МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Кафедра зоологии, физиологии и генетики

БИОМЕТРИЯ

Гомель 2015

СОДЕРЖАНИЕ

Лекция 1. ВВЕДЕНИЕ В КУРС. ДАННЫЕ В БИОЛОГИИ

1.1 Содержание науки. Биометрия − раздел вариационной статистики, с помощью методов которого производят обработку экспериментальных данных и наблюдений, а также планирование количественных экспериментов в биологических исследованиях; а также научная отрасль, связанная с разработкой и использованием статистических методов в научных исследованиях в медицине, здравоохранении, и эпидемиологии.

1.2 История. Биометрия сложилась в XIX веке, главным образом, благодаря трудам Ф. Гальтона и К. Пирсона. В 1920-30-х годах крупный вклад в развитие биометрии внес Р. Фишер.У истоков биометрии стоял Фрэнсис Гальтон (1822−1911). Первоначально Гальтон готовился стать врачом. Однако, обучаясь в Кембриджском университете, он увлекся естествознанием, метеорологией, антропологией, наследственностью и теорией эволюции. В его книге, посвященной природной наследственности, изданной в 1889 году им впервые было введено в употребление слово «biometry» и в это же время он разработал основы корреляционного анализа. Гальтон заложил основы новой науки и дал ей имя. Однако превратил её в стройную научную дисциплину математик Карл Пирсон (1857−1936). В 1884 году Пирсон получает кафедру прикладной математики в Лондонском университете, а в 1889 году знакомится с Гальтоном и его работами. Большую роль в жизни Пирсона сыграл зоолог Ф.Велдон. Помогая ему в анализе реальных зоологических данных, Пирсон ввел в 1893 г. понятие среднего квадратического отклонения и коэффициента вариации. Пытаясь математически оформить теорию наследственности Гальтона, Пирсон в 1898 г. разрабатывает основы множественной регрессии. В 1903 г. Пирсон разработал основы теории сопряженности признаков, а в 1905 г. опубликовал основы нелинейной корреляции и регрессии.

Следующий этап развития биометрии связан с именем великого английского статистика Рональда Фишера (1890−1962). Во время обучения в Кембриджском университете Фишер знакомится с трудами Менделя и Пирсона. В 1913−1915 годах Фишер работает статистиком на одном из предприятий, а в 1915−1919 годах преподает физику и математику в средней школе. С 1919 года Фишер начинает работу статистиком на опытной сельскохозяйственной станции в Ротамстеде, где он проработал до 1933 года. Затем с 1933 года по 1943 год Фишер работает профессором в Лондонском университете, а с 1943 года по 1957 год заведует кафедрой генетики в Кембридже. За эти годы им были разработаны теория выборочных распределений, методы дисперсионного и дискриминантного анализа, теории планирования экспериментов, метод максимального правдоподобия и многое другое, что составляет основу современной прикладной статистики и математической генетики.

1.3 Развитие представлений о статистике. Статистика − отрасль знаний, в которой излагаются общие вопросы сбора, измерения и анализа массовых статистических (количественных или качественных) данных.

Слово «статистика» происходит от латинского *status* − состояние дел[1]. В науку термин «статистика» ввел немецкий ученый Готфрид Ахенваль в 1746 году, предложив заменить название курса «Государствоведение», преподававшегося в университетах Германии, на «Статистику», положив тем самым начало развитию статистики как науки и учебной дисциплины. Несмотря на это, статистический учет вёлся намного раньше: проводились переписи населения в Древнем Китае, осуществлялось сравнение военного потенциала государств, велся учет имущества граждан в Древнем Риме и т. п.

Статистика разрабатывает специальную методологию исследования и обработки материалов: массовые статистические наблюдения, метод группировок, средних величин, индексов, балансовый метод, метод графических изображений и другие методы анализа статистических данных. Начало статистической практики относится примерно к времени возникновения государства. Первой опубликованной статистической информацией можно считать глиняные таблички Шумерского царства (III − II тысячелетия до н. э.).

Вначале под статистикой понимали описание экономического и политического состояния государства или его части. Например, к 1792 г. относится определение: «статистика описывает состояние государства в настоящее время или в некоторый известный момент в прошлом». И в настоящее время деятельность государственных статистических служб вполне укладывается в это определение.

Однако постепенно термин «статистика» стал использоваться более широко. По Наполеону Бонапарту, «статистика − это бюджет вещей». Тем самым статистические методы были признаны полезными не только для административного управления, но и для применения на уровне отдельного предприятия. Согласно формулировке 1833 г., «цель статистики заключается в представлении фактов в наиболее сжатой форме». Во 2-й половине XIX − начале XX веков сформировалась научная дисциплина −математическая статистика, являющаяся частью математики.

В XX веке статистику часто рассматривают, прежде всего, как самостоятельную научную дисциплину. Статистика есть совокупность методов и принципов, согласно которым проводится сбор, анализ, сравнение, представление и интерпретация числовых данных. В 1954 г. академик АН Борис Владимирович Гнеденко дал следующее определение: «Статистика состоит из трёх разделов:

1 сбор статистических сведений, то есть сведений, характеризующих отдельные единицы каких-либо массовых совокупностей;

2 статистическое исследование полученных данных, заключающееся в выяснении тех закономерностей, которые могут быть установлены на основе данных массового наблюдения;

3 разработка приёмов статистического наблюдения и анализа статистических данных. Последний раздел, собственно, и составляет содержание математической статистики».

Термин «статистика» употребляют ещё в двух смыслах. Во-первых, в обиходе под «статистикой» часто понимают набор количественных данных о

каком-либо явлении или процессе. Во-вторых, статистикой называют функцию от результатов наблюдений, используемую для оценки характеристик и параметров распределений и проверки гипотез.

1.4 Краткая история статистических методов. Типовые примеры раннего этапа применения статистических методов описаны в Библии, в Ветхом Завете. Там, в частности, приводится число воинов в различных племенах. С математической точки зрения дело сводилось к подсчёту числа попаданий значений наблюдаемых признаков в определённые градации.

Сразу после возникновения теории вероятностей (Паскаль, Ферма, XVII век) вероятностные модели стали использоваться при обработке статистических данных. Например, изучалась частота рождения мальчиков и девочек, было установлено отличие вероятности рождения мальчика от 0.5, анализировались причины того, что в парижских приютах эта вероятность не та, что в самом Париже, и т. д.

В 1794 г. (по другим данным − в 1795 г.) немецкий математик Карл Гаусс формализовал один из методов современной математической статистики − метод наименьших квадратов. В XIX веке заметный вклад в развитие практической статистики внёс бельгиец Кетле, на основе анализа большого числа реальных данных показавший устойчивость относительных статистических показателей, таких, как доля самоубийств среди всех смертей.

Первая треть XX века прошла под знаком параметрической статистики. Изучались методы, основанные на анализе данных из параметрических семейств распределений, описываемых кривыми семейства Пирсона. Наиболее популярным было нормальное распределение. Для проверки гипотез использовались критерии Пирсона, Стьюдента, Фишера. Были предложены метод максимального правдоподобия, дисперсионный анализ, сформулированы основные идеи планирования эксперимента.

Разработанную в первой трети XX века теорию анализа данных называют параметрической статистикой, поскольку её основной объект изучения − это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами. Как правило, нельзя указать каких-либо веских причин, по которым распределение результатов конкретных наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог, видимо, приближается логарифмически нормальным распределением и так далее.

Статистические методы −методы анализа статистических данных. Выделяют методы прикладной статистики, которые могут применяться во всех областях научных исследований и любых отраслях народного хозяйства, и другие статистические методы, применимость которых ограничена той или иной сферой. Имеются в виду такие методы, как статистический приемочный

контроль, статистическое регулирование технологических процессов, надежность и испытания, планирование экспериментов.

Статистические методы анализа данных применяются практически во всех областях деятельности человека. Их используют всегда, когда необходимо получить и обосновать какие-либо суждения о группе (объектов или субъектов) с некоторой внутренней неоднородностью.

Целесообразно выделить три вида научной и прикладной деятельности в области статистических методов анализа данных (по степени специфичности методов, сопряженной с погруженностью в конкретные проблемы):

а) разработка и исследование методов общего назначения, без учета специфики области применения;

б) разработка и исследование статистических моделей реальных явлений и процессов в соответствии с потребностями той или иной области деятельности;

в) применение статистических методов и моделей для статистического анализа конкретных данных.

Прикладная статистика − это наука о том, как обрабатывать данные произвольной природы. Математической основой прикладной статистики и статистических методов анализа является теория вероятностей и математическая статистика.

Описание вида данных и механизма их порождения − начало любого статистического исследования. Для описания данных применяют как детерминированные, так и вероятностные методы. С помощью детерминированных методов можно проанализировать только те данные, которые имеются в распоряжении исследователя. Например, с их помощью получены таблицы, рассчитанные органами официальной государственной статистики на основе представленных предприятиями и организациями статистических отчетов. Перенести полученные результаты на более широкую совокупность, использовать их для предсказания и управления можно лишь на основе вероятностно-статистического моделирования. Поэтому в математическую статистику часто включают лишь методы, опирающиеся на теорию вероятностей.

В простейшей ситуации статистические данные − это значения некоторого признака, свойственного изучаемым объектам. Значения могут быть количественными или представлять собой указание на категорию, к которой можно отнести объект. Во втором случае говорят о качественном признаке.

При измерении по нескольким количественным или качественным признакам в качестве статистических данных об объекте получаем вектор. Его можно рассматривать как новый вид данных. В таком случае выборка состоит из набора векторов. Есть часть координат − числа, а часть − качественные (категоризованные) данные, то говорим о векторе разнотипных данных.

Одним элементом выборки, то есть одним измерением, может быть и функция в целом. Например, описывающая динамику показателя, то есть его изменение во времени, − электрокардиограмма больного или амплитуда биений

вала двигателя. Или временной ряд, описывающий динамику показателей определенной фирмы. Тогда выборка состоит из набора функций.

Элементами выборки могут быть и иные математические объекты. Например, бинарные отношения. Так, при опросах экспертов часто используют упорядочения (ранжировки) объектов экспертизы − образцов продукции, инвестиционных проектов, вариантов управленческих решений. В зависимости от регламента экспертного исследования элементами выборки могут быть различные виды бинарных отношений (упорядочения, разбиения, толерантности), множества, нечёткие множества и т. д.

Итак, математическая природа элементов выборки в различных задачах прикладной статистики может быть самой разной. Однако можно выделить два класса статистических данных − числовые и нечисловые. Соответственно прикладная статистика разбивается на две части − числовую статистику и нечисловую статистику.

Числовые статистические данные − это числа, вектора, функции. Их можно складывать, умножать на коэффициенты. Поэтому в числовой статистике большое значение имеют разнообразные суммы. Математический аппарат анализа сумм случайных элементов выборки − это (классические) законы больших чисел и центральные предельные теоремы.

Нечисловые статистические данные − это категоризованные данные, вектора разнотипных признаков, бинарные отношения, множества, нечеткие множества и др. Их нельзя складывать и умножать на коэффициенты. Поэтому не имеет смысла говорить о суммах нечисловых статистических данных. Они являются элементами нечисловых математических пространств (множеств). Математический аппарат анализа нечисловых статистических данных основан на использовании расстояний между элементами (а также мер близости, показателей различия) в таких пространствах. С помощью расстояний определяются эмпирические и теоретические средние, доказываются законы больших чисел, строятся непараметрические оценки плотности распределения вероятностей, решаются задачи диагностики и кластерного анализа, и т. д.

В прикладных исследованиях используют статистические данные различных видов. Это связано, в частности, со способами их получения. Например, если испытания некоторых технических устройств продолжаются до определенного момента времени, то получаем т. н. цензурированные данные, состоящие из набора чисел − продолжительности работы ряда устройств до отказа, и информации о том, что остальные устройства продолжали работать в момент окончания испытания. Цензурированные данные часто используются при оценке и контроле надежности технических устройств.

Теория статистических методов нацелена на решение реальных задач. Поэтому в ней постоянно возникают новые постановки математических задач анализа статистических данных, развиваются и обосновываются новые методы. Обоснование часто проводится математическими средствами, то есть путем доказательства теорем. Большую роль играет методологическая составляющая − как именно ставить задачи, какие предположения принять с

целью дальнейшего математического изучения. Велика роль современных информационных технологий, в частности, компьютерного эксперимента.

Развитие вычислительной техники во второй половине XX века оказало значительное влияние на статистику. Ранее статистические модели были представлены преимущественно линейными моделями. Увеличение быстродействия ЭВМ и разработка соответствующих численных алгоритмов послужило причиной повышенного интереса к нелинейным моделям таким, как искусственные нейронные сети, и привело к разработке сложных статистических моделей, например обобщенная линейная модель и иерархическая модель.

1. 5 Статистическое наблюдение — это массовое (оно охватывает большое число случаев проявления исследуемого явления для получения правдивых статистических данных) планомерное (проводится по разработанному плану, включающему вопросы методологии, организации сбора и контроля достоверности информации), систематическое (проводится систематически, либо непрерывно, либо регулярно), научно организованное (для повышения достоверности данных, которая зависит от программы наблюдения, содержания анкет, качества подготовки инструкций) наблюдение за явлениями и процессами социально-экономической жизни, которое заключается в сборе и регистрации отдельных признаков у каждой единицы совокупности.

Этапы статистического наблюдения

1. **Подготовка к статистическому наблюдению** (решение научно-методических и организационно-технических вопросов).

- определение цели и объекта наблюдения;
- определение состава признаков подлежащих изучению;
- разработка документов для сбора данных;

2. **Сбор информации**

- непосредственное заполнение статистических формуляров (бланки, анкеты);
- применение стандартных методов сбора (пробная площадка, ловушки Геро и пр.)

Статистическая информация — это первичные данные о предмете изучения, формирующиеся в процессе статистического наблюдения, которые затем подвергаются систематизации, сводке, анализу и обобщению.

3. **Первичная обработка данных**

4. **Статистический анализ обработанной информации.**

5. **Разработка предложений и рекомендаций по совершенствованию статистического наблюдения**

- заключается в анализе причин, которые привели к неверному заполнению статистических формуляров и разработке соответствующих предложений по совершенствованию наблюдения.

В результате статистического наблюдения должна быть получена объективная, сопоставимая, полная информация, позволяющая на

последующих этапах исследования обеспечить научно-обоснованные выводы о характере и закономерностях развития изучаемого явления.

Виды статистического наблюдения

Статистические наблюдения подразделяются на виды по следующим признакам:

- по времени регистрации данных;
- по полноте охвата единиц совокупности;

Виды статистического наблюдения по времени регистрации:

Текущее (непрерывное) наблюдение - проводится для изучения текущих явлений и процессов. Регистрация фактов осуществляется по мере их свершения.

Прерывное наблюдение — проводится по мере необходимости, при этом допускаются временные разрывы в регистрации данных:

- **Периодическое** наблюдение — проводится через сравнительно равные интервалы времени.
- **Единовременное** наблюдение — осуществляется без соблюдения строгой периодичности его проведения.

По полноте охвата единиц совокупности различают следующие виды статистического наблюдения:

Сплошное наблюдение — представляет собой сбор и получение информации обо всех единицах изучаемой совокупности.

Несплошное наблюдение — основано на принципе случайного отбора единиц изучаемой совокупности, при этом в выборочной совокупности должны быть представлены все типы единиц, имеющихся в совокупности.

Несплошное наблюдение подразделяется на:

- **Выборочное наблюдение** - основано на случайном отборе единиц, которые подвергаются наблюдению.
- **Монографическое наблюдение** — заключается в обследовании отдельных единиц совокупности, характеризующихся редкими качественными свойствами.
- **Метод основного массива** — состоит в изучении самых существенных, наиболее крупных единиц совокупности, имеющих по основному признаку наибольший удельный вес в изучаемой совокупности.
- **Метод моментных наблюдений** — заключается в проведении наблюдений через случайные или постоянные интервалы времени с отметками о состоянии исследуемого объекта в тот или иной момент времени.

Способы статистического наблюдения

Непосредственное статистическое наблюдение — наблюдение, при котором сами регистраторы путем непосредственного замера, взвешивания, подсчета устанавливают факт подлежащий регистрации.

Документальное наблюдение — основано на использовании различного рода документов учетного характера.

Опрос - заключается в получении необходимой информации непосредственно от респондента.

Существуют следующие виды опроса:

Экспедиционный — регистраторы получают необходимую информацию от опрашиваемых лиц и сами фиксируют ее в формулярах.

Способ саморегистрации — формуляры заполняются самими респондентами, регистраторы только раздают бланки и объясняют правила их заполнения.

Корреспондентский — сведения в соответствующие органы сообщает штат добровольных корреспондентов.

Анкетный — сбор информации осуществляется в виде анкет, представляющих собой специальные вопросники, удобен в случаях, когда не требуется высокая точность результатов.

Явочный — заключается в предоставлении сведений в соответствующие органы в явочном порядке.

В зависимости от причин возникновения различают **ошибки регистрации и ошибки репрезентативности**. Ошибки регистрации характерны как для сплошного, так и для несплошного наблюдения, а ошибки репрезентативности — только для несплошного наблюдения. Ошибки регистрации, как и ошибки репрезентативности, могут быть **случайными и систематическими.**

Ошибки регистрации — представляют собой отклонения между значением показателя, полученного в ходе статистического наблюдения, и его фактическим значением. Ошибки регистрации бывают случайными (результат действий случайных факторов — перепутаны строки например) и систематическими (проявляются постоянно).

Ошибки репрезентативности — возникают, когда отобранная совокупность недостаточно точно воспроизводит исходную совокупность. Характерны для несплошного наблюдения и заключаются в отклонении величины показателя исследуемой части совокупности от его величины в генеральной совокупности.

Случайные ошибки — являются результатом действия случайных факторов.

Систематические ошибки — всегда имеют одинаковую направленность к увеличению или уменьшению показателя по каждой единице наблюдения, вследствие чего значение показателя по совокупности в целом будет включать накопленную ошибку.

