Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

В.Ф. Багинский
О.В. Лапицкая

БИОМЕТРИЯ В ЛЕСНОМ ХОЗЯЙСТВЕ

Гомель
УО «ГГУ им. Ф.Скорины»
2010

Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

В.Ф. Багинский
О.В. Лапицкая

БИОМЕТРИЯ В ЛЕСНОМ ХОЗЯЙСТВЕ

Учебное пособие для студентов высших учебных заведений, обучающихся по специальностям «Лесное хозяйство», «Лесо-инженерное дело», «Садово-парковое строительство»

Гомель
УО «ГГУ им. Ф.Скорины»
2010

УДК 630*9:519.24 (075.8)
ББК 43:22.172я73
Б 144

Рецензенты:
Профессор кафедры лесоустройства и лесной таксации учреждения образования «Белорусский государственный технологический университет», доктор сельскохозяйственных наук, профессор, заслуженный лесовод Республики Беларусь *О.А. Атрощенко*;

Заведующий кафедрой лесоустройства и лесной таксации учреждения образования «Белорусский государственный технологический университет», кандидат сельскохозяйственных наук, доцент *В.П. Машковский*;

Главный научный сотрудник Института леса НАН Беларуси, доктор биологических наук *В.Б. Гедых*

Багинский В.Ф.
Б 144 Биометрия в лесном хозяйстве: учебное пособие для студентов высших учебных заведений, обучающихся по специальностям «Лесное хозяйство», «Лесоинженерное дело», «Садово-парковое строительство» / В.Ф. Багинский, О.В. Лапицкая. – Гомель: ГГУ им. Ф. Скорины, 2010. - с.

Учебное пособие содержит изложение понятий о биометрии и вариационной статистике в их приложениях к лесному хозяйству. Показаны методы сбора, обработки и анализа биометрической информации в лесном хозяйстве. Описаны распределения случайных величин, методы установления и анализа связей между ними: регрессионный, дисперсионный анализ и др. Рассказано о планировании эксперимента. Изложены типичные случаи применения методов биометрии в лесном хозяйстве и приведены соответствующие примеры из практики.

Для студентов, магистрантов, аспирантов и преподавателей лесохозяйственных и лесоинженерных факультетов высших учебных заведений, научных и инженерно-технических работников лесного хозяйства, лесоустройства и садово-паркового хозяйства, экологов и учителей биологии.

УДК 630*9:519.24 (075.8)
ББК 43:22.172я73

ISBN 985-434-507-6

ВВЕДЕНИЕ

В лесном хозяйстве широко применяются методы лесной биометрии. Без математико-статистической обработки сегодня немыслимо проведение лесоводственных исследований. Подавляющее большинство нормативов, применяемых в лесном хозяйстве, разработано с использованием биометрических методов.

Широкое применение компьютеров в лесоустройстве и в лесхозах Беларуси сделало биометрические методы доступными широкому кругу лесоводов. В то же время лесовод не может быть бездумным пользователем компьютерных результатов биометрических измерений и вычислений. Он должен понимать суть изучаемого явления или процесса, разбираться в алгоритме и механизме вычислений, которые выполнил компьютер по заданной программе.

С этой целью в ВУЗах, которые готовят инженеров лесного хозяйства – специалистов для работы в лесхозах, в лесоустройстве, в садово-парковом хозяйстве, а также для лесопатологов, охотоведов и т.д., на втором году обучения читается курс лесной биометрии. Он включает 36 часов лекционных и 34 часа практических занятий, т.е. всего 70 часов.

Для того, чтобы курс биометрии, который в определенной степени требует знания математики, был усвоен, лесоводам в современных вузах преподают обширный курс математики. Этот курс меньше, чем на специализированных математических и технических факультетах, но тоже достаточно широк: математика по количеству отпущенных на ее изучение часов сопоставима с основными лесными дисциплинами: лесоводством, лесоустройством и др., что достаточно для понимания и усвоения курса биометрии. Для организации учебного процесса по курсу биометрии и для ее использования в практике лесного хозяйства требуется современное учебное пособие.

Математические методы в лесном хозяйстве, в т.ч. математическая и вариационная статистика, применяются уже со второй половины XIX века. Вспомним известную школу «форстматематиков», ярким представителем которой в России был А.К. Турский. Как самостоятельная дисциплина математическая статистика стала преподаваться в лесных институтах Советского Союза с 30-х годов прошлого века. Естественно, появились соответствующие учебные пособия. Их авторами с 30-х и до начала 60-х годов были известные ученые А.В. Тюрин, М.Л. Дворецкий, М.Г. Здорик, К.Е. Никитин и другие. Со второй половины 60-х годов и в 70-е годы изданы пособия, написанные О.А. Труллем, И.И. Гусевым, Э.Н. Фалалеевым, П.К. Верхуновым, Л.П.Зайченко, Г.Л. Кравченко и Е.С. Мурахтановым и др. В 1977 вышла последняя книга из этой серии – учебник «Вариационная статистика» Н.Н. Свалова.

Кроме перечисленных пособий студенты пользуются учебниками, где излагается биометрия в общебиологическом и общетехническом плане. Их авторами являются А.К. Митропольский, Г.Ф. Лакин, Н.А. Плохинский, П.В.Рокицкий, Г.Н. Зайцев, В.Л. Вознесенский, Д. Дюче, М.Г. Кенуй,

Е.И.Гурский, Б.Л. Ван дер Ванден и др. Все эти книги, выпущенные в основном 40 и более лет назад, давно стали библиографической редкостью. К тому же учебники и учебные пособия, изданные в 60-70 годы прошлого века и ранее, уже морально устарели. Многие из них сориентированы на довольно скромные знания математики лесоводами, что было характерно для того времени. В то же время чрезмерно математизированное изложение является препятствием для усвоения курса студентами-лесоводами и сегодня. Поэтому требуется, сохранив все то рациональное, что содержится в старых пособиях, подготовить новый курс лесной биометрии.

В последнее десятилетие увидели свет подобные пособия М.В. Устинова (Брянск) и львовских ученых М.П. Горошко, С.И. Миклуша и П.Г. Хомюка, но они практически неизвестны в Беларуси, а львовская книга написана на украинском языке. В конце 70-х годов К.Е. Никитин и А.З. Швиденко выпустили прекрасную книгу «Методы и техника обработки лесоводственной информации». Эта книга предназначена в основном для аспирантов и научных работников. В настоящее время она тоже уже библиографическая редкость, к тому же для студентов отдельные ее разделы несколько сложноваты.

В силу сказанного издание современного пособия по применению биометрии в лесном хозяйстве является актуальным. Скорее всего, следовало бы издать несколько пособий разного уровня сложности, чтобы у студентов и практических лесоводов был выбор доступного ему пособия по уровню подготовки и предъявляемым требованиям, как это было и раньше.

При подготовке настоящего пособия авторы, естественным образом, опирались на общеметодические разработки, которые сделаны в области математической статистики и биометрии и изложены в учебниках А.К. Митропольского, Н.А. Плохинского, П.Ф. Рокицкого, Г.Ф. Лакина, Н.Н. Свалова и других. Особое место при написании пособия заняла монография К.Е. Никитина и А.З. Швиденко, где в современной интерпретации поданы многие вопросы биометрии, обойденные в других изданиях, например, система кривых распределения Джонсона. Поэтому названное издание широко использовано в ссылках в настоящем пособии. Учтены и новые моменты в части оценки выборочных показателей, приводимых М.П. Горошко с соавторами.

Статистические таблицы (величины различных критериев, функции распределения и т.д.) обычно составлены давно и идентичны для всех учебных пособий. Поэтому они перенесены в настоящее пособие без изменений из тех книг, которые обладали лучшим полиграфическим исполнением.

Не претендуя на оригинальность изложения самих формул, их выводов и т.п., что является прерогативой специалистов-математиков, авторы пособия как лесоводы сосредоточили свое внимание на доступности изложения для понимания его лесоводами, на практические приложения к области применения биометрических методов в лесном хозяйстве.

В настоящее время биометричесикие методы широко применяются магистрантами, аспирантами, научными и научно-техническими работниками при проведении исследований в лесном хозяйстве. Поэтому настоящий курс биометрии помимо того минимума знаний, который необходим студентам

(он определяется учебной программой и регулируется преподавателем), должен удовлетворять и более высокие потребности перечисленной категории пользователей. В силу сказанного основные темы учебного пособия содержат сведения, которые обеспечивают потребность специалистов, возникающую при проведении обработки и анализа экспериментального материала.

Приводимые примеры и иллюстрации взяты в основном из практических задач, которые требовалось решать авторам в своей многолетней научно-исследовательской работе. В отдельных случаях использованы примеры других авторов, на которых сделаны ссылки.

Предлагаемое пособие состоит из 18 тем в соответствии с тематикой лекционных занятий, определяемых учебной программой по лесной биометрии. При подготовке пособия учтен многолетний опыт проведения занятий по лесной биометрии В.Ф. Багинским в Гомельском государственном университете им. Ф. Скорины и Брянской государственной инженерно-технологической академии, а также более чем 40-летний опыт математико-статистической обработки и анализа исходных материалов при проведении научных исследований.

Настоящее учебное пособие будет полезно студентам по вышеперечисленным специальностям, а также магистрантам, аспирантам, преподавателям, научным работникам и сотрудникам проектных учреждений лесного профиля. Этим пособием смогут пользоваться экологи, биологи и специалисты по лесоинженерному делу.

Авторы приносят благодарность рецензентам пособия: сотрудникам кафедры лесоустройства и лесной таксации Белорусского государственного технологического университета, доктору сельскохозяйственных наук, профессору, Заслуженному лесоводу Республики Беларусь О.А. Атрощенко и заведующему кафедрой кандидату сельскохозяйственных наук, доценту В.П. Машковскому, а также главному научному сотруднику Института леса НАН Беларуси, доктору биологических наук В.Б. Гедых за ценные советы и замечания, позволившие улучшить книгу.

1. БИОМЕТРИЯ КАК НАУКА

1.1 Определение биометрии как специальной дисциплины при подготовке инженеров лесного хозяйства, ее цели и задачи
1.2 Особенности биометрии как науки и ее место в ряду других наук
1.3 История возникновения и развития математической статистики и биометрии
1.4 Лесная биометрия как часть общей биометрии, ее значение для развития лесного хозяйства

1.1 Определение биометрии как научной дисциплины, ее цели и задачи

Слово “биометрия” происходит от греческих слов “bios” - жизнь и “metreo” - измеряю, т.е. это наука о планировании и проведении измерений в живой природе, их обобщении и анализе результатов этих измерений.

Явления жизни, как и вообще все явления материального мира, имеют две неразрывно связанные стороны: качественную, воспринимаемую непосредственно органами чувств, и количественную, выражаемую числами при помощи счета и меры.

При исследовании различных явлений природы применяют одновременно и качественные, и количественные показатели. Несомненно, что только в единстве качественной и количественной сторон наиболее полно раскрывается сущность изучаемых явлений. Однако в действительности приходится пользоваться, смотря по обстоятельствам, либо качественными, либо преимущественно количественными показателями, памятуя о том, что качество и количество материи находятся в диалектическом единстве, взаимно переходят друг в друга.

Несомненно, что количественные методы как более объективные и точные имеют преимущество перед качественной характеристикой предметов. Недаром еще в древности достоверность познания природы связывалась с математикой - наукой точной, изучающей количественные отношения и пространственные формы реальной действительности. Опираясь на количественные показатели, можно получить более достоверную информацию о предметах, что позволяет глубже постигнуть их качественное своеобразие.

Количественные методы не ограничиваются одними лишь измерениями или учетом живых существ и продуктов их жизнедеятельности. Сами по себе результаты измерений, хотя и имеют известное значение, еще недостаточны для того, чтобы сделать из них необходимые выводы. Цифровые данные, собранные в процессе массовых испытаний, т.е. измерений или учета изучаемого объекта, - это всего лишь сырой фактический материал, который нуждается в соответствующей математической обработке. Без обработки - упорядочения и систематизации цифровых

данных - не удается извлечь заключенную в них информацию, оценить надежность отдельных суммарных показателей, убедиться в достоверности или недостоверности наблюдаемых между ними различий. Эта работа требует от специалистов определенных знаний, умения правильно обобщать и анализировать собранные в опыте данные. Система этих знаний и составляет содержание биометрии - науки, занимающейся главным образом вопросами статистического анализа результатов исследований как в области теоретической, так и прикладной биологии, в том числе и в лесном хозяйстве.

1.2 Особенности биометрии как науки и ее место в ряду других наук

Одна из характерных особенностей биометрии как науки состоит в том, что она не имеет прямого и непосредственного отношения к вопросам техники измерений или учета живых существ. Это дело частных наук - таких как ботаника, зоология, антропология, лесоведение, дендрология, энтомология, агрономия и др. Они имеют свои объекты исследования и применительно к ним разрабатывают методику измерений и количественного учета изучаемых объектов. В настоящее время широко применяются биометрические способы планирования биологических экспериментов, опытов, полевых исследований. Но главной задачей биометрии по-прежнему остается обработка результатов измерений и учета результатов опытов для того, чтобы по немногим числовым показателям судить о существе изучаемых явлений. Поэтому с чисто формальной стороны биометрия представляет совокупность математических методов, применяемых к обработке результатов биологических исследований. Эти методы она заимствует преимущественно из области математической статистики и теории вероятностей, составляющих техническую основу биометрии. Следовательно, биометрия - это математическая статистика в приложении к явлениям живой природы.

Сопоставляя названные науки, надо иметь в виду, что математическая статистика и теория вероятностей являются науками сугубо теоретическими, абстрактными. Они изучают статистические совокупности безотносительно к специфике входящих в их состав элементов. Методы математической статистики и лежащей в ее основе теории вероятностей могут быть приложены к самым различным областям знания, включая и гуманитарные науки. Биометрия же - наука эмпирическая, конкретная: она исследует исключительно биологические совокупности, преследуя не математические, а биологические цели.

На этом основании нельзя ставить знак равенства между биометрией и математической статистикой, полностью их отождествлять. Биометрия имеет свой объект исследования, свое место в системе биологических наук: теория вероятностей и математическая статистика - разделы современной математики, а биометрия относится к числу биологических наук. Отношение биометрии к математике аналогично тому, какое суще-

ствует между биологией и методикой ее преподавания. Не имея собственных методов исследования, которые она заимствует из математики, биометрия занимает особое положение, являясь относительно самостоятельным разделом биологии, возникшим на стыке математических и биологических наук. Биометрия является следствием развития теории вероятностей и математической статистики. Ее часто называют вариационной статистикой.

Теория вероятности имеет дело со случайными явлениями. В научных исследованиях, технике и массовом производстве часто приходится встречаться с явлениями, которые при неоднократном воспроизведении одного и того же опыта в неизменных условиях протекают каждый раз несколько по-иному. Такие явления называют случайными. Например, при стрельбе результат каждого отдельного выстрела будет случайным. Проводя экспериментальное изучение какого-либо явления и систематизируя результаты исследования (тех же результатов стрельбы) в виде графической зависимости, мы убеждаемся в том, что при достаточно большом количестве экспериментальных точек получается не кривая, а некоторая полоса, т.е. имеет место случайный разброс экспериментальных точек.

Измеряя диаметры и высоты в лесу и нанося полученные значения на график, мы тоже увидим набор точек, но при достаточно большом числе наблюдений вырисовывается некоторая линия, близкая к параболе или к другой функции, имеющей схожий график.

При решении многих практических задач случайными отклонениями точек от кривой можно пренебречь, предполагая, что в данных условиях опыта явление протекает вполне определенно. Выявляется основная закономерность, свойственная данному явлению. По этой закономерности, применяя тот или иной математический аппарат, можно предсказать результат опыта по его заданным условиям.

По мере развития различных отраслей науки становится необходимым изучать случайные явления, с тем чтобы научиться предвидеть действия случайных факторов и учитывать их в практическом решении задач. В лесном хозяйстве такая потребность появилась в конце XIX века, что дало толчок к использованию методов математической статистики и положило начало развитию лесной биометрии.

Математическая наука, изучающая общие закономерности случайных явлений независимо от их конкретной природы и дающая методы количественной оценки влияния случайных факторов на различные явления, называется теорией вероятностей. Основой научного исследования в теории вероятностей является опыт и наблюдение. На теории вероятностей базируется вся статистика - и математическая, и экономическая, и социальная.

Мы живем в век статистики. Едва ли не в каждом своем аспекте явления природы, а также человеческая и прочая деятельность поддаются сейчас измерению при помощи статистических показателей. Суще-

ствует две диаметрально противоположные точки зрения на статистику, широко доступную в настоящее время для населения. Согласно одной из них, публикуемые статистические данные содержат в себе некое смысловое качество, нечто подобное тому, что приписывали числам пифагорейцы, и обладают такой степенью непогрешимости, что их можно принимать на веру безоговорочно. Это, конечно, так же абсурдно, как и другое, еще более распространенное, мнение о том, что можно сфабриковать статистические данные, которые докажут все что угодно, а потому, следовательно, они в действительности ничего не доказывают.

Последнее утверждение нашло отражение в юмористическом виде. Известна злая шутка. Есть три рода лжи: ложь вынужденная или просто ложь, ей есть оправдание; ложь наглая, для которой нет оправдания, и статистика.

Но ничего из сказанного не имеет отношения к математической статистике. Последней нет необходимости лгать вообще. Обе точки зрения ошибочны потому, что они основаны на незнании или непонимании целей, пределов и требований подлинной статистической теории и практики. Математическая статистика - это строгая наука и, как любая наука, она объективна и беспристрастна.

Математическая статистика - это наука о методах количественного анализа массовых явлений, учитывающей одновременно и качественное своеобразие этих явлений.

В большинстве случаев для выявления общей закономерности необходимо большее число наблюдений. Но мало выполнить наблюдения. Необходимо применить специальные способы их обработки. Более того, наблюдения должны быть спланированы и организованы специальным образом, иначе их ценность резко снизится. Методы и правила, необходимые формулы для организации наблюдений и обработки полученного материала дает математическая статистика.

В целом и теория вероятностей, и математическая статистика - это разделы математики. Связи современной биологии и лесного хозяйства с математикой многосторонни, они все больше расширяются. Но не всякие количественные методы, используемые в биологии, составляют содержание биометрии. Нельзя, например, отождествлять биометрию с кибернетикой и с так называемой “математической биологией”, у которых имеются свои, особые задачи, не совпадающие с задачами биометрического анализа совокупностей.

Другой характерной особенностью биометрии как науки является то, что ее методы могут быть приложены не к единичным объектам, не к отдельным результатам наблюдений, а к их совокупности, т.е. к явлениям массового характера. Если мы рассматриваем, например, отдельно взятую особь и сравниваем ее с популяцией, к которой она принадлежит, то можем сказать, что между данной особью и популяцией, как равно и между результатами единичных измерений и всей их совокупности в целом, существует самая тесная связь. Иначе говоря, общее и единичные явления не просто “сосуществуют”, они взаимно обусловли-

вают друг друга. Нельзя представить совокупность без ее членов, как невозможен лес без деревьев определенного ботанического вида или определенной древесной породы. Древесными породами в лесоводстве принято называть древесные виды. Каждый лесовод знает, что некоторая сумма деревьев – это еще не лес. Лес отличается от совокупности деревьев не только количественно, но и качественно. Вот это качественное отличие и должны отражать законы биометрии.

С первого взгляда кажется, что между общим и отдельным, целым и частью нет никакой разницы и что законы, действующие в сфере единичных и массовых явлений, одни и те же. Но это не так. Множество или общее не есть простая арифметическая сумма входящих в ее состав единиц. В сфере массовых явлений, т.е. в совокупностях, действуют свои, присущие им статистические законы, которые лишь в общих чертах характеризуют единичные явления. Также и законы, присущие единичным явлениям, не отражают в полной мере общих закономерностей, проявляющихся лишь в сфере статистических совокупностей.

В этом противоречивом единстве и заключается внутренняя связь между частью и целым, между единичными явлениями и их совокупностью. Биометрия помогает выявлять эту связь и оценивать значение отдельных факторов в свете общих закономерностей, присущих совокупности в целом.

Наконец, нельзя не отметить еще одну характерную черту биометрии - ее своеобразный язык знаков, символов, уравнений, графиков и формул. Все эти условные обозначения, предназначенные для “экономного” выражения мысли, биометрия заимствует из математики. Известно, что математическая логика отличается краткостью и точностью доказательных формулировок, убедительностью выводов. Благодаря символике удается сравнительно простыми и точными средствами выражать содержание сложных и многообразных явлений природы, что значительно облегчает понимание присущих им закономерностей.

Графики, уравнения и формулы, поскольку они заключают в себе наиболее существенное и типичное в явлениях, служат своего рода математическими моделями этих явлений. Математическое моделирование в данном случае аналогично схематическим построениям в виде рисунков, графиков и иных изображений, широко используемых в педагогической и научно-исследовательской работе. В наш век сплошной компьютеризации именно математическое моделирование является основным методом описания выявленных законов и закономерностей в живой природе, в том числе и в лесных науках: лесоведении, лесоводстве, лесной таксации, защите леса, лесной генетике и селекции и т.д.

Разумеется, любые схемы и модели дают лишь некоторое подобие реальной действительности. Но именно в этом и заключаются их большие методические возможности. Достаточно облечь мысли в форму символов и знаков, геометрических фигур и уравнений, как это дает

широкие возможности для глубокого и всестороннего познания явлений, быстрого движения на пути к истине.

Но символы и вообще биометрические показатели приобретают определенный смысл лишь тогда, когда они соответствуют содержанию выражаемого ими процесса, находятся в тесной связи с конкретными задачами биологического исследования. В противном случае биометрическая символика не только не оправдывает себя, но может привести исследователя к ошибкам и заблуждениям.

Дело в том, что в краткости и точности числовых характеристик, в удобстве выражать биологические явления языком математических формул и уравнений заключены не только большие методические возможности, но и опасность отрыва от конкретных явлений, а это ведет к ошибкам, создает видимость истины там, где ее на самом деле нет. Не следует выражать математическими формулами или графиками то, что очевидно само по себе. Во многих случаях биометрические данные, сведенные в статистические таблицы, оказываются настолько убедительными, что не нуждаются ни в какой дополнительной обработке.

Посмотрим на две таблицы, показывающие ход роста сосновых древостоев.

Таблица 1.1 - Динамика средних высот соснового древостоя в кисличном и сфагновом типах леса

Возраст, лет	Средняя высота в разных типах леса, (м)	
	кисличный, I^a класс бонитета	сфагновый, V класс бонитета
20	10,2	3,6
40	19,1	7,1
60	25,4	10,2
80	29,9	12,9
100	33,0	15,0

Из таблицы 1.1 сразу видно, что рост сосняка кисличного намного интенсивнее, чем сосняка сфагнового. Здесь можно привести сравнение по методикам биометрии, но результат заранее ясен. Посмотрим теперь другие материалы. Рассмотрим динамику двух однородных древостоев, растущих в одинаковых условиях, но отличающихся тем, что в одном из них провели рубки ухода, т.е. изредили древостой. В обоих участках определили среднюю высоту. Результаты измерений приведены в таблице 1.2.

Таблица 1.2 - Изменение средней высоты в неизреженном и изреженном сосняке мшистом

Возраст, лет	Средняя высота, м	
	неизреженный древостой	изреженный древостой
20	8,3	8,3
40	14,4	14,6
60	19,5	19,4
80	23,2	23,7
100	25,7	26,1

Из таблицы 1.2 вытекает, что изреженный древостой имеет несколько среднюю большую высоту, но различия небольшие. Поэтому без дополнительного анализа судить о том, что изреженные древостои имеют большую среднюю высоту чем те, где рубки ухода не проводили, нельзя. Чтобы дать ответ на этот вопрос, надо обратиться к соответствующим биометрическим методам и оценить степень достоверности наблюдаемых в опыте различий. Методика такой оценки будет показана в дальнейшем.

Из сказанного отнюдь не следует, что нужно как-то ограничивать применение математических методов в лесоводстве. Речь идет не об ограничении, а о правильном использовании этих методов в лесных исследованиях. Сама по себе биометрия не ведет к ошибкам и заблуждениям: они возникают при неумелом, механическом использовании биометрических показателей без учета их конструктивных особенностей и теоретического обоснования. В работе лесовода в равной мере неприемлемы как биометрическое жонглирование, превращающее работу исследователя в бесплодную и вредную “игру в циферки”, так и примитивизм в оценке числовых показателей, ведущий на деле к отказу от применения математических методов в лесном хозяйстве. Истина, как это всегда бывает, заключена не в крайностях, а в разумном подходе к делу.

1.3 История возникновения и развития математической статистики и биометрии

Биометрия сама по себе наука относительно молодая. Первые опыты ее применения относятся к работам Борелли, который на рубеже XVII и XVIII веков делал математические расчеты движения животных. В начале XVIII века французский ученый Реомюр (1683-1757) (вспомним, что есть температурная шкала Реомюра) искал математические законы строения пчелиных сотов.

Но как полноценная наука биометрия появилась лишь к концу XIX века. Термин “биометрия” был введен в науку английским ученым-

антропологом Фр. Гальтоном (1822-1911) в 1889 году для обозначения количественных методов, применяемых в области биологических исследований. В дальнейшем Дункер (1899) предложил другое название - “вариационная статистика”, которое тоже вошло в обиход как выражающее более точное содержание данного предмета. В настоящее время употребляются оба эти термина, хотя буквальный смысл их неодинаков. Слово “биометрия” (от лат. bios - жизнь, metron - мера) означает производство биологических измерений, а термин “вариационная статистика” (от лат. слов variatio - изменение, колебание и status - состояние, положение вещей) понимается как описание наблюдений, их математическая обработка. Гораздо более длительную историю имеет как общая статистика, так и математическая или вариационная статистика.

Развитие статистики началось в эпоху античности, т.е. эта наука имеет древние корни. Например, использование среднего значения было хорошо известно еще при жизни Пифагора (VI в. до н.э.), а упоминания о статистических обследованиях встречаются и в библейские времена. Статистика постепенно развивалась там, где в ней возникала необходимость. В первую очередь статистические методы начали применяться для анализа экономики и явлений общественной жизни, и лишь позже они проникли в биологию. Так, Исаак Ньютон (1642-1727), чей вклад в открытие дифференциального и интегрального исчисления был выдающимся событием в математике, а его теория всемирного тяготения и «ньютоново яблоко» известны любому старшекласснику, является, возможно, наиболее заметной фигурой в области развития современной статистики, хотя сам Ньютон вряд ли слышал когда-либо о существовании этой науки. Другие математики, чьи имена известны прежде всего благодаря работам в области чистой математики, косвенно сделали для развития статистики больше, чем многие из тех ученых, которые непосредственно специализировались на этой науке. Двумя наиболее выдающимися представителями таких ученых-математиков являются Абрахам де Муавр (1667-1754) и Карл Гаусс (1777-1855).

Что касается самих ученых-статистиков, то стоит упомянуть бельгийца Адольфа Кетле (1796-1874), который первым применил современные методы сбора данных. Возможно, вам покажется несколько неожиданным, что и знаменитая английская медицинская сестра середины XIX века (крымская война 1854-55 гг) Флоренс Найтингейл (1820-1910) всю свою жизнь была горячей сторонницей применения статистики. Вспомним, что высшей международной наградой медицинским сестрам, спасающим солдат на войне, является медаль Флоренс Найтингейл. Ею награждены и многие советские медицинские сестры – участницы Великой Отечественной войны 1941-1945 гг. Ф. Найтингейл, работавшая впоследствии на руководящих должностях, доказывала, что администратор может иметь успех только в том случае, если он в своей деятельности будет руководствоваться данными, получаемыми с помощью статистики, и что законода-

тели и политики часто терпели неудачи из-за того, что их статистические познания были недостаточны.

Двумя другими учеными, внесшими значительный вклад в развитие статистики, являются два англичанина - Фрэнсис Гальтон (1822-1911) и Карл Пирсон (1857-1936). Гальтон, родственник Чарльза Дарвина (1809-1882), серьезно заинтересовался проблемой наследственности, к анализу которой он вскоре применил статистические методы. Гальтон и Пирсон внесли значительный вклад в развитие теории корреляции, которую мы детально рассмотрим ниже. Наиболее известным ученым в области статистики в двадцатом веке являлся Рональд Фишер (1890-1962). Фишер продуктивно работал с 1912 по 1962 г., и многие его исследования оказали существенное воздействие на современную статистику.

В двадцатом веке статистические методы официально введены в Соединенных Штатах для обучения во всех колледжах. Там читались небольшие курсы статистики. В течение первых тридцати лет этого века постепенно возрастало значение статистики в исследовании проблем психологии. При этом отметим, что в данный период психология как наука не была самостоятельной и часто рассматривалась лишь как один из разделов философии.

Применение математической статистики не обошло и лесную отрасль. В лесном хозяйстве нашей страны (Россия, СССР) математическая статистика стала использоваться с конца XIX века в основном благодаря трудам известного лесовода и таксатора профессора А.К. Турского. Уже в 20-30 годы прошлого века математическая статистика стала неотъемлемой частью исследований в лесном хозяйстве.

Одной из причин широкого применения статистических показателей в последние годы является все возрастающая легкость обработки больших массивов чисел. Современные компьютеры позволяют в короткое время проанализировать огромное количество статистических данных, что раньше было невозможно.

Для внедрения математики в биологию и лесное хозяйство в конце XIX и начале XX века имелись серьезные основания. Одним из них был переход от описательного метода изучения явлений жизни к экспериментальному. Хотя и при описательном подходе возможно установление математических закономерностей (примером могут быть законы движения небесных тел), однако преобладает в этом случае качественная оценка. Эксперимент же неизбежно требует количественного выражения явлений и процессов. Создание физиологии, генетики, радиобиологии и других экспериментальных областей биологии повлекло за собой разработку многочисленных математических приемов и методов исследования. Большую роль сыграли и чисто практические причины. Разработанные методы стали широко применяться в зоологии, ботанике, лесоводстве, лесной таксации и других биологических науках.

Наконец, важнейшим обстоятельством, определившим использование математических и математико-статистических методов, явилось установ-

ление того факта, что многим биологическим явлениям свойственны статистические закономерности, обнаруживаемые при изучении совокупностей, но неприложимые к отдельным единицам этих совокупностей.

Когда физики перешли от изучения поведения отдельных физических тел к изучению поведения множеств молекул, электронов, они вступили в область действия статистических законов. На этой основе создалась особая область физики - статистическая физика, изучающая свойства и поведение систем, состоящих из огромного количества отдельных частиц. В основе многих физических явлений, таких, как радиоактивный распад, термодинамические явления и некоторые другие, лежат статистические закономерности. С их открытием закономерности, установленные эмпирически, например законы термодинамики, получили более глубокое обоснование и были выведены из статистических вероятностных законов.

Физики в начале XIX века долго не могли примириться, что в микромире действуют статистические законы. Ранее казалось, что в физике все детерминировано, т.е. на однозначное действие наступает однозначный результат. Квантовая механика это отвергла. Даже великий Альберт Эйнштейн (1879-1955) долго не мог воспринять такое, неоднократно повторяя: “Неужели господь Бог играет с нами в кегли”. Большинство великих физиков верили и верят в Бога.

Примерно такое же положение наблюдается и сейчас в ряде областей биологии. Когда зоологи, ботаники перешли от изучения отдельных “типичных” представителей вида к изучению многих особей одного вида, они обнаружили массовые явления статистической природы. Рыбы, рачки, моллюски, сосны, коловратки, водоросли, инфузории и другие животные и растения характеризуются изменчивостью, вариацией по самым разнообразным признакам. Такой же вариацией обладают и организмы, культивируемые человеком: колосья пшеницы различаются количеством зерен в колосе, весом отдельных зерен; звери одного вида имеют разную массу, у них варьирует экстерьер и окрас. Весьма изменчивыми объектами являются лесные насаждения. Даже в однородном древостое мы видим большое разнообразие деревьев: по высоте, диаметру, форме кроны и т.д.

При изучении биологических совокупностей, являющихся типично статистическими, оказалось целесообразным применить методы математической статистики, которую в приложении к биологии стали называть биологической статистикой. Еще ее называют вариационной статистикой.

Поле для приложения статистических методов в биологии очень значительно, так как многие экологические, генетические, цитологические, микробиологические, радиобиологические явления - массовые по своей природе. В них участвуют не одна особь или клетка, не одна α-частица, не одна бактерия или вирусная частица, не одно дерево, а множества, т.е. совокупности клеток, α-частиц, бактерий, особей вида, семей, деревьев и т.д. Осуществление событий в таких совокупностях

может быть оценено вероятностями, а анализ их требует применения статистических методов.

Статистические методы существенно необходимы и при постановке экспериментов, так как только с их помощью можно установить, зависит ли наблюдаемое различие между опытными и контрольными делянками леса от влияния изучаемого фактора или же оно чисто случайно, т.е. определяется многими другими, не контролируемыми и не поддающимися учету факторами. Понимание и учет статистических закономерностей помогают экспериментатору составить методически обоснованный план опытов, правильно их провести и, наконец, сделать из них объективные выводы. При этом надо помнить, что никакая математическая и статистическая обработка не поможет, если опыты были проведены неправильно или данные собраны небрежно.

Роль математики и математической статистики в биологии особенно возросла в связи с развитием теории информации и кибернетики в целом и многих связанных с ними областей математики, среди которых главное место занимают теория вероятности, математическая статистика и математическая логика. Применение компьютеров на порядки ускорило и расширило применение статистических методов в биологии вообще и в лесоводстве в частности.

Использование математики в современной биологии не ограничивается только статистическими методами. Поэтому биометрия (или биоматематика, как ее иногда называют) шире, нежели биологическая статистика. Она использует также приемы и методы из других областей математики: дифференциального и интегрального исчислений, теории чисел, матричной алгебры и т.д.

Внедрение математики в биологию первоначально выражалось в использовании отдельных математических и математико-статистических методов для изучения тех или иных биологических вопросов и обработки данных, полученных из природы или в лаборатории. Такие вопросы, как изменчивость морфологических, физиологических и экологических признаков животных и растений и установление влияния на них внешних и внутренних факторов, количественный учет и процессы, происходящие в популяциях, сходство и различия между видами, подвидами и иными систематическими категориями, рост индивидуальный и рост популяций, могут изучаться лишь с помощью математических и математико-статистических методов. Более того, в различных областях биологии (генетика, эволюционное учение, селекция, физиология), а также практически во всех лесных науках: лесоводстве, лесной таксации и др. соответствующие биологические процессы или явления теперь выражаются в математической форме.

Таким образом, пройдя длительный путь развития, математическая статистика и биометрия сегодня предстают перед нами как стройные и высокоразвитые науки, имеющие большое практическое значение.

1.4 Лесная биометрия как часть общей биометрии и ее значение для развития лесного хозяйства

Нас, как лесоводов, при изучении биометрии больше всего интересует тот ее раздел, который называется “лесная биометрия”.

Лесная биометрия - это раздел биометрии, содержанием которого является планирование и организация количественных экспериментов в лесоведении, лесоводстве, лесной таксации и в других лесных дисциплинах, обработка и анализ полученных результатов, используя методы математической статистики.

Лесная биометрия достаточно развитая наука. Для описания леса, лесных биогеоценозов, отдельных элементов лесного насаждения используются и дают важные практические результаты многие математические подходы, применяющиеся в общей биометрии.

Методы лесной биометрии широко используются при анализе процессов в природных явлениях, свойственных деревьям и древостоям, а также отдельным участкам леса. В качестве примера можно привести анализ наследуемости признаков деревьев и древостоев, оценка эффективности различных лесохозяйственных мероприятий.

Лесная биометрия - необходимый методический инструмент для проведения научных исследований в лесу, при разработке различных нормативных и справочных материалов. Например, все лесоводы пользуются сортиментными и объемными таблицами для учета леса на корню, применяют специальные таблицы для учета готовой лесопродукции, замеряя длину сортимента и его диаметр в верхнем отрезе и т.д. В этих таблицах приведены усредненные величины, полученные на основе обработки методами биометрии большого экспериментального (его еще называют первичным или исходным) материала. Далеко не каждое бревно определенной длины и диаметра имеет объем точно совпадающий с тем, который приведен в справочнике для искомых параметров длины и диаметра. Но, пользуясь методами лесной биометрии, ученые получили объемы с достаточной степенью приближения к реальности, и точность определения увеличивается при возрастании числа измерений.

На этом небольшом примере, а их очень много, мы видим, что ни лесная наука, ни лесохозяйственная практика не могут обойтись без использования в своей работе лесной биометрии. Методы лесной биометрии широко применяются также при обработке лесоустроительного материала в геоинформационной системе «Лесные ресурсы» с помощью компьютерных программ в системе «СОЛИ», а аткже при анализе данных аэро- и космической съемки.

В лесной биометрии широко применяются методы математического анализа, математическое моделирование. В настоящее время наиболее рациональным является применение биометрических методов с использованием компьютеров. Практически сегодня все вычисления проводятся только на компьютерах.

В то же время, чтобы правильно использовать методы лесной биометрии, необходимо хорошо знать лесоводство, лесную таксацию, лесные культуры и другие дисциплины, где применяют эти методы. Без хорошего знания той дисциплины, в которой мы работаем, биометрические методы получение корректного результата не гарантируют. Механическое, бездумное жонглирование цифрами, сведенными в модели, пусть даже и обработанными статистическими методами, недопустимо, т.к. может привести к ложным выводам.

Прежде чем выводить модель, надо представить общую биологическую закономерность. Например, известно, что дерево растет вверх, диаметр дерева не может уменьшаться и т.д. Это простейшие примеры, где все очевидно. Но есть много случаев, когда бездумное применение статистических моделей приводит даже при анализе роста деревьев к отрицательным величинам. В отношении приведенных примеров ошибку выявить легко, но далеко не все закономерности столь очевидны. Именно поэтому, начиная работу с материалом, который будет обработан биометрическими методами, надо очень хорошо знать свой предмет: лесоводство, таксацию, лесную селекцию и т.д. В дальнейшем при изучении разных методов биометрии мы это рассмотрим более подробно на конкретных примерах.

Очень часто ставят равенство между лесной биометрией и математической или вариационной статистикой. Следует помнить, что математическая статистика шире биометрии, а тем более, лесной биометрии, т.к. охватывает весь круг природных явлений живой и неживой природы, добавляя сюда технику, общество и экономику.

Когда мы в дальнейшем будем говорить о соотношениях математической статистики и лесной биометрии, то должны иметь в виду, что объектом применения статистических методов у нас будут деревья, древостои, лесные биогеоценозы и т.д. Вариационная статистика широко используется в исследованиях не только для оценки и анализа результатов измерений, но и для планирования эксперимента.

Биометрия как наука не стоит на месте. Здесь разрабатываются новые подходы и совершенствуются старые способы. Это расширяет круг решаемых задач в лесном хозяйстве.

Обобщая сказанное, можно сделать следующие выводы.

- Лесная биометрия широко используется в лесном хозяйстве.

- Без знания лесной биометрии невозможно понять суть и значение многих нормативов и величин, широко применяемых в лесном хозяйстве.

- Лесная биометрия - составная часть методики научных исследований в лесном деле. Любому исследователю ее обязательно надо знать и знать хорошо.

В дополнение ко всему отметим, что это интересная и захватывающая наука. Она имеет свою строгую логику, свои законы. В то же время лесная биометрия непростая наука, требующая для своего освоения серьезной работы и хорошей общебиологической, лесоводственной и математической подготовки.

2. СТАТИСТИЧЕСКИЕ СОВОКУПНОСТИ

2.1 Статистические совокупности и статистические наблюдения. Статистические выборки
2.2 Генеральная и выборочная совокупность и их объем
2.3 Методы сбора и обработки информации в лесной биометрии
2.4 Дедуктивный и индуктивный методы в лесной биометрии

2.1 Статистические совокупности и статистические наблюдения. Статистические выборки

Изучение биологических явлений проводится не по отдельным наблюдениям, которые могут оказаться случайными, нетипичными, неполно выражающими сущность данного явления, а на множестве однородных наблюдений, что дает более полную информацию об изучаемом объекте. Например, изучая лес, скажем ельник в определенных условиях места произрастания, мы не ограничиваемся измерениями 1-2 или 5-10 деревьев, а проводим замеры на всей пробной площади, где растет не меньше 200 еловых стволов, т.е. изучаем ту их совокупность, которая уже составляет лес.

Некоторое множество относительно однородных предметов или объектов, объединяемых по тому или иному признаку для совместного изучения, называют статистической совокупностью. При этом совершенно не обязательно, чтобы совокупность состояла из множества особей одного вида и возраста, например из однородных деревьев сосны. Она может быть образована и в результате многочисленных испытаний, т.е. проб, наблюдений и т.п., проводимых на одном и том же индивидууме. Например, статистической совокупностью будут данные наблюдений за выработкой условных рефлексов у одной собаки или кошки, фенологические наблюдения за одним или несколькими деревьями дуба и т. д.

Таким образом, совокупность объединяет какое-то число однородных наблюдений или регистраций. Совокупностями являются деревья в древостое, популяции муравьев, заготовленные во время охоты беличьи шкурки, растения на опытных участках. Понятие совокупности применимо не только к животным или растениям. Такими же совокупностями являются молекулы газа в некотором объеме, население города и т.д.

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 совокупности семян сосны - каждое семечко, для древостоя – дерево, при изучении дерева – его клетки и т. д.

Общее свойство изучаемого предмета называется его признаком. Так, при изучении лесного насаждения признаками будут древесный вид, возраст деревьев и древостоев, размер, форма или цвет семян, особенности вегетации (у дуба есть рано или поздно распускающиеся формы) и т. д. 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; семена сосны – формой, цветом, весом; деревья - толщиной, высотой 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 у разных членов совокупности, например у коров данной породы, семян одного вида, деревьев одного вида и т.п.

Элементы, входящие в состав совокупности, называются ее членами, или **вариантами.** Последний термин произошел от латинского varians - изменяющийся. Варианты - это отдельные наблюдения или некоторые числовые значения признака. Так, если обозначить признак через X (большое), то его значения или варианты будут обозначаться через x (малое), т.е. как $x_1, x_2, x_3, x_4, \ldots x_k$.

Общее число вариант, входящих в состав данной совокупности, называется ее **объемом** и обозначается буквой N или $\sum n$. Саму же варьирующую величину, т.е. величину, изменяющуюся под влиянием многих случайных причин и могущую принимать разные значения, называют **случайной переменной n_i.** Варианты являются ее числовыми значениями. Здесь значок i - порядковый номер варианты.

В то же время, несмотря на различия между вариантами, входящими в совокупность, последняя обладает внутренней однородностью. Члены совокупности сходны по ряду важных признаков. Беличьи шкурки неодинаковы по размерам, качеству меха, окраске, но все они - шкурки особей одного и того же вида - белки обыкновенной. Зерна пшеницы отличаются друг от друга по весу и другим химическим и физическим признакам, но все они - зерна пшеницы, а не ячменя, хотя обе культуры могли быть выращены на одном поле. Деревья варьируют по размеру, но все они одного вида, например, сосны. Желуди дуба отличаются по размерам, весу, цвету, но они семена одного древесного вида – дуба черешчатого и т.д.

Чаще всего, в состав совокупностей входят отдельные особи. Так, например, при характеристике плодоносящих деревьев сосны на лесосеменной плантации за единицу совокупности можно взять каждое дерево. Однако единицей совокупностей может быть не само дерево, а некоторая его характеристика. В данном случае допустимо взять урожай шишек или семян за определенные годы. Тогда при общзем количестве деревьев на лесосеменной плантации, скажем 200 штук, количество вариантов, получаемых за несколько лет семян, составит 600, 800 или другую величину.

Можно изучать вариацию того или иного признака во времени даже на одном дереве. Как известно, размер семян и их вес величина изменчивая в зависимости от ботанических и абиотических факторов. Изучая их

изменение на одном дереве за ряд лет, тоже получаем статистическую совокупность, которая изучается методами биометрии. Такой же совокупностью является время распускания почек на одном и том же дереве, но в разные годы. Совокупностью будет длина и вес иголок сосны на одной ветке и т.д.

Таким образом, сумма наблюдений или измерений есть тоже совокупность. Каждое отдельное наблюдение, при котором устанавливается значение случайной переменной, является единицей этой совокупности.

Совокупность может состоять из других, более частных совокупностей. Так, совокупность из всех диких животных одного вида распадается на частные совокупности – отдельные популяции. В пределах одной популяции можно выделить еще более частные совокупности, например, потомство определенных самцов или самок. Изучая древостои сосны, их можно разделить по областям, лесхозам или лесорастительным районам. Во всех случаях мы сталкиваемся с постоянными различиями как внутри отдельных частных совокупностей, так и между ними.

2.2 Генеральная и выборочная совокупность и их объем

Наиболее общую совокупность называют **генеральной**. Это теоретически бесконечно большая или, во всяком случае, приближающаяся к бесконечности совокупность всех единиц или членов, которые могут быть к ней отнесены. Так, если бы можно было описать все особи данного вида, например все деревья сосны в лесах Беларуси, то они составили бы генеральную совокупность.

Генеральная совокупность может состоять из такого большого количества единиц, что изучить их всех нет возможности. Поэтому практически приходится иметь дело со сравнительно небольшими **выборочными** совокупностями. Так, зоолог, изучающий в лесу тот или другой вид животных, отлавливает несколько экземпляров и по ним стремится сделать вывод обо всех особях вида. Лесовод закладывая пробную площадь, где всего-то 200 деревьев, делает выводы о всей совокупности.

Вопрос о том, в какой степени по выборочной совокупности можно судить о генеральной, принадлежит к числу важнейших теоретических и практических вопросов в биологической статистике. Он будет изложен ниже.

Задачей изучения всякой совокупности является получение статистических (или, как иногда говорят, биометрических) характеристик, или показателей. Они 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 именно тогда, когда в ее описание вносится количественный метод. Применение количественного метода изучения совокупности и позволяет получать для нее ряд статистических показателей. С их помощью мы получаем основную информацию о совокупности.

Чтобы выборочная совокупность как можно полнее отражала генеральную, необходимо учитывать следующие основные положения.

1. Выборка должна быть вполне представительной, или **типичной**, т.е. чтобы в ее состав входили преимущественно те варианты, которые наиболее полно отражают генеральную совокупность. Поэтому перед тем как приступить к обработке выборочных данных, их внимательно просматривают и удаляют явно нетипичные варианты. Например, при измерении длины колосьев нельзя включать в выборку испорченные головней или оборванные колосья, поскольку они нетипичны для такого рода выборки. При изучении хода роста деревьев в высоту надо исключить деревья, сломанные бурей, поврежденные огнем и т.д.

2. Выборка должна быть **объективной**. При образовании выборки нельзя поступать по произволу, включать в ее состав только те варианты, которые кажутся типичными, а все остальные браковать. Доброкачественная выборка производится без предвзятых мнений, по методу жеребьевки или лотереи, когда ни одна из вариант генеральной совокупности не имеет никаких преимуществ перед остальными - попасть или не попасть в состав выборочной совокупности. Иными словами, выборка должна производиться по принципу случайного отбора, без каких бы то ни было субъективных изъятий на ее состав. Например, если мы хотим для определения средней высоты измерить 20 деревьев ели, то нельзя их выбирать по своему вкусу, исключая, скажем, низкие угнетенные стволы. Надо измерять высоту у каждого 10 или 20 и т.д. елового дерева. При этом надо учитывать ограничения, упомянутые выше в п.1, например, исключать деревья, сломанные ветром.

3. Выборка должна быть качественно **однородной**. Нельзя включать в состав одной и той же выборки данные, полученные на особях разного пола, вида, возраста или физиологического состояния, так как заведомо известно, что эти факторы по-разному сказываются на величине и функциональном состоянии признаков, по которым может быть образована выборочная совокупность. Неоднородный по составу материал не дает верной информации об изучаемых явлениях. Например, нельзя объединять в одну пробную площадь древостои разных типов леса, хотя бы они росли рядом, скажем сосняк брусничный и сосняк вересковый. Все эти условия может соблюсти только специалист, хорошо знающий не только биометрию, но и предмет своего исследования. В нашем случае это лесовод.

Эмпирические, или выборочные, совокупности могут иметь самый различный объем. В зависимости от числа наблюдений принято различать **малые** выборки, содержащие не более 30 вариант, и выборки **большие >30**, включающие в свой состав до 100-200 единиц совокупности и больше. Верхний предел здесь не ограничен. Принципиальной разницы между большой и малой выборками нет. Различать их приходится на том основании, что сравнительная оценка биометрических по-

казателей, вычисляемых на малых выборках, находится в зависимости от числа наблюдений, о чем будет рассказано ниже.

Схематически совокупности можно выразить следующей схемой (рисунок 2.1).

Рисунок 2.1. Схема подразделения совокупностей

Качественная однородность совокупности определяется целью исследования. Варьирование по учетному признаку определяется единицей отсчета, которая должна быть не больше размера класса, разряда. В лесоводстве размеры разрядов или классов обычно принимается равной одной двенадцатой амплитуды ряда распределения, но допускаются в пределах 1/8 - 1/16.

Таким образом, можно сделать вывод, что совокупность статистическая - это совокупность предметов, явлений, вещей, качественно однородных и варьирующих по учетному признаку: отсюда название - вариационная статистика.

Генеральная статистическая совокупность - некая философская категория, всеобъемлющая совокупность предметов, явлений, вещей, которая может рассматриваться как конечной, так и бесконечной, но обязательно качественно однородной и варьирующей по учетному признаку.

Выборочная (частичная) статистическая совокупность - это часть генеральной совокупности, удовлетворяющая требованиям репрезентативности.

2.3 Методы сбора и обработки информации в лесной биометрии

Выбор правильных методов сбора и обработки информации определяет результат исследования. Наиболее полно они разработаны К.Е.Никитиным и А.З.Швиденко, и мы здесь будем придерживаться этих методов.

Определенная научная или производственная задача, которая ставится в соответствии с очередными потребностями и планами развития отрасли, может быть решена на основе накопленной информации или для ее решения может потребоваться сбор (полный или частичный) новой информации. В обоих случаях в большинстве лесоводственных задач на разных этапах используют нормативно-справочную информацию; иногда

доля ее в общем количестве информации и влияние на конечные результаты весьма существенны.

Если имеющейся информации достаточно, то на ее основе формулируют соответствующую гипотезу и разрабатывают модель, которую проверяют эмпирически при помощи статистических методов. Если же такой информации недостаточно, то принимают решение о конкретных путях исследования, которые определяются целью работы, финансовыми и трудовыми возможностями, имеющимися в наличии средствами сбора и обработки информации.

Одним из основных моментов при организации наблюдения является разработка правильной методики. Надо помнить, что ошибки, допущенные в методике сбора первичного материала, не могут быть потом исправлены никакой камеральной обработкой и вызовут или неточности, или ошибки в выводах.

Как известно, существует два основных способа наблюдений по охвату единиц изучаемого объекта: сплошное обследование всех единиц изучаемой совокупности и частичное обследование, когда наблюдению подвергается лишь часть единиц изучаемой совокупности. В лесном хозяйстве обычно ограничиваются частичным (выборочным) обследованием. Выборку производят для получения характеристики целого, подлежащего изучению. Та совокупность, которая подлежит изучению, называется общей, или генеральной, а отобранные из нее единицы наблюдения представляют частичную, или выборочную, совокупность. Относительная часть, которую составляет число единиц с данным значение признака от общего числа единиц совокупности, называется в генеральной совокупности - долей, а в выборочной - частностью.

Как сплошное, так и несплошное наблюдение по способу производства может быть;

- По связи с объектом - непосредственное, экспедиционное, корреспондентское и отчетное;
- По времени производства - непрерывное, по мере возникновения явления (например, фенологическое); периодическое (повторное) - через определенные промежутки времени и единовременное или однократное;
- По источникам - собственное наблюдение, устный и письменный (анкетный) опрос и по документальным материалам – литературным, служебным и архивным данным.
- По особенностям отбираемых единиц:

❖ отбор средних типичных единиц, применяемый в лесной таксации;

❖ отбор случайных единиц, производимый по специальному плану, который является в статистике основным;

❖ по виду самого отбора:

❖ бесповторное, когда отобранная единица после производства наблюдений над ней не возвращается в генеральную совокупность;

❖ повторное, когда обследованная единица возвращается в генеральную совокупность и может снова попасть в отбор.

• По принципу взятия единиц - случайный, типический (по группам, однородным по какому-либо признаку) и механический. Отбор можно производить или из всей совокупности, а также из ограниченной совокупности. В последнем случае единицы с крайними значениями признака не принимаются во внимание, отбор может вестись и из частных совокупностей, на которые общая совокупность разбита на основе количественных (например, ступени толщины) или же качественных (например, классы роста деревьев) признаков. Частичное наблюдение методом случайного отбора, при котором каждая единица изучаемого объекта обладает одинаковой с другими возможностью попасть в выборку, обеспечивает полную объективность наблюдения.

При работах в лесу случайность (объективность) отбора осуществляется применением механического отбора по принципу бесповторной выборки. В этом случае любая единица совокупности может попасть в выборку только один раз.

Принцип выборочного метода теоретически обоснован выдающимся русским математиком П.Л. Чебышевым (1821 – 1894), доказавшим, что при достаточно большой выборке выборочная средняя может быть как угодно близка к генеральной средней. Отсюда вытекает, что частость признака в выборочной совокупности может быть как угодно близка к доле этого признака в генеральной совокупности. Другими словами - достаточно большая выборка правильно воспроизводит особенности и свойства генеральной совокупности и тем лучше, чем относительно больше выборка.

Отбор единиц можно производить по методу случайной выборки, когда из совокупности наудачу, вслепую (например, по жребию) отбирают нужное число единиц. Это довольно сложно, а в условиях леса практически неприменимо. Поэтому материал наблюдения при исследованиях в лесном деле собирают путем механического отбора.

Сущность этого способа состоит в том, что всю совокупность механически разбивают на число частей, одинаковых по размеру или количеству единиц, соответствующих числу наблюдений, и в каждой части наудачу выбирают единицу для наблюдения. Можно разбить и на число частей, в несколько раз меньшее количества наблюдений. В этом случае из каждой части нужно взять наудачу не по одной единице, а во столько раз больше, во сколько нужное для наблюдения количество единиц больше числа механически образованных частей или групп.

Например, при изучении качества семян (плодов) всю партию семян можно механически разбить на несколько частей или групп, одинаковых по количеству или по весу семян, и из каждой такой части или группы выбрать наудачу партию семян для наблюдения. При отборе модельных деревьев всю совокупность стволов на пробной площади (200–300 шт.) делят по ступеням толщины (через 2–4 см) и отбор делают внутри этих ступеней.

Есть несколько широко применяемых способов отбора.

1. С п о с о б п о л о с о к. Поперек обследуемой площади (например, лесосеки), через одинаковое расстояние закладывают ленты одной и той же ширины, например, 1, 2 или 10 м, со сплошным обследованием на каждой. Площадь всех полосок или лент, выраженная в процентах от всей площади обследования, даст процент выборки. Величина процента выборки зависит от величины обследуемой площади (чем больше площадь, тем меньше может быть процент выборки) и степени колеблемости изучаемого признака: чем больше степень изменчивости признака, тем больше должна быть выборка. Поэтому число лент в разных случаях может быть разное. Значит, до составления методики нужно иметь представление о размерах обследуемых площадей и примерной изменчивости признака. Последнее определяется или по литературным данным, или по материалам лесхоза, или же путем предварительной закладки опыта. Число единиц наблюдения N (в данном случае - число лент, обеспечивающих результат с намеченной точностью) можно определить по формулам:

$$N = \frac{V^2}{p^2} \quad \text{или} \quad N = \frac{\sigma^2}{m^2},$$

где V- коэффициент изменчивости;

σ - среднее квадратическое отклонение значений признака единиц от их среднего значения;

m - намеченная точность в единицах измерения признака;

p - намеченная точность, или точность опыта, %.

Методы нахождения V, σ, m, p будут рассмотрены ниже.

Зная необходимое число лент и длину обследуемой площади, расстояние между центрами лент можно получить делением длины обследуемой площади на запроектированное число лент.

При невозможности предварительно подсчитать необходимое число лент надо брать не менее 5-10 лент и установить расстояние между ними делением длины обследуемой площади на число этих лент. Каждую ленту целесообразно разделить на три-пять частей, одинаковых по длине, которые и будут сложными единицами наблюдения, состоящими из ряда единиц совокупности (деревьев).

При обследовании лесных культур лентами могут служить ряды культур: один, два и больше в ленте. Количество отобранных рядов деревьев в лентах, выраженное в процентах от всего числа рядов на обследуемой площади, представляет процент выборки.

2. С п о с о б п л о щ а д о к. Площадки строго одинакового размера и формы закладывают через одно и то же расстояние одна от другой. Ширина и длина площадок в разных случаях может быть разной, т. е. площадки могут быть квадратными (например, 1×1,2×2 м), прямоугольными (например, 1×2,2×3,2×4 м), круговыми и т.д.. Площадки размещают по площади или рядами, когда они лежат в один ряд, т. е. на одной

линии как в продольном, так и в поперечном направлении, или же в шахматном порядке. Это зависит от наблюдателя. Расстояние измеряется от центра площадок.

Такой способ применяют, например, при учете естественного возобновления. При учете культур площадками за учетные площадки принимают посадочные или посевные места. Последние для наблюдения можно отбирать или целыми полосами, расположенными через одинаковые расстояния друг от друга, или же отдельными площадками, когда в пределах полосы обследуют не все площадки, а лишь через одно и то же их количество, например - каждую пятую, каждую десятую и т.д., независимо от того, какой по своему состоянию она окажется.

Если число площадок N, необходимых для получения результата с заданной точностью, уже известно, то расстояние l между центрами площадок в м определяется по формуле

$$l = \sqrt{\frac{П}{N}},$$

где П - размер обследуемой площади, м^2.

Наметив визиры (ходовые линии обследования), расположенные друг от друга на вычисленном расстоянии l, каждый из них разбивают на отрезки длиной тоже в l метров; в конце этих отрезков (или только справа, или только слева, или только на самом визире) закладывают площадки, где бы в натуре они ни пришлись, пусть даже на дороге или на прогалине. Передвигать их нельзя.

Так же можно закладывать ямы для определения зараженности почв, например майским хрущем, площадки на стволе для изучения степени заселенности и поражения вредителями-насекомыми и т.п.

Обследуемый участок предварительно можно разбить на типические части, однородные по какому-либо признаку, и каждую типическую часть обследовать отдельно (типическая выборка). Но разбивку целого на типические части для их характеристики можно производить камерально - по материалам наблюдения. Например, обследуя большую неоднородную площадь лесных культур ее следует разделить на относительно однородные части.

3. С п о с о б в и з и р о в. На обследуемой площади через одно и то же расстояние один от другого пробивают визиры. На каждом визире учетные единицы (площадки, экземпляры и т. п.) для наблюдения можно отбирать двумя способами.

- В каждую энную (например, десятую, двадцатую и т.д.) по ходу, но только всегда или слева, или только справа, независимо от того, какой по своему состоянию окажется отобранная единица наблюдения. Следует заметить, что принцип отбора энной (пятой, десятой и т.д.) по порядку единицы наблюдения можно применять, например, и при отборе веток деревьев для изучения плодоношения и в ряде других случаев; при сплошных перечетах - отбор для более детального изучения каждого энного дерева подряд. Так берут модельные и учетные деревья.

- Учетные единицы намечают через определенное число метров, например - 10; 20; 50 м и т.д. - в зависимости от величины обследуемой площади. Здесь обследуется та единица, которая окажется как раз в конце каждого такого отрезка визира или в ближайшей к этой точке. При этом соблюдается правило отбора, указанное в предыдущем пункте. Число отрезков, на которое будут разбиты все визиры, равняется числу отбираемых единиц наблюдения. Следовательно, зная необходимое число единиц наблюдений, задавшись расстоянием между визирами, можно подсчитать число и общую длину визиров и разделить последнюю на принятое число единиц наблюдения. В результате получим требуемое расстояние по визиру между отбираемыми единицами, или иначе - длины отрезков в конце которых и производится наблюдение. Отобранные таким путем деревья носят название модельных, если они срубаются, и учетных - если они не срубаются.

Описанным приемом можно отбирать деревья для изучения процента выхода деловой древесины, процента зараженных деревьев и т.д., подвергая глазомерному наблюдению (описанию), скажем, каждое десятое дерево. Для детального изучения со срубкой дерева можно брать, например, каждое двадцатое, сороковое и т.д., исходя из того, какое количество деревьев требуется для детального обследования.

- Задачей наблюдения может стоять выявление влияния на объект наблюдения, например на дерево, окружающей среды, скажем, сомкнутой биогруппы, открытого места, полога леса и т.д. Эти особенности среды могут также влиять на характер возобновления, на состав и состояние живого напочвенного покрова, на плодоношение разных растений, на рост и развитие леса и т.д. Тогда учетные площадки одного и того же размера закладывают в этих типических условиях, влияние которых требуется изучить; они нередко подвергаются повторным периодическим наблюдениям, например через каждые 5 или 10 лет.

При сборе материала следует создать такой фундамент из точных и бесспорных фактов, на который можно было бы опираться, с которым можно было бы сопоставлять результаты других исследователей; Поэтому необходимо анализировать не отдельные факты, а всю совокупность относящихся к рассматриваемому вопросу фактов, без единого исключения, в их связи и без вырывания отдельных фактов и цифр из общей связи явлений.

При производстве наблюдений важным является разработка формы записи данных наблюдения с четким перечнем всех признаков, подлежащих учету, и указанием единиц и точности измерения. От качества формы записи зависит и полнота получаемых сведений. Форма записи может быть списочная, когда в ведомость заносятся данные наблюдений по каждой единице наблюдения в порядке обследования, и карточная, когда все данные о каждой единице наблюдения заносятся в отдельную карточку или бланк. Характерным примером здесь будет карточка таксации модельного дерева. Карточная система записи удобнее в том отношении,

что сильно облегчает камеральную обработку при сводке и группировке материала по разным признакам. В настоящее время все чаще для записи используют носители информации для компьютеров, которые вставляют в специальные устройства. В этом случае информация сразу вводится в ПК, исключая ее набор, что многократно ускоряет процесс обработки информации.

Перед началом исследований разрабатывается методика производства наблюдения: уточняется единица наблюдения, если она сложная, определяются ее размеры и форма (ленты, пробные площадки и площади); устанавливается точность конечного результата и необходимое количество единиц наблюдения, способ производства наблюдения, отбора единиц; оформление единиц наблюдения в натуре; сроки наблюдений и единицы измерений. Составляется календарный план работ и смета расходов.

Обобщая изложенное о проведении наблюдений, можем сделать такое заключение.

- Наблюдение является опытной основой статистического исследования.
- Для того, чтобы по данным выборки можно было бы с определенной степенью уверенности делать заключения о совокупности, выборочное наблюдение должно быть правильно организовано. Здесь решают два основных вопроса:

- какое число наблюдений является достаточным;
- какие единицы совокупности должны быть выбраны для наблюдения, т.е. что (кто) будет составлять выборку.

Первый вопрос может быть решен с помощью таблицы достаточно больших чисел, которая приводится во многих пособиях по статистике, или с применением специальных формул. Они будут описаны ниже.

- Отбор единиц для наблюдения может быть спланирован различным образом в зависимости от состава совокупности и сведений о ней.
- Если совокупность варьирует не в слишком широких пределах и если выборка составляет не менее 20% объема совокупности, применяют простой случайный отбор единиц или простое выборочное наблюдение. Для этого удобно пользоваться таблицей случайных чисел.
- В лесном хозяйстве воспользоваться таблицей случайных чисел в натуре технически тяжело и практически часто невозможно. Поэтому здесь пользуются систематическим выборочным наблюдением. Например, если предстоит взять 10%-ную выборку деревьев из 800 штук, то случайным порядком выбирают первое, положим 5 дерево, и после этого берут каждое следующее через 10 номеров, т.е. за номерами: 15, 25, 35 и т.д., кончая номером 795.
- Если имеются сведения о том, что совокупность в своих частях неодинакова, например с более высоким уровнем явления в одних частях чем в других, целесообразно послойное выборочное наблюдение.

Например, известно, что запас древостоев меньше варьирует в пределах классов возраста. Тогда для получения статистических характеристик величины запаса всю совокупность древостоев расчленяют на группы по возрасту. Получим слои совокупности, для каждого из которых берут независимую выборку и вычисляют ее характеристики. Статистическая обработка материалов опыта послойной выборки несколько сложнее.

Но только наблюдениями и их статистической обработкой не ограничивается сбор информации в лесоводственных и иных исследованиях. Очень часто в дополнение к наблюдениям ставится эксперимент. В совокупности наблюдение и эксперимент практически исчерпывающие источники первичной информации в лесном хозяйстве.

Наблюдения обычно не требуют вмешательства в нормальное функционирование объекта. Во многих лесоводственных исследованиях они являются единственно возможными, например, фенологические наблюдения, изучение роста деревьев и древостоев, приживаемости лесных культур и др. Однако определенная “пассивность” наблюдения по отношению к объекту исследования не предполагает отсутствия плана или системы: наблюдение как метод научного познания предполагает наличие строгого плана. Хорошей иллюстрацией планируемых наблюдений являются выборочные методы инвентаризации лесных ресурсов на больших территориях, которые проводятся во многих странах.

Эксперимент предполагает активное и целенаправленное воздействие на изучаемый объект или явление, определенную управляемость условий его проведения. Соотношение роли наблюдения и эксперимента достаточно сложное. В научном познании задача наблюдения обычно более скромна и сводится чаще к описанию и анализу наблюдаемых явлений и процессов. В эксперименте сильнее теоретическая сторона, уровень осмысливания наблюдаемых факторов; эксперимент располагает средствами активного вмешательства в ход событий. Однако в лесном хозяйстве, особенно при изучении природных объектов, наблюдение часто играет более важную роль, чем эксперимент.

Для удобства классификации можно выделить обычный модельный эксперимент и математически спланированный или экстремальный. Обычный модельный эксперимент отличается выделением изучаемых связей и изоляцией их от внешних воздействий; при этом он может быть однофакторным и многофакторным. Например, берем сеянец, помещаем его в искусственную среду и исследуем влияние на его рост некоторого удобрения. Если же этот сеянец наблюдать в естественных условиях, то надо учесть и осадки, и температуру и другое, т.е. много факторов, а не одно удобрение. Математическое планирование эксперимента (многофакторное) позволяет оптимизировать сам процесс исследования: заранее выбрать наилучшую (с точки зрения цели работы) математическую модель, применить последовательную стратегию и скорректировать направления исследований после каждого этапа и т.д.

При любом методе сбора информации ее обработку и использование строят по схеме: информация - гипотеза - модель - проверка соответствия модели исходной информации и объекту, для которого разработана модель (рисунок 2.2).

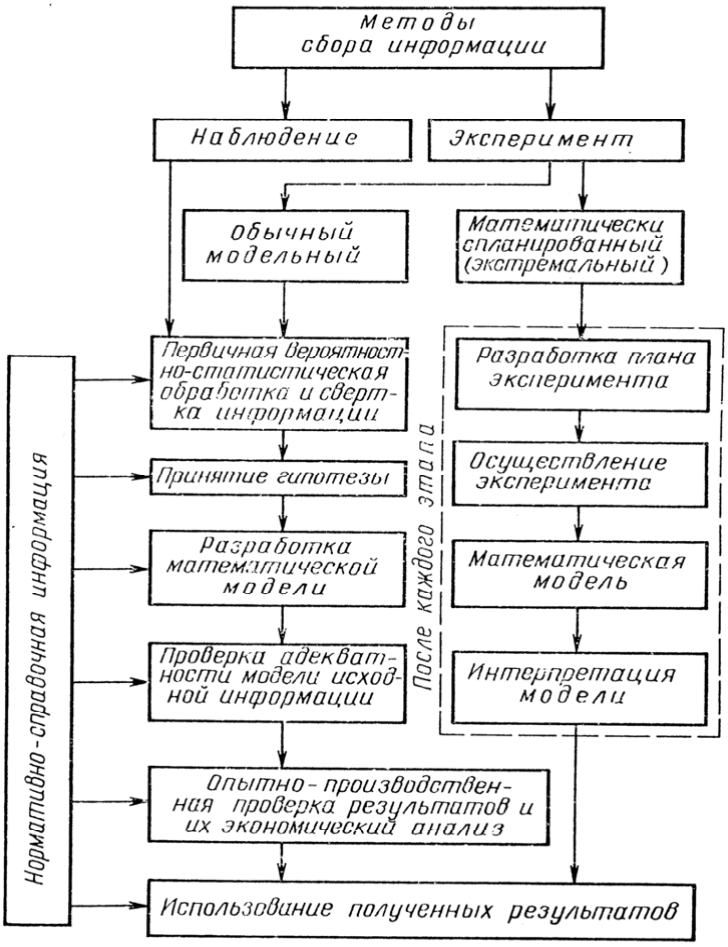


Рисунок 2.2 Упрощенная схема движения лесоводственной информации (по К. Е Никитину и А. З. Швиденко)

Следует подчеркнуть важность последнего этапа, нередко недооцениваемого. Возможность применения модели в конкретной ситуации требует обязательного доказательства, которое может быть вероятностно-статистическим, если в процессе исследования не нарушались основные статистические предпосылки организации сбора информации, или эмпирическим, т.е. на основе дополнительно собранной контрольной информации.

В основе получения первичной численной информации лежит, как правило, процесс измерения - нахождения значений физической величины опытным путем с помощью специальных технических средств. В настоящее время существует два подхода к измерительному процессу: классический и информационный. В большинстве задач лесного дела выполняются основные предпосылки классического подхода к измерениям: измеряемая величина предполагается неизменной на протяжении времени

измерений и характеризуется одним значением, для которого можно указать интервал неопределенности, т.е. ошибку измерений; время измерения практически не ограничено; внешние условия и факторы, влияющие на результат измерения, учтены полностью. Информационная модель измерительного процесса трактует измерение как случайный процесс, т.е. позволяет оценивать качество измерения величин, меняющихся во времени.

Измерение может быть прямым и косвенным. В первом случае изучаемую величину измеряют непосредственно, во втором - наблюдают не изучаемую величину, а другую, которая с ней связана и которую проще измерить. Так, объем растущих деревьев обычно определяют измерением их диаметра и высоты, а объемный прирост находится путем измерения радиального прироста, энергии роста и т.д. Переход к величине, являющейся предметом изучения, происходит при помощи математических моделей связи. Более подробно это будет изложено ниже.

2.4 Дедуктивный и индуктивный методы в биометрии

В лесной биометрии применяют как дедуктивный (от общего к частному), так и индуктивный методы исследований.

Дедуктивный метод применяется, когда, хотя бы ориентировочно, известны общие закономерности изменения случайной величины. Так, мы знаем центральную предельную теорему, доказанную русским математиком и механиком А. Ляпуновым (1857 – 1918) в 1901 году, которая гласит, что распределение суммы независимых случайных величин (i=1, 2, ..., n) стремится к нормальному распределению при неограниченном увеличении N, если все величины имеют конечные средние и дисперсии и ни одно из них по своему значению резко не отличается от других. Руководствуясь этой теоремой, можно рассматривать распределения, скажем диаметров стволов в древостое, как частный случай проявления названной закономерности и использовать кривую нормального распределения для прогноза строения древостоя, определяя параметры конкретного древостоя по проведенным наблюдениям.

Но в лесном хозяйстве чаще приходится использовать индуктивный метод исследований, т.е. от частного к общему. Выше уже упоминалось, что при статистических наблюдениях в биологии практически всегда имеют дело с выборками и по результатам их судят о совокупности. Таким образом, вариационная статистика (биометрия) применяет метод индукции, когда обобщения делают, изучив отдельные случаи. Правомерность этого метода основана на использовании важнейших понятий и положений теории вероятностей. В качестве примеров можно привести уже упомянутые зависимости между диаметрами и высотами в насаждении. Сделав анализ ряда выборок из древостоев разных древесных пород, отличающихся также и возрастом, мы придем к выводу, что в одном случае для аппроксимации и прогноза следует использовать уравнение пара-

болы 2 порядка, в другом случае -3 порядка, в третьем - некоторую более сложную кривую. Названную закономерность мы получаем, проанализировав ряд частных случаев (отдельных насаждений), т.е. идем от частного к общему, применяя индуктивный метод.

Индуктивное заключение, как общий логический процесс, идущий от большой и малой посылки, имеет такую форму:

Б о л ь ш а я п о с ы л к а : эти коричневые желуди (отборный образец) из данного хранилища.

М а л а я п о с ы л к а : эти желуди находятся в хранилище.

З а к л ю ч е н и е : все желуди в хранилище коричневые.

Очевидно, что заключение, сделаное с индуктивной аргументацией шире, чем посылки. В заключении добавляется нечто новое, расширяющее знания об изучаемом явлении. Это потенциальное расширение знаний требует осторожности. Оно может быть плодотворно, но существует некоторая опасность получить необоснованные и ложные выводы.

Логическим основанием индуктивного заключения является предположение о единообразии в системе фактов, относящихся к посылкам и заключению. Это предположение, называемое по-разному - единообразием в природе, статистической устойчивостью опыта, ограничением независимой вариации в природе, - всегда представляет как бы невысказанную посылку индукции.

Если бы единообразие в естественных процессах не проявлялось, природе был бы свойственен полный хаос. При этом никакое нагромождение фактов не могло бы оправдать индукцию. Нельзя было бы ничего сказать об условиях за пределами опыта. Но в природе существует определенное единообразие в поведении отдельных единиц, составляющих то или иное массовое явление. Однако это единообразие в природе не столь строго, чтобы можно было сделать точную оценку массового (общего) явления наблюдаемых единиц. Поэтому 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.

Следует отметить, что конструкция выборочных оценок оказывается более предпочтительной даже в тех случаях, когда все единицы, составляющие то или иное явление, могут быть измерены, т.е. относятся к ограниченным генеральным совокупностям. Это положение, затронувшее различные виды генеральных совокупностей, нуждается в более широком пояснении.

На практике встречаются обследуемые генеральные совокупности конечные и бесконечные. Примером первой может служить выборочное обследование, допустим, бюджетов семей в определенном городе. С бесконечными совокупностями имеют дело при различных экспериментальных исследованиях, когда вопрос заключается не в том, чтобы получить точный результат в данном эксперименте, но главным образом в оценке

того, каковы будут результаты массового применения данного процесса (в % от обследованных единиц) - биологического, технологического или экономического.

Предположим, производится оценка степени повреждаемости подроста на нескольких десятках лесосек при применяемой технологии лесосечных работ. В этом случае генеральная совокупность бесконечна, ибо для оценки не столь важно, сколько повреждено подроста на конкретных лесосеках, как то, сколько его будет повреждаться при подобных условиях на всех других лесосеках, не исследованных в опыте. Здесь научный эксперимент становится как бы “механизмом” получения случайной выборки.

Возможны обстоятельства, когда полезно прибегнуть к особой логической конструкции - гипотетической генеральной сверхсовокупности. Иногда мы можем располагать сведениями даже сплошного обследования реально существующей совокупности, и все же бывает полезно рассматривать эти данные как выборку из некоторой сверхсовокупности. Так поступают, когда не только нужны полученные факты, но и необходимо выявить общую закономерность, по отношению к которой статистический материал представляется лишь частным случаем.

Предположим, что из статистических обследований за 2005 – 2010 годы рождаемости в стране установлено, что 52% из числа родившихся составили мальчики. Этот материал получен путем сплошного обследования и характеризует явление однозначно. Однако, если нас интересует результат и за пределами обследованных лет или проверяется заключение о том, что мальчиков всегда рождается больше, тогда полученные данные следует рассматривать как выборку из некоторой бесконечной сверхсовокупности различных возможных пропорций рождений по полу. На основе таких сведений, пользуясь методами статистики, представляется возможным исследовать, приемлемо ли предположение о более частой рождаемости мальчиков. Заметим, что определяемая таким образом сверхсовокупность не ограничена ни численностью, ни территорией, в которой произведен эксперимент.

Из приведенных примеров видно, что в биометрии (в т.ч. в лесной биометрии) применяются оба метода (индуктивный и дедуктивный), но преобладает индуктивный.

3. ГРУППИРОВКА ИСХОДНЫХ ДАННЫХ

3.1 Количественный и качественный анализ массовых явлений

При рассмотрении массовых явлений, когда имеем дело с большими массивами информации (деревья в лесу, партии семян, явления общественной жизни и т.д.) их анализ может быть качественный или (и) количественный.

Качественный анализ изучаемого явления или процесса заключается в выделении некоторых его характерных свойств, особенностей, признаков, качественно отличающихся между собой внутри рассматриваемой совокупности. Например, мы изучаем смешанное насаждение, состоящее из сосны и березы. Заложили пробную площадь, на которой насчитали 180 деревьев сосны и 110 березы. Для дальнейшего использования этого материала необходимо сделать предварительный качественный анализ. В нашем случае разделить деревья по главному качеству - принадлежности к разным ботаническим видам, т.е. на сосну и березу. Приведенный пример прост. В реальности бывает, что сделать качественный анализ трудно, для чего применяются специальные методы, которые рассматриваются ниже.

Но качественного анализа часто бывает недостаточно, чтобы понять некоторое явление или процесс, дать ему корректное математическое описание. В этом случае необходимо использовать количественные методы исследования.

Проведение качественного и количественного исследования начинается с планирования и постановки эксперимента. В лесном хозяйстве исследование часто заключается в проведении измерений на некоторых выделенных участках леса (пробных площадях) или в измерении части дерева. Возможны и другие варианты экспериментов, что мы уже отмечали выше и к чему еще вернемся при изложении настоящего курса.

При научном или практическом исследовании некоторого явления или процесса требуется выяснить его природу, закономерности изменения во времени или связь с некоторыми параметрами. Для этого недостаточно повести наблюдение или поставить эксперимент. Вывести искомые законы и закономерности изучаемого явления, получить правильные выводы из наблюдений можно только в том случае, если будет сделан корректный количественный и качественный анализ проведенных наблюдений, или, как их еще называют, случайных явлений.

Наиболее часто для анализа используются количественные методы. Здесь смотрят наличие сходства или различия, пользуясь определенными числовыми критериями.

3.2 Систематизация и группировка исходных данных

Любой анализ проведенных наблюдений начинается с систематизации наблюдений. Первым ее этапом является группировка исходных данных или вариантов. При постановке эксперимента группировку предусматривают уже на этапе сбора экспериментального материала, т.е. на этапе наблюдений.

Например, замеряя высоты в 6-8-летних культурах сосны, можем записывать отдельно каждое измерение как приведено ниже.

Высоты в м: 0,8; 1,5; 3,1; 2,2; 0,4; 1,1; 1,6; 1,4; 1,9; 2,0; 1,8; 2,4; 2,7; 2,9; 0,9

Анализ приведенных величин, хотя их относительно немного, в представленном виде затруднен. При больших массивах информации анализ отдельных измерений перерастает в большую проблему. Для ее решения результаты наблюдений, как правило, систематизируют. Систематизация заключается в группировке измеренных величин: толщины или высоты деревьев, веса животных, размера и веса семян и т.п. Для группировки наблюдений выделяют классы (при измерениях деревьев их называют ступени), по которым разносят измеренные величины. В приведенном примере целесообразно выделить следующие классы высот:

0 - 0,50; 0,51 - 1,0; 1,01 - 1,50; 1,51 - 2,00; 2,01 - 2,50; 2,51 - 3,0; 3,01 - 3,50

Тогда запись измерений можно будет свести в таблицу (таблица 3.1).

В таблице не обязательно показывать сам интервал, достаточно привести значение его середины.

Таблица 3.1 – Распределение высот в молодняке сосны

Ступень высоты, м	Число деревьев
0,25	1
0,75	2
1,25	3
1,75	4
2,25	2
2,75	2
3,25	1
Итого	15

Данные, сведенные в таблицу, дают более наглядное представление о высоте культур сосны на исследуемом участке. К тому же обработку материала легче проводить, когда имеем систематизированные данные. Поэтому именно такая табличная форма наиболее часто используется при проведении исследований в лесном хозяйстве.

При упорядочении (систематизации) полученных данных легко обработать их математически и вывести статистические показатели, которые будут исчерпывающе характеризовать изучаемую совокупность. Проблема систематизации и группировки занимает большое место в статистике. Ошибочная группировка данных может привести к неправильным выводам о существе изучаемого явления.

Наиболее проста группировка при качественном анализе.

Так, если кора осины отличается по окраске, то распределение деревьев с разной корой может быть выражено в процентах от общего количества измеренных деревьев, как это показано в таблице 3.2.

Таблица 3.2 – Распределение деревьев осины по цвету коры

Цвет коры	Количество деревьев	Процент от общего количества
Темно-серый	160	40
Светло-серый	200	50
Зеленый	40	10
Итого	400	100

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 исследовании сосновых культур, пораженных корневой губкой, мы делим деревья по альтернативному признаку: здоровые и больные.

