вместе с тем в каждом качественном признаке, например в окраске листьев, цветков и плодов, можно обнаружить целую гамму количественных переходов, или градаций, и измерить их. И все же, несмотря на очевидную условность приведенной классификации, она необходима хотя бы потому, что количественные признаки распределяются в вариационный ряд, а качественные не распределяются. А при разных способах группировки исходных данных применяют разные способы их обработки.

На языке математики величина любого варьирующего признака является переменной случайной величиной. В отличие от постоянных величин, обозначаемых начальными буквами латинского алфавита, переменные величины принято обозначать последними в латинском алфавите прописными буквами X, У, *Z*, а их числовые значения, т. е. варианты, – соответствующими строчными буквами того же алфавита: x_1, x_2, x_3, ... , x_n или y_1, y_2, y_3, ... , y_n и т. д. Общее обозначение любой варианты отмечают символом x_i, y_i и т. д., где индекс *i* символизирует общий характер варианты (даты).

В рамках вариационной статистики признаки выступают в роли случайной величины.

Случайная величина – численная характеристика, принимающая те или иные заранее точно не известные значения. Несмотря на то, что точное описание поведения случайной величины получить нельзя, математическая статистика позволяет выполнить вероятностное описание.

Существует целый ряд методов регистрации признаков биологических объектов.

Качество (нечисловой дискретный признак) – простой, непосредственный, чувственный способ регистрации фактов; это статус, сезон, таксон, цвет, плотность, тип действия и пр. Значения таких признаков выражаются словами или символами, они не имеют количественного содержания и выражают принадлежность данного объекта к определенной обширной группе объектов (зеленый, январь, ♀, ♪).

Балл (оценка) – дискретный полуколичественный признак, численная характеристика объекта, присвоенная в соответствии с внешней заранее принятой шкалой баллов. Во время оценки объект соотносится с этими критериями и ему присваивается соответствующий балл. Баллы не обладают свойствами чисел, в частности, балл 4 не в два раза больше балла 2, для них арифметические операции применять нельзя. Для баллов многие выборочные параметры (средние, дисперсии и др.) не будут обладать свойствами статистических параметров, их нельзя статистически сравнивать, например, с помощью критерия Стьюдента. Корректно будет характеризовать выборки балльных оценок лишь с помощью частотных распределений, моды, размаха

изменчивости. Для статистической обработки балльных оценок требуются *непараметрические* методы.

Количество (число) – дискретный (счетный) количественный признак (число натурального ряда), характеризующий множество однородных объектов, черт, деталей строения, состав (например, число эмбрионов у самки, число жаберных тычинок у рыб, число тычинок в цветке, число деревьев на пробной площадке). Отдельную варианту получают, подсчитав число неких дискретных черт строения у отдельного объекта или в пробе. *Проба – ограниченная совокупность разнокачественных объектов, которая характеризуется числом объектов одного определенного качества, это значение играет роль одной варианты выборки*. Получая серию проб, мы осуществляем перевод качественных признаков в количественные.

Промер (ряд дробных или рациональных чисел) – непрерывный (мерный) количественный признак, характеризующий свойства объектов с помощью различных относительных количественных шкал – температурной, весовой, размерной, объемной и т. п. Отдельная варианта получает количественную характеристику выраженности данного признака у данного объекта (в пределах точности метода): температуру тела, его размеры, уровень глюкозы в крови и т. д. Большинство методов статистики разработано для исследования именно таких непрерывных признаков (параметрические методы).

Варьирование.

Основная особенность выборки как множества значений случайной величины – это отличие отдельных вариант друг от друга, явление *изменчивости*, варьирования, появления отличий между отдельными вариантами.

Биологу важно знать обычные причины варьирования. Один из источников, эндогенный, – это индивидуальные отличия по *статусу* и по *состоянию*. Например, животные одного возраста различны индивидуально, генетически, т. е. по статусу. Кроме того, каждое из них в разные годы, сезоны, время суток имеет разные морфофизиологические характеристики, т. е. отличается по состоянию. В наиболее точных науках (токсикология, биохимия, молекулярная биология) стремятся с помощью химической чистоты постановки опытов и выведения чистых линий подопытных животных убрать все мешающие причины «избыточного» варьирования.

Другой источник отличий между вариантами – факторы внешней среды, т. е. условия проведения наблюдений, среда существования объекта, возможная причина, определяющая текущее состояние объекта. Часто говорят про факторы эндогенные, внутренние (статус, способ существования объекта), и экзогенные, внешние (среда, условия существования объекта). Фактор всегда есть активное, действующее начало, признак – его результат, послед-

ствие. Факторы, влияющие на значения вариант, различаются по своей природе. Если фактор влияет на все варианты выборки постоянно и примерно одинаково, он называется систематическим (или доминирующим). Если фактор непостоянен, влияет на варианты не одинаково, с разной силой, он определяется как случайный. Эти рассуждения дают *модели варианты*:

$x_i = x_{дом.} \pm x_{случ.}$,

где x_i – измеренное значение варианты,

i – индекс варианты ($i = 1, 2, \ldots, n$),

n – объем (общее количество вариант) выборки,

$x_{дом.}$ – суммарный вклад j доминирующих факторов,

$x_{случ.}$– суммарный вклад k случайных факторов.

С методической точки зрения при наблюдениях или в эксперименте самым важным оказывается обязательная *регистрация* максимально возможного числа факторов (как внешних, так и внутренних). Тогда появляется возможность исследовать их раздельное действие на объект, поскольку существуют методы, которые позволяют из многокомпонентной среды вычленять эффекты действия отдельных факторов (особенно работоспособны дисперсионный, регрессионный и компонентный анализы).

При самом широком варьировании признаков разброс значений выборки не бесконечно широк, он ограничен неким диапазоном и тяготеет к определенному общему значению. Эти свойства статистических совокупностей – варьирование, но в ограниченном диапазоне, – позволяют предложить для описания две группы величин: оценку центрального значения диапазона (среднюю, моду или медиану) и оценку размаха варьирования (лимит, дисперсию, стандартное отклонение). Определение этих значений выполняется после построения вариационного ряда.

Лекция № 2. Планирование эксперимента.

Классические работы Р. Фишера открыли новую страницу в истории биометрии: они показали, что планирование экспериментов и обработка их результатов – это две тесно связанные между собой задачи статистического анализа. Это открытие легло в основу разработки теории планирования экспериментов, которая в настоящее время находит применение не только при проведении сельскохозяйственных опытов, на базе которых она возникла, но и в различных областях биологии, медицины, антропологии, в сфере других научно-практических дисциплин, включая и социально-экономические исследования.

Планирование экспериментов стало самостоятельным разделом биометрии, которому посвящено огромное число работ. Нами будут рассмотре-

ны лишь некоторые общие положения, относящиеся к этой сложной и многогранной проблеме.

Термин «эксперимент» (от лат. *experimentum* – опыт) означает искусственно организуемый комплекс условий, в которых испытывают воздействие того или иного фактора или одновременно нескольких факторов на результативный признак. В земледелии это полевые опыты; в животноводстве – опыты по кормлению животных, по уходу за ними; в педагогике – опыты по проверке новых методов обучения и воспитания учащихся; в фармакологии – испытание эффективности новых лечебных препаратов; в медицине – проверка разных способов лечения больных и т. д.

Экспериментальный подход

Эксперимент включает 5 последовательных стадий: гипотеза, планирование, реализация, статистический анализ и интерпретация. Гипотеза обладает первоочередной важностью, поскольку если она не удовлетворяет некоторым критериям качества, то даже самый правильно проведенный эксперимент будет иметь не слишком большую ценность.

Под планированием эксперимента понимается лишь «логическая структура исследования» (Fisher, 1971, p. 2). Полное описание целей эксперимента должно включать спецификацию природы используемых экспериментальных единиц, число и характер применяемых воздействий (включая "контрольные" воздействия), а также свойства или отклики (параметры экспериментальных единиц), которые предполагается измерять.

Когда решение по этим вопросам принято, план эксперимента определяет схему, согласно которой для каждой доступной экспериментальной единицы назначается уровень воздействия. При этом определяется число экспериментальных единиц (повторностей), получающих воздействие каждого уровня, устанавливается физическое расположение экспериментальных единиц, а также частота или временная периодичность, с которой реализуются воздействия и осуществляются измерения контролируемых факторов на различных экспериментальных единицах.

Реализация эксперимента включает весь комплекс процедур и операций, в отношении которых осуществлялось планирование. Успешное осуществление в равной мере зависит от искусства экспериментатора, его проницательности и рассудительности, а также от его технических навыков. Непосредственной задачей исследователя обычно является выполнение технических операций эксперимента таким образом, чтобы избежать систематических ошибок (отклонений) и минимизировать случайные ошибки. Если изучается влияние ДДТ, то препарат не должен содержать примесей иных веществ. Если изучается влияние хищника, охотящегося в приливной зоне, то

расположение клеток, блокирующих хищника, не должно иметь *прямого* влияния на поведение экосистемы, за исключением самого хищника. Если изучается влияние питательных веществ на биомассу планктона в пруду, то отбор проб должен выполняться посредством устройства, производительность которого не зависит от обилия планктона.

Систематические ошибки, допущенные как в распределении воздействий, так и в процедурах измерения или отбора проб, делают эксперимент некорректным, а его выводы неубедительными.

Субъективным образом также решается вопрос о том, какова допустимая или желательная изначальная гетерогенность между экспериментальными единицами и в какой степени следует регулировать условия среды в ходе эксперимента. Эти обстоятельства влияют на величину случайных ошибок и потому – на оценку чувствительности изучаемых объектов по отношению к воздействию. Они также влияют на конкретную интерпретацию результатов, хотя сами по себе цели исследования не определяют.

Из изложенного ясно, что планирование эксперимента и особенности его реализации в равной степени определяют обоснованность исследования и его итоги. Хотя в практическом смысле реализация – это более критичный аспект эксперимента, нежели его планирование. Действительно, ошибки при осуществлении эксперимента обычно возникают в большем числе этапов исследования, более многообразны и часто более коварны, чем ошибки при планировании. Следовательно, погрешности реализации обнаружить обычно сложнее, чем просчеты в планировании.

В экспериментальной работе основная функция статистики – увеличить четкость, выразительность и объективность, с которыми результаты представляются и интерпретируются. Статистический анализ и интерпретация – наименее критичные аспекты экспериментирования в том смысле, что если допускаются чисто статистические или интерпретационные ошибки, то данные могут быть проанализированы заново. В то время как единственным абсолютным средством исправления ошибок планирования или реализации является только повторение эксперимента.

Можно выделить два класса экспериментов: измерительные {mensurative} и манипулятивные {manipulative}. Измерительные эксперименты включают только проведение наблюдений в одной или нескольких точках пространства или времени; пространство или время – это единственные "экспериментальные" переменные или "факторы воздействия". Оценка значимости воздействия по статистическим критериям осуществляется здесь не всегда. Измерительные эксперименты обычно не включают единицы. Если они включают такое наложение (например, сравнение откликов горных и

равнинных особей дуба на экспериментальную дефолиацию), то все экспериментальные единицы подвергаются одинаковому "воздействию".

Пример 1. Мы хотим определить, как быстро разлагаются листья клена (*Acer*) на дне озера на глубине 1 м. Для этого мы делаем 8 маленьких мешков из нейлоновой сетки, наполняем каждый из них кленовыми листьями и помещаем все вместе в какой-то точке 1-метровой изобаты. Через месяц мы вынимаем мешочки, определяем потерю разложившегося органического вещества в каждом и вычисляем среднюю скорость разложения. В таком виде эта процедура удовлетворительна. Однако она не дает информации о том, как скорость может варьировать в разных точках 1-метровой изобаты. Средняя скорость, которую мы вычислили по нашим 8 мешочкам с листьями – слишком скудное основание для обобщения величины "скорости разложения на 1-метровой изобате в озере".

Такая процедура обычно называется экспериментом *просто потому, что процедура измерения достаточно трудоемка*, и часто включает вмешательство в саму систему. Если бы мы провели 8 измерений температуры или отобрали 8 проб дночерпателем, мало кто назвал бы эти процедуры и их результаты "экспериментальными".

Термин "экспериментальное" всегда использовался в контексте значений "сложное", "трудоемкое", "подразумевающее вмешательство {interventionist}".

Пример 2. Предположим, что мы хотим, используя процедуру примера 1, выяснить, отличается ли скорость разложения кленовых листьев между 1-метровой и 10-метровой изобатами. Для этого мы помещаем 8 мешочков с листьями на 1-метровую изобату и другие 8 мешочков на 10-метровую, ждем месяц, извлекаем мешочки и получаем данные. Затем мы применяем статистический критерий (например, *t*-критерий или *U*−критерий), чтобы узнать, имеется ли достоверное различие скорости разложения в двух точках.

Этот опыт можно было бы назвать *сравнительным измерительным экспериментом*. Хотя нами использовались две изобаты (или два "уровня воздействия"), полноценная проверка научных гипотез, присущих манипулятивным экспериментам, проведена не была. Мы просто измерили свойство системы в двух точках внутри нее и оценили, существует ли реальное различие ("эффект воздействия") между ними.

Чтобы достигнуть не слишком четко сформулированную цель в примере 1, любой тип пространственного размещения 8 мешочков по изобате, в принципе, был бы приемлемым. В примере же 2 мы определили нашу цель как сравнение двух изобат в отношении скорости разложения кленовых листьев. Поэтому мы не можем расположить наши мешочки в одном месте на

каждой изобате. Это не даст нам никакой информации об изменчивости скорости разложения от точки к точке вдоль изобаты. Такую информацию необходимо получить, прежде чем обоснованно применять статистический критерий для проверки нулевой гипотезы о том, что скорость разложения одинакова на двух изобатах. Поэтому мы должны рассеять наши мешочки на каждой изобате некоторым подходящим образом. Существует много путей выбора такого размещения. В идеальном случае позиции вдоль каждой изобаты должны выбираться случайно, но мешочки могут быть расположены индивидуально (8 точек), либо группами по две (4 точки) или по четыре (2 точки). Более того, мы можем решить, что достаточно работать с изобатами только вдоль одной стороны озера и т.д.

Размещение повторных выборок или измерений в пространстве (или времени) подходящим образом, соответствующим конкретной проверяемой гипотезе, – наиболее критичный аспект планирования измерительных экспериментов.