Способы контроля:

- Счетный (арифметический) — проверка правильности арифметического расчета.
- Логический — основан на смысловой взаимосвязи между признаками.

Статистическая совокупность - множество единиц, обладающих массовостью, типичностью, качественной однородностью и наличием вариации.

Статистическая совокупность состоит из материально существующих объектов, является объектом статистического исследования.

Единица совокупности — каждая конкретная единица статистической совокупности.

Одна и та же статистическая совокупность может быть однородна по одному признаку и неоднородна по другому.

Качественная однородность — сходство всех единиц совокупности по какому-либо признаку и несходство по всем остальным.

В статистической совокупности отличия одной единицы совокупности от другой чаще имеют количественную природу. Количественные изменения значений признака разных единиц совокупности называются вариацией.

Вариация признака — количественное изменение признака (для количественного признака) при переходе от одной единицы совокупности к другой.

Признак - это свойство, характерная черта или иная особенность единиц, объектов и явлений, которая может быть наблюдаема или измерена. Признаки делятся на количественные и качественные. Многообразие и изменчивость величины признака у отдельных единиц совокупности называется **вариацией.**

Атрибутивные (качественные) признаки не поддаются числовому выражению (состав населения по полу). Количественные признаки имеют числовое выражение (состав населения по возрасту).

Показатель — это обобщающая количественно качественная характеристика какого-либо свойства единиц или совокупности в целом в конкретных условиях времени и места.

Система показателей — это совокупность показателей всесторонне отражающих изучаемое явление.

Например, изучается привес коров:

- Признак — вес
- Статистическая совокупность — коровы фермы
- Единица совокупности — каждая корова
- Качественная однородность — одного возраста
- Вариация признака — ряд цифр

1.6 Генеральная совокупность и выборка из нее

Основу статистического исследования составляет множество данных, полученных в результате измерения одного или нескольких признаков. Реально наблюдаемая совокупность объектов, статистически представленная рядом наблюдений $x_1, x_2, ..., x_n$ случайной величины X, является **выборкой**, а гипотетически существующая (домысливаемая) — **генеральной совокупностью**. Генеральная совокупность может быть конечной (число наблюдений **N = const**) или бесконечной (**N** = ∞), а выборка из генеральной совокупности — это всегда результат ограниченного ряда n наблюдений. Число наблюдений n, образующих выборку, называется **объемом выборки**. Если объем выборки n достаточно велик (**n** → ∞) выборка считается **большой**, в противном случае она называется выборкой **ограниченного объема**. Выборка считается **малой**, если при измерении одномерной случайной величины X объем выборки не превышает 30 (**n <= 30**), а при измерении одновременно нескольких (**k**) признаков в многомерном пространстве отношение **n** к **k** не превышает **10 (n/k < 10)**. Выборка образует **вариационный ряд**, если ее члены

являются **порядковыми статистиками**, т. е. выборочные значения случайной величины **X** упорядочены по возрастанию (ранжированы), значения же признака называются **вариантами.**

Основные способы организации выборки

Достоверность статистических выводов и содержательная интерпретация результатов зависит от **репрезентативности** выборки, т.е. полноты и адекватности представления свойств генеральной совокупности, по отношению к которой эту выборку можно считать представительной. Изучение статистических свойств совокупности можно организовать двумя способами: с помощью **сплошного** и **несплошного наблюдения.**

Сплошное наблюдение предусматривает обследование всех **единиц** изучаемой **совокупности**, а **несплошное (выборочное) наблюдение** — только его части.

Существуют пять основных способов организации выборочного наблюдения:

1. **простой случайный отбор**, при котором nобъектов случайно извлекаются из генеральной совокупности Nобъектов (например с помощью таблицы или датчика случайных чисел), причем каждая из возможных выборок имеют равную вероятность. Такие выборки называются **собственно-случайными;**

2. **простой отбор с помощью регулярной процедуры** осуществляется с помощью механической составляющей (например, даты, дня недели, номера квартиры, буквы алфавита и др.) и полученные таким способом выборки называются **механическими;**

3. **стратифицированный** отбор заключается в том, что генеральная совокупность объема Nподразделяется на подсовокупности или слои (страты) объема $N_1,N_2,...,N_r$так что $N_1+N_2+...+N_r=N$. Страты представляют собой однородные объекты с точки зрения статистических характеристик (например, население делится на страты по возрастным группам или социальной принадлежности; предприятия — по отраслям). В этом случае выборки называются **стратифицированными** (иначе, **расслоенными, типическими, районированными**);

4. методы **серийного** отбора используются для формирования **серийных** или **гнездовых выборок**. Они удобны в том случае, если необходимо обследовать сразу "блок" или серию объектов (например, партию товара, продукцию определенной серии или население при территориально-административном делении страны). Отбор серий можно осуществить собственно-случайным или механическим способом. При этом проводится сплошное обследование определенной партии товара, или целой территориальной единицы (жилого дома или квартала);

5. **комбинированный** (ступенчатый) отбор может сочетать в себе сразу несколько способов отбора (например, стратифицированный и случайный или случайный и механический); такая выборка называется **комбинированной.**

Виды отбора

По **виду** различаются индивидуальный, групповой и комбинированный отбор. При **индивидуальном отборе** в выборочную совокупность отбираются отдельные единицы генеральной совокупности, при **групповом отборе** — качественно однородные группы (серии) единиц, а **комбинированный отбор** предполагает сочетание первого и второго видов.

По **методу** отбора различают **повторную и бесповторную** выборку.

Бесповторным называется отбор, при котором попавшая в выборку единица не возвращается в исходную совокупность и в дальнейшем выборе не участвует; при этом численность единиц генеральной совокупности **N** сокращается в процессе отбора.

При **повторном** отборе **попавшая** в выборку единица после регистрации возвращается в генеральную совокупность и таким образом сохраняет равную возможность наряду с другими единицами быть использованной в дальнейшей процедуре отбора; при этом численность единиц генеральной совокупности **N** остается неизменной (метод в социально-экономических исследованиях применяется редко). Однако, при большом **N (N → ∞)** формулы для **бесповторного** отбора приближаются к аналогичным для **повторного** отбора и практически чаще используются последние (**N = const**).

По своей природе распределения бывают **непрерывными** и **дискретными**. Наиболее известным непрерывным распределением является **нормальное**.

В зависимости от вида распределения и от способа отбора единиц совокупности по-разному вычисляются характеристики параметров распределения: теоретическое и эмпирическое распределения.

Долей выборки k_n называется отношение числа единиц выборочной совокупности к числу единиц генеральной совокупности:

$$k_n = n/N.$$

Выборочная доля w — это отношение единиц, обладающих изучаемым признаком **x** к объему выборки **n**:

$$w = n_n/n.$$

Пример. В партии товара, содержащей 1000 ед., при 5% выборке **доля выборки k_n** в абсолютной величине составляет 50 ед. (n = N*0,05); если же в этой выборке обнаружено 2 бракованных изделия, то **выборочная доля брака w** составит 0,04 (w = 2/50 = 0,04 или 4%).

Так как выборочная совокупность отлична от генеральной, то возникают **ошибки выборки**.

1.7 Шкалы измерений

Состояние объекта оценивается по критериям. В качестве критериев могут выступать: выживаемость животных, степень интоксикации, сохранение жизненно важных функций и т.д.

Оценки измеряются в той или иной шкале. *Шкала* (условно говоря, шкала – это множество возможных значений оценок по критериям) – числовая система, в которой отношения между различными свойствами изучаемых явлений, процессов переведены в свойства того или иного множества, как правило – множества чисел.

Различают несколько ***типов шкал:***

Во-первых, можно выделить ***дискретные шкалы*** (в которых множество возможных значений оцениваемой величины конечно – например, оценка в баллах – «1», «2», «3», «4», «5») и ***непрерывные шкалы*** (например, концентрация вещества в моль/л или активность фермента в сыворотке крови в мКат/л).

Во-вторых, выделяют ***шкалы отношений, интервальные шкалы, порядковые (ранговые) шкалы*** **и** ***номинальные шкалы*** (шкалы наименований).

Шкала отношений – самая мощная шкала. Она позволяет оценивать, во сколько раз один измеряемый объект больше (меньше) другого объекта, принимаемого за эталон, единицу. Для шкал отношений существует естественное начало отсчета (нуль), но нет естественной единицы измерений. Шкалами отношений измеряются почти все физические величины – время, линейные размеры, площади, объемы, сила тока, мощность и т.д. В медико-биологических исследованиях шкала отношений будет иметь место, например, когда измеряется время появления того или иного признака после воздействия (порог времени, в секундах, минутах), интенсивность воздействия до появления какого-либо признака (порог силы воздействия в вольтах, рентгенах и т.п.). Естественно, к шкале отношений относятся все данные в биохимических и электрофизиологических исследованиях (концентрации веществ, вольтажи, временные показатели электрокардиограммы и т.п.). Сюда же, например, относятся и количество правильно или неправильно выполненных «заданий» в различных тестах по изучению высшей нервной деятельности у животных.

Шкала интервалов применяется достаточно редко и характеризуется тем, что для нее не существует ни естественного начала отсчета, ни естественной единицы измерения. Примером шкалы интервалов является шкала температур по Цельсию, Реомюру или Фаренгейту. Шкала Цельсия, как известно, была установлена следующим образом: за ноль была принята точка замерзания воды, за 100 градусов – точка ее кипения, и, соответственно, интервал температур между замерзанием и кипением воды поделен на 100 равных частей. Здесь уже утверждение, что температура 300С в три раза больше, чем 100С, будет неверным. В шкале интервалов сохраняется отношение длин интервалов. Можно сказать: температура в 300С отличается от температуры в 200С в два раза сильнее, чем температура в 150С отличается от температуры в 100С.

Порядковая шкала (шкала рангов) – шкала, относительно значений которой уже нельзя говорить ни о том, во сколько раз измеряемая величина больше (меньше) другой, ни на сколько она больше (меньше). Такая шкала только упорядочивает объекты, приписывая им те или иные баллы (результатом измерений является нестрогое упорядочение объектов). Например, так построена шкала твердости минералов Мооса: взят набор 10 эталонных минералов для определения относительной твердости методом царапанья. За 1 принят тальк, за 2 – гипс, за 3 – кальцит и так далее до 10 – алмаз. Любому минералу соответственно однозначно может быть приписана определенная твердость. Если исследуемый минерал, допустим, царапает кварц (7), но не царапает топаз (8), то соответственно его твердость будет равна 7. Аналогично

построены шкалы силы ветра Бофорта и землетрясений Рихтера. Шкалы порядка широко используются в педагогике, психологии, медицине и других науках, не столь точных, как, скажем, физика и химия. В частности, повсеместно распространенная шкала школьных отметок в баллах (пятибалльная, двенадцатибалльная и т.д.) может быть отнесена к шкале порядка. В медикобиологических исследованиях шкалы порядка встречаются сплошь и рядом и подчас весьма искусно замаскированы. Например, для анализа свертывания крови используется тромботест: 0 – отсутствии свертывания в течение времени теста (а через минуту?), 1 – «слабые нити», 2 – желеподобный сгусток, 3 – сгусток, легко деформируемый, 4 – плотный, упругий, 5 – плотный, занимающий весь объем и т.п. Понятно, что интервалы между этими плохо отличимыми и очень субъективными позициями произвольны. В этом случае фраза «Тромботест у исследуемых животных повышался в среднем с 3,3 до 3,7» выглядит абсурдной. Масса подобных шкал все еще встречается в экспериментальной токсикологии, экспериментальной хирургии, экспериментальной морфологии.

Частным случаем порядковой шкалы является *дихотомическая* шкала, в которой имеются всего две упорядоченные градации – например, «выжил после эксперимента», «не выжил».

Шкала наименований (номинальная шкала) фактически уже не связана с понятием «величина» и используется только с целью отличить один объект от другого: номер животного в группе или присвоенный ему уникальный шифр и т.п.

Лекция 2. ЭЛЕМЕНТЫ ТЕОРИИ ПЛАНИРОВАНИЯ ИССЛЕДОВАНИЙ

2.1 Цели и задачи науки. Предмет биометрии – изучение свойств массовых явлений в биологии. Эти явления обычно представляются сложными вследствие разнообразия (варьирования) отдельных индивидуумов или единиц. Чтобы получить правильное представление об изучаемых свойствах массовых явлений и дать им определенные количественные оценки, их подвергают совместному рассмотрению и анализу. Отдельные единицы или индивидуумы, обладающие некоторым общим свойством, объединяют в совокупности. Наблюдаемые единицы называют вариантами (данными, датами), а образуемую совокупность единиц – статистической совокупностью. Статистическая совокупность может быть образована по одному или по нескольким признакам. Она может состоять из одной или нескольких однородных в отношении изучаемого свойства групп. Однако часто бывает целесообразно подразделить отдельные наблюдаемые единицы на группы для достижения большей однородности их внутри этих групп.

Теорию и методы изучения свойств массовых явлений, вычисления и анализа их количественных характеристик изучает биометрия. Метод изучения массовых явлений основан на теории вероятностей. Теория вероятностей устанавливает закономерности событий, наступающих случайно и называемых случайными. Статистика предполагает анализ массовых явлений, имеющих также случайный характер в распределении значений отдельных единиц, составляющих явление.

Центральной задачей биометрии как метода исследования являются заключения, выходящие за рамки изученного материала, т. е. заключения о свойствах статистических совокупностей, принимая во внимание и неизученную их часть.

Всю статистическую совокупность, в отношении которой делают статистические обобщения и заключения, называют общей, или генеральной совокупностью, а часть ее, охваченную непосредственным наблюдением, называют выборочной совокупностью.

Вариационная статистика применяет метод оценки общей совокупности на основе изученных отдельных единиц или на основе выборочных совокупностей.

2.2 Статистические заключения. Статистические заключения о свойствах генеральных совокупностей по выборочным всегда имеют вероятностный характер, т. е. делаются с определенной степенью безошибочности и никогда не делаются с полной достоверностью.

Статистические заключения, как главная составная часть метода исследования массовых явлений, имеют свои отличительные черты. Статистические заключения делают с численно выраженной определенностью. Теоретической основой для их построения является раздел математики, изучающий закономерности случайных событий и называемый теорией вероятностей. Предпосылка, что результаты статистического наблюдения

отобраны в случайном порядке из соответствующих генеральных совокупностей, дает возможность в соответствии с теорией вероятностей оценить степень отклонения результатов наблюдения от соответствующих показателей генеральной совокупности. Таким образом, вероятностная основа вариационной статистики позволяет оценить степень точности получаемых результатов опыта. Основу изучения природных процессов составляет выявление причинно-следственных связей между явлениями экспериментальным путем.

2.3 Теория вероятностей. Осуществив по своему желанию одно или несколько первоначальных явлений (в дальнейшем они называются факторами), экспериментатор получает возможность изучать появляющиеся явления − следствия. Иногда в процессе эксперимента удается сделать случайное открытие, т. е. обнаружить явление − следствие, о котором ранее ничего не было известно. Но, как правило, экспериментатор заранее намечает явления-следствия, появление которых он ожидает. При этом самое сложное явление можно разбить на частные, мелкие явления, относительно которых остается выяснить: произошла они или не произошли. Например, обрабатывая семена на всхожесть определенным препаратом, экспериментатор мог поставить задачу оценить эффект различных его доз. В качестве эффекта могло быть принято число всхожих и невсхожих семян.

Измеряя массу какого-либо вещества, в качестве отдельных частных явлений можно рассматривать всевозможные априорные значения этой массы. Задача экспериментатора, таким образом, сводится к наблюдению того, какие из значений массы осуществились.

Явления, рассматриваемые с той точки зрения, осуществились они или не осуществились, называются событиями. Применительно к событиям ставится основная задача: предсказать, появится ли изучаемое событие при осуществлении некоторого наперед заданного комплекса факторов (явлений − причин). Событие, которое при заданном комплексе факторов обязательно произойдет называется достоверным. Событие, которое при заданном комплексе факторов не может произойти, называется невозможным событием. Суждения о достоверности или невозможности некоторого события являются категорическими суждениями. Такие суждения принято, считать окончательным результатом исследования. Отсюда возникает интерес к обратной задаче: указать комплексы факторов, при которых о заданном событии можно сделать категорические суждения.

Однако каждое событие − результат действия многих факторов, часть из которых иногда нельзя предсказать или организовать в опыте. В этом случае категорическое суждение о событии невозможно. Получается ситуация: заданные факторы благоприятствуют событию, и, следовательно, оно может произойти. С другой стороны, действия этих факторов недостаточно, чтобы гарантировать появление события, и, значит, оно может и не произойти.

Событие, которое при заданном комплексе факторов может либо произойти, либо не произойти, называется случайным событием. Случайные события связаны с действием не вошедших в организованный комплекс

факторов, называемых случайными факторами в отличие от другой группы факторов, включаемых в комплекс и называемых основными, или неслучайными.

Предположим, исследуется урожайность культур. Такие факторы, как технология возделывания, внесение различных доз удобрений и т.д. можно организовать в опыте, т. е. учесть. Эти факторы являются основными. Другая группа факторов является неизвестной, или не поддающейся учету. Эти факторы при статистическом анализе получили название случайных.

Для того чтобы выяснить, произойдет или не произойдет событие при заданном комплексе факторов, нужно осуществить этот комплекс, т. е. провести испытание. Испытанием является любой эксперимент, в результате которого производят наблюдения.

Предсказать результат единичного испытания можно только для достоверных или невозможных событий. Случайность же события не видна из единичного испытания. Любое случайное событие по единичному испытанию было бы оценено как достоверное, если оно произошло, и как невозможное − если не произошло. Такие оценки, однако, были бы сами случайными, как и результат единичного испытания. Теория оценки случайных событий строится на большом числе испытаний, т. е. для массовых событий.

Важным условием при этом является неизменность комплекса основных факторов. События, происходящие при одном и том же комплексе факторов, называются однородными. Установлено, что однородные случайные события в большой их массе подчиняются некоторым закономерностям. Эти закономерности получили название вероятностных.

Характер вероятностных закономерностей можно уяснить на следующих примерах.