3.3 Составление рядов и таблиц распределения

При проведении наблюдений в лесном хозяйстве чаще всего имеют дело с непрерывным дискретным (прерывистым) распределением изучаемой величины. Так, измеряя толщину дерева, мы меряем каждое дерево. Их совокупность представляет собой некоторый ряд распределения, у которого есть минимальная и максимальная величина. Для примера в таблице 3.3 приведены результаты измерения 120 деревьев сосны II класса бонитета в типе леса сосняк мшистый возрастом 100 лет.

Для того, чтобы придать опытным материалам определенную наглядность и извлечь из них необходимую статистическую информацию о наблюденном признаке, материалы наблюдения подвергают группировке. Сгруппированные материалы именуют статистическими таблицами или статистическими рядами.

Таблица 3.3 - Результаты измерений диаметра 120 деревьев сосны на высоте 1,3 м

№ дерева	Диаметр, см	№ дерева	Диаметр, см	№ дерева	Диаметр, см	№ дерева	Диаметр, см	№ дерева	Диаметр, см
1	23	25	28	49	24	73	25	97	44
2	32	26	40	50	28	74	32	98	27
3	19	27	32	51	27	75	46	99	23
4	28	28	15	52	34	76	31	100	19
5	24	29	27	53	29	77	36	101	28
6	19	30	38	54	36	78	28	102	32
7	25	31	30	55	30	79	35	103	24
8	22	32	26	56	44	80	43	104	44
9	29	33	24	57	37	81	24	105	30
10	27	34	21	58	24	82	48	106	20
11	23	35	31	59	20	83	32	107	29
12	20	36	20	60	31	84	23	108	34
13	30	37	28	61	27	85	25	109	28
14	36	38	28	62	28	86	37	110	26
15	42	39	17	63	22	87	12	111	17
16	24	40	34	64	13	88	42	112	33
17	32	41	20	65	28	89	28	113	29
18	38	42	29	66	41	90	26	114	21
19	40	43	16	67	18	91	47	115	36
20	33	44	33	68	26	92	9	116	30
21	35	45	37	69	36	93	39	117	32
22	20	46	24	70	33	94	25	118	40
23	16	47	16	71	28	95	51	119	20
24	39	48	40	72	11	96	48	120	15

Статистическим рядом или рядом распределения называют ряд значений признака, размещенных в порядке возрастания или убывания, с указанием числа повторений.

Значения признака, сведенные в ряд, называют классовыми вариантами, а число повторений их в классах - численностями, или частотами классов. При измерении деревьев в лесу классы обычно называют ступенями толщины или ступенями высоты.

Статистический ряд значений измеренного признака получают путем определения величины класса или интервала, размещения классов и распределения в них всех единиц наблюдения.

Величину интервала (k) определяют по формуле:

$$k = \frac{X_{max} - X_{min}}{t}, \text{ где}$$

X_{max} и X_{min} - соответственно наибольшее и наименьшее значение признака или вариант; t – число принимаемых классов.

В качестве k принимают круглое число, ближайшее к полученному частному. При этом действительное число классов определится как частное от размаха вариант (X_{max} - X_{min}) на округленное значение интервала. В качестве полученного частного принимают также круглое число. Округление делается всегда в большую сторону.

При проведении биометрических исследований оптимальным количеством классов при наличии большой выборки считается 12. Допустимо увеличивать или уменьшать это количество в зависимости от объема наблюдений. Если ряд наблюдений относительно невелик (40-60 измерений), а разница между X_{min} и X_{max} невелика, то его целесообразно сузить, т.к. в противном случае ряд распределения окажется размытым, в некоторых классах может не оказаться измеренных величин.

Обычно рекомендуемое число классов равно 12±3, т.е. колеблется от 9 до 15. В отдельных случаях допустимо уменьшить количество классов до 8, как исключение – до 7. Меньшее и большее количество классов принимать не рекомендуется, если это не связано со специфическими особенностями экспермента.

Границы и срединные значения классов лучше устанавливать следующим образом. В качестве среднего значения первого класса принимают число кратное классовому промежутку k, ближайшее к наименьшей (в возрастающем ряду) или наибольшей (в убывающем ряду) вариантам ряда распределения. Срединные значения последующих классов получают путем последовательного прибавления величины интервала.

Нижние границы классов определяют путем вычитания половины величины интервала из срединных значений каждого класса, а верхние границы - путем прибавления этой половины.

В целях исключения перекрытия верхней границы предыдущего класса с нижней границей последующего класса, входящих в первый и второй классы, нижние границы классов увеличивают на некоторую величину, равную точности измерения признака. Можно также верхние границы классов уменьшить на ту же величину, но первый вариант используют чаще. Именно так получаем значения границ классов. Например, при измерении диаметров по 4 см классам (ступеням) толщины или высоты величина увеличения (уменьшения) обычно равна 0,1 см, для высот 0,1 м. Например, ступень толщины, равная 24 см, будет иметь пределы от 22,1 см до 26,0 см. Как вариант может быть от 22,0 см до 25,9 см, но первый пример предпочтительней.

Проиллюстрируем изложенное на конкретном примере. Возьмем данные измерений 120 диаметров сосны, приведенные в таблице 3.3.

В нашем примере наименьший измеренный диаметр равен 9 см (дерево №92), наибольший – 51 см (дерево №95). Разница составляет 42 см. Величина классового интервала для оптимального случая составит k = 42/12 = 3,5

см. Округлив его в большую сторону, получим величину классового интервала (ступень толщины) равную 4 см.

Руководствуясь изложенными правилами установления первого и последнего классов, получаем соответственно первый класс, равным 8 см (ближайшее число к 9, кратное 4) и последний в 52 см – ближайшее число к 51 см.

Так как величины диаметров, лежащих на границе классового промежутка, можно отнести к любому из соседних классов (например, дерево диаметром 22 см можно отнести к классовому промежутку со срединой класса и 20 см, и 24 см), то целесообразно нижние границы классов увеличивать на 0,1 см. Можно аналогично верхние границы классов уменьшать на 0,1 см, но, как отмечено выше, первый вариант удобнее. В этом случае классовый интервал со срединой 20 см будет равен 18,1-22,0 см, а со срединой в 24 см соответственно 22,1-26,0 см. В этом случае дерево толщиной 22 см однозначно будет отнесено к ступени толщины 20 см.

После установления классов разносим измеренные диаметры по классам или, как их называют в лесном деле, по ступеням толщины. При этом применяют символические отметки для учтенных деревьев от 1 до 10 – запись “конвертом”. Вид этой символической записи показан в таблице 3.4.

Таблица 3.4 – Символическая запись вариант

Величина вариант	1	2	3	4	5	6	7	8	9	10	11	12	13	…	20 и т.д.
Символы	·	· ·	· · ·	· · · ·	— · ·	⌐ ·	⌐	□	⊠	⊠	⊠ ·	⊠ · ·	⊠ · · ·	…	⊠ ⊠

Разнесенные варианты (численности) толщин деревьев с указанием классов показаны в таблице 3.5.

Таблица 3.5 – Таблица распределения диаметров 120 стволов сосен

Ступени толщины (классы), X_i, см	Численности (число деревьев), n_i, шт.
8	1
12	3
16	8
20	14
24	20
28	27
32	17
36	12
40	9
44	5
48	3
52	1
Итого (∑)	120

Вариационный ряд, рассмотренный нами в качестве примера, является одновершинным по распределению. Это значит, что он имеет один модальный класс. Возможны случаи, когда в вариационном ряду обнаруживается несколько модальных классов, и тогда полигон является многовершинным. Наиболее простой причиной многовершинности, особенно при очень растянутых рядах, является недостаточное количество вариант в изученной совокупности. При малом числе особей в некоторых классах вариационного ряда может вообще не быть ни одной варианты. Вариационный ряд окажется с перерывами, а вариационная кривая - разорванной на части.

Однако, если и при большом числе особей в изучаемой совокупности наблюдается дву- или многовершинность, причину этого надо искать в самом экспериментальном материале. Последний в таком случае обычно представляет собой смешение двух качественно различных совокупностей, которые или находились в резко отличных условиях внешней среды, или принадлежат к разным типам, например, к древостоям разных поколений. Соединение в одном ряду особей разных древостоев может дать внешнюю картину дву- или многовершинности. Известно, например, что абсолютно разновозрастные древостои ели или кедра сибирского дают многовершинное распределение. Поэтому в один вариационный ряд помещают лишь деревья одного поколения.

Правда, возможны случаи, хотя они относительно немногочисленны, когда дву- или многовершинность определяется свойствами самих изучаемых признаков и поэтому характеризует вполне однородный материал. Определение этой однородности или неоднородности – прерогатива специалиста: лесовода или биолога.

3.4 Прогнозирование случайной величины

Случайные величины и их прогнозирование являются основным объектом изучения в биометрии. Прогнозирование случайных величин основано на теории вероятности.

Теория вероятности – это одна из дисциплин математики. Подробно ее изучают на математических, физических и некоторых технических факультетах. В настоящем пособии описаны лишь некоторые положения теории вероятности, взятые из соответствующих учебников, приведенных в списке литературы. Объем изложения в данном виде хотя и невелик, но позволяет лесоводу и биологу в достаточной мере разбираться в основных понятиях, которые будут излагаться при дальнейшем изучении биометрии.

Одним из основных понятий этой теории является вероятность. **Вероятностью события А** называют отношение числа случаев, благоприятствующих появлению данного события к числу всех возможных случаев. Обозначим вероятность буквой Р с указанием в скобках индекса события, в нашем случае события А. Ее определяют по формуле:

$$P(A)=n/N, \text{ где}$$

n – число случаев, благоприятствующих событию А;

N - общее число случаев.

Так, если в урне содержится 5 одинаковых перемешанных шаров, причем 2 из них черные, а 3 – белые, то вероятность вынуть наудачу белый шар равна P(A)=3/5=0,6, а вероятность вынуть черный шар P(B)=2/5=0,4.

Вероятность изменяется от нуля до единицы. Вероятность, равная нулю, указывает, что событие является невозможным; вероятность, равная единице, означает, что событие единственно возможное или достоверное.

Если появление одного события исключает появление другого события, их называют **несовместными**.

В указанном примере события А и В – несовместные. Сумма вероятностей несовместных событий равна единице P(A)+P(B)=1.

События называют **равновозможными**, если ни одно из них не является более возможным, чем другие. Вероятности таких событий одинаковы.

В лесобиологических работах вероятность чаще всего установить невозможно, так как вся изучаемая генеральная совокупность и ее состав неизвестны, например, число деревьев разной толщины в большом участке леса. В таких случаях получают аналог вероятности на основе опыта, т.е в выборочной совокупности. Для этого подсчитывают число испытаний, в которых событие практически появилось и относят его к общему числу испытаний.

Это отношение называют относительной частотой события и выражают формулой:

$$W(A)=\frac{n}{N}, \text{ где}$$

n – число появлений события;

N – общее число испытаний.

Длительные наблюдения показали, что при одинаковых условиях испытаний и достаточно большом их числе относительная частота в различных опытах изменяется мало, причем тем меньше, чем больше объем выборки. Она колеблется (варьирует) около некоторого постоянного числа. Это замечательное свойство относительных частостей называется **устойчивостью относительной частоты**, или статистической устойчивостью.

Очень характерный пример связан с рождением людей разного пола. Раньше мы приводили пример с рождением мальчиков. Сейчас возьмем вероятность рождения девочек и приведем ее в качестве примера устойчивости относительной частоты. По месяцам за некоторый год, начиная с января, она характеризуется следующими значениями:
0,486; 0,489; 0,471; 0,478; 0,482; 0, 462; 0,484; 0,485; 0,491; 0,482; 0,473.
По мальчикам величины будут зависимы от рождения девочек, т.к. сум-

марная вероятность рождения лиц обоего пола почти равна 1,0. Правда, есть еще гермафродиты, но их очень мало, и этим показателем обычно пренебрегают.

Постоянное число, около которого варьируют относительные частоты, является вероятностью появления события. Таким образом, если опытным путем установлена относительная частота, то полученное число можно принять за приближенное значение вероятности.

Относительные частоты в указанном примере колеблются около числа 0,482, которое можно принять за приближенное значение вероятности рождения девочек.

Следует обратить внимание, что такое суждение о вероятности на основе относительной частоты тем надежнее, чем больше число испытаний или объем выборки.

Наиболее простым и убедительным примером для подтверждения этого положения является бросание монеты. Многократные опыты с монетой, в которых подсчитывали число появления герба, дали следующий результат:

число бросаний: 4040; 12000; 24000;

относительная частота: 0,5069; 0,5016; 0,5005;

Вероятность появления герба равна 0,5.

Нетрудно представить или испытать на опыте, что при малом числе наблюдений, например, при 6 бросаниях, такое приближение относительной частоты к вероятности, как правило, не получить.

Проведя анализ случайных величин необходимо выполнять ряд требований. О некоторых из них мы уже упоминали.

Требования репрезентативности:

А) выборочная совокупность должна характеризовать собой генеральную с определенной точностью;

Б) выборочная совокупность должна быть свободной от субъективных представлений о генеральной.

Иначе говоря, удовлетворять требованиям репрезентативности - значит всесторонне и соответственно характеризовать генеральную совокупность.

В зависимости от числа наблюдений (N) выборочная совокупность может называться большой или малой. Границей здесь является N, равное 30 наблюдениям.

Основные этапы изучения статистических совокупностей:

1. Составление программы, важнейшими элементами которой является цель и задачи исследования.

2. Составление методики исследования, которая содержит выбор и обоснование места, времени и учетного признака на объекте исследования. Способы учета, необходимое число наблюдений, форма записи, вы-

бор инструмента и единица отсчета, способы последующей обработки материалов исследования - все это предмет методики исследований.

3. Производство наблюдений, измерений или учета.

4. Первичная обработка результатов наблюдения.

5. Моделирование изучаемого явления.

6. Дополнительное производство наблюдений, доводка модели и исследование с помощью модели, повторная обработка результатов исследования.

7. Систематизация и анализ полученных данных.

Эта последовательность в общих чертах соблюдается как в научных исследованиях, так и при решении производственных задач. В этой последовательности необходимо выполнять и индивидуальное задание, не упуская из виду цели и физического смысла результатов исследования, которые предопределены заданием.

Правила вычисления результатов. Правила вычисления результатов представлены согласно принципу Крылова-Брадиса (П.М. Крылов – 1879-1955 – советский математик) и приводятся в сокращении.

Правила 1-5. При сложении и вычитании, умножении и делении, возведении в квадрат или куб, извлечении корня квадратного или кубического, при использовании логарифмов - в результате нужно сохранять столько десятичных знаков после запятой, сколько их имеет “слагаемое” с наименьшим количеством десятичных знаков.

Правило 6. Для промежуточного результата, получаемого по правилам 1-5, необходимо сохранить одну дополнительную “запасную” цифру; в конечном результате ее отбрасывают.

Правило 7. Если исходные данные имеют разное количество десятичных знаков или значащих цифр, то их надо предварительно округлить с сохранением одной “запасной” цифры.

Правило 8. Если результаты должны быть получены с n-значащими цифрами, то исходные данные следует брать с n+1 значащей цифрой.

Одним из существенных условий правильно и хорошо организованного вычислительного процесса является аккуратность и тщательность записи.

Курс биометрии не предусматривает детального изучения теории вероятности. В то же время те студенты, которые намерены в будущем заняться научной работой, могут факультативно проработать соответствующий материал из учебников по теории вероятностей. Некоторые из них приведены в списке литературы.

4. СРЕДНИЕ ЗНАЧЕНИЯ

4.1 Статистические показатели вариационного ряда и их классификация

Важным показателем статистического ряда является его размах. Это наиболее простой показатель, показывающий разность между наибольшими и наименьшими величинами (их еще называют лимитами) в исследуемом вариационном ряду, т.е.

$$L = x_{max} - x_{min},$$

где L – размах ряда,

x_{max}, x_{min} – максимальная и минимальная величины (лимиты).

Простота вычисления размаха ряда распределения, его наглядность и очевидность способствовали широкому употреблению этого показателя во многих исследованиях в лесном хозяйстве. Так, в лесной таксации при изучении строения древостоя размах ряда распределения – это обязательный показатель.

В биометрии размах ряда и лимиты определяются при сведении данных наблюдений или измерений в статистические совокупности, т.е. в вариационные ряды. Каждый вариационный ряд и его графическое изображение – это как бы «сгущение» исходного фактического материала, превращение его в наглядную формулу. Однако для полного анализа наблюдаемого явления или для хозяйственной оценки древостоя этого недостаточно. Дело в том, что размах ряда и лимиты подвержены значительным колебаниям от одной частичной совокупности к другой. Поэтому использование этого показателя ограничено. Следовательно, необходимо получить еще и характеристики для совокупности, которые были бы выражены более общими цифровыми показателями. С их помощью можно сравнивать разные ряды, что затруднительно сделать с помощью лимитов.

Например, если известно, что вариационный ряд распределенных деревьев в древостое по толщине у одного насаждения имеет размах от 8 до 44 см, а в другом от 12 до 52 см, то, казалось бы, можно сделать вывод о более высоком качестве деревьев второго насаждения, т.е. что во втором насаждении они толще. 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 по значению признака, вокруг которого концентрируется большинство вариант, т.е. по величине мо-

дального класса или ступени. Значение этого признака отражает центральную тенденцию, которая типична для данного ряда. Но частоты в ряду отличаются по степени вариации, т.е. по величине отклонения от центральной тенденции ряда.

Соответственно этому статистические показатели разделяются на две группы: показатели, которые характеризуют центральную тенденцию, или уровень ряда, и показатели, измеряющие степень вариации.

К первой группе относятся различные средние величины: мода, медиана, средняя арифметическая, средняя геометрическая. Ко второй - вариационный размах (коэффициент вариации), среднее абсолютное отклонение, среднее квадратическое отклонение, или варианса, дисперсия, коэффициенты асимметрии и эксцесса. Существуют еще и другие показатели.

4.2 Средние величины и способы их вычисления

Измеренные значения различных биологических совокупностей, в т.ч. в лесном хозяйстве, представляют собой варьирующие математические величины. Чтобы получить точную и объективную характеристику варьирующей величины, прибегают наряду с построением статистических таблиц, графиков и диаграмм к так называемым статистическим показателям.

Среди них наибольшее распространение и применение находит величина среднего значения исследуемого признака. Она дает суммарную характеристику любого признака, указывая на то типичное и устойчивое в явлении, что наиболее полно выражает его содержание. Так, например, принято говорить о среднем диаметре и средней высоте насаждения, о среднем весе семян, среднем размере охотничьих животных и т.д. При этом не всегда требуется вникать в глубокий смысл, который содержит понятие средней величины.

Ряды распределения численностей, приведенные ранее, показывают, что варианты концентрируются около некоторого центрального их значения. Следовательно, можно найти такое значение варианты или абстрактное среднее число, которое будет наиболее представительной характеристикой данной статистической совокупности.

Показатели центральной тенденции характеризуются различными средними величинами: средней арифметической, средней квадратической, средней геометрической, средней гармонической, модой и медианой. Назначение средних величин состоит в том, чтобы отразить какое-нибудь одно свойство совокупности, например, среднюю высоту, средний диаметр, средний запас древесины на 1 га изучаемого участка леса.

Тот признак или то свойство совокупности, которое остается неизменным при замене индивидуальных значений их средним значением, называется определяющим свойством. Средняя должна отразить определяющее свойство так, чтобы образуемая с ее помощью

выборочная абстрактная числовая совокупность при равенстве чисел по величине определенных свойств не отличалась от общей совокупности. Из этого требования средней вытекает следующее ее общее определение. ***Средняя есть величина признака, характеризующая индивидуумы в абстрактной уравненной совокупности, замещающей реальную совокупность, но при этом сохраняющей неизменным ее определяющее свойство: общую длину, массу, объем и т.д.***

К этому определению мы будем обращаться каждый раз, когда будем обсуждать реальное содержание различных средних.

Поясним сказанное примером. Допустим, мы желаем узнать средний объем дерева в еловом древостое II класса бонитета в возрасте 50 лет. Для этого выбираем некоторую выборочную частичную совокупность (закладываем одну или несколько пробных площадей), где определяем средние значения этих выборочных совокупностей. При этом методически все необходимо сделать так, чтобы средние величины выборочной совокупности соответствовали средним величинам исследуемых ельников в генеральной совокупности. Методы корректного определения этих средних величин рассмотрены ниже.

4.3 Средняя арифметическая и способы ее вычисления

Средняя арифметическая – наиболее часто употребляемый статистический показатель. Она является центром тяжести распределения. Средняя арифметическая была бы значением величины в точке равновесия кривой численностей, если бы модель кривой была сделана в виде массивной формы.

Среднюю арифметическую генеральной совокупности обычно обозначают M, а ее выборочную оценку, т.е. среднюю арифметическую выборочных наблюдений – $\overline{X}$ (или $\overline{l}$). Она имеет то же наименование, что и варианты.

Средняя арифметическая получается от деления суммы численностей всех вариант (n_1, n_2, ..., n_n) на их число (N), т.е.

$$\overline{X} = (n_1 + n_2 + \ldots + n_n)/ N = (\Sigma n)/N, \qquad (4.1)$$

где N - сумма численностей вариант; Σ - знак суммирования.

Здесь $\overline{X}$ (и в последующем его применении) без указания пределов суммирования означает, что должно быть произведено суммирование всех измеренных (наблюденных) вариант ряда от 1 до N. Для вариант (предположим, что это длина всходов сосны, см) 3, 4, 4, 4, 5

$$\overline{X} = (3 + 4 + 4 + 4 + 5)/5 = 20/5 = 4 \text{ см}$$

Реальный смысл средней арифметической и ее главное назначение лучше уяснить, если данное выше определение всем средним применить к рассматриваемому примеру ее расчета.

Благодаря полученной средней возможно реальную совокупность высоты саженцев сосны из n_1, n_2, n_3, n_4, n_5 ($N = 5$), заменить абстрактной выровненной совокупностью из $\overline{X}$, $\overline{X}$, $\overline{X}$, $\overline{X}$, $\overline{X}$ ($N = 5$), не изменяя при этом определяющего свойства, выражаемого ΣX и также $N\,\overline{X}$, откуда получено $\overline{X} = (\Sigma n)/N$.

Для ряда, разделенного на классы, т.е. для вариационного ряда, среднюю арифметическую вычисляют как взвешенную величину:

$$\overline{X} = (n_1 x_1 + n_2 x_2 + \ldots + n_n x_n)/(n_1 + n_2 + \ldots n_n) = (\Sigma n x_i)/N, \qquad (4.2)$$

где x_1, x_2, …, x_n - классовые варианты (срединные значения классов); n_1, n_2, ..., n_n - частоты соответствующих классов; N - общее число вариант (объем ряда) или общее число наблюдений.

Группируя варианты рассмотренного примера по их величине, получим следующий ряд:

x_i : 3 4 5

n_i: 1 3 1

$\overline{X} = (1 \cdot 3 + 3 \cdot 4 + 1 \cdot 5)/5 = 4$ см

Если случайная величина выражена не в виде простого набора чисел или в виде таблицы, а задана аналитически, то применяются следующие формулы вычисления средней арифметической:

Для непрерывной средней величины $\overline{O} = \int_{-\infty}^{\infty} x_i f(x)dx$ (4.3)

Для дискретной $\overline{X} = \sum_{i=1}^{k} x_i f(x_i) = \sum_{i=1}^{k} x_i p_i$ (4.4)

В дальнейшем рассмотрим другие формулы вычисления арифметической средней, основанные на использовании ее основного свойства. Это свойство состоит в том, что сумма отклонений всех вариант от арифметической средней равна нулю. Оно вытекает из содержания средней арифметической как центра тяжести ряда. Сумма вариант, которые больше средней $\overline{X}$, равна сумме вариант, которые меньше ее.

Значение средней арифметической и ее сущность. Средняя арифметическая, как и некоторые другие средние, известна давно. Она широко используется при исследовании совокупностей в науке, технике, биологии и лесном хозяйстве.

Средняя арифметическая является обобщающей величиной, которая впитывает в себя все особенности исследуемой совокупности или ряда распределения. Она отражает уровень всей совокупности в целом, дает свободную, обобщенную характеристику 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. Например, мы измеряем толщину деревьев с округлением до 1 см: 8, 15, 16, 17…30 см. Средняя арифметическая измеренной совокупности может составить дробное число, например, 25,6. В измеренной совокупности ни одного такого замера нет. Еще более разительный пример можно привести, если проанализировать приплод (количество) щенков в помете волков. В среднем может оказаться, что там 4,2 щенка. Ясно, что число волчат дробным быть не может. В этом смысле средняя арифметическая является абстрактной величиной. Но в то же время она и конкретна.

Средняя арифметическая выражается в тех же единицах измерения, что и варианты ряда. При ее определении отклонения со знаком (+) и (-) взаимопогашаются, отметаются случайные колебания, отклонения от центральной тенденции, от уровня вариационного ряда и выступает общий закон явления. Вскрывается то типичное, что характерно 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 волков 4,2 относится только ко всей группе изученных животных. Каждая же отдельная волчица приносит целое число волчат в помете. Обычно их бывает от 2 до 6-8.

Следует помнить, что средняя арифметическая имеет смысл только по отношению к качественно однородной совокупности. Так, нельзя вычислять средний вес животных разного возраста или средний объем деревьев разных древесных видов.

При изучении лесной таксации будет показано, как отдельные ученые ошибались в определении закономерностей измерения формы стволов. Они считали, что форма стволов зависит от возраста дерева. Этот вывод был получен из-за объединения в одну группу молодняков и старых насаждений. На самом деле оказалось, что форма ствола зависит от возраста лишь до 40-50 лет, а дальше при неизменной высоте остается относительно стабильной. Ошибка была доказана выдающимся белорусским ученым-таксатором Ф.П. Моисеенко (1894-1979). Он показал, что надо изучить каждую возрастную группу отдельно и для них вычислить $\overline{X}$.

Поскольку средняя арифметическая относится к конкретной совокупности, переносить ее на явления, выходящие за рамки этой совокупности, нельзя. В отдельных случаях, если такое все же требуется, то должен быть сделан специальный анализ изучаемого явления, и лишь по его результатам следует принять решение о правомерности такого перенесения.

В дальнейшем мы увидим, что особое место в вариационной статистике занимает вопрос о том, каким образом на основе данных о той или иной частной совокупности можно делать выводы о других совокупностях подобного же рода.

Наконец, средняя арифметическая относится лишь к отдельным изучаемым признакам и не может быть автоматически перенесена на их сумму.

4.4 Другие виды средних величин

Средняя геометрическая

При изучении среднего темпа роста изучаемого признака средняя арифметическая не пригодна. Вместо нее вычисляют среднюю геометрическую M_g (или $\overline{X_g}$). Ее определяем по формуле:

$$\overline{\tilde{O}_g} = \sqrt[n]{\tilde{O}_1 \tilde{O}_2 \tilde{O}_{3...} ... \tilde{O}_n}, \qquad (4.5)$$

где X_1, X_2, ..., X_n - темпы роста (величины, показывающие, во сколько раз увеличивался признак от периода к периоду); n - число периодов.

При $n>2$ формулу удобнее применять в логарифмическом виде:

$$\lg \overline{\tilde{O}_g} = \frac{1}{n}(\lg \tilde{o}_1 + \lg \tilde{o}_2 + ... + \lg \tilde{o}_n). \qquad (4.6)$$

Если данные, для которых вычисляют среднюю геометрическую, представлены разными численностями (n_i) в пределах выделенных классов (x_i), то применяется формула:

$$\lg \overline{\tilde{O}_g} = (n_1 \lg \tilde{o}_1 + n_2 \lg \tilde{o}_2 + ... + n_n \lg \tilde{o}_n) / N \qquad (4.7)$$

Исходя из содержания формул (4.5) и (4.6), среднюю геометрическую называют также средней логарифмической, так как ее логарифм есть арифметическая средняя логарифмов составляющих величин.

Применение средней геометрической поясним следующим примером. Пусть на лесокультурную площадь в сосняке мшистом высажен 1-летний саженец сосны. Измерим его объем в $см^3$ в момент посадки (1 год), а также в 5, 10, 15 и 20 лет. Пусть эти объемы составят соответственно 8, 560, 2800 и 11200 и 22400$см^3$. Тогда отношения объемов сеянцев через соседние равные промежутки времени (5 лет) будут следующие

$$x_1 = \frac{560}{8} = 70;\ x_2 = \frac{2800}{560} = 5;\ x_3 = \frac{11200}{2800} = 4;\ x_4 = \frac{22400}{11200} = 2.$$

Число рассмотренных периодов у нас равно 4, т.е. $N = 4$.

По формуле (3.1) вычислим среднюю геометрическую:

$$\overline{\tilde{O}_{ä}} = \sqrt[4]{70 \cdot 5 \cdot 4 \cdot 2} = \sqrt[4]{2800} \approx 7{,}274.$$

Это означает, что объем нашего сеянца сосны от 1 до 20 лет увеличивается в среднем за каждый период в 7,274 раза. Действительно, используя среднюю геометрическую, получаем в 20 лет объем стволика $X_5 = 8 \cdot 2800 = 22400$ см^3.

Если бы мы вычислили здесь среднюю арифметическую из наших 5 сосенок, то она составила бы 7594 см ^{3} и характеризовала бы объем стволика в возрасте примерно 15 лет, что не отвечает сути изучаемого процесса, т.к. дает объем ствола к концу 4 периода, равный 7594 см^3 $\cdot$ 5 = 30376 см^3, что на 36% больше фактических данных.

Для средней геометрической характерно равенство произведений из первоначальных данных измерений (X_1, X_2, ..., X_n) и из геометрических средних $\overline{X_g}, \overline{X_g}, ..., \overline{X_g}$, представленных n раз.

Вспомним, что для средней арифметической величины характерно постоянство суммы вариант.

Средняя квадратическая

Основная цель измерения диаметров деревьев в древостое – это определение запаса древесины. Для вычисления запаса надо знать сумму площадей поперечных сечений измеренных деревьев (g_i), а затем воспользоваться соответствующей формулой, куда $\sum g_i$ входит как сомножитель. Эта формула описана в курсе лесной таксации, изучаемой на 3 курсе. Сумму площадей сечений деревьев в древостое можно определить, умножив число деревьев на площадь сечения среднего дерева. Известно, что площадь сечения дерева на высоте 1,3 м приравнивается к площади круга, и ее находят по формуле $g_i = \pi d_i^2/4$, где π=3,14159, d_i – диаметр на высоте 1,3 метра. Нам необходимо выяснить, какую величину должен представлять средний диаметр, чтобы он был репрезентативным показателем для достижения указанной цели.

Анализ, сделанный путем использования здесь $\overline{X}$, показывает, что применение средней арифметической величины в этом случае неприемлемо, а необходимо вычислить среднюю квадратическую. Ее находят по формуле:

$$\bar{x}_{кв} = \sqrt{\frac{1}{n}\sum x_i^2 n_i} \,. \qquad (4.8)$$

В качестве иллюстрации сказанного рассмотрим определение объема среднего дерева в древостое сосны возрастом 100 лет (II класса бонитета), который измерен нами ранее и приведен в таблицах 3.1, 3.3. Результаты расчетов показаны в таблице 4.1. В этой же таблице представим исходные данные и для нахождение средней арифметической: $\frac{\sum n_i x_i}{\sum n_i}$.

Средний арифметический диаметр будет равен:

$$D_a = \frac{\sum n_i x_i}{\sum n_i} = \frac{3440}{120} = 28{,}7 \text{ (см)}$$

Таблица 4.1 – Определение среднего диаметра для совокупности 120 деревьев сосны

Ступени толщины (классы), X_i, см	Численности (число деревьев), n_i, шт.	Площадь сечения ступеней толщины, $g_i = \frac{x_i^2 \cdot \pi}{4}$, см^2	$g_i n_i$, см^2	$x_i n_i$, см
8	1	58,27	58,3	8
12	3	113,10	339,3	36
16	8	201,86	1614,9	128
20	14	314,16	4398,2	280
24	20	452,39	9047,8	480
28	27	615,75	16625,2	756
32	17	804,25	13672,3	544
36	12	1017,88	12214,6	432
40	9	1256,64	11309,8	360
44	5	1520,53	7602,6	220
48	3	1809,56	5428,7	144
52	1	2704,78	2704,8	52
Итого (∑)	120	-	85016,3	3440

Средний квадратический диаметр, который определим через среднюю площадь сечения, по правилам, принятым в лесной таксации. Формула для его нахождения соответствует средней квадратической (формула (4.8)):

$$g_{cp} = \frac{D_{ñð}^2 \cdot \pi}{4}; \qquad D_{cp} = 2\sqrt{\frac{g_{ñð}}{\pi}} \tag{4.9}$$

Средняя площадь сечения равна:

$$g_{cp} = \frac{\sum g_i n_i}{\sum n_i} = \frac{85016{,}3}{120} = 708{,}5 \text{ см}^2$$

Тогда

$$D_{cp} = 2\sqrt{\frac{708{,}5}{3{,}1416}} = 2\sqrt{225{,}5} = 2 \cdot 15 = 30 \text{ см}$$

Как видим, средний диаметр, который пригоден для нахождения запаса через среднюю площадь сечения и определенный через среднюю геометрическую, больше, чем среднеарифметический. Допустим, что среднее дерево в этом древостое имеет высоту (h) 26 м, а коэффициент формы (f), т.е. отноше-

ние объема ствола к объему цилиндра, имеющего с данным стволом одинаковые высоту и диаметр на высоте 1,3 м – 0,46.

Объем дерева (V) в этом случае определяется по формуле $V = ghf$. При этом все величины следует представить в одной размерности. Так как объем дерева обычно находят в м^3, то и все составляющие приведенной формулы выразим в м. Тогда объем дерева, имеющего среднеквадратический диаметр (V_1) будет равен

$$V_1 = \frac{3{,}1416 \cdot 0{,}3^2 \text{ì}}{4} \cdot 26\text{ì} \cdot 0{,}46\text{ì} = 0{,}84 \text{ м}^3.$$

При использовании среднеарифметического диаметра объем дерева (V_2) составит

$$V_2 = \frac{3{,}1416 \cdot 0{,}287^2 \text{ì}}{4} \cdot 26\text{ì} \cdot 0{,}46\text{ì} = 0{,}77 \text{ м}^3.$$

Как видим, в последнем случае данный показатель занижен почти на 9%. Истинный запас исследованного древостоя (M) равен произведению числа стволов ($\sum n_i = N$) на объем среднего дерева, т.е. $M=V_{ср} \cdot N$. Для нашего примера M = 0,84 м^3 · 120 = 101 м^3. Если же для расчетов брать среднеарифметический диаметр, то получим M = 0,77 м^3 · 120 = 92 м^3, т.е. недобор составил 9м^3. При средней биржевой цене древесины сосны в докризисные времена на мировом рынке 200 тыс. руб. / м^3, недобор составит 1800 тыс. рублей. Это для сравнительно небольшого участка леса, т.е. для пробной площади в 0,3-0,4 га. На 1 га недобор будет уже от 4,5 до 6 млн. рублей.

Таким образом, для получения истинного значения площади сечения или объемов всех деревьев посредством среднего дерева и числа деревьев диаметр среднего модельного дерева следует находить как среднюю квадратическую величину. В лесной таксации именно так и поступают. Поэтому средняя квадратическая величина нашла широкое распространение в лесном хозяйстве.

Средняя гармоническая

Эта величина применяется для вычисления средней характеристики признаков, которые представляют собой отношение двух других варьирующих величин. Среднюю гармоническую определяют по формуле:

$$\overline{X_h} = N/(\sum 1/X), \qquad (4.10)$$

или

$$\overline{X_h} = N/(\sum n/x), \qquad (4.11)$$

где n - веса отдельных значений.

В практике лесного хозяйства эта величина применяется очень редко.

Мода и медиана

Модой (M_o) называют наиболее часто встречающуюся варианту. В нормально распределенных совокупностях мода численно равна средней арифметической.

В положительно асимметричных рядах мода больше средней арифметической ($Ì_0 \rangle \bar{õ}$), а в отрицательно ассиметричных рядах она меньше средней арифметической ($Ì_î \langle \bar{õ}$).

Найдем моду в нашем примере, приведенном в таблице 4.1.

X	8	12	16	20	24	28	32	36	40	44	48	52	Итого
n_i	1	3	8	14	20	27	17	12	9	5	3	1	120

Мода в приведенном ряду составляет наиболее представленный класс 28см. Средняя арифметическая здесь $\bar{x} = 28{,}7$ см, т.е. $Ì_î \langle \bar{õ}$. Следовательно, ряд имеет небольшую положительную асимметрию.

Общая формула для вычисления моды (M_o) следующая:

$$M_o = x_{Mo} + \frac{C(n_{Mo} - n_{Mo-1})}{2n_{Mo} - n_{Mo-1} - n_{Mo+1}}, \qquad (4.12)$$

где x_{Mo} – середина модального (наиболее представленного) класса;

C – величина класса;

n_{Mo}, n_{Mo-1}, n_{Mo+1} – частоты соответственно модального класса; предшествующего модальному классу; класса, следующего за модальным.

Эта формула, учитывающая величину классового промежутка, более точна, т.к. наибольшее количество деревьев может быть сдвинуто вправо или влево от середины класса.

Тогда мода для нашего ряда равна:

$$M_o = 28 + \frac{4(27-20)}{2 \cdot 27 - 20 - 17} = 28 + \frac{4 \cdot 7}{54 - 37} = 28 + \frac{28}{17} = 28 + 1{,}7 = 29{,}7 \approx 30\,\text{см}.$$

Медианой (M_e) называют значение признака, занимающее срединное положение в ряду и делящее все распределение на две равные по численности части.

Среди значений: 5 6 7 8 9 $M_e = 7$.

Для вариационного ряда

$$M_e = X_o + k\,[S_1 - S_2)/n], \qquad (4.13)$$

где X_o - значение нижней границы класса, в котором содержится половина накопленных частот;

k - интервал;

$S_1=N/2$;

S_2 - накопленная частота, предшествующая группе, в которой находится медиана.

Для нашего ряда диаметров сосны (таблица 4.1) медиану вычислим следующим образом

X_i	8	12	16	20	24	28	32	36	40	44	48	52
n_i	1	3	8	14	20	27	17	12	9	5	3	1
$\sum n$	1	4	12	26	46	73	90	102	111	116	119	120

$$M_e = 26 + 4[(73 - 46)/27] = 26 + 4 \cdot 1 = 30 \text{ см}$$

В данном примере мода и медиана совпали. Это обычное явление для симметричных распределений. В сильной степени асимметричные ряды имеют несовпадающие значения моды и медианы.

Мода и медиана являются характеристиками центральной тенденции выборки. Они не имеют своего аналога в генеральной совокупности и поэтому рассматриваются как показатели относительного характера.

Верхняя и нижняя средние

В практике иногда встречаются другие виды средних. Применительно к лесохозяйственным объектам, иногда возникает необходимость применения «верхней» или «нижней» средней. Их применяют тогда, когда требуется оценить часть явления. В лесной таксации «верхнюю» среднюю вычисляют при определении верхней высоты, т.е. когда надо найти среднюю высоту самых толстых деревьев. Их может быть 50, 100, 200 или 5, 10, 20 % от всей совокупности и т.д. «Нижнюю» среднюю обычно вычисляют для определения запаса, вырубаемого при проведении уходов низовым способом.

Обобщение описанных средних величин сделано в таблице 4.2, где приведены формулы для их вычислений при малом и большом числе наблюдений. Пояснение символов было дано выше.

Как видно из таблицы 4.2, средняя геометрическая (средняя логарифмическая) равна корню n-й степени из произведения значений случайной величины, средняя гармоническая есть обратная величина среднего арифметического величин, обратных X_i; средняя квадратическая получается как корень квадратный из среднего значения квадрата X_i. Как показали советские ученые-таксаторы К.Е. Никитин (1908-1986) и А.З.Швиденко, все названные виды средних могут быть получены из формулы

$$\overline{X} = \frac{1}{n}(x_1^a + x_2^a + \ldots + x_n^a)^{1/a}. \qquad (4.14)$$

Здесь при a=1 имеем $\overline{X}$ - среднее арифметическое, a=2 – квадратическое $(\overline{X_g})$; a=–1 - гармоническое (H), при a=0 – геометрическое (G), причем по своей величине $H \le G \le \overline{X} \le \overline{X_g}$.

Таблица 4.2 – Основные виды средних, применяемых в лесном деле (по К.Е. Никитину и А.З. Швиденко)

Средняя	Формулы для вычисления средних	
	Малая выборка (невзвешенные средние)	Большая выборка (взвешенные средние)
Арифметическая	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$
Геометрическая	$G = \sqrt[n]{x_1 x_2 \ldots x_n}$ $\lg G = \frac{1}{n}\sum \lg x_i$	$G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}}$ $\lg G = \frac{1}{n}\sum n_i \lg x_i$
Гармоническая	$H = \frac{n}{\sum \frac{1}{x_i}}$	$H = \frac{n}{\sum n_i \frac{1}{x_i}}$
Квадратическая	$\overline{X_g} \sqrt{\frac{1}{n}\sum x_i^2}$	$\overline{X_g} \sqrt{\frac{1}{n}\sum x_i^2 n_i}$
“Верхняя”, $x_{i+1} \ge x_i$	$\overline{X_\rho} = \frac{1}{\rho}\sum_{i=n-\rho}^{n} x_i$	$\overline{X_\rho} = \frac{1}{\rho} \cdot \left(\sum_{i=[\rho]+1}^{k} x_i n_i - A \right)$*
“Нижняя”, $x_{i+1} \ge x_i$	$\overline{X_f} = \frac{1}{f}\sum_{i=1}^{f} x_i$	$\overline{X_f} = \frac{1}{f} \cdot \left(\sum_{i=1}^{[f]-1} x_i n_i + B \right)$**

* $A = \frac{1}{2}\left[2x_{[\rho]+1} - c - \frac{c}{n_{[\rho]}} \left(\rho - \sum_{i=[\rho]+1}^{k} n_i \right) \right] \left(\rho - \sum_{i=[\rho]+1}^{k} n_i \right)$

** $B = \frac{1}{2}\left[2x_{[f]-1} + c - \frac{c}{n_{[f]}} \left(f - \sum_{i=1}^{[f]-1} n_i \right) \right] \left(f - \sum_{i=1}^{[f]-1} n_i \right)$

Наконец, «верхние» и «нижние» средние представляют собой средние арифметические из ρ наибольших ($x_{n-\rho}$, $x_{n-\rho+1}$, …, x_n) или f наименьших (x_1, x_2, …, x_f) значений случайной величины. в частном случае, при ρ и f, равным 1, получаем просто наибольшее и наименьшее значения выборки. В формулах для определения $\overline{X_\rho}$ и $\overline{X_f}$ предполагается, что исходные данные записаны в порядке возрастания значений x_i. Для ряда распределения, где условие $x_{i+1} > x_i$ соблюдается автоматически, индексы $[\rho]$ и $[f]$ обозначают номера классов ряда распределения, в которые попадают соответственно $n - \rho$-е и f-е наблюдения.

Выбор того или иного вида среднего определяется целью работы и характером изучаемого явления. Например, если признак x_i представляет

численную характеристику некоторого элемента совокупности (объем или объемный прирост дерева, выработка бригады и др.), а $\sum x_i$ - численную характеристику совокупности в целом, то наиболее предпочтительна средняя арифметическая величина $\overline{X}$. С другой стороны, $\overline{X}$ в определенном смысле аннулирует положительные и отрицательные отклонения отдельных значений x_i от $\overline{X}$. Поэтому, если средняя величина вычислена для наблюдений, несущих в себе случайные ошибки, и эти ошибки подчиняются нормальному закону распределения (т.е. можно предполагать, что положительные и отрицательные отклонения уравновешиваются), то опять-таки наилучшая – средняя $\overline{X}$. Если влияние ошибок несимметрично, то часто более пригодна средняя геометрическая. Этот же тип среднего применяют в тех случаях, когда нужно определить скорость изменения (например, темпы роста) изучаемого показателя и пр.

Для примера возьмем данные, приведенные А.З. Швиденко, по анализу прироста по запасу в некотором древостое. Пусть M_O, M_1, ..., M_n – запасы древостоев в возрасте A_O, A_1, ..., A_n. Требуется определить относительное среднее изменение запаса древостоя в периоде от A_O до A_n лет.

Образуем величины $x_1 = M_1/M_O$, $x_2 = M_2/M_1$, …, $x_n = M_n/M_{n-1}$, представляющие относительное изменение запаса в каждом из отдельных интервалов времени $[A_1, A_O]$, …, $[A_n, A_{n-1}]$. Произведение $x_1, x_2, \ldots, x_n$ дает нам общее изменение запаса в периоде $[A_O, A_n]$. Если в этом периоде относительное изменение запаса по отдельным периодам постоянно, то должно выполняться равенство $G^n=x_1x_2\ldots x_n$, откуда следует необходимость использования средней геометрической величины, т.е.

$$\Delta M = \sqrt{\frac{M_1}{M_O}\cdot\frac{M_2}{M_1}\ldots\frac{M_n}{M_{n-1}}} - 1 = \sqrt{\frac{M_n}{M_O}} - 1.$$

Иногда целесообразно среднюю геометрическую вычислять для «загрязненных» выборок, т.е. для ситуаций, когда одно или несколько выборочных значений сильно отличаются от основной массы.

Среднюю квадратическую (как и другие виды степенных средних) можно применять в тех случаях, когда в некоторой многомерной совокупности средний элемент устанавливают по одному (обычно наиболее важному) признаку, причем для этого элемента нужно указать средние по другим признакам. Например, в однородных древостоях дереву среднего (арифметического) объема соответствует средний квадратический диаметр. Это явление объясняется просто: изменение объема пропорционального изменению не диаметра, а его квадрата.

«Верхние» («нижние») средние применяют в тех случаях, когда нужно оценить интенсивность части явления или процесса. В лесоучетных задачах обычное применение «верхней» высоты – это определение высоты n самых толстых деревьев на 1 га; могут представлять практи-

ческий интерес размеры определенной части наиболее мелких сеянцев в питомнике (в связи с достижением стандартных размеров) и т.д.

Как статистику положения, медиану обычно применяют в тех случаях, когда требуется обеспечить минимум абсолютной величины отклонений между исходными данными и используемым показателем. Например, логично потребовать соблюдение такого условия для показателя, отражающего среднюю продолжительность жизни деревьев в разновозрастном лесу. Мода представляет интерес, когда устанавливают некоторое типическое свойство явления или объекта: цвет желудей, урожайность семян на плюсовых деревьях и т.д.

Квантили являются обобщением понятия медианы. Если разделить частоты распределения на k равных частей, то $k-1$ значение случайной величины, соответствующее точкам деления, называется квантилями распределения. При k=2 единственный квантиль будет медианой, при k=4 средняя из точек деления будет медианой, а первая и третья - нижним и верхним квартилями. Аналогично девять значений случайной величины, делящих частоты распределения на 10 частей, называют децилями. Знание трех-четырех квантилей дает хорошее представление о распределении случайной величины. Квантилем x_p или $x_{1-\alpha}$, соответствующим заданной вероятности p, $p=1-\alpha$, называют такое значение случайной величины, для которого $p(x<x_p)= p=1-\alpha$.

Обобщая изложенное, отметим, что в лесном деле наибольшее применение находят средняя арифметическая и средняя квадратическая. Остальные средние применяют редко.

5. ПОКАЗАТЕЛИ ВАРИАЦИИ

5.1 Вариация как явление и ее источники
5.2 Типы варьирования
5.3 Характеристики вариационных рядов и их вычисление через моменты
5.4 Асимметрия, эксцесс, коэффициент вариации

5.1 Вариация как явление и ее источники

Ряд распределения характеризуется несколькими параметрами. Одним из них являются рассмотренные выше средние значения. Они характеризуют то значение вариационного ряда, относительно которого в определенном смысле располагаются все другие численные значения, присущие данному ряду распределения случайной величины.

В то же время существуют другие важные характеристики рядов случайных величин. В их числе одной из основных является размах (диапазон) случайных величин от минимальной до максимальной. Эта величина характеризует вариабельность, или изменчивость, ряда распределения. Приведем пример (таблица 5.1).

Таблица 5.1 – Ряды распределения числа деревьев ели по диаметру на высоте 1,3 м и их средние значения

№ классов	Средние значения классов (ступени толщины), см (x_i)	Число деревьев (численность) и их средние арифметические значения по вариантам опыта					
		Ряд 1		Ряд 2		Ряд 3	
		n_i	$x_i n_i$	n_i	$x_i n_i$	n_i	$x_i n_i$
1	8	2	16	-	-	-	-
2	12	3	36	-	-	-	-
3	16	13	208	10	160	-	-
4	20	13	260	20	400	7	140
5	24	23	552	21	504	25	600
6	28	32	896	25	700	33	924
7	32	40	1280	35	1128	65	2080
8	36	24	864	38	1368	42	1512
9	40	19	760	30	1200	22	880
10	44	12	528	16	704	6	264
11	48	7	336	4	192	-	-
12	52	5	260	1	52	-	-
13	56	4	224	-	-	-	-
14	60	3	180	-	-	-	-
Итого, $\sum n_i$	-	200	6400	200	6400	200	6400
Среднее значение, $\bar{x}$		-	32,0	-	32,0	-	32,0

В таблице 5.1 приведены 3 ряда распределения числа стволов ели по диаметру. Их количество одинаково – по 200 деревьев. Средние арифметические всех рядов тоже равны – по 32 см. Но ряды существенно отличаются. Средние показатели это выявить не могут. Так, среднеквадратическое значение этих рядов равно соответственно 37,5 см; 37,3 см; 36,6 см, т.е. практически одинаковы, т.к. разница в 0,4 см в древостоях с диаметрами такого размера при проведении замеров с помощью мерной вилки в производственных условиях лежит в пределах точности измерений.

В то же время даже наглядно видно, что ряды существенно отличаются между собой. Первый ряд имеет наибольший размах распределения, а третий – наименьший. В первом ряду крайние варианты отклоняются от средней на 6 классов в нижнюю сторону и на 8 в верхнюю, а всего ряд имеет 15 классов, или ступеней, толщины. В третьем случае соответствующие величины равны 3 и 3, а размах ряда соответствует 7 классам. Второй ряд распределения занимает промежуточное положение между 1 и 3 рядами.

Чтобы дать более полную характеристику рядов распределения случайной величины, введена характеристика их вариации, или изменчивости, которая характеризует степень рассеяния случайной величины относительно среднего значения. Для выражения вариации используют специальные величины: коэффициент вариации, среднее квадратическое отклонение, дисперсию и другие, которые рассматриваются ниже.

Возникает естественный вопрос – каковы же причины вариации. Основная причина состоит в том, что по своей природе любые биологические объекты, даже принадлежащие к однородной совокупности, отличаются друг от друга. Деревья одного вида, одного возраста, растущие в одинаковых условиях роста имеют разную высоту и толщину, что вызвано как генетическими свойствами каждой особи, так и некоторыми особенностями их территориального размещения: в тени более крупных деревьев, на микроповышении или микропонижении и т.п. Подобные примеры можно привести для любых биологических объектов: люди даже одной расы и национальности отличаются по росту, животные одного вида имеют разный вес и т.д. Основная причина этому, как уже было отмечено, - наличие биологического (в первую очередь генетического) разнообразия.

Изменчивость происходит и из-за ошибок измерений. Добиться абсолютной точности измерений очень трудно (и дорого), а часто и не нужно. Практику устраивает некоторый уровень точности, его-то и выдерживают. Например, при перечетах деревьев их измеряют по ступеням толщины. Точность измерений может повышаться, если это потребуется в отдельных случаях. Так, в условиях рынка повышаются требования к точности оценки объема древесины, особенно для ее наиболее ценной части. При проведении измерений помимо допустимых придержек в измерениях возможны и ошибки, особенно при массовых замерах. Хотя ошибки измерений и являются одной из причин варьирования, но эта причина не столь значима как естественная биологическая изменчивость.

5.2 Типы варьирования

Получаемые в результате наблюдений значения наблюдаемого признака называют **в а р и а н т а м и**. Варианты в биологических объектах обнаруживают разнообразие (или варьирование) изучаемого свойства. Например, деревья отличаются друг от друга по диаметру, высоте, объему, санитарному состоянию. Причины этого показаны выше.

В зависимости от характера изучаемого признака различают варьирование непрерывное, прерывистое (дискретное) и атрибутивное. Непрерывное и дискретное варьирование присуще количественным признакам, а атрибутивное – качественным.

При **н е п р е р ы в н о м** варьировании отдельные значения признака выражают мерой протяженности, объема и т.д. Отдельные варранты могут иметь любое, но изменяющееся в определенных пределах значения меры. Толщина деревьев в древостое, например, от самого тонкого до самого толстого может принимать самые различные значения меры протяженности. Только в зависимости от цели исследования (измерения) выражают ее в классах толщины: в несколько сантиметров, в целых сантиметрах, в десятых или сотых долях сантиметра.

При **д и с к р е т н о м** варьировании отдельные значения признака выражают отвлеченными числами, чаще всего целыми. Например, число всходов сосны на учетной площадке обладает дискретным варьированием, т.к. они, равно как и число семян в навеске, выражаются целыми числами.

При **а т р и б у т и в н о м** варьировании значения признака выражают в качественных показателях. Это может быть степень окраски, консистенции, поврежденность или устойчивость, а также форма, вид и т.д. Количественно эти признаки выражают в абсолютных числах, долях единицы, процентах, баллах и т.д. Например, различают цвет коры на деревьях, форму кроны деревьев (шаровидная, пирамидальная и т.д.), густота раствора, степень повреждения деревьев вредителями: сильная, слабая и др.

Частным случаем атрибутивного варьирования является альтернативное, при котором значения признака рассматривают в альтернативной форме, т.е. противопоставляя здоровые больным, сильные – слабым, окрашенные – неокрашенным, присутствующие – отсутствующим и т.д. В альтернативной форме можно представить и количественные признаки, противопоставляя, например, высокие деревья в древостое низким, господствующие деревья – угнетенным, здоровые – сухим или усыхающим.

5.3 Характеристики вариационных рядов и их вычисление. Пределы и размах вариации

Основными показателями вариационного ряда кроме среднего значения являются абсолютное и относительное значение его пределов, выражаемое величиной дисперсии и коэффициента вариации.

Одним из показателей амплитуды вариации служат так называемые лимиты (от лат. Limes - предел, граница), т.е. значения минимальной и максимальной вариант выборочной совокупности. Этот показатель (Lim) указывает фактические границы вариабельности признака. Поэтому его обычно приводят наряду с другими биометрическими показателями в сводных статистических таблицах. Значение лимитов заключается в их конкретности.

Величина вариации может быть оценена и по разности между максимальной и минимальной вариантами совокупности. Этот показатель получил название размаха вариации. Например, пределы первого распределения (n_1), которое только что рассматривалось (таблица 5.1, вариант 1), равны: min=8 и max=60 единицам, откуда размах этого ряда равен 60–8=52. Второй ряд (n_2) имеет пределы вариации от 16 до 52 единиц, его размах равен 52–16=36 единиц. Наиболее низким этот показатель оказывается в третьем ряду (n_3), размах которого равен 24 единицам: 44–20=24.

Рассмотренные показатели вариации вполне объективны и просты. Но в силу присущих им недостатков они мало пригодны для измерения вариабельности признаков. Дело в том, что эти показатели неустойчивы: они зависят от многих случайных причин и при повторных выборках могут резко менять свое значение. Главный же недостаток указанных показателей заключается в том, что они не отражают существенные черты варьирования.

Из сказанного следует, что лимиты и пределы вариации, хотя и дают определенное, конкретное представление о величине изменчивости признаков, не могут служить основным мерилом вариабельности биологических величин. Поэтому здесь применяют другие показатели, описываемые ниже.

Среднее линейное отклонение

Для измерения вариации можно использовать центральный момент первого порядка как одну из характеристик вариационного ряда, представляющую сумму отклонений вариант от средней арифметической, отнесенную к общему числу вариант данной совокупности.

$$\Delta = \frac{\sum \left| (x - \bar{X}) \right|}{n} \qquad (5.1)$$

Этот показатель, называемый средним линейным отклонением, может иметь значение только при условии, что отклонения вариант от средней арифметической берутся без учета знаков, так как в противном случае $\sum (x - \bar{X}) = 0$.

Используем этот показатель для характеристики взятого нами примера.

В таблице 5.2 показано вычисление Δ для третьего ряда распределения (n_3), имеющего наименьший размах. Полученное значение (Δ=1,36) необходимо сравнить с другим подобным показателем, иначе его трудно оценить. Вычислим Δ для первого ряда (n_1), имеющего наибольший размах вариации. Ход вычисления показан в таблице 5.3. Вспомним, что среднее значение для наших рядов равно 32 см.

Таблица 5.2 – Вычисление линейного отклонения для ряда распределения числа стволов по диаметру в сосновом древостое – третий вариант

Ступени толщины (варианты) (x_i)	Число деревьев (частоты), (n_i)	$x_i n_i$	$\lvert x_i \rvert - \bar{x}$	$n_i(x_i - \bar{x})$	Вычисления
20	7	140	12	84	$\Delta = \frac{\sum n_i \lvert (x_i - \bar{x}) \rvert}{\sum n_i} = \frac{822}{200} = 4{,}11$
24	25	600	8	200	
28	33	924	4	132	
32	65	2080	0	0	
36	42	1512	4	168	
40	22	980	8	176	
44	6	264	12	72	
Сумма	200	6400	-	822	

Таблица 5.3 – Вычисление линейного отклонения для ряда распределения числа стволов по диаметру в сосновом древостое – первый вариант

Ступени толщины (варианты) (x_i)	Число деревьев (частоты), (n_i)	$x_i n_i$	$\lvert x_i \rvert - \bar{x}$	$n_i(x_i - \bar{x})$	Вычисления
8	2	16	24	48	$\Delta = \frac{\sum n_i \lvert (x_i - \bar{x}) \rvert}{\sum n_i} = \frac{1252}{200} = 6{,}26$
12	3	36	20	60	
16	13	208	16	208	
20	13	260	12	156	
24	23	552	8	184	
28	32	896	4	128	
32	40	1280	0	0	
36	24	864	4	96	
40	19	760	8	152	
44	12	528	12	144	
48	7	336	16	112	
52	5	260	20	100	
56	4	224	24	96	
60	3	180	28	84	
Сумма	200	6400	-	1252	

Сравнивая первый результат (4,11) со вторым (6,26), видим, что ряд, у которого больше амплитуда изменчивости, имеет и больший показатель вариации, выражающийся величиной линейного отклонения. При меньшем размахе вариационного ряда и показатель вариации оказывается меньше. В то же время описанный показатель имеет существенные недостатки. Так, для ряда №3 от средней арифметической отклоняются три класса, а в первом ряду уже восемь. Следовательно, амплитуда изменчивости первого ряда в 2,66 раза больше, чем третьего. Если среднее отклонение для третьего ряда равно 4,11, то для первой совокупности оно должно быть в 2,66 раза больше, т.е. $4{,}11 \cdot 2{,}66 = 10{,}9$. На самом же деле этот показатель равен 6,26. Разница составляет $10{,}9 - 6{,}26 = 4{,}64$, или 74,1 %, т.е. она довольно велика. Следовательно, среднее линейное отклонение не может быть достаточно точным показателем вариации, не говоря уже о том, что этот показатель теряет всякий смысл, если брать не модули отклонений, а учитывать знаки отклонений вариант от средней арифметической, так как сумма отклонений будет близка к нулю.

Среднее квадратическое отклонение

Чтобы преодолеть недостатки линейного отклонения, принято отклонения вариант от средней арифметической возводить в квадрат и сумму квадратов отклонений относить к общему числу наблюдений, т.е. к объему выборки. Полученный таким образом показатель служит центральным моментом второго порядка, он характеризует дисперсию или рассеяние вариант около средней арифметической. Этот показатель, называемый дисперсией, или вариансой, обозначается греческой буквой σ^2 (сигма малая в квадрате)[1] и выражается следующей формулой:

$$\sigma^2 = \frac{\sum(\tilde{o}_i - \bar{x})^2}{\sum n_i} = \frac{1}{\sum n_i} \cdot \sum(\tilde{o}_i - \bar{\delta})^2 \qquad (5.2)$$

Для непрерывной случайной величины, заданной аналитически, дисперсию находят по формуле:

$$\sigma^2 = \int_{-\infty}^{\infty} (x_i - \bar{x})^2 \, fx(dx) \qquad (5.3)$$

[1] В литературе Западной Европы и в ряде отечественных публикаций дисперсию принято обозначать латинской буквой s^2, а буквой σ обозначают среднее квадратическое отклонение теоретических распределений. В отдельных случаях в дальнейшем, когда мы заимствовали некоторые формулы у других авторов, сохранена и их символика, т.е. использована буква $S^2=\sigma^2$.

При возведении отклонений вариант от средней арифметической в квадрат их сумма не превращается в нуль, так как и положительные и отрицательные отклонения получают один и то же положительный знак. Кроме того, большие отклонения от средней, будучи возведены в квадрат, получают и больший «удельный вес», оказывая большее влияние на величину показателя вариации.

Однако, возводя отклонения вариант от средней арифметической в квадрат, мы, таким образом, искусственно увеличиваем и самый показатель вариации. Чтобы преодолеть это, вместо дисперсии берут корень квадратный из указанного отношения:

$$\sigma = \pm\sqrt{\frac{\sum (x_i - \bar{x})^2}{\sum n_i}} = \pm\sqrt{\frac{1}{x_n}\sum (x_i - \bar{x})^2} \qquad (5.4)$$

Полученный таким образом показатель называется средним квадратическим отклонением. Иногда его называют основным отклонением, или просто (для краткости) сигмой. Знаки «плюс» и «минус» (+, -), поставленные перед радикалом, указывают на то, что данный показатель в равной мере характеризует отклонения вариант от средней арифметической как в сторону больших (+), так и в сторону меньших (-) значений. В дальнейшем в целях экономии эти знаки опущены, т.к. подразумевается, что они стоят впереди этого показателя.

Среднее квадратическое отклонение, как и средняя арифметическая, относится к величинам именованным и выражается в тех же величинах, что и признак. Выборка, в которой рассеяние вариант около средней арифметической больше, характеризуется и большей величиной среднего квадратического отклонения и, наоборот, при меньшей вариабельности признака среднее квадратическое отклонение оказывается меньшим.

В лесном хозяйстве чаще всего используют σ вместо σ^2. Преимущество среднего квадратического отклонения против дисперсии объясняется практическим удобством: в случае использования σ мы имеем меру рассеяния, выраженную в тех же величинах, что и среднее значение.

По сравнению со средним линейным отклонением среднее квадратическое отклонение более точно характеризует вариабельность признаков. Для подтверждения приведенного утверждения вычислим величину σ для 1 и 3 ряда распределения числа стволов по диаметру, которые показаны в таблице 5.1, и сравним ее с линейным отклонением. Техника этого вычисления показана в таблице 5.4.

На основе данных таблицы 5.4 выполним вычисления по формуле (5.2)

$$\sigma_1 = \sqrt{\frac{1}{200}\cdot 21878} = \sqrt{109{,}39} = 10{,}5\,(\text{см}); \quad \sigma_2 = \sqrt{\frac{1}{200}\cdot 6240} = \sqrt{31{,}2} = 5{,}6\,(\text{см}).$$

Таблица 5.4 – Вычисление среднеквадратического отклонения для ряда распределения числа стволов по диаметру. Среднее значение ($\bar{O}$) равно 32,0 см

Ступени (классы) толщины, x_i	Вычисления по вариантам							
	Ряд №1				Ряд №3			
	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$
8	2	-24	576	1152	-	-	-	-
12	3	-20	400	1200	-	-	-	-
16	13	-16	256	3328	-	-	-	-
20	13	-12	144	1872	7	-12	144	1008
24	23	-8	64	1472	25	-8	64	1600
28	32	-4	16	512	33	-4	16	528
32	40	0	0	0	65	0	0	0
36	24	4	16	384	42	4	16	832
40	19	8	64	1216	22	8	64	1408
44	12	12	144	1728	6	12	144	864
48	7	16	256	1782	-	-	-	-
52	5	20	400	2000	-	-	-	-
56	4	24	576	2880	-	-	-	-
60	3	28	784	2352	-	-	-	-
Итого	200	-	-	21878	200	-	-	6240

Сравнивая между собой σ_1 и σ_2, видим, что они более адекватно отражают вариацию рядов, чем Δ_1 и Δ_2.

Из теоретической статистики известно, что вариация генеральной совокупности больше вариации выборки, взятой из данной генеральной совокупности, в среднем в $\frac{\sum i_i}{\sum n_{i-1}}$ раз. Для упрощения символов обозначим $\sum n_i = N$. На этом основании в формулу (5.4) следует внести поправку, взяв в качестве множителя подкоренного выражения величину $\frac{n}{n-1}$. В результате формула (5.4) преобразуется следующим образом:

$$\sigma = \pm\sqrt{\frac{1}{N-1} \cdot \sum (x_i - \bar{x})^2 \cdot n_i} \qquad (5.5)$$

Величина $(N-1)$ называется числом степеней свободы. Она показывает, что в ограниченной совокупности (а любая выборка всегда имеет ограниченный объем) все варианты свободы принимать любые значения, кроме одной, значение которой определяется разностью между суммой всех остальных вариант и объемом выборки. В таких случаях говорят, что одна варианта не имеет степени свободы. Так, если $\sum n_i$ равна a, $\sum n_{i-1} = b$, то одна из вариант равна $n_k = \sum n_i - \sum n_{i-1}$.

Например, если четыре каких-то значения варьируют неограниченно, то их число степеней свободы 4-0=4. Когда же вариация этих значений ограничена каким-нибудь объемом, например величиной, равной 100, то три варианты (n_1, n_2, n_3) могут принимать любые значения, скажем, 27, 16, 15, или 59, 3, 37, то четвертая варианта (n_4) будет иметь только одно значение, а именно

$$n_4=100–(27+16+15) = 100–58 = 42, \text{ или } 100-(59+3+37)=100- 99=1,$$

т.е. она не имеет степени свободы. В этом случае остается только три степени свободы: 4–1=3.

В любой эмпирической совокупности всегда имеется один член, не имеющий свободы вариации. Поэтому число степеней свободы для любой выборки равно (N–1). В больших совокупностях разница между N и (N–1) неощутима, т.е. она заметно не сказывается на величине вариации и среднего квадратического отклонения. На выборках же малого объема эта разница сказывается на величине указанных показателей. Поэтому при вычислении среднего квадратического отклонения на малых выборках рекомендуется пользоваться формулой (5.5).

Способы вычисления среднеквадратического отклонения

Способов вычисления среднеквадратического отклонения вариационного ряда как и других его показателей (статистик) есть несколько. В настоящее время для вычисления статистик разработаны компьютерные программы. Они сразу дают искомый результат, но исследователь, работающий с вариационным рядом, должен понимать суть изучаемого явления. Поэтому необходимо знать алгоритм проводимых вычислений, что мы сейчас и рассмотрим.

Одним из наиболее простых способов является, так называемый, прямой или длинный. Его рационально использовать для небольшого числа наблюдений, не сгруппированных в вариационный ряд. Работа выполняется следующим образом. После того как вычислена средняя арифметическая, нужно определить отклонение от нее каждой варианты, т.е. найти значения $x_1 - \overline{X}$, $x_2 - \overline{X}$, $x_3 - \overline{X}$ и т.д. Затем каждое отклонение возводится в квадрат и, если варианта повторяются, квадраты отклонений умножаются на соответствующие частоты, и результаты суммируются. Полученная сумма $\sum \delta(\tilde{o} - \overline{\tilde{O}})^2$ делится на общее число наблюдений без единицы $(n-1)$, и из частного извлекается квадратный корень.

Возьмем для примера небольшую совокупность и вычислим для нее среднее квадратическое отклонение. Например, учитывая естественное возобновление на 10 учетных площадках, мы насчитали следующее число сеянцев: 91, 82, 76, 65, 54, 102, 94, 78, 88, 96. Их сумма ($\sum n_i$) равна 750.

Средняя арифметическая этих значений составит 750/10 = 75 сеянцев. Вычисление σ для данного ряда показано в таблице 5.5.

Таблица 5.5 – Вычисление среднего квадратического отклонения для количества сеянцев на учетных площадках. Среднее значение равно 75 шт

№ п/п	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	Вычисления
1	91	-16	256	$\sigma^2 = \frac{1}{750} \cdot 2516 = 3{,}35$ $\sigma = \sqrt{3{,}35} = 1{,}83$ шт.
2	82	-7	49	
3	76	+1	1	
4	65	-10	100	
5	54	-21	421	
6	102	+27	729	
7	94	+19	361	
8	78	+3	9	
9	88	+13	169	
10	96	+21	421	
Итого	750	-	2516	

При вычислениях на компьютере абсолютная величина чисел не имеет значения. При расчетах вручную очень крупные или очень мелкие величины удобнее увеличивать или уменьшать на 1-3 или более порядков, а итог соответственно корректировать.

При наличии больших совокупностей вычисление статистических показателей вариационного ряда чаще всего осуществляют с помощью вспомогательных величин, которые именуют моментами.

Понятие о моментах распределения

Моментом называют среднее отклонение классовых вариантов от средней величины или от любого числа.

Моменты называют начальными, ели они вычислялись от условного начала, и центральными, если вычислялись от средней ряда $\overline{X}$. Начальные моменты обозначают буквой m с индексами, указывающими на порядок момента или на степень отклонений: m_o - нулевой, m_1 - первый, m_2 - второй, m_3 - третий и m_4 - четвертый - начальные моменты. Это означает соответственно: среднее отклонение нулевой, первой степени, средний квадрат, средний куб отклонений и т.д. Причем m_o=1, так как все отклонения в нулевой степени равны единице, и, следовательно, сумма произведений их на частоты равна общему числу частот.

Центральные моменты обозначают буквой μ с теми же индексами: $\mu_0, \mu_1 \mu_2, \mu_3, \mu_4$ и т.д. - соответственно нулевой, первый, второй, третий и четвертый центральные моменты. Причем, $\mu_0 = 1$; $\mu_1 = 0$, что легко проверить, пользуясь данным понятием моментов.

Вычисление начальных моментов

Для вычисления моментов есть несколько способов. Обычно в практике применяют способ произведений и способ сумм. При проведении расчетов на компьютерах можно разрабатывать алгоритм по любому из названных способов. Более просто программируется способ произведений. Поэтому мы рекомендуем использовать именно этот способ.

Способ произведений. Техника и расчеты начальных моментов по способу произведений видны из таблице 5.6, где мы продолжаем рассматривать ряд распределения 120 деревьев сосны, показанный в таблицах 3.1 и 4.1.

В 1-м столбце таблицы вписаны классовые варианты исследуемого признака x_i, а во 2-м - соответствующие им частоты n_i. Эти два столбца цифр представляют собой исследуемый вариационный ряд. В 3-м столбце вписывают условные отклонения классовых вариант от условной средней M'. В исследуемом ряду распределения M' принято равным 32 см. Эти отклонения находят по формуле:

$$x_k = (X - M')/k, \qquad (5.6)$$

где k - величина интервала. В рассматриваемом ряду k=4 см.

Для центрального класса условное отклонение равно нулю, так как значение варианта X и условного начала M' здесь одинаковы. Начиная расчет отклонений от центрального класса, получим для классов, находящихся в стороне значений вариант меньших M' ряд чисел со знаком минус (–1, –2, –3, –4 и т.д.), а для классов, находящихся в стороне значений вариант, которые больше M' – со знаком плюс (+1, +2, +3, +4 и т.д.).

В столбцы 4-7 вписывают произведения найденных отклонений в первой, второй, третьей и четвертой степенях на частоты. Эти произведения рекомендуется находить последовательно по строкам, умножая в каждой из них число предыдущего столбца на одно и то же число – т.е., на отклонение x_k. Благодаря этому создаются условия для проверки чисел, помещенных в столбцах 4–7.

Таблица 5.6 – Вычисление начальных моментов по способу произведений для ряда распределения диаметров стволов сосны

Ступени толщины (классы) x_i	Число деревьев (частоты) n_i	Отклонение от условной средней (M'=32 см) x_k	$n_i x_k$	x_k^2	nx_k^2	x_k^3	nx_k^3	x_k^4	nx_k^4	x_k+1	$(x_k+1)^4$	$n(x_k+1)^4$
1	2	3	4	5	6	7	8	9	10	11	12	13
8	1	-6	-6	36	36	-216	-216	1296	1296	-5	625	625
12	3	-5	-15	25	75	-125	-375	625	1875	-4	256	768
16	8	-4	-32	16	128	-64	-512	256	2048	-3	81	648
20	14	-3	-42	9	126	-27	-378	81	1134	-2	16	224
24	20	-2	-40	4	80	-8	-160	16	320	-1	1	20
28	27	-1	-27	1	27	-1	-27	1	27	0	0	0
32=M'	17	0	0	0	0	0	0	0	0	1	1	17
36	12	1	12	1	12	1	12	1	12	2	16	192
40	9	2	18	4	36	8	72	16	144	3	81	729
44	5	3	15	9	45	27	135	81	405	4	256	1280
48	3	4	12	16	48	64	192	256	768	5	625	1875
52	1	5	5	25	25	125	125	625	625	6	1296	1296
Итого	120	-	$\frac{+62}{-162}$ ∑-100	-	638	-	$\frac{+536}{-1668}$ ∑-1132	-	8654	-	-	7674

Используя результаты вычислений, показанных в таблице 5.6, проведем расчеты моментов.

$$m_1 = \frac{\sum n_i x_{ki}}{\sum n_i} = -100/120 = -0{,}833;$$

$$m_2 = \frac{\sum n_i x_{ki}^2}{\sum n_i} = 638/120 = 5{,}317;$$

$$m_3 = \frac{\sum n_i x_{ki}^3}{\sum n_i} = -1132/120 = -9{,}433;$$

$$m_4 = \frac{\sum n_i x_{ki}^4}{\sum n_i} = 8654/120 = 72{,}117.$$

Моменты обычно вычисляют с точностью до 0,001. Учитывая, что при их нахождении может быть допущена ошибка, обычно проводят проверку счета по формуле:

$$m_4' = m_0 + m_1 + 6m_2 + 4m_3 + m_4, \qquad (5.7)$$

где m_4' – четвертый начальный момент, который вычислен, если начальное значение (M') сдвинуто на один разряд ниже. В нашем примере этот сдвиг сделал начальным значением величину x_i = 28 (см. таблицу 5.6).

Напомним, что $m_0 = 1$, а $m_4' = \frac{\sum n_i(x_{ki}+1)}{\sum n_i} = \frac{7674}{120} = 63{,}950$, m_4' = 1-3,332+31,908-37,732+72,117=63,961.

Из приведенных вычислений видно, что m_4', вычисленный по данным таблицы 5.6 и по формуле (5.7) практически одинаковы: разницу в 0,011 или 0,17 % можно отнести за счет округлений при расчете моментов.

Вычислив начальные моменты, приступаем к нахождению центральных моментов.

Начальные моменты как таковые не представляют самостоятельного интереса в силу условности выбора начального значения (M'). Их в основном используют для вычисления центральных моментов. Это бывает делать гораздо удобнее, чем вычислять центральные моменты непосредственно через среднее значение ($\bar{x}$) и среднее квадратическое отклонение (σ) по формуле:

$$M_k = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})n_i\,, \qquad (5.8)$$

где M_k – центральный момент порядка k;

k – 1, 2, 3, 4,…

Символы x_i, $\bar{x}$, n_i, $N=\sum n_i$ описаны выше.

Вычисления центральных моментов через начальные выполняют по формуле:

$$M_k = \sum_{i=0}^{k} C_k^j m_{k-j}(-m_1)^j\,, \qquad (5.9)$$

где $C_k^j = \frac{k!}{j!(k-j)!}$, т.е. C_k^j - это число сочетаний с k по j;

k!; *j!*; *(k-j)!* – факториалы приведенных величин, т.е. $k! = 1\cdot2\cdot3\cdot\ldots\cdot k$.

Используя (5.9) получаем формулы для **вычисления центральных моментов**:

$$\mu_0 = 1;\ \ \mu_1 = 0;\ \ \mu_2 = m_2 - m_1^2; \qquad (5.10)$$

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3\,; \qquad (5.11)$$

$$\mu_4 = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4. \qquad (5.12)$$

Для проверки центральных моментов ряда распределения применяют формулы:

$$\mu_3 = m_3 - 3\mu_2m_1 - m_1^3\,; \qquad (5.13)$$

$$\mu_4 = m_4 - 4\mu_3m_1 - 6\mu_2m_1^2 - m_1^4. \qquad (5.14)$$

Приведем теперь вычисления центральных моментов для нашего распределения 120 деревьев сосны, используя вычисленные выше начальные моменты.

Получаем следующие результаты:

μ_0=1; μ_1=0;

μ_2 = 5,317-(-0,833)^2 = 5,317-0,694 = 4,623;

μ_3 = -9,433-3(-0,833)·5,317+2·(-0,833)^3 = -9,433+13,287-1,156 = 2,698;

μ_4 = 72,117-4·(-0,833)·(-9,433)+6·(-0,833)^2·5,317-3·(-0,833)^4 = 72,117 - 31,431+ + 22,136 - 1,444 = 61,378.

Проверка (формулы (5.13), (5.14)):

μ_3 = -9,433-3·4,623·(-0,833)-(0,833)^3 = -9,433+11,553+0,578 = 2,698;

μ_4 = 72,117-4·2,698·(-0,833)-6·4,623·(-0,833)^2-(-0,833)^4 = 72,117+8,990-19,248-- 0,481 = 61,378.

Проверка показала, что центральные моменты вычислены правильно.

Моменты можно вычислять также по способу сумм. В прежние времена, когда вычисления проводили вручную, этот способ считали предпочтительным из-за меньшего количества вычислений. При расчетах на компьютере используют в основном способ произведений. Это вызвано более простым алгоритмом расчетов по способу произведений против способа сумм. Последний способ требует много сложных логических условных переходов, что реализовать на компьютере труднее, чем организовать простой счет даже больших чисел.

Хотя способ сумм не рекомендуется для современных вычислений, но его традиционно описывают в пособиях по биометрии. Здесь тоже описан этот способ в том виде как он приведен в учебнике советского ученого-таксатора Н.Н.Свалова (1925-2002), который есть в списке литературы. Технику вычислений покажем на примере ряда распределения высот сосны (таблица 5.7).

Вычисление начальных моментов по способу сумм начинается с вписывания в качестве исходных данных классовых вариантов и соответствующих им численностей (частот). Следующие за частотами столбцы нумеруют, они предназначаются для суммирования частот в определенной последовательности, которая изложена ниже. Последние два столбца таблицы не нумеруют – в них помещают вычисленные значения, необходимые для проверки верности проведенных вычислений.

Таблица 5.7 – Вычисление начальных моментов по способу сумм для ряда распределения высот сосны

	X	n_i	(1) $\sum n_i$	(2) $n_{i+1}=n_{i-1}+n_i$	(3)	(4)	(5)	(x_k+1)	$n_i(x_k+1)^2$ (из табл. прил. 2)
b	20	1	1	1	1	1	1	-6	1296
	21	2	3	4	5	6	7	-5	1250
	22	2	5	9	14	20	27	-4	512
	23	3	8	17	31	51	-	-3	243
	24	4	12	29	-	-	-	-2	64
	25	2	14	43	-	-	-	-1	2
	26	13	27 k_2	-	-	-	-	0	0
условн. средняя	M'=27	14 k_1	-	-	-	-	-	+1	14
α	28	25	53 k_3	-	-	-	-	+2	400
	29	18	28	38	-	-	-	+3	1458
	30	10	10	10	10	-	-	+4	2560
	∑	94	$\sum a$91α	48 α	10 α	-	-	-	7789
			$\sum b$70b	103 b	111 b	78	35		
		S= α+b	s_1 161	s_2 151	s_3 121	s_4 78	s_5 35	$m_4^* = \sum n_i(x_i+1)^2 / \sum n_i$	
		d= α-b	d_1- 21	d_2–55	d_3–101	d_4–78	d_5–-35	$m_4^* = 7799/94 = 82{,}968$	

Проверка суммировании (сумм *a* и *b)* приведена в таблице 5.8.