Пример 3. Предположим, что поленившись, мы расположили все 8 мешочков в одном месте на каждой из изобат. В этой ситуации все еще будет корректным применить критерий значимости к полученным данным. Однако, если достоверные различия обнаружены, это является свидетельством различий только между двумя *точками*: "так случилось", что одна из точек лежит на 1-метровой изобате, а вторая – на 10-метровой. Выявленное достоверное различие между ними не может быть корректно интерпретировано как различие между двумя *изобатами*, т.е. как свидетельство "эффекта воздействия". Такое выявленное достоверное различие не более того различия, которое мы обнаружили бы, поместив два набора по 8 мешочков в двух точках на *одной и той же* изобате.

Если мы настаиваем на интерпретации проверки гипотезы в примере 3 как "эффекта воздействия" с констатацией реальных различий между изобатами, мы совершаем ошибку, так называемая *мнимая повторность*. В целом в измерительных экспериментах мнимые повторности часто являются следствием того, что реальное физическое пространство, из которого формируются выборки (либо в котором проводятся измерения), меньше, либо более ограничено, чем то, которое фигурирует в гипотезе. В манипулятивных экспериментах мнимые повторности проявляются в результате использования статистических методов для проверки гипотезы об эффекте воздействия по данным из экспериментов, в которых либо воздействия вообще не имели повторностей (хотя могло быть несколько выборок), либо эти повторности не были статистически независимы. Таким образом, мнимые повторности относятся не к проблеме планирования эксперимента (или выборочного процесса)

как такового, а скорее к определенной комбинации планирования эксперимента (или выборочного процесса) и статистического анализа, который неадекватен для проверки поставленных гипотез.

Исследовательская работа не только сводится к экспериментам; ее проводят и вне их на основе непосредственных наблюдений.

Так что выражение «планирование исследований» оказывается более емким, а следовательно, и более подходящим, чем введенный Р. Фишером (1930) термин «планирование экспериментов ». Конечно, и термин «эксперимент» можно применять в более широком смысле, понимая под ним любые испытания, проводимые исследователем в отношении изучаемого объекта. При всем разнообразии методов исследовательской работы задача планирования сводится к тому, чтобы при возможно минимальных объемах наблюдений получать достаточно полную информацию об изучаемых объектах.

С варьированием результатов наблюдений связана повторность вариантов опыта, позволяющая повысить точность оценок генеральных параметров, надежность выводов, которые делает исследователь на основании выборочных показателей. Под повторностью в полевом опыте понимают число одноименных делянок для каждого варианта опыта. В лабораторных условиях повторность может выражаться числом одинаковых проб серий одновременных испытаний, измерений и т. п. повторений одного и того же варианта опыта. Очевидно, чем шире диапазон варьирования признака, тем больше должна быть и повторность опыта, и, наоборот, при слабом варьировании учитываемого признака число вариантов опыта, т. е. их повторность, уменьшается. В такой же зависимости от размаха варьирования признаков находится и организация планирования минимально допустимого числа испытаний.

Приближенные оценки основных статистических показателей

Прежде чем наметить необходимый объем выборки, надо определить среднюю величину и ее ошибку для варьирующего признака – характеристики, которые позволяют использовать показатель точности выборочной средней при решении этой задачи.

Приближенное значение средней арифметической $\overline{x}$ можно определить по полусумме лимитов:

$$\overline{x} = \frac{x_{min} + x_{max}}{2},$$

а среднее квадратическое отклонение s_x – по разности лимитов, отнесенной к коэффициенту К, который устанавливают в зависимости от объема выборки *(n)* с помощью табл. (по Н. А. Плохинскому, 1970), т. е. по формуле:

$$s_x = \frac{x_{max} - x_{min}}{K},$$

Пример 1. Зная лимиты x_{mln}=*9,0* мг% и $x_{\max}$ = 14,7 мг% кальция в сыворотке крови обследованной группы обезьян *(n*== 100), можно определить основные характеристики для этой выборки:

$$\overline{x} = \frac{9{,}0 + 14{,}7}{2} = 11{,}85 \text{ мг\% и } s_x = \frac{14{,}7 - 9{,}0}{5} = 1{,}14$$

Эти величины близки к фактически найденным: $\overline{x}$= 11,94 мг% и s_x= *1,26.*

n	2–5	6–15	16–49	50–200	201–1000	> 1000
K	2	3	4	5	6	7

Величину ошибки средней *s,;* можно определить по следующей приближенной формуле:

$$s_{\overline{x}} = \frac{x_{max} - x_{min}}{K\sqrt{n}},$$

Так, в данном случае $s_{\overline{x}} = \frac{14{,}7-9{,}0}{5\sqrt{100}}$ =0 114. Эта же величина получается и при использовании основной формулы $s_{\overline{x}} = \frac{s_x}{\sqrt{n}} = \frac{1{,}14}{100}$ =0,114. Отсюда показатель точности Cs выборочной среднейх $\overline{x}$ $C_x = 100\frac{s_{\overline{x}}}{\overline{x}} = 100\frac{0{,}114}{11{,}85}$ =0,96. Это очень высокая точность. Намечаемый таким образом объем выборки можно считать вполне достаточным для получения надежных оценок генеральных параметров (при условии, что совокупность, из которой взята выборка, распределяется по нормальному закону).

Определение необходимоrо объема выборки.

Элементарная логика и практический опыт подсказывают, что неразумно стремиться к неоправданно большому числу испытаний, если убедительный результат можно получить при минимально допустимом объеме выборки. Необходимая численность выборки *n,* отвечающая точности, с какой намечено получить средний результат, зависит от величины ошибки выборочной средней и определяется по формуле

$$n = \frac{t^2 s_x^2}{\Delta^2} \qquad \text{или}$$

$$n = \frac{t^2}{\Delta^2 / s_x^2} = \left(\frac{t}{K}\right)^2$$

где *t* - нормированное отклонение, с которым связан тот или иной уровень значимости (α); s_x^2- выборочная дисперсия; Δ = $ts_{\overline{x}}$-*;* – величина, опреде-

ляющая границы доверительного интервала (здесь $s_{\overline{x}} = \sqrt{\frac{s_x^2}{n}}$ – ошибка выборочной средней); К = Δ / s_x.

Пример 2. Случайная выборка девяти вариант характеризуется средней $\overline{x}$=12,1+0,68. Точность выборочной средней оказалась недостаточно высокой: *Cs*= 100 $\frac{0,68}{12,1}$=5,62=6. Какое число испытаний *n* нужно провести, чтобы ошибку средней уменьшить вдвое? В данном случае $s_x = s_{\overline{x}} \sqrt{n}$ = 0,68 $\sqrt{9}$= *2,04.* Примем *t*= 1,96 ≈ 2, что соответствует 5%-уому уровню значимости. Предварительно определим Δ = 2 $\frac{0,68}{2}$ =0,68; К =$\frac{0,68}{2,04}$ =0,33. Подставляем найденные величины в формулу *n= (2/0,33)*^{2}*=6*^{2}*=36.*

Чтобы уменьшить ошибку репрезентативности вдвое, нужно объем выборки увеличить в четыре раза (9·4=36). Обобщая эти данные, можно сделать вывод: для уменьшения ошибки выборочной средней в К раз нужно увеличить объем выборки в К^2 раз.

При определении необходимого объема выборки для получения статистически достоверной разности между средними ($\overline{x}_1$-$\overline{x}_2$) =*d* применяют формулу:

$$n_2 = \left(\frac{t}{\Delta}\right)^2 \left(\frac{s_1^2}{a} + s_2^2\right),$$

Здесь Δ =*tS*$_d$, где s_d – заданная величина ошибки для разности сравниваемых средних; s_1^2 и s_2^2 – дисперсии для сравниваемых выборок, причем s_1^2 - дисперсия для большей выборки; а=n_1 / n_2 – отношение объема большей выборки к объему меньшей выборки. При n_1 =n_2 формула (принимает следующий вид:

$$n = \left(\frac{t}{\Delta}\right)^2 (s_1^2 + s_2^2).$$

Пример 3. Изучали влияние лечебного препарата на массу тела лабораторных мышей. Были получены следующие результаты. Характеристики опытной группы (*n*$_1$ =9):

$$\overline{x}_1 = 74,1 \text{ г}; \quad s_1^2 = \frac{302,89}{9-1} = 37,86;$$

контрольной группы (*n*$_2$ = **11)** :

$\overline{x}_2$= 68,8 г; $s_2^2 = \frac{443,64}{11-1} = 44,36$.

Разность между $\overline{x}_1$ и $\overline{x}_2$, равная 5,3+2,89, оказалась статистически недостоверной. Определим число наблюдений *n*, которое необходимо провести при уменьшении ошибки разности вдвое, т. е. *s*$_d$*=2,87/2*= 1,445. Примем *t=2*. Имеем *a*= 11/9= 1,222 и Δ=2·1,445=2,89. Отсюда

$$n = \left(\frac{2}{2{,}89}\right)^2 \left(\frac{44{,}364}{1{,}22} + 37{,}861\right) = 35{,}52 \approx 36.$$

При альтернативной группировке данных, когда численность выборочных групп выражают в долях единицы, планируемый объем наблюдений определяют по формуле

$$n = \frac{t^2 p(1-p)}{\Delta^2},$$

где p – доля вариант, обладающих данным признаком; $\Delta = ts_p$. Если доли выражают в процентах от общего числа наблюдений, формула принимает следующий вид:

$$n = \frac{t^2 p(100-p)}{\Delta^2}.$$

Лекция № 3. Основные типы распределений признаков.

Начиная биологический эксперимент или приступая к наблюдению, невозможно точно сказать, каков будет результат – уровень численности животных в данном районе, вес еще не отловленных особей, количество сахара в крови через час после введения препарата и т. п. В этом смысле биологические явления случайны, точно не предсказуемы. Однако любому биологу ясно, что случайность эта не абсолютна. Несмотря на сложность точного прогноза, приблизительный результат можно предугадать, в частности, предсказав, что интересующая нас величина будет находиться в пределах некоторого интервала между конкретными минимальными и максимальными значениями. Ясно, например, что рост человека вряд ли превысит два или будет ниже полутора метров. Вариационная статистика может дать и более точный прогноз, ориентируясь на известные законы поведения случайных величин, относящихся к разным типам распределений. При этом под распределением признаков (случайных величин, объектов) понимается соотношение между их значениями и частотой встречаемости.

Среди многих известных типов распределений мы рассмотрим лишь пять (нормальное, биномиальное, Пуассона, альтернативное, полиномиальное, равномерное). Для описания природных явлений иногда реалистичные основания имеет распределение *гипергеометрическое* (безвозвратное изъятие). Распределение *негативное биномиальное* подходит для случая, когда вероятности элементарных событий (p и q) не постоянны.

Распределения *Максвелла* и *Рэлея* имеют умеренную правостороннюю асимметрию и описывают поведение непрерывных положительных случайных величин. Распределения *Парето* и *показательное* пригодны для описания резко правосторонне асимметричных вариационных рядов с перепадом частот. Распределение *логнормальное*, или логарифмически нормальное, характеризуется тем, что логарифмы исходных значений выборки образуют правильное нормальное распределение; эта модель подходит для описания признаков, имеющих распределения с умеренной правосторонней асимметрией, это в первую очередь концентрации веществ в различных средах, т. е. гидрохимические, физиологические и биохимические показатели.

Зная тип распределения, можно воспользоваться разработанными специально для него приемами математической обработки и получить наиболее полную информацию о явлении, точнее оценить различия между параметрами разных выборок.

Нормальное распределение

Наиболее характерный тип распределения *непрерывных случайных величин*, из него можно вывести (к нему сводятся) все остальные. Распределение *симметрично*, причем крайние значения (наибольшие и наименьшие) появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем оно чаще встречается (рис. 1).

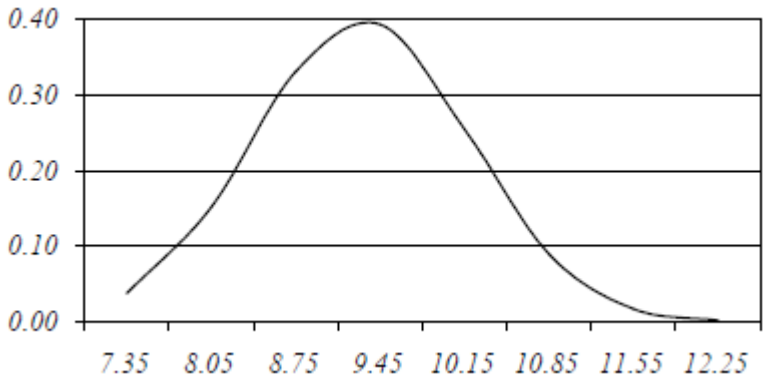

Рис. 1. *Нормальное распределение с параметрами n = 63, M = 9.3, S = 0.79. По оси абсцисс – вес тела землероек-бурозубок, по оси ординат – табличные значения для нормального распределения. Рассчитать ординаты нормальной кривой для конкретного значения xi можно по формуле:*

$$p_i = \left(1/\sqrt{2\pi}\right) \cdot e^{-(x_i - M)^2/2 \cdot S^2}$$

Среднее квадратичное отклонение примерно 4 раза укладывается в размахе изменчивости признака и по величине значительно *уступает* средней. Геометрически стандартное отклонение равно расстоянию от центра кривой распределения до точки перегиба кривой.

Биномиальное распределение

Во многом близко к нормальному. Отличие состоит лишь в том, что оно характеризует поведение *дискретных признаков*, *выраженных целыми числами*. Как правило, для описания биологических признаков подходит симметричное биномиальное распределение, у которого дисперсия много меньше средней. Распределение организуется в процессе обора *проб* (объемом больше одного, $m > 1$). Число классов больше двух, $k > 2$.