Пример. При подбрасывании монеты возможны два события: выпадение монеты гербом или решкой. События с одинаковыми возможностями осуществления называются равновозможными. Так, при симметричной монете выпадение герба и цифры − равновозможны.

Однако, если бы было произведено, например, 1000 бросаний, и из них 700 раз выпал герб, то для следующей серии испытаний можно было бы предсказывать, что герб появится в 70% случаев. Причем такое отклонение от ожидаемых 700 появлений герба из 1000 бросаний можно было бы считать связанным с несимметричностью монеты.

Установленное в результате опыта отношение числа появления события к общему числу всех испытаний называется частотой события. В указанном примере с монетой частота выпадения герба равна 0,7.

Из примера можно заключить, что частота события, выступающая как некоторая статистическая закономерность, связана с внутренними характеристиками события. Частота является мерой этих внутренних характеристик события. Она тем надежнее, чем большее число испытаний было произведено. При очень большом числе испытаний частота почти перестает изменяться, приближаясь к некоторой величине. Эту величину и можно принять за интересующую нас числовую характеристику. Так, при бросании

монеты 4, 12 и 24 тыс. раз частота появления герба соответственно равнялась 0,6080; 0,5016; 0,5005. Очевидно, что она здесь приближается к числу 0,5.

Числовая характеристика случайного события, обладающая тем свойством, что для любой достаточно большой серии испытаний частота события лишь незначительно отличается от этой характеристики, называется вероятностью события.

Из этого рассмотрения устанавливаем, что вероятность является тем теоретическим пределом, к которому стремится частота событий при увеличении числа испытаний. Вероятность − идеальное выражение частоты событий.

Данное определение вероятности называется статистическим. Это определение не является достаточно строгим с точки зрения математики. По статистическому определению трудно изучать свойства вероятности.

Однако имеется и ряд положительных его свойств. Статистический подход позволяет находить вероятности событий, структура которых неизвестна. Например, только статистический подход позволил определить вероятность рождения мальчиков, равную 0,52 и девочек − 0,48.

Существуют два других, более удобных с формальной точки зрения, определения вероятности: классическое и геометрическое. Однако для них требуется знать структуру рассматриваемых событий.

Понятие о геометрическом определении вероятности можно получить из следующего примера испытаний.

Пример. Предположим, в некотором квадрате случайным образом выбирается точка. Какова вероятность, что она окажется в области *D*. Очевидно, что вероятность эта будет тем большей, чем больше область *D*. В качестве мерила вероятности выступает здесь площадь. Вероятность того, что случайная точка попадет в область *D* (осуществление события *D*) равна: $p(D)=S_D/S$, где S_D−площадь области *D; S* − площадь всего квадрата.

Геометрическое определение вероятности пригодно не только для плоскости, но и для прямой или пространства.

В первом случае основой для определения вероятности служит некоторый отрезок, а случайным событиям соответствуют его части. Вероятность вычисляется как отношение длины частей к общей длине отрезка. Во втором, случае основой к испытанию принимают некоторый куб, случайным событиям соответствуют различные тела, расположенные в кубе. Вероятность вычисляют как отношение объемов тел к объему куба.

Наибольший интерес представляет классическое определение вероятности. С этим определением связаны основные теоремы теории вероятностей.

Вероятность здесь определяется априори, до испытаний, исходя из определенной структуры случайных событий, т. е. из разбивки на равновозможные исходы.

Пример. Пусть при подбрасывании монеты появления герба или цифры будут изучаемыми событиями ***a*** и ***b***. Причем, если при одном бросании произойдет событие ***a***, то не произойдет другого события ***b***. Такие события

называют несовместными. Каждое из событий называют исходом испытания. В силу равновозможности исходов в нашем испытании вероятность каждого события равна. При единичном бросании кубика с 6 гранями (имеющими, например, 1, 2, 3, 4, 5, 6 очков), вероятность появления любой одной грани *p= 1/6*.

Исходы испытания являются простейшими случайными событиями. Можно рассматривать более сложные события, объединяющие несколько исходов. Например, при бросании игрального кубика мы можем интересоваться таким событием, как выпадение числа очков больше 2. В таком случае говорят, что появлению события с выпадением больше двух очков, т. е. с 3, 4, 5 и 6 очками, благоприятствуют четыре исхода из шести. Вероятность этого события *p=4/6*. Таким образом, мы подошли к классическому определению вероятности. Вероятностью случайного события называется отношение числа отходов, благоприятствующих событию, к числу всех возможных исходов.

2.4 Основные теоремы теории вероятностей. Если некоторое событие *A* может произойти при *n* испытаниях и *m* − число исходов, которые благоприятствуют наступлению события, то вероятность того, что данное событие произойдет, может быть определена как *P(A)=m/n*. Тогда, сумма вероятностей двух несовместных событий равна единице.

Сложение вероятностей. $P(A+B)=P(A)+P(B)=m_1/n + m_2/n$

Если в урне с 10 шарами 6 шаров черных, 3 белых и 1 зеленый, вероятности этих событий будут равны, соответственно, 6/10, 3/10 и 1/10.

Какова вероятность вынуть белый или зеленый шар?

Благоприятствует появлению белого шара 3/10 всех исходов, а зеленого шара − 1/10 исходов. Появлению либо белого, либо зеленого шара соответствует *p*=3/10+1/10=4/10=0,25, т. е. вероятность суммы двух несовместных (взаимоисключающих случайных) событий равна сумме их вероятностей.

Умножение вероятностей. Два события называются независимыми, когда наступление одного не оказывает влияния на наступление другого. Так, результат одного метания кости не влияет на результат следующего метания.

Вероятность сложного события (т. е. наступления двух событий независимых одно от другого равна произведению вероятностей отдельных событий. $P(AxB)=P(A)xP(B)=m_1/n$ x m_2/n

Например, вероятность выпадения очка, а затем двух очков, при двух последовательных бросаниях кубиков, равна *p*==1/6x1/6=1/36.

Вычисление вероятностей. Часто возникает необходимость одновременно складывать и умножать вероятности. Например, требуется определить вероятность выпадения 5 очков при одновременном бросании 2 кубиков. Искомая сумма вероятностей может получиться как результат одной из следующих 4-х комбинаций исходов:

кубик *a* 1, 2, 3, 4;

кубик *b* 4, 3, 2, 1

Вероятность получения одного очка на кубике *a* равна 1/6 и получения четырех очков на кубике *b* −также 1/6. Вероятность получения комбинации

этих очков равна 1/36. Аналогично и вероятность трех других комбинаций равна 1/36. Но любой из этих четырех результатов, дающий в сумме 5 очков, будет считаться благоприятным исходом. Отсюда вероятность искомого исхода

p = 1/36 + 1/36 + 1/36 + 1/36 = 1/9.

Более общая форма вопроса о вероятности события является такой: какова вероятность получения не менее, например, 8 очков при бросании 2 костей? Число очков, равное и более 8, рассматривается как благоприятный исход.

Рассчитаем вероятность каждого благоприятного результата:

Вероятность появления 12 очков	1/36
Вероятность появления 11 очков	2/36
Вероятность появления 10 очков	3/36
Вероятность появления 9 очков	4/36
Вероятность появления 8 очков	5/36
Сумма вероятностей	15/36

Вероятность выпадения по меньшей мере 8 очков при бросании 2 костей равна 15/36 или 5/12.

Биномиальное разложение и измерение вероятностей

Изложенные примеры исчисления вероятностей можно обобщить на основе следующей ниже иллюстрации вывода.

Если подбрасываются одновременно 2 монеты (***a, b***), то существуют 4 возможных случая выпадения герба Т и цифры Н:

ab ab ab ab

TT TH HT HH

В первом исходе имеем 2 герба. Принимая это за 2 благоприятных исхода, получим вероятность каждого из них ***p***, а сложного события (TT) $p*p=p^2$. В данном случае, при $p=1/2$ $p^2=l/4$.

Четвертый из возможных исходов НН представляет 2 неблагоприятных исхода с вероятностью $q*q=q^2=1/4$.

Каждый из двух других исходов является комбинацией одного благоприятного и одного неблагоприятного случаев.

Вероятность каждого из этих исходов равна $1/4=pq=l/2*l/2$, а обоих вместе ТН и НТ равна их сумме, т. е. $2pq=1/2$.

Обобщенным выражением процесса получения вероятностей различных сочетаний независимых событий, когда вероятности их известны, являются последовательные члены разложения бинома.

Для рассматриваемого примера из двух событий имеем:

$(p+q)^2 = p^2 + 2pq + q^2$. , При $p = 12$ получим $(1/2+ 1/2)^2= 1/4+ 1/2 + 1/4$.

Если 3 монеты *a, b, c* подбрасываются одновременно, получим 8 возможных комбинаций:

Abc abc abc abc abc abc abc abc

TTT TTH THH THT HTT HTH HHT HHH

Вероятность выпадения 3 гербов составит 1/8, 2 гербов (в сочетании с одним случаем цифры) равна 3/8, одного герба и 2 цифр − 3/8, ни одного

герба−1/8. При 3 независимых событиях степень бинома равна 3.

Вероятности отдельных возможных исходов даются последовательными членами разложения:

$(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$.

При $p=q= 1/2$ имеем $(1/2 +1/2)^3 = 1/8 + 3/8 + 3/8 + 1/8$, т. е. то же, что и непосредственным подсчетом.

Если число независимых случайных событий n, то вероятность n, $n-1$, $n-2$ и т. д. благоприятных исходов равна последовательным членам разложения:

$(p + q)^n$

Если желаем получить вероятные численности разных исходов при данном числе испытаний n, применяем выражение:

$N(p+q)^n$.

Например, при числе испытаний $N=200$ и двух независимых событиях n в каждом испытании вероятные численности будут равны $200\ (p+q)^2=200(p^2+2pq+q^2)$. Если $p=q=1/2$, имеем последовательные вероятные численности: 50+100+50.

При подбрасывании монеты 200 раз ($N=200$) выпадения герба следует ожидать в 50 случаях, герба или цифры − в 100 случаях и цифры − 50 случаях.

При тех же p и N, но $n=3$ получим последовательные вероятные численности: 25+75+75+25, которые означают 3, 2, 1 наступление события и ненаступление его, причем сумма всех численностей равна N.

При 200 бросаниях трех монет ожидаем в 25 случаях выпадения 3 гербов (ТТТ), в 75 случаях выпадения 2 гербов и одной цифры (ТТН), в 75 случаях выпадения 2 цифр и одного герба (ННТ) и в 25 случаях − 3 цифр.

Итак, когда вероятности независимых событий известны априори, то можно определить вероятные численности любого данного числа n, $n-1$, $n-2$.... наступления события и ненаступления его. При этом неважно, равны или не равны p и q, лишь бы они оставались при испытаниях постоянными. Этот факт имеет большое значение в теории статистики и используется ниже.

При изучении природных явлений выделение элементарных событий и вообще расчленения причинного процесса, в результате которого происходят случайные события, обычно невозможно. Классический подход к определению вероятности здесь бессилен. Проблему определения вероятностей таких событий решают на основе статистического подхода.

Однако классический подход к определению вероятностей событий лежит в основе теории анализа случайных событий и теоретических (модельных) распределений исходов испытаний. В свою очередь теория математического анализа случайных событий и модели распределений исходов испытаний являются базой статистических методов, в частности, базой статистических заключений.

Лекция 3 ОПИСАТЕЛЬНАЯ СТАТИСТИКА

3.1 Характеристика совокупности. Всякое множество отдельных отличающихся друг от друга и в то же время сходных в некоторых существенных отношениях объектов составляет так называемую совокупность. (популяции рыжих полевок того или иного района, стадо коров данного хозяйства, потомство определенного быка, заготовляемые в области или крае беличьи шкурки, растения на опытных делянках, группа цыплят, на которых ставится опыт по применению антибиотиков, мальки окуня в озере и т. д.) Понятие совокупности применимо не только к животным и растениям. Такими же совокупностями являются, например, дети, родившиеся в стране в течение какого-то года пли месяца, молекулы газа в том или другом объеме.

В состав совокупности входят различные члены, или единицы: для популяции животных − каждое отдельное животное, для стада коров единицей является каждая корова, для совокупности шкурок − каждая шкурка, для потомства быка − каждый теленок, от него полученный, для совокупности зерен гречихи − каждое отдельное зерно.

Обычно число единиц совокупности называют объемом совокупности и обозначают латинской буквой n. Единица совокупности может характеризоваться определенными признаками, например: коровы − удоями за лактацию, весом, мастью; молекулы газа − скоростями их движения и т. д. Каждый изучаемый признак принимает разные значения у различных единиц совокупности, он меняется в своем значении от одной единицы совокупности к другой. Это различие между единицами совокупности называется вариацией или дисперсией (т. е. рассеянием).

Мы говорим − признак варьирует. Это означает, что он принимает различные значения совокупностей. Так, совокупность из всех животных данной у разные членов совокупности, например, у коров данной породы, мышей опытной группы поросят одного помета и т д. Значение или меру признака единицы совокупности называют вариантой и обозначают буквой x. Значок i – порядковый номер варианты. Несмотря на различия между вариантами по значению изучаемого признака, совокупность этих вариант обладает однородностью. Беличьи шкурки неодинаковы по окраске, размеру, качеству меха, но они однородны, так как все они – шкурки особей одного и того же вида – белки обыкновенной.

Различают совокупности:

1. Генеральную
2. Выборочную (выборка).

Генеральная – теоретически бесконечная совокупность всех единиц или членов, которые могут быть отнесены к ней. Из-за бесконечно большого числа членов генеральную совокупность изучить практически невозможно. Поэтому из нее выбирают часть для непосредственного изучения, т.е выборку.

Существует несколько способов отбора вариант в выборку:

1. плановый отбор (групповой отбор; гнездовой (или серийный));
2. стихийный отбор (механический).

Единственное условие – однородность отбираемого в выборку материала.

Задачей изучения всякой совокупности является получение статистических (или, как иногда говорят, биометрических) характеристик, или показателей, которые позволяют судить о данной совокупности в целом, о различиях внутри нее и об отличии ее от других, сходных с ней или близких к ней совокупностей. Совокупность становится статистической тогда, когда в ее описание вносится количественный метод. Применение количественного метода изучения совокупности и позволяет получать для нее статистические характеристики, с помощью которых получают основную информацию о совокупности.

3.2 Варьирующие признаки и их учет. При изучении единиц совокупности по признаку необходимо записать полученные данные и сгруппировать их. Способы группировки зависят от характера вариации изучаемых признаков.

Различают следующие типы вариации признаков:

- качественная;
- количественная

Если различия между вариантами выражаются в каких-то качествах, то такую вариацию называют качественной. Если совокупность животных характеризуют по масти, тогда каждая варианта должна получить качественную характеристику в соответствии с заранее принятыми обозначениями: черная, рыжая, черно-пестрая, черно-рыжая и т. д. В этом простейшем случае подсчет числа особей в каждой из выделенных групп дает представление о составе популяции в целом.

В других случаях различия между вариантами будут количественными. Количественная вариация может быть двух типов: прерывная (дискретная) и непрерывная. В первом случае различия между вариантами, отдельными значениями случайной переменной, выражаются целыми числами, между которыми нет и не может быть переходов. Например, количество детенышей в помете (поросят у свиноматок, щенков у серебристо-черных лисиц), число сосков у сриноматок, число лучей в плавниках рыб, количество лепестков в цветке, число позвонков у птиц и т. д. Для изучения подобного варьирования надо сосчитать у каждой единицы совокупности число изучаемых элементов и записать его на соответствующую карточку. При непрерывной вариации значения вариант не обязательно выражаются только целыми числами. Все зависит от того, какая степень точности принимается для характеристики данного количественного признака. Так, например, при изучении веса крупного рогатого скота можно ограничиться значениями вариант, выраженными в килограммах, отбросив граммы, но совершенно недостаточно округлять до килограммов веса рыб, так как грамм здесь имеет большое значение. В опытах же по изучению влияния гормонов на рост гребня у цыплят вес гребня придется измерять в миллиграммах. Молочную продуктивность за лактацию обычно выражают в килограммах, но общая картина удоев не изменится, если округлять ее до десятков килограммов. Оценка же жирности молока в

процентах, выраженных целыми числами, явно недостаточна, ее надо давать с учетом десятых и даже сотых долей процента. Однако во всех этих и им подобных случаях существует непрерывная вариация, выражающаяся в том, что между вариантами возможны все переходы. При изучении непрерывной вариации надо все единицы совокупности характеризовать количественно с той степенью точности, которая заранее намечена и больше всего подходит в данном конкретном случае.

3.3 Группировка данных при качественной вариации. Чтобы проанализировать ту или иную совокупность, необходимо сгруппировать полученные отдельные варианты и затем представить эту группировку в виде таблицы или ряда. При упорядочении полученных данных легко обработать их математически и вывести статистические показатели, которые будут исчерпывающе характеризовать изучаемую совокупность. Проблема группировки занимает большое место в статистике вообще (особенно в экономической), так как ошибочная группировка данных может привести к неправильным выводам о существе изучаемого явления.

Наиболее проста группировка при качественной вариации. Так, если норки различаются по окраске, то их распределение может быть выражено в количестве животных каждой окраски и в процентах, которые составляют норки каждой окраски от общего количества животных.

Частным случаем качественной вариации является альтернативная, когда в совокупности можно выделить только две группы. У членов одной группы присутствует определенное качество (или признак), у членов другой группы его нет. Так, при проверке на туберкулез животные распадаются на 2 группы − с положительной реакцией и с отрицательной. Одни коровы в данном стаде рогатые, другие − комолые и т. д.

Группировка данных при количественной дискретной вариации. При количественной вариации необходимо предварительно наметить для таблицы классы, охватывающие все полученные количественные данные от минимальных до максимальных. Это легко сделать при прерывной (дискретной) количественной изменчивости.

Допустим, что была изучена плодовитость 80 самок серебристо-черных лисиц, т. е. число родившихся у каждой самки щенков. Варианты x_1, x_2,, x_3, ..., x_n этой совокупности выражены цифрами, представленными в табл. 1.