Таблица 5.8 – Проверка суммирования *a* и *b*

a_1=38+53=91	b_1=43+27=70	b_2=60+43=103	m_o = 1,000
a_2=10+38=48	b_3=51+60=111	b_4=27+51=78	$4m_1$ = 0,892 $6m_2$ = 29,550 $4m_3$ = –38,936 m_4 = 90,457
			82,963

Проверка $N=k_1+k_2+k_3= 27+14+53=94$

Начальные моменты: m_1=21:94=0,223

m_2 = (161+2·151)/94=463/94=4,925

m_3 = [21+6(-55)+6] (-101)/94=-915/94=-9,734

m_4 = 161+14·151+36·121+24·78)/94=8503/94=90,457

Затем против частоты, соответствующей условной средней *M′*, проводим черту через все нумерованные столбцы таблицы, разделяя последнюю на две части - верхнюю и нижнюю. В столбцах 2, 3, 4, 5 добавляют сверху и снизу от проведенной общей черты дополнительные

черточки в возрастающем количестве 1, 2, 3, 4 и т.д. Таким образом, получается фигура из черточек в виде треугольника.

Составление таблицы состоит в следующем. Численности первого и последнего класса (в нашем ряду 1 и 11-го) вписывают в те же строки, т.е. первого и последнего классов, во все столбцы незанятые чертой. Каждое последующее число столбца (1), …, (5) получают как сумму двух чисел, одно из которых стоит рядом с образуемым числом слева, а другой – над ним (в верхней части таблицы) или под ним (в нижней части таблицы). Строки, занятые черточками, не заполняют. Внизу каждого столбца выписывают суммы верхней и нижней частей таблицы.

Одну из этих вспомогательных сумм, находящуюся в стороне вариант, значения которых больше условной средней M' обозначают буквой α, а другую сумму – буквой b. Алгебраические суммы этих вспомогательных сумм обозначают буквой s, а разности их буквой d (в столбцах 1, 2, 3, 4, 5 будем иметь соответственно s_1, s_2, s_3, s_4, s_5 и d_1, d_2, d_3, d_4, d_5).

Правильность суммирования в 1-м столбце проверяют, сложив наибольшие числа верхней и нижней частей этого столбца с частотой, стоящей против начального значения. Сумма этих трех чисел должна равняться объему ряда. В примере расчета, приведенном в таблице 5.7, имеем 27+14+53=94.

Проверка суммирования во 2-м и следующем столбцах состоит в сложении последнего наибольшего числа верхней или нижней части проверяемого столбца с последним числом предыдущего столбца, расположенным строкой выше (при проверке нижней суммы) или строкой ниже (при проверке верхней суммы).

Проверка сумм α и b приведена в таблице 5.8.

Начальные моменты вычисляют по формулам:

$$m_1 = d_1 / N, \qquad m_2 = (s_1 + 2s_2)/N, \tag{5.15}$$

$$m_3 = (d_1 + 6d_2 + 6d_3)/N, \tag{5.16}$$

$$m_4 = (s_1 + 14s + 36s_3 + 24s_4)/N \tag{5.17}$$

Вычисление начальных моментов приведено в той же таблице 5.7.

Вычисление центральных моментов. Формулы для вычислений те же, что приведены ранее, когда моменты нашли по способу произведений (5.10-5.14).

$$\mu_2 = 4{,}925 - 0{,}223^2 = 4{,}925 - 0{,}049 = 4{,}876.$$

$$\mu_3 = -9{,}734 - 3 \cdot 4{,}925 \cdot 0{,}223 + 2 \cdot 0{,}223^3 =$$

$$= -9{,}734 - 3{,}295 + 0{,}022 = -13{,}007$$

$$\mu_4 = 90{,}457 - 38{,}936 \cdot 0{,}223 + 29{,}550 \cdot 0{,}050 - 3 \cdot 0{,}002 =$$

$$= 90{,}457 + 8{,}683 + 1{,}477 - 0{,}006 = 100{,}611.$$

Проверка:

$$\mu_3 = -9{,}734 - 3 \cdot 4{,}876 \cdot 0{,}223 - 0{,}223^3 =$$
$$= -9{,}734 - 3{,}262 - 0{,}011 = -13{,}007.$$
$$\mu_4 = 90{,}457 - 0{,}892(-13{,}007) - 6 \cdot 4{,}876 \cdot 0{,}223^2 - 0{,}223^4 =$$
$$= 90{,}457 + 11{,}602 - 1{,}454 - 0{,}002 = 100{,}603.$$

5.4 Асимметрия, эксцесс, коэффициент вариации

Центральные моменты используют для вычисления главных статистических показателей ряда распределения: среднего значения ($\bar{\delta}$), среднего квадратического отклонения (σ), коэффициента вариации (υ), показателей асимметрии (α) и эксцесса (Е).

Асимметрию и эксцесс находят через основные моменты. Последние определяют по формулам:

$$rh = \frac{\mu h}{\sigma h}, \qquad (5.18)$$

где rh – основной момент с показателем h;

μh – центральный момент с показателем h;

σh – основное отклонение в степени h.

Из формулы (5.18) следует, что основной момент равен отношению центрального момента того или иного порядка к основному отклонению в соответствующей степени. Для распределения 120 деревьев сосны по диаметру, мы определили m_0=1; m_1=0; m_2=4,623; m_3=2,698; m_4=61,378, а σ=1,830. Тогда, проведя вычисления по формуле (5.18), учитывая значения μ_0 , μ_1, вытекающие из формул (5.9)-(5.12), получим:

$$r_0 = \frac{\mu_0}{\sigma_0} = \frac{1}{(1{,}830)^0} = \frac{1}{1} = 1;$$
$$r_1 = \frac{\mu_1}{\sigma_1} = \frac{0}{1{,}830} = 0;$$
$$r_2 = \frac{\mu_2}{\sigma_2} = \frac{4{,}623}{(1{,}830)^2} = \frac{4{,}623}{3{,}349} = 1{,}380;$$
$$r_3 = \frac{\mu_3}{\sigma_3} = \frac{2{,}698}{(1{,}830)^3} = \frac{2{,}698}{6{,}128} = 0{,}440;$$
$$r_4 = \frac{\mu_4}{\sigma_4} = \frac{61{,}378}{(1{,}830)^4} = \frac{61{,}378}{11{,}215} = 5{,}473.$$

Третий основной момент (r_3) представляет собой хаарктеристику скошенности (косости) ряда распределения и именуется асимметрией, которая обозначается α= r_3.

Четвертый основной момент служит для нахождения крутости ряда распределения, который называется эксцессом и обозначается $E=r_4-3$. Для нашего примера α=0,440, E=2,473.

Среднеквадратическое отклонение, как и средние величины ряда, является именованной величиной, выражающейся в величинах измерения ряда. Для решения многих теоретических и практических вопросов лесной биометрии нужны относительные величины, характеризующие общие особенности размаха ряда распределения. Таким показателем является коэффициент вариации, который в литературе по лесному хозяйству называют еще показателем изменчивости таксационных признаков лесных насаждений.

Коэффициент вариации представляет собой показатель изменчивости изучаемого признака, выраженный в относительных единицах, обычно в процентах. Он определяется по формуле

$$V = \frac{\sigma}{\overline{X}} \cdot 100\%, \qquad (5.19)$$

где υ – коэффициент вариации;

σ – среднеквадратическое отклонение;

$\overline{O}$ - среднее значение.

Так как коэффициент вариации не зависит от принятых единиц измерения (при делении σ на $\overline{O}$ единицы измерения взаимно уничтожаются), то он применяется при сравнительной оценке варьирования различных признаков. В лесном хозяйстве это могут быть диаметр дерева, его высота, объем, прирост и т.д.

Значение коэффициента вариации используют для вычисления точности исследования (P) по формуле

$$P = \frac{V}{\sqrt{N}}, \qquad (5.20)$$

где N – объем выборки.

Для примера вычислим υ и P по данным таблиц 5.5 и 5.6.

Для измеренного количества сеянцев (таблица 5.5) σ=1,83 шт., N=10, среднее значение ($\overline{O}$) = 75.

Коэффициент вариации сеянцев на площадках равен

$$V = \frac{1,83}{10} \cdot 100\% = 18,3\%.$$

Тогда точность исследования составит (формула (5.20))

$$P = \frac{18,3}{\sqrt{10}} = \frac{18,3}{3,162} = 5,8\%.$$

Для замеренных 120 диаметров сосны (таблица 5.6) значения υ и P следующие: $\overline{\tilde{O}}$=28,7 см (см. главу 4, таблица 4.1); $\sigma = C\sqrt{\mu_2}$ =8,83, где C – величина ступени толщины (4 см); N=120 шт.

Тогда

$$V = \frac{8{,}83\ см}{28{,}7\ см} \cdot 100\% = 30{,}8\%;$$

$$Ð = \frac{30{,}8}{\sqrt{120}} = 2{,}8\%.$$

Полученные величины варьирования диаметров соответствуют данным, приводимым в исследованиях по лесной таксации. Точность в 2,8% достаточная при проведении измерений в практике лесного хозяйства и низкая для научных исследований, где требуется точность в 1,5-2%.

Из формулы (5.20) можно определить необходимое количество наблюдений (N) при заданной точности (P).

Если $P = \frac{V}{\sqrt{N}}$, то

$$P\sqrt{N} = V;$$

$$\sqrt{N} = \frac{V}{P};$$

$$N = \frac{V^2}{P^2}.$$

Для наших 120 деревьев, где υ = 30,8%, необходимое число наблюдений при разной точности составит

Для P = 10% $N = \frac{(30{,}8)^2}{10^2} = \frac{949}{100} \approx 10$ деревьев;

Для P = 5% $N = \frac{(30{,}8)^2}{5^2} = \frac{949}{25} \approx 40$ деревьев;

Для P = 2% $N = \frac{(30{,}8)^2}{2^2} = \frac{949}{4} \approx 240$ деревьев;

Для P = 1,5% $N = \frac{(30{,}8)^2}{(1{,}5)^2} = \frac{949}{2{,}25} \approx 420$ деревьев;

Для P = 1% $N = \frac{(30{,}8)^2}{1^2} = \frac{949}{1} \approx 950$ деревьев.

Следовательно, чтобы изучить наш древостой с приемлемой точностью, необходимо измерить 240-420 стволов, в среднем 300-350.

Обобщая изложенное в настоящей главе, приведем формулы для определения статистических показателей ряда распределения через моменты:

Средняя арифметическая $\overline{\tilde{O}} = M' + \overline{\tilde{N}}m_1$ (5.21)

Среднее квадратическое отклонение:

а) в единицах интервала

(дисперсия) $\sigma = \sqrt{\mu_2}$ (5.22)

б) в единицах измерения $\sigma = \tilde{n}\sqrt{\mu_2}$ (5.23)

Коэффициент вариации $V = \frac{\sigma}{\tilde{O}} \cdot 100\%$ (5.24)

Показатель асимметрии $\alpha = r_3$ (5.25)

Показатель эксцесса $E = r_4 - 3$ (5.26)

Показатель точности исследования $P = \frac{V}{\sqrt{N}}$ (5.27)

В приведенных формулах условные обозначения ($\bar{x}$, *M'*, m_1, μ_2, σ, r_3, r_4, V, α, *E, C, P,*) показаны выше.

6. ФУНКЦИИ РАСПРЕДЕЛЕНИЯ. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

6.1 Понятие о видах распределения

Массивы случайных величин обычно распределяются не хаотично, а по некоторому закону. Эти законы распределения приводят к разным видам распределений. Все распределения можно выразить графически, что в основном и делали в лесном хозяйстве до 50-60 годов XX века. В то же время наиболее корректно выражать распределение через некоторые математические функции. В настоящее время это стало основным методом описания распределений, т.к. применение компьютеров сделало такую работу простой, быстрой и доступной. В то же время мы должны ясно понимать суть изучаемого явления. Этого не достичь, если исследователь использует компьютер только как пользователь.

Распределения могут быть дискретными и непрерывными. Так, если мы рассматриваем некоторое распределение, например, деревья в древостое, выражаемое конкретными числами: 4, 8, 12, ..., то оно будет дискретным. При описании распределения непрерывной функцией распределение становится непрерывным. Часто между ними трудно провести грань: все зависит от цели исследования, величины классового промежутка и т.д.

Все, что может быть измерено или исчислено в живой природе, называют величиной постоянной или переменной. В зависимости от обстоятельств эти величины могут принимать разные значения. Переменную величину считают определенной, если заранее, до опыта можно указать ее значение. Если же в одних и тех же условиях переменная величина может принимать разные значения, которые заранее указать нельзя, она называется с л у ч а й н о й величиной. Понятие случайной величины, как и понятие случайного события, относится к фундаментальным в теории вероятностей.

Случайные величины бывают зависимыми и независимыми. Случайные величины называют **независимыми**, если вероятность любого значения одной величины (X) не зависит от того, какие значения принимает другая величина (У). В противном случае эти величины находятся в зависимости одна от другой и называются **взаимозависимыми** случайными величинами. Например, мы изучаем некоторый участок леса. Там имеется определенная почва, на которой растут деревья. На одной и той же почве могут расти разные деревья: сосна, береза, ель и другие. Механический состав почвы не зависит от того, какой древесный вид сегодня здесь растет. Таким образом при рассмотрении системы почва-дерево характеристики почвы, скажем, процент физической глины, будет величиной независимой. В то же время высота дерева,

да и сам породный состав лесного участка определяется почвенными условиями, т.е. показатели роста деревьев зависят от почвенных характеристик и являются величиной зависимой. Если мы будем рассматривать диаметр и высоту дерева, то обнаружим, что с изменением одного показателя меняется и второй. Это величины взаимозависимые.

Существуют разные типы случайных величин. Для лесовода и биолога наиболее существенное значение имеют дискретные и непрерывные случайные величины. С ними мы уже встречались выше при рассмотрении эмпирических распределений. Дискретная случайная величина принимает лишь отдельные счетные значения, для которых можно указать вероятности, или частоты. Непрерывная случайная величина может принимать любые значения в некотором заданном интервале. Указать вероятности или частоты ее отдельных значений, вообще говоря, невозможно, поэтому они относятся к тем значениям, которые эта величина принимает в интервале (от - до), причем этот интервал может быть каким угодно - и большим и малым.

6.2 Эмпирические функции распределения

В лесной биометрии мы, как правило, имеем дело с эмпирическими функциями распределения. Это значит, что, выполняя некоторые измерения случайной величины, например, диаметры деревьев в лесу, получаем распределение и определяем вид этого распределения, руководствуясь графиком некоторой функции распределения.

Очень важно знать распределения признаков, наиболее часто встречающиеся в лесных исследованиях. Это позволяет применять различные способы обработки выборочных данных не слепо, а осмысленно, а также правильно оценивать результаты опытов и наблюдений, объективно и точно сравнивать их между собой.

В эмпирических распределениях, т. е. тех, которые наблюдаем, проводя измерения в лесу, бросается в глаза одна важная особенность - преимущественное накапливание вариант в центральных классах и постепенное убывание их числа по мере удаления от средней арифметической вариационного ряда. Эта особенность, составляющая одну из характерных черт варьирования биологических признаков вообще и лесохозяйственных в частности, факт фундаментального значения, имеющий широкое распространение в природе.

Такую картину получаем, если проанализируем распределение людей по росту, диких животных, например зайцев, по весу, величину ступни у взрослых людей одного пола и т. д. На рисунке 6.1 в качестве классического примера показано распределение размеров мужской обуви среди жителей Восточной Европы, приводимое в большинстве учебников по биометрии.

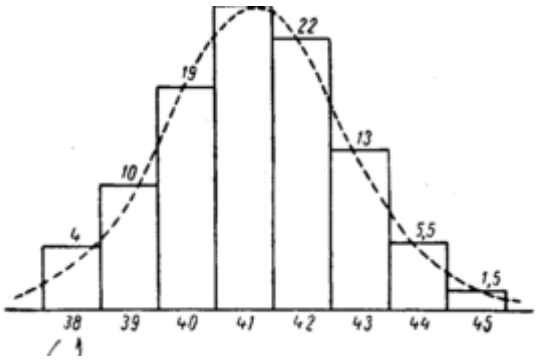


Рисунок 6.1 Гистограмма распределения размеров мужской обуви среди населения в Восточной Европе
(на оси абсцисс - номера обуви, на оси ординат - проценты)

Именно такие закономерности распределения являются основными для заказов обувным фабрикам на пошив обуви определенного размера, чтобы полностью удовлетворить спрос и избежать потерь от нераспроданных экземпляров.

Такую закономерность, т.е. концентрацию наибольших численностей в середине ряда распределения, впервые описал бельгийский статистик А.Кетле в 1835 году, исследовавший распределение нескольких тысяч солдат американской армии по росту.

Деревья в лесу распределяются по толщине, высоте и другим признакам аналогично. Если построим график по данным перечета 120 стволов сосны (таблица 4.1.), то увидим, что он одновершинный, наибольшая концентрация деревьев наблюдается в средних ступенях толщины (рисунок 6.2).

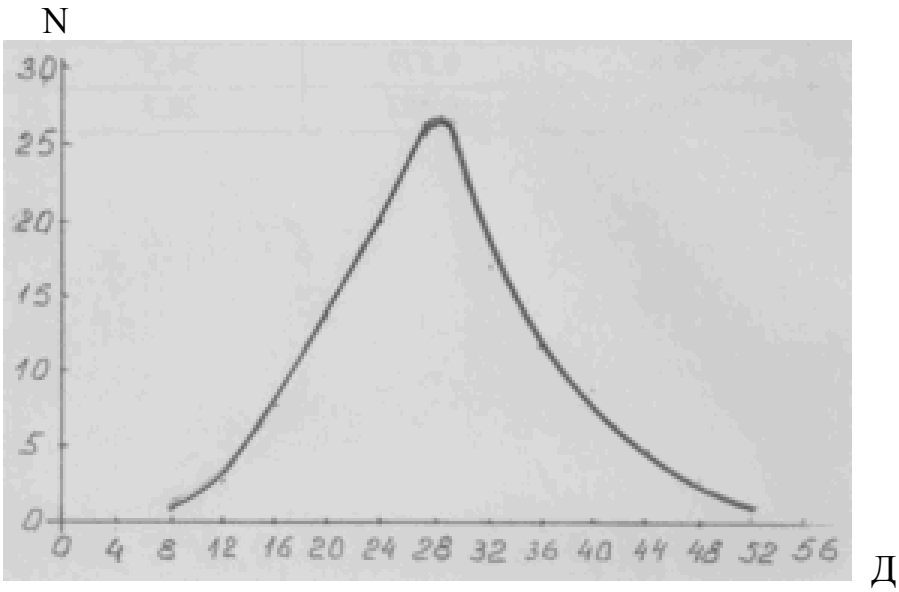


Рисунок 6.2 Распределение 120 деревьев сосны в древостое в возрасте 100 лет

6.3 Функция нормального распределения и ее параметры

Выше показано, что распределение случайных величин в биологических, в том числе и лесных, совокупностях носит закономерный характер. Есть много функций, которыми описывают названные распределения. Наиболее универсальной и используемой чаще других является уравнение кривой нормального распределения или функция Гаусса-Лапласса. Ее суть заключается в том, что частота отклонения отдельных вариант от средней арифметической данной совокупности есть функция их величины. Вероятность частоты той или иной варианты в генеральной совокупности и определяется этой функцией.

Графически функция нормального распределения схожа с графиками на рисунках 6.1. и 6.2. В то же время график, который строго соответствует уравнению кривой нормального распределения, выглядит следующим образом (рисунок 6.3.).

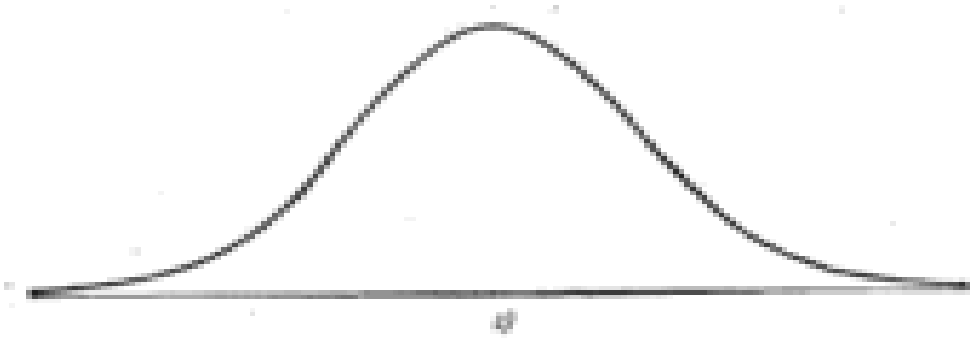


Рисунок 6.3 Кривая нормального распределения

Уравнение кривой нормального распределения выражает зависимость теоретических численностей f(x) или y от значений x, являющейся непрерывно распределяющейся случайной величиной. Есть разные формы выражения этой кривой. Основная форма написания этого выражения

$$У = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2} \qquad (6.1)$$

Здесь y – ордината или высота кривой на любом расстоянии от $\overline{X}$, т. е. центра распределения. Вправо от этого центра случайная величина x_i имеет положительные, а влево – отрицательные значения.

σ или среднеквадратическое отклонение характеризует амплитуду колебания отдельных значений случайной величины около средней арифметической;

($x_i - \overline{X}$) - отклонение варианты от средней арифметической, являющейся серединой ряда;

выражение $\frac{1}{\sigma\sqrt{2\pi}}$ - максимальная ордината, соответствующая точке $\overline{X}$.

По мере удаления от этой точки, т.е. центра распределения, плотность значений случайной величины падает и кривая асимптотически приближается к оси абсцисс;

Так как π – 3,14593 и e – основание натуральных логарифмов, равное 2,7183, являются постоянными величинами, следовательно величина $\frac{x_i - \overline{X}}{\sigma} = t$ есть не что иное как нормированное отклонение. Эта величина имеет большое значение для исследования свойств нормального распределения.

Найденные для различных значений *t* величины *У* дают ординаты нормальной кривой. Непременным условием нормирования служит требование, чтобы вся площадь, заключенная под кривой вероятности (нормальной кривой), равнялась единице.

Если принять σ=1, то уравнение (6.1) будет иметь следующий вид:

$$У = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(x_i - \overline{x}\right)^2} \qquad (6.2)$$

Кривая, описываемая этим уравнением, получила название стандартизованной кривой распределения, или кривой Гаусса-Лапласа. Она выражает закон нормального распределения с площадью под кривой, равной единице.

Чтобы ордината *У* выражала не вероятностные, а абсолютные численности, т.е. частоты вариант, нужно в числитель правой части уравнения (6.1) ввести в качестве дополнительных множителей N - общее число вариант данной совокупности и i - величину классового интервала (если вариация разбита на классы).

Тогда уравнение (6.1.) принимает следующий вид:

$$У = \frac{N \cdot i}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}t^2} \qquad (6.3)$$

Здесь У обозначает теоретическую частоту вариационного ряда, соответствующую нормированному отклонению t.

Кривая нормального распределения обладает следующими свойствами:

- Однозначно определяется двумя параметрами: $\overline{X}$ - средним значением и σ – среднеквадратическим отклонением.
- Кривая симметрична относительно среднего значения ($\overline{X}$) и имеет колоколообразную форму, которая зависит от величины σ, являющейся параметром масштаба, а положение определяется $\overline{X}$.

- Кривая имеет один максимум, равный $\frac{1}{\sigma\sqrt{2\pi}}$ и две точки перегиба на расстоянии ±σ от $\overline{X}$.
- Ветви кривой асимптотически приближаются к оси абсцисс на расстояние ±∞.

Итак, нормальный закон распределения выражает функциональную зависимость между величиной признака и его частотой в генеральной совокупности. Чем больше отклонение варианты от средней величины, тем меньше ее частота, и наоборот, чем меньше варианта отклоняется от средней арифметической, тем больше ее частота в данной совокупности. Следовательно, вероятность отклонения любой варианты от средней арифметической есть функция нормированного отклонения. Эта функция выражается с помощью асимптотических, т.е. приближенных формул (6.1), а также графически (см. рисунок 6.3) и в форме таблиц. Таблица значений вероятности, соответствующие разным значениям нормированного отклонения t приведена в приложении А. Пользуясь этой таблицей, можно по двум параметрам - $\overline{x}$ и σ - построить теоретический вариационный ряд, что имеет значение при сравнительной оценке эмпирических распределений.

Выше было показано, что нормальное распределение является широко распространенной закономерностью в живой природе, в том числе и в лесных насаждениях. Такому явлению должно быть некоторое убедительное основание. Его привел известный русский математик и механик А.М.Ляпунов (1857 – 1918), доказав в 1901 году предельную теорему Ляпунова, которая относится к области теории вероятностей.

Учитывая фундаментальность нормального закона распределения и его большую практическую значимость, приведем краткое изложение теоремы Ляпунова в интерпретации известного советского математика А.К.Митропольского.

Теорема Ляпунова утверждает, что если значения независимых случайных величин будут малы в сравнении с их суммой, то при неограниченном возрастании числа этих величин распределение их суммы становится приближенно нормальным.

Условия теоремы Ляпунова являются настолько широкими, что во многих случаях их можно предполагать выполняющимися. Поэтому, если есть основание рассматривать изучаемую величину как сумму многих независимых случайных величин, влияние каждой из которых на эту сумму практически ничтожно, то, если даже распределения составляющих величин нам неизвестны, можно часто заранее быть уверенным, что изучаемая величина имеет нормальное распределение.

Благодаря этому становится ясным, что, например, распределение случайных ошибок при измерениях будет нормальным. Точно так же нормальным будет распределение физических признаков людей, распределение механических свойств материалов и т. д. Этот вывод полностью подтверждается многочисленными исследованиями.

Таким образом, теорема Ляпунова дает объяснение тому важному положению, что во многих случаях величины имеют нормальное распределение.

Причину такого соответствия нормальной кривой полученным при наблюдении рядам распределения можно видеть в выполнении тех самых условий, на основании которых теоретически появляется нормальная кривая. Можно предположить, что отдельные наблюденные значения являются результатом бесчисленного множества весьма незначительных независимых между собой причин, каждая из которых может произвести очень малое положительное или отрицательное отклонение от среднего значения исследуемой величины.

Для наглядного выяснения тех условий, при которых возникает нормальное распределение, построен прибор Гальтона. Этот прибор представляет ящик, изображенный на рисунке 6.4. Вверху прибора устроено отверстие в виде воронки, которое ведет в узкое пространство между доской и стеклом. Внизу прибора помещены перегородки, образующие несколько отделений. Все пространство между воронкой и отделениями занято рядами игл, расставленных в шахматном порядке. Если прибор поставить в наклонном положении и сыпать в воронку мелкую дробь, то иглы будут делить эту дробь на отдельные потоки. Вследствие этого дробинки расположатся не равномерно, а образуют в своей совокупности нормальную кривую.

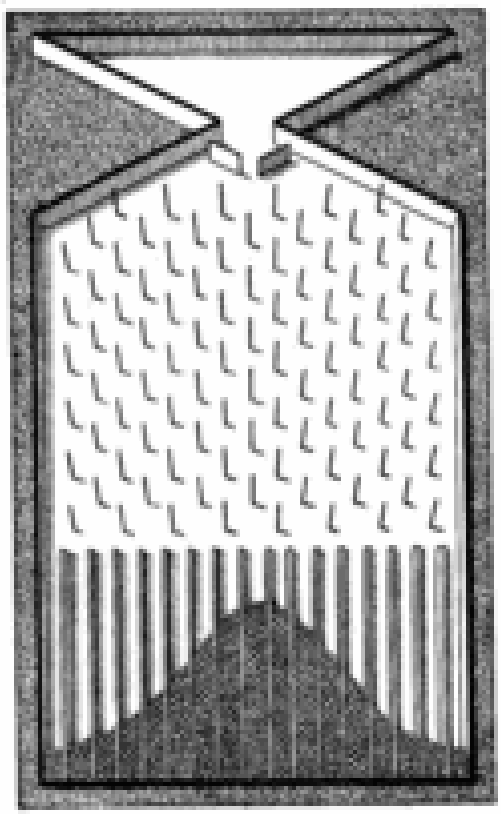


Рисунок 6.4 Прибор Гальтона

Рассматриваемый прибор ясно показывает то действие, которое оказывает множественность причин на возникновение явления. В самом деле, при

отсутствии игл в приборе весь поток дроби шел бы без отклонений вниз, и в результате дробь скопилась бы между двумя — тремя соседними вертикальными полосками против отверстия воронки. Однако препятствия в виде игл отклоняют дробинки в разные стороны, причем для большей части дробинок эти отклонения взаимно уравновешиваются, и все распределение принимает характерный вид: в отделении против воронки скопляется наибольшее количество дроби, а в остальных отделениях будет дроби тем меньше, чем дальше они отстоят от среднего отделения, так как для того чтобы отклониться от него далеко, необходимо встретить много односторонних препятствий, а это случается только с небольшим числом дробинок.

Картина распределения, подобная описанной выше, наблюдается при изучении случайных величин, взятых из любой области.

Примеров, когда применим закон, вскрытый теоремой Ляпунова, множество. Проанализируем действие этого закона в лесу. Так, на древостой, растущий без вмешательства человека, действует множество абиотических и биотических факторов. Ресурсы роста дерева определяются наследственными свойствами конкретного семени, места, где оно произрастает (условия произрастания имеют определенную неоднородность даже в пределах одного таксационного участка), соседними деревьями, случайными повреждениями (снегом, животными, молнией и т.д.) и многими другими причинами. Каждая из причин носит случайный характер и ее влияние на их совокупный результат приводит к нормальному распределению. Здесь мы исключаем антропогенный фактор, особенно рубки ухода, воздействие которых велико и целенаправленно, что имеет следствием другие распределения диаметров деревьев, о чем будет сказано ниже.

Нормальное распределение можем наблюдать, если проанализируем отклонение длины реальных сортиментов, заготовленных в лесу, от заданной их длины. Например, мы выпиливаем еловые пиловочные бревна длиной 6,5м. По действующему стандарту здесь допустимо отклонение в большую сторону (припуск) до 1%. В нашем случае это составит 6,5 см. При реальной раскрижевке из-за разных случайных причин припуски получаются неодинаковой длины. Средний припуск здесь будет отклоняться от средней величины (примерно 4 см) на ± 3 см, а весь массив чисел, характеризующих отклонения от 6,5 см будет распределен нормально.

Основные свойства нормального распределения. Теоретически случайная величина X, распределенная по нормальному закону, может принимать любые значения, меняясь от - ∞ до +∞. На самом же деле, как это видно из графика нормальной кривой (см. рисунок 6.3), значения вероятности по мере удаления от центра распределения $\bar{x}$ быстро убывают.

Если в обе стороны от $\bar{x}$ отложить отрезки, равные 3σ, то получатся точки $\bar{x}-3\sigma$ и $\bar{x}+3\sigma$; их можно выразить в нормированной форме:

$$t_1 = \frac{(\bar{x} - 3\sigma) - \bar{x}}{\sigma} = -3 \text{ и } t_2 = \frac{(\bar{x} + 3\sigma) - \bar{x}}{\sigma} = +3$$

Сделаем анализ вероятности того, что значение случайной величины окажется в данном интервале, т.е. между $\bar{x}-3\sigma$ и $\bar{x}+3\sigma$. Как утверждает теория вероятностей, искомая вероятность приближенно равна:

$$P(x_1 \le X \le x_2) \approx \frac{1}{2}[\Phi(t_2) - \Phi(t_1)], \qquad (6.4)$$

где выражение $P(x_1 \le X \le x_2)$ обозначает вероятность (P) того, что случайная величина X находится между заданными пределами ($x_1 - \bar{x}$) и ($x_2 - \bar{x}$);

t_1 и t_2 - нормированное отклонение, т.е. $t_1 = \frac{x_1 - \bar{x}}{\sigma}$ и $t_2 = \frac{x_2 - \bar{x}}{\sigma}$;

символ Φ(t), читаемый как “фи от t”, называется **интегральной функцией Лапласа**. Одно из важных свойств этой функции заключается в том, что она стремится к единице, если t неограниченно возрастает. Значения этой функции, соответствующие разным значениям t, и составляют содержание таблицы, приведенной в приложении А.

Подставим в уравнение (6.4) взятые значения t_1 и t_2:

$$P(\bar{x} - 3\sigma) \le X \le (\bar{x} + 3\sigma) = \frac{1}{2}[\Phi(3) - \Phi(-3)] = \Phi(3).$$

Вычисления показывают, что Φ(3) = 0,9973 ≈ 1. Это значит, что случайная величина, распределенная по нормальному закону, практически не отклонится от центра распределения $\bar{x}$, т.е. средней арифметической генеральной совокупности, более чем на 3σ. Этот важный для нас вывод известен в биометрии под названием “правила плюс - минус трех сигм”. Например, мы изучили расперделение по весу самцов разновозрастной популяции дикого кабана на территории некоторого лесхоза. Это распеределение будет близким к нормальному. Допустим, что средний вес кабана ($\bar{x}$) в этой популяции составил 100 кг, а среднеквадратическое отклонение (σ) равно 30 кг, т.е. изменчивость (коэффициент вариации (V) составит 30%. Тогда 68% кабанов будут иметь вес от 70 до 130 кг. Для 95% этой популяции колебания составят от 40 до 160 кг. При анализе 99,9% этой популяции вес отдельных особей может колебаться от 10 до 190 кг. Только очень малое их количество (отдельные особи) могут иметь больший вес, что мы и наблюдаем в природе.

Из таблицы в приложении А видно, что между пределами от -t до +t находится 0,6828, или 68,28%, всей площади, заключенной под кривой вероятности. Площадь этой кривой, ограниченная пределами от -2t до +2t, составляет 0,9545 долей единицы, или 95,45%, а в пределах от -3t до +3t находится 0,9973, или 99,73% всей площади нормальной кривой. Выражаясь более конкретным языком, правило “плюс - минус трех сигм” гласит, что в пре-

делах $\bar{x} \pm 1\sigma$ находится 68,28% всех вариант эмпирической совокупности, распределяющейся по нормальному закону; в пределах $\bar{x} \pm 2\sigma$ заключено 95,45%, а в пределах $\bar{x} \pm 3\sigma$ содержится 99,73% всех вариант совокупности.

Таким образом, нормальную кривую можно разделить на три участка или "сигмальные зоны", каждая из которых содержит определенное число вариант. Границы этих "зон" приблизительно совпадают с наиболее заметными изгибами нормальной кривой (рисунок 6.4.).

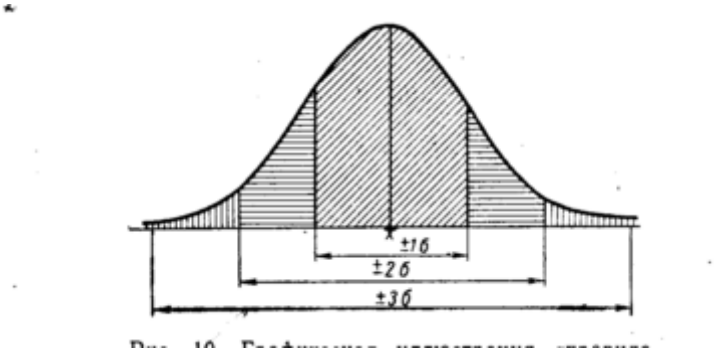


Рисунок 6.4 Графическая иллюстрация "правила плюс - минус трех сигм"

Обобщая изложенное, можно сказать, что, как указывают известные советские ученые-лесоводы К. Е. Никитин и А. З. Швиденко, в процессе статистического анализа лесоводственной информации, относящейся к некоторой случайной величине, теорию распределений применяют в двух основных направлениях:

- Как основу статистических выводов, в частности, оценки параметров и проверки статистических гипотез;
- Как средство и метод представления выборочных распределений.

В первом случае часто основополагающую роль играет нормальный закон распределения, во втором - в качестве модели можно применять самые различные типы распределений. При этом в сходных практических ситуациях при изучении одной и той же величины возможно использование разных теоретических схем, что объясняется неполнотой соответствия реальной ситуации теоретическим предпосылкам, необходимым для формирования того или иного распределения, и ограниченностью объема выборки. Последнее предопределяет приближенный характер решения задачи, необходимость статистической оценки ее результатов и объясняет обычно применяемые термины: "аппроксимация распределений", или "подгонка".

6.4 Вычисление выравнивающих частот кривой нормального распределения

Вычисление выравнивающих частот с помощью кривой нормального распределения начинают с нахождения статистик вероятностного ряда: $\bar{d}$, σ,

α, Е. Затем анализируем величины асимметрии и эксцесса. Теоретически для кривой Гаусса-Лапласа они равны 0. Но на практике такое наблюдается редко. В то же время, если α и Е относительно невелики, то аппроксимацию оправдано делать с помощью нормального распределения.

При относительной ассиметрии (α≥0), распределение скошено влево, т.е. более длинная ветвь кривой расположена слева и наоборот – при α<0 распределение скошено вправо. Знак показателя эксцесса (крутости) ряда распределения характеризует степень сосредоточения частот в центральной части распределения. При Е>0 вершина кривой будет более высокая и острая, т. е. большее число вариант сосредоточено в центральных разрядах и наоборот, при Е<0 кривая выглядит более плоской.

Для оценки возможности использования нормального распределения, когда α и $e \neq 0$, можно оценить, используя следующие признаки.

Отнесение распределения к нормальному можно оценить с помощью вычисления t-критерия Стьюдента для α и E и сравнения его с табличными значениями этого критерия при некотором числе степеней свободы ν $n = N - 1$ (приложение Е). Обычно нормальное распределение применяют, если модуль $\alpha \le 0.3$ и $E \le 0.4$. Расчет теоретических частей (n_i) проводят по формуле

$(n_i) = \left[(cN)/\left(t\sqrt{2\pi}\right)\right]e^{-\frac{t}{2}}$, где N – объем ряда распределения, с – величина интервала; t - нормированное отклонение классовых вариант x_i от среднего значения $\bar{X}$.

Остальные обозначения (t, π, e) показаны в формулах 6.1 – 6.2

При расчетах с помощью моментов

$t = (x_k - m_1)/\bar{t}$, где

x_k - условные отклонения;

m_1 - первый начальный момент;

$\bar{t}$ - среднее квадратическое отклонение в единицах интервала:

$\bar{t} = \sigma i = \sqrt{\mu^2}$.

Для класса, в котором наблюдается наибольшее количество исследуемых величин, т. е., там, где $x_i = \bar{x}$, а $t = 0$, теоретическая частота равна

$n_0 = \left[(cN/\sigma)\right] * 0.39894 * e^0 = (0.4 * c * N)/\sigma$

Множитель $f(t)$ называется функцией нормированного отклонения. Эта функция показывает значение вероятностей распределения величины x_i в зависимости от t. Значения $f(t)$=0,39894* $e^{-\frac{t^2}{2}}$ получаем из специальных таблиц (приложение Б). В результате получается следующий рабочий вид уравнения кривой нормального распределения

$n = [(c * N) * \sigma] * f(t)$ или $n = \left[\left(N / \overline{\sigma}\right) * \sigma\right] * f(t)$.

Пример расчета выравнивающих частот по кривой Гаусса-Лапласа приведен в таблице 6.1.Вычисления $\overline{X}$, σ, α, *E* и *m* сделаны по формулам, приведенным в главе 5 и здесь опущены.

Таблица 6.1 – Вычисление выравнивающих частот кривой нормального распределения для древостоя ольхи черной в возрасте 60 лет

$\overline{x} = 88.54$; $\overline{x} = 28.0$; $\sigma = 2.033$; $\alpha = -0.07$; E=-0.51; m_1=0

x_i ступени толщины	n_i численности	x*k	$\frac{x*k-m}{\sigma}$	f(x)	$\tilde{n}_j$	Округление $\tilde{n}_i$
8	2	-5	-2,459	0,0196	1,74	2
12	5	-4	-1,967	0,0570	5,05	5
16	15	-3	-1,476	0,1392	12,32	12
20	22	-2	-0,984	0,2468	21,85	22
24	28	-1	-0,443	0,3530	31,25	31
28	33	0	0	0,3989	35,5	35
32	31	1	0,443	0,3530	31,25	31
36	23	2	0,984	0,2468	21,85	22
40	15	3	1,476	0,1392	12,32	12
44	5	4	1,967	0,0570	5,05	5
48	1	5	2,454	0,0196	1,74	2
ИТОГО (Σ)	180				179,74	180

Графически настоящее распределение показано на рисунке 6.5

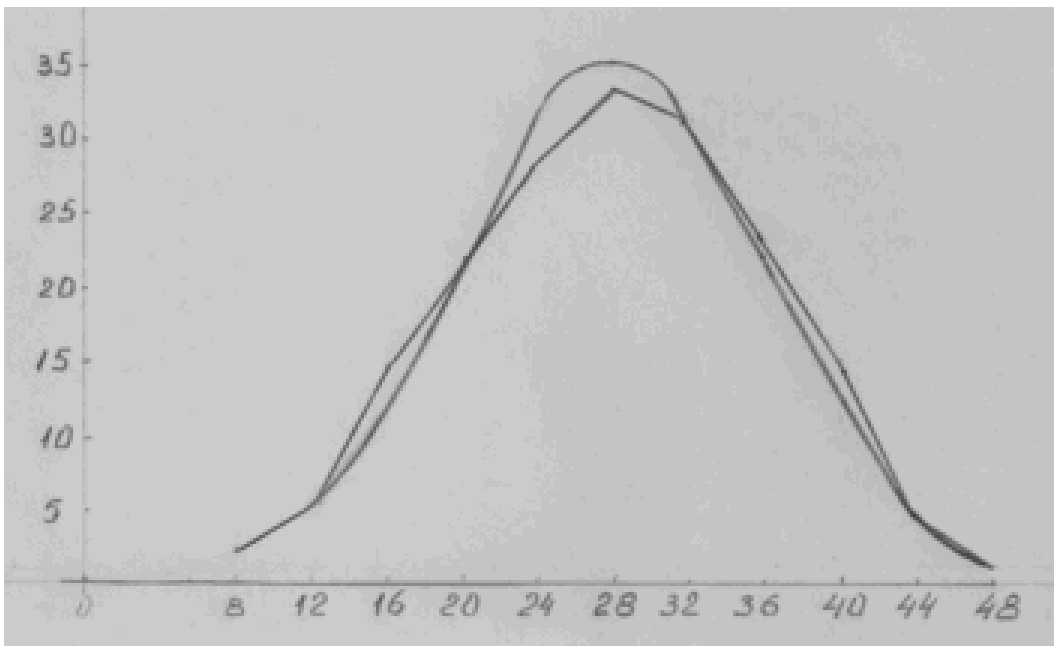


Рисунок 6.5 Экспериментальное и выравненное распределение в древостое ольхи черной

Приведенный в таблице 6.1 ряд распределения характеризуется высокой концентрацией численностей в средних ступенях толщины, о чем свидетельствует значительный отрицательный эксцесс E=-0,51. Ряд незначительно скошен вправо, т.е. справа от среднего значения насчитывается 75 дерева, а справа только 72, о чем свидетельствует небольшая величина $\alpha = -0,07$.

Приведенное распределение характерно для приспевающих и спелых древостоев, которые в незначительной степени затронуты антропогенным влиянием, особенном рубками ухода. Древостои ольхи черной, произрастающие на низинных болотах относятся именно к таким насаждениям.

Для других лесных насаждений характерны скошенные ряды распределения. Биометрические методы изучения их строения, т.е. распределения числа стволов по некоторым таксационным показателям, приводятся ниже.

7. БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ. РАСПРЕДЕЛЕНИЕ ПУАССОНА

7.1 Понятие о биномиальном распределении

7.2 Биномиальное распределение как проявление событий с двумя исходами

7.3 Распределение Пуассона как частный случай биномиального распределения

7.4 Вычисление выравнивающих частот биномиального и пуассоновского распределений

7.1 Понятие о биномальном распределении

Нормальное распределение широко распространено в природе. Но вариационные ряды распределения различных биологических и лесоводственных объектов весьма разнообразны и не могут быть описаны только нормальным распределением. Хотя большинству вариационных рядов свойственно большое количество вариант в середине ряда и уменьшение их по мере удаления от центра, кривые, описывающие это явление, могут быть различными.

Одним из достаточно распространенных видов подобных распределений является биномиальное. Для примера сделаем анализ появления всходов на лесосеке. Допустим, имеется вырубка шириной 100 м, где было 50% сосны и 50% березы, что в лесоводстве записывается в виде формулы состава 5С5Б. Стены леса, окружающие лесосеку имеют такой же породный состав. По вырубке провели плужные борозды, и следующий год появился самосев сосны и березы. Нас интересует какова доля сосны в общем количестве естественного возобновления. Для этого заложили 150 учетных площадок размером 1×1 м, на которых учли все всходы сосны и березы. На каждой площадке выразим долю сосны в процентах. Для удобства долю сосны на площади округлим до 10%. Результаты показаны в таблице 7.1.

Распределение вероятностей, показанное в таблице 7.1, описывается биномиальной кривой. Дадим пояснение этому термину. Если вероятности появления самосева сосны на учетных площадках выразить графически, то получим вариационную кривую или полигон распределения вероятностей (рисунок 7.1)

Кривая на рисунке 7.1. носит название биномиальной, так как ее численности соответствуют разложению бинома

$$(a+b)^n \qquad (7.1)$$

Таблица 7.1 – Вероятность появления сосны на учетных площадках

Доля всходов сосны в % от обычного количества самосева на учетной площадке (x_i)		0	10	20	30	40	50	60	70	80	90	100	Итого
Количество площадок фактическое (численности) - n_i		1	2	7	17	32	38	30	16	6	1	0	150
Выравненное (теоретическое) количество площадок	без округления	0,15	1,46	6,59	17,58	30,76	36,55	30,76	17,58	6,59	1,46	0,15	150
	с округлением	0	1	7	18	31	36	31	18	7	1	0	150
Вероятности	фактические (экспериментальные)	0,007	0,013	0,047	0,113	0,213	0,253	0,200	0,107	0,040	0,007	0	1,0
	теоретические без округления числа площадок	0,001	0,010	0,044	0,118	0,205	0,244	0,205	0,118	0,044	0,010	0,001	1,0
	теоретические с округлением	0	0,007	0,047	0,120	0,207	0,240	0,207	0,120	0,046	0,007	0	1,0

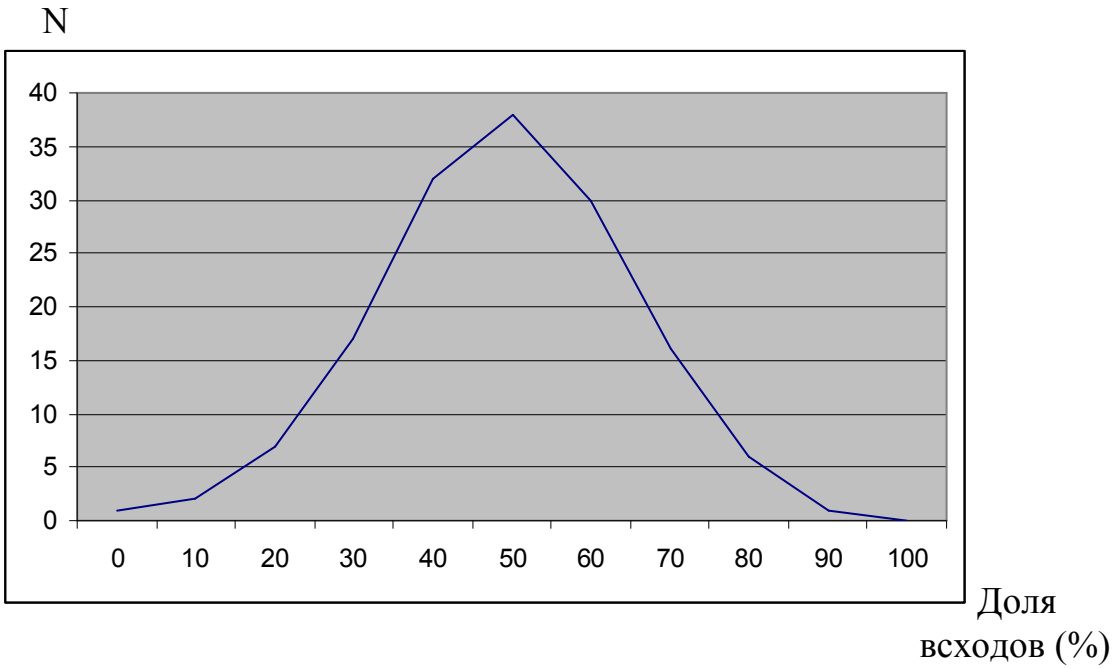


Рисунок 7.1 Полигон распределения вероятностей появления всходов сосны на учетных площадках

Из таблицы 7.1. видно, что для того, чтобы получить вероятностные численности разных результатов при некотором их количестве *N*, надо вероятности *x* умножить на *N*. Сумма вероятностей всегда равна 1. В нашем примере *N* – это общее число площадок, т. е. 150. Так как количество вариантов у нас 11, то мы имеем дело с биномом $(a+b)^{11}$. Теоретическое и практическое разложение этого выражения излагается ниже.

7.2 Биномиальное разложение как проявление событий с двумя входами

Биномиальное распределение обычно применяют, когда необходимо провести исследование событий, которые могут наступить или не наступить.

Сущность биномиальной кривой поясним на следующем примере. Известно, что количество мужчин и женщин примерно одинаково. Здесь мы не рассматриваем аномальные местности, например, в российском городе Иваново, где большинство составляют женщины (вспомним слова известной песни: «….населенье таково – незамужние ткачихи составляют большинство»), а в военных городках – мужчины. Возьмем средний белорусский город, скажем Гомель, где соотношение полов можно считать равным 1:1.

Допустим, мы стоим на улице и считаем проходящих прохожих, подразделяя их по полу. Каждые прошедшие два человека объединим в пары. Эти пары могут иметь следующие варианты: МЖ, ММ, ЖЖ, ЖМ. Вероятность появления мужчины обозначим буквой *a*, а женщины – *b*. Вероятность прохождения мужчин и женщин одинакова, т. е. $a = b = ½$. Вероятность появления один за одним двух мужчин или двух женщин в соответствии с теорией

вероятности равна $a \times a = a^2$ или $b \times b = b^2$. В нашем случае она равна $0{,}5^2 = 0{,}25$, т.е. это один случай из 4.

Сочетание появления друг за другом мужчины и женщины равна ab + ab = 2ab. Таким образом, рассматривая вероятность появления двух равновероятных событий, получаем их следующее распределение.

$$(a+b)^2 = a^2 + 2ab + b^2 \qquad (7.2)$$

Рассматривая 3,4….n сочетаний различных равновероятных случаев, приходим к выводу, что их вероятности описываются вышеприведенной формулой (7.1), т. е. биномом Ньютона. Отсюда распределение получило название биномиального.

Учитывая, что сегодня в средней школе не изучают бином Ньютона, дадим его подробное описание.

Последний выражается формулой

$$(a+b)^n = a^n C_n^1 a_b^{n-1} + C_n^2 a^{n-2} b^2 + \ldots + C_n^k a^{n-k} b^k + \ldots + C_n^{n-1} a b^{n-1} + C_n^n b^n, \qquad (7.1)$$

где C_n^k - число сочетаний из n элементов по k.

$$C_n^k = \frac{A_n^k}{P_k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{1*2*3\ldots k} = C_n^{n-k} \qquad (7.2)$$

Например, при n=8 и k=5 уравнение (7.2) будет иметь вид

$$C_8^5 = \frac{A_8^5}{P_5} = \frac{8*7*6*5*4}{1*2*3*4*5} = 56$$

$$C_8^5 = C_8^3 = \frac{8*7*6}{1*2*3} = 56$$

После подстановки выражения (7.2) (7.1) получаем рабочую формулу для вычисления разложения бинома Ньютона

$$(a+b)^n = a^n + n a^{n-1} b + \frac{n(n-1)}{1*2} a^{n-2} b^2 + \ldots + \frac{n(n-1)\ldots(n-k+1)}{1*2*\ldots*k} * a^{n-k} b^k \ldots + n a b^{n-1} + b^n.$$

Например

$$(a+b)^5 = a^5 + C_5^1 a^4 b + C_5^2 a^3 b^2 + C_5^3 a^2 b^3 + C_5^4 a b^4 + C_5^5 b^5 =$$

$$= a^5 + 5a^4 b + 10 a^3 b^2 + 10 a^2 b^3 + 5 a b^4 + b^5$$

Коэффициенты разложения бинома Ньютона можно получить с помощью треугольника Паскаля. В нем величины любого ряда являются суммой двух цифр, расположенных выше искомого ряда, что видно на рисунке 7.2.

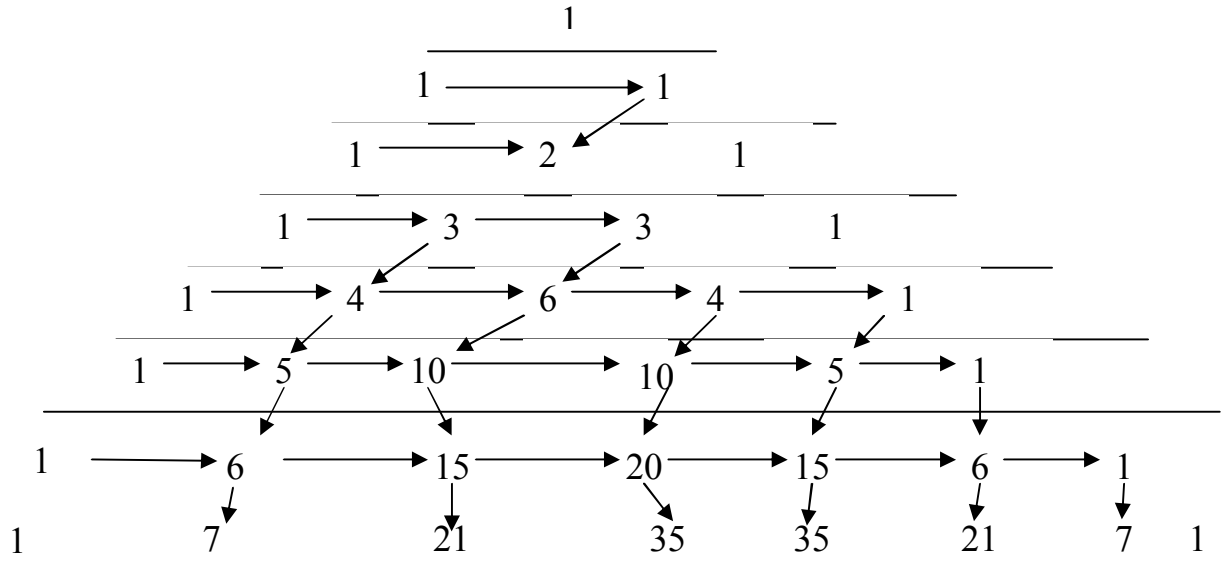


Рисунок 7.2 Треугольник Паскаля

Из школьного курса студентам известны квадрат и куб суммы двух чисел, что является частным случаем общего разложения бинома Ньютона.

Далее разложение продолжается в следующем порядке

$$(a+b)^1 = a+b$$

$$(a+b)^2 = a^2 + 2ab + b^2$$

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

$$(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$$

и т. д.

Кривая на рисунке 7.1. носит название «биномиального распределения», так как соответствует разложению бинома Ньютона. В рассмотренном вариационном ряду, как и при нормальном распределении, которое описано в главе 6, большинство вариант размещаются вблизи центральной части полигона вероятностей. Биномиальное распределение, как сказано выше, используется для математического описания событий, где имеется вероятность противоположных событий a и b.

Рассматривая это распределение с общих позиций теории вероятностей в предположении, что очевидно вероятность появления и отсутствия некоторого события равна 1,0, т. е. a+b=1. Эта простая схема называется схемой Бернулли. Она описана известным швейцарским математиком Я. Бернулли (1654 – 1705). Опубликована была лишь в 1713 г. Из схемы Бернулли вытекает следующая одноименная теорема: « В условиях схемы Бернулли вероятность $a_{n,m}$ того, что событие F появится во всех n испытаниях равно m раз, равна коэффициенту при t^m в разложении бинома $(at+b)^n$ по степени t

$$a_{n,m} = C_n^m a^m b^{n-m} = \frac{n!}{m!(n-m)!} a^m b^{n-m}$$, где

a=1-b; C_n^m - число сочетаний из n элементов по m; ! – знак факториала, т. е. a,b,c,…..,n! = a*b*c*….*n! например 5! = 1*2*3*4*5 = 120

Теорема верна для m=0 или m=n.

Доказательство этой теоремы приводится в книге А.К. Митроольского, имеющейся в списке литературы. Оно заключается в следующем.

Если m = 0 (m = n), то $p_{n,0}$ ($p_{n,n}$) есть вероятность того, что событие E обязательно не появится (появится) в каждом из n испытаний. Из независимости испытаний эта вероятность по закону сложения вероятностей оказывается равной q^n (p^n), что совпадает с коэффициентом при $t^0 = 1$ (t^n) в разложении бинома $(pt + q)^n$ по степеням t. Отсюда, в частности, следует справедливость теоремы для случая $n == 1$, ибо тогда m может равняться или нулю, или единице, т. е. n.

Далее применим метод математической индукции. Пусть теорема доказана для схемы $n - 1$ испытаний. Докажем ее для схемы n испытаний. При этом на основании предыдущего абзаца достаточно считать, что $0<m<n$.

Пронумеруем испытания числами от 1 до n. Об испытании, получившем номер k,, будем говорить как о k-м испытании. Теперь заметим, что появление E точно m раз в n испытаниях равносильно наступлению одного из двух несовместимых событий: либо E появится точно m — 1 раз в первых n — 1 испытаниях и появится также в n-м испытании — вероятность этого события будет $p \cdot p_{n-1,m-1}$ вследствие независимости испытаний, либо E появится точно m раз в и — 1 первых испытаниях и не появится в n-м испытании — вероятность этого события будет $q \cdot p_{n-1,m}$ по той же независимости испытаний. Несовместимость событий приводит к соотношению:

$$p_{n,m} = p \cdot p_{n-1,m-1} + q \cdot p_{n-1,m}.$$

Согласно же предположению о справедливости теоремы для схемы из n—1 испытаний,

$$p_{n-1,m-1} = \frac{(n-1)!}{(m-1)!(n-m)!} p^{m-1} q^{n-m},$$

$$p_{n-1,m} = \frac{(n-1)!}{m!(n-m-1)!} p^m q^{n-m-1}.$$

Подставив эти выражения в предыдущее соотношение, получим

$$p_{n,m} = \frac{n!}{m!(n-m)!} p^m q^{n-m} \left(\frac{m}{n} + \frac{n-m}{n} \right) = \frac{n!}{m!(n-m)!} p^m q^{n-m},$$

что и доказывает теорему.

Совокупность чисел $p_{n,m}$ ($m = 0, 1, ..., n$) образует распределение вероятностей случайной величины X — числа появлений события E во всех n испытаниях схемы Бернулли. Распределение величины X называется биномиальным распределением. Основанием для такого названия служит доказанная теорема, т. е. тот факт, что для определения вероятности $p_{n,m}$ надо разложить бином $(pt+q)^n$ по степеням t и взять коэффициент при t^m.

Выше мы рассмотрим случай равновероятных событий, (случайная встреча мужчины или женщины равновероятны) т. е. когда a=b=0,5. Но при проведении биометрических исследований так случается далеко не всегда. Поэтому надо сделать анализ и других вариантов. При биномиальном распределении $(a + b)^k$, (здесь a и b вероятности проявления (непроявления) некоторого события или признака, k – число классов) возможны различные значения a и b, например: a = 0,6 и b = 0,4 или a = 0,2 и b = 0,8 и т.д. При этом меняется и форма полигона распределения. По мере увеличения различий между a и b полигон становится все более скошенным, асимметричным. Однако по мере увеличения N даже при значительном различии между a и b степень симметрии полигона вновь усиливается.

Как и для других распределений, параметрами для биномиального распределения являются средняя арифметическая ($\bar{x}$) и среднее квадратическое отклонение (σ), которые можно определить с помощью приведенных выше формул (глава 4) для любого конкретного эмпирического ряда.

Теоретически их значения определяются значениями вероятностей a и b, а также значением $\sum n_i$, т.е. числа независимых событий, распределение которых изучается.

Средняя арифметическая при биномиальном распределении

$$\bar{\times} = k*a \qquad (7.6)$$

и среднее квадратическое отклонение

$$\sigma = k*a*b \quad \text{, где} \qquad (7.7)$$

k – показатель степени бинома; k = N-1.

Эти формулы дают возможность связать определенные $\bar{\times}$ и σ, вычисленные на основе данного конкретного материала, с вероятностями a и b. Сказанное поясним примером.

Пусть на некотором участке был посеян дуб в площадки по 4 шт. (желудя) на площадке. Через 5 лет провели учет выживших сеянцев. Для этого подсчитали количество деревьев на каждой площадке, а всего в учет включили 100 площадок. Результаты показаны в таблице 7.2.

Таблица 7.2 – Распределение сохранившихся сеянцев дуба на 100 площадках

Количество сохранившихся экземпляров дуба, x_i	Число площадок с сеянцами, n_i	x_i*n_i	$x_i^2*n_i$
0	7	0	0
1	24	24	24
2	37	74	148
3	26	78	234
4	6	24	96
ИТОГО (Σ)	100	200	502

Вычислим $\overline{\times}$ и σ по обычной методике (формулы 4.1 и 5.2)

$$\overline{X} = \frac{\sum x_i n_i}{\sum n_i} = \frac{200}{100} = 2$$

$$\sigma = \sqrt{\frac{\sum x^{\alpha} n_i - \frac{\left(\sum x_i n_i\right)^2}{\sum n_i}}{\sum n_i}} = \sqrt{\frac{502 - 400}{100}} = \sqrt{1.02} = 1.01.$$

Поскольку у нас сохранилось 200 сеянцев из 400, то вероятность учета сохранившихся и отпавших экземпляров одинакова, т. е. 20/400=0,5, т. е. a=b=0,5. Вычислим $\overline{\times}$ и σ по формулам (7.6) и (7.7), где k=4.

$$\overline{X} = k * a = 4 * 0.5 = 2;\ \sigma = \sqrt{k * a * b} = \sqrt{4 * 0.5 * 0.5} = \sqrt{1} = 1.$$

Величины $\overline{\times}$ и σ, вычисленных разными способами, почти одинаковы, но полного совпадения нет. Это произошло потому, что ряд в эксперименте (таблица 7.2) несколько скошен. Если его выровнять и получить теоретические величины, то при наличии строго биномиального распределения значения $\overline{\times}$ и σ, вычисленные разными способами, совпадут. Поскольку при проведении исследований никогда нет уверенности, что экспериментальное распределение строго соответствует теоретическому, то надежнее вычислить $\overline{x}$ и σ непосредственным способом, т. е. по схеме, представленной в таблице 7.2 и в формулах (4.1) и (5.2).

В приведенном примере исходим из того, что a = b = 0,5, т. е. они будут точно установленные теоретические вероятности. Но возможны и другие значения вероятностей a и b. Приведем пример, описанный известным белорусским ученым в области биометрии П.Ф. Рокицким.

Так, например, было получено следующее фактическое распределение самок в 103 пометах с 4 мышками в каждом помете (таблица 7.3).

Таблица 7.3 – Распределение самок в помете мышей

Количество самок	0	1	2	3	4
Число пометов	8	32	34	24	5

Тогда $\overline{\times} = \frac{8 \cdot 0 + 1 \cdot 32 + 2 \cdot 34 + 3 \cdot 24 + 3 \cdot 24 + 4 \cdot 5}{103} = 1{,}864.$

Но так как $\overline{\times} = kp$, а k = 4, то p = 1,864 / 4 = 0,47.

Это вероятность появления самок. Вероятность же появления самцов q=0,53.

Исходя из формулы $\sigma^2 = kpq$, можно вычислить

$$\sigma^2 = 4 \cdot 0{,}47 \cdot 0{,}53 = 1{,}0.$$

Так как данный ряд является рядом разложения бинома $(0{,}54+0{,}47)^4$ при n=103, то легко вычислить, сколько особей следует ожидать в каждом классе. Получатся следующие цифры для частот каждого класса (таблица 7.4).

Таблица 7.4 – Ожидаемое число самок в помете мышей

Количество самок	0	1	2	3	4
Ожидаемое число пометов	8	29	38	23	5

Уже на глаз видно большое совпадение фактически полученных величин с ожидаемыми.

В общем виде биномиальное распределение описывается формулой Бернулли. Ее доказательство приведено выше, а саму формулу повторим еще раз.

$$P_{m,n} = C_n^m P^m (1-P)^{n-m} = \frac{n!}{m!(n-m)!} P^m (1-P)^{n-m} \qquad (7.4)$$

Здесь P - вероятность наступления события при проведении n независимых испытаний; m - число наступивших событий; C_n^m - число сочетаний из n элементов по m.

Формула Бернулли имеет очень важное значение в теории вероятностей, так как она связана с повторением испытаний в одинаковых условиях, где как раз и проявляются законы теории вероятностей.

Обобщая изложенное, приходим к следующему выводу. Альтернативные (т.е. противоположные), дискретно варьирующие признаки распределяются так, что вероятные численности их появления могут быть найдены по формуле бинома Ньютона.

$$N(a+b)^n = N\left(a^n + na^{n-1}b + \frac{n(n-1)}{1\cdot 2}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{1\cdot 2\cdot 3}a^{n-3}b^3 + \cdots b^n\right), \qquad (7.5)$$

где n - число независимых исходов в одном испытании;

a - вероятность благоприятного исхода одного случая;

b - вероятность неблагоприятного исхода;

N - общее число испытаний (исходов);

Так, при n=5 возможны 2^5=32 исхода. При равной вероятности альтернатив, т.е. a=b=0,5 по формуле (7.5) получим следующие числовые величины

$$32(0{,}5+0{,}5)^2 = 32\left(0{,}5^5 + 5*0{,}5^4*0{,}5 + \frac{5*4*0{,}5^3 0{,}5^2}{1*2} + \frac{5*4*3*0{,}5^2*0{,}5^3}{1*2*3} + \frac{5*4*3*2*0{,}5*0{,}5^4}{1*2*3*4} + 0{,}5^5\right) =$$

$$= 32 = 32*(0{,}03125 + 5*0{,}03125 + 10*0{,}03125 + 10*0{,}03125 + 5*0{,}3125 + 0{,}3125) = 32*(0{,}03125 + 0{,}15625 +$$

$$+ 0{,}3125 + 0{,}3125 + 0{,}03125) = 32*(0{,}03125*2 + 2*0{,}15625 + 2*0{,}3125) = 32*(0{,}06250 + 0{,}31250 + 0{,}62500) =$$

$$= 32*1 = 32$$

Откладывая значения числа благоприятных исходов “m” по оси абсцисс, а по оси ординат вероятные численности “n”, получим многоугольник численностей распределения. Ломаная линия, соединяющая точки на графике, называется **кривой распределения**.

Как сказано выше, вероятность события у, которое проявляется “m“ раз в “n” независимых испытаниях описывается формулой Бернулли. Биномиальное распределение определяется двумя параметрами: средней величиной μ=N*a и дисперсией σ^2=nab или среднеквадратическим отклонением $\sigma=\sqrt{kab}$. У нас $\bar{\sigma} = N * a = 5 * 0.5 = 2.5$; σ^2=5 · 0,5· 0,5=1,25.

7.3 Распределение Пуассона как частный случай биномиального распределения

В лесохозяйственной науке и практике часто встречаются случаи, когда из двух (или более) наблюдаемых явлений одно встречается редко. Например, естественное возобновление на лесосеке, где вырубленный древостой имеет состав 6Б3Ос1С (60% березы, 30% осины и 10% сосны), идет в основном за счет самосева: семенного для С, Б, Ос и одновременно порослевого у березы и осины. Деревца сосны здесь редки, а к 4-5 годам, если не вести ухода, их практически полностью заглушат береза и осина. Поэтому вероятность встретить через 6-7 лет на такой лесосеке деревья сосны мала.

В условиях радиоактивного загрязнения больших территорий после Чернобыльской катастрофы увеличилось количество разного рода мутаций среди молодых деревьев. Эти мутации выражаются в изменении формы хвои, искривления побегов и т. д. Но их частота (во всех зонах радиоактивного загрязнения в пределах Чернобыльского следа) хотя и зависит от уровня радиоактивности, но по абсолютному значению мала и относится к редким явлениям.

Подобные примеры редких явлений часто встречаются в физике, биологии и других науках. Для описания названных и других подобных распределений обычно применяют распределение Пуассона.

Распределение Пуассона. *Распределение Пуассона*, или *пуассоново распределение*, подобно биномиальному, относится к дискретной или прерывистой изменчивости. Оно имеет самостоятельное значение, хотя его можно рассматривать и как предельный случай биномиального. При биномиальном распределении значения *a* и b могут быть близки друг к другу, при пуассоновом же *a* очень мало, т.е. события осуществляются очень редко, а b приближается к единице.

Распределение отдельных редких наблюдений является при этом чаще всего асимметричным, но симметрия возрастает с увеличением $\overline{\times}$.

При увеличении a распределение приближается к биномиальному. Пуассоново распределение характеризуется в сущности только одним параметром - средней арифметической $\overline{\times}$, так как σ^2 в этом случае обычно равна $\overline{\times}$ или близка ей по значению. Именно по этому равенству $\overline{\times}$ и σ^2 легче всего определить, что данное распределение является пуассоновым.

Средняя арифметическая для пуассонова распределения равна *na*, где *a* - вероятность обнаружения данного признака, а n - количество фактически проведенных наблюдений. Вспомним, что величина a может быть очень малой.

$$\overline{\times} = \lambda = na = \sigma^2 \qquad (7.6)$$

В формуле (7.6) среднее значение показано как $\overline{X}$ и λ, так как большинство исследователей и авторов учебников в распределении Пуассона $\overline{X}$ обозначают как λ.

Частоты распределения Пуассона представляют собой следующий ряд:

$$\frac{n}{e^{\lambda}} \text{ (нулевой член)}; \quad \frac{n\lambda}{e^{\lambda}}; \frac{n\lambda^2}{2\,e^{\lambda}}; \frac{n\lambda^3}{(2)\ (3)e^{\lambda}}; \frac{n\lambda^4}{(2)\ (3)\ (4)\,e^{\lambda}}; \text{ и т.д.} \qquad (7.7)$$

Здесь n - общее число вариант, e - основание натуральных логарифмов (2,71828…) и λ - средняя арифметическая.

Конкретные пуассоновы ряды являются конечными в силу ограниченности количества наблюдений. Но теоретически они могут продолжаться до бесконечности.

Таким образом, в общем виде распределение Пуассона описывается формулой:

$$y = C_n^m p^m q^{n-m} = \frac{n(n-1)...(n-m+1)}{m!} * \frac{\lambda}{n^m}\left(1-\frac{\lambda}{n}\right)^{n-m} \qquad (7.8)$$

где $\lambda = np$; $p = \lambda / n$.

Так как числитель первой дроби имеет m сомножителей, а в знаменателе стоит n^m каждый из сомножителей можно разделить на n. Получим:

$$y = \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{m-1}{n}\right) * \frac{\lambda^m}{m}\left(1-\frac{\lambda}{n}\right)^{n-m} \qquad (7.9)$$

При $n \to \infty$ предел любой дроби $(1-\lambda/n)$ есть 1, а предел $(1-\lambda/n)^{n-m} = e^{-\lambda}$.

При этих условиях $y = \left(\lambda^m / m!\right)e^{-\lambda}$. (7.10)

Выражение (7.10) называется функцией распределения вероятностей в распределении Пуассона. В этом выражении m – частота ожидаемого события в n испытаниях, e=2,7183; параметр $\lambda = np$ соответствует математическому ожиданию или наивероятнейшей частоте события, т. е. μ, а также дисперсии σ^2. Доказательство этого равенства здесь опускаем. Оно содержится во многих книгах по статистике.

7.4 Вычисление выравнивающих частот биномиального и пуассоновского распределений

Вычисления выравнивающих (теоретических) частот биномиального распределения проводят следующим образом.

- находят величину k, т. е. степень бинома: k=N-1.
- проводят разложение бинома по формуле Ньютона (формула 7.2). Здесь удобнее воспользоваться вышеприведенным треугольником Паскаля.
- по разложению бинома Ньютона вычисляем (i) вероятности для каждого значения x_i: i=N*a*b, где N – объем ряда распределения ($\sum x_i n_i$).
- численности ($\tilde{n}_i$) находим, умножая найденные вероятности для каждого x_i на N, т. е. $\tilde{n} = \sum x_i * n_i * a * b$.

Рассмотрим вычисление выравнивающих частот на примере (таблица 7.3), взяв исходные данные из таблицы 7.2.

Таблица 7.3 – Вычисление выравнивающих частот биноминального распределения для сохранившихся 5-летних экземпляров в посеве дуба

Количество сохранившихся экземпляров дуба, x_i	Число площадок с сеянцами в эксперименте, n_i	Вероятности a*b*k	$\tilde{n} = \Sigma n_i * a * b$		$X_i * \tilde{n}_i$	$x_i^k * \tilde{n}_i$
			без округления	с округлением		
0	7	0.0625	6,25	6	0	0
1	24	0.25	25,0	25	25	25
2	37	0.375	37,5	38	35	152
3	26	0.25	25,0	25	75	225
4	6	0.0625	6,25	6	24	96
Σ	100	1.0	100	100	200	498

Разложение $(a+b)^4$ описывается следующим уравнением:

$$a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4 \qquad (7.17)$$

$$0.5^4 + 4*(0.5)^3 * 0.5 + 6*(0.5)^2(0.5)^2 + 4*0.5*(0.5)^3 + 0.5^4 =$$

$$= 0.0625 + 0.25 + 0.375 + 0.25 + 0.0625 = 1$$

Численности ($\tilde{n}_i$) будут равны произведению вероятностей из (7.17) на $N(\sum \tilde{n}_i)$. Из таблицы 7.3 видим, что

$$\overline{X} = \frac{200}{100} = 2\,;\ \sigma = \sqrt{\frac{498 - \frac{200^2}{100}}{100}} = \sqrt{0.98} = 0.99 \approx 1.0$$

По формулам (7.6) и (7.7)

$$\overline{X} = 4 * 0.5 = 2.0\,;\ \sigma = 4 * 0.5 * 0.5 = 1.0$$

Приведенные расчеты показывают правомерность использования биноминального распределения.

Для вычисления выравнивающих частот распределения Пуассона используют его свойства, т.е.

$$\overline{X} = \sigma = \mu_3 = \lambda\,.$$

Учитывая, что это распределение редких событий по схеме Бернулли (вероятность 0,1) т. е. это вероятность того, что событие наступит 0,1,…,k раз в серии из N испытаний. При $\tilde{n} \to \infty$ вероятность $p \to 0$, то $np = const = \lambda$, т.е. единственному параметру λ распределения Пуассона. Это дает возможность вычислять вероятность p_m распределения Пуассона по составленной таблице значений функции $p_m = \frac{\lambda^m}{m!} * e^{-\lambda}$ при разных значениях λ, которая приведена в приложении В. Эта таблица дана для значений λ от 0 до 20.

Для больших величин λ соответствующая таблица приведена в книге А.К. Митропольского В силу того, что в лесном хозяйстве большие величины λ встречаются редко, мы эту таблицу опускаем.

Приведем пример вычисления выравнивающих частот при распределении Пуассона, который описан К. Е. Никитиным и А. З. Швиденко (таблица 7.4).

Вычислим выравнивающие частоты для ряда распределения числа деревьев на пробных площадях размером 0,002 га. Статистики для этого ряда $\overline{X} = 1{,}508,\ \hat{\mu}_2 = 1{,}5189,\ \hat{\mu}_3 = 1{,}5434$, что предполагает близость эмпирического распределения закону Пуассона. Такой же вывод следует и из логики рассматриваемого явления. Для вычисления выравнивающих частот вероятности *f(k)*, полученные по формуле

$$f(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} = \frac{(np)^k \cdot e^{-np}}{k!}\ , \qquad (7.18)$$

следует умножить на объем ряда распределения *n*=224. В (7.18) положим *a*=1,508. Заметим, что 0!=1, а lg *k*! вычисляют непосредственно или по таблицам логарифмов факториалов. Для вычисления по схеме таблицы 7.4 формулу (7.18) предварительно логарифмируют, т.е. $\lg f(k) = k \lg a - a \lg e - \lg k! = 0{,}0178k - 0{,}6549 - \lg k!$. В целом модель хорошо отражает эмпирический ряд.

Следующие вычисления численностей по распределению Пуассона покажем на примере учета всходов сосны на вырубке. Для этого заложили учетные площадки площадью по 10 м^2 (5×2).

Таблица 7.4 – схема вычисления выравнивающих частот распределения Пуассона (на примере ряда распределения числа деревьев на пробных площадях)

k_i	n_i	0,178 k_i	$k_i!$	-lg $k_i!$	lg $f(k)$= -0,6549 + (3)+(5)	$f(k)$	$\tilde{n}_i$
1	2	3	4	5	6	7	8
0	50	0	1	0	$\bar{1}$,345	0,221	50
1	74	0,1780	1	0	$\bar{1}$,523	0,333	75
2	57	0,3562	2	-0,3010	$\bar{1}$,400	0,251	56
3	27	0,5343	6	-0,7782	$\bar{1}$,101	0,126	28
4	12	0,7123	24	-1,3802	$\bar{2}$,577	0,048	11
5	3	0,8905	120	-2,0792	$\bar{2}$,156	0,014	3
6	1	1,0686	720	-2,8573	$\bar{3}$,557	0,004	1
∑	224	-	-	-	-	-	224

Всего заложено 8 площадок. Результаты расчетов показаны в таблице 7.5. При этом для расчетов использованы следующие формулы.

$$\tilde{n} = \left[\frac{\bar{X}^{-m}}{m!\, e^{-\bar{X}}} \right] N$$

$$y = \left(\frac{\lambda^m}{m!} e^{-\lambda} \right), \text{ где}$$

$\bar{X}, \lambda$ - среднее значение;

N - объем ряда распределения;

m – варианты;

e – 2,71

Таблица 7. 5 – Вычисление выравнивающих частот по кривой Пуассона

m	n_i	mn_i	$y(m)$	$\tilde{n}$ без округ.	$\tilde{n}$ округ.	n_i %	$\tilde{n}$ %	$\bar{X} - n_i$	$(\bar{X} - m_i)^2$	$(\bar{X} + m_i)^2 \cdot n$	$(\bar{X} + m_k)^{2N} \cdot n_i$
1	2	3	4	5	6	7	8	9	10	11	12
0	549	0	0,3329	384	385	47,5	33,3	1,1	1,21	664,2	465,9
1	271	271	0,3662	422,96	423	23,5	36,6	0,1	0,01	2,7	4,2
2	137	274	0,2014	232,62	233	11,9	20,2	-0,9	0,81	111,0	188,7
3	110	330	0,0738	85,24	25	9,5	7,4	-1,9	3,61	397,1	306,8
4	56	224	0,0203	23,45	23	4,8	2,0	-2,4	8,41	470,0	193,5
5	22	110	0,0045	5,20	5	1,9	0,4	-3,9	15,27	335,9	76,4
6	9	54	0,0008	0,92	1	0,8	0,1	-4,9	24,01	216,1	24,0
7	1	7	0,0001	0,11	0	0,1	0	-5,9	34,81	34,8	0
Σ	1155	1270	1,0000	1158	1155	100,0	100	-	88,14	2231,8	1259,6

Значения функции Пуассона со средним значением ($\overline{X}$ или λ) и показателем степени *m*, как показано выше, выписывается из специальной таблицы (приложение В).

Графы 9–12 введены в таблице 7.5 для вычисления σ. Ее близость к $\overline{X}(\lambda)$ показывает, что распределение описывается уравнением Пуассона. Тогда для эмпирического распределения

$$\sigma_1 = \sqrt{\frac{(\overline{X} - m_i)^2 n_i}{N}} = \sqrt{\frac{2232}{1155}} = \sqrt{1{,}93} = 1{,}39$$

Для теоретического распределения

$$\sigma_2 = \sqrt{\frac{1260}{1155}} = \sqrt{1{,}10} = 1{,}05$$

Близость σ_1 и σ_2 к $\overline{X}$ показывает правомерность использования для выравнивания распределения Пуассона.

Для практических расчетов, когда находят теоретические ординаты распределения $\tilde{n}$, т. е. численности распределения случайного события X, выражение (7.10) умножают на N – общее число наблюдений, вместо n_i принимают экспериментальное среднее число наблюдаемых случаев. Формула для n будет:

$$\tilde{n} = \left[(\overline{x}^m / m!)e^{-\overline{x}}\right]N$$

Распределение Пуассона с возрастанием средней λ приближается к биномиальному.

Распределение Пуассона описывает многие явления в технике и биологии. В технике оно находит широкое применение при контроле качества продукции, для аппроксимации распределения дефектных изделий. В лесном хозяйстве его применяют как модель распределения числа примесей в пробных навесках при анализе семян, допустим, сосны, при рассмотрении плодов, скажем, желудей, поврежденных вредителем. Им же описывают распределение численности возобновления, когда размер элементарных учетных площадок очень мал или условия заселения площади неблагоприятны, так что вероятность благоприятного исхода *p* мала.