Примерами описания признаков с помощью биномиального распределения могут служить число поврежденных участков на листьях, число волосков на единице площади шкурки, количество лучей в плавниках рыб, число хвостовых щитков у рептилий, плодовитость (размер выводка) самок и т. п. В основе биномиального распределения лежит альтернативное проявление качественного признака: он может присутствовать у единичного объекта или отсутствовать, проявиться или нет. Отдельный корнеплод может быть больным или здоровым (признак качественный), тогда *проба* из нескольких корнеплодов будет содержать некоторое *число* здоровых корнеплодов (признак количественный), а множество равнообъемных проб образует уже выборку чисел, для которой можно построить гистограмму распределения. Вероятность отдельного события (корнеплод больной) составляет p, а вероятность альтернативного события (корнеплод здоровый) равна $q = 1 - p$. При равенстве вероятностей событий $p = q = 0.5$. большинство проб (вариант) будет иметь около половины возможных событий (поровну больных и здоровых корнеплодов); распределение примет симметричную форму. В случае неравенства вероятностей наблюдается та или иная степень асимметрии распределения.

Рассмотрим результаты изучения плодовитости серебристо-черных лисиц (число щенков на самку). Для построения вариационного ряда берем 8 классов, классовый интервал для этого дискретного признака составит $dx = 1$.

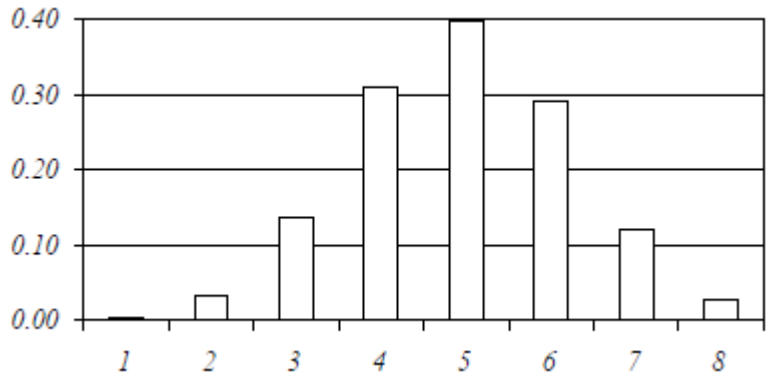

Рис.2. *Биномиальное распределение (n = 76, M = 4.95, S = 1.33).*
По оси абсцисс – число щенков лисицы на одну самку, по оси ординат – частости (относительные частоты)

Все основные параметры распределения вычисляются по рассмотренным выше формулам:

$$M = \frac{\sum x}{n} = 4.96 \text{ экз./самку},$$

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}} = 1.33 \text{ экз./самку},$$

$$m_M = \frac{s}{\sqrt{n}} = \frac{1.33}{\sqrt{63}} = 0.1676 \text{ экз./самку},$$

$$m_S = \frac{s}{\sqrt{2 \cdot n}} = \frac{1.33}{\sqrt{2 \cdot 63}} = 0.1185 \text{ экз./самку}.$$

Для расчета параметров биномиального распределения можно воспользоваться другими, более простыми формулами, если предварительно рассчитать вероятности p и q (в нашем случае p = 0.62, q = 0.38):

$$M = m \cdot p = 8 \cdot 0.62 = 4.96 \text{ экз./самку},$$

$$S = \sqrt{m \cdot p \cdot q} = \sqrt{8 \cdot 0.62 \cdot 0.38} = 1.37 \text{ экз./самку}.$$

Результаты оказываются идентичными с точностью до ошибки округления. Доверительный интервал для параметров биномиального распределения строится так же, как и для нормального распределения: $M \pm tm_M$, $S \pm tm_S$.

Распределение Пуассона

Это вариант описания стохастического поведения *дискретных количественных признаков* для случаев, когда *вероятность элементарных альтернативных событий неодинакова*, одно из них наблюдается заметно чаще другого ($p << q$) (классический пример – попадание гитлеровских авиационных бомб в разные кварталы Лондона). Закон Пуассона описывает редкие события, происходящие 1, 2, 3 и т. д. раз на сотни и тысячи обычных событий. Поведение биологических объектов, соответствующее закону Пуассона, наблюдается в том случае, когда по пробам случайно распределены редкие объекты. Примеры таких явлений – частота нарушений хромосомного аппарата на каждую тысячу митозов, встречаемость семян сорняка в большой серии навесок семян культурного растения, число повторных попаданий животных в ловушки, встречаемость животных на отрезках длинных маршрутов (или на пробных площадках обширной территории), отловы животных в отдельные промежутки времени при длительных наблюдениях.

Случайная величина, распределенная по закону Пуассона, определяется при подсчете числа элементарных событий *в пробе* (в группе, в навеске, на участке, на этапе). Число объектов в пробе больше 1 ($m > 1$), число классов больше двух ($k > 2$).

Распределение Пуассона резко асимметрично, причем *дисперсия равна средней арифметической*, что может служить критерием для оценки характера распределения изучаемого признака (рис. 3). В течение одного года (1946) пометили кольцами и выпустили на волю 32 буревестника.

Число повторных отловов, x	Число отловленных животных, a	Число случаев повторного отлова, $x \cdot a$
0	15	0
1	7	7
2	7	14
3	2	6
4	1	4
n	32	31

В последующие пять лет часть из них отлавливали повторно: 7 экз. по одному разу, 7 – по два, 2 – по три, 1 экз. – четыре раза, 15 экз. окольцованных птиц повторно не попадались. Число классов составляет $k = 4$, интервал $dx = 1$. Асимметрия в частотах встречаемости птиц позволяет предполагать распределение Пуассона.

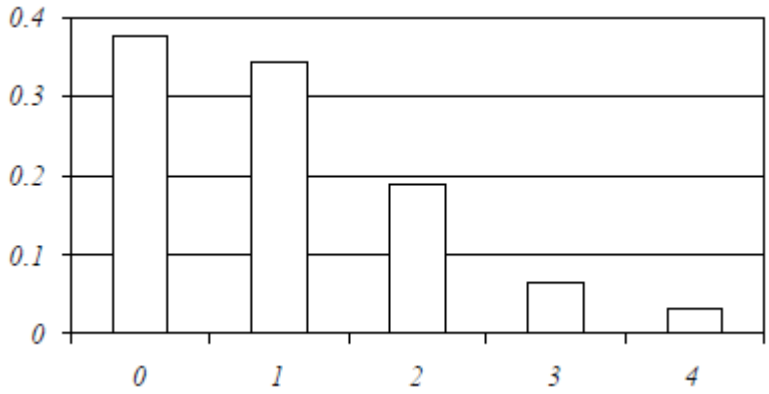

Рис. 3. *Распределение Пуассона с параметрами n = 32, M ≈ S2 = 0.968.*

По оси абсцисс – число повторных отловов, по оси ординат – частости (относительные частоты)

Расчеты показали, что средняя арифметическая (M) примерно равна дисперсии (S^2):

$$M = \frac{\sum x}{n} = \frac{31}{32} = m \cdot p = 4 \cdot 0.242 = 0.968 \text{ экз.},$$

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}} = \sqrt{\frac{69 - \frac{(32)^2}{32}}{(32-1)}} = 1.121 \text{ экз.}, S^2 = 1.257,$$

$S^2 \approx M$.

Критерий Фишера не выявил достоверных отличий между средней и дисперсией: $F = 1.257 / 0.968 = 1.157 < F(0.05,31,31) = 1.8$, что свидетельствует о соответствии наблюдаемого распределения закону Пуассона.

Возможен расчет параметров по более простым формулам:

$M = m \cdot p, S = m \cdot p .M = m \cdot p,\ S = \sqrt{m \cdot p}$.

Доверительный интервал для параметров распределения Пуассона определить несколько сложнее, чем для других типов (Ивантер, Коросов, 2003).

Альтернативное распределение

Распределение *дискретной случайной величины*, имеющей лишь два противоположных (разнокачественных) значения (два класса, $k = 2$). В одной пробе (в одном наблюдении) содержится одна варианта ($m = 1$), одно из двух возможных значений. Вероятности каждого из них могут быть равны ($p = q$) либо не равны ($p < q; p > q$). Примеры: самцы и самки, больные и здоровые организмы, сработавшие и пустые ловушки на одной учетной линии, два варианта аллельных признаков, вакцинированные и невакцинированные пациенты среди заболевших и др. (рис. 4). Вычисления констант достаточно просты и не требуют построения вариационного ряда.

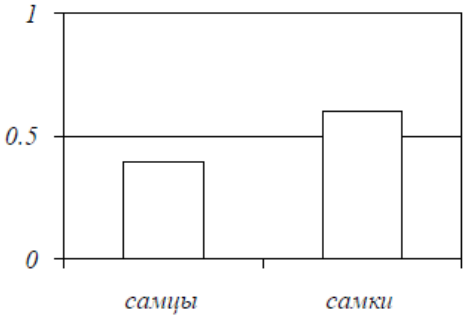

Рис. 4. *Альтернативное распределение (два класса вариант). По оси ординат – частости (доли) этих групп*

Важнейшей характеристикой является доля (p) вариант определенного вида (А), представленных общим числом nA в пределах выборки объемом n:

$$p = \frac{n_A}{n}.$$

Если исходы отдельных испытаний выразить числами 0 или 1 (что аналогично отбору проб с объемом m = 1), доля вариант совпадает со средней арифметической, рассчитанной для всех значений:

$$M = \frac{\sum x}{n}.$$

Например, результат отловов полевок из природных популяций показал, что в исследуемой группе (200 особей) было 120 самок и 80 самцов.

В данном случае мы имеем дело с альтернативным признаком (самка – самец). Из 200 проб 120 содержат самок (значение 1), 80 – не содержат (значение 0), так получаем выборку n = 200:

00000000000000000000000000000000000000000000000000000000001
11111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111.

Доля вариант со значением 1 составляет:

$$p = \frac{n_A}{n} = \frac{120}{200} = 0{,}6.$$

что совпадает со средней арифметической для всего ряда:

$$M = \frac{\sum x}{n} = \frac{120}{200} = 0{,}6.$$

Для альтернативного распределения могут применяться те же формулы расчета выборочных параметров, что и для биномиального распределения.

Средняя (доля самок):

$M = m \cdot p = 1 \cdot 0.6 = p = 0.6$.

Стандартное отклонение (при m = 1):

$$S = \sqrt{m \cdot p \cdot q} = \sqrt{p \cdot q} = \sqrt{0{,}6 \cdot 0{,}4} = 0{,}24.$$

Ошибка средней (ошибка доли самок):

$$m = \frac{S}{\sqrt{n}} = \frac{0{,}24}{\sqrt{200}} = 0{,}017.$$

Доверительный интервал для альтернативных признаков (их долей, процентов и частот) строится с помощью φ-преобразования Фишера, что дает более точные границы, особенно если доли сильно отличаются. Сначала вместо значения доли (процента) одного признака объектов берут значение φ (фи), найденное по формуле $\varphi = 2 \cdot \arcsin \sqrt{p}$ или по таблице 10 прил. Затем вычисляют ошибку: $m_{\varphi} = 1/n$, обе доверительные границы $\varphi_{лев.} = \varphi - tm_{\varphi}$, $\varphi_{прав.} = \varphi + tm_{\varphi}$, после чего с помощью таблицы переводят найденные значения обратно в проценты.

Найдем доверительные границы для доли самок полевок p = 0.6 при уровне значимости α = 0.05. Используя таблицу 10 прил. и проводя расчеты, получаем: φ(60%) = 1.772, = 1/ 200 $m_{\varphi} = 1/\sqrt{200} = 0.0707$,

$\varphi_{лев.} = 1.772 - 1.96 \cdot 0.0707 = 1.6334$,

$\varphi_{прав.} = 1.772 + 1.96 \cdot 0.0707 = 1.9106$,

$p_{лев.}(1.6334) = 53.1\%$,

$p_{прав.}(1.9106) = 66.4\%$.

Доля самок в генеральной совокупности (популяции полевок) составляет минимум 53.1%, максимум 66.4%.

Полиномиальное распределение

Наблюдается для *качественных признаков,* имеющих не два альтернативных свойства, но *несколько возможных проявлений качества*. Примеры полиморфизма популяций – из этой области. В их числе варианты окраски покровов и волос, типы рисунков в определенных областях тела, способы жилкования листьев растений или крыльев насекомых, варианты расположения и формы щитков рептилий и другие проявления множественности фенотипов особей. Формализуя описание, укажем, что в одной пробе содержится одна варианта (m = 1), но типов вариант (морф, фенотипов) больше, чем два (k > 2).

Примером полиномиального (иначе – мультиномиального) распределения может служить встречаемость 4 фенов головы живородящей ящерицы – 4 вариантов контакта лобно-носового, предлобных и лобного щитков (рис. 5).

Лучше всего выборка может быть представлена вариационным рядом – частотами (pj) встречаемости в популяции особей с данным (j-м) проявлением качественного признака и общим числом морф (k). Для более емкого представления ряда и учета характера распределения частот между разными морфами используется величина «среднее число фенотипов»: $\mu = \Sigma(p_j)^2$, статистическая ошибка которой рассчитывается так:

$$m_{\mu} = \sqrt{\frac{\mu \cdot (k - \mu)}{n}}.$$

Среднее число фенотипов (μ) равно числу фенотипов (k) только тогда, когда частоты всех фенотипов одинаковы (p1 = p2 = … = pj … = pk), и меньше во всех других случаях.

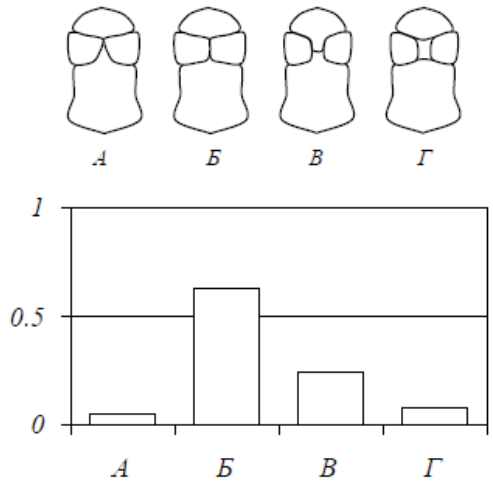

Рис. 8. *Полиномиальное распределение (4 фена головы ящерицы).*

По оси ординат – частости фенов среди 64 сеголетков живородящей ящерицы, отловленных под Петрозаводском

Равномерное распределение

Частный случай распределения альтернативного и полиномиального.

Равномерное распределение характеризуется одинаковой частотой встречаемости всех значений дискретного признака ($p = q$ для двух классов или $p1 = p2 = \ldots = pj \ldots = pk$ для нескольких классов). Такой тип распределения можно использовать для формулирования гипотез при анализе частот генов и фенов в популяциях, при подсчете тест-организмов, выживших в токсикометрическом эксперименте, можно предположить, что ветви дерева могут равномерно располагаться по сторонам света.