Таблица 1

Количество щенков у 80 самок серебристо-черных лисиц

4	5	3	4	6	7	8	3	1	4
6	4	4	3	2	5	3	4	5	4
5	3	4	5	4	4	4	6	5	7
6	4	5	4	4	4	4	2	3	4
5	5	4	5	4	4	6	4	4	4
4	8	7	5	4	9	4	3	4	4
5	4	6	4	4	3	4	4	4	2
4	4	5	4	6	4	3	3	4	2

Группировку вариант лучше всего провести по значениям отдельных

вариант. Минимальное число щенков 1, максимальное − 9. Отсюда естественно установить 9 классов: с 1 щенком, с 2, 3 и т. д.− и распределить все варианты по этим 9 классам. Наиболее простым способом разнесения вариант по классам является следующий.

Составляется таблица («классы» и «частоты») с намеченными 9 классами и в соответствующие горизонтальные строчки разносятся все варианты, начиная от первой. Обозначаются они так: первые четыре варианты данного класса − точками, а последующие − черточками, соединяющими четыре точки. (конвертик, домик, елочка).

Пример разноски:

Классы, x	Разноска	Частота, f
1	.	1
2	. .	
	. .	4
3	⊠	10
4	⊠ ⊠ ⊠ ⧅	39
5	⊠ ¨	
	.	13
6	⊏	7
7	. .	
	.	3
8	. .	2
9	.	1

Вторичная группировка данных при количественной дискретной вариации. В разобранном выше примере классов намечено столько, сколько было в изученной совокупности различных значений вариант (от 1 до 9 щенков). Однако такой способ будет нецелесообразным при очень большой вариации дискретного признака.

Так, например, у змеи *Lampropeltis getulus* количество хвостовых щитков варьировало от 40 до 58 (табл. 2). . :

Таблица 2

Количество хвостовых щитков у 60 экземпляров змеи *Lampropeltis getulus*

42	58	44	54	41	50	46	46	54	48	43	49
50	48	46	46	45	53	48	48	53	53	48	41
46	40	50	43	49	51	52	46	42	44	48	45
47	46	43	50	47	45	48	40	44	42	48	45
54	50	56	48	45	45	51	42	44	47	46	45

Если классы намечать по значениям каждой варианты, т. е. 40, 41 и т. д., то получится 19 классов, ряд окажется растянутым, труднообозримым, с перерывами в некоторых классах. Лучше наметить классы, охватывающие несколько значений вариант, например: 40−41, 42−43 и т. д. или 40−42, 43−45 и т. д. В первом случае вариационный ряд будет состоять из 10 классов, во втором − из 7. Имеем классовый промежуток – I, равен 3.

Таблица 3

Границы классов	Средний класс, x	Разноска	Частота, f
40-42	41	□	8
43-45	44	□	14
46-48	47	□□	20
49-51	50	□	9
52-54	53	□	7
55-57	56	.	1
58-60	59	.	1

3.4 Вариационный ряд и его графическое изображение. После распределения вариант по классам получаются ряды, показывающие как часто встречаются варианты каждого класса и как варьируют признак от минимума до максимума. Т.о., ВР – двойной ряд чисел, показывающий распределение вариант по их частоте или встречаемости. По ВР можно судить не только о границах, но и о характере вариации.

Класс, обладающий наибольшей частотой, получил название модального, значения же крайних классов называют лимитами или пределами.

Всякий вариационный ряд можно изобразить графически. Графическое изображение вариационного ряда в общем виде получило название кривой распределения или вариационной кривой.

Существуют два способа графического изображения конкретных вариационных рядов. Первый из них, применяющийся при дискретной вариации, но в том случае, если классы намечены по отдельным значениям вариант, носит название полигона распределения. На оси абсцисс нанесены классы, на оси ординат−частоты. Высота каждого класса, пропорциональная частоте класса, отмечается кружком. При непрерывной вариации, если классы намечены по границам, на оси абсцисс наносят нижние границы классов, на оси ординат−частоты. Такой график носит название – гистограммы.

При статистической обработке материала возникает вопрос: сколько классов необходимо намечать? Это зависит от:

- объема совокупности;
- от величины вариационного размаха.

На практике можно руководствоваться примерно следующими правилами:

Количество вариант	Число классов
25−40	5−6
40−60	6−8
60−100	7−10
100−200	8−12
более 200	10−15

Вариационный ряд при непрерывной изменчивости также может быть изображен на графике. В этом случае нужно строить гистограмму, т. е. ступенчатую диаграмму.

Характер распределения вариант в вариационном ряду.

Изучая распределение вариант в вариационных ряду легко заметить некоторые общие закономерности, а именно:

1) большинство вариант располагается в средней части вариационного ряда или около середины вариационной кривой, здесь наблюдается максимум вариант, как бы их сгущение;

2) распределение вариант в обе стороны от этого максимума более или менее симметрично;

3) частота вариант постепенно убывает к краям вариационного ряда.

Эти закономерности в той или иной степени присущи любому вариационному ряду. В дальнейшем мы увидим, что закономерности вариационного ряда основываются на закономерностях случайной вариации, изучаемых теорией вероятностей.

Лекция 4 ОПИСАТЕЛЬНАЯ СТАТИСТИКА. СРЕДНИЕ ВЕЛИЧИНЫ

4.1 Две группы показателей для характеристики вариационных рядов. В предыдущей лекции мы рассмотрели способы сведения данных, составляющих статистические совокупности, в вариационные ряды. Каждый вариационный ряд и его графическое изображение − это как бы «сгущение» исходного фактического материала, превращение его в наглядную форму. Однако этого недостаточно. Очень важно получить характеристики для совокупности, которые были бы выражены цифровыми показателями. С их помощью можно было бы сравнивать разные ряды. Одним из простейших способов количественной характеристики вариационного ряда является указание на его размах, т. е. на верхнюю и нижнюю его границы, которые обычно называют лимитами. Если, например, известно, что вариационный ряд по молочной продуктивности одного стада коров имеет размах от 2000 до 4000 кг, а другого−от 2500 до 6800 кг, то, казалось бы, можно сделать вывод о более высоком качестве второго стада. Однако лимиты не указывают на то, как распределяются по изученному признаку отдельные члены совокупности. Вот почему для характеристики совокупности нужны такие показатели, которые отражали бы свойства всех ее членов.

Вариационные ряды могут различаться: а) по тому значению признака, вокруг которого концентрируется большинство вариант. Это значение признака отражает как бы уровень развития признака в данной совокупности, или, иначе, центральную тенденцию ряда, т. е. типичное для ряда; б) по степени вариации вариант вокруг уровня, по степени отклонения от центральной тенденции ряда.

Соответственно этому статистические показатели разделяются на две группы:

- показатели, которые характеризуют центральную тенденцию, или уровень ряда,
- показатели, измеряющие степень вариации.

К первой группе относятся различные средние величины: мода, медиана, средняя арифметическая, средняя геометрическая. Ко второй − вариационный размах, среднее абсолютное отклонение, среднее квадратическое отклонение, варианса (дисперсия), коэффициенты асимметрии и вариации. Существуют еще и другие показатели, но их мы не будем рассматривать, так как они редко применяются в биологической статистике.

Мода и медиана. При изучении распределения самок лисиц по числу щенков в помете обнаружилось, что 39 самок из общего числа 80 имели по 4 щенка, т. е. класс «4 щенка» обладал наибольшей частотой. Такой класс был назван модальным. Значение же модального класса называют модой. Мода обозначается символом Мо. Величина моды является как бы типичной для всей совокупности. Действительно, в нашем примере почти половина самок из 80 имела в помете именно 4 щенка.

Для ряда распределения змей по числу хвостовых щитков (табл. 2) модальным является класс «46−48 щитков». А так как класс здесь охватывает несколько значений вариант, то для его характеристики надо вычислить

среднее значение класса. Оно равно 46 + 48/2= 47. В таком случае Мо = 47 щиткам.

К__числу средних величин относится также медиана. Медиана− это значение варианты, находящейся точно в середине ряда (обозначается Ме).

Чтобы найти такую варианту, надо сначала расположить все варианты по порядку от минимальных их значений до максимальных. Такое расположение вариант называют ранжировкой. Чтобы определить Ме при четном числе вариант, надо взять значения двух соседних срединных вариант, например при п = 80 значения вариант с порядковыми номерами 40 и 41, и разделить их сумму на 2. В примере, представленном в табл. 4, обе эти варианты будут иметь значения «4 щенка», следовательно, Ме данного ряда = 40.

Медиана и мода дают известное представление о совокупности в целом. Они характеризуют своего рода типичное в данной совокупности (конечно, речь идет только о каком-то определенном признаке).

Использование моды и медианы в биологии в настоящее время довольно ограничено, но в некоторых случаях без них очень трудно обойтись, в частности, если полученные данные не являются чисто количественными, а поэтому не могут быть представлены в виде точного вариационного ряда. Так, например, тяжесть заболевания подопытных животных или их упитанность можно условно оценивать степенями: слабая, удовлетворительная, средняя, высокая−или баллами 1, 2, 3 и т. д. Тогда мода или медиана могут достаточно хорошо характеризовать типичное в совокупности.

Обычно же, когда изучаемая совокупность достаточно_однородна и вариация внутри нее чисто количественная, выгоднее пользоваться, другими средними величинами.

2. Средняя арифметическая и ее свойства. Нахождение средней арифметической − это в сущности замена индивидуальных варьирующих значений признаков отдельных членов совокупности некоторой уравненной величиной при сохранении основных свойств всех членов совокупности. Этому условию в наибольшей степени удовлетворяет так называемая «средняя арифметическая, обозначаемая» - X (ранее обозначали М).

Представим себе, что ряд членов совокупности, т. е. ряд значений случайной переменной x_1, x_2, …, x_i заменим таким же рядом из одинаковых величин х, т. е. х, х, х, ..., х (n раз).

Тогда сумма всех вариант совокупности $x_1+x_2+x_3+\ldots+x_n$ будет равна

X+X+X+…+X (n раз), т. е. nX. Сумму всех вариант совокупности можно сокращенно обозначить х, (х − обозначает значение любой варианты; греческая буква Σ − большая сигма − обозначает суммирование; конкретные суммы часто обозначают также латинской буквой S). Тогда

Σ х, = nx, откуда

X= Σ x/n

Мы получили наиболее общую и в то же время наиболее простую формулу средней арифметической. Для того чтобы вычислить среднюю арифметическую, достаточно сложить значения всех вариант и сумму разделить на общее число вариант. В простейших случаях так и делают.

Очевидно, в таких случаях можно пользоваться данными, полученными непосредственно при анализе членов совокупности, не прибегая к группировке вариант.

Однако при большом количестве вариант этот прямой способ определения средней арифметической по указанным формулам оказывается не столь удобным. Кроме того, при его применении нет возможности вычислить некоторые другие биометрические показатели. Поэтому на практике часто пользуются окольными методами вычисления средней арифметической на основе уже сгруппированных данных. Эти методы будут разобраны позднее.

Свойства средней арифметической:

1. Если каждую из вариант совокупности, для которой вычисляется средняя арифметическая, увеличить или уменьшить на одну и ту же величину, то и средняя арифметическая соответственно увеличится или уменьшится на столько же.

X1 /A; X2 /A; X3 /A; X4 /A; Xi /A; то Σxi/A

2. Алгебраическая сумма отклонений отдельных вариант от средней арифметической (т. е. разностей между каждым кон¬кретным значением признака и средней арифметической) рав¬няется нулю:

$\Sigma(xi - X)=0$.

3. Сумма квадратов отклонений от средней арифметической меньше суммы квадратов отклонений от любой другой вели¬чины A не равной X, т. е.

$\Sigma (xi-X)2 < \Sigma (xi-A)2$, если A не равно X.

4.2 Значение средней арифметической и ее сущность. Средняя арифметическая, как и некоторые другие средние, известна издавна. Она имеет очень большое значение в науке и технике. Нет буквально ни одной биологической работы, в которой не встречались бы в той или другой форме средние арифметические. Средняя арифметическая является обобщающей величиной, которая как бы впитывает в себя все особенности данной совокупности или ряда распределения. Она отражает уровень всей совокупности в целом, дает сводную, обобщенную характеристику данного изучаемого признака.

Цифровое значение средней арифметической как таковое может не встретиться ни в одном конкретном случае в совокупности. Может оказаться, что ни одна варианта не будет ей равной. Если среднее число щенков у серебристо-черных лисиц равно 4,7, то, очевидно, фактическое число щенков никак не может быть дробным. В этом смысле средняя арифметическая является абстрактной величиной. Но в то же время она и конкретна. Она выражается в тех же единицах измерения, что и варианты ряда. При определении средней арифметической взаимопогашаются, отметаются случайные колебания, отклонения от центральной тенденции, от уровня вариационного ряда и выступает общий закон явления. Вскрывается типичное для всей совокупности в целом.

В то же время нужно предостеречь от возможных ошибок в понимании средней арифметической.

- Средняя арифметическая характеризует всю совокупность в целом,

а не отдельные члены совокупности. Среднее число щенков в помете лисиц 4,7 относится только ко всей группе, каждая же отдельная лисица характеризуется своим числом щенков в помете−от 1 до 9.

- Средняя имеет смысл только по отношению к качественно однородной совокупности. Так, нельзя вычислять средний вес животных для группы, включающей и молодняк разных возрастов и взрослых животных. Надо взять каждую возрастную группу отдельно и для них вычислить X.
- Поскольку средняя арифметическая относится к данной совокупности, перенесение ее на явления, выходящие за ее рамки, рискованно, без специального анализа вопроса о правомерности такого перенесения.
- Средняя относится лишь к отдельным изучаемым признакам и не может быть автоматически перенесена на их сумму.

4.3 Измерение вариации. Вариационный размах и средние отклонения. Средняя арифметическая указывает на то, какое значение признака наиболее характерно для данной совокупности. Но сама по себе она еще недостаточна для характеристики совокупности, так как главной особенностью совокупности является наличие разнообразия между ее членами, т. е. вариации. Если бы не было вариации, то информацию о совокупности можно было бы получить по одному члену совокупности. При наличии же вариации эта информация должна быть основана на учете характера и степени вариации.

Учет вариации того или другого признака в совокупности имеет очень большое значение для биолога, так как всякая вариация в популяции животных или растений в конечном счете отражает различия между организмами − в их наследственной природе и в тех условиях, при которых они выращивались. Приемы работы с животными должны меняться в зависимости от характера их вариации. Без оценки вариации невозможно и сравнение двух совокупностей.

Два стада коров могут иметь очень близкие средние удои, но в одном величины удоев сильно различаются, в другом же коровы представляют собой довольно однородную группу с небольшим размахом колебаний. Определение вариационного размаха, т. е. разницы между максимальным и минимальным значениями вариант, может в известной степени указывать на степень вариации, но оно недостаточно. Во-первых, крайние величины в рядах не очень устойчивы, и при изменении количества изучаемых особей они легко сдвигаются. Во-вторых, при одних и тех же пределах вариации распределение вариант в рядах может быть различным. Вот почему для характеристики различий между отдельными значениями случайной переменной x, иначе говоря, вариации между членами совокупности нужен такой показатель, который обобщал бы колеблемость всех вариант. Для этого надо сравнивать варианты илу друг с другом, или с какой-то одной постоянной величиной. В качестве последней лучше всего взять среднюю арифметическую.

Варианса и среднее квадратическое отклонение. Более совершенными показателями, характеризующими вариацию, являются средний квадрат отклонений вариант от средней арифметической, иначе называемый вариансой,** и среднее квадратическое отклонение, или, иначе, стандартное отклонение. Вариансу обозначают $\sigma 2$ (греческая буква сигма) или s 2

(латинская буква эс), а среднее квадратическое отклонение−σ.

Словами это можно формулировать так: варианса − это сумма отклонений отдельных значений вариант от средней арифметической, деленная на общее количество вариант, а среднее квадратическое отклонение − корень квадратный из этого частного. Хотя после извлечения корня квадратного получаются значения со знаками плюс и минус, обычно берут только положительное значение.

Степени свободы. Величина n − 1 получила особое название− число степеней свободы (точнее, число степеней свободы вариации). Мы будем обозначать ее буквами df. Так как во многих разделах статистики приходится пользоваться числом степеней свободы, то следует объяснить его значение.

Существуют различные способы вычисления статистических показателей:

а) прямой через значения вариант (без ВР, при малом n);

б) прямой через значения вариант для ВР;

в) непрямой способ (способ условной средней)

Из всего сказанного видно, что для определения статистических показателей требуется довольно большая вычислительная работа, но объем ее может быть сокращен правильным выбором метода, наиболее подходящего для обработки данного материала, и применением имеющихся технических средств для вычислений ЭВМ, лучше всего пользоваться прямым способом вычислений, так как он дает наиболее точные результаты. Непрямому же способу в силу искусственной разбивки материала на классы всегда сопутствует известная неточность.

Средняя геометрическая. Средняя арифметическая − наиболее часто применяемый статистический показатель, в том числе в биологии. Однако в некоторых случаях (например, при изучении темпов роста организмов или роста целых популяций приходится пользоваться другой средней величиной − средней геометрической.

Формула для ее вычисления следующая:

$$X_g = \sqrt{x1 \cdot x2 \cdot x_n} = \sqrt{\Pi x_i}$$

Очевидно, что при ее определении надо исключать варианты выражающиеся нулем или отрицательным числом.

На практике вычисление средней геометрической производится с помощью логарифмов по следующей рабочей формуле

$$\text{Log } X_g = 1/n (\log x_1 + \log x_2 + \log x_n)$$

т. е. логарифм средней геометрической равен арифметической средней суммы логарифмов отдельных значений x. По значению log X затем определяется величина x .

Лекция 5 СТАТИСТИЧЕСКАЯ ГИПОТЕЗА. ВЫБОРОЧНЫЙ МЕТОД

5.1 Проблема достоверности в статистике. Приемы и методы, изложенные в предыдущих лекциях дают возможность исчерпывающе охарактеризовать биологические совокупности. Каждая совокупность может быть представлена в виде ряда распределения. Для ряда распределения можно определить статистические показатели, указывающие на наиболее типичный уровень развития изучаемого в совокупности признака и на степень вариации отдельных единиц совокупности вокруг этого уровня.