В лесном хозяйстве к распределению Пуассона прибегают, когда исследуют мутации при проведении генетико-селекционных исследований, при анализе появления альбиносов среди лесных зверей и птиц, для оценки выживаемости самосева сосны под пологом леса и т.д., при анализе других редких событий.

8. ДРУГИЕ РАСПРЕДЕЛЕНИЯ. СИСТЕМА КРИВЫХ ПИРСОНА И ДЖОНСОНА

8.1 Распределение типа А или Грама-Шарлье

Система кривых распределения далеко не исчерпывается нормальным и биномиальным распределениями. Видов распределений много, и некоторые, наиболее часто встречающиеся, мы здесь рассмотрим.

Еще в XIX веке немецкими лесоводами и таксаторами (Вейзе и др.), а затем и в России (А.В. Тюрин (1882-1979) и др.) было доказано, что распределение числа деревьев в древостое по таксационным показателям соответствует закону нормального распределения. Но более глубокие исследования, проведенные во второй половине XX века, показали, что это не совсем так. Реальная кривая похожа на кривую нормального распределения, но чаще всего бывает скошенной вправо или влево, более плоской или более остроконечной. Оказалось, что кривой Гаусса-Лапласа соответствуют, да и то не всегда, древостои естественного происхождения, растущие без вмешательства человека и имеющие возраст, близкий к спелому (для сосны это 80 лет и старше). В большинстве случаев кривая, которой хорошо моделируются распределения деревьев в лесах Беларуси, лишь близка к нормальному распределению. Называется она кривой обобщенного нормального распределения или кривая типа А. Ее еще называют кривой Грама-Шарлье, иногда просто кривой Шарлье.

В том случае, когда статистики распределения α и E или один из них оказались значимыми, или когда при интервальной оценке параметров α и E интервал перекрывает нуль, выравнивающие частоты бывает целесообразно рассчитать именно по уравнению обобщенного нормального распределения. Это распределение является разложением в ряд уравнения кривой нормального распределения. Оно учитывает имеющиеся асимметрию и эксцесс. Выравнивание по этому уравнению, как показали многочисленные исследования, дает лучшую аппроксимацию экспериментального ряда числа стволов по диаметру и высоте в антропогенных лесах, чем нормальное распределение. Последнему отдают предпочтение лишь в тех случаях, когда $\alpha < t_{0,05}$ и $E < t_{0,05}$. Исходят при этом из положения, что при недоказанном отклонении распределения от нормального, надежнее считать это распределение следующим модели нормального распределения, а найденные показатели α и E относить за счет случайного состава выборочной совокупности.

Теоретические частоты кривой типа А вычисляют по формуле

$$\tilde{n} = \left[f(x) - \frac{r_3}{6} f^3(x) + \frac{r_4 - 3}{24} f^4(x) - \frac{r_5 - 10r_3}{120} f^5(x) + \frac{r_6 - 15r_4 + 30}{720} f^6(x) - \cdots \right] \frac{N}{\sigma'} \quad (8.1)$$

где $f(x)$ - значения функции нормальной кривой при $N/\sigma' = 1$, $f^3(\chi)$, $f^4(x)$, $f^5(x)$, $f^6(x)$ - 3, 4, 5, 6-я производные функции $f(x)$; r_3 (асимметрия), r_4 (r_4-3 – эксцесс), r_5, r_6… – основные моменты; N – общая численность вариант ряда; σ^2 - основное отклонение в единицах интервала, $\sigma^2 = \mu_2$.

Из формулы 8.1 мы видим, что распределение Грамма-Шарлье описывается бесконечным рядом, использующим функцию нормального распределения и ее производные. В практике достаточно ограничиться тремя первыми членами разложения. Увеличение числа членов обычно не улучшает (может и ухудшить) аппроксимацию эмпирического распределения. Применение распределения Грамма-Шарлье требует знания четырех параметров: $\bar{x}$, σ, α, E. Вспомним, что для вычисления кривой нормального распределения достаточно двух параметров ($\bar{X}$, σ), а биномиального и пуассоновского – одного, λ.

При выборе в качестве аппроксимирующей кривой уравнения типа А следует помнить, что близость этой кривой к нормальному распределению накладывает достаточно жесткие ограничения на величины α и E. Они не должны быть слишком велики, например, α не должна превышать величины 0,5 – 0,7, а E – 0,6 – 0,8.

Приведем пример расчета кривой типа А. Для этого возьмем ряд распределения числа стволов по диаметру на высоте 1,3 м в сосновом насаждении I^a класса бонитета в возрасте 50 лет, где проведены рубка ухода и древостой имеет полноту 0,7. Пользуясь формулами, приведенными выше, предварительно сделаем расчеты необходимых параметров кривой (таблица 8.1).

Для вычисления частот кривой типа А требуется знать $\bar{x}$, σ, α, E.

Таблица 8.1 – Вычисление показателей для определения статистик ряда распределения числа стволов по диаметру в древостое сосны I класса бонитета в возрасте 50 лет

Ступени толщины x_i	Число стволов n_i	x_k	$x_k n_i$	$x^2_k n_i$	$x^3_k n_i$	$x^4_k n_i$	x_{k+1}	$(x_{k+1})^4 n_i$	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}^2) n_i$
8	2	-4	-8	32	-128	512	-3	162	16	-17	289	578
12	10	-3	-30	90	-270	810	-2	160	120	-13	169	1690
16	22	-2	-44	88	-176	352	-1	22	352	-9	81	1782
20	43	-1	-43	43	-43	143	0	0	860	-5	25	1075
24(M′)	39	0	0	0	0	0	1	39	936	-1	1	39
28	32	1	32	32	32	32	2	512	896	3	9	288
32	24	2	48	96	192	384	3	1944	768	7	49	1176
36	18	3	54	162	486	1458	4	4608	648	11	121	2178
40	9	4	36	144	576	2304	5	5625	360	15	215	2025
44	1	5	5	25	125	625	6	1296	44	19	136	361
Итого (Σ)	200	-	50	712	794	6520	-	14368	5000	-	-	1192

$$\bar{x} = \frac{\sum n_j x_j}{\sum n_j} = \frac{5000}{200} = 25;\ \bar{\sigma} = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{\sum n_i}} = \sqrt{\frac{11192}{200}} = \sqrt{55{,}96} = 7{,}481$$

$$m_1 = \frac{\sum x_k n_j}{\sum n_j} = \frac{50}{200} = 0{,}25;\ m_2 = \frac{\sum x_k^2 n_j}{\sum n_j} = \frac{712}{200} = 3{,}55.$$

$$m_3 = \frac{\sum x_k^3 n_j}{\sum n_j} = \frac{794}{200} = 3{,}97;\ m_4 = \frac{\sum x_k^4 n_j}{\sum n_j} = \frac{6520}{200} = 32{,}6.$$

Проверка: $\frac{\sum (x_{k+1})^4 n_j}{\sum n_j} = m_4 + 4m_3 + 6m_2 + 4m_1 + 1{,}0 = \frac{14368}{20} =$
$32{,}6 - 4 \cdot 3{,}97 + 6 \cdot 3{,}55 + 4 \cdot 0{,}25 + 1{,}0;$
71,84 = 32,6+15,88+21,36+2;
71,84 = 71,84.

$\mu_2 = m_2 - m_1^2 = 3{,}56 - 0{,}0625 = 3{,}4975.$

$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3 = 3{,}97 - 3 \cdot 3{,}56 \cdot 0{,}25 + 2 \cdot 0{,}015625 =$
$3{,}97 - 2{,}67 + 0{,}03125 = 1{,}33125.$

$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4 =$
$32{,}6 - 4 \cdot 0{,}25 \cdot 3{,}97 + 6 \cdot 3{,}56 \cdot 0{,}0625 - 6 \cdot 0{,}00390625 =$
$32{,}6 - 3{,}97 + 1{,}335 - 0{,}0234372 = 29{,}941563.$

Проверка: $\mu_3 = m_3 - 3\mu_2 m_1 - m_1^3 = 3{,}97 - 3 \cdot 3{,}4975 \cdot 0{,}25 - 0{,}015625 =$
$3{,}97 - 2{,}623125 - 0{,}015625 = 1{,}33125.$

$\mu_4 = m_4 - 4\mu_3 m_1 - 6\mu_2 m_1^2 - m_1^4 =$
$32{,}6 - 4 \cdot 1{,}33125 \cdot 0{,}25 + 6 \cdot 3{,}4975 \cdot 0{,}0625 - 4 \cdot 0{,}00390625 =$
$32{,}6 - 1{,}33125 - 1{,}3115625 - 0{,}015625 = 29{,}9415.$

$\bar{X} = \mu' + 4 \cdot m_1 = 24 + 4 \cdot 0{,}25 = 25.$

$\sigma = \sqrt{\mu_2} = \sqrt{3{,}4975} = 1{,}87;\quad \bar{\sigma} = 4 \cdot 1{,}87 = 7{,}481$

$r(\alpha) = \frac{\mu_3}{\sigma^3} = \frac{1{,}33125}{6{,}54089} = 0{,}2035 \approx 0{,}203 = \infty$

$r_4 = \frac{\mu_4}{\sigma^4} = \frac{29{,}94159}{12{,}2325} = 2{,}448 \approx 2{,}45$

$E = r_4 - 3 = -0{,}55.$

Проведя вышеприведенные вычисления, получили $\bar{X}$=25,0; $\bar{\sigma}$=7,48; σ=1,87; α=0,20; E = -0,55; m_1=0,25. Значения α и E показывают на значительные отличия ряда распределения от нормального, т. к. требования к «нормальности» α=E=0 (во всяком случае α<0.05; E<0,05), недопустимы для использования кривой типа А. Схема вычисления выравнивающих частот с использованием кривой типа А показана в таблице 8.2.

Таблица 8.2 – Вычисление частот распределения типа А (Грамма-Шарлье) для ряда распределения числа стволов по диаметру в древостое сосны в возрасте 50 лет

Параметры для вычисления

$\bar{x}$=25,0; σ=1,8701; $\bar{\sigma}$=7,4086; α = 0,203; E = -0,55; m_1=0,25

Ступени толщины, x_i	Число стволов	x_k	$x_k - m$	$x=\frac{x_k - m}{\sigma}$	$f(x)$	$f^3(x)$	$f^4(x)$	$-\frac{\alpha}{6} f^3(x) = -0.0338 * f^3(x) = k_1$	$\frac{E}{24} f^4(x) = -0.02298 * f^4(x) = k_2$	$\sum f(x)_1 k_1 k_2 = k_3$	$\frac{\sum n_i}{\sigma} \div k_3 = 106.952 * k_3 = \tilde{n}$	$\tilde{n}$ округление
1	2	3	4	5	6	7	8	9	10	11	12	13
8	2	-4	-4,25	-2,27253	0,0302	+0,14582	-0,3971	-0,00493	-0,00091	0,02336	2,4	2
12	10	-3	-3,25	-1,73782	0,0881	+0,00316	-0,52857	-0,00011	-0,0121	0,07588	8,1	8
16	22	-2	-2,25	-1,20311	0,1935	+0,36143	-0,69435	+0,01223	-0,01591	0,18982	20,3	20
20	43	-1	-1,25	-0,66839	0,3192	-0,54445	0,16653	+0,01842	+0,00382	0,34124	36,5	36
24=m^1	39	0	-0,25	-0,13368	0,3955	-0,15799	1,14360	+0,00534	+0,02621	0,42705	45,6	46
28	32	1	0,75	0,40104	0,3681	+0,41905	0,75870	+0,01418	+0,01739	0,37131	39,7	40
32	24	2	1,75	0,93575	0,2577	0,51175	-0,38329	+0,01297	-0,00785	0,26282	28,1	28
36	18	3	2,75	1,47046	0,1355	0,16670	-0,71696	-0,00564	-0,01643	0,11343	12,1	12
40	9	4	3,75	2,00518	0,0539	+0,00485	0,26498	+0,00016	+0,00607	0,06013	6,4	6
44	1	5	4,75	2,53990	0,0157	-0,13900	0,09370	+0,00470	+0,00245	0,02255	2,4	2
ИТОГО (Σ)	200	-	-			+0,035	0,06	-	-	-	201,6	200

В 1-й столбец таблицы 8.2. вписаны классовые варианты X, в 3-й – их условные значения x_k, выраженные в долях интервала, x_k=(X-M')/k; во 2-м столбце записаны частоты ряда; в 4-м столбце – разности между условными значениями вариант и первым начальным моментом ряда, т. е. значения $x_k - m_1$, в 5-м – значения аргумента x, т. е. частное от разностей ($x_k - m_1$) на основное отклонение не именованное σ. В 6, 7 и 8-м столбцах записывают значения функции *f(τ)* и ее производных. Величину этих функций берут из специальных таблиц, приведенных в приложении Г.

Значение $f^3(x)$ дается для положительных значений x. При отрицательных значениях *x* знак, указанный в таблицах, нужно изменить на обратный. Для *f(x)* и $f^4(x)$ знаки остаются без изменения, т.е. теми же, что и в таблицах, независимо от знака при x.

В 9-й столбец записывают значения $\alpha/6$ $f^3(t)$, в 10-й – значения (E/24) $f^4(\tau)$. Данные 11-го столбца представляют алгебраическую сумму цифр 6, 9 и 10-го столбцов. В 12-м столбце помещены выровненные частоты, полученные путем умножения данных 11-го столбца на N/σ (без округления), а 13-м столбце – округленные величины теоретических частот с округлением. Графически выравнивание распределения числа стволов по диаметру из нашего примера показано на рисунке 8.1.

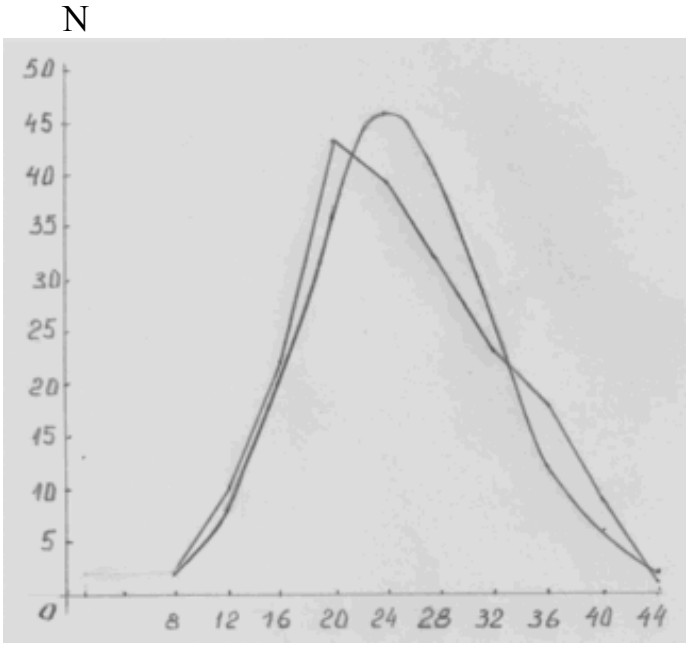


Рисунок 8.1 Распределение числа стволов сосны по диаметру, выровненное с помощью кривой типа А

Правильность расчетов теоретических частот ряда проверяют сравнением общей их суммы с суммой фактических частот. Видно, что экспериментальная кривая согласуется с моделью (8.1), т.к. N в обоих случаях равно 200.

Распределение Грама-Шарлье широко применяется в лесоводственных исследованиях. Наиболее часто его используют, изучая древостои, где имеет-

ся значительное антропогенное воздействие, особенно рубки промежуточного пользования. В этом случае нарушаются требования предельной теоремы Ляпунова о том, чтобы воздействие каждого фактора было относительно мало и примерно соотносимо с другими.

Рубки промежуточного пользования могут существенно изменить характер распределения деревьев в древостое. В то же время при их правильном проведении в чистых насаждениях сохраняется одновершинность кривой и ее соответствия обобщенному нормальному распределению, т.е. распределение из нормального трансформируется в тот его вид, который описывается кривой типа А. Это основная причина того, что названную кривую широко применяли для описания реальных распределений многие ученые-лесоводы: А.Патацкас, К.Е. Никитин (1908-1987), Н.Н. Свалов (1918-1995), Л.Н. Толкачев (1938-2008). Кривая типа А взята за модель распределения В.Ф. Багинским при описании строения древостоев Беларуси, на основе которого составлены действующие товарные таблицы для насаждений сосны, ели, дуба, березы, осины и ольхи черной.

8.2 Другие распределения

В лесных исследованиях используется ряд других распределений. Поскольку настоящее учебное пособие предназначено также для магистрантов и аспирантов, то целесообразно привести описание этих распределений. Для студентов-лесоводов при изучении основного курса биометрии можно ограничиться распределением Грама-Шерлье, но в ряде случаев, особенно при выполнении НИИРС нужно обращаться к другим видам распределений.

Сводку распределений, которые находят применение в лесном хозяйстве, сделали К.Е. Никитин и А.З. Швиденко в капитальном труде «Методы и техника обработки лесоводственной информации», которая приведена в списке литературы. Поскольку книга вышла более 30 лет назад и уже стала библиографической редкостью, приведем здесь материалы о других распределениях в том виде как они изложены в упомянутой монографии.

Сводка основных распределений, используемых в лесном хозяйстве, приведена в таблице 8.3.

Опишем основные из приведенных распределений.

Логнормальное распределение. Логнормальное распределение формируется в условиях, подобных тем, где применяется кривая типа А. Величина x распределена логнормально, если логарифмы её значений $U = \bar{l}$, n, x имеют нормальное распределение. Теоретически это распределение можно рассматривать как распределение величины, полученной умножением примерно одинаковых случайных величин при большом их числе, и в этом смысле оно является некоторым аналогом нормального закона.

Это распределение случайной величины x_i зависит от двух параметров (среднего и дисперсии логарифмов значений с. в. X), хотя можно ввести один

или два параметра, ограничивающие размах распределения с одной или двух сторон.

Таблица 8.3 – Основные непрерывные распределения, используемые в лесном деле

Формула	Распределение	Плотности распределения	Среднее значение	Дисперсия
8.0	Нормальное	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right]$	$\bar{x}$	σ^2
8.1	Грамма-Шарлье («обобщенное нормальное»)	$f(x)-\frac{r_3}{6}f^{(3)}(x)+\frac{r_4-3}{24}\times f^{(4)}(x)-\ldots$, $f(x), f^{(3)}(x), f^{(4)}(x)$ - плотность нормального распределения и её 3 и 4-я производные	$\bar{x}$	σ^2
8.6	Логарифмически-нормальное	$\frac{1}{\sigma_u\sqrt{2\pi}}\exp\left[-\frac{(\ln x_i-\bar{u})^2}{2\sigma_u{}^2}\right]$, где $\sigma_u{}^2$ - дисперсия lnx; u - среднее $(\ln \bar{X})$, $\bar{X}\geq 0$	$\bar{u}$	$\sigma_u{}^2$
8.11	Показательное (экспоненциальное)	$\lambda t\exp(-\lambda x)$ при $x\geq 0$ 0 при $x<0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
8.12	Вейбулла	$abx^{a-1}e^{-bx^a}$	$b^{-1/a}г(\frac{1}{a}+1)$	$b^{-2/a}\left[г(\frac{a}{2}+1)-r^2(\frac{1}{a}+1)\right]$
8.13	Равномерное (прямоугольное)	$(b-a)^{-1}$ при $a\leq x\leq b$ 0 при $x<a$. $x>b$	$\frac{b-a}{2}$	$\frac{(b-a)^2}{12}$
8.14	Гамма-распределение (III тип Пирсона)	$\frac{a^b x^{b-1e^{-ax}}}{\Gamma(b)}$; $x\geq 0$ $a>0$, $b>0$	$\frac{b}{a}$	$\frac{b}{a^2}$
8.15	Бета-распределение (I тип Пирсона)	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$	$\frac{ab}{(a+b)_2(a+b+1)}$	$\frac{2(b-a)(a+b+1)^{1/2}}{(ab)^{1/2}(a+b+2)}$

Кривая распределения имеет правостороннюю асимметрию, которая возрастает с увеличением σ_u (рисунок 8.2.), поэтому хорошо аппроксимирует распределения с положительной косостью.

Рисунок 8.2 Кривые логнормального распределения с различными параметрами σ_u

Если для величин x_i известно среднее $\overline{X}$ и дисперсия σ_x^2, то параметры логнормального распределения можно вычислить непосредственно по формулам

$$\sigma_u^{\,2} = \ln\left[\sigma_x^{\,2} / \overline{X}^2 + 1\right], \qquad (8.7)$$

$$\overline{U} = \ln \overline{X} - \frac{\sigma_u^{\,2}}{2} = \ln \overline{X} - \frac{1}{2}\left[\frac{\sigma_x^{\,2}}{\overline{X}^2} + 1\right], \qquad (8.8)$$

а плотность логнормального распределения величины x

$$f(x) = \frac{1}{x\sqrt{2\pi \ln(\sigma^2 / \overline{X}^2 + 1)}} \times \exp\left\{-\frac{\ln x - \ln \overline{X} + \frac{1}{2}(\sigma^2 / \overline{X}^2 + 1)}{2\ln(\sigma^2 / \overline{X}^2 + 1)}\right\}, \qquad (8.9)$$

Уравнение задано на интервале [0, ∞]. Если кривая распределения ограничена слева точкой x_1, то имеем трехпараметрическое логнормальное распределение.

$$f(x) = \frac{1}{\sigma_u (x - x_1)\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_u^{\,2}}\left|\ln(x - x_1) - \overline{u}\right|\right\}, \qquad (8.10)$$

которое заменой $x' = x - x_1$ приводится к (8.6.).

Имеются многочисленные примеры использования логнормального распределения как модели при свертке лесоводственной информации.

Пример. Вычислим выравнивающие частоты для ряда распределения числа стволов по диаметру в древостое ели (таблица 8.4) по уравнению логнормального распределения (8.6). Среднее значение и дисперсия этого ряда соответственно равны $\overline{X}$ = 18,36 (см), σ^2 = 5,85 (см). Если бы эти показатели были неизвестны, то в таблице 8.4. следовало бы добавить 3 колонки для отклонений от x_0 и вычисления первых двух моментов. По (8.7) и (8.8) находим

$\sigma_u^2 = 0{,}311$, $\overline{U} = 2{,}862$. Дальнейшая схема вычислений полностью соответствует схеме таблицы 8.4.

Таблица 8.4 – Схема вычисления выравнивающих частот с применением логнормального распределения для диаметров древостоя ели (таблицы 8.2 – 8.3)

x_i	n_i	$b = x_i + \frac{c}{2}$	$\ln b$	$\ln b - \overline{n}$, где $n = \ln \overline{X}$	$z = \frac{\ln b - \overline{n}}{\sigma_u}$	Ф/z	nФ/z	$\widetilde{n}_i$
8	11	10	2,3026	-0,559	-1,80	0,036	20	20
12	18	14	2,6391	-0,223	-0,72	0,236	132	112
16	181	18	2,8904	+0,029	+0,09	0,536	209	167
20	124	22	3,0910	0,229	0,74	0,770	430	130
24	67	26	3,2581	0,396	1,27	0,898	501	72
28	31	30	3,4012	0,539	1,73	0,958	535	34
32	17	34	3,5264	0,665	2,14	0,984	549	15
36	5	38	3,6376	0,776	2,49	0,994	555	5
40	3	42	3,7377	0,876	2,82	0,998	557	2
44	1	46	3,8286	0,967	3,11	1,000	558	1
Σ	558	-	-	-	-	-	-	558

Сравнение эмпирических и вычисленных частот свидетельствует о хорошем соответствии принятой модели ряду распределений.

Логнормальное распределение использовано для выравнивания рядов распределения диаметров литовскими учеными-лесоводами: В. В. Антанайтисом, А. А. Кулешисом, Ю. Ф. Можейкой и др.

Распределение Вейбулла. Распределение Вейбулла (формула (8.12), таблица 8.5) обычно применяют как модель распределения времени ожидания, например, времени работы системы, состоящей из совокупности последовательно соединенных элементов, т.е. распределение времени работы до выхода из строя первого элемента. Его можно рассматривать как обобщение показательного распределения, если интенсивность отказов меняется во времени. Параметр α характеризует скорость выхода системы из строя. Это распределение хорошо описывает время выхода из строя отремонтированного оборудования, срок работы машин, тракторов и т.д.; его можно использовать как модели распределения крайних (экстремальных) значений или распределения биометрических характеристик деревьев, в частности диаметра. Более подробное описание этого распределения здесь опускаем. Оно приведено в книге К.Е. Никитина и А.З. Швиденко, имеющейся в предложенном списке литературы.

Гамма- и бета- распределения. Гамма- и бета- распределения принадлежат к числу основных моделей, используемых при изучении распределений. Оба они связаны с одним из наиболее общих распределений – распределением Маркова, из которого можно получить практически все встречаемые

в приложениях распределения как предельные стохастические кривые. Условия, при которых формируются гамма- и бета – распределения, весьма широки в зависимости от величины входящих в них параметров. Как правило, они могут описывать любую практическую ситуацию из приведенных в настоящем параграфе, а ряд рассмотренных распределений может быть получен как частные случаи гамма- и бета- распределений.

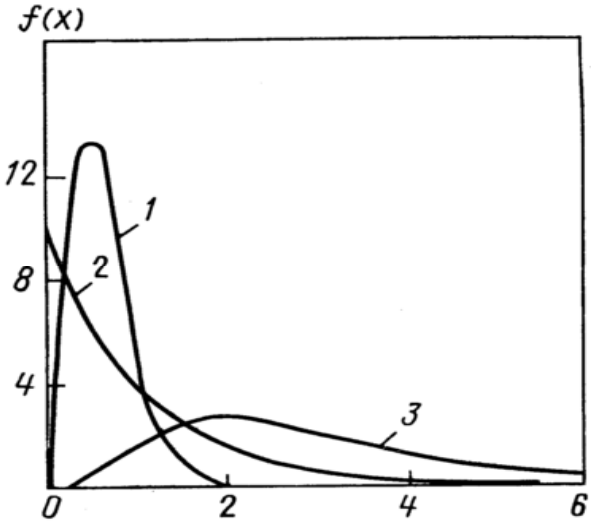


Рисунок 8.3 Кривые гамма-распределения:
1-α=5; b=3; 2-α=b=1 3-α=1; b=3

Гамма- распределение (формула (8.14), таблица 8.3) – одна из основных статистических моделей для представления распределений случайной величины, ограниченных с одной стороны. Это распределение хорошо описывает время, необходимое для появления ровно b независимых событий, если они происходят с равной интенсивностью α. Форма и масштаб кривых распределения зависят от величины и соотношения параметров α и b: b- параметр формы, α – параметр масштаба. Если $b \le 1$, то плотность гама- распределения убывающая кривая (рисунок 8.3.), если $b > 1$, то распределение представлено одновершинной кривой с максимумом в точке $(b-1)/\alpha$.

Обычно в лесном хозяйстве интервал значения случайной величины ограничен с обоих концов. Плотность гамма-распределения для случая, когда величина задана на интервале [x, ∞], выражается формулой

$$f(x) = \frac{1}{\Gamma(b)} a^b (x - x_1)^{b-1} e^{-a(x-x_1)}, \; x \ge x_1, \qquad \alpha>0, \; b>0 \qquad (8.16)$$

Переход к распределению в формуле (8.15.) обеспечивается заменой $x' = x - x_1$. При аппроксимации гамма-рапределением вероятности для очень больших значений случайной величины невелики, ими пренебрегают как и в случае нормального распределения.

Для практического вычисления параметров α и b используют метод моментов, дающий приближенные, но, как правило, вполне приемлемые результаты. Среднее значение Г-распределения $\overline{X} = b/a$, а дисперсия $\sigma^2 = b/\alpha^2$.

Вычислив на основании выборки значения $\overline{X}$ и σ^2 и приравняв их соответствующим соотношением параметров, находим выборочные оценки $\hat{\alpha}$ и $\hat{b}$

$$\hat{a} = \frac{\overline{X}(n-1)}{\sum (x_i - \overline{X})^2} = \frac{\overline{X}}{\sigma^2} = (n-1)\left[\frac{\sum x_i}{n\sum x_i^{\,2} - (\sum x_i)^2}\right], \qquad (8.17)$$

$$\hat{b} = \hat{a}\overline{X} \qquad (8.18)$$

При использовании гамма-распределения приходится вычислять значения Г-функци, называемой иначе интегралом Эйлера первого рода.

Г-функция – обобщенное понятие факториала для какого-либо положительного числа ρ, в том числе и для дробного

$$\Gamma(\rho) = \int_0^\infty x^{\rho-1} e^{-x} dx, \quad \rho > 0. \qquad (8.19)$$

Для ρ целого положительного Г (ρ +1) = ρ! =1 · 2 · 3 … ρ.

Так как Г – функция имеет один минимум (в точке 1,4616 … со значением Г (1,4616 …) = 0,8856 …), а с увеличением ρ значения Г (ρ) резко возрастают, то обычно пользуются lg Г (ρ). Обычно в таблицах приводят значения lg Г (ρ) для ρ от 1 до 2. Для других значений lg Г (ρ) находят с использованием соотношения Г (ρ + 1) = ρ Г (ρ), откуда при ρ >2

lg Г (ρ) = lg (ρ - 1) + lg (ρ - 2) + … + lg (ρ - k) + lg Г (ρ - k) (8.20)

а при 0 < ρ < 1

lg Г (ρ) = lg Г (ρ + 1) - lg ρ. (8.21)

Многие известные распределения являются частными случаями гамма-распределения. Если b = 1, то получаем эспоненциальное распределение: при α = ½ и b кратном 1/2 имеем x - квадрат-распределение, а при b целом положительном – распределение Эрланга, широко используемое в теории массового обслуживания.

Для примера аппроксимируем при помощи гамма-распределения ряд распределения диаметра (из таблицы 8.4.). Следует иметь в виду, что, выбирая подходящим образом начало кривой и ее масштаб, можно значительно упростить вычисления. Поскольку для гамма-распреления в форме (8.16) начало кривой находится в точке x_1 (в нашем примере x_1 = 6 см), Перенесем начало координат в эту точку и произведем замену $x_i' = (x_i - x) / c$, чтобы можно было пользоваться статистиками, вычисленными для ряда в «рабочих единицах».

Для нашего примера имеем $\overline{X}$ = 3,0896, $\sigma^2 = \mu_2$ = 2,1385. Тогда по (8.17) и (8.18) $\hat{a}$ = 3,0896 / 2,135 = 1,4448; $\hat{b}$ = 1,4448 · 3,0896 = 4,4637, и уравнение (8.14) приобретает вид

$$f(x) = \frac{(1{,}4448)^{4{,}4637}}{\tilde{A}(4{,}4637)} x^{3{,}4637} e^{-1{,}4448\,x}, \qquad (8.22)$$

а выравнивающие частоты $\tilde{n}_i = f(x)n$, (8.23)

где n – количество наблюдений.

Заметим, что при вычислениях непосредственно для исходного ряда (а не для рабочего) среднее = 18,36 см, $\hat{\sigma}^2$ = 34,216 (см^2) и параметр формы b

практически не меняется, а в формуле (8.23) n_i необходимо умножить на величину c. Подставив (8.23) в (8.22) и прологарифмировав, получаем

lg f (x) = lg 558 + 4,4637 lg 1,4448 – lg Г (4,4637) + 3,4637 lg x – 1,4448 x lg e = 2,4157 + 3,4637 lg x – 0,6275 x,

где логарифм Г-функции вычислен по (8.20) с использованием таблиц lgГ (ρ), которые на отрезке от x=1 до x=2 приведены в приложении Д. Для больших величин x эти значения опущены. На отрезке от x=1 до x=50 они имеются в книге А.К. Митропольского, приведенной в списке литературы.

lgГ (4,4637) = lg 3,4637 + lg 2,4637 + lg 1,4637 + lg (1,4637) =

= 0,5395 + 0,3916 + 0,1654 + 1,9472 = 1,0438

Схема вычисления выравнивающих частот по кривой Г-распределения приведена в таблице 8.5. Контролем вычислений служит совпадение сумм колонок 2 и 8; небольшие расхождения объясняются погрешностями округлений в процессе расчета.

Таблица 8.5 – Схема вычисления выравнивающих частот по кривой гамма-распределения

x_i	n_i	x_i'	$0{,}6275\,x_i'$	$\lg x_i'$	$3{,}4637\,\lg x_i'$	$\lg \tilde{n}_i = 2{,}4159 + (4) + (6)$	$\tilde{n}_i$
1	2	3	4	5	6	7	8
8	11	0,5	-0,3138	-0,3010	-1,0426	1,061	11,7
12	118	1,5	-0,9412	+0,1761	0,6099	2,084	121,4
16	181	2,5	-1,5688	0,3979	1,3783	2,225	168,0
20	124	3,5	-2,1962	0,5441	1,8845	2,104	127,1
24	67	4,5	-2,8238	0,6532	2,2625	1,854	71,5
28	31	5,5	-3,4512	0,7404	2,5644	1,529	33,8
32	17	6,5	-4,0788	0,8129	2,8157	1,153	14,2
36	5	7,5	-4,7062	0,8751	3,0309	0,740	5,5
40	3	8,5	-5,3338	0,9294	3,2192	0,301	2,0
44	1	9,5	-5,9612	0,9777	3,3865	1,841	0,7
∑	558						555,9

Бета-распределение (формула (8.15) в таблице 8.3) часто называют основным распределением для величин, ограниченных с двух сторон. Это удобная модель для многочисленных приложений, поскольку кривая бета-распределения может принимать самую различную форму в зависимости от величины параметров (рисунок 8.4). Кроме того, посредством бета-распределения можно вычислять другие важные распределения.

Если α > b > 1 или b > α > 1, то распределение одновершинное с максимумом в точке x = (α – 1) / (α + b – 2) с левосторонней асимметрией в первом и правосторонней во втором случае; если α < 1, b < 1, то распределение имеет U-образную, а при α ≥ 1, b < 1 i-образную форму. При α < 1, b ≥ 1 кривая распределения убывающая.

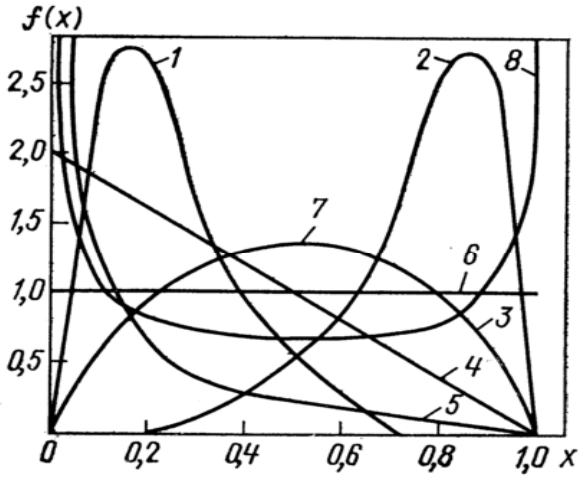


Рисунок 8.4 Кривые бета-распределения
1-α=1,5, b=5; 2-α=5, b=1,5; 3-α=b=2; 4-α=1, b=2;
5-α=0,2, b=2; 6-α=b=1; 7-α =b=2; 8-α=b=0,5

Если α = b, то распределение симметрично. В качестве примеров случайной величины, подчиняющихся бета-распределению, можно привести выработку бригады, цеха и др. за определенный срок (смену, сутки), распределение большинства биометрических признаков деревьев и древостоев и др.

Название бета-распределения следует из того, что выражение из (8.15) в таблице 8.3, имеющие вид

$$\frac{\tilde{A}(a)\tilde{A}(b)}{\tilde{A}(a+b)} = \beta(a+b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx \qquad (8.24)$$

при α > 0, b > 0 называют β-функцией или интегралом Эйлера II рода. Так как B-функция выражается через Г-функцию, то её обычно вычисляют по таблице значений Г-функции.

Формула плотности задает бета-распределение на интервале (0,1). В конкретных задачах интервал обычно ограничен некоторыми значениями [x_1, x_2]. Тогда плотность задается формулой.

$$f(x) = \frac{1}{x_2 - x_1} \cdot \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \left(\frac{x - x_1}{x_2 - x_1}\right)^{a-1} \left(1 - \frac{x - x_1}{x_2 - x_1}\right)^{b-1} \qquad (8.25)$$

Заменив в (8.25) x' = (x – x_1) /(x_2 – x_1), получим плотность бета-распределения на [0, 1], т.е. формулу (8.24).

Частными случаями бета-распределения являются равномерное (α = b = 1), треугольное (α = 2, b = 1) и параболическое (α = b = 2). В задачах свертки информации эти частные случаи находят применение для приближенного представления более сложных распределений.

Соотношения между параметрами бета-распределения и моментами, в частности средним и дисперсией, можно использовать для аппроксимации бета-распределения:

$$\hat{b} = \frac{1-\overline{X}}{\hat{\sigma}^2}\left[\overline{X}(1-\overline{X}) - \hat{\sigma}^2\right], \qquad (8.26)$$

$$\hat{a} = \overline{X}\hat{b}/(1-\overline{X}), \qquad (8.27)$$

где $\overline{X}$ - выборочное среднее; $\hat{\sigma}^2$ - выборочная дисперсия.

Пример. Вычислим выравнивающие частоты по кривой бета-распределения ряда распределения второго коэффициента формы q_2 (таблица 8.6). Напомним, что $q_2=d_{0.5}/d_{1.3}$, где $d_{0.5}$ - диаметр ствола на 0,5 высоты дерева, $d_{1.3}$ – диаметр ствола на высоте 1,3 м. Коэффициент q_2 широко применяется для оценки полнодревесности деревьев.

Если имеем $x_1 = 0{,}54$, $x_2 = 0{,}78$ (начало и конец ряда), среднее и дисперсия (в рабочих единицах) $\overline{X} = 0{,}6596$, $\hat{\sigma}_1^2 = 3{,}7712$. Замена x' = (x – 0,54) / (0,78 – 0,54) дает $x_1' = 0$, $x_2' = 1$, $\overline{X}' = 0{,}4983$, величина разряда c = 0,02/0,24 = 0,08333, дисперсия $\hat{\sigma}^2 = (c\hat{\sigma}_1)^2 =$ $(0{,}08333)^2 \cdot 3{,}7712 = 0{,}02619$. Значения α и b находим из (8.26) и (8.27)

$$\hat{b} = \frac{0{,}5017}{0{,}02619}(0{,}4983 - 0{,}5017 - 0{,}02619 = 4{,}2873,$$

$$\hat{a} = \frac{0{,}4983}{0{,}5017}4{,}2873 = 4{,}2582$$

Для перехода к выравнивающим частотам значения f (x) из (8.25) следует умножить на величину n/c, т. е.

$$n_i = \frac{n}{c}\cdot\frac{\tilde{A}(a+b)}{\tilde{A}(a)\tilde{A}(b)}x^{a-1}(1-x)^{b-1}, \qquad (8.28)$$

Прологарифмировав 8.28 и вычислив коэффициенты, получим

$$\lg\left[\frac{n}{c}\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right] = \lg n + \lg\Gamma(a+b) - \lg c - \lg\Gamma(a) - \lg\Gamma(b) =$$

$$= 2{,}3997 + 4{,}1883 - (-1{,}0792) - 0{,}9398 - 0{,}9231 = 3{,}6459$$

$\ln n_i = 3{,}6459 + 3{,}2582 \lg x'_i + 3{,}2873 \lg (1 - x'_i)$.

Схема вычисления n_i по последнему выражению приведена в таблице 8.6

В колонках 4 и 7 записаны одинаковые числа, только в обратном порядке (в колонке 4 сверху вниз, в колонке 7 снизу вверх. Поэтому при вычислениях значения колонки 7 могут быть получены по данным колонки 4. Контролем правильности вычислений служит близкое значение сумм колонок 2 и 10.

Таблица 8.6 – Схема вычисления выравнивающих частот распределения числа деревьев по коэффициенту формы q_2 по кривой бета-распределения

x_i (q_2)	n_i лет	x_i'	lgx_i'	3,2582 lgx_i'	1-x_i'	$lg(1-x_i')$	3,2873 $lg(1-x_i')$	$lg\,\tilde{n}_i$	$\tilde{n}_i$
0,55	1	0,0417	-1,3799	-4,4960	0,9583	-0,0185	-0,0608	1,089	0,1
0,57	3	0,1250	-0,9031	-2,9425	0,8750	-0,0580	-0,1907	0,513	3,3
0,59	12	0,2080	-0,6813	-2,2198	0,7917	-0,1014	-0,3333	1,093	12,4
0,61	23	0,2917	-0,5351	-1,7435	0,7083	-0,1498	-0,4924	1,410	25,7
0,63	38	0,3750	-0,4260	-1,3880	0,6250	-0,2041	-0,6709	1,587	38,6
0,65	49	0,4583	-0,3388	-1,1039	0,5417	-0,2662	-0,8751	1,667	46,5
0,67	51	0,5417	-0,2662	-0,8673	0,4583	-0,3388	-1,1137	1,665	46,2
0,69	36	0,6250	-0,2041	-0,6650	0,3750	-0,4260	-1,4004	1,581	38,1
0,71	21	0,7083	-0,1498	-0,4881	0,2917	-0,5651	-1,7590	1,399	25,0
0,73	14	0,7919	-0,1014	-0,3304	0,2083	-0,6813	-2,2396	1,076	11,9
0,75	2	0,8750	-0,0580	-0,1889	0,1250	-0,9031	-2,9087	0,489	3,1
0,77	1	0,9583	-0,0189	-0,0603	0,0417	-1,3799	-4,4960	0,090	0,1
∑	251								251

Описанная β-функция является основной моделью, которую применяют проф. О.А. Атрощенко и его ученики для решения большого ряда прикладных задач: материально-денежная оценка лесосек, описание строения древостоев и т. д.

8.3 Семейство кривых распределения Джонсона

В лесоводственной литературе упомянутое семейство кривых впервые описали К.Е. Никитин и А.З. Швиденко. Приведем здесь его в изложении названных авторов.

Джонсон предложил для аппроксимации эмпирических распределений семейство кривых, полученных преобразованием исходного распределения $\bar{X}$ с плотностью f(x) к нормировано нормально распределенной величине z. Это семейство включает три типа кривых, представляющих распределение неограниченных случайных величин (тип S_U), ограниченных с одной стороны (S_L) и ограниченных с двух сторон (S_B). В общем виде семейство кривых Джонсона требует знания параметра положения ξ, параметра масштаба λ и двух параметров формы – γ и δ. В таблице 8.9 приведены формулы преобразований к кривым Джонсона и соответствующие плотности распределения. Для простоты записи формулы приведены для нормированных значений величины X путем замены $y = (x - \xi)/\lambda$.

Для выбора конкретной кривой необходимо установить, к какому типу принадлежит эмпирическое распределение. Для этой цели на основании ряда распределения вычисляют выборочные значения β_1 и β_2 и по графику (рисунок 8.5) устанавливают тип распределения.

Таблица 8.9 – Типы преобразований и плотность распределения кривых семейства Джонсона

Тип кривой	Преобразование/ Плотность распределения	
S_L	$z = \gamma + \delta \ln y$	8.29
S_B	$z = \gamma + \delta \ln[y/(1-y)]$	8.30
S_U	$Z = \gamma + \delta \operatorname{arcsh} y = \gamma + \delta \ln(y + \sqrt{y^2+1})$	8.31
S_L	$\frac{\delta}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \times (\gamma + \delta \ln y)^2\right\}, y \geq 0$	8.50
S_B	$\frac{\delta}{\sqrt{2\pi}} \cdot \frac{1}{y(1-x)} \times \exp\left\{-\frac{1}{2}(\gamma + \delta \ln \frac{y}{1-y})^2\right\}, 0 \leq y \leq 1$	8.51
S_U	$\frac{\delta}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{y^2+1}} \times \exp\left\{-\frac{1}{2}\left[\gamma + \delta \ln\left(y + \sqrt{y^2+1}\right)\right]^2\right\}, -\infty \leq y \leq \infty$	8.52

Выборочные распределения, для которых β_1 и β_2 лежат вблизи и на линии S_L, относятся к этому типу; лежащие ниже линии S_L - к типу S_U, а выше, исключая критическую область, - к типу S_B .Затем по выборке оценивают параметры данного типа распределения и вычисляют выравнивающие частоты.

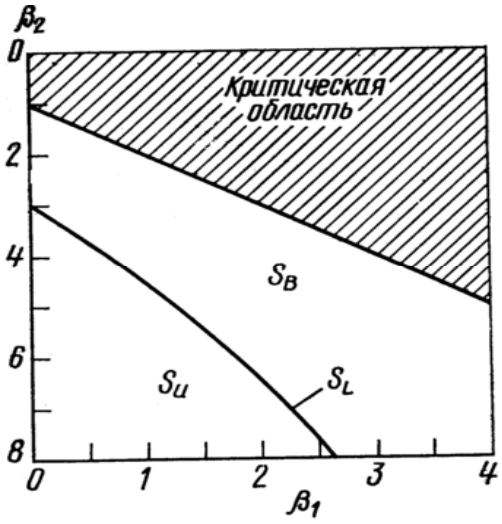


Рисунок 8.5 График для выбора типа кривых Джонсона

Вычисление оценок параметров ξ, λ, γ и δ через моменты очень сложно. Поэтому обычно для аппроксимации кривыми Джонсона используют метод приравнивания теоретических (выравнивающих) и выборочных квантилей. Тип S_L. Кривая распределения ограничена слева точкой ξ , а значения $x \geq \xi$. Можно показать, что этот тип распределения зависит только от трех парамет-

ров δ, ξ и $\gamma^* = \gamma - \delta \ln \lambda$, поэтому, положив $\delta = 1/\sigma$ и $\gamma^* = \overline{X}/\sigma$, получаем трехпараметрическое логнормальное уравнение, приведенное выше (8.10), т.е. тип S_L имеет такую же форму распределения как на рисунке 8.2. Аппроксимация этого распределения возможна методом, рассмотренном в таблице 8.4.

В практике использование типа S_L может встретиться два случая величина $\xi = x_1$ известна; величина ξ не известна. В большинстве задач по аппроксимации величин в лесном хозяйстве, особенно для свертки информации, начало кривой распределения, т.е. величина ξ, как правило, может быть установлена. Так, в качестве ξ можно использовать нижнюю границу первого класса ряда распределения. Если значение ξ известно, то, вычислив значения

$$\overline{X} = \frac{1}{n}\sum \ln\left(x_i - \xi\right), \tag{8.32}$$

$$\hat{\sigma} = \left\{\frac{n\sum\left[\ln\left(x_i - \xi\right)\right]^2 - \left[\sum \ln\left(x_i - \xi\right)\right]^2}{n\left(n-1\right)}\right\}^{1/2}, \tag{8.33}$$

находят два неизвестных параметра формулы (8. 50):

$$\hat{\sigma} = 1/\hat{\sigma}, \tag{8.34}$$

$$\hat{\gamma}* = -\overline{X}/\hat{\sigma}. \tag{8.35}$$

Если значение ξ не известно, то из (8.29) заменой $\gamma^* = \gamma - \delta \ln (x - \xi)$ находят

$$Z = \gamma* + \delta \ln (x-\xi). \tag{8.36}$$

Так как по выборке необходимо оценить параметры γ*, δ и ξ, составляют три уравнения, приравнивающих три выборочных квантиля трем соответствующим квантилям нормированной нормально распределенной величины Z.

$$Z_a = \gamma* + \delta \ln (x_a - \xi), \tag{8.37}$$

где z_a и x_a - соответственно теоретические и выборочные квантили.

Целесообразно выбирать два симметричных квантиля – это упрощает расчеты, в противном случае приходится решать нелинейное уравнение. Вполне приемлемо брать *a* = 0,05; 0,5 и 0,95. Выбор других близких квантилей мало меняет результаты. Тогда, поскольку для нормированного нормального распределения $z_{0,05}$ = - 1,645, $z_{0,5}$ = 0 и $z_{0,95}$ = 1,645, решением системы трех уравнений (8.37) находят:

$$\hat{\delta} = 1{,}645\left[\ln\left(\frac{x_{0,95} - x_{0,5}}{x_{0,5} - x_{0,05}}\right)\right]^{-1}, \qquad (8.38)$$

$$\hat{\gamma}^* = \hat{\delta}\ln\left(\frac{1 - e^{-1{,}645/\hat{\delta}}}{x_{0,5} - x_{0,05}}\right), \qquad (8.39)$$

$$\hat{\xi} = x_{0,5} - e^{-\hat{\gamma}^*/\hat{\delta}}, \qquad (8.40)$$

где $x_{0,05}$, $x_{0,5}$, $x_{0,95}$ – квантили выборочного распределения.

В качестве I и III квантилей можно взять и другие произвольные симметричные квантили x_a и x_{1-a}.

Поскольку техника вычисления выравнивающих частот по уравнению типа S_L по сути одинакова для обоих описанных выше случаев, рассмотрим ее на примере аппроксимации при неизвестном ξ.

В качестве примера вычислим выравнивающие частоты для ряда распределения диаметра стволов в 110-летнем древостое (таблица 8.10). Имеем $\beta_1 = 1{,}0300$ и $\beta_2 = 4{,}3262$, т.е. распределение принадлежит к типу S_L семейства кривых Джонсона. Для определения параметров $\hat{\delta}$, γ^* и ξ выберем квантили соответствующие вероятностям 0,05; 0,5; 0,95. По нашим данным они соответственно равны 10,57; 17,31 и 29,74. По (8.38) – (8.40) получаем сразу:

$$\hat{\delta} = 1{,}645\left[\ln\left(\frac{29{,}74 - 17{,}31}{17{,}31 - 10{,}57}\right)\right]^{-1} = 2{,}688 \quad ;$$

$$\hat{\gamma}^* = 2{,}688\ln\left(\frac{1 - e^{-1{,}645/2{,}688}}{17{,}31 - 10{,}57}\right) = -7{,}229$$

$$\hat{\xi} = 17{,}31 - e^{7{,}229/2{,}688} = 2{,}588\,.$$

Теоретические частоты по вычисленным выборочным оценкам определяют по формуле (8.36) с использованием в качестве квантилей границ классов эмпирического распределения, т.е.для вычисленных квантилей

$$\hat{Z}_p = \hat{\gamma}^* + \hat{\delta}\ln\left(x_p - \hat{\xi}\right). \qquad (8.41)$$

Напомним, что *z* по условию нормированная нормально распределенная случайная величина. По таблицам *Ф (z)* (приложение Б) находят накопленную вероятность соответствующую верхним границам классов ряда распределения, откуда умножением на общее количество наблюдений N = 558 переходят к выравнивающим (накопленным) частотам. Вычисление проводят по схеме таблицы 8.10. В колонках 10, 11 таблицы 8.10 приведены частоты, вычисленные по значениям параметров, оцененных на основании квантилей

a = 0,01; 0,5 и 0,99 (колонка 10) и a = 0,1; 0,5 и 0,9 (колонка 11). В целом расхождения между теоретическими и эмпирическими частотами относительно невелики. Правда, следует учесть, что для рядов с большими значениями частот в крайних классах линейная интерполяция при расчете квантилей и x_a и x_{1-a} может давать заметные погрешности, особенно при малом a (этим объясняется частота 31 в колонке 10 таблицы 8.10). В таких случаях лучшие результаты можно получить построением для нахождения квантилей графика накопленных частот или (что, конечно, лучше) вычислением квантилей для не сгруппированных данных.

Таблица 8.10 – Вычисление выравнивающих частот для типа S_L семейства кривых Джонсона

x_i	n_i	x_p	$\mathrm{Ln}(x_p-\hat{E})$	$\hat{\delta}\ln(x_p-\hat{E})$	z_p	$\Phi(z)$	$\Sigma\tilde{n}_i$	$\tilde{n}_i$	$\tilde{n}_i$ для квантилей	
									α=0,01 1-=0,99	α=0,1 1-=0,9
1	2	3	4	5	6	7	8	9	10	11
-	-	6	1,227	3,299	-3,930	0,001	0,6	0,6	1	1
8	11	10	2,003	5,384	-1,845	0,032	17,9	17,3	31	20
12	118	14	2,435	6,544	-0,685	0,246	137,3	120,0	122	114
16	181	18	2,735	7,352	0,123	0,548	305,8	168,5	166	164
20	124	22	2,956	7,972	0,743	0,771	430,2	124,4	121	127
24	67	36	3,153	8,476	1,247	0,894	498,9	68,7	67	72
28	31	30	3,311	8,900	1,671	0,953	531,8	32,9	32	35
32	17	34	3,447	9,266	2,037	0,979	546,3	14,5	14	15
36	5	38	3,567	9,588	3,359	0,991	553,0	6,7	6	6
40	3	42	3,674	9,876	2,647	0,996	555,8	2,8	2	3
44	1	46	3,771	10,136	2,907	1,0	558,0	1,2	1	1
Σ	558							558	558	558

Тип S_B. Этот тип (формулы (8.30), (8.51)) представляет случайные величины, ограниченные с двух сторон значениями $\xi \equiv x_1$ и $x_2 = \xi + \lambda = (x_1 + \lambda)$, т.е. параметр ξ - здесь начало кривой, а λ - размах распределения, или расстояние между крайними значениями. Кривая типа S_B может принимать различную форму (рисунок 8.6).

При аппроксимации кривыми типа S_B возможны три случая: известны оба крайних значения; одно из них; ни одного. Если оба крайних значения известны, то остается оценить γ и δ. Как и для типа S_L, оценки находят приравниванием теоретических и выборочных квантилей. Выбирают два симметричных квантиля z_a и z_{1-a}, например, 0,05 и 0,95 или 0,1 и 0,9. Оценки γ и δ получают из системы уравнений

$$z' = \gamma + \delta \ln\left(\frac{x' - \xi}{\lambda + \xi - x'}\right), \qquad (8.42)$$

где z' и x' – соответственно квантили z_a и z_{1-a}, x_a и x_{1-a}.

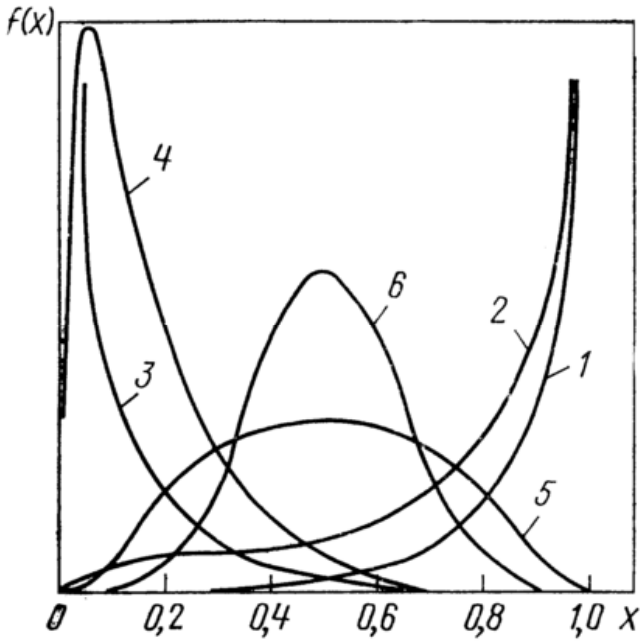


Рисунок 8.6 Кривые распределения Джонсона типа S_B при ε=0, λ=1
1-δ=1,5, γ=-2; 2-δ=0,5, γ=-1;
3-δ=0,5, γ=2; 4-δ=1, γ=2; 5-δ=γ=1; 6-δ=2, γ=0

Отсюда

$$\hat{\delta} = \frac{z_{1-a} - z_a}{kn\left[\frac{(x_{1-a} - \xi)(\xi + \lambda - x_a)}{(x_a - \xi)(\xi + \lambda - x_{1-a})}\right]}; \qquad (8.43)$$

$$\hat{\gamma} = z_{1-a} - \hat{\delta} \ln\left(\frac{x_{1-a} - \xi}{\xi + \lambda - x_{1-a}}\right). \qquad (8.44)$$

Если известно одно крайнее значение ξ, то для оценки трех параметров составляют систему типа (8.42) из трех уравнений, в третьем из которых обычно в качестве соответствующих квантилей используют медиану. Тогда оценки δ, γ и λ находят решением системы (8.42) с добавлением третьего уравнения для определения

$$\lambda = (x_{0,5} - \xi) \times \frac{(x_{0?5} - \xi)(x_a - \xi) + (x_{0,5} - \xi)(x_{1-a} - \xi) - 2(x_a - \xi)(x_{1-a} - \xi)}{(x_{0,5} - \xi)^2 - (x_a - \xi)(x_{1-a} - \xi)} \qquad (8.45)$$

Техника аппроксимации типом S_B такая же, как и типом S_L: 1) находят выборочные значения квантилей для некоторых выбранных заранее вероятностей (один квантиль – медиана, два других – симметричные); 2) по формулам (8.43) – (8.45) находят выборочные значения $\hat{\delta}, \hat{\gamma}^* u \hat{\xi}$ (или по формулам (8.43) и (8.44), если ξ известно); 3) с использованием (8.42) вычисляют квантили z_p, соответствующие квантилям x_p – граница классов аппроксимируемого ряда распределения; 4) по таблице нормированного нормального распределения (приложение А) вычисляют значения накопленных вероятностей, от которых переходят к теоретическим частотам $\sum \tilde{n}_i$ и $\tilde{n}_i$.

Пример. В первых двух колонках таблицы 8.11 приведен ряд распределения высоты (в сантиметрах) жизнеспособного 3-летнего самосева ели. В качестве жизнеспособного учтем самосев высотой от 0,5 см. Аппроксимируем этот ряд при помощи кривых семейства кривых Джонсона.

Определим для ряда моменты. Если в качестве начального значения взять $x_0 = 4$, то начальные моменты $m_1 = -0{,}7442$, $m_2 = 6{,}2053$, $m_3 = -0{,}8768$, $m_4 = 69{,}3000$; Центральные $\mu_2 = 5{,}6515$, $\mu_3 = 12{,}1529$, $\mu_4 = 86{,}3899$; Основные $r_3 = 0{,}9045$, $r_4 = 2{,}7048$. Отсюда $\beta_1 = 0{,}8181$, $\beta_2 = 2{,}7048$, т.е. по графику рисунке 8.6 имеем тип S_B. Примем, что начало кривой $\xi = 0{,}5$. Для вычисления $\hat{\delta}$, $\hat{\lambda}$ и $\hat{\gamma}$ используем квантили $\alpha = 0{,}05$; 0,5 и 0,95. Имеем:

$$x_{0,05} = 0{,}5 + 1\frac{0{,}05 \cdot 950 - 0}{310} = 0{,}65;$$

$$x_{0,50} = 1{,}5 + 1\frac{0{,}05 \cdot 950 - 310}{168} = 2{,}48;$$

$$x_{0,95} = 7{,}5 + 1\frac{0{,}95 \cdot 950 - 872}{43} = 8{,}21$$

Подставив значения квантилей x_α и $x_{1-\alpha}$ и использовав табличные значения z_α и $z_{1-\alpha}$ находим:

$$\hat{\lambda} = (2{,}48 - 0{,}5)\frac{(2{,}48 - 0{,}5)(0{,}65 - 0{,}5) + (2{,}48 - 0{,}5)}{(2{,}48 - 0{,}50)^2 - (0{,}65 - 0{,}50)} \times \frac{(8{,}21 - 0{,}5) - 2(0{,}65 - 0{,}5)(8{,}21 - 0{,}5)}{(0{,}21 - 0{,}50)} = 9{,}5$$

$$\hat{\delta} = \frac{1{,}645 - (-1{,}645)}{\ln\left[\frac{(8{,}21 - 0{,}5)(0{,}5 + 9{,}5 - 0{,}65)}{(0{,}65 - 0{,}5)(0{,}5 + 9{,}5 - 8{,}21)}\right]} = 0{,}59;$$

$$\hat{\gamma} = 1{,}645 - 0{,}59\frac{8{,}21 - 0{,}5}{0{,}5 + 9{,}5 - 8{,}21} = 0{,}78.$$

Дальнейшие вычисления идут по схеме таблицы 8.11.

Таблица 8.11 – Вычисление выравнивающих частот для типа S_B кривых Джонсона (на примере ряда распределения высоты самосева ели)

x_i	n_i	$X_p = x_{i+}\frac{c}{2}$	$\frac{x_p - \xi}{\lambda + \xi - x_p}$	Ln(4)	$\hat{\gamma} + \hat{\delta}(5)$	Φ(6)	$\Sigma n_i = n(1)$	$\tilde{n}_i$	Для квантилей	
									α=0,01 1-α =0,99	α=0,1 1-α =0,9
-	0	0,5		-	-	0,000	0	0	0	0
1	310	1,5	0,118	-2,140	-0,483	0,315	299,2	299,2	291	318
2	168	2,5	0,267	-1,322	0,000	0,500	475,0	175,8	185	159
3	122	3,5	0,462	-0,773	0,324	0,627	595,6	120,6	125	109
4	95	4,5	0,727	-0,318	0,592	0,723	686,8	91,2	92	85
5	70	5,5	1,111	0,105	0,842	0,800	760,0	73,2	72	71
6	57	6,5	1,714	0,539	1,098	0,864	820,8	60,8	58	62
7	50	7,5	2,800	1,030	1,388	0,917	871,2	50,4	46	55
8	43	8,5	5,333	1,674	1,768	0,961	912,9	41,7	37	50
9	35	9,5	18,000	2,880	2,479	0,994	944,3	31,4	28	40
-	0	10,0				1,0	950	5,7	16	1
Σ	950							950	950	950

В целом аппроксимацию можно считать приемлемой. Отметим, что использование квантилей α = 0,01; 0,5 и 0,99 (колонка 10) и квантилей α = 0,1; 0,5 и 0,9 (колонка 11) дало худшие результаты из-за приближенности определения квантилей для сгруппированного ряда. Первая и последняя строки таблицы 8.11. соответствуют границам ряда для колонки (10) x_{max} = 10,82, для колонки (11) x_{max} = 9,6.

Если не известны все крайние значения, то по числу неизвестных параметров составляют четыре уравнения, в которых 4 выборочных квантиля приравнивают к соответствующим квантилям нормированного нормального распределения. Оценки параметров $\hat{\delta}$, $\hat{\gamma}$, ξ и $\hat{\lambda}$ определяют из системы нелинейных уравнений. Случаи с четырьмя неизвестными параметрами в лесных исследованиях крайне редки, а в задачах свертки информации практически исключаются. Поэтому более детально тип S_B (формулы (8.31), (8.52)) здесь не рассматривается.

Тип S_U. Преобразование и плотность кривых распределения типа S_U (рисунок 8.7) приведены в таблице 8.9. В задачах аппроксимации кривыми этого типа требуется оценка всех четырех параметров. Как и в аналогичном случае типа S_B, составляют систему уравнений, в которых приравнивают четыре выборочных и четыре теоретических квантиля. Это приводит к нелинейной относительно искомых параметров системе. Поэтому Н. Джонсон предложил таблицы для нахождения по значениям β_1 и β_2 оценок параметров $\hat{\gamma}$ и $\hat{\delta}$. Величины $\hat{\lambda}$ и $\hat{\xi}$ вычисляют по формулам

$$\hat{\lambda} = \frac{\hat{\sigma}}{\left\{\frac{1}{2}(\omega - 1)\left[\omega ch\left(2\frac{\hat{\gamma}}{\hat{\delta}}\right) + 1\right]\right\}^{1/2}}; \qquad (8.46)$$

$$\hat{\xi} = \widetilde{X} + \hat{\lambda}\omega^{1/2} sh\left(\frac{\hat{\gamma}}{\hat{\delta}}\right). \qquad (8.47)$$

Техника вычислений иллюстрируется следующим примером.

Пусть имеем ряд распределения, составленный по результатам 250 случайных замеров диаметра поперечного среза по высоте груди ствола ели (таблица 8.12). Статистики ряда следующие: $\overline{X}$ = 29,43 см; $\hat{\sigma}$ = 0,4454 см; β_1 = 0,0527; β_2 = 4,96; $\sqrt{\beta_1}$ =0,23. По таблицам Джонсона находим $\hat{\gamma}$ = -0,1972; $\hat{\delta}$ = 1,86. По (8.46) и (8.47) вычисляем $\hat{\lambda}$ и $\hat{\xi}$, учитывая, что ω = exp $[(1,86)^{-2}]$ = 1,336, а $\omega^{1/2}$ = 1,156.

Таблица 8.12 – Схема аппроксимации типа S_U кривых Джонсона для ряда распределения диаметра на 1,3 м деревьев ели

x_i	n_i	$x_i'=x_i+\frac{c}{2}$	$y=\frac{x_i'-\xi}{\lambda}$	arcsh y	$z=\gamma+\delta(5)$	$\Phi(z)$	f(z)	$\tilde{n}_i$
1	2	3	4	5	6	7	8	9
	0	27,6	-2,463	-1,630	-3,23	0	0	0
27,8	1	28,0	-1,897	-1,394	-2,79	0,003	0,003	0,8
28,2	3	28,4	-1,331	-1,100	-2,24	0,013	0,010	2,5
28,6	11	28,8	-0,764	-0,707	-1,51	0,066	0,053	13,2
29,0	51	29,2	-0,198	-0,196	-0,56	0,288	0,222	55,5
29,4	96	29,6	0,368	0,360	0,47	0,681	0,393	98,2
29,8	67	30,0	0,934	0,834	1,35	0,912	0,231	57,8
30,2	15	30,4	1,501	1,195	2,21	0,986	0,074	18,5
30,6	5	30,8	2,067	1,474	2,54	0,994	0,008	2,0
31,0	1	31,2	2,633	1,694	2,95	0,998	0,004	2,0
	0	31,2	-	-		1,0	0,002	0,5
Σ	250							250

$$\hat{\lambda}=\frac{0,4454}{\left\{\frac{1}{2}0,336\left[1,336ch\left(-0,212\right)+1\right]\right\}^{1/2}}=0,7064$$

$$\hat{\xi}=29,43+0,7064\cdot 1,156\ \mathrm{sh}\ (-0,106)=29,34.$$

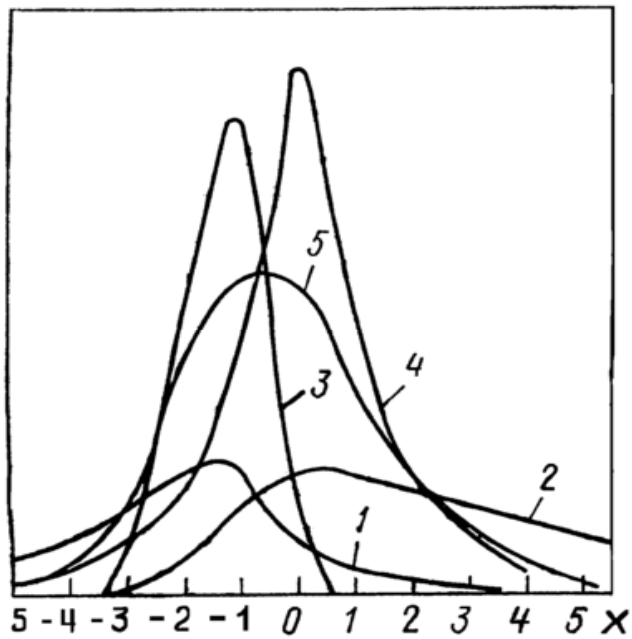


Рисунок 8.7 Кривые распределения Джонсона типа S_U при ε=0, λ=1: 1-δ=1, γ =2; 2-δ=0,5, γ=-2; 3-δ=2, γ=1; 4-δ=0,5, γ=0; 5-δ=γ=2

8.4 Семейство кривых Пирсона

К. Пирсон в цикле работ начала прошлого века предложил цельное семейство кривых. Эта система распределений Пирсона столь же обширна, как и семейство кривых Джонсона. Семейство кривых Пирсона полностью определяется первыми четырьмя моментами. Оно насчитывает 12 типов кривых, из которых обычно используют 7. Конкретный тип кривой устанавливают в зависимости от величин א–критерия, который называют критерием Пирсона и определяют через величины r_3 и r_4:

$$א = \frac{r_3^2(r_4 - 3)}{4(4r_4 - 3r_3^2)(2r_4 - 3r_3^2 - 6)} = \frac{r_3^2(S+2)^2}{16(S+1)}, \qquad (8.48)$$

где r_3 и r_4 – третий и четвертый основные моменты,

$$S = \frac{6(r_4 - r_3^2 - 1)}{3r_3^2 - 2r_4 + 6}. \qquad (8.49)$$

Семейство кривых Пирсона в отличие от кривых Джонсона широко используется в исследованиях по лесному хозяйству, для чего разработаны стандартные компьютерные программы, входящие в сертифицированное математическое обеспечение персональных компьютеров. Поэтому детальное описание выравнивающих частот по кривым Пирсона опускаем.

В то же время студенты и, особенно магистранты и аспиранты, должны понимать математическое содержание этого семейства кривых. Поэтому дадим их краткое описание. Подобное изложение семейства кривых Пирсона приводит А.К. Митропольский.

Обобщение критериев для определения типа кривых Пирсона 1-7, их основные уравнения и некоторые особенности показаны в таблице 8.13.

С точки зрения практической аппроксимации наибольший интерес представляют следующие четыре типа ассиметричных кривых (все они, за исключением кривых IV типа, могут быть U – и I- образные): I – случайная величина имеет размах, ограниченный с двух сторон, критерий א< 0, это – бета-распределение, рассмотренное выше; III –размах ограничен слева; критерий א=±∞, практическое применение возможно при א>4, это – гамма-распределение; IV – размах неограничен с обеих сторон, критерий א заключен между 0 и 1; VI – размах ограничен с одной стороны (как правило справа). Последний тип заменой y = α/x можно привести к первому типу, а непосредственное применение IV типа сопровождается определенными вычислительными трудностями. В качестве основных обычно используют кривые I, IV и VI типов со значением критериев соответственно 0<א<1 א и >1, обеспечивающих по разнообразию форм кривых распределений потребности лесоводственной практики. Остальные типы – переходные на границах между I, IV и VI. Если א = x и r_4 = 0, то получают кривую нормального распределения, если א = 0, r_4 < 3 – типа II, если r_4 > 3 – VII, если א = 1 – V.

Таблица 8.13 – Определение типов кривых Пирсона (по А.К. Митропольскому)

Тип кривой	Критерий определения типа кривой	Уравнение кривой	Примечания
I	א <0 ордината кривой в начальной точке равна - ∞; в конечной – 0	$f_I(x) = f_{I,0}(1+\frac{x}{l_1})^{q_1} \cdot (1-\frac{x}{l_2})^{q_2}$ где l_1, l_2 – пределы размаха ряда распределения; q_1, q_2 - показатели, которые могут быть положительными и отрицательными	
II	א=0 r_3=0 r_4<3	$f_{II}(x) = f_{II,0}(1-\frac{x^2}{l^{1,2}})^q$	Частный случай кривой I. Симметрична вокруг оси O_y и имеет конечный размах распределения
III	א=±∞	$f_{III}(x) = f_{III,0}(1+\frac{x}{l_1})^p \cdot l^{-\frac{px}{l_1}}$ $p = \frac{4}{r_3^2} - 1$	Асимметричная кривая, ограниченная в одном направлении точкой $x=-l$
IV	0<א<1	$f_{IV}(x) = f_{IV,0}(1+\frac{x^2}{l^2})^{-q} \cdot l^{-y \cdot arctg\frac{x}{l}}$ $q = \frac{r+2}{2}$ $r=-S$	Асимметричная кривая, ограниченная от -∞ до +∞
V	א=1	$f_V(x) = f_{V,0} \cdot x^{-p} \cdot l^{-\frac{\gamma}{x}}$ $p = 4 + \frac{8+4\sqrt{4+r_3}}{r_3}$ $\gamma = \overline{\sigma}(p-2)\sqrt{p-3}$	x не принимает отрицательных значений. Кривая асимметрична в одном направлении 0<x<1
VI	1<א<∞	$f_{VI}(x) = f_{VI,0}(x-l)^{q_1} \cdot x^{q_2}$	Кривая асимметрична и ограничена в одном направлении точкой $x=l$ $l \le x < \infty$
VII	א=0 r_3=0 r_4>3	$f_{VII}(x) = f_{VII,0}(1+\frac{x^2}{l^{1,2}})^{-q}$	при א=0, r_3=0, r_4=3 имеем нормальную кривую

		$q = \frac{5r_4 - 9}{2(r_4 - 3)} \rangle 0 \; ; l = \overline{\sigma}\sqrt{\frac{2r_4}{r_4 - 3}}$	

Аппроксимацию кривыми Пирсона проводят в определенной последовательности:

1) по результатам наблюдений вычисляют первые четыре момента эмпирического распределения, на основе которых определяют критерий Пирсона א и выбирают тип кривой распределения;

2) через эмпирические моменты выражают параметры кривой выбранного типа;

3) полученные параметры подставляют в уравнение соответствующего типа и вычисляют теоретические частоты.

Систематическое изложение техники аппроксимации по каждому типу семейства Пирсона достаточно громоздко. В настоящее время эти задачи, как и применение семейства кривых Джонсона, решаются по стандартным программам.

Кривые семейства Пирсона неоднократно применяли многие исследователи в качестве универсальной модели распределения при решении многочисленных задач обработки лесоводственной информации.

Для упрощения нахождения типа кривой Пирсона для конкретного распределения В.Ф. Багинским рассчитана таблица значений критерия Пирсона (א) при заданных α и Е. (таблица 8.14).

Таблица 8.14 – Критерии Пирсона (א) при заданных α и ε

Величина эксцесса *E*	Величина ассиметрии α							
	2,2	1,8	1,4	1	0,8	0,4	0,2	0,1
-2,2	0,08	0,13	0,26	2,44	-0,29	-0,04	-0,01	0
-2,0	0,10	0,17	0,42	-0,57	-0,21	-0,04	-0,01	0
-1,6	0,15	0,29	3,73	-0,3	-0,16	-0,04	-0,01	0
-1,2	0,22	0,61	-1,03	-0,25	-0,16	-0,05	-0,01	0
-1,0	0,28	1,00	-0,73	-0,25	-0,17	-0,05	-0,01	0
-0,8	0,35	2,1	-0,61	-0,25	-0,18	-0,06	-0,02	0
-0,6	0,46	18,02	-0,54	-0,26	-0,19	-0,08	-0,02	-0,01
-0,4	0,60	-3,55	-0,51	-0,28	-0,22	-0,10	-0,03	-0,01
-0,2	0,82	-1,82	-0,49	-0,3	-0,25	-0,14	-0,06	-0,02
-0,1	0,98	-1,51	-0,49	-0,32	-0,27	-0,18	-0,09	-0,03
0,0	1,19	-1,32	-0,49	-0,33	-0,30	-0,26	-0,25	-0,25
0,1	1,48	-1,18	-0,49	-0,35	-0,33	-0,45	0,38	0,04
0,2	1,92	-1,08	-0,49	-0,38	-0,37	-1,56	0,11	0,02
0,4	3,93	-0,96	-0,51	-0,49	-0,5	-0,39	0,05	-0,01
0,6	33,0	-0,88	-0,53	-0,53	-0,78	0,17	0,03	0,01
0,8	-6,36	-0,84	-0,57	-0,68	-1,79	0,11	0,02	0
1,0	-3,20	-0,82	-0,61	0,94	6,96	0,08	0,02	0
1,2	-2,27	-0,81	-0,67	-1,56	1,16	0,07	0,01	0
1,6	-1,83	-0,83	-0,84	4,68	0,44	0,05	0,01	0

2,0	-1,34	-0,88	-1,18	0,94	0,27	0,04	0,01	0
2,2	-1,28	-0,92	-1,49	0,67	0,23	0,03	0,01	0

Из таблицы 8.14 следует, что большая часть всех возможных распределений выравнивается кривой Пирсона I типа. Другие кривые нужны при значительной асимметрии (α>1,0). С учетом того, что II тип Пирсона можно рассматривать как модификацию I типа, последний занимает около 70% всего поля таблицы. Остальную часть в основном занимает IV тип. Нормальная кривая – лишь небольшая часть общего поля таблицы. Остальные типы распределения редки.

Достаточно широкое применение кривых Пирсона для исследования различных распределений обусловлено следующим.

- Применяя систему кривых Пирсона, исследователь не связывает себя никакими предварительными условиями в отношении характера распределения.
- Достаточно простое определение типа кривой.
- Наличие практически в любой системе матобеспечения для ПК программы расчетов по вычислению выравнивающих частот по кривым Пирсона.

9. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

9.1 Ошибки выборочных статистических показателей. Их теоретическое объяснение

Значение основных статистических показателей ($\bar{o}$, σ, α, E) и их распределение очень важно для характеристики статистической выборки, но недостаточно для ее полной оценки. Если бы мы работали с численностями генеральной совокупности, например, замеряли все деревья лесхоза, области, государства, то вычисленных значений среднего, среднеквадратического отклонения и других показателей было бы достаточно. Но в подавляющем большинстве случаев замеры проводят для выборочной совокупности: на пробной площади, на опытном участке и т.д. Поэтому нам надо знать, насколько наши статистические показатели соответствуют аналогичным данным из генеральной совокупности, т.е. достоверны ли они и каковы их ошибки. Такую оценку нам дают вычисленные ошибки статистик.

В основе оценок статистик лежит знание об их распределении. Так, если из одной генеральной совокупности взять некоторое число выборок и для каждой из них определить статистики, например, $\bar{o}$ и σ, то можно выявить распределение этой статистики, которая тоже является случайной величиной. Знание закона распределения искомой статистики позволяет делать ее вероятные оценки. Следовательно, оценку искомого статистического параметра можно сделать только с определенной вероятностью.

Теоретическую основу таких оценок дает теория вероятности и описываемый ею закон больших чисел. Здесь важными для решения наших задач являются теоремы Маркова, Чебышева, Пуассона и Бернулли. Их подробное изложение с приведением доказательств есть в учебниках по теории вероятностей, а также в книге А.К. Митропольского. Учитывая ограниченность курса лесной биометрии, приведем здесь лишь сами теоремы без их доказательства в том порядке, в каком они взаимосвязаны.

Доказательства названных теорем базируются на использовании леммы Маркова. Поэтому приведем ее определение и доказательство.

Лемма Маркова формулируется следующим образом. Если случайная величина x может принимать только положительные значения, то вероятность P того, что эта величина не превзойдет своего математического ожидания $M(x)$, умноженного на некоторое положительное число t^2, больше разности между единицей и числом, обратным данному положительному числу.

Обозначая вероятность соотношения как $P\{x\}$, лемму Маркова можно записать следующим образом:

$$P\left\{\tilde{o} \le M(x)t^2\right\} > 1 - \frac{1}{t^2}. \qquad (9.1)$$

Для доказательства (9.1) допустим, что x принимает только некоторые положительные значения, т.е. $x_1, x_2, \ldots x_k$ с вероятностями $P_1, P_2, \ldots P_k$.

Из ранее изложенного материала мы знаем, что $\sum \delta_i = 1$.

Тогда $M(x)$ будет равно

$$M(x) = \sum P_i x_i. \qquad (9.2)$$

Если возьмем любое положительное число t^2 (при возведении в квадрат любое число будет положительным), допустим, что первые i значений этого ряда не больше $M(x)\,t^2$, а остальные больше $M(x)\,t^2$.

Тогда по теории сложения вероятностей имеем

$$P\left\{(x) \le M(x)t^2\right\} = \sum P_i \text{ и} \qquad (9.3)$$

$$Q\left\{x > M(x)t^2\right\} = P_{i+1} + \ldots \mathrm{P_k}. \qquad (9.4)$$

Следовательно, $P+Q=1$.

Так как вероятности представляют числа неотрицательные, то, опуская в правой части равенства (9.2) члены с индексами

$$1, 2, \ldots, j,$$

мы получим неравенство

$$M(X) \ge p_{j+1}x_{(j+1)} + \ldots + p_k x_{(k)} \qquad (9.5)$$

Так как далее все значения

$$x_{(j+1)}, \ldots, x_{(k)} \qquad (9.6)$$

больше $M(X)t^2$, то, подставляя в (9.5) эту последнюю величину, вместо каждого из значений (9.6), получим более строгое неравенство

$$M(X) > M(X)t^2[p_{j+1} + \ldots + p_k], \qquad \text{т.е.}$$

$$M(X) > M(X)t^2 Q.$$

Разделив левую и правую части этого неравенства на положительную величину $M(X)t^2$, находим

$$\frac{1}{t^2} > Q.$$

И так как на основании того, что P+Q=1 имеем

$$Q = 1 - P, \qquad \text{то}$$

$$\frac{1}{t^2} > 1 - P, \qquad \text{т.е.}$$

$$P\left\{X \le M(X)t^2\right\} > 1 - \frac{1}{t^2},$$

что и выражает лемму Маркова.