Лекция № 4. Оценка различий двух выборок

В любых биологических экспериментах и наблюдениях особое значение имеют различия, на основании которых судят об эффективности действия тех или иных факторов, например, по разности между опытной и контрольной группами делают заключение о результатах опыта. Точно так же по соответствующим изменениям морфофизиологических показателей определяют возрастные, сезонные и популяционные особенности животных. При этом особенно важно оценить статистическую *достоверность разности*, т. е. определить, можно ли данное различие считать закономерным, *характерным для всей генеральной совокупности* и рассматривать его как результат действия особенных факторов, или же оно случайно и является следствием недостаточного количества данных и в следующих опытах может не проявиться.

Обнаружение достоверных отличий статистических параметров – первый шаг к познанию новых биологических закономерностей, причем количественно доказанных. Ответ на вопрос о достоверности или случайности отличий дают статистические критерии, среди которых самые распространенные критерии *t* Стьюдента и *F* Фишера. Вычисление их ведется по специальным формулам (различным в зависимости от сравниваемых параметров и типов распределения). Полученные этим способом значения критериев (для чего в формулы подставляются экспериментальные данные) сравнивают с табличными при выбранном уровне значимости (обычно 0.05) и числе степеней свободы (объемы выборок без числа ограничений). Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны, в разных выборках действительно проявилось действие разных факторов или разных уровней одного фактора. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Последнее говорит о том, что различия случайны, никакого определенного вывода о побудительных причинах отличий сделать нельзя, нулевая гипотеза остается неопровергнутой.

При сравнении выборок по степени выраженности признака говорят о достоверности (недостоверности) отличий средних арифметических и долей, а при сравнении по уровню изменчивости показателей – о достоверности (недостоверности) отличий стандартных отклонений (дисперсий) и коэффициентов вариации. Особый случай представляет сравнение двух выборок по характеру распределения (достоверность отличия частот), а также общее отличие выборок без указания определенных параметров (для признаков в полуколичественных единицах).

Сравнение средних арифметических

Задача сравнения выборочных средних – это вопрос о том, действовал ли при составлении одной из выборок новый систематический фактор по сравнению с другой выборкой. В терминах статистики отличия между средними могут иметь два противоположных источника:

1) Обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности.

2) Выборки взяты из разных генеральных совокупностей, отличие средних вызвано в основном действием разных доминирующих факторов (а также и случайно).

Статистическая задача состоит в том, чтобы сделать обоснованный выбор. Исходно предполагается (Но): «Достоверных отличий между средними нет». Отличить закономерное от случайного можно только на основе знания законов поведения случайной величины. Для исключения чужеродных («выскакивающих») вариант мы применяли закон нормального распределения: в диапазоне четырех стандартных отклонений, $M \pm 1.96 \cdot S$, отклонение вариант от средней происходит по случайным причинам; за границами этого диапазона лежат чужеродные для данной выборки значения. Поскольку выборочные средние имеют нормальное распределение, критерий отличия двух выборочных средних также базируется на *свойствах нормального распределения*: в границах $M_{общ.} \pm 1.96 \cdot m$ (или приблизительно $M_{общ.} \pm 2 \cdot m$) выборочные средние арифметические отличаются от общей (генеральной) средней по случайным причинам. Тогда рабочая формула для t критерия отличия средних будет:

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} \sim t_{(a,df).}$$

Следует помнить, что разность средних нужно брать по модулю, т. е. без учета знака. Полученное этим способом значение критерия t Стьюдента сравнивают с табличным при выбранном уровне значимости (обычно для $\alpha = 0.05$) и числе степеней свободы (*объемы выборок без числа ограничений*, $df = n1 + n2 - 2$). Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Последнее говорит о том, что различия случайны, никакого определенного вывода сделать нельзя, нулевая гипотеза остается неопровергнутой.

При сравнении выборочных параметров нормального и биномиального распределений используется одна и та же формула. Например, в процессе специальных исследований было установлено, что у стариков до лечения инсулином среднее содержание белков в крови составляло 81.04 ± 1.7, а после лечения – 79.33 ± 1.6. Нетрудно видеть, что полученные величины неодина-

ковы. Но достоверно ли это различие, закономерно ли оно? Можно ли на его основании утверждать, что лечение инсулином понижает содержание белков в крови? Ответ на этот вопрос может дать критерий достоверности различий средних арифметических. Согласно общей нулевой гипотезе, средние не отличаются. Проверим ее с помощью критерия Стьюдента:

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} = \frac{81{,}04 - 79{,}33}{\sqrt{1{,}7^2 + 1{,}6^2}} = 0{,}7.$$

По таблице граничных значений критерия (табл. 6*П)* находим, что для уровня значимости α = 0.05 и числа степеней свободы *df* = 20 + 20 − 2 = 38 величина критерия составляет *t*(0.05,39) = 2.03. Поскольку полученное значение (0.7) меньше табличного (2.03), нулевая гипотеза сохраняется, различия между средними величинами статистически недостоверны (незначимы). Следовательно, влияние инсулина на содержание белков в крови приведенными выше данными не подтверждается и остается недоказанным, возможно, из-за недостаточного числа определений.

Сравнение долей

При сравнении достоверности различия долей или процентов (*p*) признаков, характеризующихся альтернативным распределением, применяют критерий Фишера с φ-преобразованием. Вместо процентов берут физначения φ = arcsin $\sqrt{p}$ (или по таблице 10*П)* и подставляют их в формулу:

$$F = \frac{(\varphi_1 - \varphi_2)^2 \cdot n_1 \cdot n_2}{n_1 + n_2} \sim F_{(a,df1,df2)},$$

где φ_1 и φ_2 – преобразованные доли,

n_1 и n_2 – объемы выборок.

Полученное значение сравнивают с табличным в соответствии с заданным уровнем значимости, α = 0.05, и числом степеней свободы: *df*1 = 1, *df*2 = $n_1 + n_2 - 2$.

Например, в процессе учетов мелких млекопитающих в двух разных биотопах, где стояло по 200 ловушек, попалось соответственно 5 и 15 зверьков. Отличается ли численность животных на этих площадках? Если рассматривать ловушку как варианту, способную принимать два значения – «пустая» и «сработавшая» (со зверьком), то получаем выборку вариант (ловушек) с альтернативным распределением. Число пойманных особей можно пересчитать в процент сработавших ловушек: M_1 = 100% ⋅ 5 /200 = 2.5%,

M_1 = 100% ⋅ 15 / 200 = 7.5%. По таблице 10*П* находим значения φ и вычисляем значение критерия:

$$F = \frac{(0.318 - 1.555)^2 \cdot 200 \cdot 200}{200 + 200} = 5.62.$$

Полученная величина (5.62) больше критической $F_{(0.05,\ 1,\ 398)}$ = 3.9, значит, численность мелких млекопитающих во втором биотопе достоверно выше, чем в первом.

Сравнение показателей изменчивости

Наиболее точным методом определения достоверности различий между выборочными дисперсиями служит критерий *F* Фишера в форме отношения дисперсий (большее значение должно стоять в числителе):

$$F = \frac{S_1^2}{S_2^2} \sim F_{(a,df1,df2)},$$

где $S_1 > S_2$, $df_1 = n_1 - 1$, $df_2 = n_2 - 1$.

Если полученная величина *F* больше табличного значения при принятом уровне значимости (табл. 7*П* для α = 0.05 и табл. 8*П* для α = 0.01) и числе степеней свободы (*df*1 и *df*2), то различие между дисперсиями признается достоверным; если она меньше, то расхождение между ними может считаться несущественным, случайным, т. е. нулевая гипотеза не отвергается.

Рассмотрим такой пример. При сравнении по показателю плодовитости (число эмбрионов на самку) двух популяций красной полевки с разным уровнем численности (у первой, горной, популяции плотность населения в два раза выше, чем у равнинной) оказалось, что при очень близких средних арифметических (соответственно M_1 = 5.8 и M_2 = 5.4, разница статистически недостоверна) стандартные отклонения значительно различаются: S_1 = 1.82, S_2 = 0.52 (при n_1 = 27, n_2 = 12). Отсюда

$$F = \frac{S_1^2}{S_2^2} = \frac{3.3124}{0.2704} = 12.25.$$

Полученное значение критерия (*F* = 12.2) больше табличного *F*(0.05, 26, 11) = 2.6, следовательно, нулевую гипотезу о случайности отличий можно отбросить, сделав вывод о том, что показатели изменчивости плодовитости в разных по численности популяциях достоверно отличаются. С биологических позиций это понятно, поскольку генетические отличия между особями практически по всем признакам, включая плодовитость, в больших популяциях выше, чем в малых. Новым фактором, усиливающим изменчивость особей в выборке, становится возможность появления аберрантных форм в условиях более свободной панмиксии.

Коэффициенты вариации также можно использовать для сравнения изменчивости разных показателей. Достоверность отличий коэффициентов оценивается с помощью критерия Стьюдента по формуле:

$$t = \frac{|CV_1 - CV_2|}{\sqrt{m_1^2 + m_2^2}}$$

где CV_1, CV_2 и m_1, m_2 – значения и ошибки коэффициентов вариации.

Вывод о достоверности отличий делается в том случае, если рассчитанное значение превысит табличное при заданном уровне значимости α = 0.05 и числе степеней свободы $df = n1 + n2 - 2$. Сравним по критерию Стьюдента изменчивость веса тела землероек и плодовитости лисиц:

$CV_1 = 8.6 \pm 0.77\%$, $n_1 = 63$; $CV_2 = 26.7 \pm 2.2\%$, $n_2 = 76$, отсюда

$$t = \frac{|8.6 - 26.7|}{\sqrt{0.77^2 + 2.2^2}} = 7.76.$$

Поскольку полученное значение (7.8) больше табличного (t(0.05, 137) = 1.96), изменчивость плодовитости лисиц достоверно выше, чем изменчивость веса тела землероек.

Сравнение выборок с помощью непараметрических критериев

Описанные выше статистические критерии (t, F и др.) относятся к *параметрическим*, т. к. *используют стандартные параметры распределений* (M, S, n). Они связаны с законом нормального распределения и применяются для оценки расхождения между генеральными параметрами по выборочным показателям сравниваемых совокупностей. Существенным достоинством параметрических критериев служит их большая статистическая мощность, т. е. широкие разрешающие возможности, а недостатком – трудоемкость расчетов, неприменимость к распределениям, сильно отклоняющимся от нормального, а также при исследовании качественных признаков.

Наряду с параметрическими критериями для ориентировочной оценки расхождений между выборками (особенно небольшими) применяются так называемые непараметрические критерии, ориентированные в первую очередь на исследование соотношений *рангов* исходных значений вариант. *Ранг – это число натурального ряда, которым обозначается порядковый номер каждого члена упорядоченной совокупности вариант.* Эта замена позволяет сравнивать выборки как по количественным, так и по качественным признакам, значения которых не имеют числового представления, но которые можно ранжировать. Конструкции непараметрических критериев отличаются простотой.

Вся процедура состоит из трех этапов – упорядочивание и ранжирование вариант, подсчет сумм рангов в соответствии с правилами данного критерия, сравнение полученной величины с табличным значением критерия.

При этом с параметрическими критериями их роднит общая идеологическая подоплека. Нулевая гипотеза, как правило, состоит в том, что сравниваемые выборки взяты из одной и той же генеральной совокупности, значит, характер распределения вариант в этих выборках должен быть сходным. Поскольку вместо самих значений вариант используются ранги, все непараметрические методы исследуют один вопрос, насколько равномерно варианты разных выборок «перемешаны» между собой. Если варианты разных выборок более или менее регулярно чередуются в общем упорядоченном ряду, значит, они распределены сходным образом и отличий между совокупностями нет.

Если же выборки пересекаются неполно (смешиваются только краями распределений, либо одна поглощает другую), то становится ясно, что эти выборки взяты из разных генеральных совокупностей (со смещенными центрами или разными дисперсиями).

Среди множества известных методов мы рассмотрим два метода: Уилкоксона – Манна – Уитни (довольно точный, но не самый простой для вычислений) и критерий *Q* Розенбаума. (простой для расчетов, но не очень точный).

Критерий U Уилкоксона – Манна – Уитни

Этот метод сравнения двух выборок признается наиболее чувствительным и мощным среди прочих непараметрических критериев. Согласно нулевой гипотезе, сравниваемые совокупности имеют одинаковые распределения.

Техника метода состоит в том, что все варианты сравниваемых совокупностей ранжируют в одном общем ряду: каждому значению присваивают ранг, порядковый номер. При этом одинаковым (повторяющимся) значениям вариант должен соответствовать один и тот же средний ранг (они как бы «делят места»). После этого ранги вариант суммируют отдельно по каждой выборке: $R_1 = \Sigma r_i$, $R_2 = \Sigma r_j$, $i = 1, 2, \ldots, n_1$, $j = 1, 2, \ldots, n_2$; $n = n_1 + n_2$ и вычисляют величину критерия:

$$t = \frac{U - 0.5 \cdot n_1 \cdot n_2}{\sqrt{(n_1 \cdot n_2(n+1)/12)}},$$

где $U = \max(U_1, U_2)$ – максимальное значение из двух величин:

$U_1 = n_1 \cdot n_2 + 0.5 \cdot n_1 (n_1 + 1) - R_1$,

$U_2 = n_1 \cdot n_2 + 0.5 \cdot n_2 (n_2 + 1) - R_2$.

Если выборка достаточно велика ($n > 20$), величина статистики t сравнивается с табличным значением критерия Стьюдента для $df = \infty$ и $\alpha = 0.1$ (т. е. только для верхней 95%-й области нормального распределения). Считается, что метод хорошо работает для выборок объемом больше 10. В случае с меньшими выборками нужно пользоваться таблицами Уилкоксона – Манна – Уитни (табл. 11*П*).

В качестве примера сравним 5- и 35-дневных щенков песцов по активности фермента каталазы в сердце (E):

5-дневные: 41, 44, 31, 38, 43, 29, 71, 45, $M = 42.6$, $S = 12.8$, $n1 = 8$,

35-дневные: 52, 51, 62, 52, 52, 50, 54, 62, 31, $M = 51.7$, $S = 9.0$, $n2 = 9$.