Большинство из них − именованные величины (средняя арифметическая, мода, медиана, среднее квадратическое отклонение), некоторые выражаются в процентах (коэффициент вариации) или, наконец, являются неименованными числами (варианса, коэффициент асимметрии). Но так как все они − статистические величины, то есть основаны на изучении массовых явлений, возникает очень важный теоретически и практически вопрос о том, насколько они достоверны.

Проблема достоверности занимает видное место в статистической теории. Выборочные и генеральные совокупности. Напомним, что генеральная совокупность − это вся подлежащая изучению совокупность данных объектов. В пределе она рассматривается как состоящая из бесконечно большого количества отдельных единиц. Та часть объектов, которая подвергается исследованию, называется выборочной совокупность или просто выборкой.

Оба типа совокупностей в общем характеризуются одинаковыми закономерностями случайной вариации. Для их характеристик могут быть вычислены статистические показатели: средняя арифметическая и среднее квадратическое отклонение. Среднюю арифметическую мы обозначали ранее символом x. Условимся теперь что x обозначает среднюю арифметическую выборочной совокупности. Среднюю арифметическую генеральной совокупности будем обозначать μ. Каково же соотношение между x и μ?

Допустим, что для совокупности, состоящей из 168 коров симментальской породы, была получена средняя арифметическая глубины груди 73,8 см. 168 коров представляют собой выборку из генеральной совокупности, охватывающей популяцию всех коров симментальской породы. Если бы мы взяли ряд выбора из популяции симментальской породы, то обнаружилось бы, что x этих выборок будут различными. Одни из x будут несколько больше чем 73,8 см, другие − меньше.

Очень важно, что распределение выборочных средних при достаточном их количестве близко к нормальному, поэтому к нему относятся указанные в предыдущей лекции закономерности. Оказывается, что отдельные значения средних арифметических выборок (x) варьируют вокруг средней арифметической генеральной совокупности. Вариация же выборочных средних вокруг μ может быть измерена своим средним квадратическим отклонением, своей сигмой. Эта сигма получила название сред¬ней ошибки или средней квадратической ошибки. Иногда ее называют также стандартной ошибкой. Именно она указывает на степень близости x и μ.

Вопрос 2. Формула для средней ошибки. Средняя ошибка для x может быть вычислена по формуле

$m_x = \delta/\sqrt{n}$

В знаменателе формулы под корнем n − объем выборочной совокупности. Это значит, что величина средней ошибки обратно пропорциональна численности выборочной совокупности.

В примере с глубиной груди у симментальских коров n = 168 и δ == 2,45. Отсюда средняя ошибка для средней арифметической глубины груди изученных 168 симментальских коров

$\delta = 2{,}45/168 = 0{,}17$

5.2 Средняя ошибка − ошибка выборочности Термин «ошибка» часто вводит в заблуждение начинающих, которые предполагают, что она является результатом недостаточной аккуратности в работе. Это не так. Средняя ошибка − это статистическая ошибка. Она не имеет ничего общего с ошибкой точности. Само собою разумеется, что все измерения (веса и промеров рыб, удоев коров и жирности их молока, настригов шерсти овец и ее длины) надо делать точно и добросовестно. Но статистические показатели для выборочной совокупности всегда имеют так называемые ошибки выборочности (их также называют ошибками репрезентативности), которые представляют собой среднюю величину расхождения между средними значениями изучаемых признаков в выборках и генеральной совокупности. Так как

$m_x = \delta/\sqrt{n}$

то, очевидно, что размер определяемой средней ошибки зависит от сигмы выборочной популяции и от ее объема. Чем лучше взята выборка и, чем больше ее размеры, тем меньше и средняя ошибка, тем меньше расхождение между значениями признаков в выборочных и генеральной совокупностях.

Биолог почти всегда имеет дело с выборками − и при проведении опытов с животными или растениями, и при изучении материала, взятого из природы, генеральные же совокупности остаются неизвестными. Поэтому он должен постоянно помнить о том риске, который сопутствует его выводам. Часто эти выводы основываются на изучении небольшого материала, поэтому полученные в опытах или наблюдениях статистические показатели могут иметь значительные статистические ошибки. Легко видеть, что в силу колеблемости выборочных средних вокруг средней генеральной совокупности один какой-либо опыт может дать результат, отклоняющийся от истинного на 2 или даже 3 ошибки. Но при значительном количестве опытов их результаты будут группироваться близко к центру распределения генеральной совокупности, т. е. к μ , что дает возможность уверенно сделать правильный вывод.

Некоторая погрешность органически присуща результатам всякого наблюдения, проведенного на основе выборки. Эту погрешность и измеряет средняя ошибка, которая поэтому и называется ошибкой выборочности (или, иначе, ошибкой репрезентативности). Вместе с тем совершенно необходимо, чтобы выборочная совокупность достаточно хорошо отображала генеральную совокупность, иначе суждение о генеральной совокупности по выборке будет

неправильным, несмотря на правильность статистических вычислений. Добиться правильного отображения генеральной совокупности можно при одном непременном условии − отборе вариант для выборки на основе случайности. Чем в большей степени этот отбор будет случайным, тем более правильными будут выводы, делаемые на основе выборочной совокупности. Именно тогда можно полагаться на результаты выборочного наблюдения.

Наиболее простой способ получения случайных выборок −отбирать экземпляры с помощью таблицы случайных чисел. На принципе случайности основываются различные схемы отбора вариант для выборки: случайная бесповторная выборка, когда взятые для выборки варианты уже не возвращаются обратно в генеральную совокупность, случайная повторная выборка с возвратом взятых для выборки вариант обратно в генеральную совокупность и т. д. Все они подробно рассматриваются в специальных пособиях.

5.3 Закон больших чисел. В связи между статистическими показателями выборочных и генеральных совокупностей выражается так называемый закон больших чисел. В наиболее общем виде этот закон заключается в том, что чем больше число n некоторых случайных величин, тем их средняя арифметическая ближе к средней арифметической генеральной совокупности, тем меньше разница между x и μ. По мере увеличения n вероятность осуществления приближения x к μ становится все большей, стремясь при $n = \infty$ к единице, т. е. к полной достоверности.

В этом заключается теорема одного из основоположников математической статистики русского математика П. Л. Чебышева.

Так как всякое явление, как правило, складывается из массы единичных, случайных явлений, то закон больших чисел выступает как реальный закон объективной действительности. Именно он лежит в основе нормального распределения вариант в вариационном ряду, т. е. распределения значений случайной переменной xi вокруг X, а также в основе распределения выборочных X вокруг μ.

Выборочные средние, для которых вычисляются средние ошибки, являются такими же случайными величинами, как и значения вариант в обычном вариационном ряду. С возрастанием объемов выборок их вариация вокруг генеральной средней становится все меньше. Средняя же арифметическая из всех выборочных средних должна быть равна средней арифметической генеральной совокупности, т. е. μ .

Таким образом, основное содержание закона больших чисел состоит в том, что при увеличении n отдельных выборок происходит взаимное погашение индивидуальных отклонений от некоторого уровня, характерного для всей совокупности в целом. Именно тогда проявляется закономерность, лежащая в основе биологического процесса. Закон больших чисел − одно из выражений диалектической связи между случайностью и необходимостью.

Распределение X малых выборок. Когда выборки являются достаточно большими по объему, распределение их средних арифметических является нормальным. Однако если выборки малы ($n < 30$), то возникает большое

сомнение в возможности суждения по таким выборкам о генеральной совокупности. В значение t может вкрасться значительная неточность.

В биологических исследованиях нередко приходится встречаться с выборочными совокупностями, состоящими из очень ограниченного количества вариант или наблюдений.

Возникает вопрос о том, каковы в этих случаях закономерности распределения выборочных средних арифметических. Ответ на него практически дал английский математик Госсет, который писал под псевдонимом Стьюдент. Поэтому изученное им распределение вероятностей получило название t-распределения по Стьюденту.

Теоретическое обоснование закона распределения, открытого Стьюдентом, было дано Фишером. Существенно то, что оно может быть использовано и при очень малых количествах вариант.

Критерий t по Стьюденту − Фишеру представляет собой следующее:

$t = X - \mu / m_x$

Оказалось, что распределение значений t отличается от нормального, при этом тем сильнее, чем меньше n. Поэтому и вероятности нахождения выборочных средних в пределах определенных значений n значительно снижаются по сравнению с нормальным распределением. В практической работе надо исходить из определенных уровней значимости, поэтому были составлены рабочие таблицы, по которым можно определять минимальное значение, обязательно требующееся для данной вероятности (табл. III, Рокицкий).

5.4 Определение необходимого объема выборочной совокупности. В практике биологических исследований часто возникает вопрос о том, сколько животных (или растений) данного вида надо взять, чтобы получить достаточно правильное представление о популяции вида (по изучаемому признаку). Вообще говоря, следует стремиться к большему числу наблюдений, однако очевидно, что численность выборки не может возрастать бесконечно. Она должна иметь какие-то рациональные границы, которые будут зависеть прежде всего от желаемой точности наблюдения, т. е. допустимого расхождения между средней арифметической (по данному признаку) выборки и средней арифметической генеральной совокупности, а также от заданной вероятности и от степени однородности популяции. Желаемая точность (обозначим ее Δ)− это возможное при принятой вероятности отклонение X от μ, т. е.

$\Delta = tm$.

А так как $m=\delta/n$. то $\Delta = t\,\delta/n$. Отсюда $n=t\,\delta / \Delta$

Значение t определяется ожидаемой вероятностью результата выборочного обследования. При p = 0,997 t должно быть равно 3. При p = 0,95 можно ограничиться t= 2. Величина Δ берется заранее. Так, например, изучая вес зайцев, можно принять, что желаемая точность должна быть в пределах 0,2 кг, т. е. Δ == 0,2 кг.

Несколько труднее решить вопрос о величине среднего квадратического отклонения изучаемой популяции вида, заранее неизвестной. В качестве ее приблизительной оценки можно взять сигму по данным проводившихся ранее

исследований или попытаться вычислить ее по максимальным и минимальным значениям изучаемого признака, имея в виду, что вариационный размах должен охватывать примерно шесть средних квадратических отклонений.

ЛЕКЦИЯ 6 СТАТИСТИЧЕСКАЯ ГИПОТЕЗА. РЕПРЕЗЕНТАТИВНОСТЬ ВЫБОРОЧНЫХ ПОКАЗАТЕЛЕЙ

6.1 Оценка достоверности статистических показателей с помощью средней ошибки. Оценка достоверности Х. Роль средней, или статистической, ошибки в статистическом анализе очень велика. С одной стороны, как было показано выше, она позволяет определить границы для показателей генеральной совокупности, например для у, а с другой стороны, дает возможность оценить степень достоверности самих статистических показателей, в частности средней арифметической данной выборочной совокупности.

Что же следует понимать под достоверностью средней арифметической? Фактическая средняя арифметическая всегда является выборочной. Поэтому для суждения о ее достоверности надо сравнить ее со средней арифметической генеральной совокупности. Мерилом достоверности явится нормированное отклонение, для вычисления которого можно использовать приведенную выше формулу.

Возникает вопрос о том, откуда же взять величину р.? Возможны два случая. В первом у. представляет собой определенную, отличающуюся от нуля, величину, значение которой можно примерно предположить по другим данным. Допустим, что изучали жирность молока 10 коров. Были получены следующие показатели; x= 3,7%; o =0,28%; m=0,09%. Если при этом ранее изучали жирность молока в других выборках и получали различные значения выборочных средних, то можно вычислить среднюю из этих средних. Допустим, что она оказалась равна 4,0%. Можно принять ее за мю. Тогда t=3,7-4,0/0,009=3,3

При малом n (= 8) следует проверить достоверность по табл. II.(Рокицкий) Вероятность достоверности (p= 0,987) вполне достаточная.

В общем можно сказать, что x, вычисленные для большинства биологических показателей даже на сравнительно малых по размерам выборочных совокупностях, чаще всего будут достаточно достоверными, если только ряд не слишком растянут. Однако может получиться иначе, если приходится оперировать экспериментальными данными, в которых фигурируют какие-либо условные или относительные величины, часть последних может иметь и отрицательный знак. Тогда установление достоверности x совершенно необходимо.

6.1 Нулевая гипотеза. Метод средней ошибки позволяет сравнивать между собой любые две группы животных или растений, например: две выборочные совокупности, взятые из природной, неизученной популяции; выборку из какой-то уже известной группы и группу, из которой эта выборка взята; опытную и контрольную группы при постановке опытов − и установить, насколько достоверны различия между их статистическими показателями (средними арифметическими, вариансами и др.).

Общие принципы сравнения основываются на анализе так называемой нулевой гипотезы. Согласно этой гипотезе, первоначально принимается, что

между данными показателями (или группами, на основе которых они получены) достоверного различия нет, т. е. что обе группы вместе составляют один и тот же однородный материал, одну совокупность. Статистический анализ должен привести или к отклонению нулевой гипотезы, если доказана достоверность полученных различий, или к ее сохранению, если достоверность различий не доказана, т. е. различия признаны случайными. Но так как все статистические показатели и различия между ними характеризуются определенными уровнями значимости, то отбрасывание нулевой гипотезы должно быть связано с принятием определенного уровня значимости. Так, если признан необходимым уровень значимости 0,01 и если вероятность достоверности данного статистического показателя или разницы между показателями не удовлетворяет этому условию, т. е. она ниже 0,99 (например, 0,97, 0,91, 0,88), то нет оснований для отбрасывания нулевой гипотезы. Ее надо считать правильной по крайней мере до тех пор, пока новые данные не дадут возможности ее опровергнуть, доказав, что существующие различия не являются чисто случайными.

Конечно, и в том случае, когда нулевая гипотеза считается опровергнутой, какой-то шанс, что она в действительности верна, остается. При уровне значимости 0,01 этот шанс составляет 1 на 100, т. е. в 1 % случаев отбрасывание нулевой гипотезы было ошибкой. Если достигнут уровень значимости не 0,01, а 0,001, то уверенность в том, что нулевая гипотеза действительно отвергнута правильно, резко возрастает (лишь 1шанс на 1000 случаев, что она все же верна). При P = 0,05 уверенность правильности вывода составляет лишь 95 случаев из 100, а в 5 возможен неправильный вывод.

Таким образом, если полученные данные характеризуются уровнем значимости $P<0,05$, то нет оснований отклонять нулевую гипотезу. Если $P > 0,05$ – нулевая гипотеза опровергнута.

Но значительно неопределеннее положение вещей, если результаты анализа или сравнения удовлетворяют уровню значимости 0,05, но не удовлетворяют уровню значимости 0,01. Надежное суждение оказывается невозможным. Очевидно, что в таких случаях должны быть проведены дополнительные опыты, чтобы решить, следует ли отбрасывать нулевую гипотезу. Вообще надо иметь в виду, что сохранение нулевой гипотезы еще не означает ее правильности. Может оказаться все же, что она неправильна. Сохранение же нулевой гипотезы оставляет вопрос открытым.

Приведенная выше оценка достоверности средней арифметической выборочной совокупности также являлась проверкой нулевой гипотезы. Согласно нулевой гипотезе, X=0. Надо было доказать, что X достоверно отличается от нуля. При достаточном доказательстве, удовлетворяющем принятому уровню значимости, нулевая гипотеза отбрасывается, т. е. признается достоверность X. Если это не у дается сделать, остается правильной нулевая гипотеза (недостоверность x) впредь до новых опытов.

6.3 Оценка достоверности разницы между средними арифметическими двух выборочных совокупностей. Если была получена разница между средними арифметическими двух генеральных совокупностей,

то, очевидно, не может стоять вопрос о статистической ошибке этой разницы. Эта разница всегда достоверна, даже если она и очень мала. Иное дело, если сравниваются две выборочные совокупности, например: две группы морских свинок, подвергавшихся воздействию химических веществ или физических факторов, две группы коров, сравниваемые по удою и взятые из одной породы, хозяйства н т. д. В этих случаях разница между средними имеет свою статистическую ошибку, с которой ее можно сравнить и установить, достоверна эта разница или нет. Нулевая гипотеза в данном случае будет сводиться к тому, что две изучаемые выборочные совокупности происходят из одной и той же генеральной совокупности и что разница между их средними арифметическими случайна, т. е. лежит в пределах ошибки выборочности.

Чтобы иметь право отвергнуть нулевую гипотезу, надо доказать, что разница между средними арифметическими достоверна, т.е. удовлетворяет требуемому уровню значимости. Для установления достоверности разницы между средними арифметическими надо воспользоваться нормированным отклонением. Нормированное отклонение примет следующую форму:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_{(\overline{x_1} - \overline{x_2})}}.$$

На самом деле формула для t должна быть несколько сложнее, а именно;

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_{(\overline{x_1} - \overline{x_2})}}.$$

Но, так как надо исходить из нулевой гипотезы о том, что две выборочные средние арифметические взяты из одной генеральной совокупности, то $\mu_1 = \mu_2$ и правая часть числителя обращается в нуль.

Числителем является разница между средними арифметическими двух групп (знак разницы не имеет значения). Ее можно обозначить сокращенно буквой d.. В знаменателе же−средняя ошибка этой разницы, т. е. $m_{x1} - m_{x2}$ или более сокращенно m_d. Тогда

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

Существует два способа определения средней ошибки разницы. Первый из них применяется, когда обе сравниваемые группы обладают достаточно большой численностью, большей чем по 30 особей в каждой. Средняя ошибка разницы определяется тогда по формуле

$$s_d = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}.$$

Допустим, что мы хотим сравнить по удою 2 группы коров. В одной группе n_1==50. В другой n_2=40. Средние удой и ошибка для первой группы: $X_1 \pm m_{x1}$== 2100 ± 120 кг; для второй группы: $X_2 \pm m_{x2}$= 2635 ±140 кг. Разница между средними удоями 2 групп

$$d = \bar{x}_2 - \bar{x}_1 = 2635 - 2100 = 535 \text{ кг.}$$

Ошибка разницы

$$s_d = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}} = \sqrt{140^2 + 120^2} = 184 \text{ кг.}$$

Таким образом, $d \pm m_d$ = 535 ± 184 кг, а t= 2,91,

По таблице нормального интеграла вероятности (табл. I, Рокицкий) находим, что в этом случае вероятность достоверности очень велика - 0,9963.