Заметим, что хотя лемма Маркова имеет место при любом положительном числе t^2, однако, в силу того, что величина вероятности не может быть меньше нуля, мы, рассматривая (9.1) заключаем, что имеет смысл брать лишь те t^2, которые не меньше 1. При $t^2 = 1$ имеем

$$P\{X \le M(X)\} > 0 .$$

Чем больше P, тем больше будет вероятность того, что

$$X \le M(X)t^2 .$$

Как следствие из леммы Маркова, находим, что

$$Q\{X > M(X)t^2\} \le \frac{1}{t^2}, \qquad (9.7)$$

т.е. вероятность неравенства $X > M(X)t^2$ не больше $\frac{1}{t^2}$.

Написав неравенство (9.1) в виде

$$P\{X \le M(X)t^2\} > 1 - \frac{M(X)}{M(X)t^2}$$ и положив $M(X)t^2 = \tau$, имеем

$$P\{X \le \tau\} > 1 - \frac{M(X)}{\tau} . \qquad (9.8)$$

Поэтому лемма Маркова может быть выражена также следующим образом.

Если случайная величина X может принимать только положительные значения, то вероятность того, что она получит значения, не превосходящие некоторого положительного числа τ, больше

$$1 - \frac{M(X)}{\tau} .$$

Приведем геометрическую иллюстрацию леммы Маркова (рисунок 9.1).

Левая часть неравенства (9.8) выражает вероятность того, что величина X не превосходит некоторого положительного числа τ. По определению, вероятность изменяется от нуля для событий невозможных до единицы для событий достоверных. Так как по условию величина X может принимать только положительные значения, то при $\tau = 0$ очевидно, что

$$P\{X \le \tau\} = 0$$

как вероятность невозможного события. При возрастании τ вероятность P будет возрастать, стремясь к 1 при стремлении τ к бесконечности. На рисунке вероятность P будет изображена линией, которая при $\tau \le 0$ будет совпадать с осью абсцисс, а при $\tau > 0$ будет подниматься над этой осью, стремясь слиться с прямой, параллельной этой оси и проходящей от нее на расстоянии, равном единице (рисунок 9.1).

Рисунок 9.1 Геометрическая иллюстрация леммы Маркова

Правая часть неравенства (9.8) представляет при $\tau > 0$ ветвь гиперболы, которая при $\tau < M(X)$ расположена под осью абсцисс,

при $\tau = M(X)$ пересекает эту ось, при $\tau > M(X)$ проходит над осью абсцисс и при неограниченном увеличении τ стремится слиться с прямой, указанной выше.

На основании неравенства (9.8), кривая вероятностей P расположена над ветвью гиперболы; следовательно, эта кривая не может проникать в области, заштрихованные на чертеже.

Лемма Маркова является основным предложением статистического исчисления. Замечательным свойством этой леммы является ее независимость от природы распределения положительной случайной величины. На лемме Маркова основано доказательство многих теорем статистического исчисления и, в частности, важнейшей из этих теорем – закона больших чисел.

Неравенство Чебышева. Лемма Маркова дает возможность найти вероятность того, что положительная случайная величина примет значение, не превышающее некоторого данного числа; при этом требуется только знание математического ожидания этой величины.

Определенное заключение о случайной величине дает также неравенство Чебышева, которое приложимо к каким угодно (не обязательно положительным) случайным величинам, причем требуется только знание математического ожидания и дисперсии случайной величины.

Неравенство Чебышева. Если случайная величина X может принимать и положительные и отрицательные значения, то, каким бы ни было положительное число τ,

$$P\{-\tau \le X \le \tau\} > 1 - \frac{M(X^2)}{\tau^2}. \qquad (9.9)$$

При установлении этого неравенства применяется лемма Маркова. Доказательство (9.9) опускаем.

Неравенство (9.9) дает нижнюю границу

$$1 - \frac{\sigma^2}{\tau^2}$$

вероятности того, что отклонение значений случайной величины от ее математического ожидания не превзойдет некоторого заданного числа $\pm \tau$. Если дисперсия σ^2 уменьшается, то нижняя граница вероятностей этих отклонений возрастает. Это показывает, что значения случайной величины тем более сосредоточиваются около ее математического ожидания, чем меньше дисперсия. Таким образом, выясняется смысл дисперсии σ^2 как меры рассеяния значений случайной величины около ее математического ожидания.

Полагая в неравенстве (9.9)

$$\tau = t\sigma$$

имеем

$$P\{-t\sigma \le x - M(x) \le t\sigma\} > 1 - \frac{1}{t^2}$$

и, следовательно,

$$Q\{|x - M(x)| > t\sigma\} \le \frac{1}{t^2}.$$

Эти неравенства справедливы для любого распределения случайной величины с конечной дисперсией.

При постоянном основном отклонении σ неравенство показывает, что если t будет увеличиваться, то вероятность того, что значения случайной величины будут находиться в увеличивающемся промежутке $(-t\sigma+M(x), t\sigma+M(x))$, будет увеличиваться. В частности, если t=2, то

$$P\{-2\sigma \le x - M(x) \le 2\sigma\} > 0{,}75;$$

если t=3, то

$$P\{-3\sigma \le x - M(x) \le 3\sigma\} > 0{,}889.$$

Пусть теперь величина t будет постоянной. Тогда при уменьшающемся основном отклонении σ, т.е. уменьшающемся промежутке $(M(x)-t\sigma, M(x)+t\sigma)$, нижняя граница вероятности значений $x-M(x)$, заключающихся в этом промежутке, будет оставаться постоянной. Отсюда опять-таки следует, что чем меньше основное отклонение, тем теснее отдельные значения случайной величины сосредоточиваются около ее математического ожидания.

Таким образом, основное отклонение служит мерой рассеяния значений случайной величины.

После усвоения леммы Маркова и неравенства Чебышева приведем (без доказательства) теорему Маркова.

Она выражается уравнением

$$P = \left\{ -v < \frac{\sum_{s=1}^{n} x_s}{n} - \frac{\sum_{s=1}^{n} a_s}{n} < v \right\} > 1 - w, \qquad (9.10)$$

где v, w – некоторые произвольно заданные положительные числа;

a_s – математическое ожидание.

Теорема Маркова выполняется при условии $\frac{\sigma^2}{n^2} \to 0$; $n \to \infty$.

Теорема Маркова представляет собой наиболее общее выражение закона больших чисел. Ее частным случаем является теорема Чебышева. Она формулируется следующим образом.

Т е о р е м а Ч е б ы ш е в а. Если число n попарно независимых случайных величин

$$x_1, x_2, \ldots, x_n$$

можно увеличивать беспредельно и математические ожидания их квадратов все не превосходят одного и того же постоянного числа, то при достаточно большом числе этих величин будет сколь угодно близкой к достоверности

вероятность того, что их среднее арифметическое отличается произвольно мало от среднего арифметического их математических ожиданий:

$$P\left\{-\varepsilon < \frac{\sum_{s=1}^{n} x_s}{n} - \frac{\sum_{s=1}^{n} a_s}{n} < \varepsilon\right\} > 1-\eta\,. \qquad (9.11)$$

Одним из важных следствий теоремы Чебышева является применение ее к случаю, когда все попарно независимые величины

$$x_1,\ x_2, \ldots,\ x_n$$

имеют одно и то же математическое ожидание, т. е. когда все

$$a_s = a \qquad (s = \overline{1,n})$$

и, кроме того, все

$$M(x_s^2) = b_s = b \qquad (s = \overline{1,n})\,,$$

причем b существует, т.е. конечно. Иначе говоря, независимые величины

$$x_1,\ x_2, \ldots,\ x_n$$

можно рассматривать как значения, полученные в n независимых испытаниях относительно одной и той же случайной величины x. В таком случае, согласно теореме Чебышева имеем

$$P\left\{-\varepsilon < \frac{\sum_{s=1}^{n} x_s}{n} - a < \varepsilon\right\} > 1-\eta\,.$$

Таким образом, из теоремы Чебышева получается как следствие важная т е о р е м а:

Если с величиной x, имеющей конечную дисперсию, производится достаточно большое число независимых испытаний, то с вероятностью, сколь угодно близкой к достоверности, можно ожидать, что среднее арифметическое наблюденных значений величины x будет произвольно мало отличаться от ее математического ожидания.

Теоремы Пуассона и Бернулли мы рассмотрели ранее, когда описывали биноминальное распределение.

Таким образом, рассмотренные теоремы доказывают, что чем больше объем выборки, тем точнее средний результат, т.е. выборочная средняя ($\overline{X}$) в меньшей мере отклоняется от средней арифметической (M) генеральной совокупности, и наоборот, чем меньше выборка, тем меньше и

шансов на то, что выборочная средняя совпадет по величине со средней арифметической генеральной совокупности. Действие этого закона основано на свойстве самих случайных величин, отрицательные и положительные значения которых способны компенсировать друг друга и тем полнее, чем большему числу испытаний подвергается случайная величина. На этом свойстве случайных величин компенсировать друг друга и основана относительная устойчивость средних значений. Поэтому описанные в предыдущей главе закономерности распределения, наблюдаемые в ранжированных совокупностях вариант, следует рассматривать как проявление наиболее общего закона поведения случайных величин - закона больших чисел.

Закон больших чисел утверждает, что практически маловероятно значительное отклонение средней арифметической выборочной совокупности ($\overline{X}$) от средней арифметической генеральной совокупности (M), если число наблюдений достаточно велико.

9.2 Основные задачи статистического оценивания. Смещенные и несмещенные оценки

Мы рассмотрели теоретические аспекты статистического оценивания. Теперь рассмотрим его практическое приложение. К.Е. Никитин и А.З. Швиденко рассматривают оценки в следующей интерпретации.

Статистическое оценивание информации включает три основные задачи: нахождение по выборке наиболее вероятных значений оценок некоторых параметров («точечное» оценивание); оценку интервалов, относительно границ которых с определенной вероятностью можно утверждать, что они заключают неизвестный параметр (интервальное оценивание); проверку справедливости тех или иных утверждений относительно изучаемого явления (проверка статистических гипотез). Эти задачи тесно взаимосвязаны, но можно решать каждую отдельно или все одновременно в зависимости от цели статистического анализа. Например, мы, определив запас древесины на 1 га в древостое дуба, нашли его равным 250 м^3. Эту величину надо оценить следующим образом.

- Насколько точно значение 250 м^3/га (% ошибки) и какие допустимы отклонения от нее в обе стороны: точечное оценивание и оценка интервала.
- Насколько найденная величина соответствует среднему запасу дубовых древостоев в данном возрасте при определенном классе бонитета.

Оценки параметров получают различными методами. Поэтому, естественно, выбирают те, которые дают наилучшие результаты. Для этой цели вводят понятия несмещенности, состоятельности, эффективности и достаточности оценок.

Оценку называют несмещенной, если она не дает систематической ошибки при оценивании некоторого параметра θ, другими словами, если среднее значение оценки, полученное по множеству выборок, практически совпадает с θ. В принципе может быть несколько несмещенных оценок одно-

го и того же параметра; например, в качестве оценки среднего значения генеральной совокупности при некоторых условиях можно взять выборочное среднее или выборочную медиану. В этом случае целесообразно брать оценку с меньшей изменчивостью.

Оценку называют состоятельной, если по мере увеличения объема выборки она все больше приближается к оцениваемому параметру. Состоятельную оценку с наименьшей дисперсией называют эффективной. Наконец, достаточность оценки понимают в том смысле, что не существует другой оценки параметра θ, вычисленной на основании данной выборки и содержащей дополнительную информацию об этом параметре. Достаточность оценки влечет за собой ее эффективность и состоятельность.

Для нахождения оценок с заданными свойствами существует ряд методов, из которых наиболее часто применяют метод максимального правдоподобия и метод моментов. Метод максимального правдоподобия предполагает использование в качестве оценки параметра θ такого значения, которое (по данным выборки) наиболее вероятно с точки зрения возможности замены им θ. Этот метод дает эффективные, но не всегда несмещенные оценки, а для выборок большого объема оценки имеют нормальное распределение.

Метод моментов обеспечивает состоятельные, но не всегда эффективные оценки. В этом методе, использованном выше при аппроксимации выборочных распределений, параметры распределения представляют через моменты. По выборке вычисляют оценки моментов и подставляют их в полученные уравнения, по которым находят неизвестные параметры.

Целесообразность практического использования оценок с теми или иными свойствами должна обсуждаться в контексте конкретных задач. Зачастую «хорошие» свойства оценок совпадают. Так, выборочное среднее всегда является несмещенной и состоятельной оценкой среднего генеральной совокупности, а при нормальности исходной совокупности – еще и достаточной, в то время как медиана в качестве выборочной оценки среднего таковой не является, так как ее дисперсия больше дисперсии выборочного среднего. В некоторых случаях лучше иметь оценку несколько смещенную, но состоятельную; иногда несмещенность можно устранить. Так, при нахождении дисперсии, знаменатель n-1 объясняется именно тем, что выборочная дисперсия $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \overline{X})^2$ является смещенной оценкой дисперсии генеральной совокупности, а сомножитель $n/(n - 1)$ это смещение устраняет.

9.3 Ошибки статистик и их определение. Доверительный вариант

Для практического использования нам надо знать ошибки основных статистик распределения: среднего значения, среднего квадратического отклонения, коэффициента вариации, асимметрии и эксцесса.

В литературе часто встречается выражение «ошибка репрезентативности». Поэтому начнем рассмотрение с нее.

Ошибки репрезентативности. Расхождение между величиной средней арифметической ($\overline{X}$) выборки и величиной средней арифметической генеральной совокупности (M) принято называть ошибкой репрезентативности, т.е. ошибкой, допускаемой не в самом процессе измерительной и вычислительной работы, а в результате случайного отбора вариант из генеральной совокупности при образовании выборки. Репрезентативная ошибка – это не техническая, а статистическая ошибка. Она указывает на величину отклонения выборочной средней ($\overline{X}$) от средней (M) генеральной совокупности.

Величина репрезентативной ошибки определяется по разности между средними величинами выборки и генеральной совокупности, т.е. как ($\overline{X}$)-M. Однако этот показатель в практике использовать невозможно, так как средняя арифметическая генеральной совокупности обычно остается неизвестной. Если же средняя (M) генеральной совокупности известна, то указанная разность ($\overline{X}$-M) теряет свое значение. Поэтому ошибки репрезентативности определяются не прямым, а косвенным путем – через отклонения вариант от выборочной средней.

Ошибка отдельно взятой варианты. Если судить о величине статистической ошибки отдельно взятой варианты, то она равна среднему квадратическому отклонению, так как любое эмпирическое распределение, следующее нормальному закону, практически укладывается в пределах плюс-минус трех сигм, т.е. $\overline{X} \pm 3\sigma$. Ошибку репрезентативности называют поэтому средней квадратической ошибкой, или просто средней ошибкой. Будем ее обозначать через m, указывая при этом и характеристику, которую она сопровождает. Таким образом, средняя квадратическая ошибка отдельно взятой варианты выразится в виде

$$m_{x_i} = \pm\sigma .$$

Среднее квадратическое отклонение имеет двоякое значение: во-первых, оно основное мерило изменчивости, показатель вариабельности признаков, а во-вторых, среднее квадратическое отклонение служит в качестве статистической ошибки отдельно взятой варианты.

Ошибка средней арифметической. Математическая статистика утверждает, что выборочная средняя ($\overline{X}$) отклоняется от своего математического ожидания или средней арифметической (M) генеральной (теоретически рассчитанной) совокупности меньше в $\sqrt{n}$ раз по сравнению с отдельными вариантами данного распределения. Отсюда следует, что средняя квадратическая ошибка выборочной средней ($\overline{X}$) равняется частному от деления среднего квадратического отклонения на корень квадратный из числа всех вариант данной совокупности, т.е.

$$m_x = \frac{\sigma}{\sqrt{n}} \qquad (9.12)$$

Приведем пример. Возьмем распределение числа деревьев по толщине и сделаем нужные нам вычисления (таблица 9.1.).

Таблица 9.1 – Вычисление среднего значения и его ошибки на примере распределения диаметров в сосновом 80-летнем древостое II класса бонитета

Ступени толщины (классовые варианты, x_i)	Число стволов (частоты, n_i)	$x_i \cdot n_i$	$a=x_i-\overline{X}$	a^2	$n_i a^2$
12	4	48	-14	196	784
16	8	128	-10	100	800
20	26	520	-6	36	936
24	43	1032	-2	4	172
28	31	868	+2	4	124
32	22	704	+6	36	792
36	11	396	+10	100	1100
40	4	160	+14	196	784
44	1	44	+18	324	324
Итого	150	3900	-		5816

По данным таблицы 9.1. вычислим статистики распределения и их основные ошибки.

$$\overline{X} = \frac{\sum x_i n_i}{\sum n_i} = \frac{3900}{150} = 26$$

$$\sigma = \sqrt{\frac{\sum x_i a_i^2}{\sum n_i}} = \sqrt{\frac{5816}{150}} = \sqrt{38,77} = 6,23 \approx 6,2$$

$$m_{\overline{X}} = \frac{\sigma}{\sum n_i} = \frac{6,23}{\sqrt{150}} = \frac{6,23}{12,247} = 0,509 \approx 0,51 \approx 0,5$$

Принято записывать среднюю величину вместе с ее основной ошибкой, т.е. $\overline{X} = 26 \pm 0,5$.

Средняя ошибка указывает наиболее вероятные границы, в которых возможны случайные колебания величины средней арифметической в зависимости от объема выборки. При увеличении числа испытаний сред-

няя ошибка уменьшается. Когда число наблюдений неограниченно возрастает, средняя ошибка стремится к нулю, т.е. при $N \to \infty$ и $m \to 0$. Следовательно, средняя ошибка есть мера точности, или относительной достоверности, нашего суждения о возможных колебаниях средних показателей варьирующих величин.

Поскольку весь вариационный ряд нормально распределяющейся случайной величины X практически укладывается в пределах между $\overline{X}+3\sigma$ и $\overline{X}-3\sigma$ на 99,9 %, то можно сказать, что генеральная средняя (M) таких распределений не выходит за пределы утроенного значения средней ошибки средней арифметической любой выборки, взятой из данной генеральной совокупности, т.е. она всегда заключена между пределами от $\overline{X} - 3m_{\bar{x}}$ до $\overline{X} + 3m_{\bar{x}}$ или в пределах $\overline{X} \pm 3m_{\bar{x}}$. Поэтому утроенное значение средней квадратической ошибки называется предельной ошибкой средней арифметической выборочной совокупности. Выражение $\overline{X} \pm 3m_{\bar{x}}$ заключает в себе содержание так называемого "правила утроенной ошибки" и правила трех сигм.

При вычислении ошибки средней арифметической на малых выборках число наблюдений (N) берется с "числом степеней свободы", и формула (9.12) принимает следующий вид:

$$m_{\bar{x}} = \sqrt{\frac{\sum(x-\overline{X})^2}{N(N-1)}} = \frac{\sigma}{\sqrt{N-1}} \qquad (9.13)$$

При большом N разница между N и N-1 несущественная, и формулу (9.13) можно записать как $m_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$ (9.13а)

Когда средняя арифметическая вычисляется прямым способом на материале, не сгруппированном в классы, ошибку можно определить по следующей формуле:

$$m_{\bar{x}} = \sqrt{\frac{\left(\frac{\sum x^2}{N} - \overline{X}^2\right)}{N-1}} \qquad (9.14)$$

Например, имеются следующие десять вариант, выражающих толщину деревьев на опытном участке в культурах сосны возрастом 10 лет:

5, 6, 6, 7, 4, 4, 2, 5, 5, 5, 6

Вычисления здесь будут проведены по схеме, приведенной в таблице 9.2.

Если же средняя арифметическая определяется на несгруппированном в классы материале коротким способом моментов (способом условной средней), то средняя ошибка вычисляется по формуле (9.11):

$$m_{\bar{x}} = \sqrt{\left[\frac{\sum a^2}{N} - \left(\frac{\sum a}{N}\right)^2\right]/(N-1)}\ , \text{ где} \qquad (9.15)$$

$a=x-A$, т.е. отклонение варианты от условной средней. Продемонстрируем применение этой формулы на предыдущем примере (таблица 9.3).

Таблица 9.2 – Вычисление ошибки среднего при несгруппированных вариантах, формула (9.14)

№ п/п x_i	Толщина деревьев, варианты n_i, см	n_i^2	Вычисление
1	5	25	$\bar{x} = \frac{\sum n_i \cdot x_i}{n_i} = \frac{50}{10} = 5\text{см}$ Из уравнения (9.14) $m_{\bar{x}} = \sqrt{(\frac{268}{10} - 25) + 9} = \sqrt{1{,}68 + 9} = \sqrt{10{,}68} = 3{,}27\text{ñì}$ $\bar{\sigma} = \sqrt{\frac{\sum x_i a_i^2}{\sum n_i}} = \sqrt{\frac{268}{50}} = \sqrt{5{,}36} = 2{,}31$
2	6	36	
3	6	36	
4	7	49	
5	4	16	
6	4	16	
7	2	4	
8	5	25	
9	5	25	
10	6	36	
Сумма	50	268	

Таблица 9.3 – Схема вычисления ошибки среднего по формуле (9.15)

№ п/п x_i	Численности n_i	Отклонения от условной средней a	a^2	Вычисление по формуле (9.15)
1	5	+1	1	$\bar{x} = 4 + \frac{10}{10} = 5\text{ñì}$ $m_x = \sqrt{\frac{28}{10} - (\frac{10}{10})^2 + 9} = \sqrt{2{,}8 - 1 + 9} = \sqrt{10{,}8} =$ $= 3{,}29$
2	6	+2	4	
3	6	+2	4	
4	7	+3	9	
5	4=k	0	0	
6	4	0	0	
7	2	-2	4	
8	5	+1	1	
9	5	+1	1	
10	6	+2	4	
Сум-	50	+10	2	

ма			8	

Из приведенных расчетов следует, что полученный результат практически (с учетом округления) идентичен предыдущему.

Ошибка среднего квадратического отклонения. Понятие ошибки репрезентативности относится не только к средней арифметической, но и к другим средним показателям, в частности к среднему квадратическому отклонению, характеризующему варьирование признака в данной совокупности. Средняя ошибка среднего квадратического отклонения вычисляется по формуле

$$m_\sigma = \frac{\sigma}{\sqrt{2N}} \qquad (9.16)$$

$$m_{\sigma^2} = \frac{\sigma^2}{\sqrt{\frac{N}{2}}} \quad \text{или} \quad m_{\sigma^2} = \sigma^2\sqrt{\frac{2}{N}} \qquad (9.16\ а)$$

В нашем примере (таблица 9.1), где N=150, σ=6,2, ошибка m_σ составит

$$m_\sigma = \frac{6,2}{\sqrt{300}} = \frac{6,2}{17,32} = 0,356.$$

Следовательно, m_σ лежит в пределах (с достоверностью 99,9 %) 6,2±3·0,356=6,2±1,068.

Для примера, рассчитанного по таблице 9.2 $m_\sigma = \frac{2,31}{\sqrt{20}} = \frac{2,31}{4,47} = 0,517$.

Тогда пределы m_σ следующие: 2,31±1,55.

Ошибка коэффициента вариации. Средняя ошибка коэффициента вариации (V) определяется по следующей приближенной формуле:

$$m_V = \frac{V}{\sqrt{2N}} \cdot \sqrt{1+\left(\frac{V}{100}\right)^2} \qquad (9.17)$$

В формулу (9.17) можно ьакже представить в следующем виде

$$m_v = V\sqrt{\frac{0.5 + 0.0001V^2}{N}} \qquad (9.17\ а)$$

Вычислим по данным таблицы 9.1 коэффициент вариации и его ошибку.

$$V = \frac{\sigma}{\overline{X}} \cdot 100\% = \frac{6,2}{26} \cdot 100\% = 23,8\% \approx 24\%$$

Тогда по формуле (9.17)

$$m_V = \frac{23,8}{\sqrt{300}} \cdot \sqrt{1 + \left(\frac{23,8}{100}\right)^2} = \frac{23,8}{17,32} \cdot \sqrt{1 + 0,113} = 1,37 \cdot 1,05 = 1,44$$

По формуле (9.17а) $m_v = 23.8\sqrt{\frac{0.5 \cdot 0.0001 \cdot 23.8^2}{150}} = 23.8 \cdot 0.0609 = 1.449$.

Значения m_v по обоим формулам с учетом округления практически одинаковы.

Следовательно, коэффициент вариации генеральной совокупности, к которой принадлежит наша выборка не выйдет за пределы m_V =23,8±4,32.

Средняя ошибка доли. В отношении качественных признаков, когда средняя арифметическая показывает относительную численность одной из альтернатив и выражается либо в абсолютных значениях, либо в долях единицы или в процентах, средняя ошибка выражается в тех же значениях, что и альтернативы. Так, если признак выражен в абсолютных значениях, то средняя ошибка, называемая ошибкой относительной частоты, равняется:

$$m_D = \pm \sqrt{\frac{p(N - p)}{N}} \qquad (9.18)$$

Здесь p - частота одной из альтернатив; N - общее число наблюдений, т.е. $p_1+p_2=N$.

Например, при посадке 100 семян березы карельской получено 44 дерева карельской березы и 56 деревьев березы повислой (обычной). Определим среднюю ошибку этого отношения:

$$m_D = \sqrt{\frac{44 \cdot 56}{100}} = 4,96 \approx 5 .$$

Можно сказать, что среди потомства карельской березы есть 44±5 особей.

Если альтернативы выражаются долями единицы, то ошибка относительной частоты определяется по аналогичной формуле (9.19).

$$m = \sqrt{\frac{p(1 - p)}{N}} \qquad (9.19)$$

Формула достаточно ясна и без числовых примеров.

Если численность одной из альтернатив близка к нулю, то среднюю ошибку относительной частоты можно определить из отношения

$$m=\frac{N}{N+1}, \quad \text{или} \quad m=\frac{100}{N+1}\%,$$

когда альтернативы выражены в процентах.

Ошибки показателей асимметрии и эксцесса. Средняя ошибка показателя асимметрии определяется по следующей формуле:

$$m_{as} = \sqrt{\frac{6}{N}}, \tag{9.20}$$

или более точно по формуле

$$m_{as} = \sqrt{\frac{6N(N-1)}{(N-2)(N+1)(N+3)}} \tag{9.21}$$

Ошибку коэффициента эксцесса можно вычислить по следующим аналогичным формулам:

$$m_{Ex} = 2\sqrt{\frac{6}{N}} \quad \text{или} \quad m_{Ex} = \sqrt{\frac{24}{N}} \tag{9.22}$$

или по более точной формуле

$$m_{Ex} = \sqrt{\frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)}} \tag{9.23}$$

Определим ошибки асимметрии и эксцесса, использовав ранее рассмотренный пример (таблица 8.2) ряда распределения диаметров сосны.

Вспомним, что по данным таблицы 8.2 мы получили α=0,203; E=-0,55. Число стволов в нашем примере составило 200.

Тогда

$$m_{as} = \sqrt{\frac{6}{200}} = \sqrt{0,03} = 0,173$$

или более точно

$$m_{as} = \sqrt{\frac{6 \cdot 200 \cdot 199}{198 \cdot 201 \cdot 203}} = \sqrt{\frac{238800}{8078994}} = \sqrt{0,0296} = 0,172.$$

Результаты получились практически одинаковые. Поэтому при достаточно большом N>30-40 рациональнее использовать формулу (9.16).

Ошибка эксцесса для нашего примера равна (формула (9.18))

$$m_{Ex} = 2\sqrt{\frac{6}{200}} = 0{,}343\,.$$

По уточненной формуле (9.19)

$$m_{Ex} = 2\sqrt{\frac{6\cdot 200\cdot 199^2}{197\cdot 198\cdot 203\cdot 205}} = 2\sqrt{\frac{47521200}{162323470}} = 2\sqrt{0{,}293} = 0{,}342\,.$$

Результат получился подобным тому, что видели при вычислении $m_{\alpha s}$, т.е. при N>30 лучше пользоваться формулой (9.18).

Таким образом, повторим все сказанное об ошибках статистик. Они определяются по формулам

$$m_{\bar{x}} = \frac{\sigma}{\sqrt{N}}; \quad \text{а при } N<30 \quad m_{\bar{x}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N(N-1)}} = \frac{\sigma}{\sqrt{N-1}};$$

или при больших N $\qquad m_x = \dfrac{\sigma}{\sqrt{N}}$

$$m_{\sigma^2} = \sigma^2\sqrt{\frac{2}{N}} \quad \text{или} \quad m_{\sigma^2} = \frac{\sigma^2}{\sqrt{\frac{N}{2}}}; \qquad m_\sigma = \frac{\sigma}{\sqrt{2n}};$$

$$m_v = \frac{V}{\sqrt{2n}}\left\{1 + (\frac{V}{100})^2\right\}^{\frac{1}{2}} \quad \text{или} \quad m_v = V\sqrt{\frac{0.5 + 0.0001V^2}{N}};$$

$$m_\alpha = \sqrt{\frac{6}{N}}; \qquad m_E = 2\sqrt{\frac{6}{N}} = 2m_\alpha\,.$$

Понятие об ошибках статистик тесно связано с величиной доверительного интервала. Ранее мы уже упоминали доверительные интервалы в 1σ(0,68), 2σ(0,95), 3σ(0,999). Здесь дадим более обстоятельное описание доверительных интервалов.

Доверительные вероятности и уровни значимости. В математической статистике, а также и в биометрии принято существенность того или иного результата оценивать по значению трех вероятностей, близких к достоверности: P_1=0,95, или 95%, P_2=0,99, или 99%, P_3=0,999, или 99,9%. Эти вероятности получили название *доверительных*. Вероятности же, которыми решено пренебрегать, т.е. P_1=0,05, или 5%, P_2=0,01, или 1%, и P_3=0,001, или 0,1%, получили название уровней *значимости*, или уровней *существенности*. И те и другие вероятности обозначаются символами, как $P_{0,95}$ или $P_{0,05}$ и т.д.

Каждой доверительной вероятности соответствует определенное значение нормированного отклонения (t):

вероятности	P_1=0,95	соответствует	t_1=1,96
-″-	P_2=0,99	-″-	t_2=2,50
-″-	P_3=0,999	-″-	t_3=3,30

Доверительный интервал и границы доверия. Выше было сказано, что величина ошибки выборочной средней определяется по разности между этой средней ($\overline{X}$) и средней генеральной совокупности (M), т.е. как $\overline{X}-M$. Можно ли по эмпирическим данным определить наиболее вероятные границы, в которых находится средняя (M) генеральной совокупности? Математическая статистика дает на этот вопрос положительный ответ. Интервал, в котором с заданной вероятностью или уровнем значимости заключена средняя арифметическая генеральной совокупности, называется доверительным интервалом. Границы этого интервала получили название доверительных границ, или границ доверия. Как же определить доверительный интервал и его границы? Это достигается нормированием отклонения варианты, или выборочной средней, от средней генеральной совокупности. Так, если взять нормированное отклонение варианты от выборочной средней

$$t=\frac{x_i-\overline{X}}{\sigma},$$

то можно преобразовать его следующим образом: $x_i-\overline{X}=t\sigma$. Аналогично отклонение выборочной средней $\overline{X}$ от средней генеральной совокупности (M) выражается через

$$\overline{X}-M=t\frac{\sigma}{\sqrt{N}},\ \text{или}\quad \overline{X}-M=t\,m_{\overline{X}}.$$

Величина этого отклонения зависит, следовательно, от степени вариабельности признака, а также от уровня вероятности, с которой определяется доверительный интервал. Заменив в этом уравнении знаки на обратные и переставив $\overline{X}$ в правую часть уравнения получим

$$M=\overline{X}-tm_{\overline{X}}.$$

А так как $\overline{X}$ может быть и больше и меньше M, то указанное выражение можно написать в таком виде: $M=\overline{X}\pm tm_{\overline{X}}$. Отсюда доверительный интервал для средней арифметической генеральной совокупности выразится следующим неравенством:

$$\overline{X}-tm_{\overline{X}}\le M\le \overline{X}+tm_{\overline{X}},$$

где $\overline{X}-tm_{\overline{X}}$ и $\overline{X}+tm_{\overline{X}}$ - границы доверительного интервала.

Из изложенного вытекает много важных практических приложений для лесного хозяйства. Например, требуется, чтобы материально-денежная оцен-

ка лесосек проводилась достаточно точно на каждой из протаксированных делянок, т.е. достоверность здесь должна равняться 3σ.

Приемлемая технология таксации лесосек обеспечивает достоверность в 3σ при точности в ±10%. Правда, 1-2 лесосека из 1000 могут не вложиться в точность ±10%, но на это пришлось пойти, так как более высокая точность учета требует иной, гораздо более дорогой технологии, и экономически не оправдана.

Но уже совокупность 3-5 лесосек при отсутствии систематической ошибки должна таксироваться с точностью 5-6%, а при наличии 10 лесосек точность должна составлять 3-4% и т.д.

9.4 Ошибка суммы или разности средних значений

Часто нам надо знать, различаются ли между собой средние величины. Например, спустя *n* лет после внесения удобрений средний диаметр опытного древостоя составил 24 см, а контрольного – 22 см. Возникает вопрос, насколько эта разница существенна и является результатом проведенных мероприятий или зависит от случайных причин. Для оценки таких различий используют t-критерий Стьюдента.

t-распределение Стьюдента. Прежде чем приступить к рассмотрению вопросов, связанных с методикой оценки достоверности различий, наблюдаемых между выборочными средними, необходимо рассмотреть еще одну сторону нормированного отклонения, имеющую прямое отношение к статистике малой выборки. В данном случае имеется в виду закон распределения величин нормированного отклонения выборочной средней ($\overline{X}$) от средней арифметической генеральной совокупности (M), открытый английским математиком Вильямом Госсетом в 1908 году. Этот ученый печатался под псевдонимом Стьюдент (Student). Стьюдент установил, что вероятность нормированного отклонения $\overline{X}$-M: $\sigma = t$ выражается следующим уравнением:

$$P(t) = C\left(1+\frac{t^2}{N-1}\right)^{-\frac{1}{2}N},$$

которое носит название распределения Стьюдента. Здесь $P(t)$ обозначает вероятность указанного нормированного отклонения, а C – некоторый множитель, зависящий лишь от объема выборки (N).

В практике (при независимых x_i) для оценки достоверности разницы между средними используют следующие приемы.

Так, если x_i независимы, то, например, для $y=x_1-x_2$ имеем

$$m_y = m_{x_1-x_2} = \sqrt{m_{x_1}^2 + m_{x_2}^2}\ . \qquad (9.24)$$

Выражение (9.24) – основная ошибка разности двух случайных величин. Этот показатель можно применять для оценки значимости различия между средними двух выборок, например, для суждения о том, можно ли считать, что данные выборки принадлежат к одной генеральной совокупности. Для этой цели вычисляют величину

$$t^* = \frac{|\tilde{x}_1 - \tilde{x}_2|}{m_{x_1 - x_2}} = \frac{|\tilde{x}_1 - \tilde{x}_2|}{\sqrt{m_{x_1}^2 + m_{x_2}^2}}. \qquad (9.25)$$

Если t*>2, то с вероятностью 0,95, а при t*>3 с вероятностью, практически не отличающейся от 1, можно утверждать, что различие между средними значимо.

Приведем пример практического использования t-критерия Стьюдента в лесохозяйственной практике.

Пусть на некотором участке лесных культур сосны возрастом 30 лет II класса бонитета внесли минеральные удобрения. Через 10 лет измерили опытный и контрольный участок, которые 10 лет назад были аналогичны, т.е. имели средние диаметры 10,9 (контроль) и 11 см (опыт), а через десять лет соответственно 14,0 и 15,1 см. Ошибка среднего значения составила в 40 лет 0,3 и 0,4 см. Используя формулу (9.25) найдем

$$t = \frac{15{,}1 - 14{,}0}{\sqrt{0{,}3^2 + 0{,}4^2}} \frac{1{,}1}{\sqrt{0{,}09 + 0{,}16}} = \frac{1{,}1}{\sqrt{0{,}25}} = \frac{1{,}1}{0{,}5} = 2{,}2.$$

Таким образом, с вероятностью 0,95 мы можем утверждать, что удобрения дали положительный эффект.

В практике лесного хозяйства и, особенно при проведении научных исследований, t-критерий применяется повсеместно. Критические значения t-критерия Стьюдента при определенном числе степеней свободы для разных уровней вероятности приведены в приложении Е.

10. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

10.1. Статистические гипотезы. Простые и сложные гипотезы
10.2. Параметрические методы оценки гипотез
10.3. Непараметрические методы оценки гипотез
10.4. Проверка статистических гипотез в практике лесного хозяйства и их использование а в лесном хозяйстве

10.1 Статистические гипотезы. Простые и сложные гипотезы

Гипотеза – это научно обоснованное предположение о вероятности некоторого события, явления, закона и т.д. Гипотезой может считаться не любое предположение, а только такое, которое имеет некоторое научное обоснование, хотя еще недостаточно доказанное и проверенное. Поэтому гипотеза может как подтвердиться, так и быть отвергнутой. Но предварительное научное обоснование сделать надо, чтобы не выполнять явно ненужную работу, проверяя предположения, которые ни на чем не основаны.

Гипотеза, которая находит подтверждение и обоснование, перерастает в теорию, закон, закономерность и т.д. Примером подтвердившихся гипотез, ставших теориями и законами, служит периодическая система элементов Д.И. Менделеева (1834-1907), атомное строение материи, строение атома, наличие кварков. В лесном хозяйстве теория строения древостоев выдвинутая в 19 веке, являлась лишь гипотезой, которая затем подтвердилась работой известных лесоводов А.В. Тюрина (1882-1979), В.К. Захарова (1882-1966), Н.В. Третьякова (1880-1957), Ф.П. Моисеенко (1894-1979) и других.

В настоящем пособии мы будем рассматривать не все гипотезы, а только статистические, т.е. относящиеся к области математической статистики. Для их проверки существует стандартная процедура. Для того, чтобы ее пояснить, возьмем простой пример с появлением перед наблюдателем в крупном городе мужчин или женщин. Мы уже видели, что вероятности появления лиц одного пола равны и составляют $P_м = P_ж = 0,5$. Такое соотношение будет наблюдаться, если общее число лиц, прошедших мимо наблюдателя N, достаточно велико > 30, а лучше > 100, а в соотношении мужчин и женщин нет асимметрии, т.е. соблюдена симметрия, что обычно для крупных белорусских городов. Допустим, что такая симметрия нарушена, скажем проводим наблюдение в военном городке или в российском «городе невест» - Иваново. В этом случае соотношение полов нарушится, что будет свидетельствовать о некорректности опыта.

При анализе гипотез в начале необходимо остановиться на следующем важном факте: результат эксперимента, подтверждающий справедливость выдвинутой гипотезы, почти никогда не может служить основанием для принятия этой гипотезы. В то же время результат, несовместный с выдвинутой гипотезой, вполне достаточен для отклонения ее как ложной. Ко-

нечно, ошибки всегда могут иметь место, однако высказанное положение является важным и требует тщательного обоснования. Причина того, что результат, подтверждающий выдвинутую гипотезу, не обязательно может являться основанием для ее принятия, состоит в том, что данный результат может быть совместным также и с другими гипотезами и, следовательно, не обязательно может служить доказательством справедливости данной гипотезы против других выдвинутых альтернатив. Например, встреча 53 мужчин из 100, встреченных лиц обоих полов совместно с гипотезой о том, что количество мужчин и женщин в городе примерно одинаково. Этот результат совместен и с предположением, что мужчин в городе не намного, но больше. Таким образом, результат, совместный с первоначально выдвинутой гипотезой, не является стопроцентным доказательством ее достоверности. Даже встреча 50 мужчин не может служить основанием для заключения, что в городе имеется строго одинаковое количество лиц обоих полов. С другой стороны, встреча 90 мужчин при 100 прошедших прохожих могло бы на практике служить для опровержения гипотезы об одинаковом количестве мужчин и женщин в исследуемом населенном пункте, скажем в вахтовом поселке нефтяников в Сибири.

В следующем примере предположим, что средний коэффициент умственного развития некоторой совокупности людей составляет 100. Результат выборки, показавший, что средний показатель равен 102, совместен с выдвинутой нами гипотезой. Однако этот результат совместен также и с предположением, что средний коэффициент умственного развития равен 101 или 99, и, конечно же, совместен с гипотезой о том, что средний показатель исходной совокупности составляет 102. Следовательно, данный результат никоим образом не может являться свидетельством предпочтительности гипотезы, в соответствии с которой средний коэффициент равен 100. Допустим теперь, что некоторая выборка дала средний коэффициент умственного развития, равный 135. Предположив, что объем выборки был достаточно большим, мы могли бы показать следующее: если исходная гипотеза является достоверной, то мы практически никогда не получили бы подобного результата. На основании этого вывода полученный результат вполне обоснованно может быть использован нами в качестве свидетельства ложности выдвинутой гипотезы, причем риск совершения ошибки в данном случае был бы минимальным.

Все вышесказанное основывается на уже высказанном факте, что обычно результат эксперимента, совместный с выдвинутой гипотезой, оказывается также совместным и с рядом других гипотез. В итоге подобный результат не может быть принят в качестве обоснования предпочтительности некоторой гипотезы перед другими гипотезами. Однако мы всегда можем получить расходящийся с выдвинутой гипотезой результат, который может вызвать существенные сомнения в ее достоверности. Гипотезу можно сравнить с показаниями обвиняемого в суде. Он не может доказать истинности своих слов. В то же время некоторые

приведенные им факты оставляют открытой возможность для выдвижения предположений о том, что он мог действовать не так, как описывал. И прокурор может подвергнуть сомнению точность его рассказа, показав, что можно интерпретировать приведенные факты по-другому.

В то же время для доказательства виновности обвиняемого должны быть предоставлены неопровержимые доказательства. Во всех цивилизованных странах действует принцип «презумпции невиновности». Это значит, что все сомнительные гипотезы трактуются в пользу обвиняемого с целью исключения осуждения невиновно, т.е. соблюдается правило, что для общества менее вредно не осудить виновного, чем наказать невиновного. К сожалению этот принцип в нашей стране долгое время (с 20 и до середины 50-х годов 20 века) нарушался, а господствовала презумпция виновности, что привело к тяжелым последствиям и гибели миллионов граждан.

В биометрии для доказательства некоторого утверждения часто применяют метод, известный в математике, как «доказательство от противного». Для этого в качестве рабочего инструмента используют так называемую «нулевую гипотезу». Поясним ее суть.

Нулевая гипотеза. Когда мы не в состоянии отвергнуть гипотезу, мы тем самым признаем, что эта гипотеза может оказаться верной. С другой стороны, если мы можем отвергнуть выдвинутую гипотезу, то тем самым делаем вполне определенный вывод о ее ложности.

Последнее положение является очень важным. При проверке гипотез мы можем сделать окончательный вывод только в случае, когда в состоянии отвергнуть выдвинутую гипотезу. Следовательно, цель проводимого нами эксперимента должна заключаться в *опровержении* проверяемой гипотезы. Это означает, что в качестве гипотезы мы должны сформулировать предположение, *альтернативное* тому, во что верим.

Например, если надо показать, что деревья дуба в целом выше, чем деревья граба, то выдвинем гипотезу об *отсутствии различий* в их росте. Затем попробуем отвергнуть эту гипотезу. В другом случае чтобы доказать, что между анатомическим строением древесины дуба и березы есть существенные различия, нужно подвергнуть проверке гипотезу, что между ними *не существует различий*. И вновь мы должны попытаться отвергнуть эту последнюю гипотезу, чтобы тем самым установить истинность нашего исходного предположения.

О п р е д е л е н и е . *Гипотеза, в соответствии с которой отсутствуют различия между различными совокупностями, называется нулевой гипотезой.*

Гипотеза, которую мы будем в состоянии проверить, не может быть сформулирована на основе любого суждения. Мы это уже наблюдали на примере встречи с лицами разных полов. Суждение о том, что мужчин и женщин в городе имеется одинаковое количество при встрече 50 мужчин и столько же женщин, недостаточно, чтобы на его основе можно было сформулировать некоторую определенную гипотезу. Подобные случаи обычны в практике исследований. Из сказанного следует, что экспериментатор дол-

жен сформулировать альтернативу тому, что он пытается доказать, в виде вполне определенной гипотезы. Только в случае, если это возможно, он может попытаться отвергнуть ее с тем, чтобы доказать справедливость своих исходных предположений.

Таким образом, первый шаг, предпринятый экспериментатором, должен состоять в формулировке статистической гипотезы, которую он надеется *опровергнуть* с тем, чтобы показать истинность своего исходного предположения. После этого он будет в состоянии применить процедуру проверки гипотезы.

В биометрии (статистике) применяются достаточно конкретные гипотезы, связанные с проведением числовых вычислений. Из этого следует, что понятие статистической гипотезы уже, чем понятие научной гипотезы вообще, и предполагает возможность статистического эксперимента для объективного подтверждения (или отклонения) рассматриваемого предположения. Иначе говоря, статистические гипотезы относятся к статистическим моделям.

Примерами гипотез такого рода являются предположения относительно параметров распределения - среднего, дисперсии и т.д. (параметрические гипотезы), либо относительно типа распределения или связи - непараметрические гипотезы. Так, параметрическими гипотезами являются утверждения: среднее значение в некоторой генеральной совокупности равно числу a (обозначается $H_0 : \overline{X} = a$), среднее и (или) дисперсии двух выборок равны (не равны) между собой ($H_0 : \overline{X}_1 = \overline{X}_2$, $H_0 : \sigma_1^2 = \sigma_2^2$ и т.д.). Непараметрическая гипотеза, например, одна из следующих: распределение диаметра данного древостоя подчиняется нормальному закону, рост древостоя в высоту есть экспоненциальная кривая и т.д. Гипотезы называют простыми, если они относятся к конкретному значению параметра (числу); сложные гипотезы представляют объединение простых.

На основании эксперимента, т.е. выборочных данных, решают вопрос: принять или отвергнуть гипотезы, т.е. свидетельствуют полученные данные “за” или “против” испытуемой гипотезы. Для решения этого вопроса мало рассматривать только проверяемую гипотезу H_0; необходимо знать и область “против” гипотезы H_0 - некоторую (их может быть несколько) исключающую ее альтернативную гипотезу H_a.

Для проверки необходимо выбрать статистическую характеристику критерия - показатель, разделяющий зоны, каждая из которых свидетельствует в пользу гипотезы H_0 или H_a. Если речь идет о параметрических гипотезах, то в качестве статистической характеристики обычно используют определенные значения рассматриваемого параметра, а основой для заключений служит распределение статистики, оценивающей данный параметр. Поэтому проверка гипотез теснейшим образом связана с интервальным оцениванием, но позволяет делать более глубокие заключения.

Рассмотрим в качестве примера гипотезу о том, что количество осадков в мае – июле влияет на текущий прирост древостоев. Для этого рас-

смотрим текущий прирост за разные годы с разным количеством осадков, выпавших за исследуемый период, например, 200 мм и 600 мм. Мы считаем, что при 600 мм осадков прирост будет выше. Проверке подлежит утверждение, что среднее значение текущего прироста (обусловленного влиянием осадков) в генеральной совокупности равно нулю, т.е. испытуемой является гипотеза $H_0: \overline{X} = 0$ против альтернативной $H_a: \overline{X} \neq 0$. Выбор такой альтернативы говорит о том, что нас интересуют как положительные отклонения от проверяемой гипотезы (осадки в количестве 600 мм увеличивают прирост), так и отрицательные. В этом случае проверку гипотезы называют двусторонней. Если бы нас интересовали отклонения в одну сторону - только положительные (тогда альтернативная гипотеза $H_a: \overline{X} > 0$) или отрицательные ($H_a: \overline{X} < 0$), то проверка была бы односторонней. В данном случае испытуемая гипотеза простая, а все три альтернативные - сложные.

Основные идеи проверки гипотез рассмотрим на примере среднего значения. Пусть проверяется гипотеза $H_0: \overline{X} = X_0$ и распределение статистики $\widetilde{X}$ известно (рисунок 10.1).

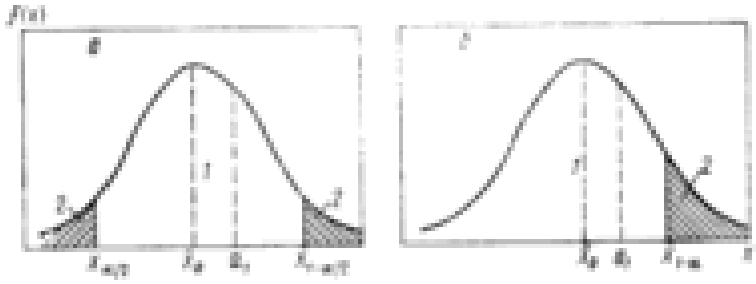


Рисунок 10.1 Зоны принятия (1) и отклонения (2) гипотез при уровне значимости α:
а - двусторонняя проверка; б - односторонняя

На основании выборки получено выборочное среднее $\widetilde{X}=a_1$. Если a_1 не отличается сильно от X_0, то естественно считать, что экспериментальные данные не противоречат проверяемой гипотезе, в противном случае ее отклоняют. В оценку величины различия вкладывается более конкретный смысл, а именно: еще до получения выборки задаются некоторой вероятностью α, которая делит распределение статистики на две зоны. В первую относят все те значения статистики, которые признаются практически возможными (область допустимых значений), в другую - те значения статистики, появление которых в отдельном испытании (на основе одной выборки) признаются практически невозможными при условии, что проверяемая гипотеза верна. Поэтому величина α должна быть достаточно мала, например 0,05 и 0,01. При двусторонней проверке критическая область имеет вид

$$p(|\tilde{X}|>x_{1-\alpha/2}) = p(\tilde{X}<x_{\alpha/2}) + p(\tilde{X}>x_{1-\alpha/2}) = \alpha/2+\alpha/2=\alpha, \qquad (10.1)$$

а при левосторонней и правосторонней односторонних проверках соответственно

$$p(\tilde{X}<x_{\alpha}) = \alpha; \qquad p(\tilde{X}>x_{1-\alpha}) = \alpha; \qquad (10.2)$$

где $\tilde{X}$ - выборочное значение статистики;

x_{α} и $x_{1-\alpha}$ - соответствующие квантили распределения данной статистики (статистические характеристики критерия).

Далее вычисляют конкретное выборочное значение $\tilde{X}$. Если оно попадает в область допустимых значений - гипотеза не отклоняется, если в критическую - гипотеза отклоняется, в связи с чем эти две области называют соответственно **областью принятия и непринятия гипотез.**

Число α называют уровнем значимости критерия. От его величины зависит решение относительно испытуемой гипотезы. Если гипотеза H_0 верна, то α дает нам вероятность того, что статистика попадет в критическую область, и правильная гипотеза ошибочно будет отвергнута: при α=0,001 - в одном случае из 1000, при α=0,05 - в пяти случаях из 100 и т.д. Следует различать уровни значимости (α), уровень достоверности (p) и доверительный коэффициент (t). Между ними есть тесная связь, которая видна из таблицы 10.1

Таблица 10. 1 – Соотношение между различными критериями оценки статистических величин

Уровень значимости, α	Уровень достоверности, P	Доверительный коэффициент, t
32%	68%	1,00
5%	95%	1,96
1%	99%	2,58
0,1%	99,9%	3,39

Дадим пояснение к таблице 10.1. Уровень значимости (α) – это значение вероятности, которое показывает, что различия между средними значениями можно считать несущественными.

Уровень достоверности (P) – это случайная величина, для которой известен закон ее распределения. Обычно используют его критические значения для определенного уровня значимости (α) и числа степеней свободы (γ). Например, t = I - критическое значение t-критерия Стьюдента.

Методы оценки достоверности подразделяются на параметрические и непараметрические, о чем речь пойдет ниже.

Чем меньше уровень значимости, тем меньше вероятность ошибочного отклонения правильной гипотезы. Однако уменьшение величины уровня значимости не всегда целесообразно. Если нулевая гипотеза неверна (например, среднее генеральной совокупности в действительности отличается от X_0), то с уменьшением α уменьшается критическая область и увеличивается область допустимых значений, т.е. статистика при очень малых α попадет в область допустимых значений проверяемой гипотезы H_0 : $\overline{X} = X_0$ и последняя не отклоняется, являясь в действительности ложной. Поэтому мало проверить гипотезу H_0, одновременно нужно испытывать альтернативную гипотезу H_a. Только в таком случае можно оценить риск отклонения гипотезы, когда она верна, или принятия гипотезы, когда она неверна, а верна альтернативная.

Итак, при оценке гипотез возможны ошибки двух типов:

1) гипотеза верна, но отвергается; вероятность этой ошибки дается уровнем значимости и равна α. Величина $1-\alpha$ дает нам вероятность принять гипотезу, если гипотеза верна;

2) гипотеза не верна, но принимается; если обозначить вероятность ошибки второго рода β, то $1-\beta$ (мощность критерия) есть вероятность отклонить гипотезу, если она не верна, а верна альтернативная.

Соотношение между ошибками первого и второго рода иллюстрирует рисунок 10.2 применительно к среднему значению $\overline{X}$.

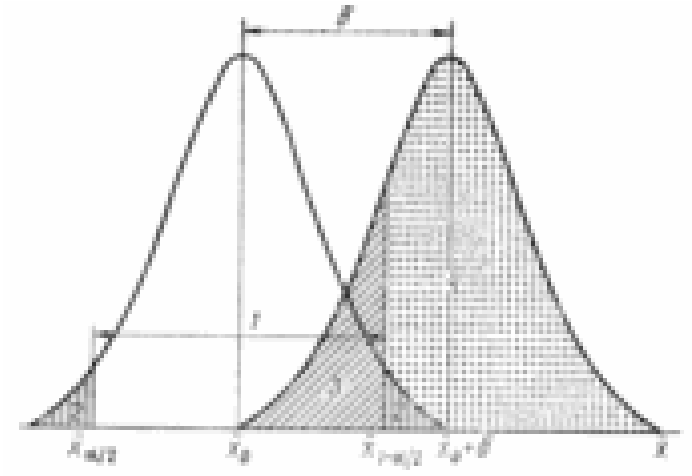


Рисунок 10.2 Ошибки первого и второго рода
1 - зона принятия H_0; 2 - вероятность ошибки 1-го рода;
3 - вероятность ошибки 2-го рода; 4 - мощность критерия.

Если гипотеза верна, то площадь $1-\alpha$ дает вероятность принять гипотезу H_0, $\alpha=\alpha/2+\alpha/2$ - уровень значимости или вероятность ошибки 1-го рода. Пусть в действительности среднее генеральной совокупности

равно $\overline{X}+\delta$. Кривая распределения статистики $\tilde{X}$ не изменяется, но центр распределения сдвигается на величину δ. Тогда заштрихованная площадь β, соответствующая области допустимых значений проверяемой гипотезы, даст вероятность принять гипотезу $H_0: \overline{X} = X_0$, в то время как в действительности верна гипотеза $H_a: \overline{X} = X_0+\delta$, а площадь $1-\beta$ дает величину мощности критерия.

Критическую область для проверяемой гипотезы выбирают так, чтобы обеспечивалась максимальная мощность используемого критерия; в таком случае при заданном уровне значимости α гарантирована минимальная вероятность ошибки 2-го рода β. Критерии, для которых обеспечивается это условие, называют наиболее мощными.

Из рисунка 10.2 видно, что мощность критерия при прочих равных условиях есть функция δ, а величины α и β взаимосвязаны; для выборки фиксированного объема N (от которого также зависит распределение данной статистики) уменьшение вероятности одной из ошибок и ведет к увеличению вероятности другой.

Уменьшить одновременно обе вероятности можно только путем увеличения объема выборки. Но это связано с техническими и экономическими проблемами и потому не всегда возможно. Поэтому с учетом того, что практические последствия ошибок 1 и 2-го рода неодинаковы, поступают следующим образом. Если, например, практический риск, связанный с ошибками 1-го рода, больше, чем риск, связанный с ошибками 2-го рода, то следует уменьшить α за счет увеличения $1-\beta$. Если же представляется возможность оценить последствия ошибочных решений численно (в кубометрах древесины, в стоимостном выражении или как-нибудь иначе), то соотношение величин α и β может быть установлено на этой основе.

Так, при проверке гипотезы о влиянии удобрений на прирост древостоев ошибка 1-го рода приводит к отклонению гипотезы о том, что удобрения не влияют на увеличение текущего прироста, хотя в действительности это может быть не так, т. е. влияния нет или оно незначительно. Эта ошибка влечет за собой неоправданные затраты на внесение удобрений, которые не увеличивают продуктивности древостоев. Ошибка 2-го рода приводит к потере некоторого дополнительного количества древесины.

Очевидно, что последствия ошибки 1-го рода более существенны. Соотношение ошибок 1 и 2-го рода можно оценить с учетом, с одной стороны, стоимости удобрений и затрат на их внесение и, с другой, - стоимости дополнительно полученной древесины, с учетом того, что эта древесина может быть использована лишь через А лет, а А достигает и 30, и 60 лет.

Конечно, далеко не всегда стоимость определяет уровень допускаемых ошибок. Приведем несколько отвлеченный пример. Если в системе ПВО проверяют гипотезу о наличии в зоне обороны вражеской ракеты, то ошибка 1-го рода приведет к пропуску ракеты к цели, а ошибка 2-го

рода - к объявлению ложной тревоги, и обе ошибки нежелательны, но следствия первой ошибки более значимы.

10.2 Параметрические методы оценки гипотез

Параметрические методы оценки достоверности статистических гипотез базируются на основе анализа некоторых параметров выборочной совокупности. Для применения таких оценок вычисляют среднее значение ($\overline{X}$), среднеквадратическое отклонение (σ) или дисперсию (σ^2).

Наиболее часто употребляемым методом параметрической оценки является уже упомянутый выше критерий Стьюдента. Этот критерий всегда обозначается латинской буквой t, в интерпретации автора критерия.

Стьюдент установил, что закон распределения случайной величины зависит от объема выборки и основного отклонения. Опуская достаточно сложные описания распределения вероятностей, которые вывел Стьюдент, т.к. это выходит за пределы относительно небольшого курса лесной биометрии, значение t можно определить по формуле

$t = \frac{\overline{X} - M}{\sigma} * \sqrt{N}$, где

$\overline{X}$ - среднее значение выборочной совокупности

σ - стандартное (среднеквадратическое) отклонение

M – среднее значение в генеральной совокупности.

N – объем ряда распределения.

В практике наибольшее значение t-критерия Стьюдента выбирают из специальных таблиц (приложение Е), где критическая величина t-критерия определяется уровнем значимости (P) и числом степеней свободы υ, где υ = $N_1 + N_2 - 2$, где N_1 и N_2 - величина выборок.

Критерий Стьюдента (t) используют для сравнения существенности разницы между средними значениями двух выборок в следующих вариантах

1. При $N_1 = N_2$

$$|t| = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{m_1^2 + m_2^2}}$$

2. При $N_1 \neq N_2$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma^2 * \frac{N_1 + N_2}{N_1 - N_2}}}, \text{ где}$$

$$\sigma^2 = \frac{\sum (X_i^1 - \overline{X}_1)^2 + \sum (X_i^2 - \overline{X}_2)^2}{N_1 + N_2 - 2};$$

N_1,N_2 – объем соответствующих выборок;

m_1 и m_2 – основные ошибки средних значений ($\overline{X}_1$ и $\overline{X}_2$) исследуемых двух выборок;

σ^2 – объединенная дисперсия двух выборок.

$$m_{\bar{X}_i} = \frac{\sigma_i}{\sqrt{N}}$$

Интересующиеся описанием уравнений, выведенных Стьюдентом, могут найти их в книге Митропольского А. К. «Техника статистических вычислений», а также в других пособиях, например, М. П. Горошко, С. И. Миклуш, П.Г. Хамюк «Біаметрія», которые приведены в списке литературы.

Наряду с проверкой нулевой гипотезы равенства средних величин в генеральной совокупности выполняется проверка равенства среднеквадратического отклонения (σ) и коэффициента вариации (V). Это вызвано тем, что при $\bar{X}_1 = \bar{X}_2$ σ_1 и σ_2, а, соответственно V_1 и V_2 могут отличаться и целесообразно знать, насколько эти отличия существенны.

Для этого используют следующие формулы

$$|t| = \frac{\sigma_1 - \sigma_2}{\sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}}}$$

$$|t| = \frac{V_1 - V_2}{\sqrt{\frac{V_1^2}{N} + \frac{V_2^2}{N}}}.$$

Критические значения t при заданном уровне значимости берут из уже упомянутых специальных таблиц (приложение Е) для υ степеней свободы, где $\upsilon = N_1 + N_2 - 2$. Сравнивая вычисленные и табличные величины t выбираем нулевую (H_0) (нулевую) или альтернативную гипотезу, а именно

$$t_i \le t_{кр} = H_0;\ t_i > t_{кр} = H_p.$$

Примеры использования t критерия будут приведены ниже в 10.4.

Критерий Фишера. Помимо критерия Стьюдента в ряде случаев проверка нулевой гипотезы проводится по критерию Фишера. Этот критерий считается более точным при оценке равенства дисперсий в генеральной и выборочной совокупностях или двух генеральных совокупностей.

Р. Фишер открыл закон F-распределения, который описал специальной F-функцией. Учитывая краткость курса биометрии для лесоводов, мы описание этой функции опускаем. Отметим только, что функция Ф. Фишера (F) является непрерывной и зависит только от числа степеней свободы.

При выборках не очень малого размера (n>30) значимость различия между стандартными отклонениями σ_1 и σ_2 оценивают с помощью

$$t=(\sigma_1 - \sigma_2) / \sqrt{\sigma_{\sigma_1}^2 + \sigma_{\sigma_2}^2} \qquad (10.4)$$

где σ_{σ_1} и σ_{σ_2} - ошибки стандартного отклонения, определяемые по формуле $p=(\sigma_{\bar{x}} / \bar{x}) \cdot 100\%$.

При выборках малого объема разности стандартных отклонений имеют распределение, отличающееся от нормального, и рассмотренный метод оценки этих разностей (по доверительным границам или проверкой H_0 на основе t-критерия) является неточным. Здесь можно воспользоваться формулой

$t=\frac{\bar{x}-a}{\sigma}\sqrt{n}$, которая подчиняется t-распределению Стьюдента с k=n-1 степенями свободы.

Р.А. Фишер предложил вместо разностей σ_1 и σ_2 оценивать разность $Z=\ln \sigma_1 - \ln \sigma_2$, которая имеет нормальное распределение и при выборках среднего объема. При вычислении Z можно пользоваться десятичными логарифмами $Z=1{,}15131 \lg (\sigma_1^2 / \sigma_2^2)$.

Критерий Фишера (F) для оценки различия между выборочными дисперсиями обычно применяют в виде, который предложил Д. Снедекор.

$$F = \sigma_1^2 / \sigma_2^2 \qquad (10.2)$$

В уравнении (10.2) значение $\sigma_1^2 > \sigma_2^2$. Критические значения F для разных уровней значимости в практике определяют по специальным таблицам в зависимости от числа степеней свободы υ_1 и υ_2. При этом $\upsilon_1 = N_1 - 1$, а $\upsilon_2 = N_2 - 1$. Первой совокупностью (N_1) будет та, где величина σ^2 больше ($\sigma_1^2 > \sigma_2^{\kappa}$).

Для уровней достоверности 0,95, 0,99 и 0,999 критические значения F приведены в приложении Ж.

После сравнения вычисленного и критического (табличного) значений F выбирают нулевую (H_0) или рабочую (альтернативную) – H_p-гипотезу.

$F_i \le F_{кр} = H_0$; $F_i > F_{кр} = H_p$.

Примеры применения F критерия приведены ниже в 10.4.

Значимость различий качественных признаков. Качественные признаки, распределяющиеся по модели биномиального распределения, оценивают на основе долей. Методы оценки аналогичны вышерассмотренным для средних, выраженных в количественной мере.

Ошибка разности выборочных долей p_1 и p_2 определяется по формуле:

$$\sigma_{p1\text{-}p2} = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}} \qquad (10.6)$$

Для критерия t с числом степеней свободы $\nu = N_1 - N_2 - 2$ имеем

$$t=(p_1-p_2) / \sigma_{p1\text{-}p2} \qquad (10.7)$$

Когда имеется одна выборка, значение средней ее доли может быть оценено путем сравнения с гипотетической (теоретической) долей. Например, в отношении теоретической доли рождаемости мальчиков может быть выдвинута нулевая гипотеза $H_0 : P=0{,}5$.

В этом случае критерий

$$t=P-p / \sqrt{[p(1-p)] / N}, \qquad (10.8)$$

где P - теоретическая доля или вероятность, p - выборочная доля, N - численность выборки.

При $H_0 : P = 0$, $t=p / \sqrt{[p(1-p)] / N}$. (10.9)

10.3 Непараметрические критерии

Непараметрические критерии оценки статистических гипотез не требуют вычисления показателей ($\bar{X}$, σ и v) в выборочной совокупности. Они не базируются на нормальном распределении случайных величин в совокупности, и здесь часто применяют другие законы распределения. В ряде случаев для определения этих оценок используют условные значения, порядковые номера и т. д.

В современной практике математической статистики из непараметрических критериев чаще всего применяют следующие критерии: $\bar{x}$-критерий Ван-дер-Вардена, Т-критерй Уайта, Z-критерий знаков и W-критерий Вилкоксона

Не приводя доказательств соответствующих теорем, дадим формулы для их применения в практике.

$\bar{x}$-критерий Ван-дер-Вардена находят по формуле

$$\bar{x} = \sum \psi\left(\frac{R}{N_1 + N_2 + 1}\right), \text{ где}$$

R – порядковый номер (ранг);

N_1, N_2 – объемы выборок

Ψ – значение функции, определенное по специальной таблице в зависимости от величины $R/(N_1+N_2+1)$ (приложение З).

$\bar{x}$-критерий Ван-дер-Вардена используется для выборок с несвязанными парными вариантами. Для этого варианты обоих рядов ранжируют по мере их возрастания. В результате ранжирования каждое значение x_i получает порядковый номер (ранг). Затем находят отношение

$$K = \frac{R}{N_1 + N_2 + 1}.$$

Критическое значение $\bar{x}$-критерия находят по специальной таблице (приложение И) для разных уровней значимости (5%, 1%) с числом степеней свободы ($Y \neq N_1 + N_2 - 2$) и при разнице между объемами выборок: N_1-N_2. По $\bar{x}$-критерию выбираем нулевую (H_0) и рабочую (H_p) гипотезу.

$x \leq x_{кр} = H_0$;

$x > x_{кр} = H_p$.

Поясним изложенное примером. Пусть мы измерим высоты в двух дубовых древостоях II класса бонитета в возрасте 80 лет в типе леса дубрава кисличная (Д. кис) и дубрава папоротниковая (Д. пап). (таблица 10.2)

Таблица 10.2 – Результаты измерений высот в двух древостоях дуба

Тип леса	Высота измеренных деревьев (x_i), м										Средне-нее значе-ние ($\bar{x}$)
Д. кис.	22,0	22,5	21,7	17,9	20,8	28,4	25,6	19,6	23,6	20,9	22,3
Д. пап.	26,3	18,8	20,6	21,9	17,5	22,4	22,6	28,3	17,8	-	21,8

Проведем ранжирование данных таблицы 10.2 и найдем К и ψ (таблица 10.3).

Таблица 10.3 – Расчет $\bar{x}$ -критерия Ван-дер-Вардена

Высоты (м) по типам (x_i) леса		Ранг дерева (R)	$K_1 = \frac{R}{N_1 + N_2 + 1}$	$\psi(\frac{R}{N_1 + N_2 + 1})$
Д. кис.	Д. пап.			
-	17,5	1	0,050	-1,64
-	17,8	2	0,100	-1,28
17,9	-	3	-	-
-	18,8	4	0,200	-0,84
19,6	-	5	-	-
-	20,6	6	0,300	-0,53
20,8	-	7	-	-
20,9	-	8	-	-
21,7	-	9	-	-
-	21,9	10	0,500	0,00
22,0	-	11	-	-
-	22,4	12	0,600	0,25
22,5	-	13	-	-
-	22,6	14	0,700	0,52
23,6	-	15	-	-
25,6	-	16	-	-
-	26,3	17	0,850	1,04
-	28,3	18	0,900	1,28
28,4	-	19	-	-
$N_1 = 10$	$N_2=9$			-1,20

В таблице 10.3 мы вычислили K_1 = R/ N_1+N_2+1 для второй (меньшей) совокупности (N_2). Затем, используя таблицу в приложении И, где даны величины x, ψ в зависимости от величины R, нашли критерий Ван-дер-Вардена для наших рядов. Он оказался равен -1,2. По таблице в приложении И нашли критические значения критерия x при N_1-N_2=1. Он равен (при 1% уровне значимости) для Y=19 (Y = N_1+N_2) 4.77. Так как у нас 1,20 < 4,77, то принимается нулевая гипотеза, т. е., что различие при разнице в средних значениях на 0,5 м не значимы. Это значит, что обе выборки принадлежат к одной статистической совокупности высот дуба II класса бонитета, а тип леса не оказал существенного влияния на величину средней высоты.

Критерий Уайта. Этот критерий тоже используют для оценки разницы между средними значениями ($\bar{x}_1$ и $\bar{x}_2$) двух выборок с попарно несвязанными вариантами. Схема вычисления Т-критерия Уайта показана в таблице 10.5. Для примера воспользуемся вышеприведенными данными замеров высот дуба (таблица 10.2). Приведем парные величины высот в порядке их возрастания и покажем объем каждой выборки (таблица 10.4)

Таблица 10. 4 – Попарные замеры высот (H_i) дуба по мере возрастания

№ ряда	Высоты, м										Объем выборки	Сумма x_i	Среднее значение
1	17,9	19,6	20,8	20,9	21,7	22,0	22,5	23,6	25,6	28,4	10	223	22,3
2	17,5	17,8	18,8	20,6	21,9	22,4	22,6	26,3	28,3	-	9	196,2	21,8

На основе таблицы 10.4 построим таблицу 10.5.

В таблице 10. 5 мы выписали ранги деревьев по мере возрастания высот: X ранги равны от от 1 до 19, т.к. N_1+N_2 = 19. Затем против каждого дерева указываем номер выборки, из которой оно взято (графа 2). В графе 3 суммируем ранги деревьев (графа 1), которые попарно размещены в таблице 10. 4. Например, ранг дерева с высотой 20,8 м в первом ряду равен 7, а парного ему дерева в ряду 2 (18,8 м) – 4. Средний ранг этой пары будет (7+4)/2=5,5. В графы 5 и 6 выписываем средние ранги, принадлежащие первой и второй выборкам.

Таблица 10.5 – Порядок расчетов для нахождения Т-критерия Уайта

Ранги	Номер выборки	Высота, м	Совместный ранг	Ранги для первой выборки	Ранги для второй выборки
1	2	3	4	5	6
1	2	17,5	2	-	2
2	2	17,8	3,5	-	3,5
3	1	17,9	2	2	-
4	2	18,8	5,5	-	5,5
5	1	19,6	3,5	3,5	-
6	2	20,6	7,0	-	7,0
7	1	20,8	9,5	5,5	-
8	1	20,9	9,5	7,0	-
9	1	21,7	11,5	9,5	-
10	2	21,4	11,5	-	9,5
11	1	22,0	13,5	11,5	-
12	2	22,4	13,5	-	11,5
13	1	22,5	16	13,5	-
14	2	22,6		-	13,5
15	1	23,6	16	16	-
16	1	25,6	17	17	-
17	2	26,3	19	-	16
18	2	28,3		-	17
19	1	28,4		19	-
Итого			173	104,5	85,5

Проверка правильности вычислений проводится по формуле $\sum R = \frac{(N+1)*N}{2}$, где $N = N_1 + N_2$. В нашем примере $\sum R - 104.5 + 85.5 - 190$. $\frac{N+1}{2} = \frac{20*19}{2} = \frac{380}{2} = 190$. Таким образом, требуемое равенство соблюдено, т. е. вычисления сделаны правильно.

За значение Т-критерия Уайта принимается меньшая сумма рангов. В нашем примере это 85,5. По таблице в приложении К находим критическое значение Т-критерия для большего (N_1) и меньшего (N_2) объемов выборки при 1 % уровне значимости. Для $N_1 = 10$ и $N_2 = 9$ Т-критерий Уайта равен 58. Сравнив вычисленную нами величину ($R_{min} = 85.5$) с критическим значением Т-критерия при 1% уровне значимости, видим, что $T_{факт} > T_{крит}$, т. е. 85,5 > 58. Таким образом, Т-критерий Уайта тоже подтверждает, что измеримые высоты дуба в возрасте 80 лет из типов леса кисличный и папоротниковый принадлежат к одной статистической совокупности, определяемой вторым классом бонитета.

Z-критерий знаков используется для сравнения попарно связанных вариант, которые можно обозначить знаками (+) или (-). Обычно этот критерий

используют при сравнении опытных и контрольных измерений после проведения лесохозяйственных опытов.

Применение Z-критерия знаков основано на предположении, что количество опытов со знаком (+) и (-) одинаково при однородной выборке, и наоборот – неоднородная выборка показывает различное количество знаков как следствие влияния исследуемого фактора.

Опыт применения названного критерия покажем на следующем примере. Пусть имеем 2 однорядных участка ольхи черной в типе леса черноольшанник болотно-папоротниковый в возрасте 20 лет, класс бонитета – III. Разница в высотах этих участков, выявленная путем замеров, несущественна. В одном из участков провели гидротехническую мелиорацию. Через 10 лет провели повторные замеры высот как на контрольном (без мелиорации), так и на опытном участке и получили следующие результаты (таблица 10.6).

Таблица 10.6 – Расчет Z-критерия знаков для двух черноольховых древостоев

Вариант	Высоты измеренных деревьев, м															∑	$\bar{X}$
Опыт	16,3	17,0	17,9	10,9	10,8	14,3	13,0	18,2	10,6	11,5	16,4	14,1	14,3	11,9	12,7	210,8	14,1
Контроль	15,4	17,2	16,8	10,6	11,8	13,2	13,4	17,0	10,8	11,0	13,6	14,1	14,5	10,7	10,9	200,8	13,4
Эффект с (+)	+		+	+		+		+		+	+		+	+	+	10	
Эффект с (-)		-					-		-							3	
Нет эффекта (0)					0							0				2	

Значение Z-критерия знаков соответствует большему количеству эффектов без учета нулевых. В нашем примере положительный эффект наблюдали в 10 случаях, отрицательный – в 3, нулевой – 2. Объем выборки без нулевых эффектов равен 15 – 2 = 13. Число степеней свободы – 12*(N – 1).

По специальной таблице (приложение Л) находим величину Z-критерия для N = 12 при 1% уровне значимости он равен 11. Нулевая гипотеза предполагает отсутствие влияния исследуемого фактора, т. е.

$Z < Z_{кр} = H_0$ – нулевая гипотеза

$Z \geq Z_{кр} = H_0$ – рабочая (альтернативная) гипотеза.

Поскольку Zфактическое > Zкритическое, то применяется альтернативная гипотеза о том, что мелиорация за 10 лет (с возраста 20 до 30 лет) повлияла положительно (знаков “+” больше, чем “-“) на рост черноольшанника болотно-папоротникового.

W-критерий Вилкоксона. Определение этого критерия базируется на ранжировании положительных и отрицательных эффектов (без учета нулевых) при сравнении попарно связанных вариантов в опытной и контрольной выборке. За вычисленное (фактическое) значение W-критерия принимается сумма рангов, которая имеет наименьший знак.

Критическое значение W-критерия при заданном уровне значимости берем из специальных таблиц (приложение М). Нулевая (влияния нет) и альтернативная (рабочая) гипотезы определяются по методу, описанному выше.

$W_{факт} \geq W_{кр} = H_0$

$W_{факт} < W_{кр} = H_р$

Сказанное последним примером. Возьмем те же два насаждения ольхи черной (таблица 10. 6) и проведем вычисления W-критерия. Результаты показаны в таблице 10. 7.

Таблица 10.7 – Вычисление W-критерия Вилкоксона

№ п/п	Значение вариант		Ранги		Общий попарный ранг	Ранги с	
	Опыт	Контроль	Опыт	Контроль		+	-
1	16,3	15,4	10	12	11	11	
2	17,0	17,2	12	15	13,5	13,5	
3	17,9	16,8	13	13	13	13	
4	10,9	10,6	2	1	1,5	1,5	
5	11,8	11,8	4	6	5	-	-
6	14,3	13,2	9	7	8	8	
7	13,0	13,4	7	8	7,5		7,5
8	18,2	17,0	14	14	14	14	
9	10,6	10,8	1	3	2		2
10	11,5	11,0	3	5	4	4	
11	16,4	13,6	11	9	10	10	
12	14,1	14,1	8	10	9	-	-
13	14,3	14,5	9	11	10		10
14	11,9	10,7	5	2	3,5	3,5	
15	12,7	10,9	6	4	5	5	
	∑					83,5	19,5

По таблице (приложение М) необходим критический W-критерий при $N = N_ф – N_0 = 15 – 2 = 13$, который при 1% уровне значимости равен 11.

В нашем случае сумма рангов с (-) меньше, и ее принимаем за фактический W-критерий, который равен 19,5. Тогда $W_{факт} = 19$, $W_{кр} = 13$, т. е. $W_{факт} > W_{кр}$.

Таким образом, W-критерий Вилкоксона показывает недостоверность имеющихся различий.

В практике чаще применяются параметрические критерии как более научно обоснованные. Как видно из приведенных примеров применение двух разных непараметрических критериев для одного и того же опыта с влиянием мелиорации на черноольховый древостой (Z-критерий знаков и W-критерий Вилкоксона) показывает противоположные результаты.

Если бы мы для приведенных двух примеров (таблицы 10.2 и 10.6) применили параметрические (верность вычисления $\bar{x}$, σ и $m_{\bar{x}}$ предлагается проверить самостоятельно), то получили бы следующие результаты.

При сравнении высот древостоев дуба.

Для первого ряда: $N = 10$; $\bar{x} = 22{,}3$; $\sigma = 3{,}03$; $m_{\bar{x}} = 0{,}96$.

Для второго ряда: N = 9; $\bar{x} = 21{,}8$; $\sigma = 3{,}68$; $m_{\bar{x}} = 1{,}23$.

При сравнении высот древостоев черной ольхи.

Для первого ряда: $N = 15$; $\bar{x} = 14{,}1$; $\sigma = 2{,}55$; $m_{\bar{x}} = 0{,}92$.

Для второго ряда: $N = 15$; $\bar{x} = 13{,}4$; $\sigma = 2{,}33$; $m_{\bar{x}} = 0{,}60$.

Сделаем сравнение по критерию Стьюдента и Фишера. Сравнение высот дуба

$$t = \frac{x_1 - x_2}{\sqrt{m_{x_1}^2 + m_{x_2}^2}} = \frac{22.3 - 21.8}{\sqrt{0.96^2 + 1.23^2}} = \frac{0.5}{\sqrt{0.93 + 1.51}} = \frac{0.5}{\sqrt{2.44}} = \frac{0.5}{1.56} = 0.32$$

$$\bar{t} = \frac{\sigma_2^2}{\sigma_1^2} = \frac{1.51}{0.92} = 1.64$$

Критерий Стьюдента меньше его критического значения (приложение Е) при уровне значимости в 5% и 10%, т. е. различия несущественные. То же показывает и критерий Фишера, который оказался ниже порогового значения (приложение Ж). Следовательно, оба критерия подтверждают случайность расхождений в величинах высот.

Сравнение высот ольхи черной

$$t = \frac{14.1 - 13.4}{\sqrt{0.92^2 + 0.60^2}} = \frac{0.7}{\sqrt{0.85 + 0.36}} = \frac{0.7}{\sqrt{1.21}} = \frac{0.7}{1.1} = 0.64$$

$$F = \frac{2.55^2}{2.33^2} = \frac{6.5}{5.43} = 1.20$$

Полученные данные также свидетельствуют о несущественной разнице.

Оценка двух выборок при качественных признаках. Если применить количественную шкалу для оценки свойств того или иного явления невозможно, применяют оценки качественные. Можно, например, распо-

ложить отдельные единицы в ранжированный ряд от худших к лучшим, допустим, по форме, вкусу, запаху или другим свойствам. Если подобного рода ранжирование ряда объектов или вариантов эксперимента будут проведены при помощи случайной выборки из числа экспертов, то можно сделать определенные выводы о ранжированном ряде в генеральной совокупности. Предположим, шесть случайных экспертов оценивают лекции, сделанные двумя лекторами. В основу оценки положен учет ряда не измеряемых количественно факторов: содержание прочитанного материала, форма его подачи, культура чтения и пр. Предположим, независимые оценки с подразделением их на два ранга (1-лучше, 2-хуже) были такие: первый лектор получил пять оценок 1-го ранга и одну 2-го ранга, второй лектор, наоборот.