Высокие коэффициенты вариации (30 и 17%) говорят о том, что распределения признаков, скорее всего, не соответствуют нормальному. Поэтому сравнивать средние следует с помощью непараметрического критерия.

Ранжируем всю совокупность – упорядочим значения выборок по возрастанию:

E_5	29	31	38	41	43	44	45	71	
E_{35}	31	50	51	52	52	52	54	62	62

Затем упорядочим все значения вместе, но так, чтобы значения каждой выборки располагались в двух отдельных рядах ($E5$, $E35$). Такое расположение упрощает назначение рангов (ряды $r5$, $r35$) и суммирование рангов (R):

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	R
E_5	29		31	38	41	43	44	45									71	
E_{35}		31							50	51	52	52	52	54	62	62		
r_5	1		2.5	4	5	6	7	8									17	50.5
r_{35}		2.5							9	10	12	12	12	14	15.5	15.5		102.5

$U_1 = 9.8 + 0.5 \cdot 9 \cdot (9+1) - 50.5 = 66.5$,

$U_2 = 9.8 + 0.5 \cdot 8 \cdot (8+1) - 102.5 = 5.5$,

$U_{max} = (U_1, U_2) = 66.5$, $n = 8+9 = 17$,

$$t = \frac{66.5 - 0.5 \cdot 9 \cdot 8}{\sqrt{(9 \cdot 8 \cdot 18/12)}} = 2.93.$$

Полученное значение (2.93) больше табличного ($t(0.1, \infty) = 1.65$, табл. 6*П*), т. е. активность каталазы с возрастом меняется. Раз выборки малы, воспользуемся точными таблицами Уилкоксона – Манна – Уитни (табл. 11*П*). Получаем $t(0.05, n1, n2) = t(0.05, 8, 9) = 51$. Полученное значение (66.5) больше табличного (51), следовательно, различия между выборками достоверны.

Лекция № 5. Оценка влияния фактора

При изучении и анализе сложных и многообразных причинно-следственных отношений между объектами и явлениями биологу приходится учитывать целый комплекс внешних и внутренних факторов, от которых в конечном итоге зависят уровень и ход наблюдаемых процессов, те или иные биологические свойства живых организмов, их динамика и разнообразие.

При этом зачастую важно оценивать не только роль одного из многочисленных внешних факторов, но и их взаимодействие при констелляционном влиянии на популяцию или организм.

Идейная база для изучения действия факторов содержится уже в методе сравнения двух выборок. Биологическим содержанием операции сравнения двух выборок, в конце концов, выступает поиск факторов, ответственных за смещение средних арифметических или усиление изменчивости признаков. Развивая это направление биометрического исследования, можно не ограничиваться только двумя «дозами» фактора, но изучить серию ситуаций, в которых фактор проявлял разную силу действия на результативный признак – от самого слабого до самого сильного. При этом каждому уровню фактора будет соответствовать отдельная выборка и общая задача получит формулировку «сравнить несколько выборок». В терминах факториальной биометрии вопрос о влиянии фактора на признак звучит так: сказывается ли отличие условий получения разных выборок на качестве (значениях) вариант? В терминах статистики вопрос звучит несколько иначе: из одной ли генеральной совокупности отобраны все выборки, оценивают ли выборочные средние арифметические одну и ту же генеральную среднюю? Вариантов ответа может быть только два:

1. Все выборки отобраны из одной генеральной совокупности, условия возникновения вариант одни и те же.
2. Выборки отобраны из разных генеральных совокупностей, условия возникновения вариант выборок различаются.

В постановке вопроса можно уловить противоречие. Выше было сказано, что по условию задачи выборки формировались в разных условиях, и тут же предполагается, что условия были одинаковые. На самом деле противоречия нет, поскольку речь идет об определении чувствительности признака к действию фактора. Условия формирования выборок могут отличаться, но они могут никак и не сказаться на величине изучаемого признака, не отразиться на значениях вариант. Смысл статистического сравнения в том и состоит, чтобы оценить эффективность действия фактора на признак, доказать реальность реакции вариант выборок на разные условия их формирования. В сферу исследования можно вовлекать как один, так и два признака,

как количественные, так и качественные характеристики. В каждом случае процедура анализа несколько отличается.

Однофакторный дисперсионный анализ количественных признаков

Дисперсионный анализ позволяет оценить степень и достоверность отличия нескольких выборочных средних одновременно, т. е. изучить влияние одного контролируемого фактора на результативный признак путем оценки его относительной роли в общей изменчивости этого признака, вызванной влиянием всех факторов. Сущность дисперсионного анализа заключается в расчленении общей вариации (дисперсии) изучаемого признака, вычисляемой по сумме квадратов отклонений отдельных вариант (x) от средней арифметической всего комплекса наблюдений (M), на его составные части – дисперсию, вызванную организованными, учитываемыми в исследовании факторами (факториальную дисперсию), оценивающую межгрупповую изменчивость, и дисперсию, обусловленную остальными, неорганизованными в данном исследовании факторами (внутригрупповую, или случайную, дисперсию) отклонения отдельных значений от средней в группе.

Общая вариация (сумма квадратов) признака рассчитывается как сумма квадратов отклонений всех вариант (xi) от общей средней (M):

$C_{общ.} = \Sigma\,(x_i - M)^2$.

Факториальная (межгрупповая, межвыборочная) сумма квадратов рассчитывается как сумма квадратов отклонений частных средних (Mi) для каждой выборки (всего k выборок) от общей средней:

$C_{факт.} = \Sigma\,(M_j - M)^2$.

Остаточная (случайная, внутригрупповая) сумма квадратов есть сумма квадратов отклонений вариант каждой выборки (x_i) от своей средней (M_j):

$C_{случ.} = \Sigma\,(x_i - M_j)^2$.

Очевидно, что в общем комплексе наблюдений должно выполняться равенство $C_{общ.} = C_{факт.} + C_{случ.}$

Отношение сумм квадратов к соответствующему числу степеней свободы дает оценку величины дисперсии, или средний квадрат, иногда ее именуют варианса. Влияние изучаемого фактора отражает факториальная, или межгрупповая, дисперсия $S^2_{факт.}$, а влияние случайных неорганизованных в данном исследовании причин – случайная $S^2_{случ.}$, или внутригрупповая, остаточная дисперсия $S^2_{остат.}$:

$$S^2_{\text{факт.}} = \sum^{k} \left(M_j - M_{\text{общ.}}\right)^2 / df_{\text{факт.}},$$

где $df_{факт.} = k - 1$, $j = 1, 2, \ldots, k$, k – число сравниваемых средних.

$$S^2_{\text{случ.}} = \sum^{k} \sum^{n_j} (x_y - M_j)^2 / df_{\text{случ.}},$$

где $df_{случ.} = n - 1$, $i = 1, 2, \ldots, n$, n – число вариант всех выборок.

Сила влияния фактора определяется как доля частной суммы квадратов в общем варьировании признака. Показатель силы влияния изучаемого фактора составляет: $\eta^2_{факт.} = C_{факт.} / C_{общ.}$, неорганизованных (случайных):

$\eta^2_{случ.} = C_{случ.} / C_{общ.}$; сумма этих показателей, естественно, равна единице: $\eta^2_{факт.} + \eta^2_{случ.} = 1$. Заметим, что показатель силы влияния дисперсионного комплекса есть не что иное, как квадрат пирсоновского корреляционного отношения, которым и оценивается относительная доля влияния организованного (изучаемого) фактора в общем суммарном статистическом влиянии всех факторов, определяющих развитие данного результативного признака.

О достоверности оценок влияния факторов судят по уже знакомому нам критерию Фишера:

$$F = \frac{S^2_{\text{факт.}}}{S^2_{\text{случ.}}} \sim F_{(a, df1, df2)}$$

где $df1 = k - 1$, $df2 = n - k$, k – число градаций, n – общий объем всех выборок.

Проверяется нулевая гипотеза: «влияние фактора на признак отсутствует». Влияние считается доказанным, если величина расчетного критерия равна или превышает свое табличное значение с принятым уровнем значимости (обычно $\alpha = 0.05$) (F определяется по табл. 7*П*). Все параметры однофакторного дисперсионного анализа и порядок их вычислений представлены в таблице:

Составляющие дисперсии	Суммы квадратов (*SS*), *C*	Сила влияния, η^2	Степени свободы, *df*	Дисперсии (средний квадрат, *MS*), S^2	Критерий влияния, *F*
Факториальная	$C_{факт.} = \Sigma (M_j - M)^2$	$\frac{C_{факт.}}{C_{общ.}}$	$k - 1$	$S^2_{факт.} = \frac{C_{факт.}}{df_{факт.}}$	$F = \frac{S^2_{факт.}}{S^2_{случ.}}$
Случайная	$C_{случ.} = \Sigma (x_i - M_j)^2$		$n - k$	$S^2_{случ.} = \frac{C_{случ.}}{df_{случ.}}$	
Общая	$C_{общ.} = \Sigma (x_i - M)^2$				

Однофакторным называется анализ, изучающий действие на результативный признак только одного организованного фактора *A*. Для примера оценим

влияние растворенного в воде вещества на плодовитость дафний, используемых в качестве тест-объектов в водно-токсикологических экспериментах. В ходе предварительного исследования были получены четыре выборки, четыре группы значений плодовитости животных, выращенных в средах с разным содержанием химической добавки.

Сначала необходимо сгруппировать выборочный материал в комбинативную таблицу (организовать дисперсионный комплекс). Для этого варианты каждой выборки записываются в отдельные графы, именуемые градациями (табл.). Результативным признаком служит средняя плодовитость дафний за неделю (для иллюстративности расчетов она дана в целых числах). В нашем примере организованы 4 градации – чистая вода (контроль, градация *A*1; значения плодовитости 6, 5, 5, 7), слабая концентрация вещества (5 мг/л, *A*2; 8, 7, 6, 6), средняя (15 мг/л, *A*3; 8, 8, 7) и сильная (30 мг/л, *A*4; 8, 7, 9).

Предлагаемый ниже алгоритм расчетов позволяет использовать неравное число вариант в градациях. Расчеты показаны в таблице :

	Градации фактора									
	*A*1		*A*2		*A*3		*A*4			
	x	x^2	x	x^2	x	x^2	x	x^2		
	6 5 5 7	36 25 25 49	8 7 6 6	64 49 36 36	8 8 7	64 64 49	8 7 9	64 49 81	Σ	
Σx^2		135		185		177		194	691	$H1 = \Sigma\Sigma x^2 = 691$
Σx	23		27		23		24		97	$H2 = (\Sigma\Sigma x)^2/n =$
n	4		4		3		3		14	$= (97)^2/14 = 672$
$\Sigma x^2/n$	132		182		176.3		192		682.8	$H3 = \Sigma\Sigma x^2/n =$ $= 682.8$
M	5.8		6.8		7.67		8		6.93	

$$C_{факт.} = H3 - H2 = 682.8 - 672 = 10.76$$
$$C_{случ.} = H1 - H2 = 691 - 672 = 8.17$$
$$C_{общ.} = H1 - H3 = 691 - 682.8 = 18.93$$

Полученные значения позволяют вычислить дисперсии, определить силу влияния фактора и критерий достоверности Фишера.

Составляющие дисперсии	Суммы квадратов, C	Сила влияния, η^2	Степени свободы, df	Дисперсии, S	Критерий, F
Факториальная	10.76	57%	3	3.59	4.39
Случайная	8.17		10	0.82	
Общая	18.93			4.39	

Поскольку полученное значение критерия (F = 4.39) больше табличного (F(0.05,3,10) = 3.7) (табл. 7*П*), отличие факториальной и случайной дисперсий достоверно, влияние фактора значимо.

Отсюда следует биологический вывод: стимулирующее влияние изучаемого фактора (вещества) на плодовитость дафний относительно велико (57%) и достоверно (с вероятностью $P > 0.95$).

Непараметрический однофакторный дисперсионный анализ

Рассмотренные выше схемы дисперсионного анализа исходили из предположения о нормальном распределении изучаемого результативного признака. Когда для какого-либо признака нет уверенности, что выполняется предположение о его нормальном распределении, когда требуется провести анализ быстро и без особой точности, когда мало данных или они выражены *качественными признаками*, можно использовать схему непараметрического дисперсионного анализа. Этот метод более неприхотлив, но менее точен, нежели параметрический анализ. Он исследует распределения вариант в нескольких выборках. Нулевая гипотеза состоит в том, что распределения одинаковы, т. е. выборки взяты из одной генеральной совокупности.

Порядок вычислений состоит в том, что все варианты ранжируются в порядке возрастания. Затем суммируются ранги вариант по каждой выборке отдельно и рассчитывается критерий:

$$H = \frac{12}{n \cdot (n-1)} \cdot \left(\frac{R_1^{\,2}}{n_1} + \ldots + \frac{R_j^{\,2}}{n_j} + \ldots + \frac{R_k^{\,2}}{n_k} \right) - 3 \cdot (n+1) \sim \chi^2_{(\alpha,\, k-1)},$$

где n – число всех вариант,

n_j – объем j-й градации фактора,

R_j – сумма рангов для каждой j-й градации фактора,

k – число градаций фактора (j = 1, 2, …, k).

При объеме выборок больше 5 вариант статистика H имеет распределение хи-квадрат с $df = k - 1$ степенями свободы и сравнивается со значениями из табл. 9*П*.

Применим эту схему (табл. 10) к нашим данным из табл. 9, расположив их в строку.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Градация	1	1	1	1	2	2	2	2	3	3	3	4	4	4
Значение	5	5	6	7	6	6	7	8	7	8	8	7	8	9

Затем упорядочим и ранжируем их. Для нескольких одинаковых значений берется средний ранг.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Градация	1	1	1	2	2	1	2	3	4	2	3	3	4	4
Значение	5	5	6	6	6	7	7	7	7	8	8	8	8	9
Ранг	1.5	1.5	4	4	4	7.5	7.5	7.5	7.5	11.5	11.5	11.5	11.5	14

Наконец, разнесем ранги по градациям и подсчитаем необходимые суммы.