При отсутствии таблицы можно исходить из правила трех сигм: если разница превышает свою ошибку почти в три раза, она достоверна с вероятностью не менее 0,991. Но из сказанного выше очевидно, что в таком высоком значении t нет надобности. Если n > 30, то t=2,58 гарантирует достоверность разницы с вероятностью 0,99.

При сравнении двух групп с малыми, и, особенно с неодинаковыми обьемами, ошибка разницы определяется по формуле:

$$s_d = \sqrt{\frac{\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} \left(\frac{n_1 + n_2}{n_1 \cdot n_2} \right)}.$$

Смысл этой формулы заключается в том, что нельзя пользоваться просто готовыми средними ошибками, вычисленными заранее для двух сравниваемых групп, как это было при применении формулы, а нужно сначала сложить суммы квадратов отклонений по обеим группам, т.е. т. е. получить объединенную сумму квадратов отклонений, затем определить вариансу объединенных рядов (путем деления объединенной суммы квадратов на сумму чисел степеней свободы обеих групп) и, наконец, после умножения на $n_1 + n_2$ / n_1 x n_2 и извлечения квадратного корня получить ошибку разницы.

Для иллюстрации возьмем следующий дример. На двух группах крыс был поставлен опыт по сравнению влияния разных рационов на рост. Крысы 1 группы (12 шт.) получали рацион с высоким содержанием белка, крысы второй (7) − с низким. Привесы за 56 дней опыта для каждой крысы составляли (в г): первая группа − 134, 146, 104, 119, 124, 161, 107, 83, 113, 129, 97. 123;

вторая группа−70, 118, 101, 85, 107, 132,94.

После обработки данных с помощью одной из формул для сумм квадратов получим: $d = X_1 - X_2 = 19$ г.; $\sum(x - X_1)^2 = 5302$, $\sum(x - X_2)^2 = 2575$, тогда общая сумма квадратов равна 7584, а степени свободы df=17. Применив

указанные выше формулы получим t=1,89. По табл. III (Рокицкий) находим, что (при df = 17 и уровне значимости 0,05) t должно быть не менее 2,11 , полученное значение t ниже табличного. Для уточнения вероятности достоверной разницы воспользуемся табл. II. Из нее видно, что t = 1,89 соответствует вероятности 0,92, т,е. уровень значимости 0,08. Т.о. можно считать, что разные рационы не привели к разделению популяции крыс по привесам на две достоверно отличающиеся друг от друга популяции, иначе говорят нулевая гипотеза не может быть отвергнута. Конечно, опытные группы были слишком малы. Возможно, что при их увеличений будет получена более достоверная разница между группами крыс, находившимися на разных рационах кормления.

Лекция 7. ОСНОВЫ ДИСПЕРСИОННОГО АНАЛИЗА

7.1 Сущность и метод дисперсионного анализа. Ранее были рассмотрены методы оценки различия двух выборок путем сравнения их средних μ_1 и μ_2 и стандартных отклонений. В исследованиях часто приходится иметь дело не с двумя, а с большим числом выборок. Обычно эти выборки относятся к различным совокупностям. Например, это могут быть группы растений, получивших разные удобрения или уход, когда в опыте ставится цель статистически оценить эффект мероприятия. В начале 1950-х годов Р. А. Фишер разработал критерий и метод для такой оценки. Это привело к значительному последующему развитию теории планирования опыта и статистической оценки его эффекта.

Статистический смысл задачи по оценке эффекта мероприятия в многогрупповом опыте состоит в проверке значимости различия в групповых средних оцениваемого на основе сравнения дисперсий.

Для раскрытия сущности метода оценки эффекта мероприятия, т. е. дисперсионного анализа, рассмотрим сначала анализ нескольких выборок, взятых из общей совокупности. Такой опыт называют условным экспериментом.

Дж. У. Снедекор (1961) произвел 4 выборки (a=4) из общей совокупности данных по привесу 511 животных. Каждая из групп включала n=5 наблюдений (повторений). Средняя для совокупности μ=30, а дисперсия σ^2=100. Результаты опыта приведены в табл. 1.

Таблица 1. Привесы (в фунтах) 4 групп по 5 животных в группе.

Группа	Привес X	Сумма ΣX	Сред-ние μ	ΣX^2	$\frac{\Sigma(X)^2}{n}$	Σx^2
1	40,. 24, 46, 20, 35	165	33	5917	5445	472
2	29, 27, 20, 39, 45	160	32	5516	5120	396
3	11, 31, 17, 37, 39	135	27	4261	3645	616
4	17, 21, 28, 33, 21	120	24	3044	2880	164
По опыту в целом		580	29	18738	16820	1918

Данные таблицы позволяют получить три оценки дисперсии в совокупности σ^2==100. Первая оценка получается на основе всех 20 наблюдений.

$$\sigma = \frac{\sum x^2}{N-1} = \frac{1918}{19} = 100.9, (N = a \cdot n)$$

Вторая оценка получается из сумм квадратов внутри четырех групп. Она отражает варьирование «отдельных групп».

$$\sigma_1 = \frac{\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2}{a \cdot n - n} = \frac{472+396+616+164}{20-4} = 103$$

Групповые средние приводят к третьей оценке дисперсии совокупностей.

Средний квадрат средних будет равен:

$$\frac{(\mu_1-\mu)^2+(\mu_2-\mu)^2+(\mu_3-\mu)^2+(\mu_4-\mu)^2}{n-1}=\frac{(33-29)^2+(32-29)^2+(27-29)^2+(24-29)^2}{4-1}=18$$

Число 18 является оценкой $\sigma^2/5$, т. е. оценкой 20.

Каждая средняя представляет 5 наблюдений. Следовательно, третья оценка σ^2 будет равна $\sigma^2=18\cdot5=90$. Она основана на 4 групповых средних при n–11=4–1=3 степенях свободы. Сумма квадратов всех групповых средних составит 90*3=270.

Результаты произведенного подразделения общего варьирования на части и его анализ называют дисперсионным анализом (табл. 2).

Таблица 2. Дисперсионный анализ данных о привесе животных

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Объекты отдельных групп	16	1648	103
Групповые средние	3	270	90
Итого	19	1918	100,9

1 Сумма всех наблюдений:

2 ΣX =40+ 24+... +21=580.

3 Общая сумма квадратов:

$$\sum x^2=\sum X^2-\frac{\left(\sum X\right)^2}{N}=40^2+24^2+\ldots+21^2-\frac{580^2}{20}=1918$$

4 Сумма квадратов для групповых средних:

$$\frac{\sum\left(\sum X\right)^2}{n}-\frac{\left(\sum X\right)^2}{a\cdot n}=\frac{165^2+160^2+\ldots+120^2}{5}-\frac{580^2}{20}=17090-16820=270$$

Сравнение среднего квадрата групповых средних (90) и среднего квадрата для объектов внутри отдельных групп (103) показывает незначительное их расхождение.

Прежде чем делать окончательные выводы, приведем схему расчетов и таблицу анализа в общепринятом виде.

Результат пунктов 2, 3 вносят в таблицу и на их основе получают данные для объектов (табл. 3).

Таблица 3. Дисперсионный анализ данных р привесе животных (общепринятая форма)

Источник варьирования	ν	Σx^2	σ
Общее	19	1918	–
Групповые средние (факториальное)	3	270	90
Объекты отдельных групп (случайное)	16	1648	103

7.2 Дисперсионный анализ случайных выборок из двух или большего числа совокупностей. В большинстве приложений дисперсионного анализа изучаемые варианты опыта (например, данные дозы удобрения) влияют на средние. Группы становятся выборками из различных совокупностей. Считается, что эти совокупности имеют различные средние μ, но общую дисперсию, не зависимую от вариантов опыта. При дисперсионном анализе средний квадрат для объектов оценивает σ^2, как ранее было показано, но средний квадрат групповых средних оказывается преувеличенным в связи с различиями между μ. Табл. 4 и 5 представляют данные такого эксперимента.

Таблица 4. Высота тополевых саженцев, полученных из черенков особей с разными потомственными данными (от высоты каждого саженца отнято 50 см)

Группа	Высота, см						Сумма	Средняя
1	64	72	68	77	56	95	432	72
2	78	91	97	82	85	77	510	85
3	75	93	78	71	63	76	456	76
4	55	66	51	64	70	66	372	62

Вычисления:

1 $\sum X = 64^2 + 78^2 + \ldots + 66^2 = 1770$

2 $\sum x^2 = 64^2 + 78^2 + \ldots + 66^2 - \frac{1770^2}{24} = 3586.5$

3 Для средних:

$$\frac{432^2 + 510^2 + \ldots + 372^2}{6} - \frac{1770^2}{24} = 1636.5$$

Таблица 5. Дисперсионный анализ данных о высоте саженцев

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	23	3586,5	
Между группами (факториальное)	3	1636,5	545,5
Варианты (случайное)	20	1950	97,5

7.3 Критерий *F*–отношение дисперсий. Заключение о равенстве μ. Полученные данные, приводят к вопросу: обусловливается ли значительное различие между средними квадратами σ_1^2 и σ_2^2 обычным варьированием случайных выборок из одной совокупности или оно настолько велико, что следует его приписать влиянию выборочных средних. Соответствующая такой постановке вопроса нулевая гипотеза такова. H_0: $\mu_1 = \mu_2 = \ldots = \mu_0$ (средние групп одинаковы). Для ответа на подобные вопросы Р. А. Фишер предложил критерий − отношение дисперсий, распределение которого получено на основе случайных выборок из одной общей совокупности. Выше применение критерия *F* рассматривалось для проверки различия в дисперсиях двух малочисленных выборок.

Дж. У. Снедекор знакомит с распределением, полученным на основе 100

выборок по 10 наблюдений в каждой, взятых из уже упоминавшейся общей совокупности по привесу животных. Для каждой выборки по методу, изложенному выше, найдены *F:*

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

Распределение 100 значений *P* (число степеней свободы 9 и 90):

Интервал *F*	0–	0,25–	0,50–	0,75–	1,00–	1,25–
Число случаев	7	16	16	26	11	8
Интервал *F*	1,50–	1,75–	2,00–	2,25–	2,50–	2,75–
Число случаев	5	2	4	2	2	1

Распределение *F* несимметрично. 65 значений *F* меньше 1. Однако среднее значение $\overline{F} = 0.96$, т. е. близко к ожидаемой единице. 5% значений *F* превосходят 2,25, а 1 % выше 2,75.

Такой таблицей распределения *P* можно пользоваться для практических целей. Можно, например, сказать, что при выборках в 10 единиц значение *F*>2,75 может встретиться вследствие случайных причин 1 раз на 100 случаев.

На основе исследований Р. А. Фишера получено теоретическое распределение *F*-критерия для разных уровней значимости и для различного числа степеней свободы.

В таблицах приложений практически всех изданий по статистическим методам приведен 5%-ный уровень в распределении *F*.

При числе степеней свободы *v*=3 и *v*=20 имеем 5%-ный уровень критерия *F*=3,10. Полученное в опыте с саженцами отношение дисперсий $F = \frac{\sigma_1^2}{\sigma_2^2} = \frac{545.5}{97.5} = 5.6 > F_{0.05}$. Оно превышает даже $F_{0.01}$=4,9.

На основании сопоставления *F,* полученного в опыте, с табличными значениями можно сказать, что вследствие случайных причин из одной общей совокупности имеется менее одной возможности из 100 получить выборку, дающую значение *F* больше, чем наблюденное. Очевидно, что данные анализируемой выборки принадлежат к совокупности с различными μ. Следовательно, должен быть дан положительный ответ на поставленный выше вопрос о влиянии материнских наследственных качеств на рост нового поколения. Нулевая гипотеза H_0: $\mu_1 = \mu_2 = ... = \mu_0$ отвергается.

Такой вывод получен на основе установленного значимо более высокого варьирования между групповыми средними, измеряемого σ_1^2 по сравнению с варьированием высот растений внутри групп, измеряемым σ_2^2.

7.4 Дисперсионный анализ с классификацией по двум признакам. В рассмотренном выше примере с высотой саженцев была использована классификация только по одному признаку. Дисперсионный анализ применим и при классификации по нескольким признакам. Ниже рассмотрим пример группировки по двум признакам (факторам), значимость которых проверяют. Имеем следующие результаты наблюдений *X* относительно влияния удобрений (*B_1* и *B_2)* на почвах с разным качественным составом *(A_1* и *A_2)* (табл. 6).

Таблица 6. Результаты наблюдений *X*

Удобрение	Почва			
	A_1	A_2		
B_1	8, 12 $\mu_{11}=10;\ldots\sum x_{11}^2=8$	1, 3 $\mu_{21}=2;\ldots\sum x_{21}^2=2$	$\mu_{B1}=\frac{24}{4}=6$	$\sum x_B^2=0$
B_2	3, 4, 5 $\mu_{12}=4;\ldots\sum x_{12}^2=2$	6, 8, 10 $\mu_{22}=8;\ldots\sum x_{12}^2=8$	$\mu_{B2}=\frac{36}{6}=6$	
Вся группа	$\mu_{A1}=\frac{32}{5}=6.4$	$\mu_{A2}=\frac{28}{5}=5.6$	$\mu=6;\ldots\sum x^2=108$	
	$\sum x_A^2=1.6$			

Числа 8, 12, 3, 4, 5, 1, 3, 6, 8, 10 –значения результативного признака – *X*.

$\mu_{11}, \mu_{12}, \mu_{21}, \ldots\mu_{22}$ − частные средние в клетках; они получены по формуле:

$$\mu=\frac{\sum X_i}{n}$$

μ_{A1}, μ_{A2} −средние для 1 и 2-й групп почв;

μ_{B1}, μ_{B2}− то же, для соответствующих групп удобрений.

$\sum x_{11}^2 \ldots \sum x_{22}^2$ –суммы квадратов отклонений вариант от средних в клетках.

Проверяемые гипотезы. В опытах, подобных рассматриваемому, интересуют вопросы:

1 Различаются ли значимо по своему эффекту на рост растений почвы A_1 и A_2?
2 Значительно ли различен эффект двух удобрений B_1 и B_2?
3 Влияют ли удобрения на рост растений в одинаковой мере на обоях почвах?

Ответ на первый вопрос содержат средние для двух групп почв μ_{A1}=6,4 и μ_{A2}= 5,6.

Различия этого рода, связанные с неотъемлемыми качественными факторами среды, в литературе о дисперсионном, анализе называют *эффектом среды*.

Ответ на второй вопрос содержится в итогах двух строк $\mu_{B1} = \mu_{B2}$=6,0.

Различия, связанные с процессом производства, в данном случае с удобрением, называют *эффектом обработки*.

Ответ на третий вопрос следует искать в средних по клеткам $\mu_{11}, \mu_{12},$

$\mu_{21},\ldots\mu_{22}$. Видно, что удобрение B_1 на почве A_1 привело к средней μ_{11}= 10, тогда как на почве A_2 средняя μ_{21}=2. Удобрение B_2 характеризуется обратным указанному результатом: μ_{12}=4; μ_{22}=8. Ответ на третий вопрос выявляет взаимодействие, факторов *AB*.

В поисках заслуживающего доверия ответа на поставленные 3 вопроса выдвигаются 3 нулевые гипотезы:

гипотеза H_a– средние столбцов не отличаются друг от друга

гипотеза H_b– средние строк не отличаются друг от друга

гипотеза H_{ab}– взаимодействие *ab* отсутствует.

Компоненты общей суммы квадратов.

Общая сумма квадратов:

$$\sum x = \sum\left(8^2 + 10^2 + \ldots + 6^2 + 8^2 + 10^2\right) - \frac{60^2}{10} = 468 - 360 = 108.$$

Эту сумму квадратов разделяем на компоненты, измеряющие влияние двух испытываемых факторов, их взаимодействие, а также влияние большого числа случайных факторов, т. е. «компонента ошибки» – меры колебаний вследствие игры случая.

Сумма квадратов, соответствующая каждому из принципов классификации, вычисляется так же, как и при однофакторном комплексе, – как сумма квадратов отклонений каждой групповой средней от общей средней (с учетом веса n_i каждой средней):

для фактора почвы – $\sum x_A^2 = (6.4-6)^2 \cdot 5 + (5.6-6)^2 \cdot 5 = 1.6$

для фактора удобрения – $\sum x_B^2 = (6-6)^2 \cdot 5 + (6-6)^2 \cdot 5 = 0$.

«Компонент ошибки», независимый от двух положенных в основу классификации принципов, представляет собой сумму квадратов внутри всех четырех клеток.

$\sum\sum x^2 = 8 + 2 + 2)8 = 20$

Эта сумма квадратов, разделенная на соответствующее число степеней свободы, принимается в качестве меры влияния случайных факторов.

Сумма трех компонентов $\sum x_A^2 + \sum x_B^2 + \sum\sum x^2 = 1.6 + 0 + 20 = 21.6$.

Вычитая этот результат из общей $\sum x^2 = 108$, получим остаток равный 86,4. Этот остаток можно определить как «остаточную межгрупповую изменчивость». Он будет измерять взаимодействие *AB*.

Для степеней свободы найденных 4-х компонентов имеем следующие зависимости (*a* − число столбцов, *b* − число строк).

Степень свободы	
Между строками	*b–1*
Между столбцами	*a−1*
Для взаимодействия	*(a−1) (b−1)*
Внутри клеток	*N–ab*
Для итога	*N–1*

Дисперсионный анализ показан в табл. 7.

Данные анализа подтверждают нулевую гипотезу H_a*=0;* H_b*=0*, но не

согласуются с нулевой гипотезой $H_{ab}=0$. Эта гипотеза отвергается на *1%*-м уровне значимости, т. е. при вероятности p=0,99.