Нулевая гипотеза состоит в том, что нет значимого различия в качественной оценке лекций. Для оценки используют критерий χ^2 (хи-квадрат), формула для которого при малых выборках имеет выражение:

$\chi^2=(n_1-n_2-1)^2 / N$, где
n_1 - число однородных оценок;
n_2 - число неоднородных оценок.

Для рассматриваемого примера имеем $\chi^2=(5-1-1)^2 / 6=1,5$. Число степеней свободы при двух группах оценок равно 1. Из таблицы приложения Н находим $\chi^2_{0,05}=3,8$.

Нулевая гипотеза на уровне значимости 5% (т.е. при вероятности безошибочности заключения p=0,95) не отвергается. Она отвергается с вероятностью 0,90, которую в подобных случаях можно было бы считать достаточной, если повторение эксперимента было бы найдено затруднительным.

10.4 Проверка статистических гипотез в практике лесного хозяйства

В практике лесного хозяйства и, особенно при проведении научных исследований в лесном хозяйстве, часто возникают вопросы оценки эффекта от проведенных лесохозяйственных мероприятий или от причиненного вреда: рубки ухода, применение удобрений, селекция, мелиорация, вредители и болезни леса, пожары и т. д.. Примеры таких оценок с помощью критериев Стьюдента и Фишера, а также при использовании непараметрических оценок приведены выше. Наиболее часто, как уже отмечено, для этих целей используют параметрические оценки как более строгие. Здесь мы опишем типичную методику проведения таких оценок, которая принципиально применима к большинству оценок гипотез в лесном хозяйстве.

В опытах, чаще всего, возникают проблемы оценки эффектов, например, между высотами, диаметрами, приростами деревьев, получивших разные дозы удобрений, остающиеся при проведении рубок ухода разной интенсивно-

сти, при повреждениях пожарами или вредителями. При проведение таких исследований образуются парные наблюдения, где одно из них относится к первому варианту опытов (обычно это контроль), а другое ко второму. Разности между значениями признаков по парам образуют выборку, анализируя которую с помощью t-критерия Стьюдента, F-критерия Фишера или непараметрических критериев, делают соответствующее заключение.

Разности в опытах могут быть следствием достигнутого эффекта, но бывают из-за случайных причин, которые обычно остаются неизвестными. Если бы действовали только случайные причины, то по законам теории вероятности они имели бы разные знаки и их средняя в одной выборке равнялась бы нулю. Если же средняя здесь не равна нулю, то ее значимость требуется оценить.

Методику такой оценки покажем на искусственной модели двух выборок из одной совокупности. Допустим, заложена пробная площадь насаждений дуба в возрасте 95 лет II класса бонитета на площади в 1 га – 100×100 м. На этой пробной площади замерено 238 диаметров и высот (таблица 10.8). Данные замеры примем как генеральную совокупность и найдем ее статистические оценки. Для упрощения сделаем группировку диаметров по ступеням толщины через 4 см (таблица 10.8), а высот по ступеням высоты через 2 м (таблица 10.10). На основе распределения сгруппированных данных вычислим статистические показатели.

Для распределения диаметров схема вычислений показана в таблице10.9 Методика расчетов (вычисление $\bar{x}, \sigma$, начальных моментов) ранее излагалась в главе 3.

Таблица 10.9 – Исходные данные для вычисления статистических показателей ряда распределения диаметров в древостое дуба

Ступени толщины, x_i	Численности, n_i	$x_i * n_i$	Отклонения, x_i'	$x_i' * n_i$	$(x_i')^2 * n_i$	$(x_i')^3$	$(x_i')^4 *$	Вычисления для правила		$x_i - \bar{x} = k$	k^2	$k^2 * n_i$
								$x' + 1 = x''$	$(x'')^4 * n_i$			
12	3	36	-4	-12	48	-192	768	-3	243	-18.7	349,7	1049,1
16	9	144	-3	-27	81	-243	729	-2	144	-14.7	216,1	1944,9
20	21	420	-2	-42	84	-168	336	-1	21	-10.7	114,5	2404,5
24	30	720	-1	-30	30	-30	30	0	0	-6.7	44,9	1347,0
28	44	1232	0	0	0	0	0	1	44	-2.7	7,3	321,2
32	54	1728	1	54	54	54	54	2	864	1.3	1,7	91,8
36	35	1260	2	70	140	280	560	3	2835	5.3	28,1	983,5
40	23	920	3	69	207	621	1863	4	5888	9.3	86,5	1989,5
44	17	748	4	68	272	1088	4352	5	10625	13.3	176,9	3007,3
48	2	96	5	10	50	250	1250	6	2592	17.3	219,2	598,4
∑	238	7304	-	160	966	1660	9942	-	23256	-	-	13737,2

Таблица 10.8 – Ведомость измерения диаметров (Д) и высот (Н) дуба

№ дер.	Д, см	Н, м	№ дер.	Д, см	Н, м	№ дер	Д, см	Н, м	№ дер	Д, см	Н, м	№ дер	Д, см	Н, м	№ дер	Д, см	Н, м	№ дер	Д, см	Н, м
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	44,3	25,4	35	22,6	22,1	69	16,3	17,1	103	40,0	28,1	137	11,8	16,5	171	45,9	28,8	205	28,2	24,4
2	23,6	22,7	36	25,4	24,1	70	21,5	21,0	104	28,1	23,3	138	40,5	28,3	172	25,9	24,0	206	32,7	25,8
3	18,2	17,4	37	24,4	25,7	71	28,8	26,0	105	44,0	28,3	139	40,0	27,5	173	32,0	24,6	207	24,4	25,3
4	36,6	18,7	38	10,5	29,3	72	28,0	26,2	106	33,5	24,6	140	36,3	26,6	174	24,9	24,1	208	38,8	27,6
5	30,5	26,0	39	36,7	28,4	73	24,3	20,5	107	36,4	26,7	141	33,0	25,7	175	39,6	27,3	209	24,0	23,1
6	33,4	26,5	40	30,1	24,7	74	30,2	25,8	108	28,7	24,1	142	44,1	29,6	176	33,8	25,6	210	36,2	25,6
7	40,1	29,1	41	33,0	26,6	75	36,0	27,4	109	32,4	25,8	143	30,7	25,8	177	28,8	26,1	211	27,0	24,2
8	29,3	24,7	42	32	27,1	76	34,1	28,1	110	45,6	29,3	144	40,5	30,2	178	26,1	23,3	212	30,3	27,0
9	28,4	24,4	43	27,6	22,3	77	33,2	27,5	111	29,4	25,6	145	28,4	25,1	179	44,3	27,2	213	31,9	27,2
10	27,5	24,2	44	28,1	24,6	78	38,5	27,7	112	29,6	25,5	146	42,6	30,8	180	31,7	26,0	214	27,4	25,0
11	15,5	16,0	45	31,4	26,4	79	40,9	28,8	113	24,3	23,3	147	18,7	20,3	181	24,8	23,3	215	43,3	28,5
12	32,2	25,5	46	16,3	25,1	80	49,9	30,6	114	18,1	20,7	148	36,0	26,1	182	39,0	26,3	216	23,3	24,0
13	30,6	25,3	47	21,7	20,2	81	45,7	30,1	115	44,8	30,0	149	42,7	30,3	183	28,0	23,0	217	31,1	26,6
14	28,5	24,7	48	34,5	26,3	82	12,2	16,1	116	26,1	27,1	150	33,1	25,7	184	30,6	24,0	218	22,1	22,7
15	27,1	24,2	49	37,7	26,8	73	14,7	15,5	117	24,5	20,9	151	34,0	25,6	185	24,0	23,5	219	36,0	25,8
16	26,6	23,7	50	19,5	19,5	84	26,6	24,8	118	40,6	27,3	152	36,2	26,8	186	48,1	31,0	220	23,5	23,0
17	36,8	27,1	51	18,2	18,4	85	32,0	26,3	119	25,6	23,5	153	38,3	27,0	187	32,0	25,1	221	30,6	25,1

Продолжение таблицы 10.8

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
18	34,1	26,5	52	24,4	22,3	86	33,4	26,4	120	32,0	25,0	154	33,4	26,6	188	23,5	23,1	222	28,7	25,0
19	32,5	28,8	53	41,2	30,5	87	34,5	26,8	121	39,7	28,7	155	36,1	26,9	189	36,9	26,2	223	45,7	28,5
20	30,6	25,4	54	40,1	29,3	88	37,8	26,6	122	26,8	24,4	156	27,7	24,4	190	30,1	21,4	224	24,0	22,8
21	33,8	26,0	55	20,5	18,4	89	42,2	29,3	123	36,0	27,8	157	44,0	30,0	191	32,8	25,8	225	33,3	26,1
22	23,0	21,7	56	20,4	18,9	90	33,5	27,2	124	19,4	21,5	158	32,1	25,0	192	35,1	24,3	226	28,8	26,0
23	17,7	17,0	57	32,5	27,0	91	16,6	19,1	125	40,1	27,1	159	25,0	22,1	193	20,6	20,5	227	40,3	28,2
24	31,5	25,6	58	31,9	25,2	92	36,1	26,5	126	20,0	18,8	160	36,0	27,3	194	22,1	22,0	228	33,8	27,6
25	11,3	14,6	59	30,6	25,8	93	18,0	20,3	127	33,7	24,7	161	29,1	25,3	195	18,1	20,0	229	32,2	26,8
26	39,8	28,8	60	29,5	24,1	94	36,5	32,0	128	36,1	25,8	162	32,3	25,4	196	40,2	19,1	230	29,5	26,2
27	22,5	20,5	61	26,1	23,5	95	30,8	25,4	129	22,0	22,0	163	37,0	26,7	197	25,7	24,1	231	40,6	28,5
28	20,0	21,6	62	27,0	24,0	96	40,6	27,5	130	40,1	28,8	164	32,0	24,4	198	28,0	23,4	232	43,9	30,4
29	16,4	18,0	63	28,1	24,7	97	26,6	18,2	131	20,5	20,4	165	31,6	24,6	199	19,9	18,6	233	34,0	27,7
30	37,2	28,4	64	34,6	27,4	98	36,0	25,7	132	34,4	26,5	166	38,0	26,7	200	34,9	25,3	234	32,1	27,6
31	35,4	28,1	65	36,6	28,2	99	37,2	28,9	133	40,8	27,4	167	28,2	25,0	201	24,0	25,0	235	38,8	23,5
32	31,1	26,5	66	40,8	29,0	100	16,2	19,7	134	20,7	20,8	168	32,0	25,0	202	26,3	22,4	236	40,1	29,4
33	30,8	25,8	67	23,0	22,4	101	15,1	16,8	135	19,4	17,6	169	36,2	25,5	203	22,8	20,5	237	28,4	26,6
34	29,1	25,5	68	28,5	26,6	102	44,8	29,5	136	20,2	19,6	170	28,3	24,4	204	36,1	26,7	238	36,3	27,7

На основе таблицы 10.9 вычислим $\bar{x}$, $\sigma, m_{\bar{x}}$, V и начальные моменты.

Среднее значение $\bar{x} = \frac{\sum x_i * n_i}{\sum n_i} = \frac{7304}{238} = 30.7$ см.

Среднеквадратическое отклонение

$$\bar{\sigma} = \sqrt{\frac{\sum k^2 * n_i}{\sum n_i}} = \sqrt{\frac{13737.2}{238}} = \sqrt{57.719} = 7{,}60.$$

Основная ошибка среднего значения $m_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{7.60}{\sqrt{238}} = \frac{7.60}{15.43} = 0.49.$

Коэффициент вариации $V = \frac{\sigma}{\bar{x}} * 100\% = \frac{7.60}{30.7} * 100\% = 24.8.$

$m_{\bar{\sigma}} = \frac{\bar{\sigma}}{\sqrt{2n}} = \frac{7.6}{21.8} = 3.5$; $m_{\sigma} = \frac{1.9}{21.8} = 0.09.$

Для полной характеристики статистического ряда необходимо определить показатели асимметрии (α) и эксцесса (Е). Их найдем через моменты.

Начальные моменты, используя таблицу 10.9 определим по формулам.

$$m_1 = \frac{\sum x'_i * n_i}{\sum n_i} = \frac{160}{238} = 0.672;$$

$$m_2 = \frac{\sum (x'_i)^2 * n_i}{\sum n_i} = \frac{966}{238} = 4{,}059;$$

$$m_3 = \frac{\sum (x'_i)^3 * n_i}{\sum n_i} = \frac{1660}{238} = 6{,}975;$$

$$m_4 = \frac{\sum (x'_i)^4 * n_i}{\sum n_i} = \frac{9942}{238} = 41{,}773;$$

Сделаем проверку правильности вычисления начальных моментов.

$$m'_4 = \frac{\sum (x'_1 + 1)^4 * n_i}{\sum n_i} = \frac{23256}{238} = 97.71;$$

$$m'_4 = m_4 + 4m_3 + 6m_2 + 4m_1 + m_0.$$

Учитывая, что $m_0 = 1$, запишем

41,773+4*6,975+6*4,059+4*0,672+1=41,773+27,9+24,354+2,688+1=97,71

Равенство величин m'_4 вычисленного разными способами, выдерживается, т. е. начальные моменты верны.

$$\mu_2 = m_2 - m_1{}^2 = 4{,}059 - 0{,}452 = 3{,}607$$

$$\mu_3 = m_3 - 3m_2m_1 + 2m_1^3 = 6{,}975 - 3 * 4{,}059 * 0{,}672 + 2 * 0{,}672^3 = 6{,}975 - 3 * 4{,}059 * 0{,}672 + 2 * 0{,}672^3 = 6{,}975 - 8{,}183 + 0{,}607 = -0{,}601$$

$$\mu_4 = m_4 - 4m_3m_1 + 6m_2m_1^2 - 3m_1^4 = 41{,}773 - 4 * 6{,}975 * 0{,}672 + 6 * 4{,}059 * 0{,}672^2 - 3 * 0{,}672^4 = 41{,}773 - 18{,}749 + 10{,}998 - 0{,}612 = 33{,}411$$

Сделаем проверку вычисления центральных моментов.

$$\mu_3 = m_3 - 3m_2m_1 - m_1^3 = 6{,}975 - 3 * 3{,}607 * 0{,}672 - 0{,}672^3 = 6{,}975 - 3 * 3{,}607 * 0{,}672 - 0{,}672^3 = 6{,}975 - 7{,}272 - 0{,}304 = -0{,}601$$

$$\mu_4 = m_4 - 4m_3m_1 - 6m_2m_1^2 - m_1^4 = 41{,}773 - 4 * (-0{,}601) * 0{,}672 - 6 * 3{,}607 * 0{,}672^2 - 0{,}672^4 = 41{,}773 + 1{,}615 - 9{,}773 - 0{,}204 = 33{,}411$$

Равенство μ_3 и μ_4, вычисленных разными способами, говорит о верности расчетов.

С помощью моментов сделаем проверку вычисления $\bar{x}$ и $\bar{\sigma}$.

$\bar{x} = M' + km_1$. Учитывая, что величина ступени толщины (k) равна 4 см, запишем $\bar{x} = 28 + 4 * 0{,}672 = 28 + 2{,}69 \approx 30{,}7$, $\sigma = \sqrt{\mu_2} = \sqrt{3{,}607} = 1{,}9$, $\bar{\sigma} = k\sigma = 4 * 1{,}9 = 7{,}6$.

Как видим, величины $\bar{x}$ и $\bar{\sigma}$, вычисленные разными способами, совпали в пределах точности округления.

Теперь находим третий и четвертый основные моменты, помня, что $r_0 = 1$; $r_1 = 0$; $r_2 = 1$.

$r_i = \frac{\mu_i}{\sigma^i}$; $r_3 = \frac{\mu_3}{\sigma^3} = \frac{-0{,}601}{1{,}9^3} = -\frac{0{,}601}{6{,}859} = -0{,}088$; $r_4 = \frac{\mu_4}{\sigma^4} = \frac{33{,}411}{13{,}032} = 2{,}564$.

Тогда $\alpha = r_3 = -0{,}09$; $E = r_4 - 3 = 2.564 - 3 = -0.44$; $m_\alpha = \frac{6}{\sqrt{n}} = \frac{6}{15{,}4} = 0{,}39$; $m_E 2m_\alpha = 0{,}78$.

Теперь проведем аналогичные вычисления для распределения высот в исследованном древостое дуба (таблица 10.10)

Таблица 10.10 – Исходные данные для вычисления статистических показателей для распределения высот в дубовом древостое

Ступени высоты, x_i	Численности, n_i	$x_i n_i$	Отклонение, x'_i	$x'_i n_i$	$(x'_i)^2 n_i$	$(x'_i)^3 n_i$	$(x'_i)^4 n_i$	Данные для проверки			Данные для вычисления		
								$(x'+1)=k_i$	k_i^4	$k_i^4 n_i$	$x_i - \bar{x}_i =$	$(k'_i)^2$	$k'_i 2$
16	8	128	-4	-32	128	-512	2048	-3	81	648	-9	81	648
18	10	180	-3	-30	90	-270	810	-2	16	160	-7	49	490
20	15	300	-2	-30	60	-120	240	-1	1	15	-5	25	375
22	18	396	-1	-18	18	-18	18	0	0	0	-3	9	162
24=M'	46	1104	0	0	0	0	0	1	1	46	-1	1,0	46
26	75	1950	1	75	75	75	75	2	16	1200	1,0	11,0	75
28	46	1288	2	92	184	368	736	3	81	3726	3,0	9,0	414
30	18	540	3	54	162	486	1458	4	256	4608	5,0	5,0	450
32	2	64	4	8	32	128	512	5	625	1250	7,0	49	98
ИТОГО (∑)	238	5950	-	119	749	137	5897	-	-	11653	-	-	275 8

Среднее значение $\bar{x} = \frac{\sum x_i n_i}{n_i} = \frac{5950}{238} = 25.0$.

Вычисленные коэффициенты вариации для ряда распределения диаметров и высот в древостоях дуба соответствуют величине варьирования этих показателей для приспевающих и спелых насаждений, которые установлены крупнейшими нашими учеными-таксаторами: В. К. Захаровым, М. Л. Дворецким, Ф. П. Моисеенко, А. Г. Мошкалевым и др.

На основе таблицы 10.10 найдем начальные, центральные и основные моменты для определения асимметрии и эксцесса рядов распределения, используя формулы, приведены выше.

$m_1 = \frac{119}{238} = 0{,}500$; $m_2 = \frac{749}{238} = 3{,}147$; $m_3 = \frac{137}{238} = 0{,}576$; $m_4 = \frac{5897}{238} = 24{,}777$.

Сделаем проверку

$m'_4 = \frac{\sum(kr_i)^4}{\sum n_i} = \frac{11653}{238} = 48{,}962$.

$m'_4 = m_4 + 4m_3 + 6m_2 + 4m_1 + 1 = 24{,}777 + 4 * 0{,}576 + 6 * 3{,}147 + 4 * 0{,}5 + 1 = 24{,}477 + 2{,}304 + 18{,}882 + 2 + 1 = 48{,}963$

Таким образом, величины m'_4, вычисленные разными способами, равны (в пределах точности округления), т. е. расчеты сделаны верно.

Находим центральные моменты.

$\mu_2 = 3{,}147 - 0{,}5^2 = 2{,}897$;

$\mu_3 = 0{,}576 - 3 * 3{,}147 * 0{,}5 + 2 * 0{,}5^3 = 0{,}576 - 4{,}721 + 0{,}25 = -3{,}895$

$\mu_4 = 24{,}777 - 4 * (-3{,}895) * 0{,}5 - 6 * 2{,}897 * 0{,}5^2 - 0{,}5^4 = 24{,}777 + 7{,}79 - 4{,}345 - 0{,}062 = 28{,}16$

Вычисленные вторым способом величины μ_3 и μ_4 подтверждают верность расчетов.

Проверим правильность нахождения $\bar{x}$ и $\bar{\sigma}$.

$\bar{x} = M' + km_1 = 24 + 2 * 0{,}5 = 25$; $\sigma = \sqrt{\mu_2} = \sqrt{2{,}897} = 1{,}702$;

$\bar{\sigma} = k\sigma = 2 * 1{,}702 = 3{,}404$.

Как видим величины $\bar{x}$ и $\bar{\sigma}$ тоже совпадают с ранее найденными непосредственным способом.

$r_3 = \frac{\mu_3}{\sigma^3} = \frac{-3{,}895}{1{,}702^3} = -\frac{3{,}895}{4{,}930} = -0.79$

$r_4 = \frac{\mu_4}{\sigma^4} = \frac{2{,}816}{8{,}391} = 3{,}35$

Тогда $\alpha = r_3 = -0{,}79$; $E = r_4 - 3 = 0{,}35$.

Из приведенных статистик следует, что ряд сильно скошен влево.

Вычислив данные для всей совокупности, сделаем ее оценку по двум частичным выборкам и проверим их принадлежность к одной или нескольким совокупностям, т. е. проведем проверку статистических гипотез. Для этого из совокупности, представленной в таблице 10.8 возьмем по 2 выборки для диаметров и высот. Для этого отберем каждое 10 дерево и выпишем в таблицу 10.11 данные парных выборок для диаметров и высот. Первую выборку начнем с дерева №1, вторую с дерева №5. Тогда отбираемые деревья будут иметь следующие номера.

Для диаметров: №№1, 11, 21, 31, … , 231.

Для высот: №№5, 15, 25, 35, … , 235.

Таблица 10.11 – Частичные выборки деревьев с замеренными диаметрами и высотами из 238 деревьев на пробной площади в дубовом насаждении

Выборка №1			Выборка №2		
№ деревьев	Диаметры, см x_i (Д)	Высоты, м y_i (Н)	№ деревьев	Диаметры, см x_i (Д)	Высоты, м y_i (Н)
1	44,3	25,4	5	30,5	26,0
11	15,5	16,0	15	27,1	24,2
21	33,8	26,0	25	11,3	14,6
31	35,4	28,1	35	22,6	22,1
41	33,0	26,6	45	31,4	26,4
51	18,2	18,4	55	20,5	18,4
61	26,1	23,5	65	36,6	28,2
71	28,8	26,0	75	36,0	27,4
81	45,7	30,1	85	32,0	26,3
91	16,6	19,1	95	30,8	25,4
101	15,1	16,8	105	44,0	28,3
111	29,4	25,6	115	44,8	30,0
121	39,7	28,7	125	40,1	27,1
131	20,5	20,4	135	19,4	17,6
141	33,0	25,7	145	28,4	25,1
151	34,0	25,6	155	36,1	26,9
161	29,1	25,3	165	31,6	24,6
171	45,9	28,8	175	39,6	27,3
181	24,8	23,3	185	24,0	23,5
191	32,8	25,8	195	18,1	20,0
201	24,0	25,0	205	28,2	24,4
211	27,0	24,2	215	43,3	28,5
221	30,6	25,1	225	33,3	26,1
231	40,6	28,5	235	38,8	23,5
ИТОГО (∑)	731,7	591,6	-	738,4	594,3

Так как выборки малые ($\sum n_i < 30$), то вычисление $\bar{x}$ и $\bar{\sigma}$ проведем непосредственно с использованием ранее упомянутых формул.

$$\bar{x}_1^Д = \frac{731,4}{24} = 30,5 \quad \bar{x}_1^Н = \frac{591,6}{24} = 24,7 \quad \bar{x}_2^Н = \frac{591,3}{24} = 24,6$$

Вычисление $\bar{\sigma}$, Д, Н предлагается провести самостоятельно по данным таблицы 10.10. Приведем результаты счета.

$$\bar{\sigma}_1^Д = \sqrt{\frac{\sum(x_i-\bar{x})^2}{\sum n_i - 1}} = 9,20; \quad \bar{\sigma}_2^Д = 7,61; \quad m_{\bar{x}_{Д_1}} = \frac{\bar{\sigma}}{\sqrt{N-1}} = \frac{9,20}{\sqrt{23}} = \frac{9,20}{4,8} = 1,92;$$

$$m_{\bar{x}_{Д_2}} = \frac{7,61}{4,8} = 1,59; \quad m_{\bar{x}_{Н_1}} = 0,80; \quad m_{\bar{x}_{Н_2}} = 0,84; \quad \bar{\sigma}_1^Н = 3,85; \quad \bar{\sigma}_2^Н = 4,02.$$

Имея данные о средних значениях ($\bar{x}$) двух выборок по диаметру, и их среднеквадратических ошибок ($\bar{\sigma}$), а также об основных ошибках средних величин ($m_{\bar{x}}$) можем сравнить выборки по критериям Стьюдента и Фишера.

t-критерий Стьюдента:

$$t_Д = \frac{\bar{x}_{Д_1} - \bar{x}_{Д_2}}{\sqrt{m^2_{\bar{x}_{Д_1}} + m^2_{\bar{x}_{Д_2}}}} = \frac{30,8-30,5}{\sqrt{1,92^2+1,59^2}} = \frac{0,3}{\sqrt{3,69+2,53}} = \frac{0,3}{\sqrt{6,22}} = \frac{0,3}{2,49} = 0,12;$$

$$t_H = \frac{\bar{x}_{H_1} - \bar{x}_{H_2}}{\sqrt{m^2_{\bar{x}_{H_1}} + m^2_{\bar{x}_{H_2}}}} = \frac{24,7-24,6}{\sqrt{0,84^2+0,80^2}} = \frac{0,1}{\sqrt{0,71+0,64}} = \frac{0,1}{\sqrt{1,34}} = \frac{0,1}{1,16} = 0,09.$$

Величина t-критерия для диаметров и высот значительно меньше его критических значений (приложение Е), т. е. $t_{фак} < 2$, что говорит о том, что различия между выборками носит случайный характер, и обе выборки взяты из одной генеральной совокупности. Это же подтверждает и сравнение по F-критерию Фишера (r=23).

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

$$F_Д = \frac{9,2^2}{7,61^2} = \frac{84,64}{57,91} = 1,46$$

$$F_H = \frac{4,02^2}{3,85^2} = \frac{16,16}{14,82} = 1,09$$

Величины $F_Д$ и F_H меньше их критических значений при 5% уровне значимости (Y=N_1=23, N_2=23). Полученные величины $\bar{x}$, $m_{\bar{x}}$ и $\bar{\sigma}$ целесообразно сравнить по t и F-критериям с данными для нашей генеральной совокупности, т. е. с теми статистиками, которые вычислены по замерам 238 деревьев.

$$t_{Д_1} = \frac{30,8-30,7}{\sqrt{1,92^2+0,49^2}} = \frac{0,1}{\sqrt{3,69+0,24}} = \frac{0,1}{\sqrt{3,93}} = \frac{-0,1}{1,92} \approx 0,05$$

$$t_{Д_2} = \frac{30,7-30,5}{\sqrt{1,59^2+0,49^2}} = \frac{0,2}{\sqrt{2,53+0,24}} = \frac{0,2}{\sqrt{2,77}} = \frac{0,2}{1,66} = 0,12$$

$$\frac{\sigma^2_{Д_2}}{\sigma^2_Д} = \frac{7,61^2}{7,60^2} = \frac{57,91}{57,76} = 1,00$$

$$\frac{\sigma^2_{Д_1}}{\sigma^2_Д} = \frac{9,2}{7,60^2} = \frac{84,64}{57,76} = 1,46$$

$$t_{H_1} = \frac{25,0-24,7}{\sqrt{2,21^2+0,8^2}} = \frac{0,3}{\sqrt{4,88+0,64}} = \frac{0,3}{2,35} = 0,13$$

$$t_{H_2} = \frac{25,0-24,6}{\sqrt{3,21^2+0,84^2}} = \frac{0,4}{\sqrt{4,88+0,17}} = \frac{0,4}{2,36} = 0,17$$

$$F_{H_1} \frac{3,85^2}{3,404^2} = \frac{14,82}{11,59} = 1,28$$

$$F_{H_2} \frac{4,02^2}{3,404^2} = \frac{16,16}{11,59} = 1,39$$

Из приведенных расчетов видно, что обе частичные выборки соответствуют генеральной совокупности, т. к. значения t- и F-критериев меньше их критических значений.

По подобным схемам проводят оценку влияния различных факторов, а также проверяют соответствие разных выборок, взятых в разных древостоях одной генеральной совокупности.

Обобщая изложенное в настоящей главе, приведем алгоритм проверки статистических гипотез, который используется в лесном хозяйстве.

При проведении исследований - определении целей и конкретных путей его проведения - следует рассматривать возможность статистического подтверждения результатов, т.е. исследование в итоге должно сводиться к проверке (одной или нескольким) статистических гипотез, явно сформулированных.

Для лесоводственных исследований предварительный анализ такого рода тем более важен, что во многих ситуациях математическая формализация достаточно сложна. Кроме отмеченных выше трудностей перехода от объекта к модели (зачастую проверка гипотез относится именно к обоснованию возможности представления данного объекта при помощи определенной статистической модели), дальнейшие трудности вытекают из необходимости знания выборочных распределений используемых статистик. Здесь для простых случаев могут быть предложены стандартные методы, например использование предпосылки о нормальности распределения статистик, вычисленных на основании больших выборок.

Проверку статистических гипотез в исследованиях по лесному хозяйству обычно проводят для готовых результатов наблюдений. Порядок работы при этом следующий.

- Выбирают статистическую характеристику критерия, выяснив соответствие структуры задачи предпосылкам, лежащим в основе применяемых критериев. В частности, это могут быть выборочные распределения статистик, независимость наблюдений и пр.
- Выбирают и формулируют H_0 и H_a - проверяемую и альтернативную гипотезы. Нулевую гипотезу обычно принимают таким образом, чтобы последствия ошибки 1-го рода были более существенны, чем последствия ошибки 2-го рода.
- На основе последствий ошибок 1 и 2-го рода устанавливают допустимый уровень значимости α (односторонний или двусторонний в зависимости от типа альтернативной гипотезы) и вычисляют критические значения статистической характеристики, т.е. те значения, которые разделят распределение статистической характеристики на область допустимых значений и критическую.
- Вычисляют выборочное значение статистической характеристики и проверяют испытуемую гипотезу; если она не отклонена (полученное выборочное значение принадлежит области допустимых значений), то вычисляют мощность критерия $1-\beta$; при достаточной мощности гипотезу принимают, иначе заключение остается неопределенным и требуется увеличение объема выборки.

При планировании выборки объемом N можно вычислить мощность критерия или соответствующее этой мощности различие между параметрами, которое критерий может уловить. Если мощность оказывается недостаточной, то повышают объем выборки, обеспечивающий необходи-

мую мощность. При уровне значимости, равном или меньшем величине ошибки 2-го рода, неотклонение гипотезы означает ее принятие, ибо в таком случае критическая область альтернативной гипотезы при данном уровне значимости соответствует области допустимых значений проверяемой гипотезы.

Необходимо подчеркнуть, что термин “принятие” гипотезы не является синонимом абсолютной истинности H_0, поскольку H_0 и H_a, как правило, не исчерпывают всех влияний и обстоятельств, сказывающихся на изучаемом явлении. Они справедливы лишь в том смысле, что данные эксперимента не противоречат (либо противоречат) проверяемой гипотезе. Если для признания неверным какого-либо положения достаточно привести один пример, то в таком случае отклонение гипотезы (с вероятностью наличия ошибки 1-го рода), является свидетельством того, что она является ложной. В этом случае любое число подтверждений может оказаться недостаточным для того, чтобы гипотеза была справедливой.

При проведении исследований часто ограничиваются лишь частичным использованием вышеприведенных положений, применяя один из методов сравнения. Но при сложных случаях и при высокой практической важности исследований необходимо делать полный цикл проверки статистических гипотез.

11. КРИТЕРИИ СОГЛАСИЯ. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ В ЛЕСНОМ ХОЗЯЙСТВЕ

11.1 Критерии согласия

Статистические критерии являются важнейшей частью оценивания рядов распределения. Во многих случаях на основании некоторых гипотез или каких-то данных делается предположение о виде законов распределения интересующей нас случайной величины X. Мы часто имеем опытные данные в виде эмпирических дискретных рядов распределения, которые аппроксимируем некоторой кривой распределения: нормальной, типа А, β- распределения, логнормальной, Пуассона и т.д. Но здесь без дополнительных исследований нельзя утверждать, что примененная для аппроксимации кривая отвечает имеющемуся закону распределения искомой случайной величины X.

Например, сделав перечет диаметров деревьев, мы в зависимости от возраста древостоя, его происхождения, густоты, условий произрастания можем аппроксимировать эмпирическое распределение разными кривыми. Заранее нельзя с полной уверенностью заявить, которое распределение соответствует нашему ряду распределения замеренных диаметров. Поэтому появляется потребность проверить соответствие теоретического распределения опытным данным.

В силу ограниченного числа наблюдений теоретическая кривая будет в какой-то мере отличаться от фактического распределения опытных данных, даже если предположение о законе распределения сделано правильно. В связи с этим возникает необходимость решать следующую задачу: является ли расхождение между опытным законом распределения и предполагаемым законом распределения следствием ограниченного числа наблюдений, т.е. зависит от случайных причин, существенно не влияющих на характер распределения, или оно является существенным и связано с тем, что действительное распределение случайной величины отличается от предполагаемого. Для решения поставленной задачи служат критерии согласия.

Идея этих критериев заключается в том, что на основании определенного статистического материала они позволяют проверить гипотезу H, состоящую в том, что случайная величина X имеет функцию распределения F(x).

Для того чтобы принять или опровергнуть гипотезу H, будем рассматривать случайную величину Y, характеризующую степень расхождения теоретического и статистического распределений. Величину Y можно выбирать различными способами. Например, в качестве Y можно взять максимальное отклонение статистической функции распределения $F^*(x)$. Очевидно, закон

распределения случайной величины Y зависит от закона распределения случайной величины X, над которой производились опыты, и от числа опытов n.

Предположим, что закон распределения случайной величины нам известен. Тогда пусть в результате проведенных n опытов над случайной величиной X величина Y приняла некоторое значение y. Спрашивается, можно ли объяснить принятое значение Y=y случайными причинами или же это значение слишком велико и указывает на наличие существенной разницы между теоретическим и статистическим распределениями, т.е. непригодность гипотезы H.

Для ответа на этот вопрос допустим, что верна гипотеза H, и вычислим вероятность того, что случайная величина Y за счет случайных причин, связанных с ограниченным объемом опытного материала, примет значение не меньше, чем наблюдаемое значение y, т.е. вычислим вероятность P($Y \geq y$). Если эта вероятность мала, то гипотезу H следует опровергнуть как мало правдоподобную, а если же эта вероятность значительна, то экспериментальные данные не противоречат гипотезе H.

Для вычисления вероятности P($Y \geq y$) необходимо знать закон распределения случайной величины Y, который, как мы уже отмечали, зависит от закона распределения случайной величины X (функции распределения F(x)) и от числа опытов N. Оказывается, что при некоторых способах выбора случайной величины Y ее закон распределения при достаточно большом N практически не зависит от закона распределения случайной величины X. Именно такими мерами расхождения и пользуются в математической статистике и в биометрии в качестве критериев согласия, по которым оценивают соответствие распределения, полученного в опыте, теоретическому.

Критерии согласия можно разделить на две группы: в первой не используют значения выборочных статистик, во второй – критерии строят с использованием параметров генеральной совокупности, т.е. оценок последних на основе статистик. Из критериев второй группы наиболее интересны критерии Колмогорова-Смирнова и ω^2 для сравнения выборочной и теоретической функций распределения. Однако использование статистик вместо параметров, которые неизвестны в задачах аппроксимации, обычно приводит к завышению (иногда значительному) вероятностей согласия.

Из параметрических критериев наиболее употребляем и научно обоснован критерий согласия Пирсона.

11.2 Критерий согласия Пирсона

Этот критерий был предложен К. Пирсоном в 1900 году. Он базируется на определении некоторой статистики, которую автор определил как χ^2**.** Опустим доказательства для обоснования величины χ^2, которые сводятся к исследованию распределения случайной величины χ^2 с **ν**-степенями свободы, в силу краткости курса лесной биометрии. Здесь же отметим, что крите-

рий χ^2 представляет сумму отношений между квадратами разностей эмпирических и вычисленных или ожидаемых частот к ожидаемым частотам:

$$\chi^2 = \sum \frac{(p - p')^2}{p'}. \qquad (11.1)$$

Здесь $\sum$ - знак суммирования; p – эмпирическая частота;

p′ - ожидаемая или теоретически вычисленная частота.

Если разность между эмпирическими и вычисленными частотами обозначить через d, т.е. принять p - p′ = d, то формула (11.1) приобретает более простой вид:

$$\chi^2 = \sum \frac{d^2}{p'}. \qquad (11.2)$$

Чтобы вычислить критерий хи – квадрат, необходимо каждое отклонение эмпирического ряда от его ожидаемого значения возвести в квадрат и разделить на величину ожидаемого значения, затем полученные результаты сложить.

Когда эмпирические и вычисленные численности полностью совпадают друг с другом, разность p - p′ = d равна нулю. Чем больше различия между наблюдаемыми и ожидаемыми численностями, тем больше и указанная разность, а следовательно, и величина критерия хи – квадрат, которая может возрастать до бесконечности. Поскольку различия между ожидаемыми и эмпирическими частотами возводятся в квадрат, то значения критерия хи – квадрат могут быть только положительными. Поэтому при установлении разности p - p′ = d знаки можно не учитывать.

Преимущественное значение этого критерия состоит в том, что он применим к оценке опытных и ожидаемых значений в самых различных случаях – как при сравнении данных по одному, так и по нескольким независимым признакам, а также и для сравнения данных опыта и контроля. Особенно часто критерий хи–квадрат используется в генетике для сравнительной оценки результатов расщепления. Находит свое широкое применение этот критерий во многих областях лесного хозяйства, в частности, в лесной таксации.

Критерием «хи-квадрат» часто руководствуются для оценки выборочных распределений с их теоретически вычисленными частотами. Методика расчета теоретических значений эмпирических частот вариационного ряда изучена нами ранее.

Критерий согласия χ^2 Пирсона для сравнения рядов распределений определяют по одной из эквивалентных формул, первую из которых применяют при сравнении эмпирических n_i и теоретических $\tilde{n}_i$ частот, вторую – частостей p_i^1 и вероятностей p_i,

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - \tilde{n}_i)}{\tilde{n}_i} = N \sum_{i=1}^{m} \frac{(p_i - \tilde{p}_i)^2}{\tilde{p}_i}, \qquad (11.3)$$

где N – общее число наблюдений;

m – количество сравниваемых групп частот.

Критические значения χ^2 берут по таблицам (приложение Н) с учетом числа степеней свободы (V).

Если верна нулевая гипотеза H_0, заключающаяся в том, что распределение в генеральной совокупности соответствует выбранному теоретическому закону распределения, т.е. $H_0 : \chi^2 > 0$ (против альтернативной, или рабочей $H_д : \chi^2 = 0$, - проверку проводят одностороннюю. Так как χ^2 не может быть отрицательным, то величина χ^2 из (11.3) асимптотически подчиняется распределению χ^2 с $\nu = m - l - 1$ степенями свободы, где l – число параметров, использованных при расчете теоретического распределения. Количество параметров (l) зависит от вида распределения. Для распределений, которые наиболее часто применяются в лесном хозяйстве, число параметров следующее:

- распределение Пуассона – 1 ($\bar{\delta}$);
- нормальное, логнормальное, биномиальное – 2 ($\bar{\delta}$, σ);
- обобщенное нормальное (типа А или Грамма-Шарлье), Вейбула, система кривых Пирсона или Джонсона – 4 ($\bar{\delta}$, σ, α, Е).

Из разности $m - l$ вычитают единицу, поскольку одна дополнительная связь накладывается на частоты, чтобы суммы частот выборочного и теоретического распределений были равными.

Вывод распределения критерия χ^2 получают при условии, что все из сравниваемых частот не очень малы, поэтому при использовании (11.3) частоты n_i должны быть больше 5, т.е. $n_i > 5$. В связи с этим в практических задачах для одновершинных распределений, сходных с нормальным, частоты крайних интервалов рядов распределения нужно объединять. Отсюда следует, что одним из недостатков критерия χ^2 является его малая чувствительность к отклонениям частот для крайних значений.

Второй его недостаток – зависимость от способа группировки частот. В некоторых работах по математической статистике утверждается, что лучшие результаты дает использование вместо интервалов равной длины интервалов равных вероятностей. Последний способ пока не привился в практике из-за усложненной сводки материалов, т.к. необходимо рассматривать упорядоченную по возрастающим значениям случайных величин выборку. Однако, несмотря на приближенный характер критерия χ^2, он удобен и достаточно надежен для получения обоснованных заключений в задачах для лесного хозяйства.

Порядок применения χ^2 обычен. По (11.2) вычисляют статистическую характеристику критерия $\chi^2_{выч}$ и сравнивают ее с табличным (критическим)

значением $\chi^2_{1-\alpha}(\gamma)$ при уровне значимости α. Если $\chi^2_{выч} < \chi^2_{1-\alpha}(\gamma)$, то гипотезу принимают, т.е. можно считать, что эмпирический ряд подчиняется выбранному закону распределения.

При выборе величины α следует проверять гипотезу H_0 на тот предмет, что между выборочным и теоретическим рядами нет существенных различий, т.е. ошибка 1-го рода заключается в том, что в действительности имеющееся согласие признается ложным. Тогда ошибка 2-го рода сводится к тому, что, хотя в действительности ряд распределения не подчиняется данному теоретическому закону, критерий признает согласие. В этом случае ошибка 2-го рода более важна, чем ошибка 1-го рода, т.е. при формулировке гипотезы о согласии мы отступили от правил, изложенных выше. Однако здесь «поменять местами» H_0 и H_p невозможно, так как существуют разные теоретические законы распределения, по которым можно получить практически одинаковые частоты - значения χ^2. Поэтому вычисление мощности критерия согласия не имеет большого практического значения и обычно не производится, а уровень значимости выбирают относительно большим, чтобы уменьшить вероятность ошибки 2-го рода. Для задач в лесном хозяйстве можно рекомендовать α = 0,1 или даже 0,2.

Приведем пример применения критерия согласия χ^2.

Возьмем распределение диаметров в древостое дуба (таблица 10.9), которое попытаемся аппроксимировать кривой нормального распределения.

Вычислим выравнивающие частоты для этого ряда распределения (схема вычислений описана в главе 5). В таблице 11.1 приведена схема нахождения χ^2 по данным о численностях – фактическим и аппроксимированным кривой нормального распределения, и показана схема нахождения χ^2.

Таблица 11.1 – Схема вычисления χ^2 для распределения 238 диаметров в древостое дуба

Ступени толщины x_i	Численности		$n_i - \tilde{n}_i$	$(n_i - \tilde{n}_i)^2$	$\chi^2 = \frac{(n_i - \tilde{n}_i)^2}{n_i}$	Проверка $\frac{n_i^2}{\tilde{n}_i}$
	фактические, n_i	теоретические, $\tilde{n}_i$				
1	2	3	4	5	6	7
≥ 16	12	10	2	4	0,4	14,40
20	21	18	3	9	0,4	24,50
24	30	33	-3	9	0,3	27,27
28	44	45	-1	1	0,0	43,02
32	54	48	6	36	0,7	60,75
36	35	37	-2	4	0,1	33,11
40	23	24	-1	1	0,1	22,04
≥ 44	19	20	1	1	0,1	18,05
∑	238	235	3	69	2,1	243,14

В таблице 11.1 мы величины численности ≤5 объединили с соседними значениями, о чем было сказано выше.

Вычисленный критерий χ^2 равен 2,1. Колонка 7 служит для контроля. При условии, если $N = \widetilde{N}$, то

$$\chi^2 = \sum_{i=1}^{m} \frac{n_i^2}{\widetilde{n}_i} - N \qquad (11.4)$$

должен быть равен вычисленному по схеме таблицы 11.1. Поскольку у нас $N \neq \widetilde{N}$, то следует внести поправки, тогда формула приобретет вид

$$\chi^2 = \sum_{i=1}^{m} \frac{n_i^2}{\widetilde{n}_i} - N - (N - \widetilde{N}) \qquad (11.5)$$

Для нашего примера

$$\chi^2 = 243{,}14 - 238 - 3 = 2{,}1.$$

Сравним найденный критерий χ^2=2,1 с его табличным критическим значением (α) для разных уровней значимости (α), взятых из специальной таблицы (приложение Н). Число степеней свободы у нас равно: m-*l*-1 = 8-2-1 = 5. Для γ=5 имеем $\alpha_{0,9}$ = 9,24; $\alpha_{0,8}$ = 7,29; $\alpha_{0,5}$ = 4,35; $\alpha_{0,2}$ = 2,67; $\alpha_{0,1}$ = 1,62; $\alpha_{0,05}$ = 1,15 и т.д. Следовательно, по критерию χ^2 наш ряд распределения диаметров в древостое дуба соответствует нормальной кривой Гаусса-Лапласа с вероятностью более 80% (примерно 85-86%).

Рассмотренные выше критерии (t-критерий Стьюдента, χ^2) вводились в предположении, что выборки взяты из нормально распределенной генеральной совокупности, а сами наблюдения независимы. Между тем во многих практических случаях распределение случайной величины неизвестно, а в некоторых – известно то, что оно не соответствует кривой нормального распределения. Поэтому нам надо знать, какие отклонения от нормальности не искажают заключений, получаемых при помощи критериев, и как улучшить условия их применения. Не приводя системы доказательств по причинам, отмеченным выше (краткость курса изучаемой дисциплины), скажем лишь следующее.

В целом, критерии, относящиеся к средним генеральной совокупности (t-критерий Стьюдента), достаточно нечувствительны к отклонениям от нормальности распределений, особенно, если последние не очень асимметричны. Критерии, относящиеся к дисперсиям χ^2 и F, наоборот, очень чувствительны, при этом отклонения от нормального эксцесса играют гораздо большую роль, чем отклонения от симметричной формы. Применение критериев о дисперсии требует осторожности и проверки исходного распределения на нормальность, особенно при малых выборках. Если не подтверждается гипо-

теза о нормальности, то хорошие результаты дает применение преобразования распределений в нормальные. Так, если необходимо выровнять дисперсии выборок с приближенно равными коэффициентами изменчивости, то хорошие результаты дает логарифмическое преобразование y = lnx. Преобразование Джонсона хороший пример преобразования к нормальному виду. Для распределений, близких биномиальному, хорошие результаты дает преобразование $y = \arcsin\sqrt{x}$, для пуассоновского $y = \sqrt{x}$ или $y = \sqrt{x + 0{,}5}$ (для малых значений), для распределений со значительной левой асимметрией преобразование Фишера $y = 1/2\ln[(1+x)/(1-x)]$ и т.д.

Одним из критериев для проверки отклонения от нормальности (в случае больших выборок) могут служить основные ошибки асимметрии и эксцесса. В соответствии со стандартными методами распределение следует считать приближенно нормальным, если $\alpha / m_\alpha < 2$ и $E/m_E < 2$. Для более точного суждения о значимости α и E в зависимости от уровня значимости и числа наблюдений можно использовать специальные таблицы для α (приложение О) и E (приложение П). При пользовании этими таблицами надо иметь в виду, что распределение α симметрично, т.е. (1- α) = -(α) при данном N. Так, если α = 0,05, а выборочные значения α < 0,389 и E < 0,77 при N = 100, то гипотезу о нормальности распределения принимают.

Для вышеприведенного примера с распределением 238 стволов дуба по диаметру и высоте (глава 10) мы получили следующие величины.

Для ряда распределения диаметров:

$\bar{\delta}$=30,7; σ=1,9; $\bar{\sigma}$=7,6; v=24,7%; α=-0,01; E=-0,44;

$m_{\bar{x}}$=0,49; m_σ=0,09; $m_{\bar{\sigma}}$=0,35; m_α=0,39; m_E=0,78.

Для ряда распределения высот:

$\bar{\delta}$=25,0; σ=1,7; $\bar{\sigma}$=3,4; v=13,6%; α=-0,79; E=0,35;

$m_{\bar{x}}$=2,21; m_σ=0,078; $m_{\bar{\sigma}}$=1,56; m_α=0,39; m_E=0,78.

В нашем случае:

$$\frac{\alpha_d}{m_{\alpha D}} = \frac{0{,}01}{0{,}39} = 0{,}03 < 2\,;\quad \frac{\alpha_i}{m_{\alpha i}} = \frac{-0{,}79}{0{,}39} = 2{,}02 \approx 2\,;$$

$$\frac{E_d}{m_{ED}}\ \frac{0{,}44}{0{,}78} = 0{,}56 < 2\,;\quad \frac{E_i}{m_{Ai}} = \frac{0{,}35}{0{,}78} = 0{,}45 < 2\,.$$

На основе проведенных вычислений можем утверждать, что распределение диаметров соответствует нормальному закону (это подтверждается критерием согласия χ^2), а распределение высот этому распределению не отвечает.

Величина критических значений α и E (приложение О, П, Р) для распределения диаметров и высот (объем совокупности, от которой зависят β_1 и β_2, одинаков) β_1 = 0,246; β_2 = 0,55. Величины β_1 и β_2 подтверждают сделанный выше вывод о соответствии кривой нормального распределения ряда распределения по диаметру и отклонению от него (по α) для ряда высот.

11.3 Критерий согласия Колмогорова-Смирнова

Сравнительную оценку двух однородных вариационных рядов, как и сопоставление частот эмпирического и вычисленного распределений, можно произвести и с помощью так называемых непараметрических, или порядковых, критериев. В отличие от критерия хи-квадрат и критерия t Стьюдента, применение которых основано на использовании выборочных характеристик (параметров) $\bar{x}$ и σ, при вычислении непараметрических критериев этого не требуется. Но для их применения необходимо упорядочение в виде кумуляции эмпирических и теоретических распределений, т.е. получение рядов накопленных частот.

Для непараметрических критериев характерно то, что они в равной мере пригодны для оценки выборочных распределений любого вида, тогда как применение параметрических критериев исходит из положения о нормальности распределения оцениваемых рядов.

Один из наиболее простых и удобных при сопоставлении эмпирических совокупностей большого объема – критерий, предложенный советсткими математиками А.Н. Колмогоровым (1903-1987) и Н.В. Смирновым (1900-1966). Этот непараметрический показатель, обозначаемый греческой буквой λ (лямбда), представляет собой максимальную разность (d_{max}) между значениями накопленных частот эмпирического и вычисленного рядов (без учета знаков d), отнесенную к корню квадратному из суммы всех вариант совокупности:

$$\lambda = \frac{d_{\max}}{\sqrt{n}}. \qquad (11.6)$$

В отличие от критерия хи-квадрат критерий «лямбда» не только прост по конструкции, но не требует и специальных таблиц, хотя такие таблицы имеются (приложение С) и применяются при уточненных расчетах, но пользуются ими редко. Для упрощенной оценки критерия Колмогорова-Смирнова используют предельные значения критерия лямбда, соответствующие трем уровням доверительной вероятности – $P_1 = 0{,}95$, $P_2 = 0{,}99$ и $P_3 = 0{,}999$, которые соответственно равны 1,36, 1,63 и 1,95. Этот вывод вытекает из следующего расчета. Предельное значение критерия $\lambda = \sqrt{\frac{1}{2}\ln\frac{2}{P}}$, где P – соответствующий уровень значимости. Если принять P_1=0,05, то $\lambda = \sqrt{\frac{1}{2}\ln 40} = 1{,}36$. При P_2=0,01 λ=1,63 и т.д. При вычислении критерия «лямбда» отпадает необходимость определения числа степеней свободы.

В тех случаях, когда сравниваются два эмпирических распределения, взятых из одной и той же генеральной совокупности, но имеющих разный объем, критерий «лямбда» вычисляется по следующей формуле:

$$\lambda = d_{max}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \qquad (11.7)$$

Здесь $d_{max} = \sum\frac{p_1}{n_1} - \sum\frac{p_2}{n_2}$, т.е. эта максимальная разность между значениями первого $\left(\frac{p_1}{n_1}\right)$ и второго $\left(\frac{p_2}{n_2}\right)$ рядов накопленных частот.

Вычисление критерия Колмогорова-Смирнова продемонстрируем на примере распределения 238 диаметров в дубовом древостое (таблица 10.9). Результаты приведены в таблице 11.2

Таблица 11.2 – Вычисление критерия λ для распределения 238 диаметров дуба, аппроксимированных кривой нормального распределения

Ступени толщины (разряды) (x_i)	Численности		Накопленные частоты (P_i)		$d = P_ф - P_т$
	фактические (n_i)	выровненные	фактические $P_ф$	теоретические $P_т$	
12	3	3	3	3	0
16	9	7	12	10	2
20	21	18	33	28	5
24	30	33	63	61	2
28	44	45	107	106	1
32	54	48	161	154	7
36	35	37	196	191	5
40	23	24	219	215	4
44	17	16	236	231	5
48	2	4	238	235	3
Сумма	238	235	-	-	-

Максимальная величина разности $d = P_ф - P_т$ (без учета знаков) равна 7. Тогда $\lambda = \frac{d}{\sqrt{N}} = \frac{7}{\sqrt{238}} = \frac{7}{15{,}7} = 0{,}45$.

Полученная величина λ (0,45) значительно меньше его предельного значения (1,36) для P=0,05. Таким образом, можно утверждать, что расхождения между теоретически вычисленными частотами (по кривой нормального распределения) и фактическими носят случайный характер, и наш ряд распределения можно описать с помощью кривой Гаусса-Лапласа (нормального распределения).

Теперь рассмотрим пример, когда сопоставлены два ряда распределения, и необходимо оценить, принадлежат ли они к одной генеральной совокупности. Для этого сравним данные замеров диаметров на 2 пробных площадках, заложенных в двадцатитилетних сосновых насаждениях в типе леса сосняк мшистый II класса бонитета, различных по происхождению: естественный

древостой и лесные культуры. При этом объемы выборок отличаются (таблица 11.3).

Вычисление λ проводится по формуле (11.7):

$$\lambda = d_{max}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

$$\lambda = 0{,}232 \cdot \sqrt{\frac{320 \cdot 434}{320 + 434}} = 0{,}232 \cdot \sqrt{\frac{138880}{754}} = 0{,}232 \cdot \sqrt{184{,}2} = 0{,}232 \cdot 13{,}57 = 3{,}149\,.$$

Вычислить λ можно также по формуле

$$\lambda^2 = d^2\,\frac{N_1 \cdot N_2}{N_1 + N_2} = 0{,}232^2 \cdot \frac{138880}{754} = 0{,}05382 \cdot 184{,}2 = 9{,}9136\,;$$

$$\lambda = \sqrt{9{,}9136} = 3{,}15\,,$$

т.е. получили одинаковые величины в пределах точности округлений.

Таблица 11.3 – Вычисление критерия Колмогорова-Смирнова для двух пробных площадей, заложенных в сосновых молодняках

Ступени толщины (классовые промежутки) (x_i)	Численности		Частоты		Накопленные частоты		$d_i = \lvert P_1 - P_2 \rvert$	max d_i
	n_i^1 л/к	n_i^2 естеств.	$P_1 = \frac{n_i^1}{\sum n_1}$	$P_2 = \frac{n_i^2}{\sum n_2}$	$\sum P_{i1}$	$\sum P_{i2}$		
2	-	86	-	0,198	-	0,148	-	-
4	61	98	0,191	0,225	0,191	0,423	0,232	0,232
6	76	102	0,237	0,235	0,428	0,658	0,230	-
8	91	69	0,284	0,150	0,712	0,816	0,104	-
-10	46	35	0,144	0,080	0,856	0,896	0,040	-
12	29	27	0,091	0,062	0,947	0,958	0,011	-
14	17	12	0,053	0,027	-	0,985	-	-
16	-	5	-	0,015	-	1,000	0	-
∑(N)	320	434	1,000	1,000	-	-	-	-

Анализ вычисленного значения λ показывает, что сравниваемые совокупности принадлежат к разным генеральным совокупностям, и описываются различающимися кривыми: λ=3,15>2,1, т.е. достоверность различий превышает 99% уровень. Это соответствует материалам, приводимым многими учеными, которые исследовали строение искусственных и естественных сос-

новых молодняков: И.И. Григалюнас, В.Ф. Багинский, В.С.Моисеев, А.А.Макаренко и другие.

Из опыта применения критериев λ и χ^2 следует, что критерий Колмогорова-Смирнова менее чувствителен, чем χ^2. В спорных случаях (при граничных значениях) λ обычно подтверждает соответствие теоретическому распределению, а χ^2 может это утверждение опровергать. В этом случае исследователь в силу своей классификации, особенностей задачи, ее важности и сложности принимают нулевую или альтернативную (рабочую) гипотезу.

При применении критерия λ непременным условием применения должно быть относительно большое число (не менее 100) наблюдений. Поэтому простой по своей конструкции, он не приложим к оценке малочисленных совокупностей.

11.4 Применение статистического оценивания в лесном хозяйстве

Статистические критерии находят широкое применение в лесном хозяйстве, особенно при проведении научных исследований. Про применение t-критерия Стьюдента и F-критерия Фишера уже описано в главе 10.

Критерии согласия используют при многочисленных исследованиях товарности древостоев. Товарность насаждения зависит от многих факторов (их изучает лесная таксация и древесиноведение), но одним из главных являются закономерности распределения числа стволов по ступеням толщины в зависимости от среднего диаметра древостоя. Для того, чтобы убедиться, что распределение выбрано верно, применяют критерии согласия χ^2 и λ. Ошибка в выборе верного распределения может дорого стоить в прямом смысле этого слова, т.к. товарность древостоя определяет цену 1 $м^3$ и стоимость древостоя на выделе.

Критерии согласия используют также для оценки относительной однородности разных выборок. Это часто имеет значение для доказательства принадлежности древостоев на разных пробных площадях к одной или разным генеральным совокупностях как в примере, представленном выше. Такое сравнение важно при подборе серии пробных площадей, закладываемых с разными целями в древостоях примерно одинакового возраста, но отличающихся происхождением, полнотой и густотой, режимом ухода, проведенными мелиоративными мероприятиями и т.д.

Заканчивая рассмотрение про статистическое оценивание приведем основные статистические оценки в компактном виде для удобства пользователя.

Ошибка среднего значения $m_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, где

σ - среднее квадратическое отклонение

$P(|x_i - \bar{x}|) \le m) = 0{,}683$

$P(|x_i - \bar{x}|) \le 2m = 0{,}95$

$P\left(\left|x_i - \bar{x}\right|\right) \le 3m = 0{,}99$

Ошибка среднего квадратического отклонения

$$m_\sigma = \frac{\sigma}{\sqrt{2n}}$$

Ошибка коэффициента вариации

$$m_V = \frac{V}{\sqrt{2N}} \cdot \sqrt{1 + \left(\frac{V}{100}\right)^2} \quad \text{или} \quad m_V = V \cdot \sqrt{\frac{0.5 + 0.0001V^2}{N}}$$

Ошибка асимметрии

$$m_\alpha = \sqrt{\frac{6}{n}}$$

Ошибка эксцесса

$$m_E = 2m_\alpha$$

Точность опыта

$$P = \frac{V}{\sqrt{n}}$$

Количество наблюдений при заданной точности

$$n = \frac{V^2}{P^2}$$

Достоверность вывода

$$t = \frac{\bar{x}}{m_x}$$

Ошибка суммы средних величин ($\bar{x}$)

$$m_{\bar{x}_1 = \bar{x}_2} = \sqrt{m_{\bar{x}_1}^2 + m_{\bar{x}_2}^2}$$

$$m_{\bar{x}_1 + \bar{x}_2 + \ldots \bar{x}_n} = \sqrt{m_{\bar{x}_1}^2 + m_{\bar{x}_2}^2 + \ldots m_{\bar{x}_n}^2}$$

Средняя ошибка разности двух средних величин при $N_1 = N_2$

$$m_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{x_1}^2 + \sigma_{x_2}^2}$$

при $N_1 \neq N_2 \quad m_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{об}^2 \left(N_1 + N_2\right)} / N_1 N_2$

$$\sigma_{об}^2 = \left[\sum \left(x_{i1} - \bar{x}_1\right)^2 + \sum \left(x_{i2} - \bar{x}_2\right)^2\right] / \left(N_1 + N_2 - 2\right)$$

Ошибка произведения средних величин

$$m_{\bar{x}_1 \cdot \bar{x}_2} = \bar{x}_1 \bar{x}_2 \sqrt{\left(\frac{\sigma_{\bar{x}_1}}{\bar{x}_1}\right)^2 + \left(\frac{\sigma_{x_2}}{\bar{x}_2}\right)^2}$$

Средняя ошибка частного средних величин

$$m_{\frac{\bar{x}_1}{\bar{x}_2}} = \frac{\bar{x}_1}{\bar{x}_2} \sqrt{\left(\frac{m_{x_1}^2}{\bar{x}_1}\right)^2 + \left(\frac{m_{x_2}^2}{\bar{x}_2}\right)^2}$$

Оценка существенности различий между средними

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{m_{x_1}^2 + m_{x_2}^2}} = \frac{D}{m_D}$$

t – критерий достоверности

$|t| \geq 3$

Достоверность различия между дисперсиями

$$t = \frac{\sigma^2{}_1 - \sigma^2{}_2}{m^2{}_\sigma} = \frac{D}{m^2{}_\sigma}$$

$$m_\sigma = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(p - p')^2}{p}$$, где

p – эмпирические частоты

p´ - теоретические частоты

$$\lambda = d_{\max} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$

$$d_{\max} = \sum \frac{p_1}{N_1} - \sum \frac{p_2}{N_2}$$, где

N_i - объем ряда распределения

p_i - накопленные частоты

12. СТАТИСТИКИ СВЯЗИ

12.1 Понятие о корреляции

Исследуя лесные насаждения, мы замечаем, что в них гармонично сочетаются различные признаки. В лесу нет хаоса, а существуют законы и закономерности, по которым лесной биогеоценоз развивается и на основе которых существует. Подобные закономерности присущи как другим биологическим объектам, так и человеческому сообществу.

Взаимная зависимость двух величин (или явлений), когда изменение одной из них ведет к закономерному изменению другой называется **корреляцией**. Это понятие широко используется в науке, особенно в биологии, физике, химии. Например, в лесном хозяйстве известны тесные связи между диаметром и высотой дерева. Для одной породы, например березы, при большей высоте, как правило, наблюдается и больший диаметр. В свою очередь высота дерева в определенном возрасте зависит от плодородия почвы. С высотой, значит, и с плодородием почвы, тесно связана производительность древостоев. Форма ствола зависит от таксационной характеристики насаждения: его густоты, высоты дерева.

Зависимости и связи в природе и обществе, имеющие общие методы их статистического измерения, называют **корреляцией, связью или зависимостью**. Последние два слова - это синонимы термина “корреляция”.

Известны функциональные и корреляционные связи. Последние еще называют стохастическими. К функциональным относят законы математики и физики. Например $E = mc^2$; $L = 2\pi R$; $V = at$; $S = gt^2/2$ и другие известные из школьного курса математики и физики. Здесь изменение одной из величин всегда ведет к обязательному, четко обозначенному и всегда определенному изменению другой величины. Общие связи такого рода называются физическими или природными законами. Закон, примененный к некоторому частному случаю, называется закономерностью. Например, есть общий физический закон - металлы при нагревании расширяются. В отношении же конкретного металла, скажем меди, конкретные коэффициенты температурного расширения – это уже закономерность. Применительно к лесному хозяйству можно привести следующий пример. Рост дерева в высоту – биологический закон роста растений. Параметры роста (скорость увеличения высоты дерева с возрастом) для конкретного древесного вида, скажем, березы, - это закономерность.

В природе явления развиваются под воздействием различных факторов внешней среды. Поэтому связь между признаками проявляется в виде корреляционной связи, или корреляции.

В этом случае каждому значению одного признака здесь соответствует не одно, а несколько значений другого признака, т.е. его распределение. Один из признаков (обычно легче или точнее измеримый) принимают за факториальный, а другой – за результативный. Иногда, в условном значении, один называют независимым, а другой – зависимым от первого.

Статистическое исследование корреляции сводится к установлению факта связи, определению ее формы, направленности и тесноты. Установление факта связи специалисты в определенной отрасли производят сначала на основе общего анализа явления. Например, можно сказать о наличии корреляции между размерами дерева: толщиной и высотой еще до их измерения. В других случаях наличие корреляции между изучаемыми признаками нельзя предсказать столь определенно. Например, без измерений и последующего анализа трудно оценить связь формы ствола с его высотой. В этом случае решают вопрос о наличии корреляции на основе измерения и статистического анализа его результатов.

Поясним сказанное примерами. Так, мы уже говорили, что с увеличением диаметра дерева становится больше и его высота. Пусть у нас есть деревья ели I бонитета с диаметрами 20, 24, 28, 32 см. Высоты этих стволов равны 23, 25, 29, 32м. Но, если мы измерим, скажем, по 20 деревьев каждого из названных диаметров, то окажется, что высоты колеблются в таких пределах

Диаметр, см	Высота (от-до), м
20	21-25
24	23-27
28	26-32
32	29-35

Общая закономерность (увеличение высоты с ростом диаметра) выдерживается, среднее значение высоты тоже, но высота будет соответствовать вычисленному значению (23, 25, 29, 32) с определенной вероятностью (0,68) и среднеквадратической ошибкой. Это и есть вероятностная или стохастическая связь.

Корреляцию называют п р о с т о й, если она измеряется на основе двух признаков, или м н о ж е с т в е н н о й, если изменение результативного признака изучают в связи с влиянием или изменением нескольких факториальных признаков. Например, когда мы рассматриваем связь диаметр дерева – высота, то это простая корреляция. Связь, когда высоту дерева изучают в зависимости от почвенного плодородия, густоты древостоя, древесной породы, будет множественной.

По форме различают корреляцию л и н е й н у ю, когда зависимость между признаками отражается прямой линией, и к р и в о л и н е й н у ю, когда ее отражает уравнение какой-нибудь кривой. Во многих случаях форму корреляции можно предсказать еще до опыта. Например, между

длиной и толщиной корней молодых деревьев в древостое можно ожидать линейную корреляцию. Но нельзя ожидать такой же формы корреляции у деревьев, растущих в старых древостоях. Статистический анализ дает ответ о форме связи и в тех случаях, когда на основе биологического анализа ее установить трудно или вообще невозможно.

По направленности различают корреляцию прямую, когда с увеличением одного признака в среднем увеличиваются и значения другого, а с уменьшением - уменьшаются, и обратную, когда с увеличением значений одного признака значения другого в среднем уменьшаются и наоборот. Пример прямой связи приведен выше – диаметр и высота дерева. Обратную связь мы можем наблюдать, изучая влияние вредителей (допустим, обыкновенного соснового пилильщика) на прирост сосны. С увеличением количества гусениц на одном дерева прирост уменьшается.

Типичные картины статистических связей наблюдаются в двух практически разных случаях:

а) изучается связь между случайными величинами;

б) в действительности изучается функциональная связь, но погрешности измерений порождают изменчивость и создают видимость статистической связи.

С точки зрения статистического анализа на этапе выяснения существования зависимости эту разницу можно не учитывать; на этапе построения моделей возникают существенные различия в ее интерпретации и использовании.

Можно дать следующую трактовку стохастическим связям. При изучении связи между двумя величинами нельзя гарантировать, что одна величина полностью определяет значения другой, т.е. что учтены все основные факторы, общие для обеих величин. Кроме того, может существовать различная доля основных факторов, общих для обеих величин. Кроме основных существуют и случайные факторы, влияющие на изучаемые величины и затушевывающие имеющуюся закономерность. Поэтому чем больше общих факторов для изучаемых величин и чем полнее они учтены, тем отчетливее связь между изучаемыми величинами и тем уже “зона рассеивания”, которая в пределе превращается в некоторую линию, отражающую функциональную зависимость.

Из вышеприведенных рассуждений вытекает очевидная необходимость в статистических показателях, отражающих наличие и степень тесноты связи. Такие показатели (для выборки) называют статистиками связи. Важность статистик связи в моделировании очевидна: прежде чем конструировать модель, следует выяснить, в каких отношениях между собой находятся интересующие показатели. По смыслу здесь наблюдается полнейшая аналогия со статистиками распределений: статистики связи являются выборочными оценками параметров связи в генеральной совокупности.

Исходными данными для статистического анализа при большом числе наблюдений служат таблицы распределения, составляемые с со-

блюдением правил, идентичных правилам для составления рядов распределения, но наблюдения "разносят" с учетом обоих признаков. Пример распределения такого показан в таблице 12.1. Наглядно подобные распределения видны на графиках. Для примера такие графики приведены на рисунке 12.1. На рисунке 12.1а четко просматривается закономерная зависимость исследуемых признаков, а на рисунке 12.1б она выражена гораздо слабее.

Таблица 12.1 – Распределение диаметров и высот для 238 деревьев дуба

Ступени толщины, см (середины классов)	Число стволов, шт. (частоты)									Сумма частот	Условные средние
	Ступени высоты, м (середины классов)										
	16	18	20	22	24	26	28	30	32		
12	3	-	-	-	-	-	-	-	-	3	16,0
16	3	3	3	-	-	-	-	-	-	9	18,0
20	2	5	10	4	-	-	-	-	-	21	19,5
24	-	2	2	11	12	3	-	-	-	30	22,8
28	-	-	-	3	28	11	2	-	-	44	24,5
32	-	-	-	-	5	38	11	-	-	54	26,2
36	-	-	-	-	1	22	11	-	1	35	26,7
40	-	-	-	-	-	1	15	7	-	23	28,5
44	-	-	-	-	-	-	7	10	-	17	29,2
48	-	-	-	-	-	-	-	1	1	2	31,0
Сумма частот	8	10	15	18	46	75	46	18	2	238	-
Условные средние	15,5	19,6	19,7	23,8	27,6	32,3	37,2	42,7	42,0	-	-

Таблицы, составляемые подобно 12.1, используют затем для расчета тесноты связи между искомыми величинами. При малом числе наблюдений тесноту связи можно вычислять непосредственно, т.е. без сводки исходных данных в таблицы.

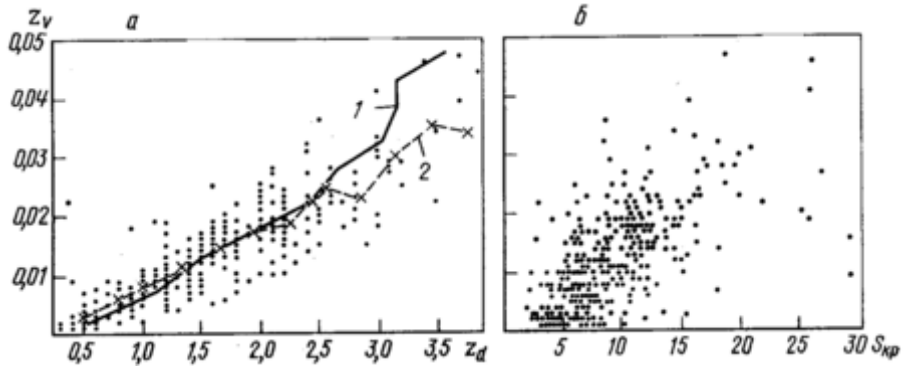


Рисунок 12.1 – Точечные диаграммы (по К.Е. Никитину и А.З. Швиденко):
а - текущего прироста по диаметру z_d и по объему z_v;
(1 - регрессия y на x; 2 - регрессия x на y)
б - площади проекций крон $S_{кр}$ и z_v

Требуется пояснить смысл терминов “связь” и “зависимость”. Как и во всем статистическом анализе, основным здесь является выяснение причинной сути установленных корреляций. Так, если рассматривается связь между диаметром дерева и его высотой, то очевидно, что y_i (D_i) зависит от x_i (H_i), но можно считать и наоборот. Поэтому в некоторых задачах возможна симметрия воздействия. Но если рассматривается связь между количеством выпадающих осадков и производительностью древостоев, то здесь может быть только зависимость «в одном направлении», хотя статистически можно попытаться установить и зависимость осадков от производительности древостоев, и даже получить какое-то ее цифровое выражение, но бесплодность подобных упражнений очевидна.

Тесноту корреляции, или степень сопряженности между значениями одного и другого признака, выражают в виде отвлеченных статистических характеристик (показателей) связи - коэффициента корреляции r и корреляционного отношения - η.

12.2 Коэффициент корреляции как мера линейной связи

Коэффициент корреляции – это статистика, которая является численной характеристикой связи между признаками, когда она имеет линейный характер. Это значит, что связь между величинами x и y выражается общей формулой $y=a_0+a_1x$, где a_0 и a_1 – коэффициенты, которые определяют на основе выборочных наблюдений. Коэффициент корреляции численно выражает отношение числа факторов, действующих на изменение обоих признаков к общему числу факторов.

Указанное содержание коэффициента корреляции достаточно хорошо выражает формула

$$r_{yx} = \frac{\sum n_{ij}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} \quad . \qquad (12.1)$$

Это выражение также можно записать

$$r = [\sum (X_i - \bar{X})(Y_i - \bar{Y})] / N\sigma_x \sigma_y \, . \qquad (12.2)$$

В формуле (12.1) x_i и y_i – случайные (исследуемые) величины; $\bar{x}, \bar{y}$ - средние значения для x и y.

Величина σ_x и σ_y в формуле (12.2) – среднеквадратические отклонения распределений X и Y; N - число сопоставляемых пар или число наблюдений.

Из формулы видно, что при независимом варьировании признаков, когда любое из отклонений $X_i - \bar{X}$ может сочетаться с любыми $Y_i - \bar{Y}$ (как с положительными, так и с отрицательными, притом одинаково часто),

числитель ее будет равен нулю или близкой к нулю величине. Следовательно, и r≈0. При сопряженном варьировании отклонения $X_i-\overline{X}$ сочетаются только с некоторыми отклонениями $Y_i-\overline{Y}$, например, положительные в основном только с положительными (при прямой связи) или положительные с отрицательными (при обратной связи). В этом случае сумма произведений будет иметь положительное (при прямой связи) или отрицательное (при обратной связи) значение, притом тем большее по своей величине (при данном N), чем связь сопряженнее.

Делением суммы произведений отклонений на число коррелирующих пар получают среднюю величину произведения, а делением на стандартные отклонения σ_x и σ_y выражают это произведение отвлеченным числом, характеризующим тесноту связи.

В целом коэффициент корреляции представляет собой эмпирический первый основной смешанный момент, т.е.

$$r = \frac{m_{1/1}}{\sigma_1, \sigma_2} \tag{12.3}$$

Для его вычисления надо найти первый начальный смешанный момент $m_{1/1}$ и первые два начальных момента каждого из рядов распределения. Вычисление моментов для одномерных рядов распределения описано ранее (глава 5). Здесь же покажем вычисление требуемых нам моментов для двухмерной совокупности.

Эмпирическим смешанным начальным моментом порядка (k_1, k_2) двух случайных величин (x_i, y_i), которые сведены в таблицу и сгруппированы по определенным разрядам, называется сумма произведений каждой пары отклонений $x_{1(i2)}$ и $x_{2(j2)}$. Отклонения берут от начальных значений $x_{1(a)}$ и $x_{2(a)}$ в h_1 и h_2 степени и умножают на соответствующую частность $P'_{i1/j2}$ (формула (12.4)).

$$m_{k_1/k_2} = \sum_{i_1=1}^{n_1} \sum_{j_2=1}^{n_2} P'_{k_1/k_2} x_{1(i/1)}^{k_1} x_{2(j/2)}^{k_2} \tag{12.4}$$

Придавая величинам k_1 и k_2 разные значения, получаем смешанные моменты $m_{1/1}$, $m_{2/1}$, $m_{1/2}$, $m_{1/3}$, $m_{2/2}$.

Соотношения между определенными выше смешанными центральными и начальными моментами даются формулами:

$$\mu_{1|1} = m_{1|1} - m_{1|0} m_{0|1},$$
$$\mu_{2|1} = m_{2|1} - 2m_{1|1} m_{1|0} - m_{0|1}(\mu_{2|0} - m_{1|0}^2),$$
$$\mu_{1|2} = m_{1|2} - 2m_{1|1} m_{0|1} - m_{1|0}(\mu_{0|2} - m_{0|1}^2),$$
$$\mu_{3|1} = m_{3|1} - 3m_{2|1} m_{1|0} + 3m_{1|1} m_{1|0}^2 - m_{0|1}(\mu_{3|0} - m_{1|0}^3),$$
$$\mu_{1|3} = m_{1|3} - 3m_{1|2} m_{0|1} + 3m_{1|1} m_{0|1}^2 - m_{1|0}(\mu_{0|3} - m_{0|1}^3),$$

$$\mu_{2|2} = m_{2|2} - 2(m_{2|1}m_{0|1} + m_{1|2}m_{1|0}) + 4m_{1|1}m_{1|0}m_{0|1} + m_{1|0}^2(\mu_{0|2} - m_{0|1}^2) + m_{0|1}^2\mu_{2|0}.$$

Для проверки вычислений применяются формулы:

$$\mu_{1|1} = m_{1|1} - m_{1|0}m_{0|1},$$
$$\mu_{2|1} = m_{2|1} - 2\mu_{1|1}m_{1|0} - m_{2|0}m_{0|1},$$
$$\mu_{1|2} = m_{1|2} - 2\mu_{1|1}m_{0|1} - m_{0|2}m_{1|0},$$
$$\mu_{3|1} = m_{3|1} - 3\mu_{2|1}m_{1|0} + 3\mu_{1|1}m_{1|0}^2 - m_{3|0}m_{0|1},$$
$$\mu_{1|3} = m_{1|3} - 3\mu_{1|2}m_{0|1} + 3\mu_{1|1}m_{0|1}^2 - m_{0|3}m_{1|0},$$
$$\mu_{2|2} = m_{2|2} - 2(\mu_{2|1}m_{0|1} + \mu_{1|2}m_{1|0}) + 4\mu_{1|1}m_{1|0}m_{0|1} + m_{0|2}m_{1|0}^2 + \mu_{2|0}m_{0|1}^2$$

Эмпирические смешанные основные моменты порядка (h_1; h_2) находятся при помощи центральных:

$$r_{h_1|h_2} = \frac{\mu_{h_1|h_2}}{\sigma_1^{h_1}\sigma_2^{h_2}}.$$

В частности,

$$r_{1|1} = \frac{\mu_{1|1}}{\sigma_1\sigma_2}, \qquad r_{2|1} = \frac{\mu_{2|1}}{\sigma_1^2\sigma_2}, \qquad r_{1|2} = \frac{\mu_{1|2}}{\sigma_1\sigma_2^2},$$
$$r_{3|1} = \frac{\mu_{3|1}}{\sigma_1^3\sigma_2}, \qquad r_{1|3} = \frac{\mu_{1|3}}{\sigma_1\sigma_2^3}, \qquad r_{2|2} = \frac{\mu_{2|2}}{\sigma_1^2\sigma_2^2}.$$

Смешанный основной момент первого порядка $r_{1|1}$ называется коэффициентом корреляции и обозначается через r:

$$r = \frac{\mu_{1|1}}{\sigma_1\sigma_2}. \qquad (12.5)$$

Смешанные моменты различных порядков могут быть вычислены как по способу произведений, так и по способу сумм.

По способу произведений вычисляется смешанный момент $m_{1|1}$ порядка (1,1). Для этого применяется схема, в которой наряду с таблицей распределения составляется еще вспомогательная таблица.

Схема вычисления $m_{1/1}$ показана в таблице 12.2. Для примера взято распределение 238 деревьев дуба по диаметру и высоте (таблица 12.1). Схема вычисления $m_{1/1}$ представляет собой таблицу распределения, в заголовках которой разрядные значения первой величины (диаметра) и разрядные значения второй величины (высоты) заменяются отклонениями $x'_{1(i/1)}$ и $x'_{2(j/2)}$ значений от соответствующих начальных величин начальных значений.