Градация	1	1	1	1	2	2	2	2	3	3	3	4	4	4
Значение	5	5	6	7	6	6	7	8	7	8	8	7	8	9
Ранг, R	1.5	1.5	4	7.5	4	4	7.5	11.5	7.5	11.5	11.5	7.5	11.5	14
Сумма, R				14.5				27			30.5			33
n				4				4			3			3
R^2/n				52.56				182.3			310.1			363

Общий объем выборки равен $n = 14$. Величина критерия H составит:

$$H = \frac{12}{14 \cdot 13} \cdot (52.56 + 182.3 + 310.1 + 363) - 3 \cdot 13 =$$
$$= 0.065934 \cdot 907.8958 - 45 = 14.86.$$

По таблице распределения статистики $\chi 2$ для $\alpha = 0.05$ и $df = 4 - 1 = 3$ находим $\chi 2(0.05, 3) = 7.81$. Полученное значение критерия (14.86) больше табличного (7.81), значит, отличие выборочных распределений достоверно. Химическая добавка действительно изменяет плодовитость дафний.

Двухфакторный дисперсионный анализ количественных признаков

Двухфакторный дисперсионный анализ исследует влияние на результативный признак двух факторов как порознь, так и совместно. Учет эффекта влияния каждого фактора по отдельности теоретически ничем не отличается от описанных выше схем. И там и тут оценивается изменчивость средних по градациям на фоне случайной изменчивости вариант внутри градаций, с помощью критерия Фишера устанавливается достоверность отличий межгрупповых дисперсий от внутригрупповых.

Двухфакторный дисперсионный анализ, естественно, требует более сложных вычислительных операций, чем однофакторный, но в принципе ничем не отличается от описанных выше схем. Однако это относится лишь к ортогональным (равномерным, или пропорциональным) комплексам, характеризующимся равной или, по крайней мере, пропорциональной численностью групп (в градациях содержатся одинаковые или пропорциональные числа вариант). Что же касается неортогональных многофакторных комплексов, то их анализ принципиально возможен, но имеет свои особенности, существенно усложняющие технику вычислений, и в настоящем пособии не рассматривается.

На практике вполне допустим и такой способ избегнуть сложностей обработки неравномерных комплексов, как искусственное превращение их в равномерные. Для этого нужно составить выборки одинаковой или пропор-

циональной численности, используя только часть имеющихся данных. Следует, однако, помнить, что такой отбор не должен быть субъективным. Чтобы не допустить возможной тенденциозности, лучше всего прибегнуть к жеребьевке.

Важным преимуществом двухфакторного дисперсионного анализа перед однофакторным служит то, что с его помощью удается определить варьирование по сочетанию градаций $C_{сочет.} = C_{AB}$, позволяющее получить новый и весьма ценный в биологическом отношении показатель – оценку влияния сочетанного действия (взаимодействия) факторов.

Общая вариация (сумма квадратов) признака теперь состоит из четырех компонентов за счет более детального разложения факториальной дисперсии.

Правило разложения вариаций предстает как:

$C_{общ.} = C_A + C_B + C_{AB} + C_{случ.}$,

$C_{факт.} = C_{общ.} - C_{случ.} = C_A + C_B + C_{AB}$.

Для расчетов используются следующие смысловые формулы:

$C_{общ.} = \Sigma(x_i - M)^2$,

$C_A = \Sigma(M_{Aj} - M)^2$, j – число градаций фактора A, MAj – групповые средние по градациям фактора А,

$C_B = \Sigma(M_{Bk} - M)^2$, k – число градаций фактора B, M_{Bk} – групповые средние по градациям фактора В,

$C_{случ.} = \Sigma(x_i - M_{xi})^2$,

$C_{AB} = C_{общ.} - (C_A + C_B + C_{случ.})$.

Сочетанное действие (взаимодействие) каждого из двух факторов проявляется в усилении или ослаблении непосредственного действия другого фактора на объект исследования. К примеру, неурожай кормов усугубляет негативное действие зимнего холода на численность популяций мелких млекопитающих.

Рассмотрим числовой пример – испытания стимулятора многоплодия при разной полноценности рационов. Полноценность рациона (первый фактор) представлена двумя градациями: $A1$ – рацион с недостатком минеральных веществ, $A2$ – рацион, полностью сбалансированный по всем питательным веществам, включая и минеральные. Стимулятор (второй фактор) был испытан в трех дозах: $B1$ – одинарная, $B2$ – двойная, $B3$ – тройная. Результативный признак – плодовитость самок, измерявшаяся числом детенышей в помете. Для каждого сочетания градаций рациона и стимулятора были подобраны три одновозрастные самки.

Комбинативная таблица двухфакторного равномерного дисперсионного комплекса с трехкратной повторностью ($ni = 3$) включает две градации по фактору A и три градации по фактору B (табл. 11). Варианты размещаются по

градациям, определяется объем градации, вычисляются суммы вариант, частные средние, затем вспомогательные величины (*H*1, *H*2, *H*3, *HA*, *HB*) и суммы квадратов отклонений (дисперсий) по рабочим формулам. В завершение всего заполняют таблицу дисперсионного анализа (табл. 12), находят показатель достоверности влияния Фишера и, сопоставляя его с табличным для соответствующих степеней свободы и принятого уровня значимости, делают статистический вывод.

Градации факторов		*A*1		*A*2				Для	*B*
		x	x^2	x	x^2	Σ	M_B	$\Sigma\Sigma x^2/n$	$\Sigma(\Sigma x^2/n)$
*B*1		5 6 7	25 36 49	1 4 1	1 16 1				
	Σx^2		110		18	$\Sigma\Sigma x^2 = 128$			
	Σx	18		6		$\Sigma\Sigma x = 24$	4	96	
	n	3		3		$n_{B1} = 6$			
	$\Sigma x^2/n$	108		12		$\Sigma(\Sigma x^2/n) = 120$			
*B*2		4 3 5	16 9 25	10 9 11	100 81 121				$\boldsymbol{H_B}$ =
	Σx^2		50		302	$\Sigma\Sigma x^2 = 352$			$\Sigma(\Sigma x^2/n)$
	Σx	12		30		$\Sigma\Sigma x = 42$	7	294	= 486
	n	3		3		$n_{B2} = 6$			
	$\Sigma x^2/n$	48		300		$\Sigma(\Sigma x^2/n) = 348$			
*B*3		2 3 1	4 9 1	7 4 7	49 16 49				
	Σx^2		14		114	$\Sigma\Sigma x^2 = 128$			
	Σx	6		18		$\Sigma\Sigma x = 24$	4	96	
	n	3		3		$n_{B3} = 6$			
	$\Sigma x^2/n$	12		108		$\Sigma(\Sigma x^2/n) = 120$			
ΣΣ	$\Sigma\Sigma x^2$		174		434	$\boldsymbol{H1} = \Sigma\Sigma\Sigma x^2 = 608$			
	$\Sigma\Sigma x$	36		54		$\Sigma\Sigma\Sigma x = 90$	$\boldsymbol{H2}$ =	$(\Sigma\Sigma\Sigma x)^2/N = 450$	
	$n_A = \Sigma n$	9		9		$N = \Sigma\Sigma n = 18$			
	$\Sigma x^2/n$	168		420		$\boldsymbol{H3} = \Sigma\Sigma(\Sigma x^2/n) = 588$			
Для А	$M_A = \Sigma\Sigma x/n$	2		6		j = 2 – число градаций фактора А k = 3 – число градаций фактора В			
	$\Sigma x^2/n$	144		324					
	$\boldsymbol{H_A} = \Sigma(\Sigma x^2/n) = 468$								

$C_{общ.} = H1 - H2 = 608 - 450 = 158$
$C_{случ.} = H1 - H3 = 608 - 588 = 20$
$C_{факт.} = C_{A+B+AB} = H3 - H2 = 588 - 450 = 138$
$C_A = H_A - H2 = 468 - 450 = 18$
$C_B = H_B - H2 = 486 - 450 = 36$
$C_{AB} = C_{факт.} - C_A - C_B = 138 - 18 - 36 = 84$

В нашем примере все факториальные влияния оказались достоверными с доверительной вероятностью $P > 0.95$ (табл. 12). Это позволяет сделать определенные выводы относительно действия стимулятора на плодовитость

самок. Влияние каждого фактора в отдельности (качества рациона и дозы стимулятора) и их суммарного эффекта достаточно существенно, но особенно результативно действие стимулятора в сочетании с полноценным рационом (величина $\eta 2AB$ выше, чем $\eta 2A$ и $\eta 2B$). Более того, при недостатке в корме минеральных веществ двукратные и трехкратные дозы стимулятора могут даже снизить плодовитость животных.

Составляющие дисперсии	Суммы квадратов, C	Сила влияния, η^2 (%)	Степени свободы, df	Дисперсии, S	Критерий, F ($F_{(\alpha,df1,df2)}$)
Фактор A	18	11	$j-1=1$	18	10.8 (4.7)
Фактор B	36	23	$k-1=2$	18	10.8 (3.9)
Взаимодействие AB	84	53	$df_A \cdot df_B = 2$	42	25.2 (3.9)
Факториальная (всего)	138	87	$j \cdot k - 1 = 5$	27.6	16.5 (3.1)
Случайная	20	13	$N - j \cdot k = 12$	1.67	
Общая	158	100	$N-1=17$		

Таблица двухфакторного дисперсионного анализа имеет ту же структуру, что и таблица для однофакторного анализа, только факториальная дисперсия разложена на три компоненты (для факторов A, B и их взаимодействия). Для каждой из них требуется вычислить число степеней свободы с учетом числа градаций фактора A (j, количество столбцов) и числа градаций фактора B (k, количество рядов), значения дисперсий, а также критерий Фишера. Поскольку каждому из расчетных значений критерия соответствует свое число степеней свободы, табличные значения окажутся разными.

Лекция № 6. Оценка зависимости между признаками

Рассмотренные выше методы статистического анализа дают возможность изучать изменчивость биологических объектов по отдельным признакам – весу, размерам, плодовитости, физиологическим показателям и др. Однако в ряде случаев важно знать, какова зависимость между вариацией двух или нескольких признаков, изменяются ли две переменные самостоятельно, независимо друг от друга, или варьирование одного признака в какой-то степени связано с изменчивостью другого. В качестве второй переменной часто выступает какой-либо фактор среды.

Задачу исследования зависимостей можно рассматривать как развитие метода дисперсионного анализа, решающего задачу сравнения нескольких выборок, т. е. изучающего влияния фактора на признак. Техника дисперсионного анализа имеет две особенности. Фактор (или факториальный признак) задан дискретно, в виде градаций, или «доз». Когда исследуется фактор, за-

данный *качественно*, то разбиение на градации всего диапазона его действия оказывается очень эффективным способом создания подобия количественной переменной. Но при изучении количественно заданного фактора в грубой градуальной схеме дисперсионного анализа утрачивается часть информации, которая содержится в исходных выборках и которую можно было бы использовать. Кроме этого, дисперсионный анализ явным образом не учитывает тенденции изменения среднего уровня признака при изменении уровня фактора, не содержит показателя характера (знака) зависимости признака от фактора. Все эти «недостатки» дисперсионного анализа не характерны для методов изучения *сопряженной изменчивости* – корреляционного и регрессионного анализов.

Способ представления отдельных наблюдений здесь меняется: каждая варианта рассматривается как носитель двух численных характеристик объекта измерения, двух *зависимых* значений случайной величины. Если выше мы отождествляли отдельное значение с отдельной вариантой, то теперь мы рассматриваем варианту как некоторое тело, обладающее минимум двумя зарегистрированными качествами, различными у разных вариант:

Например, для любого животного можно определить массу (*M*) и длину (*L*) тела; отдельная варианта будет нести два значения (*L, M*). При этом множество вариант выборки можно отобразить графически как точки на плоскости осей двух признаков *M* и *L*. Вся выборка предстанет в виде множества точек на плоскости (двумерное рассеяние). Как видно на диаграмме (рис. 6), «облако» вариант вытянуто в направлении диагонали облака точек. Справа вверху находятся варианты с высокими значениями и размеров, и массы тела, в левом нижнем углу – с наименьшими значениями. В центре расположены варианты с промежуточными, средними значениями.

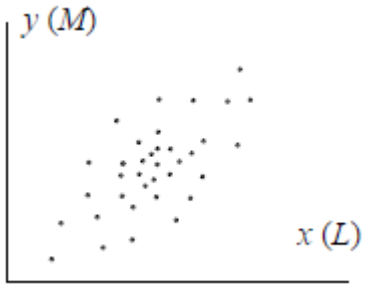

Рис. 6. *Область рассеяния вариант*

В первом приближении можно сказать, что *двумерное распределение – это ординация вариант на плоскости осей двух признаков*. Помимо рассея-

ния на плоскости, в определение двумерного распределения входит и частота встречаемости отдельных значений (*a*). Если признаки *x* и *y* теоретически подчиняются нормальному закону, тогда скопление вариант в трех осях (оси признаков *x*, *y* и частоты *a*) образует весьма странный «гребень», растянутое в пространстве *выпуклое нормальное распределение* (рис. 7). Однако в реальности такой идеальной картины получить никогда не удается, приходится ориентироваться только на плоскую фигуру рассеяния немногочисленных вариант. Если область, занятую вариантами, очертить по периферии плавной линией, мы получим вытянутую фигуру, эллипс, ограничивающий область рассеяния вариант, эллипс рассеяния. *Эллипс рассеяния – это область распространения вариант одной совокупности*.

Можно видеть, что в нашем примере признаки связаны друг с другом – есть общая тенденция: чем больше длина тела, тем больше вес; эта зависимость не очень жесткая, она размыта индивидуальными особенностями объектов (вариант).

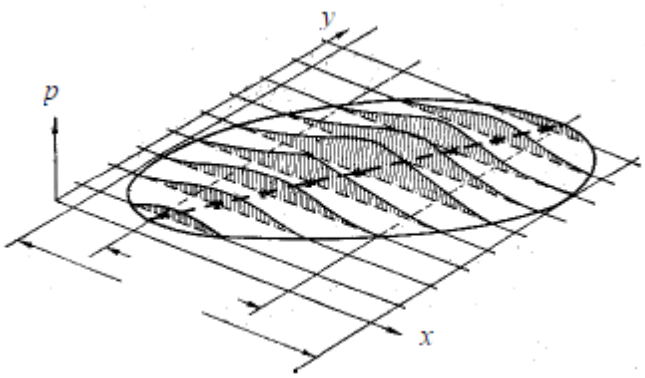

Рис. 7. *Двумерное распределение*

В двумерном распределении проявляются два эффекта: синхронное изменение двух признаков и размывание этой синхронности, т. е. действие факторов сопряжения признаков вдоль оси эллипса и действие случайных факторов – поперек нее.