Таблица 7. Дисперсионный анализ

Источник варьирования	Степень свободы v	Сумма квадратов $\sum x^2$	Средний квадрат (дисперсия) σ	F
Фактор *A* (почвы)	1	1,6	1,6	$0,5<F_{005}=6,0$
Фактор *B* (удобрение)	1	0		
Взаимодействие *(AB)*	1	86,4	86,4	$26,2>F_{005}>F_{001}=13.4$
Внутри клеток (ошибка)	6	20	3,3	–
Итого	9	108		

Из этих результатов анализа делаем вывод, что 2 вида удобрения B_1 и B_2 тесно взаимодействуют с почвой, т. е. Производят эффект в зависимости от почв. Можно сказать и так почвы A_1и A_2 по-разному реагируют на удобрения.

Лекция 8. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

8.1 Понятие о корреляции. Изложенные в предыдущих главах методы анализа дают возможность изучать вариацию животных по каждому отдельному признаку − весу, промерам, плодовитости и т. д. Однако в ряде случаев важно знать, какова зависимость между вариацией двух или даже нескольких признаков изменяются ли два признака самостоятельно, независимо от друга или, может быть, вариация одного признака в какой степени связана с вариацией другого.

Существуют две категории связей, или зависимостей между признаками: функциональные и корреляционные, или статистические. При функциональных зависимостях каждому значению одной переменной величины соответствует одно вполне определённое значение другой переменной. Такие зависимости наблюдаются в математике и физике. Различные измерительные приборы основаны на функциональных зависимостях. Так, высота ртутного столбика в термометре дает точный и однозначный ответ о температуре воздуха или воды. Между радиусом окружности К и ее длиной С существуй функциональная зависимость по известной из элементарной геометрии формуле $C=2\pi R$. Иначе говоря, каждому значений Х соответствует строго определенное значение У. Точно так же накал нити в электрической лампочке определяется напряжением

Наряду с функциональными существуют статистические связи, при которых численному значению одной переменной соответствует много значений другой переменной. Например, между количеством внесенных на поле удобрений и урожайностью пшеницы существует бесспорная зависимость. Это не значит, что определенному количеству удобрений соответствует строго определенная величина урожая. В формировании урожая на данном участке поля много влияет факторов (состава и структуры почвы, способа внесения удобрений, глубина их заделки, различий в методах посева). Во многих исследованиях требуется изучить несколько признаков в их взаимной связи. Если вести такое исследование по отношению к двум признакам, то можно заметить, что изменчивость одного признака находится в некотором соответствии с изменчивостью другого.

В некоторых случаях такая зависимость проявляется настолько сильно, что при изменении первого признака на определенную величину всегда изменяется и второй признак на определенную величину, поэтому каждому значению первого признака всегда соответствует совершенно определенное, единственное значение второго признака. Такие связи называются функциональными.

Встречаются функциональные связи в физических и математических обобщениях. Площадь треугольника точно определяется его высотой и основанием, длина окружности − радиусом, скорость падения есть функция времени падения и ускорения силы тяжести, скорость протекания определенной химической реакции находится в зависимости от температуры.

Необходимо учесть, что функциональные связи встречаются только в

идеальных условиях, когда предполагается, что никаких посторонних влияний нет.

При изучении живых объектов − диких и культурных растений, животных, микроорганизмов − приходится иметь дело со связями другого рода. Живой организм развивается в связи с условиями его жизни, под действием бесконечно большого числа факторов, которые по-разному определяют развитие разных признаков.

У живых объектов связь между любыми двумя признаками настолько часто и сильно нарушается и модифицируется, что не всегда даже может быть легко обнаружена. У растений, животных и микроорганизмов связь между признаками обычно проявляется особым образом. Каждому определенному значению первого признака соответствует не одно значение второго признака, а целое распределение этих значений при вполне определенных основных показателях этого частного распределения− средней величины и степени разнообразия. Такая связь называется корреляционной связью или просто корреляцией.

Корреляционная связь, например, между весом животных и их длиной выражается в том, что каждому значению длины соответствует определенное распределение веса (а не одно значение веса), и с увеличением длины увеличивается и средний вес животных.

Корреляционная связь не является точной зависимостью одного признака от другого, поэтому она может иметь различную степень − от полной независимости до очень сильной связи. Кроме того, характер связи между разными признаками может быть различен. Поэтому возникла необходимость определять форму, направление и степень корреляционных связей.

По форме корреляция может быть прямолинейной и криволинейной, по направлению − прямой и обратной. Степень корреляции измеряется различными показателями, введенными для установления силы связи между количественными и качественными признаками. Такими показателями являются коэффициент корреляции r, корреляционное отношение η.

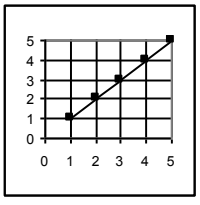
Изобразить корреляционную связь двух признаков можно тремя способами:

- При помощи корреляционного ряда, состоящего из ряда пар значений, из которых одно относится к первому признаку, а другое в этой паре относится ко второму признаку, связанному с первым. На рис. 7.1 показаны схемы корреляционных рядов при пяти степенях корреляционной связи.

5					**1**
4				**1**	
3			**1**		
2		**1**			
1	**1**				
	1	2	3	4	5

X_1	1	2	3	4	5
X_2	1	2	3	4	5

Прямая полная связь; r=+1,0



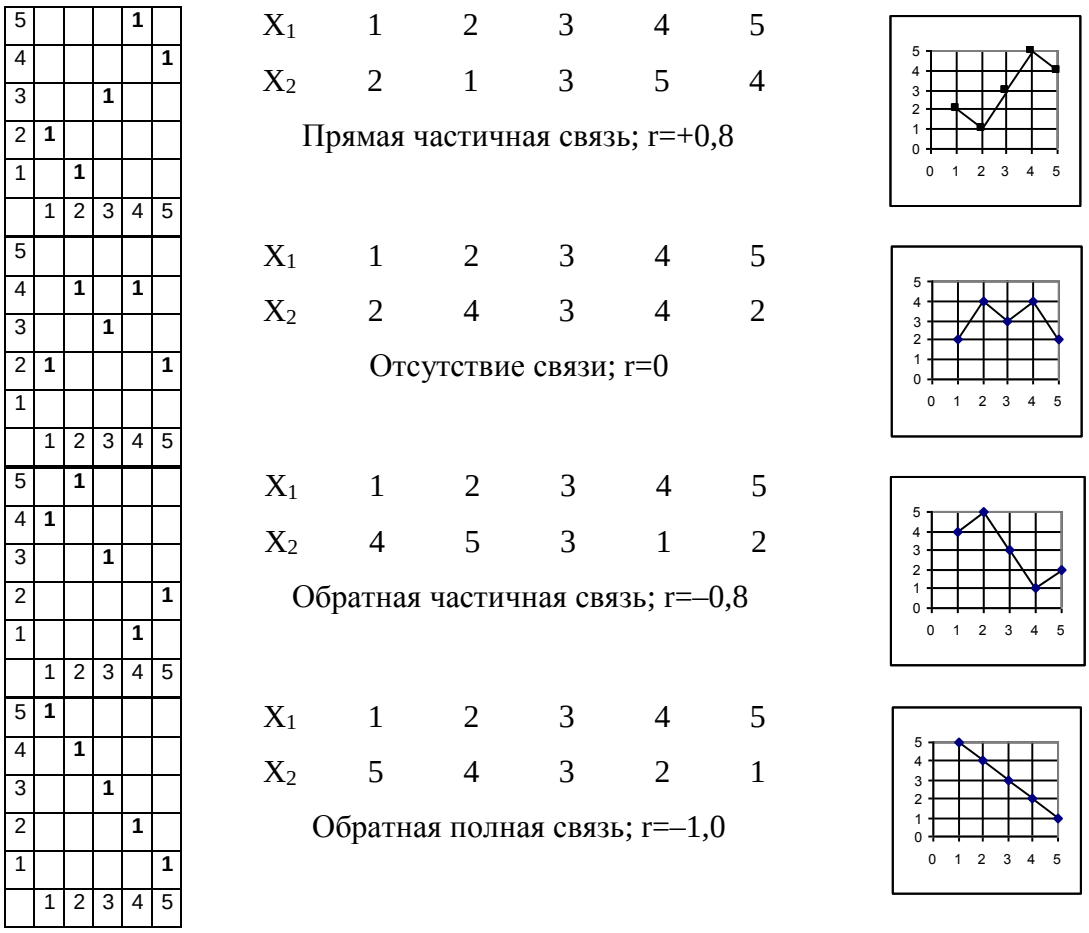


Рис. 1 Схема прямолинейных корреляционных связей

- При помощи корреляционной решетки, в которой каждой особи соответствует определенная клетка. На рис. 7.1 показана схема корреляционных решеток для пяти степеней корреляционной связи между двумя признаками Значения первого признака нанесены по оси абсцисс, значения второго − по оси ординат.
- При помощи линии регрессии, абсциссы которой пропорциональны значениям первого признака, а ординаты − значениям второго признака, корреляционно связанного с первым. На рис. 7.1 показаны схемы линий регрессии для пяти степеней корреляционной связи между двумя признаками.

8.2 Коэффициент корреляции. Коэффициент корреляции измеряет степень и определяет направление прямолинейных связей.

Прямолинейная связь между признаками − это такая связь, при которой равномерным изменениям первого признака соответствуют равномерные (в среднем) изменения второго признака при незначительных и беспорядочных отклонениях от этой равномерности. Например, при увеличении длины тела на каждый сантиметр ширина увеличивается в среднем на 0,7 см

При графическом изображении прямолинейных связей (см рис. 7.1) (если по оси абсцисс отложить значения первого признака, по оси ординат − второго и полученные точки соединить) получается прямая или такая кривая, среднее течение которой проходит по прямой

При изображении прямолинейных корреляционных связей в форме корреляционных решеток (см рис. 1) частоты внутри располагаются в форме эллипса Большая ось этого эллипса проходит или по диагонали от угла наименьших значений (при положительной корреляционной связи), или по диагонали от угла, где сходятся наименьшие значения одного признака и наибольшие значения другого, к противоположному углу (при отрицательной корреляционной связи).

При измерении степени связи между разными признаками приходится сравнивать величины, выраженные в разных единицах измерения. Например, при измерении связи между весом животного и его длиной надо сопоставить килограммы веса с сантиметрами длины. В других случаях изменения объема сопоставляются с изменениями возраста, изменения веса руна в килограммах с изменениями содержания в нем жиропота в процентах, длина ног в сантиметрах со скоростью бега в минутах и т. д.

Проводить такие сравнения оказалось возможным путем использования нормированного отклонения, вычисляемого по формуле:

$$\bar{x}_i = \frac{X_i - \mu}{\sigma}.$$

Нормированное отклонение служит универсальной и неименованной мерой развития признаков. Эти свойства нормированного отклонения и позволили сконструировать основной показатель корреляционной связи − коэффициент корреляции.

Основная формула, которая вскрывает сущность этого показателя, имеет совсем простую структуру:

$$r = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{\nu}$$

где r – коэффициент корреляции;

$\bar{x}_1 \cdot \bar{x}_2$ – нормированные отклонения дат по первому и второму признаку;

ν – число степеней свободы, равное в данном случае числу сравниваемых пар без одной.

Сумма произведений нормированных отклонений, входящая в формулу для коэффициента корреляции, обладает следующими тремя особыми свойствами

Если оба признака изменяются параллельно, то сумма произведений их нормированных отклонений дает положительную величину. Если при увеличении одного признака другой уменьшается, то приходится умножать положительные числа на отрицательные и вся сумма произведений нормированных отклонений дает отрицательную величину. Поэтому коэффициент корреляции может определять направление связи: при прямых связях он положителен, а при обратных связях отрицателен.

При полных связях, когда изменения обоих признаков строго соответствуют друг другу и корреляционная связь превращается в функциональную, сумма произведений нормированных отклонений становится равной числу степеней свободы:

$$\sum \bar{x}_1 \cdot \bar{x}_2 = \nu = n - 1$$

Поэтому максимальное значение коэффициента корреляции равно 1; для положительных, или прямых связей:

$$r_{\max} = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{\nu} = \frac{+\nu}{\nu} = +1.0$$

для отрицательных, или обратных связей:

$$r_{\min} = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{\nu} = \frac{-\nu}{\nu} = -1.0$$

- При полном отсутствии корреляционной связи между признаками сумма произведений нормированных отклонений равна нулю, и поэтому коэффициент корреляции в этих случаях тоже равен нулю:

$$r_{\min} = \frac{\sum \bar{x}_1 \cdot \bar{x}_2}{\nu} = \frac{0}{\nu} = 0$$

Предельные значения коэффициента корреляции (*r=+1,0; r=0,0; r= –1,0*) на практике встречаются крайне редко.

Пять основных видов прямолинейной корреляционной связи, соответствующие коэффициентам корреляции +1,0; +0,8; 0,0; –0,8 и –1,0, показаны на рис. 7.1.

Основная формула коэффициента корреляции хорошо вскрывает сущность этого показателя, но для работы крайне неудобна, особенно при многочисленных группах. Поэтому разработаны разнообразные рабочие формулы для практических расчетов в разных условиях − для малых и больших групп при малозначных и многозначных вариантах.

Все эти формулы дают одинаковый результат и применение любой из них обусловливается только удобством и простотой необходимых вычислений.

Наиболее приемлемы в биологических работах две формулы, предложенные дли малых групп:

$$r = \frac{\sum X_1 \cdot X_2 - \frac{\sum X_1 \sum X_2}{n}}{\sigma_1 \cdot \sigma_2}$$

где X_1, X_2 − даты первого и второго признаков; N− число сравниваемых пар дат, или число объектов, у которых измерено по два признака;

σ_1, σ_2− стандартные отклонения по первому признаку и по второму признаку.

Применяется коэффициент корреляции в тех случаях, когда необходимо знать направление и силу связи между признаками, причем заранее известно, что эта связь может считаться прямолинейной, или когда требуется выяснить степень именно прямолинейной связи. При этом лучше проводить два этапа исследования: 1) рассмотрение корреляционной решетки; 2) расчет коэффициента корреляции или по этой же решетке, или непосредственно по датам.

Уже самый вид корреляционной решетки позволяет приблизительно установить направление и степень прямолинейных связей, а также характер криволинейных связей. При известном опыте по виду корреляционной решетки

можно получить первое представление об особенностях и силе связи между изучаемыми признаками. Облегчает решение этой задачи схема степеней прямолинейной корреляции, показанная в табл. 7.1. В этой схеме приведены стандартные корреляционные распределения 50 особей при различных степенях прямолинейной связи по девяти градациям от *r=+1,0 до r=-1,0.*

Схемой степеней прямолинейной корреляции можно пользоваться как эталоном для первоначального ориентировочного отнесения изучаемой связи к одной из условных степеней («сильная», «средняя», «слабая») только по одному виду корреляционной решетки. В некоторых случаях такая грубая оценка бывает достаточна для выяснения предварительных вопросов исследования.

Таблица 1. Схема степеней прямолинейной корреляции

Прямая корреляция сильная

			2	2
		5	2	2
	5	8	5	
2	5	5		
2	2			

r=+0.75

Прямая корреляция средняя

		1	2	1
	2	4	4	2
1	4	8	4	1
2	4	4	2	
1	2	1		

r=+0.5

Прямая корреляция слабая

	1	1	1	1
1	2	3	5	1
1	3	10	3	1
1	5	3	2	1
1	1	1	1	

r=+0.25

Прямая корреляция полная

				4
			12	
		18		
	12			
4				

r=+1.0

Отсутствие корреляции

	1	2	1	
1	3	4	3	1
2	4	6	4	2
1	3	4	3	1
	1	2	1	

r=+0.0

Обратная корреляция полная

4				
	12			
		18		
			12	
				4

r=-1.0

Обратная корреляция слабая

1	1	1	1	
1	5	3	2	1
1	3	10	3	1
1	2	3	5	1
	1	1	1	1

r=-0.25

Обратная корреляция средняя

1	2	1		
2	4	4	2	
1	4	8	4	1
	2	4	4	2
		1	2	1

r=-0.5

Обратная корреляция сильная

2	2			
2	5	5		
	5	8	5	
		5	5	2
			2	2

r=-0.75

8.3 Ошибка коэффициента корреляции

Как и всякая выборочная величина, коэффициент корреляции имеет свою ошибку репрезентативности, вычисляемую для больших выборок по формуле:

$$s_r = \frac{1-(\bar{r})^2}{\sqrt{n-1}},$$

Где $\bar{r}$ −коэффициент корреляции в генеральной совокупности, из которой взята выборка;

n − численность выборки, т. е. число пар значений, по которым

вычислялся выборочный коэффициент корреляции.

Поскольку в числителе формулы ошибки выборочного коэффициента корреляции стоит квадрат генерального коэффициента корреляции, то эта формула может применяться лишь в исключительных случаях, когда заранее известна или предполагается степень корреляции в генеральной совокупности.

Пример. Для проверки гипотезы о том, что коэффициент корреляции между детьми и родителями r =+0,5, была сопоставлена плодовитость 226 лисиц и их дочерей в соответствующем возрасте и в сходных условиях. Коэффициент корреляции оказался равным +0,45. Подтверждает или опровергает этот результат гипотезу?

В данном случае разность между выборочным и генеральным коэффициентами *d=+0,45–(+0,50)=–0,05*, а ее ошибка равна ошибке выборочного коэффициента, так как генеральные величины не имеют ошибок репрезентативности. Для вычисления ошибки коэффициента корреляции имеется возможность применить точную формулу с генеральным коэффициентом в числителе:

$$s_r = \frac{1-0.5^2}{\sqrt{225}} = \frac{0.75}{15} = 0.05$$

Оказалось, что критерий достоверности разности $t_{(r-r)} = \frac{0.05}{0.05} = 1$ не превышает даже первого порога достоверности ($t_1=2,0$ $\beta_1 = 0,95$).

Гипотеза в данном исследовании не опровергнута, так как эмпирический коэффициент корреляции недостоверно отличается от гипотетического.