Таблица 12.2 – Схема вычисления $m_{1/1}$ для распределения диаметров и высот по способу произведений для 238 деревьев дуба

Ступени толщины, см	Число стволов, шт. (частоты)										Сумма частот
	Ступени высоты, м (середины классов)										
		16	18	20	22	24	26	28	30	32	
		-4	-3	-2	-1	0	1	2	3	4	
12	-5	+20 3	-	-	-	-	-	-	-	-	3
16	-4	+16 3	+12 3	+8 3	-	-	-	-	-	-	9
20	-3	+12 2	+9 5	+6 10	+3 4	-	-	-	-	-	21
24	-2	-	+6 2	+4 2	+2 11	12	-2 3	-	-	-	30
28	-1	-	-	-	+1 3	28	-1 11	-2 2	-	-	44
32	0	-	-	-	-	5	38	11	-	-	54
36	1	-	-	-	-	1	+1 22	+2 11	-	+4 1	35
40	2	-	-	-	-	-	+2 1	+4 15	+6 7	-	23
44	3	-	-	-	-	-	-	+6 7	+9 10	-	17
48	4	-	-	-	-	-	-	-	+12 1	+16 1	2
Сумма частот	-	8	10	15	18	46	75	46	18	2	238

Таблица 12.3 – Вспомогательная таблица для вычисления $m_{1/1}$ в древостое дуба

x_2 x	Четверти				∑стр. 1+4	∑стр. 2+3	∑стр. 5+6	Стр.0*стр.7	Проверка
	I	II	III	IV					
0	1	2	3	4	5	6	7	8	9
1	3	11	-	22	25	11	14	14	1.127+16+ +46+54- -5=238
2	11	3+2	-	11+1	23	5	18	36	
3	4	-	-	-	4	-	4	12	
4	2	-	-	1+15	18	-	18	72	
6	10+2	-	-	7+7	26	-	26	156	2.14+18+ +4+18+ +26+3+ +15+6+4+ +3=111
8	3	-	-	-	3	-	3	24	
9	5	-	-	10	15	-	15	135	
12	2+3	-	-	1	6	-	6	72	
16	3	-	-	1	4	-	4	64	
20	3	-	-	-	3	-	3	60	
∑	-	-	-	-	127	16	111	645	

Частоты строки и столбца таблицы, расположенных против отклонений, равных нулю, будучи умножены на нуль, дадут в результате нуль. Выделив эти нулевые строку и столбец при помощи жирных линий, мы разделим всю таблицу распределения на четыре четверти.

В I (левой верхней) четверги таблицы частоты должны быть умножены на отрицательные отклонения $x_{1\,(j1)}$ и отрицательные отклонения $x_{2\,(j2)}$; во II (правой верхней) четверти отклонения $x_{1\,(j1)}$ отрицательны; а отклонения $x_{2\,(j2)}$ положительны; в III (левой нижней) четверти отклонения $x_{1\,(j1)}$ положительны, а отклонения $x_{2\,(j2)}$ отрицательны; наконец, в IV (правой нижней) четверти как отклонения $x_{1\,(j1)}$ так и отклонения $x_{2\,(j2)}$ положительны.

Таким образом, в I и IV четверги частоты должны быть умножены на положительные произведения отклонений $x_{1\,(j1)}\,x_{2\,(j2)}$, а во II и III четверти – на отрицательные произведения отклонений $x_{1\,(j1)}\,x_{2\,(j2)}$. Каждое такое произведение отклонений с указанием знака выписывается в правом верхнем углу соответствующей клетки таблицы распределения цифрами меньшего размера.

Вспомогательная таблица (таблица 12.3) состоит из десяти столбцов (0)-(9). В столбце (0) выписывают в возрастающем порядке, не обращая внимания на знак и на число повторений, все произведения отклонений $x_{1\,(j1)}\,x_{2\,(j2)}$, помеченные маленькими цифрами в клетках таблицы распределения. В столбце (1) против соответствующих абсолютных произведений отклонений выписывают из I четверти таблицы распределения все частоты, соединяя их знаком плюс. Подобным же образом, в столбцах (2), (3) и (4) выписывают все частоты из II, Ш и IV четвертей таблицы распределения. В столбце (5) записывают суммы чисел каждой строки (1) и (4) столбцов, а в столбце (6) – суммы чисел каждой строки (2) и (3) столбцов. Затем находят итоги чисел столбца (5) и столбца (6). В рассматриваемом примере итог столбца (5) равен 127, а итог столбца (6) равен 16.

Прежде чем делать дальнейшие вычисления, необходимо произвести проверку правильности предыдущей работы. Для этого складывают итоги столбцов (5) и (6) вспомогательной таблицы 12.3 с итогами нулевой строки и нулевого столбца таблицы 12.2 распределения и из полученной суммы вычитают число, стоящее в центральной нулевой клетке таблицы распределения 12.2, так как это число, как нетрудно видеть, было подсчитано дважды. В результате таких действий должно получиться число, равное объему таблицы распределения. В рассматриваемом примере имеем:

$$127+16+46+54-5=238.$$

Проверка записывается в верхней части столбца (9) вспомогательной таблицы 12.3. Это проверка 1.

Убедившись, таким образом, в правильности проделанной вычислительной работы, составляют затем числа столбца (7), которые получаются путем вычитания чисел столбца (6) из чисел столбца (5); при этом необходимо проставлять знак полученной разности. Для проверки этого шага вы-

числительной работы заметим, что сумма чисел столбца (7) вспомогательной таблицы 12.3 должна равняться разности итогов столбцов (5) и (6) этой же таблицы. В рассматриваемом случае

$$127\text{-}16=111.$$

Эта проверка записывается в нижней части столбца (9) вспомогательной таблицы 12.3. Это проверка 2. У нас равенство получилось.

Последний шаг работы при вычислении смешанного момента $m_{1/1}$ делается в столбце (8) вспомогательной таблицы. В этом столбце помещаются произведения чисел столбцов (0) и (7). Сумма чисел столбца (8) представляет числитель смешанного момента $m_{1/1}$. Разделив эту сумму на объем таблицы распределения, получим искомый момент. В рассматриваемом примере имеем

$$m_{1/1} = \frac{645}{238} = 2{,}170\,.$$

При ручном счете большой объем вычислительной работы и отсутствие возможности полной проверки в самом ходе вычислений смешанного момента $m_{1/1}$ были существенными недостатками способа произведений. Для вычисления же смешанных моментов более высокого порядка ручной способ оказывается совершенно непригодным.

В настоящее время все подобные расчеты автоматизированы и, естественно, практически никто вручную коэффициент корреляции не вычисляет, но специалист должен знать суть этого процесса и алгоритм вычислений.

Для ручного счета ранее применялся способ сумм как более легкий в расчетах. При компьютерных вычислениях проще применять способ произведений (легче составлять программу), поэтому описание громоздкого способа сумм здесь опускаем.

Вычислив первый смешанный момент, определяем коэффициент корреляции по приведенной выше формуле (12.5). Для этого необходимо вычислить первый смешанный центральный момент $\mu_{1/1}$. Формула для его вычисления приводилась выше:

$$\mu_{1/1} = m_{1/1} - m_{1x} \cdot m_{2y}, \qquad (12.6)$$

где $\mu_{1/1}$ – первый основной смешанный момент;

m_{1x} и m_{2y} – соответственно первые начальные моменты распределения по x и y. У нас это диаметр и высота. Эти моменты вычислены ранее (глава 10, табл. 10.9 и 10.10). Они равны $m_{1x(D)} = 0{,}672$; $m_{2y(M)} = 0{,}500$.

Тогда $\mu_{1/1} = 2{,}71 - 0{,}672 \cdot 0{,}500 = 2{,}76 - 0{,}336 = 2{,}374$.

Среднеквадратическое отклонение для рядов распределения по диаметру и высоте вычислены ранее (глава 10) и равны

$$\sigma_D = 1{,}9\ (\overline{\sigma} = 7{,}6),\ \sigma_M = 1{,}702\ (\overline{\sigma} = 3{,}403)\ ;$$

Тогда $r = \dfrac{2{,}374}{1 \cdot 9 \cdot 1{,}7} = \dfrac{2{,}374}{3{,}23} = 0{,}74$.

При малых выборках коэффициент корреляции можно вычислить непосредственно по приведенным выше формулам (12.1-12.2).

Пример вычисления коэффициента корреляции, который приводит Н.Н. Свалов, для малых выборок показан в таблице 12.4. Подставив в формулу величины, вычисленные в 12.4, получим

$$r = \frac{227 - (55 \cdot 40)/10}{\sqrt{(313 - \frac{55^2}{10})(166{,}26 - \frac{40^2}{10})}} = \frac{7{,}0}{\sqrt{10{,}50 \cdot 6{,}26}} = 0{,}86.$$

Таблица 12.4 – Вычисление коэффициента корреляции между длиной стволиков и длиной корней сеянцев сосны

Длина стволика X, см	Длина корня Y	Отклонение					Вычисление
		x	y	xy	x^2	y^2	
5	3,5	-0,5	-0,5	+0,25	0,25	0,25	
6	4,0	+0,5	0	0	0,25	0	
5	4,1	-0,5	+0,1	-0,05	0,25	0,01	$\overline{X} = \sum X/N = 55/10 = 5{,}5$ см
7	5,0	+1,5	+1,0	+1,50	2,25	1,00	
6	3,5	+0,5	-0,5	0,25	0,25	0,25	
4	3,1	-1,5	-0,9	+1,35	2,25	0,81	$\overline{Y} = \sum Y/N = 40/10 = 4{,}0$ см
5	3,5	-9,5	-0,5	+0,25	0,25	0,25	$r = \sum xy / \sqrt{\sum x^2 \sum y^2} =$
4	3,0	-1,5	-1,0	+1,50	2,25	1,00	$= 7{,}00 / \sqrt{10{,}50 \cdot 6{,}26} = 0{,}86$
7	5,3	+1,5	+1,3	+1,95	2,25	1,69	
6	5,0	+0,5	+1,0	+0,50	0,25	1,0	
$\sum$ 55	40,0	0	0	+7	10,50	6,26	

Оценка показателей связи при малых выборках. При оценке коэффициента корреляции при малой выборке возникают некоторые новые проблемы в связи с тем, что при высоких значениях корреляции в генеральной совокупности (ρ) выборочные коэффициенты корреляции (r) имеют

не нормальное, а позитивно (правостороннее) асимметричное распределение. В таких случаях на основе выборочного r и его ошибки σ_r можно применить лишь оценку значимости r при гипотезе ρ=0, т.е. лишь одну форму оценки из двух.

Для данных таблицы 12.4 при r=0,86 имеем следующую его оценку

$$\sigma_r = \sqrt{(1-r^2)/(N-2)} = \sqrt{(1-0{,}86)(10-2)} = 0{,}18.$$

Затем вычислим t-критерий и сравним его с табличными значениями (приложение Е)

$$t_r = r / \sigma_r = 0{,}86 / 0{,}18 = 5{,}33.$$

Полученное значение t_r превышает даже $t_{0{,}001} = 5{,}0$. Следовательно, r значимо на высоком уровне достоверности.

Эту форму оценки нельзя применять при других гипотезах, кроме ρ=0. Равным образом критерий t нельзя использовать для построения доверительного интервала. Р.А. Фишер предложил для этих целей z-преобразование величины r.

Величина z

$$z = 1/2\ [ln\ (1+r) - ln\ (1-r)] \qquad (12.7)$$

имеет нормальное распределение с дисперсией

$$\sigma_z^2 = 1/(N-3) \qquad (12.8)$$

Для получения z имеются номограммы и таблицы, которые здесь не приводятся.

Для совокупности длин стволиков и корней при r=0,86 z=1,293. Ошибка этой величины $\sigma_z = \sqrt{1/(10-3)} = 0{,}378$. Доверительный интервал для z_r в генеральной совокупности равен $z \pm t_{0{,}05} S_z$=1,293±2,3 · 0,378, т.е. от 0,424 до 2,162. Тогда интервал для ρ равен от 0,40 до 0,97.

Преобразование z применяется также при сравнении двух выборочных коэффициентов корреляции и при нахождении обобщенного для них коэффициента, когда оказывается, что выборки взяты из одной совокупности. Пусть взято 2 выборки всходов сосны, предположительно из одной совокупности: r_1=0,86, S_{r1} = 0,18, N_1=10 и r_2=0,70; σ_{r2} = 0,25, N_2=10. Можно ли считать различие в коэффициентах значимым?

Квадратическая ошибка разности z_1 и z_2 следующая:

$$\sigma_{z1\text{-}z2} = \sqrt{[1/(N_1-3)] + [1/(N_2-3)]} \qquad (12.9)$$

Для сравниваемых выборок имеем

$$\sigma_{z1\text{-}z2} = \sqrt{[1/7 + 1/7} = 0{,}54,$$

$$t_{z1\text{-}z2} = (z_1 - z_2) / S_{z1\text{-}z2} = (1{,}293 - 0{,}867) / 0{,}54 = 0{,}78 < t_{0{,}05}.$$

Имеющееся различие между коэффициентами корреляции следует считать случайным, а выборки – взятыми из одной генеральной совокупности.

Объединение выборочных коэффициентов корреляции целесообразно, если корреляция между признаками в выборках существенно не различается, как в рассмотренном примере. Обобщенный коэффициент корреляции является более надежной оценкой коэффициента корреляции ρ в общей совокупности.

Для получения r также пользуются значениями z, находя из них среднюю взвешенную величину. В качестве веса принимают число наблюдений в каждой выборке, уменьшенное на 3 единицы, т.е.

$$z=[z_1 (N_1 - 3) + z_2 (N_2 - 3)] / [(N_1 - 3) + (N_2 - 3)]$$

Для двух выборок сеянцев сосны имеем

$$z = (1{,}293 \cdot 7 + 0{,}867 \cdot 7) / (7 + 7) = 1{,}080.$$

Квадратическая ошибка среднего взвешенного z

$$\sigma_z = 1 / \sqrt{(N_1 - 3) + (N_2 - 3)}$$

Для нашего примера σ_z =0,071, $t_z = z / S_z$ =1,080/0,071 = 15 > $t_{0{,}05}$ и даже $t_{0{,}01}$, т.е. z является значимым на высоком уровне значимости.

Найденному значению z=1,080 соответствует r=0,79, который является наиболее надежной оценкой ρ в генеральной совокупности.

Коэффициент корреляции может принимать значения от +1 до -1. При полной прямой корреляции r = +1, при полной обратной r = -1. При r = 0 или близкой к 0 прямолинейная связь отсутствует (криволинейная связь при этом может быть, например, окружность). Обычно считают, при величине до 0,1 – связи нет; r = 0,11-0,30 – связь слабая, при r = 0,5-0,6 – средняя; при r = 0,7-0,9 – сильная или тесная; r = 0,9-1,0 – очень высокая.

Таким образом, в нашем примере связь между диаметрами и высотами в древостое дуба оказалась сильной. Наличие или отсутствие связи между различными явлениями или признаками можно подтвердить и другими примерами. Так, нет связи между классом бонитета древостоя и наличием лесных дорог. Слабая связь между водообеспеченностью древостоев и уровнем грун-

товых вод на гидроморфных почвах: она проявляется только в засушливые годы. Средняя связь между величиной текущего прироста ($м^3$/га) и полнотой, т.к. здесь идут два взаимно противоположные процесса: уменьшение числа деревьев и увеличение прироста на оставшихся стволах. Сильная связь наблюдается между классом бонитета древостоя и уровнем почвенного плодородия. Очень высокая связь существует в древостое между диаметрами и высотами, высотами и видовыми числами.

Детальные исследования в теории корреляции показали, что степень сопряженности случайных величин x и y более точно описывает квадрат коэффициента корреляции, получивший название коэффициента детерминации *(d)*, определяемый как $d=r^2$. Коэффициент детерминации, выраженный в процентах, показывает ту часть изменчивости зависимой переменной (y), которая вызвана влиянием независимой переменной (x), т.е. он более отчетливо выражает зависимость $y=f(x)$.

12.3 Корреляционное отношение как мера нелинейной связи

Связь между x и y далеко не всегда выражается линейно. Поэтому для любой формы связи К. Пирсон предложил статистику зависимости между случайными величинами x и y, которую назвал корреляционным отношением. Статистическое обоснование этого показателя имеет определенную сложность и здесь не приводится. Наиболее важно знать вычисление корреляционного отношения. Особенно если с помощью коэффициента корреляции установлено, что связь слабая или средняя (r < 0,7-0,8). В этом случае связь может быть даже сильной, но иметь криволинейный характер. Например, для зависимости $y=f(x)$, имеющей вид окружности, r=0, хотя связь есть и весьма тесная. Поэтому и был предложен новый показатель – корреляционное отношение, – который с успехом отражает как линейную, так и нелинейную связь. Корреляционное отношение выражают греческой буквой эта – η. Предварительную оценку криволинейной связи можно сделать, построив график функции $y=f(x)$, на котором вид связи (прямолинейная или криволинейная) будет виден.

Для вычисления корреляционного отношения строят таблицу распределения. Подобную мы составляли при определении коэффициента корреляции (таблица 12.5)

Таблица 12.5 – Общий вид таблицы исходных данных для вычисления корреляционного отношения

x	y	$\overline{\gamma}$
x_1	y_{11} y_{12} y_{13} … y_{1n}	$\overline{\gamma_1}$
x_2	y_{21} y_{22} y_{23} … y_{2n}	$\overline{\gamma_2}$
x_3	y_{31} y_{32} y_{33} … y_{3n}	$\overline{\gamma_3}$
…	……	…
x_k	y_{k1} y_{k2} y_{k3} … y_{kn}	$\overline{\gamma_k}$

По данным таблицы 12.5 вычисляем следующие суммы квадратов

$$\sigma_y^2 = \sum (y_i - \overline{\gamma_i})^2 ,$$

где $\overline{\gamma_i}$ - условные средние, вычисленные для каждой строки;

$$\sigma_{xy}^2 = \sum (\overline{\gamma_i} - \overline{\gamma})^2 ,$$

где $\overline{\gamma}$ - средняя по γ для всей совокупности;

$$\sigma_0^2 = \sum (\overline{\gamma_i} - y_i)^2 ,$$

где σ_0 – остаточная сумма квадратов, обусловленная иными причинами, чем регрессия y на x.

Тогда

$$\sigma_y^2 = \sigma_{yx}^2 + \sigma_0^2 , \qquad (12.10)$$

$$\text{а} \quad \eta^2 = 1 - \frac{\sigma_0^2}{\sigma_y^2} . \qquad (12.11)$$

После элементарных математических преобразований формулы (12.11) с учетом (12.10) получаем

$$\eta^2 = \frac{\sigma_y^2 - \sigma_0^2}{\sigma_y^2} = \frac{\sigma_{yx}^2}{\sigma_y^2} . \qquad (12.12)$$

Выражение (12.12) есть отношение суммы квадратов отклонений, обусловленной зависимостью y от x к общей сумме квадратов отклонений. Для линейной связи это совпадает с определением коэффициента детерминации, но форма связи возможна любая.

Поясним сказанное примером вычисления корреляционного отношения. Возьмем уже ранее рассматриваемое распределение 238 деревьев дуба по диаметру и высоте (таблица 12.1). Для удобства пользования представим ее в виде таблица 12.6, т.к. нам требуется ввести дополнительные графы, а в том виде, как имеется в таблице 12.1, все графы затруднительно разместить на одном листе.

В таблице 12.6 знаком $\overline{y_2}$ обозначены групповые средние, вычисленные в таблице 12.1. Общее среднее (для совокупности) обозначено $\overline{y}$.

По данным таблицы 12.6 определяем

$$\overline{x} = \frac{5952}{238} = 25 .$$

Общая дисперсия

$$\sigma^2 = \frac{2758}{238-1} = \frac{2758}{237} = 11{,}64 .$$

Дисперсия групповых средних

$$\sigma_{yx}^2 = \frac{2305{,}1}{237} = 9{,}73 .$$

Таблица 12.6 – Вычисление корреляционного отношения для связи диаметров и высот в древостое дуба

Ступени толщины x_i	Ступени высот y_i	n_i	$y_i n_i$	$\bar{y}_2^{\text{í}}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(y_i - \bar{y})^2 \cdot n_i$	$\bar{y}_2 - \bar{y}$	$(\bar{y}_2 - \bar{y})^2$	$(\bar{y}_2 - \bar{y})^2 n_i$
12	16	3	48	16	-9	81	243	-9,0	81,0	243,0
16	16	3	48	18,0	-9	81	243	-7,0	49,0	441,0
	18	3	54		-7	49	177			
	20	3	60		-5	25	75			
	∑	9	162	-	-	-	465			
20	16	2	32	19,52	-9	81	162	-5,48	30,03	630,1
	18	5	90		-7	49	245			
	20	10	200		-5	25	250			
	22	4	88		-3	9	36			
	∑	21	410	-	-	-	693			
24	18	2	36	22,8	-7	49	98	-2,2	4,44	133,2
	20	2	40		-5	25	50			
	22	11	242		-3	9	99			
	24	12	288		-1	1	22			
	26	3	78		+1	1	3			
	∑	30	684	-	-	-	272			
28	22	3	66	24,54	-3	9	27	-0,46	0,21	9,2
	24	28	672		-1	1	28			
	26	11	286		+1	1	11			
	28	2	56		+3	9	18			
	∑	44	1080	-	-	-	74			
32	24	5	120	26,22	-1	1	5	+1,22	1,49	88,5
	26	38	988		+1	1	38			
	28	11	308		+3	9	99			
	∑	54	1416	-	-	-	142			
36	24	1	24	26,74	-1	1	1	+1,74	3,03	106,1
	26	22	572		+1	1	22			
	28	11	308		+3	9	99			
	32	1	32		+7	49	49			
	∑	35	936	-	-	-	171			
40	26	1	26	28,52	+1	1	1	+3,52	12,39	285,0
	28	15	420		+3	9	135			

	30	7	210		+5	25	175			
	∑	23	656	-	-	-	311			
44	28	7	196	29,18	+3	9	63	+4,18	17,47	297,0
	30	10	300		+5	25	250			
	∑	17	496	-	-	-	313			
48	30	1	30	31,0	+5	25	25	+6,0	36,0	72,0
	32	1	32		+7	49	49			
	∑	2	64	-	-	-	74			
∑	-	238	5952	25,0	-27		2758	-	-	2305,1

Корреляционное отношение

$$\eta_{yx} = \sqrt{\frac{9,73}{11,64}} = \sqrt{0,836} = 0,914 .$$

Оценка достоверности корреляционного отношения.
Основная ошибка:

$$m_\eta = \sqrt{\frac{1-n^2}{N-2}} = \sqrt{\frac{0,164}{236}} = \sqrt{0,0007} = 0,03 .$$

Критерий Стьюдента:

$$t = \frac{0,914}{0,03} = 30,5 .$$

Для $\gamma = N-2 = 236$ критическое значение t (приложение Е) даже при достоверности 99,9% равно 3,29<30,5, т.е. наше корреляционное отношение достоверно с высоким уровнем значимости.

Вспомним, что коэффициент корреляции (раздел 12.2) для этих же исходных данных составил 0,74. Значительная разница между величиной коэффициента корреляции и корреляционного отношения показывает, что связь между диаметрами и высотами в древостое носит криволинейный характер. Это хорошо известно лесоводам, которые выражают эту связь уравнением параболы 3 порядка $y=a_1+a_1x+a_2x^2+a_3x^3$ или иными функциями, имеющими схожие графики.

Для строгого установления, является ли связь криволинейной, существует мера криволинейности (K), выражаемая через соотношение r и η.

$$K = \eta^2 - r^2$$

При $K<0,1$ связь прямолинейная; $K \geq 0,1$ связь криволинейная.

Для нашего примера $K = 0,836 - 0,548 = 0,288$, что подтверждает криволинейность связи: 0,288>0,1.

Основную ошибку меры криволинейности (K) определяют по формуле

$$m_K = \frac{2}{\sqrt{N}}\sqrt{K - K^2(2-K)},$$

а ее достоверность проверяется по t-критерию Стьюдента

$$t = \frac{K}{m_K},$$

где t критическое берут по таблицам (приложение Е) для $\gamma=N$-1. При $t<t_{таб}$ мера криволинейности достоверна, а при $t\geq t_{таб}$ мера криволинейности недостоверна.

Для нашего примера

$$m_K = \frac{2}{238}\sqrt{0{,}288 - 0{,}288^2(2-0{,}288)} = 0{,}0084\sqrt{0{,}288 - 0{,}08294\cdot 1{,}712} = 0{,}0084\sqrt{0{,}288 - 0{,}142}$$
$$= 0{,}0084\sqrt{0{,}146} = 0{,}0084\cdot 0{,}382 = 0{,}032.$$

Для γ=237 $t_{таб}$=3,29, т.е. мера криволинейности достоверна в 99,9%.

Криволинейность связи может быть установлена и через F-критерий Фишера.

$$F = \frac{\eta^2 - r^2}{1-\eta^2}\cdot\frac{N-K}{K-2},$$

где N – объем выборки,

K – число классов вариационного ряда.

Критические значения F берут по таблице (приложение Ж) при числе степеней свободы $\gamma_1=K$-2; $\gamma_2=N$-K.

При $F_{факт}<F_{таб}$ связь прямолинейная; при $F_{факт}\geq F_{таб}$ связь криволинейная.

Для нашего примера

$$F = \frac{0{,}836 - 0{,}548}{1-0{,}914}\cdot\frac{238-10}{10-2} = \frac{0{,}288}{0{,}086}\cdot\frac{228}{8} = 3{,}3488\cdot 28{,}5 = 95{,}4.$$

Табличное $F_{таб}$ (при γ_1=236, γ_2=8) для достоверности 99,9% равно 9,6. Вычисленная нами величина $F_{факт}$=95,4. Она больше табличного значения, что доказывает криволинейность связи.

Заметим, что η, как правило, имеет большее значение, чем r. Лишь в случае, когда $r\approx$1,0, величина η приближается к r. Поэтому, если окажется, что при вычислениях $r \geq \eta$ (такое у неопытных студентов случается), то, значит, в вычислениях допущена ошибка.

Есть и другие критерии, показывающие вид связи (прямолинейность или криволинейность): критерии Бокмана, Романовского и др., но здесь мы их не рассматривали. Интересующихся отсылаем к учебнику «Биометрия» (авторы М.П. Горошко и др.), который приведен в списке литературы.

Как уже отмечено, корреляционное отношение показывает, какую часть общей дисперсии (вариации) результативного признака (y) составляет дисперсия частных средних этого признака, т.е. измеряет относительную степень варьирования групповых средних $\bar{y}_i$.

Можно вычислить два корреляционных отношения $\eta_{y/x}$ и $\eta_{x/y}$. Однако реальное значение имеет, как правило, один из них. Корреляционное отношение имеет всегда положительное значение, изменяющееся от 0 до 1. Когда групповые средние одинаковы (не варьируют), η=0. Связь отсутствует. В случае строго прямолинейной связи (все точки лежат на прямой) $\eta = r = 1$. В других случаях $\eta > r$. Чем это различие больше, тем связь более криволинейна. В предельном случае, когда связь строго криволинейна и кривая проходит через групповые средние, так что $\sigma_{\bar{y}i} = \sigma_y$, корреляционное отношение равно 1, а r=0.

Вычисление по формуле (12.12) значения η возможно лишь в том случае, когда выборка большая и данные расположены в виде корреляционной таблицы, как в таблицах 12.5 и 12.6.

При малом числе наблюдений показатель η недостаточно надежен. В группах может оказаться по одному значению y. Тогда $\sigma_{\bar{y}i} \approx \sigma_y$, а $\eta \approx 1{,}0$.

При малой выборке следует вводить корректирование η по формуле:

$$\eta_2 = 1 - \frac{(1-\eta^2)(N-1)}{N-m}, \qquad (12.13)$$

где m - число групп.

Ниже для примера по данным, приведенным в таблице 12.6, показан расчет η по названной формуле.

$$\eta_2 = 1 - \frac{(1-0{,}914^2)(238-1)}{238-10} = 1 - \frac{(1-0{,}8354)237}{228} = 1 - \frac{0{,}1646 \cdot 237}{228} =$$

$$1 - \frac{39{,}010}{228} = 1 - 0{,}17 = 0{,}83$$

Отсюда $\eta_2 = \sqrt{0{,}83} = 0{,}911$. Разница между величинами η, вычисленными по формулам 12.12 и 12.13 составила 0,3%, что предпставляет незначительную величину, т.е. значения η, полученные по обоим формулам совпали. Это произошло потому, что наше N=238 велико. При N<30 разница может быть существенной. Допустим, что при том же η=0,914 имеем N=29, а m=12. То-

гда η_2 будет равно η_2=1-0,243=0,757. $\bar{\eta}$=0,870. Как видим, разница составила $P = \frac{0{,}914 - 0{,}870}{0{,}914} \cdot 100\% \approx 5\%$, что уже значимо.

12.4 Другие статистические показатели корреляции. Использование корреляции в лесном хозяйстве

При проведении исследований не всегда ограничиваются нахождением коэффициента корреляции и корреляционного отношения. Бывает, что при наблюдении встречаются случаи качественно неизмеримые или те из них, когда проводились отметки только наличия или отсутствия признаков. Тогда производят ранжирование признаков, т.е. присваивают им определенные номера и определяют показатель корреляции рангов.

В практике есть примеры, когда характер распределения неизвестен, но он явно отличается от нормального. В этом случае, если показатели поддаются ранжированию, применяются ранговые коэффициенты корреляции. Опишем нахождение показателя корреляции рангов, в редакции А.К.Митропольского.

Ранг указывает то место, которое занимает данная единица частичной совокупности среди других ее единиц. Если бы каждая из этих единиц отличалась в отношении рассматриваемого признака от всех других единиц совокупности, тогда ранги представляли бы порядковые номера от 1 до числа *n*, равного объему частичной совокупности. Если же некоторые из единиц совокупности оказываются в отношении рассматриваемого признака одинаковыми, тогда ранг всех этих единиц принимается равным среднему из соответствующих номеров.

Показатель корреляции рангов ρ равен

$$\rho = 1 - \frac{6\sum_{h=1}^{n} d_h^2}{n(n^2-1)}, \qquad (12.14)$$

где величины d_h представляют разности между рангами единиц, извлеченных совместно из двух общих совокупностей.

Подобно коэффициенту корреляции *r*, показатель корреляции рангов ρ изменяется от -1 до + 1. Чем теснее будет связь между величинами, тем ближе к единице по своей абсолютной величине будет показатель корреляции рангов; знак же показателя указывает, является ли связь прямой или обратной.

В качестве примера вычислим показатель корреляции рангов, выражающий связь между объемным весом X_1 *(γ, г/см^3)* и пределом прочности при сжатии X_2 *(σ_B, кг/см^3)* древесины дуба (таблица 12.7, которую приводит А.К. Митропольский).

Для проверки вычислений заметим, что сумма рангов должна равняться сумме натуральных чисел от 1 до n, т.е. $\frac{n(n+1)}{2}$. В рассматриваемом примере

$$\frac{19(19+1)}{2}=190.$$

Кроме того, сумма отрицательных разностей d_h должна равняться сумме положительных разностей.

Применяя формулу (12.14), находим

$$\rho = 1-\frac{6\cdot 171{,}5}{19(19^2-1)} = 0{,}85$$

Таблица 12.7 – Вычисление показателя корреляции рангов

№	X_1	X_2	h_1	h_2	d_h	d_h^2
1	758	676	19	19	0	0
2	739	646	17	13	4	16
3	719	642	13	12	1	1
4	717	635	10	9,5	0,5	0,25
5	678	567	5	5	0	0
6	613	476	2	2	0	0
7	753	635	18	9,5	8,5	72,25
8	724	665	14,5	18	-3,5	12,25
9	718	661	12	17	-5	25
10	693	610	7	7	0	0
11	658	531	4	3	1	1
12	737	649	16	14	2	4
13	707	635	8	9,5	-1,5	2,25
14	717	635	10	9,5	0,5	0,25
15	653	544	3	4	-1	1
16	724	650	14,5	15	-0,5	0,25
17	717	653	10	16	-6	36
18	685	578	6	6	0	0
19	612	462	1	1	0	0
∑	-	-	190	190	-17,5 17,5	171,5
$\rho = 1-\frac{6\cdot 171{,}5}{19(19^2-1)} = +0{,}85$						

Ранговый коэффициент корреляции Спирмена изложен здесь в редакции М.П. Горошко с соавторами. Уравнение коэффициента корреляции рангов имеет вид

$$r_s = 1 - \frac{6\sum d_i^2}{N^3 - N}, \qquad (12.15)$$

где d – разница между рангами;

x_i и y_i – ранги значений x_i и y_i $(1 \le x_i \le N;\ 1 \le y_i \le N)$;

N – объем выборки.

Приведенный коэффициент является непараметрическим показателем связи и используется при изучении как количественных, так и качественных значений. При этом закон распределения этих величин и форма связи нам могут быть неизвестны, как и обычный коэффициент корреляции r_s может иметь величину от -1 до +1.

При независимости величин x и у r_s распределяется ассимптотично нормально при $N \to \infty$, а $\bar{\delta} = 0$, где дисперсия $\sigma^2 = \frac{1}{N-1}$.

Как и другие выборочные показатели, коэффициент корреляции рангов есть оценка параметра в генеральной совокупности. Его достоверность устанавливают на основе нулевой гипотезы, которую принимают для условия, что $r_s = 0$ в генеральной совокупности. Эта гипотеза проверяется путем вычисления t-критерия Стьюдента и сопоставления его с критической величиной $t_{кр}$.

$$t = r_s \sqrt{\frac{N-2}{1 - r_s^2}}.$$

Обычно используют 95% или 99% уровень достоверности, где применяют следующие формулы.

Для уровня достоверности 99%:

$$t_{кр\%} = \frac{2{,}58}{\sqrt{N-1}}\left(1 - \frac{0{,}69}{N-1}\right).$$

Для уровня достоверности 95%:

$$t_{кр\%} = \frac{1{,}96}{\sqrt{N-1}}\left(1 - \frac{0{,}16}{N-1}\right).$$

При $t < t_{таб}$ принимают нулевую гипотезу (H_0), т.е. r_s недостоверен.

При $t \ge t_{таб}$ принимается рабочая (альтернативная) гипотеза (H_a) и r_s является достоверным.

Достоверность r_s можно оценить и по специальным таблицам критических значений r_s (приложение У), где оно дается в зависимости от уровня значимости (5%, 1%) и объема выборки (N). При $r_s < r_{таб}$ принимается нулевая гипотеза, а $r_s \ge r_{таб}$ – альтернативная или рабочая.

Приведем пример вычисления r_s (таблица 12.8). Пусть у нас есть замеры 15 высот и диаметров в древостое осины с полнотой 0,7 I^a класса бонитета в 40 лет.

$\overline{\delta_d} = 21{,}3$, $\overline{x_i} = 21{,}5$.

Сумма рангов (по x и y) контролируется по формуле арифметической прогрессии $s = \frac{a_1 + a_n}{2} \cdot n$. У нас $s = \frac{1+15}{2} \cdot 15 = 120$.

Величину r_s определяем по формуле (12.15):

$$r_s = 1 - \frac{6 \cdot 28}{3375 - 15} = 1 - \frac{168}{3360} = 1 - 0{,}05 = 0{,}95.$$

Как видим, коэффициент ранговой корреляции высок, что соответствует известной закономерности в отношении связи диаметров и высот деревьев в древостое. В этом убедимся, вычислив достоверность r_s.

Таблица 12.8 – Вычисление коэффициента ранговой корреляции для 15 стволов осины

№ п/п	Диаметр, см x_i	Высота, м y_i	Ранги (P)		$d_1=P_1-P_2$	d_1^2	P_1P_2
			P_1 по x_i	P_2 по y_i			
1	13,8	17,5	1,0	1,0	0	0	1,0
2	16,9	18,4	2,5	2,0	0,5	0,25	5,0
3	16,8	19,5	2,5	4,5	-2,0	4,0	11,25
4	17,9	19,2	4,0	3,0	1,0	1,0	12,0
5	18,5	20,1	5,0	6,0	-1,0	1,0	30,0
6	19,2	19,5	6,5	4,5	2,0	4,0	29,25
7	19,2	21,5	6,5	8,0	-1,5	2,25	52,0
8	20,3	22,0	8,0	9,5	-1,5	2,25	76,0
9	21,4	22,0	9,0	9,5	-0,5	0,25	85,5
10	22,5	21,3	10,0	7,0	3,0	9,0	170,0
11	23,0	23,6	11,5	12,0	-0,5	0,25	138,0
12	23,0	22,9	11,5	11,0	0,5	0,25	126,5
13	27,6	24,0	13,0	13,5	-0,5	0,25	175,5
14	28,5	26,5	14,0	15,0	-1,0	1,0	210,0
15	30,9	24,0	15,0	13,5	1,5	2,25	202,5
∑	319,4	322,0	120	120	0	28,0	1324,5

Основная ошибка r_s определяется по формуле:

$$m_r = \sqrt{\frac{1 - r^2}{N - 2}} = \sqrt{\frac{1 - 0{,}95^2}{15 - 2}} = \sqrt{\frac{1 - 0{,}9025}{13}} = \sqrt{\frac{0{,}0975}{13}} = \sqrt{0{,}0075} = 0{,}0866.$$

$$t = \frac{r_s}{m_{rs}} = \frac{0{,}95}{0{,}0866} = 10{,}97 \approx 11{,}0.$$

По таблицам (приложение У) для $\gamma=N=15$ для $P=0{,}001$ находим $t_{таб}=3{,}95$.

Поскольку $t_{факт} > t_{таб}$ (11,0>3,95) устанавливаем, что величина r_s достоверна на высоком уровне значимости.

Есть и другие показатели ранговой корреляции, коэффициенты взаимной сопряженности, описанные в учебниках А.К. Митропольского, М.П. Горошко с соавторами, но эти показатели здесь опущены из-за их редкого использования.

Использование корреляции в лесном хозяйстве. В лесном хозяйстве приведенные величины, определяющие корреляцию между различными признаками, применяются очень широко. Уже более 100 лет коэффициент корреляции и корреляционное отношение постоянно вычисляют при проведении научных исследований в лесном хозяйстве. Эти показатели мы встречаем в трудах классиков лесоводства и лесной таксации М.М. Орлова, А.К. Турского, А.В. Тюрина, В.К. Захарова и др. Применение корреляции мы видим в классических трудах видного белорусского (гомельского) ученого, лесовода и таксатора Ф.П. Моисеенко. Он установил высокую степень корреляции между формой ствола и средним диаметром дерева и доказал небольшую изменчивость средней формы ствола в древостое. Это стало научной основой для составления объемных и сортиментных таблиц по средней форме ствола, что в несколько раз упростило составление этих нормативов и пользование ими.

В настоящее время вычисление r и η является обязательным условием проведения исследований в случае наличия связи между изучаемыми величинами. Применение ПК сделало эту работу простой и рутинной.

При исследовании лесных биогеоценозов очень часто влияние одних показателей биоценоза на другие не очевидно. Для доказательства (или опровержения) этого влияния пользуются показателями корреляции. Например, влияет ли процент физической глины в разных почвенных горизонтах (A_1, A_2, B, C) на величину класса бонитета без вычисления показателей корреляции для каждого типа леса (типа условий, местопроизрастания) сказать трудно, т.к. в ряде случаев (сухие боры, мокрые условия произрастания) велико влияние и других факторов, например, увлажнения почвы. Таких примеров очень много.

В то же время использование законов корреляции должно сопровождаться логическим анализом сути явления, что могут сделать только специалисты. В лесном хозяйстве это лесоводы, в отдельных случаях биологи (ботаники, зоологи). Механическое сопоставление логически несвязанных величин с вычислением показателей корреляции может привести к неверным, а иногда и абсурдным выводам. Например, мы можем связать величину, характеризующую количество лунного света с уровнем производительности древостоя и даже получим какие-то величины коэффициента корреляции, но результат таких расчетов не будет отражать реальность.

Другой пример. Мы можем связать количество сухостоя на 1 га древостоя только с плодородием почвы, но эти сопоставления будут некорректны,

т.к. на количество сухостоя (отпада) на 1 га влияют (более значимо) другие факторы: густота насаждения, его возраст, рубки ухода.

Здесь, в частном случае, проявляются общие требования к определению статистических показателей – необходимость их логической верификации специалистами.

13. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ КАК ИНСТРУМЕНТ НАУЧНОГО ИССЛЕДОВАНИЯ

13.1 Цель и задачи корреляционного анализа

Ранее (глава 12) мы разобрали связи между двумя величинами. Установили, что они бывают прямолинейные и криволинейные. Для характеристики тесноты связи используется коэффициент корреляции (r) и корреляционное отношение (η).

Но, установив сам факт зависимости одной величины от другой, получили еще недостаточную информацию. Необходимо знать форму этой связи, ее числовое выражение. Тогда, имея одну из коррелирующих величин, обычно ту, которая легче и проще определяется, можно с определенной долей вероятности предсказывать значение другой (искомой) величины $y=f(x)$. Для решения этой задачи используют корреляционные уравнения или, как их еще называют, корреляционные модели.

Мы уже отмечали, что определений моделей есть много, но лучше других их выражает следующая формулировка, данная В.А. Штоффом: «Модель – это мысленно представляемая или материально реализуемая система, которая отображая или воспроизводя объект исследования, способна замещать его так, что ее изучение дает нам новую информацию об этом объекте». Именно это определение используют в лесных исследованиях К.Е. Никитин и А.З. Швиденко.

Следовательно, модель - это некоторая абстракция, возможная на определенной стадии изучения некоторого предмета или явления, для дальнейшего его познания или используемая для решения практических задач.

Корреляционные уравнения представляют собой разновидность стохастических моделей. При исследовании моделей, выражающих некоторую зависимость между изучаемыми величинами, еще раз подчеркнем, что это может быть и взаимозависимость, либо только зависимость. Примером взаимозависимости служит соотношение диаметра и высоты в древостое, а зависимости – изменение средней высоты или прироста при разном плодородии почвы, но не наоборот.

Теория корреляции разработана в основном в конце XIX и начале XX века Карлом Пирсоном и Юлом. Она позволяет описать разные связи, но не вскрывает причины их происхождения. Здесь нужен специальный анализ: биологический, лесоводственный, генетический и т.п. При этом причинную связь, изучая корреляцию, мы должны знать, т.к. иначе можем совершить ошибку, найдя связь там, где ее нет. Об этом хорошо

сказал великий английский писатель Бернард Шоу (1856-1956) еще в 1906 году в предисловии к «Доктору на распутье»: «Даже опытные статистики часто оказываются не в состоянии оценить, до какой степени смысл статистических данных искажается молчаливыми предположениями их интерпретаторов... Легко доказать, что ношение цилиндров и зонтиков расширяет грудную клетку, удлиняет жизнь и дает относительный иммунитет от болезней... Математик, чьи корреляции привели бы в восхищение Ньютона, может собирая данные и делая из них выводы впасть в совершенно грубые ошибки на основе таких популярных заблуждений, как описанные выше». Здесь Б. Шоу подчеркнул, что не сами цилиндры и зонтики приводят к описанным следствиям, а образ жизни их обладателей, которыми в те времена были богатые люди.

Ранее уже отмечено, что и в начале широкого распространения статистики, а затем в 60-70-е годы, когда математические методы стали широко применять в лесном хозяйстве, математики, слабо разбирающиеся в причинных связях предметов, которые они описывали с помощью корреляционных уравнений, совершили много ошибок. Об этой опасности предупреждали основатели учения о корреляции. Так, Юл в 1926 году напугал ученых примерами высоких корреляций между количеством самоубийств в Англии и принадлежностью к англиканской церкви. Причинной связи здесь нет, а высочайшая корреляция есть. Дело здесь объясняется просто – подавляющее большинство жителей Англии в те годы принадлежало к англиканской церкви.

Поэтому еще раз напомним о важности проведения профессионального анализа причинно-следственных связей, прежде чем взяться за конкретные вычисления.

13.2 Множественная корреляция

При рассмотрении корреляции часто встречаются случаи, когда две величины, вроде бы взаимозависимы, но более подробный анализ показывает, что эта «взаимозависимость» есть отражение того факта, что они обе коррелированны с некоторой третьей величиной или с совокупностью величин. В природе мы часто встречаемся с такими явлениями, когда изменение одной величины (функции) определяется изменением не одного, а нескольких аргументов.

Например, высота дерева зависит от почвенного плодородия, количества влаги в корнеобитаемом слое почвы, древесной породы, возраста древостоя и т.д. Величина суммы площадей сечения или запас древостоя зависят от высоты (H) и полноты насаждения. Включение в это уравнение диаметра древостоя (D) будет лишним из-за высокой корреляции H-D. Аналогичный пример можно привести, рассматривая коэффициент формы древесного ствола (q_2), который равен частному от деления диаметра на половине высоты дерева к диаметру на 1,3 м: $q_2 = D_{0,5м}/D_{1,3м}$. Рассматривая корреляцию q_2-D, мы можем придти к выводу о ее наличии. На самом деле есть корреляция q_2-H и H-D.

Такие закономерности приводят нас к понятию множественной корреляции. Ее суть заключается в следующем. Коэффициент множественной корреляции – это показатель тесноты связи (линейной) между одной зависимой величиной и совокупностью независимых.

Если у нас есть 3 сопряженные величины *X, Y, Z*, то коэффициент множественной корреляции определяется из матрицы

$$R_{XYZ} = \begin{vmatrix} 1 & r_{XY} & r_{XZ} \\ r_{YX} & 1 & r_{YZ} \\ r_{ZX} & r_{ZY} & 1 \end{vmatrix}$$

Решение этой матрицы приводит к уравнениям для определения R_{XYZ}, R_{YXZ}, R_{ZXY}:

$$R_{XYZ} = \sqrt{\frac{r_{XY}^2 + r_{XZ}^2 - 2r_{XY}r_{XZ}r_{YZ}}{1 - r_{YZ}^2}} \qquad (13.1)$$

$$R_{YXZ} = \sqrt{\frac{r_{XY}^2 + r_{YZ}^2 - 2r_{XY}r_{XZ}r_{YZ}}{1 - r_{XZ}^2}} \qquad (13.2)$$

$$R_{ZXY} = \sqrt{\frac{r_{XZ}^2 + r_{YZ}^2 - 2r_{XY}r_{XZ}r_{YZ}}{1 - r_{XY}^2}} \qquad (13.3)$$

где r_{xy}, r_{xz}, r_{yz} – парные коэффициенты корреляции между величинами *X, Y, Z*.

Вместо коэффициента множественной корреляции, который обычно обозначают как R в отличие от парного – r, часто при моделировании удобнее использовать коэффициент множественной детерминации – $D=R^2$ (в отличие от парного $d=r^2$). Он измеряет ту долю общей дисперсии зависимой переменной y, которая может быть объяснена влиянием изменения аргументов.

Значимость коэффициента множественной корреляции оценивают по F-критерию Фишера

$$F = \frac{R^2}{1 - R^2}\left(\frac{N - k}{k - 1}\right),$$

где N – объем выборки;

k – количество факторов влияния (число независимых переменных).

Критические значения F берут из таблиц (приложение Ж) при числе степеней свободы $\gamma_1=k$-1; $\gamma_2=N-k$. При $F_{факт}<F_{таб}=H_0$, т.е. принимается нулевая гипотеза об отсутствии корреляции и, наоборот, принимают альтерна-

тивную или рабочую гипотезу, т.е. о наличии корреляции, при $F_{факт} \geq F_{таб}$. Величина R (как и частные r) колеблется в пределах (-1)-(0)-(1).

Пределы (l), в которых может находиться R_{YZ}, если известны R_{XY} и R_{XZ} определяют по формуле

$$l = R_{XY} \cdot R_{XZ} - \sqrt{(1 - r_{XY}^2)(1 - r_{XZ}^2)} < R_{YZ} < R_{XY} R_{XZ} + \sqrt{(1 - r_{XY}^2)(1 - r_{XZ}^2)} \qquad (13.4)$$

Для того чтобы выяснить долю влияния одной из величин на другую в общей системе нескольких взаимовлияющих показателей, введено понятие частного коэффициента корреляции. Графически это можно выразить следующим образом.

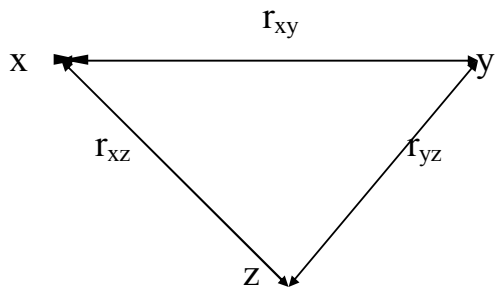


Рисунок 13.1 – Графическая интерпретация частной корреляции

Если существует тесная корреляция между x-z и y-z, то связь между x-y может создаваться за счет одновременного влияния на x и y третьего признака z. Чтобы найти то влияние, которое оказывает x на y (или y на x), надо исследовать зависимость x-y при постоянном z, т.е. x и y будут изменяться, а $z=const$, т.е. z мы элиминируем.

Вычисленный в этом случае коэффициент корреляции x-y называется частным коэффициентом корреляции. Формула его вычисления следующая

$$r_{xy\text{-}(z)} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \qquad (13.5)$$

$$r_{xz\text{-}(y)} = \frac{r_{xz} - r_{xy} \cdot r_{zy}}{\sqrt{(1 - r_{xy}^2)(1 - r_{zy}^2)}} \qquad (13.6)$$

$$r_{zy\text{-}(x)} = \frac{r_{zy} - r_{zx} \cdot r_{yx}}{\sqrt{(1 - r_{zx}^2)(1 - r_{yx}^2)}} \qquad (13.7)$$

Статистиками множественной нелинейной корреляции являются множественные и частные корреляционные отношения.

Эмпирическое множественное корреляционное отношение $\eta_{1.23}$ величины X_1 к величинам X_2 и X_2 дается выражением

$$\eta^2_{1.23} = \sum_{j_2=1}^{k_2}\sum_{j_3=1}^{k_3} p'_{j_2 j_3} \cdot r^2_{1(j_2 j_3)} \qquad (13.8)$$

или равносильными ему выражениями

$$\eta^2_{1.23} = 1 - \frac{\sum_{j_2=1}^{k_2}\sum_{j_3=1}^{k_3} p'_{|j_2|j_3} \cdot \mu_{2|(j_2)|(j_3)}}{\mu_{2/0/0}}, \qquad (13.9)$$

$$\eta^2_{1.23} = \frac{\sum_{j_2=1}^{k_2}\sum_{j_3=1}^{k_3} p'_{/j_2/j_3} \cdot (m_{1/(j_2)/(j_3)} - m_{1/0/0})^2}{\mu_{2/0/0}}. \qquad (13.10)$$

Эмпирическое частное корреляционное отношение $\eta_{12.3}$ величины X_1 и X_2 при данном значении X_3 вычисляется по формуле

$$\eta^2_{12.3} = 1 - \frac{\sum_{j_2=1}^{k_2} p'_{/j_2/(j_3)} \cdot \mu_{2/(j_2)/(j_3)}}{\mu_{2/0/(j_3)}}, \qquad (13.11)$$

или

$$\eta^2_{12.3} = \frac{\sum_{j_2=1}^{k_2} p'_{/j_2/(j_3)} \cdot (m_{1/(j_2)/(j_3)} - m_{1/0/(j_3)})^2}{\mu_{2/0/(j_2)}}, \qquad (13.12)$$

где m_{ij}, μ_{ij} – соответствующие моменты, описанные в разделе 12;

p'_{ij} – частости, описанные там же.

В качестве примера применения частного коэффициента корреляции в лесном хозяйстве приведем установленные зависимости q_2 от d и h, где q_2 – второй коэффициент формы, d – диаметр дерева на 1,3 м, h – высота дерева.

Коэффициент множественной корреляции здесь достигает 0,95-0,96. В то же время частные коэффициенты корреляции оказались r_{HD} – 0,75-0,90; r_{DH} – 0,73-0,92; r_{Hq2} – 0,90-0,92; r_{Dq2} – 0,16-0,20.

Долгое время считалось, что q_2 зависит от обеих величин d и h. Подобную ошибку совершил даже крупный ученый-таксатор Матвеев-Мотин. Лишь Ф.П. Моисеенко, изучив частную корреляцию q_2 - d и q_2 - h, доказал, что величина коэффициента корреляции в связи q_2 - d опре-

деляется высокой корреляцией d - h. При фиксированных h коэффициент корреляции q_2 - d оказался не выше 0,2, т.е. этой связи практически не было. Почему же до Ф.П. Моисеенко ученые впадали в ошибку? Ведь как методически построить исследование большинство из них знало. Но для этого нужен был огромный экспериментальный материал. Имея около 18 тысяч срубленных и обмеренных деревьев, Ф.П. Моисеенко такой материал собрал, смог вычислить частные r, а у других столь обширных замеров не было. В то же время в те годы, когда проводилась эта работа (1936-1941 гг.), подобные вычисления требовали длительного времени, на что не все шли. Теперь на ПК такая работа делается очень быстро.

13.3 Корреляционные модели

Корреляционные модели представляют собой уравнения, где на основе связи или взаимозависимости двух величин строится уравнение, в котором одна из величин выступает в качестве аргумента, а другая - функции, т.е. $y = f(x)$. Построение таких моделей связано с вычислением коэффициента корреляции, т.е. здесь предполагается линейная зависимость. Из курса математики мы знаем, что линейная зависимость реализуется в виде уравнения прямой. Ее формула

$$y = a + вx \qquad (13.13)$$

где a, $в$ – некоторые коэффициенты.

Здесь коэффициент a показывает величину отступления от начала координат, а $в$ – угол наклона прямой к оси ox.

Приведем пример. Для этого воспользуемся результатами замеров диаметров и высот сосны в молодом возрасте (таблица 13.1).

Таблица 13.1 – Результаты измерения диаметров и высот молодых деревьев сосны

№ п/п	Диаметр, см	Высота, м	№ п/п	Диаметр, см	Высота, м	№ п/п	Диаметр, см	Высота, м
1	7,6	8,7	6	8,8	9,8	11	8,2	9,2
2	6,4	7,5	7	9,0	10,1	12	4,0	5,6
3	5,2	6,6	8	6,2	7,2	13	5,6	7,0
4	4,1	5,7	9	7,0	8,1	14	4,8	6,5
5	3,7	5,2	10	7,4	8,5	15	6,9	8,0

Для того, чтобы оценить вид модели и общую закономерность изменения функции в зависимости от изменения аргумента построим график. На оси абсцисс будем откладывать диаметры, на оси ординат - высоты (рисунок 13.2)

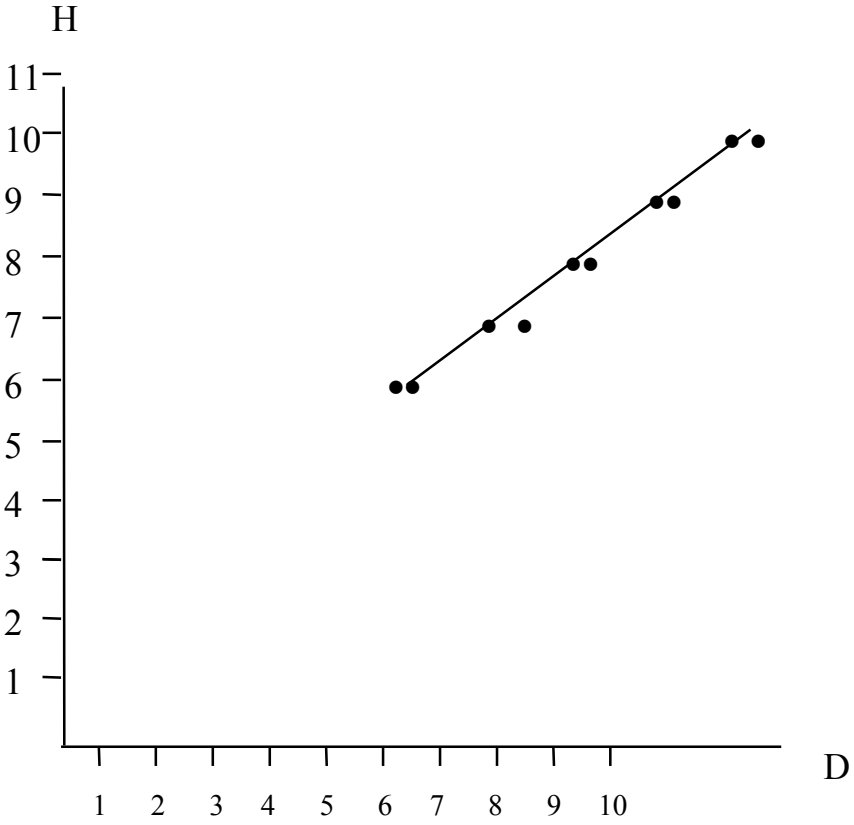


Рисунок 13.2 Зависимость высот и диаметров в сосновых молодняках

Из рисунка 13.2 видно, что точки, показывающие соотношение *D-H* лежат практически на прямой линии. Ранее в лесных исследованиях ограничивались графическим построением прямой. Достоинства графического построения очевидны: простота и наглядность. Его недостатками является определенный субъективизм. Даже анализируя убедительный рисунок 13.2, можно разойтись во мнении, как провести прямую линию, хотя расхождения у разных исполнителей будут невелики.

Корреляционные уравнения бывают не только двухмерные (парные), но и многомерные. Их вид следующий:

$$y=ax_1 + bx_2 + cx_3 + ... + nx_n \qquad (13.14)$$

Графическое изображение множественных связей (после трех аргументов) затруднительно. Поэтому графические построения здесь обычно не делают.

Вычисление коэффициентов корреляционных уравнений осуществляют по специальным методикам, которые мы рассмотрим в главе 14.

Здесь же еще остановимся на некоторых особенностях применения корреляционных уравнений.

В общую формулу корреляционного уравнения все показатели проставляются в своих единицах измерения: действия производятся над их

абсолютными величинами, не обращая внимания на наименования. Ответ получается в единицах измерения зависимой величины. Для проверки правильности полученного уравнения в него подставляется среднее значение независимого признака в его единицах измерения, а в ответе должно получиться среднее значение зависимого переменного, равное исходному.

Корреляционное уравнение применяется для получения точного среднего значения зависимого переменного, за которое принимается трудно измеримый признак, зная точное среднее значение независимого переменного, за которое принимается более быстро и легко измеримый признак, корреляционно взаимосвязанный с первым. Так при изучении диаметров и высот древостоя за независимую переменную принимают диаметр на высоте 1,3 м, который значительно легче, проще и точнее измеряется, чем высота.

В этом случае необходимое число наблюдений для зависимого признака требуется значительно меньше того, какое было бы необходимо при самостоятельном анализе этого признака, и вычисляется по формуле (если точность задана в процентах)

$$n = \frac{V^2}{p^2}(1 - r^2), \qquad (13.15)$$

а точность опыта

$$p = \frac{V}{\sqrt{n}} \cdot \sqrt{1 - r^2} \qquad (13.16)$$

Вспомним, что для одного статистического ряда $n = \frac{V^2}{p^2}$.

Пусть требуется определить среднюю высоту с точностью 1 %, зная точный средний диаметр на отведенном участке леса. По данным, имеющимся в лесотаксационной литературе, известно, что коэффициент вариации высот в пределах спелого древостоя равен 12 - 15%. Значит, для получения средней высоты с точностью 1% потребовалась бы измерить $n = \frac{V^2}{p^2} = 15^2/1^2 = 225$ стволов, что трудно выполнимо.

Поскольку есть высокая корреляционная связь между диаметрами и высотами, то для получения средней высоты с той же точностью требуется значительно меньшее количество измеренных высот. Так, при обычном коэффициенте корреляции между D и H, равном 0,90, оно составит всего $225(1\text{-}0{,}90^2) = 225(1\text{-}0{,}81) = 43$. Измерив у 43 деревьев высоты и диаметры, обрабатываем полученные данные и составляем корреляционное уравнение, в которое подставляем точную величину среднего диаметра. Среднюю высоту затем находим по корреляционному уравнению с заданной точностью.

Так, пусть по данным измерения высот у 43 стволов уравнение связи получилось h=0,40d+14,2 м, а точный средний диаметр 32,2см; отсюда искомая высота равна h = 0,40·32,2+14,2 = 12,9+14,2 = 27,1 м. При этом оказалось: V = 15%, а r = 0,90. Точность опыта, т.е. полученной средней высоты, будет равняться

$$- p = \frac{15}{\sqrt{43}} \cdot \sqrt{1 - 81} = 2{,}27 \cdot 0{,}436 = 0{,}99 = 1\%$$

Угловой коэффициент корреляционного уравнения $R = r \frac{\sigma_y}{\sigma_x}$ называется **коэффициентом регрессии**; он показывает, на сколько единиц в среднем изменяется зависимый признак, если независимый изменился на одну единицу.

В вышеприведенном примере ($\sigma_д$=6,96 и $\sigma_н$=2,69 вычислены ранее) угловой коэффициент получился

$$R = 0{,}90 \cdot \frac{2{,}69}{6{,}96} = 0{,}90 \cdot 0{,}386 = 0{,}35 \text{ м.}$$

Это значит, что при изменении диаметра на 1 см высота деревьев в среднем изменяется на 0,36 м.

В общем виде линейная регрессия между двумя переменными величинами может быть выражена с помощью следующего общего уравнения

$$y = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}), \qquad (13.17)$$

где y, x – переменные вариационных рядов;

$\bar{y}$, $\bar{x}$ - средние величины рядов;

σ_y, σ_x - среднеквадратические отклонения вариационных рядов;

r – коэффициент корреляции.

Основная ошибка уравнения корреляционной связи определяется по формуле

$$m_{y/x} = \sigma_y \sqrt{1 - r^2}.$$

13.4 Применение корреляционных уравнений в лесном хозяйстве

Корреляционные уравнения очень широко применяются как в лесных исследованиях, так и в практической работе, особенно в лесоустройстве и при проведении проектных работ. Примеров этому очень много.

Так, связь диаметров и высот и конкретное выражение этой связи самым непосредственным образом используют при построении сорти-

ментных и товарных таблиц. Первые обычно строят по разрядам высот. Для определения разряда высот используют замеры диаметров и высот у 9-12 деревьев. Далее для вычисления объема деревьев применяют специальные сортиментные таблицы, где заложены уравнения связи *H-D*, выведенные при проведении научных исследований. Ранее, до выведения таких связей, были таблицы, которые требовали измерения высоты у каждого дерева. Это очень трудоемко. Использование здесь корреляционных уравнений снизило затраты труда на отводах лесосек в десятки и сотни раз.

Продолжая тему сорментации отметим, что для вычисления объемов стволов и выхода сортиментов у сосны после подсочки, когда форма ствола на высоте 1,3 м деформирована карами, с которых собирают живицу, использована связь диаметра на 1,3 м, измеренного по ремням, т.е. по неповрежденной части ствола и диаметра на 0,5 *H*.

Запас древостоя при проведении лесоинвентаризации инженер-таксатор определяет по таблице, где входами служат высота и полнота, т.е. используется корреляция. Очень широко используются корреляционные зависимости в охотоведении. Например, по форме следов и их размерам опытные охотоведы определяют пол, возраст, состояние животного.

Велико значение корреляции в защите леса от вредителей. Есть уравнения описывающие характер размножения вредителя, ожидаемый ущерб от состояния популяции в некоторый момент времени, погодных условий и состояния лесного насаждения. Использование таких уравнений позволит сделать прогнозы массового размножения вредителей и принять превентивные меры защиты насаждений.

Нахождение величин *y* по корреляционным уравнениям из формулы $y = a + вх$ называется **выравниванием или аппроксимацией**. Выравненные значения функции отражают закономерное ее изменение (с некоторой основной ошибкой), где исключены случайные влияния на отдельные измеренные значения вариационного ряда.

В научных исследованиях необходимо не просто измерить или определить какие-то величины, но и найти закономерные связи между ними, определить то общее, что связывает функцию и аргументы и выразить это математически.

Более общие случаи определения зависимости $y = f(x)$ будут рассмотрены в следующей главе.

14. РЕГРЕССИОННЫЙ АНАЛИЗ

14.1. Сущность регрессионного анализа. Регрессионные модели

Рассмотренные в предыдущей главе корреляционные уравнения являются частным случаем более общего вида вероятностных связей, которые выражаются методами регрессионного анализа. Поясним его суть. Пусть у нас имеется некоторая функция $y = f(x)$. Она может приобретать разные выражения в зависимости от того, что подразумевается под $f(x)$. Например, уже известное нам выражение уравнения прямой $y = a + bx$ или уравнение параболы второго порядка $y = a + bx + cx^2$, или гиперболы $y = a + b / x$ и т.д.

Для обозначения разных связей в биологической статистике, т.е. в биометрии английским ученым Ф. Гальтоном предложен термин р е г р е с с и я. Смысл этого термина состоит в том, что коррелирующие пары в биологических объектах, обнаруживающие в потомствах отклонения от средней линии, определяющей корреляцию признаков совокупности, имеют тенденцию возврата к этой средней, если только действуют одни случайные причины. В дальнейшем мы будем пользоваться этим термином как более общим, говоря об уравнениях стохастической связи между случайными величинами.

Регрессионные модели - это уравнения стохастической связи вида $y = f(x)$, где $f(x)$ может быть выражено в любом виде.

Корреляционные модели - частный случай регрессионных, когда связь носит прямолинейный характер.

Регрессионные модели обычно используют для выражения разного рода связей в лесной таксации, лесоводстве и в других лесных дисциплинах. Чаще всего они применяются для нахождения общей зависимости по экспериментальным данным. В этом случае выведенное уравнение служит для выравнивания материала, полученного при постановке опыта. При этом сохраняется главная тенденция изменения функции в зависимости от изменения аргументов, и устраняются случайные отклонения. Такое уравнение удобно использовать для получения величин функции через равный шаг аргумента, хотя в опытном материале этот равный шаг не всегда выдерживается. Например, нам удобнее, изучая ход роста древостоев, иметь данные через 10 лет: в 10, 20, 30, ... 100 лет, а наши пробные площади имеют возраст 9, 22, 29, 44, 57, ... 102 года. Поэтому необходимо найти промежуточные значения в 10, 20, 30 лет по данным заме-

ров пробных площадей, что обычно называют сглаживанием опытных данных.

14.2 Методы определения вида регрессионных уравнений и их параметров

При обработке опытных данных очень часто приходится решать задачу, в которой необходимо исследовать зависимость одной физической или биологической величины y от другой физической или биологической величины x. Например, зависимость продуктивности древостоя от количества осадков и средних температур, размеров лося от величины его следа, объема дерева от количества физической глины в почвенных горизонтах A_1, A_2, B_1, C, коэффициента формы ствола от его высоты и т.д.

Пусть производится опыт с целью исследования зависимости величины y от величины x, которая в общем случае может быть записана в виде

$$y = f(x).$$

Вид этой зависимости и требуется определить из опыта.

Допустим, что мы исследуем зависимость видового числа древостоя (F_q) от его средней высоты (H). В результате опыта получен ряд экспериментальных точек (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) , которые приведены в таблице 14.1. По данным таблицы 14.1 построен график изменения переменной величины $y(F_q)$ пр разных независимых переменных x(H) (рисунок 14.1). На нем показана связь рассматриваемого видового числа (зависимая переменная – y) от высоты H (независимая переменная - x). Эта зависимость обычно выражается уравнением гиперболы

$$y = a + \frac{b}{x}.$$

Таблица 14.1 – Средние видовые числа (F_g) древостоя ели в зависимости от средней высоты (H)

H	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36
$F_g \cdot 0{,}001$	615	568	541	522	509	495	492	486	482	476	472	469	467	465	464	463

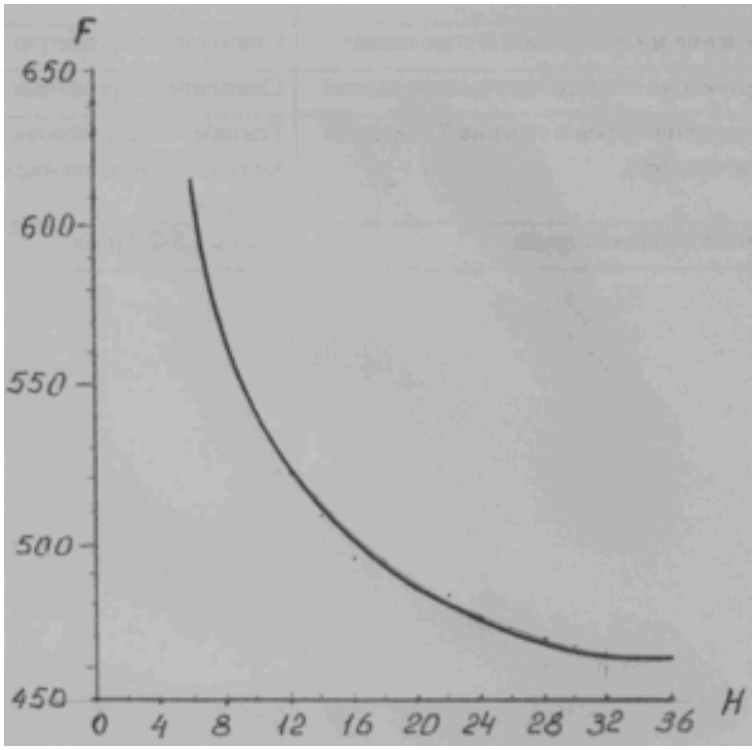


Рисунок 14.1 График изменения видового числа древостоя ели в зависимости от средней высоты

Так как проводимые в ходе опыта измерения связаны с ошибками случайного характера, то обычно экспериментальные точки на графике имеют некоторый разброс относительно общей закономерности. В силу случайности ошибок измерения этот разброс или уклонения точек от общей закономерности также являются случайными.

Следовательно, задача состоит в такой обработке экспериментальных данных, при которой по возможности точно была бы отражена тенденция зависимости y от x и возможно полнее исключено влияние случайных, незакономерных уклонений, связанных с погрешностями опыта. Такая задача является типичной для практики и называется задачей сглаживания экспериментальной зависимости.

Очень часто бывает так, что вид зависимости $y = f(x)$ до опыта известен из физических, биологических или лесоводственных соображений, связанных с существом решаемой задачи. Тогда на основании опытных данных требуется определить только некоторые параметры этой связи, которые входят в нее линейно, или по другой модели. Например, известно, что зависимость диаметра и высоты в молодняках до 10 – 15 лет можно выразить уравнением прямой.

Встречаются и более сложные случаи. Так, запас древесины (M) в лесном насаждении определяют через сумму площадей сечений стволов на высоте 1,3 м (G), среднюю высоту (H) и видовое число (F) по формуле $M=G \cdot H \cdot F$. Только G достаточно просто вычислить непосредственно, измерив

диаметры стволов. Среднюю высоту определяют после измерения 12 -15 деревьев (D и H у каждого) по связи H = f(D). Эта связь выражается полиномами разных степеней, логарифмической кривой и другими моделями, которые подбирают по виду их графиков. Видовое число вычисляют по моделям F=f(H), которые описывается простой или усложненной гиперболой. Подобных примеров в лесном хозяйстве множество.

Конечно, если вид связи хотя бы приблизительно известен, это упрощает и облегчает работу по проведению регрессионного анализа. Если же вид связи неизвестен, то его надо установить хотя бы ориентировочно, руководствуясь логической верификацией и профессиональными знаниями. В обычных задачах лесного хозяйства часто бывает достаточно построить график и сопоставить его с графиками известных функций. Последние в большом количестве представлены в специальных альбомах. Иногда (для недостаточно очевидных явлений) приходится перебирать ряд известных моделей или выводить новые. Последнее в лесном хозяйстве происходит редко и доступно лишь ученым с хорошей профессиональной лесоводственной и математической подготовкой. В числе таких наших ученых можно назвать А.З. Швиденко, О.А. Атрощенко, В.В. Севастьянова, В.В. Кузмичева, В.П. Машковского, В.Б. Гедых, В.А. Усольцева, из старшего поколения К.Е. Никитина, В.С. Чуенкова, А.Г. Мошкалева, В.В. Антанайтиса, Я.А. Юдицкого, Н.Т. Воинова и других.

Некоторые уравнения связи относительно просты, другие же отличаются повышенной сложностью с участием нескольких независимых переменных. Например, для нахождения текущего прироста древостоя в качестве независимых переменных для определенной древесной породы должны использоваться такие показатели как возраст, полнота, класс бонитета (высота) и иные аргументы.

При решении описанных задач, когда вид зависимости y = f(x) известен, применяют различные методы нахождения параметров таких уравнений, т.е. коэффициентов a, b, c, ... Наиболее общий подход разработан здесь русским математиком Пафнутием Львовичем Чебышевым (1821-1894), создателем петербургской научной школы математиков. Он создал теорию наилучшего приближения функций с помощью многочленов. Хотя метод Чебышева наиболее подходит для решения названных задач, но он сложен и используется редко, в основном профессиональными математиками. В Белорусском НИИ лесного хозяйства его использовал в 1971 г. Н.Т. Воинов (1934-1988), работавший впоследствии доцентом в ГГУ им. Ф.Скорины, для описания кривых, характеризующих образующую древесного ствола осины.

Более прост метод чисел Чебышева, но он требует равных интервалов между опытными данными, т.е. интервалы между

$$x_1, x_2, x_3, \ldots x_n = k$$

должны быть одинаковы (k), что можно обеспечить в опыте очень редко. Чаще всего, для нахождения коэффициентов регрессионных уравнений используют метод наименьших квадратов.

14.3 Метод наименьших квадратов

Метод наименьших квадратов применяется для решения различных задач, связанных с обработкой результатов опыта. Наиболее важным приложением этого метода является решение задачи сглаживания экспериментальной зависимости, т.е. изображения опытной функциональной зависимости аналитической формулой. При этом метод наименьших квадратов не решает вопроса о выборе общего вида аналитической функции, а дает возможность при заданном типе аналитической функции $y = f(x)$ подобрать наиболее вероятные значения для параметров этой функции.

Сущность метода наименьших квадратов при решении поставленной задачи заключается в следующем.

Пусть получено n экспериментальных точек с абсциссами

$$x_1,\ x_2,\ \ldots,\ x_n$$

и соответствующими им ординатами

$$y_1,\ y_2,\ \ldots,\ y_n$$

Зависимость y от x, изображаемая аналитической функцией,

$$y = f(x), \qquad (14.1)$$

которая обычно полностью не совпадает с экспериментальными значениями y_i во всех n точках. Это означает, что для всех или некоторых точек разность

$$\Delta_i = y_i - f(x_i) \qquad (14.2)$$

будет отлична от нуля.

Требуется подобрать параметры функции (14.1) таким образом, чтобы сумма квадратов разностей (14.2) была наименьшей, т.е. требуется обратить в минимум выражение

$$z = \sum_{i=1}^{n} \Delta_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \qquad (14.3)$$

Таким образом, при методе наименьших квадратов приближение аналитической функции $y = f(x)$ к экспериментальной зависимости счита-

ется наилучшим, если выполняется условие минимума суммы квадратов отклонений искомой аналитической функции от экспериментальной зависимости.

Следует заметить, что выражение (14.3) представляет собой полином второй степени относительно неизвестных параметров. Эти неизвестные параметры в зависимость $y = f(x)$ входят линейно, и выражение (14.3) не может принимать отрицательных значений. Поэтому существуют такие значения неизвестных параметров, при которых функция (14.3) достигает минимума, и этот минимум в зависимости от значений x_i и y_i будет положительным или равным нулю.

При решении многих практических задач функциональную зависимость y от x ищут в виде

$$y = \sum_{k=1}^{m} a_k f_k(x) \qquad (14.4)$$

где $f_1(x)$, $f_2(x)$, ..., $f_m(x)$ - известные функции,

$a_1, a_2, ..., a_m$ - неизвестные параметры.

Так, например, при исследовании колебательных процессов функциями $f_k(x)$ (k=1, 2, ..., m) являются тригонометрические функции

$$f_k(x) = \cos kx, \qquad f_k(x) = \sin kx.$$

При исследовании во многих областях техники, а также в лесном хозяйстве часто встречаются степенные функции

$$f_k(x) = x^{k-1} \quad (k = 1, 2, ..., m).$$

Есть и другие виды функций. Обычно гипотезу о виде требуемой функции принимают, анализируя по экспериментальным данным их графическое изображение и сравнивая его с графиками различных функций, которые приводятся в специальных альбомах.

Таким образом, $f_k(x)$ в равенстве (14.4) являются известными элементарными функциями аргумента x.

Исходя из принципа наименьших квадратов, мы должны подобрать такие значения неизвестных параметров $a_1, a_2, ..., a_m$, при которых обращается в минимум выражение

$$z = \sum_{i=1}^{n} \left[y_i - \sum_{k=1}^{m} a_k f_k(x_i) \right]^2 \qquad (14.5)$$

Выражение (14.5) является функцией неизвестных параметров a_k, поэтому для отыскания минимума этой функции нужно согласно правилам дифференциального исчисления найти частные производные функции z по всем параметрам a_k (k = 1, 2, ..., m) и приравнять их нулю:

$$\left.\begin{aligned}
&\frac{dz}{da_1} = 2\sum_{i=1}^{n}\left[y_i - \sum_{k=1}^{m} a_k f_k(x_i)\right]\cdot\left[-f_1(x_i)\right] = 0,\\
&\frac{dz}{da_2} = 2\sum_{i=1}^{n}\left[y_i - \sum_{k=1}^{m} a_k f_k(x_i)\right]\cdot\left[-f_2(x_2)\right] = 0,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\frac{dz}{da_m} = 2\sum_{i=1}^{n}\left[y_i - \sum_{k=1}^{m} a_k f_k(x_i)\right]\cdot\left[-f_m(x_i)\right] = 0.
\end{aligned}\right\} \qquad (14.6)$$

Подставляя в систему (14.6) опытные значения x_i и y_i, мы получим систему m линейных уравнений относительно неизвестных параметров a_k, решение которой может быть получено с помощью определителей или последовательным исключением неизвестных.

Рассмотрим применение метода наименьших квадратов, когда для изображения экспериментальной зависимости выбрана парабола второго порядка

$$y = ax^2 + bx + c.$$

Пусть в результате независимых опытов получено n значений величины y:

$$y_1, y_2, \ldots, y_n,$$

соответствующих значениям величины x:

$$x_1, x_2, \ldots, x_n.$$

Для определения неизвестных параметров a, b и c методом наименьших квадратов составляем сумму квадратов отклонений искомой аналитической функции от наблюдаемых значений в данных точках

$$z = \sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)^2 \qquad (14.7)$$

Дифференцируя функцию (14.7) по неизвестным параметрам a, b, и c и приравнивая производные к нулю, получим следующую систему уравнений:

$$\left.\begin{aligned} -\frac{1}{2}\frac{dz}{da} &= \sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)\cdot x_i^2 = 0,\\ -\frac{1}{2}\frac{dz}{db} &= \sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)\cdot x_i = 0,\\ -\frac{1}{2}\frac{dz}{dc} &= \sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right) = 0 \end{aligned}\right\} \qquad (14.8)$$

или несколько преобразовав уравнения (14.8), получим систему уравнений:

$$\left.\begin{aligned} a\cdot\sum_{i=1}^{n}x_i^4 + b\cdot\sum_{i=1}^{n}x_i^3 + c\cdot\sum_{i=1}^{n}x_i^2 &= \sum_{i=1}^{n}x_i^2 y_i,\\ a\cdot\sum_{i=1}^{n}x_i^3 + b\cdot\sum_{i=1}^{n}x_i^2 + c\cdot\sum_{i=1}^{n}x_i &= \sum_{i=1}^{n}x_i y_i,\\ a\cdot\sum_{i=1}^{n}x_i^2 + b\cdot\sum_{i=1}^{n}x_i + c\cdot n &= \sum_{i=1}^{n}y_i. \end{aligned}\right\} \qquad (14.9)$$

Система (14.9) представляет собой систему трех линейных уравнений относительно неизвестных параметров a, b и c. Решая систему (14.9) с помощью определителей третьего порядка или последовательным исключением неизвестных, мы и получим значение параметров a, b и c по методу наименьших квадратов.

Здесь приведен метод наименьших квадратов в строгой математической форме, что для студентов университета, изучающих курс высшей математики: матанализ, дифференциальное и интегральное исчисление, вполне понятно.

Но желательно показать и более наглядную форму вычисления коэффициентов уравнений с помощью наименьших квадратов. Для этого приведем следующий пример, где для простоты вычислим коэффициенты для уравнения прямой y = a+bx. Возьмем пример, приводимый Н. Н. Сваловым (1977), который представляет собой зависимость длины корней сеянца сосны от протяженности его стволика. Измеренные величины сведем в таблице. 14.2.