Корреляционный анализ

Взаимная связь (взаимная зависимость) двух признаков при их изменчивости, т. е. сопряженность их вариации, называется корреляцией. Корреляция имеет место в тех случаях, когда признаки изменяются не автономно, а согласованно. Если с увеличением одного признака происходит соответствующее увеличение другого, говорят о положительной корреляции, и коэффициент корреляции имеет в этом случае положительный знак (+). Если же по мере увеличения первого признака второй уменьшается, то это отрицательная корреляция, коэффициент корреляции пишется со знаком минус (−).

Полная положительная корреляция выражается единицей $r = 1$, полная отрицательная – $r = -1$. В природе такая ситуация встречается редко, и степень связи выражается той или иной долей единицы. При этом о тесной (сильной) корреляции обычно говорят в тех случаях, когда коэффициент корреляции не ниже ±0.6; значения ниже ±0.6 указывают на среднюю связь, а ниже ±0.3 – на слабую.

Коэффициент корреляции призван численно выражать долю сопряженной вариации двух признаков в общей их вариации:

$$r = \sqrt{\frac{ковариация}{изменчивость}} = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{\sum (y - M_y)(x - M_x)}{\sqrt{\sum (y - M_y) \cdot \sum (x - M_x)}},$$

где Cxy – характеристика сопряженной изменчивости признаков,

Cx, Cy – характеристика общей изменчивости признаков.

При большом количестве данных коэффициент корреляции имеет смысл вычислять на компьютере (например, с помощью функции КОРРЕЛ в среде программы Excel), но для небольших выборок его можно быстро найти и при ручном счете. Рабочая формула для расчетов имеет вид:

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{\sum xy - (\sum x \cdot \sum y)/n}{\sqrt{(\sum x^2 - (\sum x)^2/n) \cdot (\sum y^2 - (\sum y)^2/n)}}.$$

Способ вычисления коэффициента корреляции показан в таблице на примере зависимости между живым весом коров (x) и их приплода (y, кг). По таблице рассчитываются квадраты вариант и их произведения, а также суммы вариант, квадратов и произведений. Вычисления ведутся по точным рабочим формулам.

i	y	x	y^2	x^2	$x \cdot y$
1	25	352	625	123904	8800
2	26	376	676	141376	9776
3	31	402	961	161604	12462
4	32	453	1024	205208	14496
5	34	484	1156	234256	16456
6	38	528	1444	278784	20064
7	38	555	1444	308025	21090
Σ	224	3150	7330	1453158	103144

Проведем последовательные расчеты. Сначала определим вспомогательные величины:

$$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y)/n = 103144 - 3150 \cdot 224/7 = 2344,$$
$$Cy = \Sigma y^2 - (\Sigma y)^2/n = 7330 - 224^2/7 = 162,$$
$$Cx = \Sigma x^2 - (\Sigma x)^2/n = 1453158 - 3150^2/7 = 35658;$$

затем – коэффициент корреляции:

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975.$$

Далее найдем его ошибку:

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.975^2}{7-2}} = 0.099,$$

и, наконец, критерий *t* Стьюдента для проверки значимости коэффициентов: $t_r = r / m_r = 0.975 / 0.099 = 9.84$.

Нулевая гипотеза предполагает отсутствие связи: «коэффициент корреляции значимо от нуля не отличается», $r = 0$. В нашем примере для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = n - 2 = 5$ находим табличное значение критерия Стьюдента $t(0.05, 5) = 2.57$. Полученная величина (9.84) значительно превышает табличную (2.57), что говорит о высокой статистической значимости коэффициента корреляции, о достоверности его отличия от нуля. Признаки *положительно* коррелируют, масса тела теленка действительно *возрастает* вслед за ростом массы тела коровы.

Выборный коэффициент корреляции в той или иной степени соответствует генеральному параметру. Определить диапазон, где лежит генеральное значение, можно с помощью доверительного интервала, хотя его *нельзя* построить непосредственно по формуле $r \pm t_{(\alpha,\, df)} \cdot m_r$. Дело в том, что область изменений коэффициента ограничена рамками ±1, поэтому распределение выборочных коэффициентов корреляции в общем не соответствует нормальному (с диапазоном изменчивости ± ∞). Поэтому перед расчетом коэффициент корреляции преобразуют в величину *z*, имеющую нормальное распределение, и уже для нее отыскивают границы доверительного интервала, после чего выполняют обратное преобразование.

Доверительный интервал для нашего случая ($r = 0.975$, $\alpha = 0.05$, $n = 7$, $df = n - 2 = 5$, $t_{(0.05,5)} = 2.57$) рассчитывается так. Преобразуем *r*:

$$z = 0.5 \cdot \ln\left(\frac{1+r}{1-r}\right) = 0.5 \cdot \ln\left(\frac{1+0.975}{1-0.975}\right) = 2.184$$

или берем его более точное значение из таблицы 13*П*, тогда $z = 2.0923$.

Определяем ошибку

$$m_z = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{7-3}} = 0.5.$$

Находим верхнюю границу: $_{\max}z = z + t_{(\alpha,df)} \cdot m_z = 2.09+2.57 \cdot 0.5 = 3.375$ и нижнюю границу: $_{\min}z = z - t(\alpha,df) \cdot m_z = 2.09-2.57 \cdot 0.5 = 0.805$. Обратное преобразование (по табл. 14*П*) дает: $_{\max}r \approx 1.00$, $_{\min}r \approx 0.67$. Истинное значение коэффициента корреляции находится в диапазоне от 0.67 до 1.00.

Множественная корреляция

Разобранные выше примеры корреляционных зависимостей касались главным образом взаимосвязи двух сопряженных процессов, явлений или варьирующих признаков. Между тем в практике биологических исследований нередко приходится сталкиваться с более сложными случаями, например, когда сопряжены не два, а три или более изменчивых фактора (признака). В такой ситуации возникает необходимость изучить множественные связи между большим числом взаимодействующих переменных, выступающих как в виде целой системы коррелированных признаков организма, так и в форме совместного влияния сложной совокупности факторов на определенное явление. Корреляционная зависимость нескольких переменных носит название множественной корреляции и оценивается коэффициентом, определяемым на основе корреляций между всеми парами признаков. Например, коэффициент множественной корреляции между тремя признаками *A*, *B* и *C* вычисляется по формуле:

$$r_{A.BC} = \sqrt{\frac{r_{AB}^2 + r_{AC}^2 - 2 \cdot r_{AB} \cdot r_{AC} \cdot r_{BC}}{1 - r_{AB}^2}}.$$

Полученная величина характеризует связь первого признака (*A*) с двумя другими (*B* и *C*). Покажем этот способ на примере совокупного действия двух факторов, *B* и *C* (температуры и влажности), на суточную активность травяных лягушек (*A*). Определение парных корреляций дало следующие результаты (n = 110): r_{AB} = +0.58; r_{AC} = +0.80; r_{BC} = −0.45. Отсюда

$$r_{A.BC} = \sqrt{\frac{0.58^2 + 0.8^2 - 2 \cdot 0.58 \cdot 0.8 \cdot 0.45}{1 - 0.45^2}} = 0.86.$$

Сводный коэффициент корреляции оказался довольно высоким и, как показывает его сопоставление со стандартным значением по таблице 15*П*, вполне достоверным (при $\alpha < 0.001$).

С другой стороны, если обнаружена значительная корреляция между признаками *A* и *C* и между *B* и *C*, то не исключена возможность мнимой корреляционной зависимости между *A* и *B*, которая создается за счет одновременного влияния на них третьего признака *C*. Например, установленная по исследованиям в Карелии корреляция между численностью лесных полевок и урожаем семян сосны, скорее всего, объясняется не значением последних в питании грызунов (т. е. прямой причинной связью), а тем, что оба эти явления (численность полевок и урожай семян) контролируются одними и теми же экологическими факторами (прежде всего метеорологическими) и поэтому изменяются параллельно, хотя непосредственно между собой не связаны.

В этом и подобных случаях (например, когда настоящие зависимости между признаками животных маскируются влиянием возраста или когда свя-

зи между отдельными промерами организма создаются за счет влияния живого веса и т. д.) возникает задача изучить корреляцию между двумя признаками (*A* и *B*), исключив влияние на эту связь третьего признака (*C*), как бы элиминировав его.

Регрессионный анализ

Коэффициент корреляции указывает лишь на степень (тесноту) связи в изменчивости двух переменных величин, но не позволяет судить о том, как меняется одна величина по мере изменения другой. Ответ на этот вопрос дает вычисление коэффициента регрессии, показывающего, на какую величину в среднем изменяется один признак при изменении другого на единицу измерения. Регрессионный анализ, в отличие от корреляционного, изучает эффект *влияния одного признака на другой*, зависимость признака от фактора, характер влияния фактора на признак. Его основные результаты таковы:

1. Таблица дисперсионного анализа, в которой показаны сила и достоверность влияния на признак изучаемого фактора или другого признака.

2. Уравнение регрессии, выражающее пропорциональность сопряженного изменения признаков, тенденции их взаимосвязанной изменчивости или динамики.

3. Оценки значимости коэффициентов уравнения регрессии.

Регрессионный анализ методически ориентирован односторонне – на изучение зависимости одного признака от другого (зависимость *y* от *x* или, напротив, зависимость *x* от *y*), хотя может применяться к случаям, когда фактически имеется взаимозависимость двух переменных.

Основную тенденцию взаимосвязанного изменения двух признаков можно отобразить с помощью простого графического приема. Разобьем ось *x* на несколько интервалов. Найдем для каждого из них частные средние значения признака *y* (*My*). Теперь проведем через эти средние точки ломаную линию. Это будет линия регрессии *Y* по *x*. *Регрессия – изменение среднего уровня одного признака при изменении другого* (рис. 8).

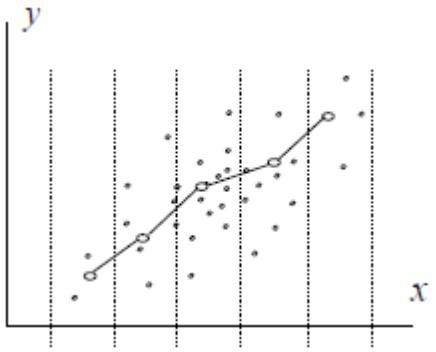

Рис. 8. *Эмпирическая линия регрессии*

Линейная регрессия

К сожалению, ход ломаной линии нельзя передать простым уравнением, к тому же на нем сказываются способ интервального разбиения оси абсцисс, а также уровень репрезентативности в разных областях распределения.

В этом смысле предпочтительнее единственная прямая линия регрессии, подчеркивающая основные тенденции зависимости признаков, которая может быть выражена простым уравнением линии: $y = ax + b$.

Судить о том, как меняется одна величина по мере изменения другой, позволяет коэффициент регрессии (a), показывающий, на какую величину в среднем изменяется один признак (y) при изменении другого (x) на единицу измерения (точнее, на какую величину один признак отклоняется от своей средней при некотором отклонении другого признака от своей средней):

$y - My = a \cdot (x - Mx)$.

Простые преобразования:

$y = a \cdot x + My - a \cdot Mx$, $b = My - a \cdot Mx$

и приводят к уравнению линии: $y = ax + b$.

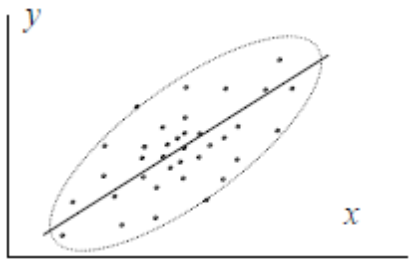

Рис. 9. *Линейная регрессия*

Рассчитать коэффициенты уравнения регрессии позволяет *метод наименьших квадратов*, основная идея которого состоит в том, чтобы линия регрессии прошла на наименьшем удалении от каждой точки, т. е. чтобы сумма квадратов расстояний от всех точек до прямой линии была наименьшей. В математической статистике показано, что для случая двумерного нормального распределения лучшей (эффективной, несмещенной и пр.) линией, описывающей зависимость одного признака от другого, может быть только линия частных средних арифметических.

Вычисления коэффициентов линейной регрессии $y = ax + b$ ведутся по следующему алгоритму. Сначала найдем вспомогательные величины:

$$Cx = \Sigma x^2 - (\Sigma x)^2 / n,$$
$$Cy = \Sigma y^2 - (\Sigma y)^2 / n,$$
$$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n,$$
$$M_y = \Sigma y / n,\ M_x = \Sigma x / n.$$

Затем рассчитаем коэффициенты: $a = Cxy / Cx$, $b = My - a \cdot Mx$.

Оценить значимость коэффициента регрессии позволяет критерий *t* Стьюдента, проверяющий нулевую гипотезу Но: $a = 0$, коэффициент регрессии значимо от нуля не отличается. С этой целью рассчитывается ошибка коэффициента регрессии m_a:

$$m_a = \frac{S_y}{S_x} \cdot m_r,$$

где m_r – ошибка коэффициента корреляции, и вычисляется значение критерия:

$t = (a - 0) / m_a = a / m_a \sim t(0.05, n - 2)$.

Смысл этого критерия состоит в следующем. Коэффициент регрессии *a* характеризует сопряженность пропорционального изменения двух признаков, т. е. отвечает за то, что линия регрессии имеет некоторый угол относительно оси абсцисс. Значение $a = 0$ означает, что линия регрессии идет параллельно оси OX, что при изменении признака *x* признак *y* не меняется, т. е. что *y* не зависит от *x*. Значения коэффициента, отличные от нуля, говорят о том, что взаимосвязь признаков имеет место, при $a > 0$ зависимость положительная, при $a < 0$ – отрицательная.