В большинстве исследований значение коэффициента корреляции в генеральной совокупности неизвестно, поэтому вместо точного значения ошибки коэффициента корреляции берут приближенное значение:

$$s_r = \frac{1-r^2}{\sqrt{n-1}}$$

Где *r*−выборочное значение коэффициента корреляции,

n−число сравниваемых пар данных или число объектов, у которых измерены два признака.

Ошибка коэффициента корреляции используется для определения: 1) достоверности выборочного коэффициента корреляции; 2) доверительных границ генерального коэффициента корреляции; 3) достоверности разности двух выборочных коэффициентов корреляции; 4) достоверности разности между выборочным и генеральным коэффициентом корреляции.

8.4 Достоверность выборочного коэффициента корреляции

Критерий выборочного коэффициента корреляции определяется по формуле:

$$t_r = \frac{r}{s_r} \geq t_{st}\{\nu = n-2\}$$

где t_{st} − критерий достоверности коэффициента корреляции;

r− выборочный коэффициент корреляции;

n−число коррелированных пар дат;

t_{st} – стандартное значение критерия Стьюдента, находимое по таблице для установленного числа степеней свободы и порога вероятности безошибочных прогнозов.

При $t \geq t_{st}$ коэффициент корреляции достоверен. В этом случае с определенной вероятностью можно считать, что между коррелируемыми признаками имеется связь и в генеральной совокупности такая же по знаку, какая получилась в выборке (прямая или обратная).

При $t < t_{st}$ выборочный коэффициент корреляции недостоверен, что не дает возможности сделать какое-либо заключение о связи признаков в генеральной совокупности. Для выяснения этого вопроса требуется провести повторные исследования на более многочисленном материале.

Пример. При проверке гипотезы о связи крупноплодности с жирномолочностью был рассчитан коэффициент корреляции между процентом жира в молоке у 50 коров и весом при рождении телят от этих же коров. Получено:

коэффициент корреляции $r = +0{,}21$;

его ошибка $s_r = \sqrt{\frac{1-0.21^2}{50-2}} = 0.14$;

критерий достоверности:

$t_r = \frac{0.21}{0.14} = 1.5$; $\nu = 48$; $t_{st} = \{2{,}0\text{-}2{,}7\text{-}3{,}5\}$.

Выборочный коэффициент оказался явно недостоверным. На основе проведенного исследования нельзя ожидать связи между крупноплодностью и жирномолочностью у всех коров вообще.

Определение достоверности коэффициента корреляции можно значительно упростить, используя свойства особой функции предложенной Фишером:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

При помощи этой функции можно заранее определить, при каком объеме выборки коэффициент корреляции определенной величины будет достоверен по требуемому порогу вероятности безошибочных прогнозов, по следующей формуле:

$$\hat{n} = \frac{t^2}{z^2} + 3$$

где $\hat{n}$ – количество пар значений, достаточное для достоверности выборочного коэффициента корреляции,

t – критерий Стьюдента для каждого из трех порогов вероятности безошибочных прогнозов ($\beta_1 = 0{,}95$, $\beta_2 = 0{,}99$, $\beta_3 = 0{,}999$), для больших групп: $t_1 = 1{,}96$, $t_2 = 2{,}58$, $t_3 = 3{,}30$.

z – функция Фишера $z = \frac{1}{2}\ln\frac{1+r}{1-r}$

По этой формуле рассчитано значение z и количество пар значений, достаточное для достоверности выборочного коэффициента корреляции для каждого из трех порогов вероятности безошибочных прогнозов.

В примере в выборке объемом n = 50 получен коэффициент корреляции $\boldsymbol{r}$= +0,21.

При r = 0,21, рассчитаны три числа: 87−149−242. Это значит, что выборочный коэффициент корреляции, равный r=0,21, может стать достоверным в том случае, если объем выборки (число коррелируемых пар данных) будет: для первого порога вероятности 87, для второго− 149, для третьего − 242. Так как фактический объем выборки n =50 далеко не достигает первого, максимальною порога, то полученный коэффициент корреляции оказался недостоверным, что было найдено и обычным способом.

8.5 Доверительные границы коэффициента корреляции

Доверительные границы генерального значения коэффициента корреляции находятся общим способом по формуле:

$\bar{r} = r \pm \Delta$,

где $\bar{r}$ и r − генеральное и выборочное значения коэффициента корреляции;

$\Delta = t^* s_r$ − возможная погрешность при определении генерального параметра;

t_{st} − критерий Стьюдента при числе степеней свободы $\nu = n-2$;

s_r − ошибка коэффициента корреляции.

Пример. При разработке способов определения веса устриц определенного вида по их длине было измерено и взвешено 200 экземпляров и определен коэффициент корреляции между весом и длиной $r = +0{,}85$.

Ошибка этого коэффициента

$$s_r = \sqrt{\frac{1-0.85^2}{200-2}} = 0.037$$

Число степеней свободы и критерий Стьюдента

ν= n− 2 ==198, t_{st} = {2,0 − 2,6 − 3,3}.

Возможная погрешность при прогнозе генерального параметра

$\Delta = t^* s_r$ = 2,0*0,037 = 0,074.

Доверительные границы:

r =+0.85±0.074[−не более + 0.85+0.074 = 0.92; −не менее 0.85-0.074 = 0.78]

Даже минимальная граница (гарантированный минимум) оказалась достаточно высокой. Это указывает на возможность практического использования вскрытой закономерности путем разработки формулы регрессии для определения веса устриц по их длине с практически достаточной точностью.

Достоверность разности двух коэффициентов корреляции

Достоверность разности коэффициентов корреляции определяется так же, как и достоверность разности средних, по обычной формуле

$$t_d = \frac{d}{s_d} \geq t_{st}\{\nu = n_1 + n_2 - 4\},$$

где t_d −критерий достоверности разности коэффициентов корреляции;

$d = r_1 - r_2$−разность коэффициентов корреляции;

$s_d = \sqrt{s_1^2 \cdot s_2^2}$ −ошибка разности, равная корню квадратному из суммы квадратов ошибок обоих сравниваемых коэффициентов корреляции; $s^2 = \frac{1-r^2}{n-2}$;

t_{st} − стандартные значения критерия Стьюдента;

ν − число степеней свободы для разности коэффициентов корреляции, равное сумме чисел степеней свободы обоих коэффициентов:

$\nu = n_1 - 2 + n_2 - 2 = n_1 + n_2 - 4$.

Пример. При разработке способов определения высоты дерева по его обхвату (на высоте груди измеряющего) получены коэффициенты корреляции между этими признаками для двух пород деревьев:

$n_1 = 200$, $r_1 = 0{,}60$, $s_1^2 = \frac{1-0.6^2}{198} = 0.0032$;

$n_2 = 150$, $r_2 = 0{,}80$, $s_2^2 = \frac{1-0.8^2}{148} = 0.0024$.

Для выяснения возможности применения единой формулы пересчета обхвата на высоту потребовалось выяснить: достоверно ли различие связи высоты с обхватом между двумя изучаемыми породами деревьев. Получены следующие результаты:

$d = 0{,}80 - 0{,}60 = 0{,}20$;

$s_d^2 = 0{,}0032 + 0{,}0024 = 0{,}0056$, $s_d = \sqrt{0{,}0056} = 0{,}075$;

$t_d = \frac{0.200}{0.075} = \underline{\underline{2.7}}$, $\nu = 200 + 150 - 4 = 346$, $t_{st} = \{2{,}0 - 2{,}6 - 3.3\}$.

Оказалось, что сравниваемые породы достаточно достоверно (по второму порогу вероятности) различаются по степени связи между высотой и обхватом дерева. Поэтому для этих пород нельзя пользоваться единой формулой пересчета обхвата на высоту.

Лекция 9. РЕГРЕССИОННЫЙ АНАЛИЗ

9.1 Многообразие методов изучения связи. Известно, что различные зависимости широко распространены как в органической, так и в неорганической природе. Их изучение проводилось уже давно и привело к разработке большого количества методов их математической характеристики. И первым из них являлся разобранный в предыдущей лекции корреляционный метод, или метод корреляций.

Коэффициент корреляции указывает лишь на степень связи в вариации двух переменных величин или, как иногда говорят, на меру тесноты этой связи, но не дает возможности судить о том, как количественно меняется одна величина по мере изменения другой. На этот последний вопрос позволяет ответить другой метод определения связи между варьирующими признаками, носящий название метода регрессии.

В современной статистике, в том числе биологической, коэффициентами корреляции пользуются реже, чем прежде, Метод же регрессии приобретает все большее значение. Анализ взаимоотношения двух изменчивых величин с помощью метода регрессии часто может дать очень ценные результаты, особенно в практическом отношении. В некоторых случаях для освещения различных сторон вопроса надо применять и корреляционный, и регрессионный методы анализа.

При простой корреляции изучается зависимость между изменчивостью двух признаков x и y. С помощью регрессии ставится дополнительно задача установить, как количественно изменяется одна величина при изменении другой на единицу. Так как изменчивых величин две, то регрессия, очевидно, может быть двусторонней:

1. определение изменения y по изменению x
2. определение изменения x по изменению y.

В этом заключается главное отличие метода регрессии от метода корреляции. Регрессия может быть выражена несколькими способами:

- путем построения так называемых эмпирических линий регрессии,
- путем составления уравнений регрессии и построения теоретических линий регрессии,
- с помощью вычисления коэффициента регрессии.

Первые два способа позволяют выразить регрессию графически. Для построения эмпирических линий регрессии можно воспользоваться обычной корреляционной решеткой. Но в ней следует заменить границы классов средними значениями классов.

9.2 Коэффициент прямолинейной регрессии

Прямолинейная корреляция отличается тем, что при этой форме связи каждому из одинаковых изменений первого признака соответствует вполне определенное и тоже одинаковое в среднем изменение другого признака, связанного с первым или зависящего от первого.

Та величина, на которую в среднем изменяется второй признак, при изменении первого на единицу измерения, называется коэффициентом

регрессии. Рассчитывается он по формуле:

$$R_{2/1} = \frac{\sigma_2}{\sigma_1} \cdot r_{12},$$

где $R_{1/2}$ − коэффициент регрессии второго признака по первому;

σ_2 − среднее квадратическое отклонение второго признака, который изменяется в связи с изменением первого;

σ_1 − среднее квадратическое отклонение первого признака, в связи с изменением которого изменяется второй признак;

r_{12} − коэффициент корреляции между первым и вторым признаками.

Ошибка коэффициента регрессии равна ошибке коэффициента корреляции, умноженной на отношение сигм:

$$s_R = \frac{\sigma_2}{\sigma_1} \cdot s_r = \frac{\sigma_2}{\sigma_1} \cdot \sqrt{\frac{1-r^2}{n-2}}.$$

Критерий достоверности коэффициента регрессии равен критерию достоверности коэффициента корреляции:

$$t_R = \frac{R}{s_R} = \frac{\frac{\sigma_2}{\sigma_1} \cdot r_{12}}{\frac{\sigma_2}{\sigma_1} \cdot s_r} = \frac{r}{s_r} = t_r,$$

Пример. Для разработки способа определения веса лошадей без взвешивания по обхвату груди было взвешено 1618 лошадей и у каждой из них измерен обхват груди. Получены следующие показатели: x − обхват груди, n = 1618, μ_x=174 см, σ_x=7,9 см;

y− вес, n=1618, μ_y=424 кг, σ_y = 56,8 кг.

Коэффициент корреляции $r_{x/y}$=*+0,89±0,011*.

Коэффициент регрессии веса по обхвату равен:

$$R_{y/x} = \frac{\sigma_y}{\sigma_x} \cdot r_{y/x} = \frac{56.8}{7.9}(+0.89) = +6.4.$$

Ошибка коэффициента регрессии веса лошадей по обхвату их груди равна:

$$s_R = \frac{\sigma_y}{\sigma_x} \cdot s_r = \frac{56.8}{7.9} \cdot 0.011 = 0.08.$$

Достоверность этого коэффициента регрессии определяется следующим образом:

$$t_R = \frac{6.4}{0.08} = \underline{\underline{80.0}},\ \nu=1618-2=1616,$$

t_{st}={2.0−2.6−3.3}

Возможная максимальная погрешность при прогнозе генерального параметра

Δ =t_m = 2,0*0,08 = 0,16.

Доверительные границы

$$\overline{R}_{y/x} = = +6{,}4 \pm 0{,}16 = \{6{,}24 - 6{,}56\}.$$

Таким образом, можно ожидать, что при увеличении (или уменьшении)

обхвата груди на 1 см вес лошадей увеличится (или уменьшится) в среднем на *R*=+6,4 кг при гарантированном минимуме изменения +6,24 кг и возможном максимуме +6,56 кг, если учитывать изменения признаков в обе стороны от их средней величины.

Коэффициент прямолинейной регрессии показывает, на сколько от своей средней отклоняется второй признак, если первый признак от своей средней отклонился на единицу измерения. Это можно выразить следующей формулой:

$(X_2-\mu_2)=R_{2/1}\ (X_1-\mu_1)$

Обозначая X_1 через x, X_2 через y, $R_{1/2}$ через b и произведя необходимые преобразования этого выражения, можно получить рабочую формулу прямолинейной регрессии:

$y=a+bx$

$$\left\{\frac{a=\mu_y-b\mu_x}{b=R_{y/x}}\right\}.$$

По этой формуле, зная значение x (аргумент), можно определить значение y (функция) без непосредственного его измерения: нужно аргумент x помножить на коэффициент регрессии и к полученному произведению прибавить (или отнять) свободный член a.

Для предыдущего примера (определение веса лошадей по обхвату груди) уравнение регрессии может быть выведено следующим образом:

$a = \mu_y - R_{y/x} * \mu_x = 424 - (+6{,}4)\ 174 = -690,$

$b = R_{y/x} = +6{,}4,$

$y = a + bx = -690 + 6{,}4x = 6{,}4x - 690.$

Следовательно, чтобы определить (без взвешивания живой вес лошади по этому способу, надо обхват груди лошади умножить на постоянный коэффициент 6,4 и из полученного произведения вычесть постоянное число –690.

На основе уравнения прямолинейной регрессии можно заранее рассчитать значение функции для каждого значения аргумента.

По обхвату груди можно определить живой вес лошадей.

Если эти цифры нанести на график, по оси абсцисс которого отложить через равные интервалы значения аргумента (обхвата), а по оси ординат − значения функции (веса), то получится номограмма для определения веса лошадей без взвешивания и без вычислений.

Ошибки элементов уравнения прямолинейной регрессии.

В уравнении простой прямолинейной регрессии:

$y_x = a + bx$

возникают три ошибки репрезентативности.

4 Ошибка коэффициента регрессии:

$$s_b = \frac{\sigma_y}{\sigma_x}\cdot\sqrt{\frac{1-r^2}{n-2}} = \frac{\sigma_y}{\sigma_x}\cdot s_r$$

5 Ошибка уравнения регрессии, т. е. ошибка средней величины функции для каждого значения аргумента:

$$m_{y_x} = \sigma_y \cdot \sqrt{\frac{1-r^2}{n-2}}$$

По данным примера

$s_{y_x} = 56.8 \cdot 0.011 = 0.62$.

Следовательно, максимальная погрешность в определении уровня точек линии регрессии при первом пороге вероятности безошибочных прогнозов (β_1 = 0,95, t_1=2,0) будет равна:

$\Delta = t*s = 2*0{,}62 = \pm 1{,}24$ кг.

6 Ошибка индивидуальных определений функции:

$s_y = \sigma_y \sqrt{1-r^2}$

Для примера:

$s_y = 56.8\sqrt{1-0.89^2} = 26.2$.

Следовательно, индивидуальная погрешность в определении веса лошадей по обхвату груди по найденной формуле регрессии, принимая первый порог вероятности безошибочных прогнозов (β_1 = 0,95, t_1 = 2,0), в крайних случаях не будет превышать

$\Delta = 2 * 26 = \pm 52$ кг.

ЛИТЕРАТУРА

Рекомендуемая литература (основная):

1. Лакин Г.Ф. «Биометрия». М. Высшая школа, 1990.
2. Бейли Н. «Математика в биологии и медицине». М., Мир, 1970.
3. Урбах В.Ю. Статистический анализ в биологических и медицинских исследованиях. – М.: Медицина, 1975.
4. Гланц С. Медико-биологическая статистика. М.: Практика, 1998.
5. Мюллер П., Нойман П., Шторм Р. Таблицы по математической статистике. – М.: Финансы и статистика, 1982.
6. Носов В.Н. «Компьютерная биометрика». МГУ, 1990.
7. Боровиков В.П. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. – СПб.: Питер, 2003.
8. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. – М.: МедиаСфера, 2002.
9. Рокицкий П. Ф. Биологическая статистика, 1976(1980).

Рекомендуемая литература (дополнительная):

1. Плохинский Н.А. Биометрия. - М.: МГУ, 1970. – 368 с.
2. Свалов Н.Н. Вариационная статистика. - М.: Лесная промышленность, 1977. – 177 с.
3. Справочник по прикладной статистике. В 2-х т. / Под ред. Э. Лойда, У. Ледермана, Ю.Н. Тюрина. – М.: Финансы и статистика, Т.1: 1989; Т.2: 1990.
4. Глас Дж., Стенли Дж. Статистические методы в педагогике и психологии. – М.: Прогресс, 1976.
5. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: ИНФРА-М, Финансы и статистка, 1995.
6. Боровиков В.П. Популярное введение в программу STATISTICA.- М.: КомпьютерПресс, 1998.
7. Козлов А.Ю., Мхитарян В.С., Шишов В.Ф. Статистические функции MS Excel в экономико-статистических расчетах: Учеб. пособие для вузов. – М.: ЮНИТИ-ДАНА, 2003.
8. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистические методы в медико-биологических исследованиях с использованием Excel. – К.: МОРИОН, 2000.
9. Макарова Н.В., Трофимец В.Я. «Статистика в Excel». М. Финансы и статистика, 2002.
10. Глотов Н. В., Животовский Л. А.,Хованов Н. В., Хромов-Борисов Н. Н. Биометрия. Л., 1982.
11. Терентьев П. В. Истоки биометрии. Из истории биологии. М., 1971.