Вернемся к примеру с описанием зависимости между живым весом коров и их приплода. Расчеты для построения уравнения регрессии показаны в таблице. Сначала вычисляются квадраты вариант и их произведения, а также суммы вариант, квадратов и произведений. Вычисления ведутся по точным рабочим формулам. Проще всего это делать в среде Excel, с помощью команды Сервис \ Анализ данных \ Регрессия.

i	y	x	y^2	x^2	$x \cdot y$	Y	$(y-Y_i)^2$	$t \cdot m_Y$	min$Y$	max$Y$
1	25	352	625	123904	8800	25.6	0.31	2.0	23.6	27.5
2	26	376	676	141376	9776	27.1	1.29	1.7	25.5	28.8
3	31	402	961	161604	12462	28.8	4.65	1.4	27.4	30.2
4	32	453	1024	205208	14496	32.2	0.04	1.2	31.0	33.4
5	34	484	1156	234256	16456	34.2	0.06	1.3	32.9	35.5
6	38	528	1444	278784	20064	37.1	0.76	1.7	35.4	38.9
7	38	555	1444	308025	21090	38.9	0.81	2.1	36.8	41.0
Σ	224	3150	7330	1453158	103144		7.92			

Проведем последовательные расчеты вручную. Сначала определим вспомогательные величины:

$n = 7$,

$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n = 103144 - 3150 \cdot 224 / 7 = 2344$,

$Cy = \Sigma y2 - (\Sigma y)2 / n = 7330 - 2242 / 7 = 162$,

$Cx = \Sigma x2 - (\Sigma x)2 / n = 1453158 - 31502 / 7 = 35658$,

затем – параметры:

$$M_y = \Sigma y / n = 224 / 7 = 32,$$
$$M_x = \Sigma x / n = 3150 / 7 = 450,$$
$$S_y = \sqrt{\frac{Cy}{n-1}} = \sqrt{\frac{162}{6}} = 5.2,$$
$$S_x = \sqrt{\frac{Cx}{n-1}} = \sqrt{\frac{35658}{6}} = 77.1,$$
$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975,$$
$$a = \frac{Cxy}{Cx} = \frac{2344}{35658} = 0.0657,$$
$$b = M_y - a \cdot M_x = 32 - 0.0657 \cdot 450 = 2.419.$$

Получено уравнение линейной регрессии $Y = 0.0657x + 2.419$, которое позволяет рассчитать теоретические значения Y (табл., графа 7).

Далее найдем ошибку коэффициента регрессии:

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.975^2}{7-2}} = 0.099,$$

$$m_a = \frac{S_y}{S_x} \cdot m_r = \frac{5.2}{77.1} \cdot 0.099 = 0.00667,$$

и, наконец, критерий t Стьюдента для *проверки значимости коэффициента регрессии*: $ta = a / ma = 0.0657 / 0.00667 = 9.84$.

Для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = n - 2 = 5$ находим табличное значение критерия Стьюдента $t(0.05,5) = 2.57$. Полученная величина (9.84) превышает табличную (2.57), что говорит о статистической значимости коэффициента регрессии (a), о достоверности его отличия от нуля. Масса тела теленка действительно возрастает вслед за ростом массы тела коровы.

Рассчитаем *доверительную зону* (интервал), *в которой с той или иной вероятностью заключены теоретические средние значения* веса новорожденных. Критерий Стьюдента (нормированное отклонение) для уровня значимости $\alpha = 0.05$, и числа степеней свободы $df = n - 1 = 6$ составит 2.45. Далее находим границы. Так, для значения $x = 352$ кг прогноз по уравнению регрессии равен $Y = 25.56$, а возможное отклонение средней составит:

$$t \cdot m_Y = t \cdot m_y \cdot \sqrt{\frac{1}{n} + \frac{(x_i - M_x)^2}{Cx}} = 2.45 \cdot 1.2582 \cdot \sqrt{\frac{1}{7} + \frac{(352-450)^2}{35658}} =$$
$$= 2.45 \cdot 0.81 = 1.98.$$

Отсюда находим границу доверительного интервала (табл.):

верхнюю: $\max Y = Yi + t \cdot m_Y = 25.56 + 1.98 = 27.54$

и нижнюю: $\min Y = Yi - t \cdot m_Y = 25.56 - 1.98 = 23.58$.

Средняя масса новорожденного теленка для коров весом 352 кг с вероятностью $P = 0.95$ должна находиться в диапазоне от 23.6 до 27.5 кг (рис. 10).

Регрессионный анализ позволяет проверить *значимость* и второго коэффициента уравнения регрессии, *свободного члена* b. Математический смысл свободного члена уравнения линии состоит в том, что этому значению равна функция (y) при условии, что аргумент равен нулю ($x = 0$):

$y = ax + b = a \cdot 0 + b = b.$

В рамках регрессионного анализа рассматривается именно эта гипотеза. Но: $b = 0$, т. е. что линия регрессии проходит через начало осей координат, точку пересечения осей координат, через нуль. Если гипотеза опровергается, значит, линия регрессии не пересекает ось ординат. Если гипотеза не опровергается, мы можем считать, что между признаками существует простая пропорция ($Y = ax$) и расчет коэффициента регрессии a упрощается: $a = \Sigma(x \cdot y) / \Sigma x2$. Нулевая гипотеза Но: $b = 0$ проверяется по критерию Стьюдента: $t = (b - 0) / m_b = b / m_b \sim t_{(0.05,\, n-2)}$, где m_b – ошибка коэффициента b.

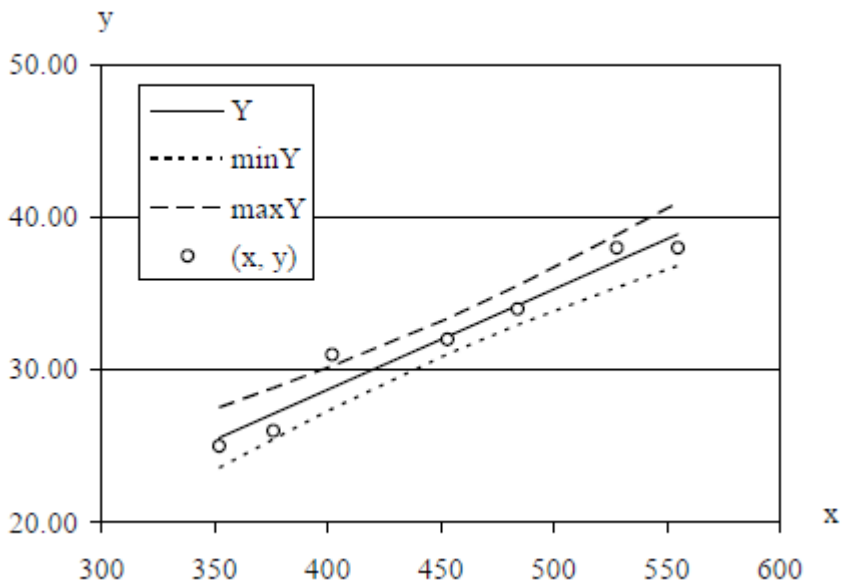

Рис. 10. *Линия регрессии $Y = 0.0657 \cdot x + 2.1347$ и ее доверительный интервал*

Ошибка второго коэффициента регрессии рассчитывается в два этапа.

Сначала находим общую ошибку регрессионной средней (или остаточное стандартное отклонение), которая может вычисляться по-разному.

Точная формула *для небольших выборок* дает величину:

$$m_y = S_y \cdot \sqrt{\frac{(n-1)\cdot(1-r^2)}{n-2}} = 5.2 \cdot \sqrt{\frac{(7-1)\cdot(1-0.975^2)}{7-2}} = 1.2582.$$

Общая точная формула показывает практически такой же результат:

$$m_y = \sqrt{\frac{\sum^{n}(y_i - Y_i)^2}{n-2}} = \sqrt{\frac{C_{остат.}}{n-2}} = \sqrt{S^2_{остат.}} = \sqrt{\frac{7.92}{5}} = \sqrt{1.5832} = 1.2582$$

(величина $C_{остат.}$ – это сумма квадратов разности между расчетными и реальными значениями признака, она найдена в табл., внизу 7 графы, $C_{остат.}$= 7.92).

Теперь вычисляем ошибку коэффициента *b*:

$$m_b = m_y \cdot \sqrt{\frac{1}{n} + \left(\frac{M_x}{Cx}\right)^2} = 1.2582 \cdot \sqrt{\frac{1}{7} + \left(\frac{450}{35658}\right)^2} = 3.0359$$

и критерий *t* Стьюдента: *tb* = *b* / *mb* = 2.419 / 3.0359 = 0.797.

Для уровня значимости α = 0.05 и числа степеней свободы *df* = *n* − 2 = 5 табличное значение составляет *t*(0.05, 5) = 2.57. Анализ показал, что критерий Стьюдента для свободного члена уравнения (0.797) оказался ниже табличного значения (2.57), т. е. коэффициент *b* значимо от нуля не отличается (при данном объеме собранных материалов). Это позволяет пересчитать коэффициент регрессии: *a* = Σ(*x* ⋅ *y*) / Σ*x*2 = 0.071. Теперь можно пользоваться уравнением регрессии вида: *Y* = 0.071 ⋅ *x*.

Оценить *достоверности взаимодействия признаков* можно и с помощью дисперсионного анализа (табл. 1). В этом случае общая дисперсия зависимого признака *y* ($C_{общ.}$) разлагается на две составляющие – регрессионную дисперсию (изменчивость признака *y*, связанная с влиянием признака *x* ($C_{регр.}$), и случайную, или остаточную, дисперсию (изменчивость признака *y*, связанная с влиянием неучтенных случайных факторов ($C_{остат.}$) (рис. 11, табл. 1, 2).

Общую сумму квадратов ($C_{общ.} = C_y = \Sigma(y_i - M_y)^2 = \Sigma y_i{}^2 - (\Sigma y_i)^2 / n$) находят непосредственно как сумму квадратов отличий между значением y_i для каждой варианты и общей средней признака *y*. Остаточную сумму квадратов ($Cостат. = \Sigma(y_i - Y_i)^2$) находят также непосредственно как сумму квадратов отличий между значением y_i для каждой варианты и значением, предварительно рассчитанным по уравнению регрессии $Y_i = ax_i + b$ (для соответствующих значений x_i). Модельную сумму квадратов ($C_{мод.} = \Sigma(Y_i - M_y)^2$) рассчитывают как разность между общей и остаточной ($C_{мод.} = C_{общ.} - C_{остат.}$).

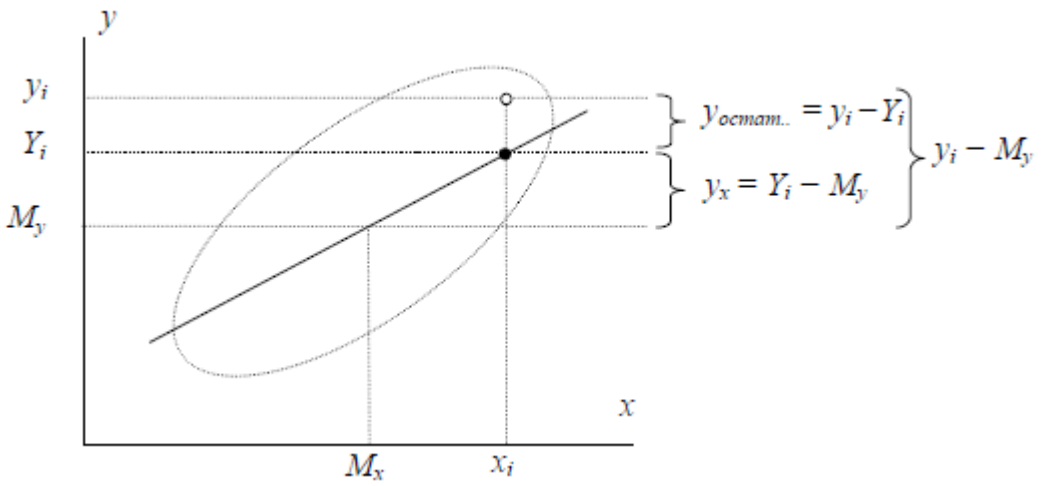

Рис. 11. *Модель варианты в регрессионном анализе*

Таблица 1

Составляющие дисперсии	Суммы квадратов, C	Формулы расчета сумм квадратов	df	S^2	F
Регрессия	$C_{регр.} = \Sigma(Y_i - M_y)^2$	$C_{общ.} - C_{остат.}$	1	$S^2_{регр.} = \frac{C_{регр.}}{df_{регр.}}$	$\frac{S^2_{регр.}}{S^2_{остат.}}$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma(y_i - Y_i)^2$		$n - 2$	$S^2_{остат.} = \frac{C_{остат.}}{df_{остат.}}$	$F_{(0.05,\,1,\,n-2)}$
Общая (всего)	$C_{общ.} = \Sigma(y_i - M_y)^2$	$(\Sigma y_i^2 - \Sigma y_i)^2 / n = C_y$			

Таблица 2

Составляющие дисперсии	C		df	S^2	F
Регрессия	$C_{регр.} = \Sigma(Y_i - Y)^2$	154.08	1	$S^2_{регр.} = 154.08$	$F = \frac{154.08}{1.58} = 97.3$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma(y_i - Y_{xi})^2$	7.92	5	$S^2_{остат.} = 1.58$	$F_{(0.05,\,1,\,5)} = 6.6$
Общая (всего)	$C_{общ.} = \Sigma(y_i - Y)^2$	162			

Показателем «силы влияния признака на признак» служит *коэффициент детерминации, отношение регрессионной суммы квадратов к общей сумме квадратов* (принимает значения от 0 до 1):

$$R^2 = \frac{C_{мод.}}{C_{общ.}} = \frac{154.08}{162} = 0.95.$$

Между коэффициентом детерминации и коэффициентом корреляции существует простое соответствие: $r = R = 0.95 = 0.975$.

Построив таблицу дисперсионного анализа с помощью критерия Фишера, можно проверить нулевую гипотезу Но: предсказания регрессионной модели в целом неадекватно описывают исходные данные, зависимости между признаками нет. Конструкция критерия исследует вопрос, превышает ли варьирование, учтенное моделью, случайное (остаточное) варьирование?

Критерий Фишера вычисляется как отношение модельной и остаточной дисперсии:

$$F = S^2_{мод.} / S^2_{остат.} = 154.08 / 1.58 = 97.3.$$

Табличное значение $F(0.05, 1, 5) = 6.6$. Поскольку полученное значение критерия оказалось выше табличного, дисперсия реального признака y приближается по величине к дисперсии расчетных значений признака Y, т. е. существенно превышает (случайные) отличия между ними. Регрессионная модель в целом адекватно описывает исходные данные.