Таблица 14.2 – Зависимость длины корней от высоты ствола сеянцев сосны

Показатели	Величины, см									
Длина корней (y)	4	4	5	5	5	6	6	6	7	7
Длина стволиков (x)	3,0	3,1	3,5	3,5	4,1	3,5	4,0	5,0	5,0	5,3

Графически это изображено на рисунке 14.2

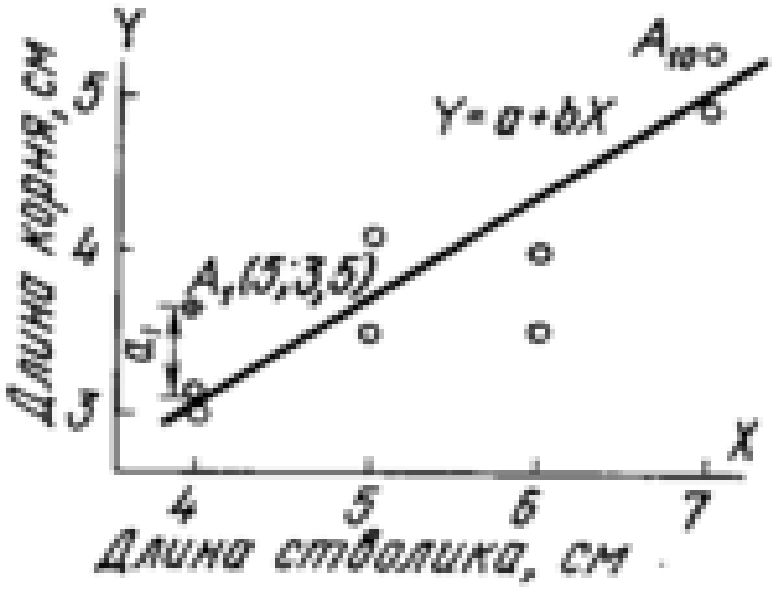


Рисунок. 14.2 Регрессия длины корней на длину стволиков всходов сосны

Здесь десять точек A_1, A_2, ..., A_{10}, изображенных на рисунке. 14.2, имеют соответственно абсциссы X_1, X_2, ..., X_{10}, и ординаты $У_1$, $У_2$, ..., $У_{10}$. Проведем визуально для целей рассмотрения метода искомую теоретическую прямую с уравнением $\hat{y} = a + bX$, где $\hat{y}$ - теоретические (выравненные) ординаты.

Интересующие нас разности между теоретическими и экспериментальными ординатами будут такими:

$$\left.\begin{aligned} d_1 &= \hat{Y}_1 - Y_1 = a + bX_1 - Y_1 \\ d_2 &= \hat{Y}_2 - Y_2 = a + bX_2 - Y_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ d_n &= \hat{Y}_n - Y_n = a + bX_n - Y_n \end{aligned}\right\} \qquad (14.10)$$

При этом d_i будут иметь и положительные и отрицательные значения. Вследствие этого, в сумме они могут компенсировать друг друга, так что $\sum_{i=1}^{n} d_i$ может оказаться весьма малой или даже равной нулю, хотя отдельные отклонения будут и большими.

Мы имеем дело с двумя признаками, для каждого из которых может быть найдено среднее квадратическое отклонение, обозначаемое соответственно символами σ_y и σ_x. В данном случае (в регрессионном

анализе) нас интересуют отклонения вариант не от средней ряда, а от выравненных значений $\hat{y}$, или от линии регрессии. Эти отклонения обозначают σ_{yx}. Подстрочные значки читают: “игрек по икс”. Они означают, что находят разности σ_{yx} величины y для соответствующих значений x. Как и при вычислении среднего квадратического отклонения σ, для исключения влияния знаков будем находить сначала средние квадраты σ_{yx}^2. Таким образом, решение поставленной задачи по нахождению теоретической регрессии (в нашем случае линейной) сводится к получению такой линии, для которой сумма квадратов отклонений всех экспериментальных значений y_i от вычисленных $\hat{y}_i$ является наименьшей, отсюда и название метода.

Для нахождения минимума $\sum\sigma_{yx}^2$ или в более подробной записи суммы

$$\sum (a+bX_i-Y_i)^2 \qquad (14.11)$$

по правилам дифференциального исчисления надо приравнять нулю частные производные от формулы (14.11) по a и по b.
Получим уравнения

$$2\sum(a+bX_i-У_i) = 0 \qquad (14.12)$$

и

$$2\sum(a+bX_i-У_i)\,X_i = 0 \qquad (14.13)$$

Сокращая обе части этих уравнений на 2, раскрывая скобки и замечая, что a+a+ ... +a = Na (N - число наблюдений или исходных уравнений), получим:

$Na + \sum bX_i - \sum У_i = 0$, $\sum aX_i + \sum bX_i^2 - \sum X_i У_i = 0$.

Вынося в суммах общие множители a и b за знак суммы и перенося последние члены в правую часть, получим уравнения:

$$\left.\begin{aligned} aN + b\sum X_i &= \sum Y_i \\ a\sum X_i + b\sum X_i^2 &= \sum X_i Y_i \end{aligned}\right\} \qquad (14.14)$$

Суммы распространены на все i от 1 до n. В полной записи следовало бы, например, вместо $\sum X_i$ написать $\sum_{i=1}^{n} X_i$. В дальнейшем в целях упрощения записи не указываются пределы для $\sum$, а также и подстрочный знак i при переменных x, y, означающий “любое” x, y. Это было сделано при написании формулы для $\bar{x} = (\sum_{i=1}^{n} x_i) / N$, которая сведена до выражения $\bar{x} = (\sum X) / N$. При таком условии уравнения (14.14), которые называют нормальными, будут:

$$\left.\begin{aligned} aN + b\sum x &= \sum y \\ a\sum x + b\sum x^2 &= \sum xy \end{aligned}\right\} \qquad (14.15)$$

Для нахождения коэффициентов a и b необходимо иметь конкретные значения N, $\sum x$, $\sum x^2$, $\sum y$, $\sum xy$.

При вычислении показателей корреляции между длиной стволиков x и корней y указанные суммы получим из таблицы 14.3, построенной на базе таблицы 14.2.

Таблица 14.3 – Исходные данные для вычисления показателей корреляции длины корней и высоты сеянцев сосны

№ п/п	1	2	3	4	5	6	7	8	9	10	∑
Длина корней (x)	4	4	5	5	5	6	6	6	7	7	55
Высота сеянцев (y)	3,0	3,1	3,5	3,5	4,1	3,5	4	5	5	5,3	40
x^2	16	16	25	25	25	36	36	36	49	49	313
y^2	9	9,61	12,25	12,25	16,81	12,25	16	25	25	28,09	227

Здесь имеем: N=10; $\sum x$=55; $\sum x^2$=313; $\sum y$=40; $\sum xy$=227.

Следовательно, нормальные уравнения в конкретном виде будут:

$$10a+55b=40 \qquad (14.16)$$

$$55a+313b=227 \qquad (14.17)$$

Поделив все члены на коэффициенты при b, равные 55 и 313, получим:

$$0{,}1757a+b = 0{,}7252 \qquad (14.18)$$

$$0{,}1818a+b = 0{,}7273 \qquad (14.19)$$

Вычитая уравнение (14.18) из (14.19), имеем 0,0061a = 0,0021, откуда a=0,344. Подставляя a в уравнение (3), получим b= 0,665.

Уравнение регрессии будет таким: $\hat{y} = 0{,}344+0{,}665x$.

Рассмотренный способ решения нормальных уравнений называют способом исключения. Преимущество этого способа состоит в его универсальности, т.е. применимости для регрессий любой формы и для любого числа коэффициентов.

Из (14.15) можно получить и другие уравнения и способ их решения. Выражение (14.15) преобразуется, если перенести начало отсчета x и y в точку O $(\bar{x}, \bar{y})$ (рисунок 14.3), которую называют центром распределения. В качестве значений исследуемых признаков при таком рассмотрении регрессии принимают не сами значения вариант x и y, а центральные отклонения их от своих средних $\bar{x}$ и $\bar{y}$, т.е. $\hat{x} = x_i - \bar{x}$ и $\hat{y} = y_i - \bar{y}$.

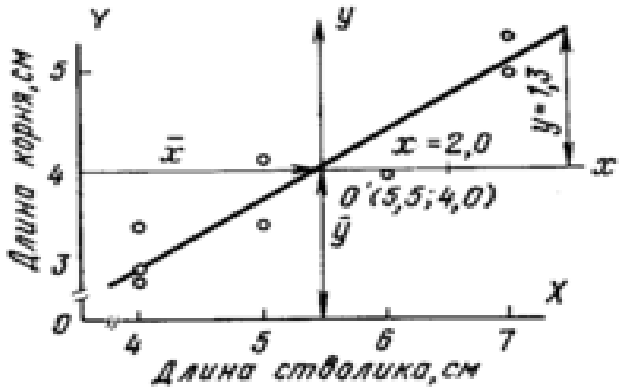


Рисунок 14.3 Регрессия длины корней на длину стволиков всходов сосны

Благодаря указанной замене, имеем:

$$\sum x = \sum(x+\bar{x}) = \sum x + \sum \bar{x} = \sum x + N\bar{x} \qquad (14.20)$$

Но так как из формулы средней величины $\bar{x} = \sum x_i / N$, $N\bar{x} = \sum x_i$, то очевидно, что в равенстве (14.20) $\sum x = 0$. Аналогично и $\sum y = 0$, если подобные действия провести для переменной y.

Из (14.20)

$$\sum x^2 = \sum(x+\bar{x})^2 = \sum(x^2 + 2x\bar{x} + \bar{x}^2) = \sum x^2 + 2\bar{x}\sum x + N\bar{x}^2 = \sum x^2 + N\bar{x}^2 \qquad (14.21)$$

$$\sum xy = \sum(x+\bar{x})(y+\bar{y}) = \sum(xy + \bar{x}y + \bar{y}x + \bar{x}\bar{y}) =$$
$$= \sum xy + \bar{x}\sum y + \bar{y}\sum x + N\bar{x}\bar{y}) = \sum xy + N\bar{x}\bar{y} \qquad (14.22)$$

Подставив выражения (14.20), (14.21) и (14.22) в уравнение (14.15) получим

$$\left.\begin{aligned} aN + b(\sum x + N\bar{x}) &= N\bar{y} \\ aN\bar{x} + b(\sum x^2 + N\bar{x}^2) &= (\sum xy + N\bar{x}\bar{y}) \end{aligned}\right\} \qquad (14.23)$$

Умножив (14.16) на $\bar{x}$ и вычитая из (14.17), получим одно нормальное уравнение

$$b\sum x^2 = \sum xy \qquad (14.24)$$

Из (14.24) следует $b = (\sum xy) / \sum x^2$ (14.25)

Величина b называется **коэффициентом регрессии**. Она показывает, на сколько единиц принятой меры изменяется y при изменении x на единицу ее меры.

Принимая во внимание общее уравнение линейной регрессии $y = a + bx$, имеем в частном случае при $x = \bar{x}$,

$$\bar{y} = a + b\bar{x} \qquad (14.26)$$

Из этого уравнения получим выражение для $a=\bar{y}-b\bar{x}$ (14.27)

Этот способ нахождения коэффициентов уравнения называется способом определителей. Алгоритм этого способа следующий.

Пусть имеем исходное уравнение

$$\hat{y}=\bar{y}+bx, \quad \text{где } x=X-\bar{x}. \qquad (14.28)$$

Нормальное уравнение здесь следующее: $b\sum x^2 = \sum xy$, $b = (\sum xy) / \sum x^2$. При уравнении $y=a+bx$, где x - варианты в первоначальных единицах измерения, нормальные уравнения будут

$$\left.\begin{aligned} aN + b\sum x &= \sum y \\ a\sum x + b\sum x^2 &= \sum xy \end{aligned}\right\} \quad \text{(см. 14.15)}$$

Определитель $D = N\sum x^2 - (\sum x)^2$ (14.29)

$$a = (\sum y \sum x^2 - \sum x \sum xy) / D, \qquad (14.30)$$

$$b = (N\sum xy - \sum x \sum y) / D, \qquad (14.31)$$

Применим метод определителей для нахождения коэффициентов a, b в регрессии длины стволиков сосны на длину корней. Исходные данные помещены в таблице 14.1. Они следующие:

$N=10$; $\bar{x}=5,5$; $\bar{y}=4,0$; $\sum x^2=10,50$; $\sum y^2=6,26$; $\sum xy=7,0$.

Исходное уравнение $\hat{y}=\bar{y}+bx$ (14.28), нормальное уравнение будет $b\sum x^2 = \sum xy$, откуда $b=(\sum xy) / (\sum x^2) = 7,0 / 10,50 = 0,667$.

$\hat{y}= 4,0+0,667\ x$. Заметим, что в качестве переменной здесь участвуют отклонения вариант x от средней $\bar{x}$. Если требуется найти выражение регрессии с вариантами x и y, следует поставить в исходное уравнение (14.28) вместо x его значение $(x-\bar{x})$.

Получим $\hat{y}=\bar{y}+b(x - \bar{x})$ (14.32)

Для нашего примера имеем

$\hat{y}=4,0+0,667(x-5,5)$

или

$\hat{y}=0,332+0,667x$.

Сравнивая результат с полученным ранее другим способом решения (способ исключения постоянных), видим их практически полное совпадение.

14.4 Вычисление значений зависимого признака на основе уравнений регрессий в лесном хозяйстве

Уравнение регрессии дает возможность найти значения $\hat{y}$, которые называют вычисленными или выравненными (иногда - наиболее вероятными значениями).

Для примера с сеянцами сосны, применяя уравнение $\hat{Y}=0{,}332+0{,}667\,x$, для x: 4, 5, 6, 7 см, получим $\hat{y}$, соответственно равные 3,0; 3,7; 4,3 и 5,0 см; или наиболее точно: 3,000; 3,667; 4,334; 5,001 см. Вычисленные значения $\hat{y}$ следует понимать как выравненные (усредненные) с помощью регрессии величины y, которые наиболее близки к истинным значениям этой величины при данных x, если бы истинные значения были найдены по большому числу наблюдений. На этом основании обоснованно считать $\hat{y}$ и наиболее вероятными значениями величины y.

Чаще всего, величины, которые вычислены по уравнениям регрессии, представляют собой общую закономерность для некоторой совокупности. Например, разрабатывая таблицы хода роста для основных лесообразующих пород Беларуси, В. Ф. Багинским были выведены различные уравнения, описывающие связи таксационных показателей: H=f(A); D=f(A); G=f(H); F=f(H) и т. д.

Так, для березовых древостоев зависимость величины суммы площадей сечений деревьев на высоте 1,3 м (Σq) от средней высоты (H) описана уравнением

$$\sum q = 5{,}151 + 1{,}3208 * H - 0{,}00842 * H^2 - 0.0000979 * H^3 ,$$

где H – высота древостоя в пределах от 5 до 32 м. Для вычисления средних видовых чисел древостоя (F) использовано уравнение $F = 0.398 + \frac{0.9845}{H}$

Решая это уравнение, т. е. подставляя последовательно значения H, равные 5, 6, 7, …,32 м, получили следующие величины $\sum q$. (таблица 14.2)

Таблица 14.2 – Величины $\sum q$ и F для березовых древостоев, вычисленные по уравнениям регрессии

H, м	ΣG, м^2	F
8	15,1	0,521
10	17,4	0,497
12	19,6	0,481
14	21,7	1,469
16	23,7	0,460
18	25,6	0,453
20	27,4	0,448
22	29,1	0,443
24	30,6	0,440
26	32,1	0,437
28	33,4	0,434
30	34,6	0,432

Приведенные величины являются средними для совокупности всех березовых древостоев Беларуси. Для отдельных насаждений отклонения могут достигать ±4 – 5%, но уже для совокупности в 5 – 10 выделов (участков однородного березового древостоя) отклонения не выходят за пределы 1 – 1,5%.

Поэтому при таксации лесного фонда лесничества и лесхоза, где однородных выделов обычно не меньше 50 – 100 для лесничества или 300 – 700 для лесхоза, общая ошибка определения $\sum q$ при отсутствии систематических отклонений составит незначительную величину 0,1 – 0,3%, а видового числа еще меньше.

Подобным образом в лесном хозяйстве используется большинство данных, полученных по уравнениям регрессии.

15. ОЦЕНКА РЕГРЕССИОННЫХ УРАВНЕНИЙ

15.1. Ошибки регрессионных уравнений
15.2. Оценки коэффициентов уравнений регрессии
15.3. Остаточная дисперсия и ее анализ
15.4. Взаимокоррелирующие аргументы. Выбор аргументов в уравнении регрессии при их взаимной корреляции в лесном хозяйстве

15.1. Ошибки регрессионных уравнений

Установить наличие связи между зависимой и независимой (независимыми) переменными, конечно, важно, но недостаточно. В наше время исследователь, вооруженный современным компьютером, имеющим мощное программное оснащение (например, систему «MathCAD») за несколько минут получает коэффициенты заданного уравнения. Сегодня для исследования на первый план выходит задача оценки и интерпретации полученных уравнений. Вызвано это тем, что все регрессионные уравнения дают некоторое приближение к тому закону или закономерности, которая существует в природе. Уравнение регрессии характеризует эту закономерность с некоторыми ошибками. Они вызваны в основном недостатками в экспериментальном материале, а иногда и невысоким качеством модели.

Может оказаться, что выборка не охватывает все особенности изучаемого явления, бывают ошибки измерений, на величину исходных данных влияют случайные причины и т.д. Но бывают ошибки и иного рода. Мы не всегда знаем причинно-следственную связь изучаемых явлений, иногда неправильно подбираем форму связи, выражаемую конкретным уравнением. Например, немецкий ученый проф. Продан еще в 60-е годы прошлого века отметил, что уравнение параболы второго порядка $y = a+bx+cx^2$, которое в то время часто использовали для описания хода роста древостоев по высоте, является слишком “жестким”: оно занижает значения высот в молодом возрасте и завышает их для спелых и перестойных древостоев. Поэтому очень важно установить, насколько вычисленная линия регрессии достоверна, т.е. найти основную ошибку регрессионного уравнения.

15.2 Оценки коэффициентов уравнений регрессии

Подобно коэффициенту корреляции и ряду других статистических показателей, коэффициент регрессии всегда определяется на основе выборочной совокупности. Значит, он является выборочным показателем, само же конкретное уравнение регрессии также может быть названо выборочным. Уравнение истинной линии регрессии будет выглядеть следующим образом: $y = \alpha + \beta x$

Параметры a и b выборочного уравнения регрессии служат для оценки истинных значений α и β, т.е. их значений в генеральной совокупности.

Нулевой гипотезой является отсутствие связи, т.е. признание того, что коэффициент регрессии (b) не отличаются от нуля. Для того, чтобы иметь право отбросить нулевую гипотезу, необходимо установить достаточную достоверность этого коэффициента σ_b по отношению σ_b к t критерию Стьюдента, что может быть сделано путем сопоставления b с его ошибкой

$$\sigma_{b/t} = b_p / \sigma_b). \qquad (15.1)$$

О достоверности b можно судить по величине ошибки σ_b, а также оценить и степень близости b к β.

Поскольку в определении линии регрессии участвуют два параметра a и b, следует отдельно рассмотреть, как могут они варьировать в выборочных совокупностях, взятых из одной и той же генеральной совокупности.

Теоретическая линия регрессии обычно расположена под большим или меньшим углом по отношению к оси абсцисс. Этот угол определяется величиной (коэффициентом) b. В геометрическом смысле b есть тангенс угла между линией регрессии и осью абсцисс (или ординат - если рассматривать вторую линию регрессии). При отсутствии регрессии b =0. Тогда линия регрессии y по x должна идти горизонтально по отношению к оси абсцисс, а линия регрессии x по y - вертикально. Место их пересечения соответствует средним значениям обоих признаков. Таким образом, каждая линия регрессии обязательно пройдет через точку K (рисунок 15.1), координаты которой $\bar{x}$, $\bar{y}$.

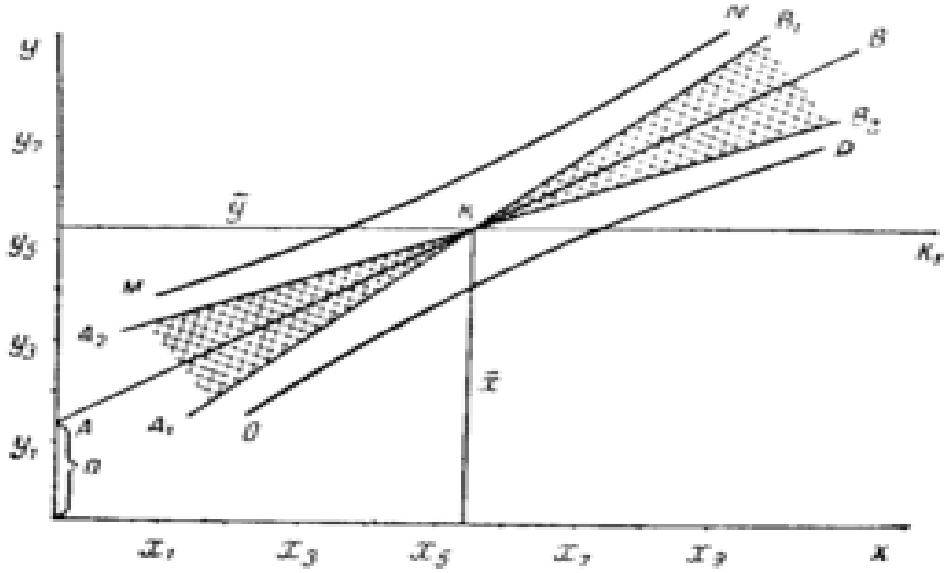


Рисунок 15.1 Доверительные границы для выборочной линии регрессии AB

Это важно для понимания возможной колеблемости линии регрессии. Так как b имеет ошибку σ_b, то, очевидно, значения выборочных b могут находиться в границах, определяемых этой ошибкой. Это значит, что угол наклона линии регрессии может быть или большим, или меньшим.

Как показано на рисунке 15.1, линия регрессии AB пересекает точку K и имеет угол наклона по отношению к горизонтали BKK_1. Но истинная линия регрессии заключена внутри пары углов, образованных пересечением линий A_1B_1 и A_2B_2. Если углы B_1K и B_2K построены по верхней и нижней границам b с учетом только одной ошибки, то вероятность нахождения истинной линии регрессии в этих границах равна 0,68.

Однако уравнение регрессии имеет еще свободный член a. Он определяет величину отрезка, отсекаемого на оси y линией регрессии AB. Величина a также имеет свои границы колеблемости, поэтому линия регрессии при том же значении b, т.е. при том же угле наклона ее к оси абсцисс, может проходить или несколько ниже линии AB, или несколько выше. Так как надо учитывать оба параметра уравнения регрессии, то установление доверительных границ для линии регрессии не так просто. В общем можно считать, что границы доверительного интервала представляют собой кривые линии типа гипербол (линии MN и OP на рисунке 15.1). Это значит, что по мере отдаления от средней точки $(\bar{x}, \bar{y})$ они расширяются. Крайние точки, по которым строится линия регрессии, обладают большей ошибкой.

Однако при проведении специальных опытов можно добиться достаточно больших n на всех частях интервала изменений x (или соответственно y) и принять, что дисперсия отдельных значений y (или x) будет примерно одинаковой на всех частях интервала. В таком случае можно применять сравнительно простые методы для оценки достоверности коэффициента и линии регрессии.

Основой для определения возможной вариации линии регрессии является сумма квадратов отклонений фактических значений y_i от вычисленных теоретически $\hat{y}_i$ по тем же значениям ряда x_i.

Формула для определения основной ошибки регрессии y по x, т.е. $y \cdot x$ (если регрессия x по y, то записывается $x \cdot y$) следующая

$$\sigma^2_{y \cdot x} = \frac{\Sigma(y_i - \hat{y}_i)^2}{N-2} \qquad (15.2)$$

$$\sigma_{y \cdot x} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{N-2}} \qquad (15.3)$$

Здесь N-2 - число степеней свободы. В данном случае число степеней свободы определяется как N -2, потому что при вычислении откло-

нений используются две величины y_i и $\hat{Y}_i$, а не одна, как это делалось при анализе вариационного ряда, когда число степеней свободы определялось как N -1.

Поясним изложенное примером вычисления основной ошибки регрессионного уравнения прямой, описывающей зависимость изменения высоты от диаметра в молодых сосновых древостоях. Исходные данные приведены в таблице 15.1.

Таблица 15.1 – Диаметры и высоты в молодых древостоях сосны

№ п/п	Д (x_i)	Н (y_i)	$\overline{Н}$ $\hat{y}_i$	Отклонения		Пределы прохождения линии регрессии при достоверности		
				$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	0,68 (1σ)	0,95 (2σ)	0,997(3σ)
1	3,9	4,8	5,1	-0,3	0,09	4,79-5,41	4,47-5,73	4,16-6,04
2	5,0	6,0	5,8	0,2	0,04	5,49-6,11	5,17-6,43	4,86-6,74
3	5,5	6,4	6,3	0,1	0,01	5,99-6,61	5,67-6,93	5,36-7,24
4	6,0	6,1	6,7	-0,6	0,36	6,39-7,01	6,07-7,33	5,76-7,64
5	6,5	7,1	7,1	0	0	6,79-6,41	6,47-7,73	6,16-8.04
6	6,8	7,7	7,3	0,4	0,16	6,99-7,61	6,67-7,93	6,36-8,24
7	7,4	7,6	7,9	-0,3	0,09	7,59-8,21	7,27-8,53	6,96-8,84
8	7,7	8,2	8,0	0,2	0,04	7,69-8,31	7,37-8,63	7,06-8,94
9	8,2	8,2	8,5	-0,3	0,09	8,19-8,81	7,87-9,13	7,56-9,44
10	8,9	9,1	9,0	0,1	0,01	8,69-9,31	8,37-9,63	8,06-9,94
11	9,7	9,7	9,7	0	0	9,39-10,11	9,07-10,33	8,76-10,64
∑	75,1	80,9	81,4	-0,5	0,89	-	-	-
Среднее	6,87	7,35	7,40	-	-	-	-	-

Уравнение прямой, соответствующее данным таблицы 15.1. равно

$$y = 1,9+0,8\ x \qquad (15.4)$$

Подставив значения из табл. 15.1 в (15.2) и (15.3), получим

$$\sigma^2_{y\cdot x} = \frac{0{,}89\text{м}}{9} = 0{,}099\text{м}$$

$$\sigma_{y\cdot x} = \sqrt{0{,}099}\text{м} = 0{,}314\text{м}$$

Здесь величина $\sigma_{y\cdot x}$ имеет такое же значение как и σ в вариационном ряду. В пределах одной $\sigma_{y\cdot x}$ отклонения распределяются вверх и вниз от линии регрессии в 68% случаев. В 95% они лежат в пределах $2\sigma_{y\cdot x}$, а в 99,7% случаев отклонения от теоретической линии регрессии составляют величину $3\sigma_{y\cdot x}$, т.е. в нашем примере это 0,942 м. Пределы изменения $\hat{y}_i$ при разной величине основного отклонения (1σ, 2σ, 3σ) показаны в графах 7-9 таблицы 15.1. На рисунке 15.2 наглядно видна математическая сущность основной ошибки регрессии - это та зона, ограниченная сверху и снизу значениями, соответствующими $\hat{y}\pm\sigma$, $\hat{y}\pm2\sigma$ и $\hat{y}\pm3\sigma$, внутри которой с заданной вероятностью могут располагаться искомые величины в генеральной совокупности.

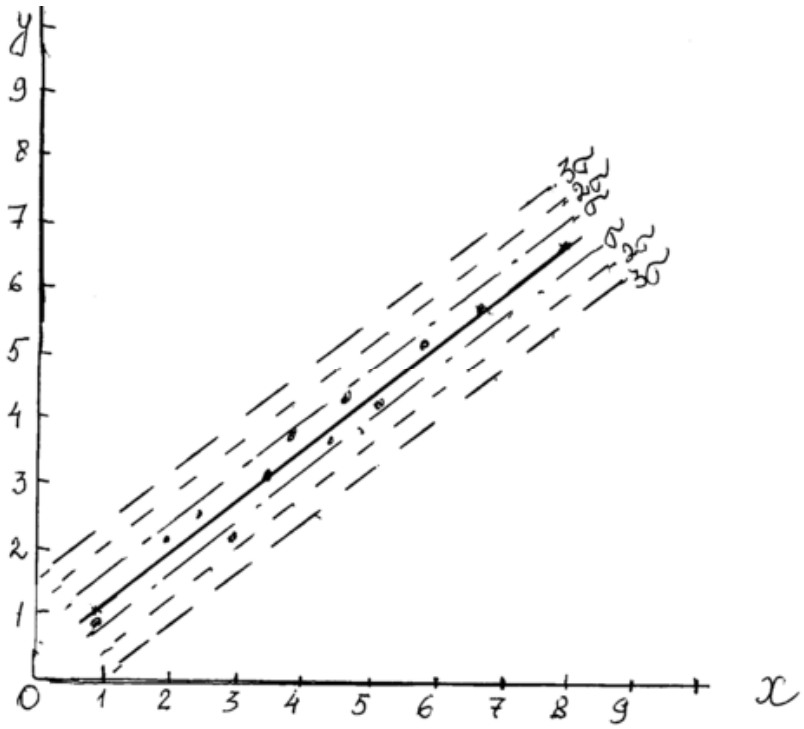


Рисунок 15.2 Математическая сущность основной ошибки регрессии

Мы помним из предыдущей главы, что коэффициент регрессии $R_p = r\frac{\sigma_y}{\sigma_x}$. По данным таблицы 15.1. вычисляем величины r, σ_y и σ_x (таблица 15.2).

Таблица 15.2 – Вычисление статистик распределения и связи по исходным данным таблицы 15.1.

№ п/п	x_i	y_i	$K_1=x_i-\bar{x}$	$K_2=y_i-\bar{y}$	K_1^2	K_2^2	K_1K_2
1	3,9	4,8	-2,97	-2,55	8,82	6,50	7,57
2	5,0	6,0	-1,87	-1,35	3.50	1,82	2,52
3	5,5	6,4	-1,37	-0,95	1,88	0,90	1,30
4	6.0	6,1	-0,87	-1,25	0,76	1,56	1,09
5	6,5	7,1	-0,37	0,25	0,14	0,06	0,09
6	6,8	7,7	-0,07	0,35	0,05	0,12	0,02
7	7,4	7,6	0,53	0,25	0,28	0,06	0,13
8	7,7	8,2	0,83	0,85	0,69	0,72	0,71
9	8,2	8,2	1,33	0,85	1,77	0,72	1,13
10	8,9	9,1	2,03	1,75	4,12	3,06	3,55
11	9,7	9,7	2,83	2,35	8,01	5,52	6,65
∑	75,1	80,9	0,03	0,05	30,02	21,04	24,76
Среднее	6,87	7,35	-	-	-	-	-

$$\sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = 1{,}65$$

$$\sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}} = 1{,}38$$

$$r = \frac{\sum K_1K_2}{\sqrt{K_1^2K_2^2}} = \frac{24{,}76}{\sqrt{30{,}02 \cdot 21{,}04}} = \frac{24{,}76}{25{,}13} = 0{,}985$$

Тогда $R_p = 0{,}985 \cdot \frac{1{,}38}{1{,}65} = 0{,}824$

Теперь найдем ошибку коэффициента регрессии

$$\sigma_R = \frac{\sigma_{y \cdot x}}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{0{,}099}{\sqrt{30{,}02}} = \frac{0{,}099}{5{,}48} = 0{,}02$$

Небольшая величина ошибки говорит о достоверности линии регрессии.

Это проверяется по формуле

$$t = \frac{R_p}{\sigma_R} = \frac{0{,}824}{0{,}02} = 41{,}2$$

По специальным таблицам t-критерия Стьюдента (приложение Е) находим, что критическая величина t при числе степеней свободы 9 равна 5,04 при p=0,001, т.е. вычисленное значение t превышает p при 0,001. Поэтому нулевая гипотеза о том, что коэффициент регрессии в генеральной совокупности (β) равен 0, должна быть отброшена.

Вычисленное уравнение достоверно.

15.3 Остаточная дисперсия и ее анализ

Для корректного решения вопроса об адекватности принятой модели, описывающей некоторую закономерность, недостаточно знать ее основную ошибку и определить значимость коэффициентов уравнения. Очень большое значение имеет анализ остаточной дисперсии. Правда, часто этот анализ биологи и лесоводы не делают, т.к. он труден. Просто предполагается, что остаточные величины, которые выходят за пределы уравнения регрессии, т.е. те значения x_i, что вызваны случайными причинами (их обычно называют просто остатками) распределены нормально и не влияют на результат. Но при строгом исследовании, которое и должны выполнять выпускники университета, делать анализ остатков необходимо. Опишем эту работу в интерпретации К. Е. Никитина и А. З. Швиденко.

Для названного анализа рассмотрим три величины, а именно $\sigma^2_y, \sigma^2_{yx}, \sigma^2_0$, где $\sigma^2_y = \sum (y_i - \bar{y})$, здесь $\bar{y}$ - общее среднее зависимой переменной y, т.е. σ^2_y - сумма квадратов отклонений y_i от средней.

$\sigma^2_{y \cdot x} = \sum (\hat{y} - \bar{y})^2$, здесь $\bar{y}_i$ - значения y_i, вычисленные по уравнению регрессии, т.е. $\sigma^2_{y \cdot x}$ - сумма квадратов отклонений, обусловленных регрессией y по x.

$\sigma_0^2 = \sum(\hat{y} - \bar{y})^2$, т.е. σ_0^2 - остаточная сумма квадратов отклонений, объясняемая иными причинами, чем зависимость y от x. Эти три суммы квадратов отклонений находятся в соотношении

$$\sigma_y^2 = \sigma_{y \cdot x}^2 + \sigma_0^2 \qquad (15.5)$$

Это выражение представляет разложение общей суммы квадратов отклонений на две составляющие: обусловленную регрессией $\sigma_{y \cdot x}^2$ и остаточную, объясняющуюся иными причинами, чем зависимость y от x. Очевидно, что чем наша модель лучше описывает изучаемую зависимость, тем меньшей должна быть остаточная сумма квадратов σ_0^2 и тем ближе к единице коэффициент детерминации $r^2 = \sigma_{y \cdot x}^2 / \sigma_y^2$. Поэтому в основу оценки уравнения регрессии может быть положен анализ дисперсий σ^2 с учетом того, что сумме σ_y^2 соответствует N -1 степень свободы, $\sigma_{y \cdot x}^2$ - имеет одну степень свободы, поскольку является единственной функцией от y_i, а σ_0^2 имеет N -2 степени свободы, так как в соответствии с (15.5) (N -1)=(N -2)+1. Сумма квадратов отклонений, деленная на число степеней свободы, дает соответствующую оценку дисперсии, еще именуемый средний квадрат отклонений.

Обозначим оценку дисперсии, обусловленной регрессией, через $\sigma_{y \cdot x}^2$. Если применяемая модель адекватна, то $\sigma_{y \cdot x}^2 = \sigma^2$, в противном случае $\sigma_{y \cdot x}^2 > \sigma^2$. Оценка остаточной дисперсии $\hat{\sigma}_0^2 = \sigma_0^2 / (N - 2)$. Если в генеральной совокупности связь между величинами x и y линейна, то σ_0^2 является несмещенной оценкой дисперсии σ^2 ошибок ε_i или, что то же самое, дисперсии величины $\tilde{y}_i$, обозначенной через $\sigma_{\tilde{y}_i}^2$. Если связь нелинейна, т.е. модель выбрана неверно, то последнее утверждение неверно. Исходная таблица для анализа линейного уравнения регрессии имеет вид (таблица 15.3).

Таблица 15.3 – Исходные данные для анализа дисперсий регрессионного уравнения

Источник	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Регрессии	$\sigma_{y \cdot x}^2$	1	$\hat{\sigma}_{y \cdot x}^2$
Остаточная	σ_0^2	N -2	$\hat{\sigma}_0^2 = \sigma_0^2 / (N - 2)$
Общая	σ_y^2	N -1	

При названных выше предпосылках регрессионного анализа и при условии, что истинный коэффициент регрессии $\beta_1=0$, отношение

$$F_{1-\alpha}(1, n-2) = \frac{\hat{\sigma}^2_{y \cdot x}}{\hat{\sigma}^2_{\bar{y}}} \qquad (15.6)$$

подчиняется F-распределению с 1 и N -2 степенями свободы. Если расчетное значение F превышает табличное $F_{1-\alpha}$ [1; (N -2)], то гипотезу $H_0 : \beta_1 =0$ отклоняют с уровнем значимости α. Табличную величину F берут из приложения Ж.

Предположив, что линейная модель адекватна, можно вывести формулы основных ошибок коэффициентов b_0 и b_2 и уравнения регрессии в целом, а также построить доверительные интервалы для истинных β_0 и β_1. Применив формулу основной ошибки функций нескольких случайных величин, можно получить

$$m_{b_1} = \left[\frac{\hat{\sigma}^2_0}{\sum(x_i - \bar{x})^2}\right]^{1/2}, \qquad (15.7)$$

где $\hat{\sigma}^2_0$- оценка дисперсии σ^2, т.е. величина m_{b1} зависит как от остаточной дисперсии, так и от суммы $\sum\left(x_1 - \bar{X}\right)^2$.

Отсюда следует, что для увеличения надежности выборочного коэффициента регрессии следует брать по возможности больший интервал $x_{max} - x_{min}$.

Доверительные интервалы для β_1 имеют вид

$$b_1 - t_{1-\alpha/2}\,(N-2)\,m_{b1} \le \beta_1 \le b_1 + t_{1-\alpha/2}\,(N-2)\,m_{b1}, \qquad (15.8)$$

где $t_{1-\alpha/2}$ (N -2) - табличное значение t для двустороннего уровня значимости α и при числе степеней свободы k= N -2.

Для проверки гипотезы $H_0 : \beta_1 = \beta_*$ против альтернативы $H_0 : \beta_1 \neq \beta_*$ достаточно убедиться, что β_* находится в доверительном интервале (15.8), если же это не так, то альтернативную гипотезу отвергают.

Основную ошибку b_0 находят из формулы

$$m_{b_0} = \left[\hat{\sigma}^2_0 * \frac{x^2}{n\sum(x_i - \bar{x})^2}\right]^{1/2}. \qquad (15.9)$$

доверительные интервалы

$$b_0 - t_{1-\alpha/2}\,(N-2)\,m_{b1} \le \beta_0 \le b_0 + t_{1-\alpha/2}\,(N-2), \qquad (15.10)$$

а проверку гипотезы $H_0 : \beta_0 = \beta^*$ против $H_\alpha : \beta_0 \neq \beta^*$ проводят аналогично коэффициенту регрессии b_0.

Основная ошибка уравнения линейной регрессии $y = \hat{Y} + b_1(x - \hat{x})$ есть ошибка каждого расчетного значения $\hat{Y}_k$, т.е. условного среднего $\hat{Y}_k$ для некоторого x_k. Ее определяют по формуле

$$m_y = \sigma_0 * \left[\frac{1}{n} * \frac{(x_k - \hat{x})^{1/2}}{\sum(x_i - \hat{x})^2}\right]^{1/2}$$

Если $x_k = \hat{x}$, то ошибка имеет наименьшее значение и возрастает при увеличении разности $x_k - \hat{x}$. Основная ошибка уравнения регрессии указывает ту зону, которая в 68 случаях из 100 накрывает истинное значение $\hat{y}_k$. Удвоенная ошибка соответствует приближенно доверительной вероятности 0,95 и т.д.

Для примера при проведении описанных расчетов возмем исходные данные, показанные в таблице 15.4. Для определения суммы квадратов воспользуемся следующими формулами

$$\sigma_y^2 = \sum y_i^2 - (\sum y_i)^2 / N$$
$$\sigma_{yx}^2 = b_1\left[\sum x_i y_i - (\sum x_i) \cdot (\sum y_i) / N\right]$$
$$\sigma_0^2 = \sigma_y^2 - \sigma_{yx}^2$$

Приведенные формулы получены непосредственными преобразованиями соответствующих сумм квадратов.

Таблица 15.4 – Исходные данные для регрессионного анализа

Источник	Сумма квадратов	Число степеней свободы ν	Оценки дисперсии (средний квадрат)
Регрессии	148,31	1	148,31
Остаточная	14,00	N-2=9	1,556
Общая	162,31	N-1=10	

Вычисленное значение F-критерия равно 148,31 / 1,556 = 95,3. Если α=0,05, $F_{0,05}$ при числе степеней свободы 1 и 9 (приложение Ж) равно 5,12, следовательно $F_{выч} > F_{табл}$, то гипотезу $H_0 : \beta_1 = 0$ отвергают с уровнем значимости α=0,05.

Вычислим основные ошибки (15.7) и (15.8) и доверительные интервалы для α=0,05 при 5%-ном уровне значимости. Для нахождения соответствующих t-коэффициентов получаем $m_{b1}=(1,556/236,97)^{1/2}$ =0,081; $0,791-0,081 \cdot 2,26 \le \beta_1 \le 0,791+0,081 \cdot 2,26$, или $0,608 \le \beta_1 \le 0,974$, т.е. с вероятностью 0,95 истинный коэффициент регрессии находится между 0,608 и 0,974. Если взять для проверки альтернативную гипотезу со значением β_* из этого интервала, то такая H_α не может быть отклонена.

Аналогичные рассуждения относительно b_0 предоставляем студентам для самостоятельного решения.

Ранее рассматривался коэффициент детерминации r^2, вычисляемый (если линейная модель верна) как отношение суммы квадратов, обусловленной регрессией, к общей сумме квадратов и измеряющий ту часть изменчивости зависимой переменной y, которая определяется зависимостью y от x. В нашем примере r^2 =148,31/162,31 = 0,913, т.е. на 91% изменчивость y зависит от изменения x.

Соответствие модели исходным данным можно оценить по графику, нанеся исходные точки и вычисленное уравнение регрессии. Для линейного уравнения с одной переменной такой путь вполне приемлем, но если модель сложная (много переменных, сложный вид зависимости), то для суждения об адекватности требуются объективные статистические критерии. Обычно для этой цели применяют F-критерий. Следует подчеркнуть, что во всех случаях, когда имеется возможность проверить адекватность, эта процедура необходима.

Если модель адекватна, то остаточная сумма квадратов $\sigma^2_{ост}$ объясняется только дисперсией ошибок ε_1 σ^2; в противном случае $\sigma^2_{ост}$ дополнительно включает сумму квадратов, порожденную неадекватностью, т.е. сумму квадратов расстояний между вычисленным и истинным уравнением регрессии. Для адекватной модели различие между оценкой дисперсии неадекватности $\sigma^2_{неад}$ и оценкой дисперсии ошибок σ^2 объясняется только случайными причинами, т.е. отношение

$$F = \frac{\sigma^2_{неад}}{\sigma^2} \qquad (15.14)$$

должно быть незначимым при выбранном уровне значимости. Для применения (15.14) необходима оценка σ^2, которая, как правило, неизвестна. Вычислить ее можно только в тех случаях, когда есть параллельные наблюдения при x_i; тогда разброс y_{ij} при одинаковых i дает оценку дисперсии σ^2, так как все остальные влияния, кроме случайной изменчивости ε_i, исключаются. Поэтому сумму квадратов, по которой оценивают σ^2, называют суммой квадратов, связанной с чистой ошибкой, мы ее обозначим через $\sigma^2_{\div.o}$.

Анализ остатков, полученных для различных моделей, позволяет выбрать наилучшую модель. Основные вопросы, на которые должен быть получен ответ следующие:

1) подтверждение нормальности распределения остатков;

2) постоянство дисперсии σ^2 и независимость ее от величины x_i;

3) адекватность модели на всех отрезках интервала изменения зависимой переменной или возможность ее улучшения добавлением нелинейных членов.

Остатки можно исследовать при помощи специальных критериев. Однако вполне достаточные результаты дает графический анализ, использующий общий график остатков (для суждения о нормальности ε_i), графики зависимости остатков от x_i, $\bar{y}_i$ и др. Если модель адекватна, то точки ε_i должны располагаться в полосе, параллельной оси абсцисс на графиках зависимости ε_i от x_i и y_i (или $\bar{y}_i$). Некоторая субъективность заключений при этом, понятно, остается. В настоящее время полные лицензионные программы матобеспечения для ПК включают анализ остатков. Поэтому эту процедуру надо обязательно применять при проведении регрессионного анализа. Про необходимость проведения анализа остатков в своих работах неоднократно упоминают проф. О.А. Атрощенко, А.З. Швиденко и др.

15.4. Взаимокоррелирующие аргументы. Выбор аргументов в уравнении регрессии при их взаимной корреляции в лесном хозяйстве

Если мы определяем функцию от нескольких аргументов, то надо знать как влияет каждый аргумент, его значимость, что мы разобрали выше. Но остается вопрос о необходимом количестве аргументов. На первый взгляд может показаться, что чем больше аргументов, тем точнее вычисление функции, так как учитывается большее количество влияющих факторов.

Например, если мы изучаем зависимость высоты дерева от качества условий местопроизрастания, то можем взять в качестве аргумента количество физической глины в горизонтах A_1, С. Если же эти аргументы дополним процентом гумуса в горизонтах A_1 и A_2, то прогноз функции будет точнее. Можем добавить увлажненность, скажем записать глубину уровень грунтовых вод (УГВ), что даст еще большее уточнение. Но бесконечно добавлять аргументы нельзя. Во-первых, это просто сложно реализуется, особенно на стадии эксперимента. Главное же - аргументы могут быть взаимно коррелированы. Так, если у нас УГВ расположен на глубине 1,0 м, то нет смысла брать процент влаги в горизонте С. Эти два аргумента взаимозависимы. Такое же явление мы наблюдаем в ранее приведенном примере по определению коэффициента формы q_2. Ранее использовали выражение $q_2 = f(Н,Д)$, но позже (после исследований Ф.П. Моисеенко) поняли, что достаточно будет принять $q_2 = f(Н)$.

Все дело здесь в том, что аргументы в уравнении регрессии не должны коррелировать друг с другом. Так УГВ и процент физической глины в горизонте А, если и коррелируют, то слабо. Поэтому применение их обоих оправдано. Но уже увлажненность горизонта С прямо зависит от УГВ. Взаимно коррелируют диаметр и высота дерева.

Применение множественного регрессионнного анализа предполагает наличие некоррелированных аргументов. В случае, если аргументы регрессии коррелированы, что часто бывает в лесоводственных исследова-

ниях, то коэффициенты, стоящие при них (a, b, ...) определяются в зависимости друг от друга и тоже коррелированы. Корреляция между коэффициентами регрессии искажает оценку индивидуального изменения каждого аргумента на величину функции. Это значит, что вычисленные коэффициенты a, b, ... не определены строго, а могут взаимно изменяться: увеличивается a, тогда уменьшается b и т.д.

Взаимосвязь между аргументами находится через внутреннюю меру определенности $d_{bH} = 1 - (1 - r_{ji})$, которая лежит в пределах от 0 до 1. Здесь r_{ji} - коэффициент корреляции между аргументами j и i.

При увеличении d_{bH} возрастает неопределенность относительно коэффициентов регрессии. Границу для d_{bH}, где интерпретация отдельных факторов еще возможна, определяют из условий задачи и сути изучаемого явления. При $d_{bH} \leq 0,5$ коэффициенты регрессии искажены не очень сильно и поддаются интерпретации.

Выбор конкретных аргументов в уравнении регрессии определяется из условий задания. В первую очередь делается логический анализ и выбираются те аргументы, которые лучше соответствуют биологическим и лесоводственным особенностям и законам. Среди выбранных аргументов, если они взаимокоррелированы, оставляют те, где большие частные коэффициенты корреляции $x - y_0$. Два аргумента, корреляция между которыми больше чем 0,5, не должны быть в одном уравнении. Без ущерба для результата вычислений один из них опускается.

16. ВЫБОР УРАВНЕНИЙ РЕГРЕССИИ, ИХ ВЕРИФИКАЦИЯ И ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ В ЛЕСНОМ ХОЗЯЙСТВЕ

16.1 Принципы и основные критерии выбора регрессионной модели

В предыдущих главах описано как вычислить коэффициенты регрессионного уравнения и как оценить это уравнение. Но для проведения анализа биологических и лесоводственных явлений этого недостаточно. Нам надо выбрать вид связи, определиться с конкретной ее моделью с тем, чтобы она отражала суть изучаемого явления, описывала закономерности его изменения.

Наилучшим образом модель будет выбрана тогда, когда она не просто сгладит экспериментальные данные, но и будет отражать суть изучаемого явления. В этом случае с определенными оговорками эту модель можно использовать для экстраполяции опытных данных. Последнее представляет собой наиболее сложную и неоднозначную часть по выбору и использованию моделей. Если модель выбрана не совсем верно и не отражает всей сути изучаемого явления, то, хотя на некотором отрезке кривой сглаживание может быть удовлетворительным, но для экстраполяции такая модель не годится.

Например, если мы описываем рост дерева или древостоя в высоту, то модель, исходя из основных законов роста древостоя, должна отвечать следующим условиям:

- кривая роста должна начинаться в начале координатных осей, т.к. рост начинается от семени, т.е. от нулевого значения высоты;
- в начальный период жизни (у разных древесных пород этот период неодинаков) рост идет медленно;
- постепенно происходит ускорение роста, достигая некоторого максимума прироста по высоте. Последний наблюдается в разном возрасте в зависимости от древесной породы и условий произрастания;
- постепенное замедление прироста по высоте продолжается до старшего возраста, который тоже бывает разным, аналогично предыдущему условию;
- в самом старшем возрасте рост в высоту практически прекращается (по диаметру дерево еще растет) и крона деревьев приобретает своеобразную уплотненную форму;

- в конце жизни дерева (древостоя) оно погибает и начинается новый цикл.

Исходя из описанных условий, кривая роста должна отвечать следующим условиям:

- начинаться в начале координат;
- иметь не менее 2 (иногда 3) точки перегиба: начало большого роста, переход к замедленным темпам прироста, прекращение роста. В практике обычно не используют последнее условие, т.к. при интенсивном хозяйстве до естественной спелости и распада древостои могут дожить только в зоне полной заповедности в заповедниках и национальных парках;
- абсолютное максимальное значение высоты, которое дает модель, не должно превышать максимально достижимую высоту, до которой вырастают деревья определенной породы в заданных условиях произрастания.

Поскольку для моделирования роста деревьев (древостоев) в высоту часто используют экспериментальный материал (пробные площади), где возраст древостоя колеблется от 20-30 до 60-80 лет, то этот отрезок может быть описан различными кривыми, даже параболой 2 порядка – $y=a_0+a_1x+a_2x^2$. Но эта кривая из-за ее «жесткости» совершенно непригодна для экстраполяции. Поэтому, применив кривую, не отвечающую сути изучаемого явления, мы не можем сделать суждения о действительной траектории изменения некоторого признака под воздействием разных факторов. Так, в приведенном примере нельзя сказать, какова будет высота за пределами изученных возрастов – в 10 или 100 лет. Кривая может здесь уйти далеко в сторону, даже вниз в 100 лет, просто исказив суть явления. Поэтому в лесной таксации для этих целей еще с XIX века применяют достаточно гибкие модели. Наиболее известны и хорошо «работают» следующие.

Функция Вебера: $H_a = H_{\max}(1-\frac{1}{1.0p^c})$; (16.1)

Уравнение гомельских ученых В.Н. Дракина, работавшего в Гомельском пединституте (ныне ГГУ им. Ф. Скорины) и Д.И. Вуевского научного сотрудника БелНИИЛХа: $H_a = H_{\max}(1-e^{-kt})^m$; (16.2)

Формула Я.А. Юдицкого (УСХА, г. Киев): $H_a = b_1\Phi[b_2(A-b_3)]+b_4$. (16.3)

В формулах (16.1-16.3) применены следующие обозначения:

H_a – высота в возрасте «*a*»;

H_{max} – максимальная высота данной породы в определенных условиях произрастания;

Φ – функция Маркова.

Другие обозначения характеризуют параметры уравнений.

Несколько иная тенденция наблюдается в модели роста деревьев по диаметру. Объясняется это тем, что в толщину деревья растут до момента их отмирания, хотя темп прироста в старости замедляется. Здесь зависимый признак – диаметр (независимый – возраст) – имеет положительное приращение практически в течение всей жизни, тогда как рост в высоту в старовозрастном лесу прекращается.

На основе знаний биологии можно предсказать, что кривые, отражающие рост в высоту в связи с возрастом и диаметром $H=f(A,D)$, будут отличаться друг от друга. Это обусловлено тем, что изменение возраста характеризуется равными приращениями в течение всей жизни. Приращение же диаметра неодинаково. В первые годы жизни оно небольшое, затем достигает максимума в молодые годы, после чего прогрессивно убывает до периода отмирания дерева. Здесь мы характеризуем среднее поведение роста, не обращая внимания на годовые или краткопериодические колебания прироста высоты или диаметра, которые возможны в связи с влиянием различных факторов среды, например, засуху, но не биологических законов.

Из этого далеко не полного перечня моделей видно, что они постепенно совершенствуются, сохраняя основные требования к моделям роста. Существуют определенные правила, приемы и критерии, которые в совокупности определяют выбор моделей. В большинстве случаев выбор уравнения регрессии производят на основе анализа, используя профессиональные знания, как это делали выше при рассмотрении роста деревьев по высоте и диаметру.

К настоящему времени накоплены знания о конкретных уравнениях регрессии для описания важнейших биологических процессов, подобных рассмотренным. Однако следует заметить, что при изучении многих явлений возникают большие затруднения в выборе подходящего уравнения регрессии. Даже установление общей ее формы (прямолинейна она или криволинейна) на основе профессиональных знаний часто не может быть сделано. Статистические методы в таких случаях дают основание для принятия решений о форме регрессии и выборе уравнения.

Для примера возьмем ранее рассмотренную зависимость длины корней сеянцев сосны и высоты ее всходов. Размещение точек на графике часто указывает на форму кривой. Если точки расположились, как на рисунке 16.1, есть основания принять связь за линейную. Теоретический анализ существа явления не доставляет аргументации, противоречащей этому выводу.

Действительно, для всходов сосны нет ограничивающих факторов к развитию обоих изучаемых признаков - длины стволиков и длины корней. В отношении регрессии длины стволиков и корней всходов или молодых растений можно сказать, что она линейна. Если точки на графике расположились так, что указывают на изгиб обобщающей их кривой, есть основания проверить гипотезу о линейности регрессии, т.е. рассчитать и оценить достоверность меры криволинейности.

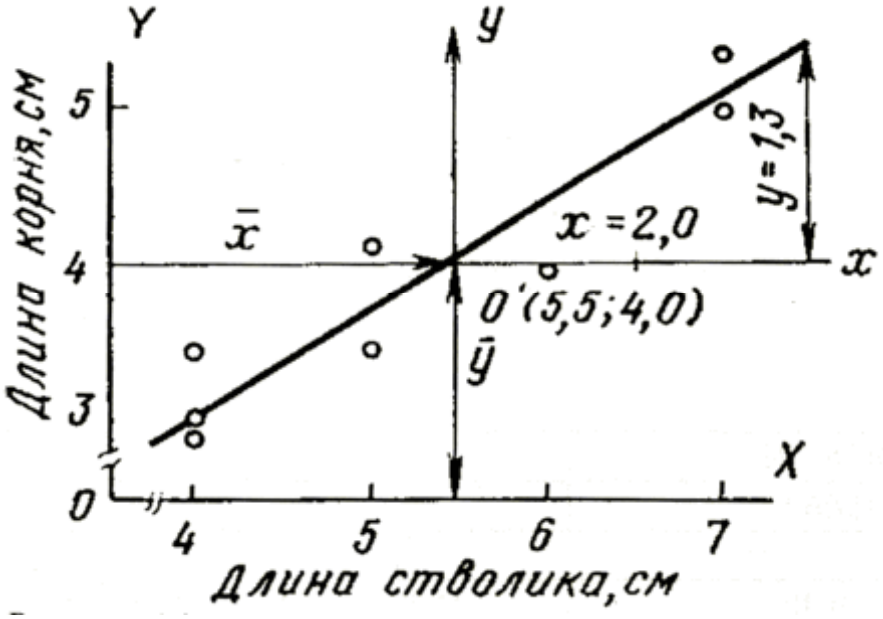


Рисунок 16.1 Регрессия длины корней на длину стволиков всходов сосны

Часто при исследованиях связи между признаками регрессионный анализ следует за корреляционным или осуществляется вместе с ним. Тогда определенные статистические заключения о линейности регрессии делают на основе t-критерия Стьюдента, но лучше на основе F-критерия Фишера.

Отметим, однако, что сами по себе критерии не дают исчерпывающего ответа о выборе уравнения, а лишь о форме регрессии: прямолинейна она или криволинейна. Если получены данные, свидетельствующие, что связь прямолинейна, этого достаточно, чтобы перейти к следующему шагу регрессионного анализа. Указание на криволинейный характер регрессии обязывает вести дальнейший поиск функции среди многих функций этого вида. В таком случае следует испытать наиболее простые и доступные исследователю функции. Останавливают выбор на функции, дающей лучшее приближение к опытным данным, или имеющую наименьшее среднее квадратическое отклонение вычисленных данных (σ_{yx}) или (σ'_{yx}). Здесь часто помогают уже упомянутые альбомы, где помещены графики различных функций.

В лесном хозяйстве во многих случаях удовлетворительную аппроксимацию опытных данных получают на основе парабол 2, 3-й и более высоких степеней, оценивая точность каждой из регрессий. При малом числе групп (классов) зависимой переменной y можно получить параболу с числом коэффициентов, равным числу групп, и проходящую через все точки, характеризующие групповые средние. Однако ценность регрессии в этом случае снижается. Кривая не выражает в таком случае закономерности связи, а отражает случайности выборочных наблюдений. Применять такую кривую можно только для аппроксимации и нахождения промежуточных значений в конкретном случае, не выходя за пределы изме-

рений. Например, если мы измерили на срубленном дереве диаметры в 6 точках, то для получения промежуточных диаметров, лежащих между измерениями, правомерно использовать уравнения вида целых полиномов 5-6 степени. Но они будут описывать образующую только этого дерева.

Криволинейный характер зависимости между переменными иногда удается заменить на прямолинейный путем преобразования x или y. Логарифмирование часто дает существенное уточнение выражения связи. Логарифмические параболы вследствие растянутости осей, вообще, более гибки.

16.2 Основные виды закономерностей в лесоводстве, лесной таксации и других лесных дисциплинах, выражаемые с помощью регрессионных моделей

В лесном хозяйстве, как видно из изложенного, применяются разные модели. Они выбираются, исходя из условий проведенного эксперимента и из сущности изучаемого явления. Если мы изучаем рост чего-либо (дерева, насаждения, ветви, животного и т.д.), то должны определиться каким условиям будет соответствовать наша модель. Для описания роста в высоту и по диаметру такие подходы изложены выше. Если же мы изучаем другие показатели роста дерева или древостоя, то должны учитывать уже их особенности. Например, изучая изменение запасов (M) древостоя, мы должны помнить следующие условия:

- кривая начинается в начале координат;
- она должна отражать медленный рост в самом молодом возрасте;
- необходимо, чтобы нашло отражение замедление увеличения запаса древостоя после определенного возраста (у сосны после 60-70 лет, у березы после 40-50 лет и т.д.), которое будет видно из наличия точки перегиба;
- в самом старшем возрасте идет процесс распада древостоя, появляется много сухостоя, и величина запаса уменьшается. При этом он не может быть отрицательным – предельный случай при полном распаде M=0.

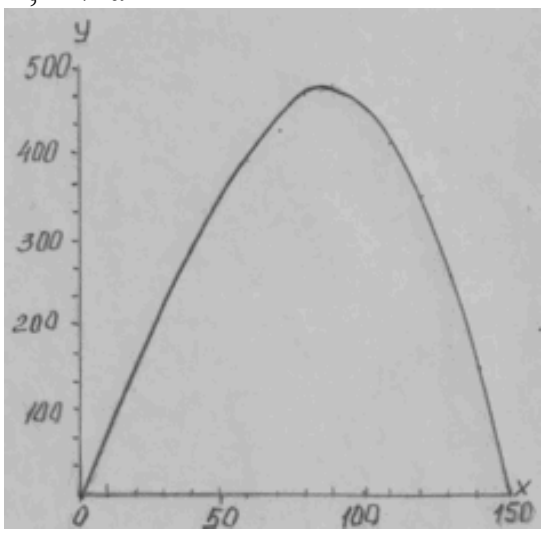
Таким образом, модель изменения запаса насаждения будет иметь примерно такой вид, как кривая на рисунке 16.2.

Если же мы будем исследовать прирост запаса древостоя ($Z_M^{Тек}$), то он должен соответствовать следующим условиям:

- начало кривой приходится на точку 0;
- медленный рост в начале жизненного цикла;
- резкое увеличение в период «большого роста» - сосна в 20-50 лет – и т.д.;
- достижение максимума прироста и постепенное его снижение;
- в перестойных древостоях, когда преобладают процессы распада, текущий прирост может быть отрицательным за счет отмирания деревьев;
- для отдельного дерева прирост может быть равен 0, но отрицательные величины исключаются.

Графически изменение прироста запаса древостоя (M) и наличного древостоя (Z) показано на рисунке 16.2.

M, м^3/га



A

Рисунок 16.2
Ход роста по наличному запасу древостоя осины I класса бонитета и его текущее изменение запаса с увеличением возраста

Z

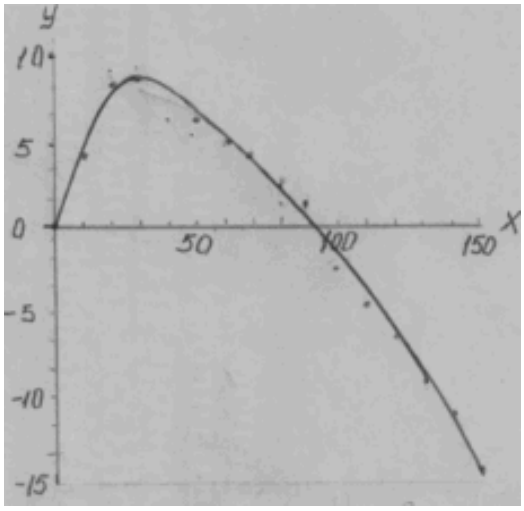


Рисунок 16.3
Изменение прироста наличного древостоя

Исследование других закономерностей в лесном хозяйстве приводит к использованию других кривых. Выбор формы зависимости во всех случаях идет от простого к сложному.

Модели, отражающие причинно-следственные взаимосвязи и взаимодействия в системах (или модели связи), - основной тип моделей, применяемых в практике и исследованиях по лесному хозяйству. Ниже мы рассмотрим модели связи двух классов: эмпирические и структурные или функциональные.

В качестве математической формы эмпирических моделей связи в основном используют регрессионные уравнения и реже - интерполяционные многочлены. В первом случае для нахождения коэффициентов уравнений применяют различные модификации метода наименьших квадратов, позволяющие просто и достаточно надежно оценить статистическим путем разрабатываемую модель. Методом регрессионного анализа получены, пожалуй, практически все наиболее содержательные биометрические закономерности в лесном хозяйстве.

В то же время метод наименьших квадратов имеет существенные недостатки чисто познавательного плана: во-первых, он не учитывает природной сущности изучаемого явления и допускает известный произвол в выборе конкретных типов уравнений. Во-вторых, этот метод предполагает детерминированный характер изучаемого процесса. Поэтому в последнее время все больше внимание привлекают вероятностные модели (особен-

но для отражения процессов, протекающих во времени), использующие методы теории случайных функций.

Основное средство объективной разработки моделей эмпирических связей – использование методов биометрии, основанных на законах математической статистики. Это позволяет принять обоснованные решения, оценить структуру, рабочий диапазон и надежность моделей. Однако в полной мере методы статистики можно применять только к материалу, полученному с соблюдением статистических предпосылок. Во всех остальных случаях применение относительно трудоемких методов статистического анализа взаимосвязей неоправданно. То же относится и к материалу, из которого произвольно исключена значительная часть только на том основании, что она “не укладывается” в некоторые представления экспериментатора, ибо простые графические методы позволяют в таких случаях получить те же результаты.

Если для разработки модели связи информация еще не собрана, то планирование эксперимента позволяет значительно улучшить результаты, так как лучше потратить часть времени и средств для предварительной оценки ситуации, выбора независимых переменных и их анализа. Главное здесь, как и во многих других случаях применения математических методов, - это точная формулировка задачи и преследуемых целей, исходя из сути изучаемого явления.

В данном разделе будем применять следующие термины: адекватность модели – соответствие исходным данным, подтвержденное статистическими критериями; корректность – ее приемлемость (с точки зрения пользователя), соответствие моделируемому процессу или системе. Эти термины являются обычными в математике.

Статистические методы могут подтвердить высокую вероятность адекватности модели, но особенности информации могут привести к результатам, неприемлемым с точки зрения существа явления; иначе говоря, корректная модель есть в известном смысле и лучшая. Особенно возрастает опасность ошибок при малых выборках.

Модели, связывающие более двух переменных, называют многомерными. К ним относятся множественные регрессионные уравнения, многомерные случайные процессы. По форме модели связи могут быть в табличном, графическом или аналитическом виде, т.е. в виде уравнений.

Регрессионные уравнения бывают линейные и нелинейные, причем этот термин может относиться как к коэффициентам уравнения, так и к независимой переменной. Например, уравнение $y=a+bx+cx^2+dx^3$ является линейным по коэффициентам a, b, c, d и нелинейным по x, а множественное уравнение $y = ax_1^b + cx_2^{dx_3}$ нелинейно как по коэффициентам, так и по переменным x_1, x_2, x_3. Мы рассмотрим линейные уравнения относительно коэффициентов, т.к. модели такого рода вполне достаточны для моделирования связей в лесных исследованиях, а теория нелинейного (по коэффициентам) оценивания очень сложна. Многие нелинейные моде-

ли можно привести к линейному виду, например, логарифмированием. Такие модели называют внутренне линейными.

Модели могут быть представлены не только уравнениями, но и в графической и табличной форме. Табличное и графическое представления закономерностей связи традиционно свойственны лесному хозяйству и широко распространено, особенно для данных, полученных без надлежащего статистического обоснования. Так, лесотаксационные таблицы (объемные, сортиментные, хода роста и др.), разработанные до 60-х годов XX века, представлены преимущественно в виде числовых массивов, полученных графическим выравниванием опытных материалов. Недостатки такого рода моделей известны: субъективизм конечных результатов, невозможность статистической оценки соответствия модели изучаемому явлению, неудовлетворительность формы для целей механизации обработки информации.

В то же время модели, выраженные в табличном виде, где величины получены путем графического выравнивания, разработанные весьма квалифицированными и ответственными специалистами на основе большого (даже огромного) экспериментального материала, часто оказывались адекватными и корректными и находили (находят и сегодня) широкое применение в практике. К таким моделям можно отнести таблицы объемов стволов А. Крюденера и объемные таблицы Союзлеспрома (М.М. Орлов (1867-1932), Д.И. Товстолес, В.К. Захаров, Б.А. Шустов, А.В. Тюрин), сортиментные таблицы Ф.П. Моисеенко, таблицы хода роста А.В. Тюрина и другие. Эти табличные данные обычно точнее результатов, полученных в 60-80 годы прошлого века путем использования математических моделей, имеющих недостаточное лесоводственное обоснование.

В принципе любой численный или графический материал можно выразить в виде формул. В 60-70 годы прошлого века машинный счет больших массивов информации из-за несовершенства ЭВМ того времени был затруднен. Поэтому таблицы выражали в виде математических моделей. Для этого был выполнен ряд работ по моделированию существующих таблиц. Однако опыт моделирования численных массивов, выровненных графически, свидетельствует, что достаточно точное их отражение требует использования громоздких и разнородных математических выражений, что также создает определенные неудобства в их использовании в системах автоматизированной обработки данных. В настоящее время возможности компьютеров позволяют пользоваться огромными массивами информации, выраженной в табличной форме, не прибегая к ее упрощению в виде уравнений.

При обработке лесоводственной информации нередки случаи, когда графические методы выравнивания наиболее целесообразны. Это бывает тогда, когда методы сбора исходных данных неизвестны либо статистически несостоятельны, а сбор информации (метод сбора, объем) не согласован с конечной целью работы. В силу этого применение статистиче-

ских методов может дать здесь результаты, противоречащие сути изучаемого явления. Проиллюстрируем второе положение примером.

Пусть у нас есть некоторый массив данных, который характеризует изменение запаса древостоя с возрастом. Известно, что запас древостоя зависит от породы, условий произрастания, возраста и полноты. При сборе экспериментального материала наиболее сложно строго выдержать единообразие по полноте. Если мы хотим определить ход роста по запасу нормальных древостоев (с полнотой 1,0), но в опытном материале есть пробные площади с полнотой от 0,6 до 1,0, то математическое выравнивание с использованием всего имеющегося материала даст искаженную картину, что видно на рисунке 16.4.

Из рисунка 16.4 видно, что использование всего массива пробных площадей, которые представляют разнородный материал (по полноте) ведет к ошибкам. В этом случае правильно предварительно построить график, определить главную закономерность (ход роста при полноте 1,0), даже сделать графическое выравнивание, а потом разрабатывать математическую (регрессионную) модель.

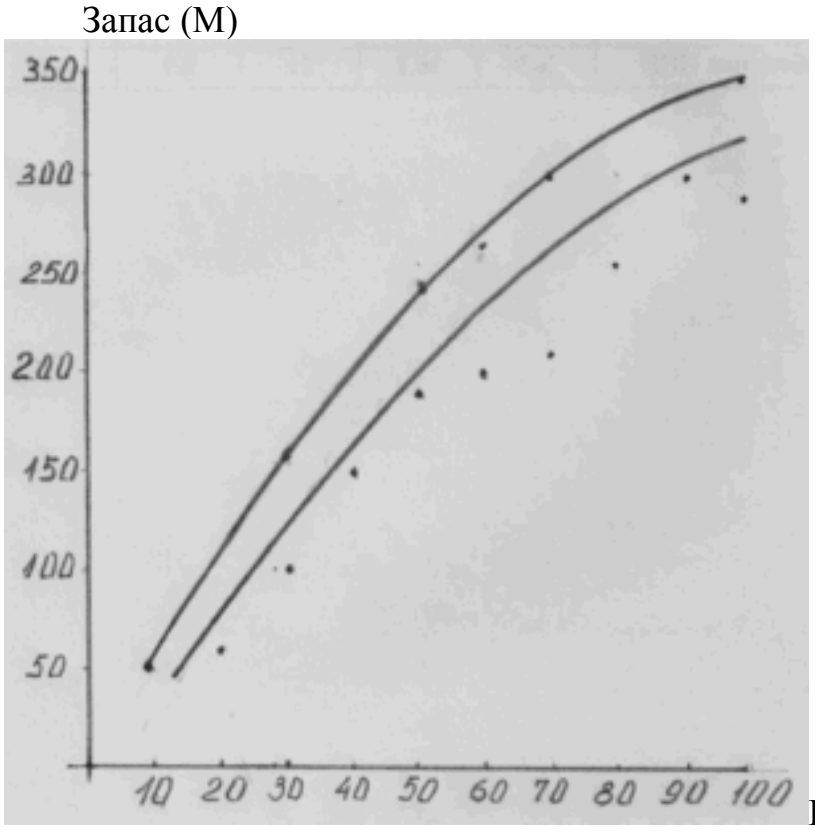


Рисунок 16.4 Модель изменения запаса (М) древостоя от возраста (А) при неверно организованных исходных данных для березовых древостоев II класса бонитета

1 – линия, построенная по данным из разнополнотных древостоев;

2 – графическое выравнивание по данным древостоев с полнотой 1,0

Подобный пример приводят К.Е. Никитин и А.З. Швиденко (рисунок 16.5). На нем показано 15 случайным образом замеренных высот деревьев в разновозрастном насаждении бука. Допустим, что требуется установить зависимость высоты от возраста. Применение стандартных статистических методов дает зависимость, показанную штриховой линией: начиная с некоторого возраста, высота деревьев постепенно уменьшается, т.е. модель адекватна, но не корректна.

Такой результат получен, конечно, из-за недостаточности исходных данных. В разновозрастном древостое деревья, имеющие одинаковый возраст, но отличающиеся расположением в пологе, имеют разные высоты. Если нет возможности собрать дополнительный материал, то графическое выравнивание - единственная возможность решения задачи. Надежность такого выравнивания нельзя оценить, поскольку в основе его лежит единственная предпосылка: высота деревьев с возрастом должна, как правило, увеличиваться, но не может уменьшаться.

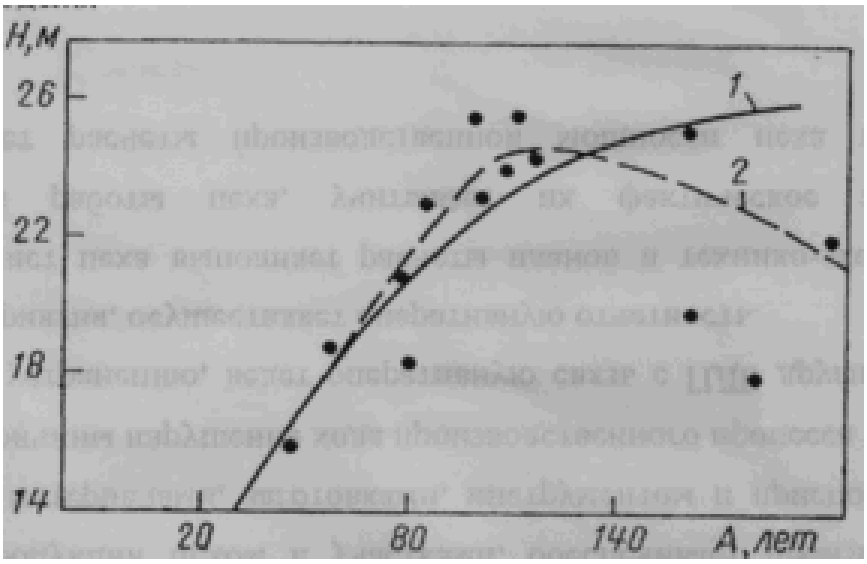


Рисунок 16.5 Выравнивание плохо организованных данных:
1 – графическое; 2 – аналитическое

Иногда возникает необходимость графические модели представить в аналитическом виде. Для этого применяют два практических приема: снимают с графика значения зависимой переменной и применяют один из статистических методов вычисления уравнения регрессии или, если конкретный вид уравнения известен, то коэффициенты уравнений целесообразно вычислять решением систем уравнений. Статистическая оценка моделей в обоих случаях может дать только соответствие графика полученному уравнению.

В качестве примера возьмем данные из рисунка 16.4. Здесь необходимо подобрать коэффициенты уравнения, отражающего зависимость высоты от возраста. Исходя из графика, форма связи здесь (сплошная линия) может быть в виде параболы второй степени

$$y=b_0+b_1x+b_2x^2, \qquad (16.4)$$

где x – возраст; y – высота.

Для нахождения коэффициентов b, b_1, b_2 возьмем три точки, например, со значениями x, равными 20, 80 и 140 годам. По графику находим y_1=12,3; y_2=21,1; y_3=25,2. Поскольку уравнение (16.4) справедливо для любой точки, лежащей на кривой, то подставив значения x и y, в уравнение (16.4) и решаем систему из трех уравнений для нахождения коэффициентов. Переменную x для упрощения расчетов можно маштабировать в 20 раз и записать:

$x'_1 = x_1/20 = 1$, $x'_2 = x_2/20 = 4$, $x'_3 = x_3/20 = 7$, т.е.

$$\left.\begin{aligned} 12{,}3 &= b_0 + b_1 + b_2; \\ 21{,}1 &= b_0 + 4b_1 + 16b_2; \\ 25{,}2 &= b_0 + 7b_1 + 49b_2. \end{aligned}\right\} \qquad (16.5)$$

Решая систему (16.5), находим b_0=8,32, b_1=4,239, b_2=-0,2611; вернувшись к исходной переменной, окончательно получим

$y = 8{,}32 + 4{,}239x/20 - 0{,}2611(x/20)^2 = 8{,}32 + 0{,}212x - 0{,}000652x^2$.

Легко убедиться, что сплошная линия на рисунке 16.4 есть графическое изображение полученного уравнения. Для этого вычислим для некоторого значения x, например 80 лет, , получим

$y=8{,}32+0{,}212*80+0{,}000652*80^2=8{,}32+16{,}96-4{,}17=21{,}11$, т.е. искомый результат (21,1 м) в 80 лет. Аналогично рассчитываем значения для любого возраста на имеющемся интервале наблюдений.

Приведенный пример есть частный случай так называемых интерполяционных многочленов. При многочисленных приближениях интерполяционный многочлен степени n однозначно определяется n+1 его значением. В нашем примере для нахождения коэффициентов многочлена второй степени мы взяли 3 точки. Применение интерполяционных многочленов такого типа удобно, если используемая модель должна предсказывать поведение функции только в точках (x_i, y_i) - узлах интерполяции, поскольку оценка поведения многочлена между узлами затруднена.

Если статистический анализ (или другие соображения) позволяет ограничиться линейным (относительно независимой переменной) уравнением, то выбор модели связи однозначен. Но если связь нелинейна, то процедуры однозначного выбора конкретного аналитического выражения в качестве уравнения регрессии не существует: одна и та же зависимость y от x может быть отражена многими формулами. Здесь требуется, с одной стороны, знание общих свойств моделируемой зависимости, с другой – математических свойств используемых выражений. В качестве примера рассмотрим материал, приводимый К.Е. Никитиным и А.З. Швиденко. Пусть нам требуется найти зависимость коэффициента вариации (υ_M) запаса древостоя, определенного по материалам измерений на круговых пробных площадях, от величины этих круговых проб (b). Результаты показаны на рисунке 16.6.

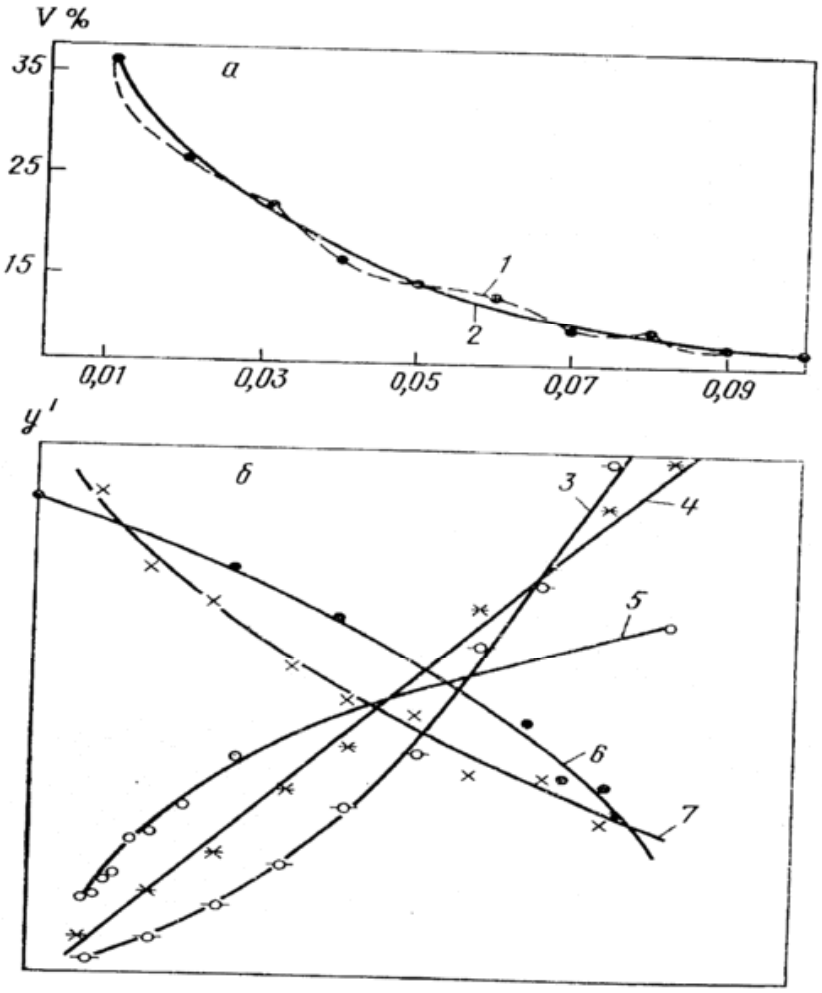


Рисунок 16.6

а) Модель $\upsilon_M = f(l)$: 1 – исходные данные; 2 – выравнивающая кривая гиперболического вида: $y = x(a + \frac{1}{bx})$ (16.8);

б) Зависимость коэффициента изменчивости запаса от величины круговых проб и преобразование ее к линейному виду. Здесь кривые 3-7 – показывают преобразование данных соответственно по формулам (16.7), (16.8), (16.9), (16.10), (16.11).

Можно подобрать такое уравнение регрессии, которое пройдет через все точки (так, интерполяционный многочлен n-1 степени пройдет точно через n точек), но бессмысленность подобного подхода очевидна. Например, нам требуется получить закономерность изменения показателей формы ствола (q_2) в зависимости от высоты дерева в среднем, т.е. учесть изменение условных средних (а не каждого конкретного значения), которые являются случайными величинами, подвержены изменчивости и содержат ошибки измерений. Анализ ситуации позволяет предположить, что уменьшение q_2 с увеличением высоты должно быть представлено монотонно убывающей кривой. Следовательно, выбираемая модель не должна содержать внутри своего “рабочего” интервала (диапазона

значений независимой переменной) особых точек минимума, максимума, точек перегиба.

Важное значение имеет простота модели и количество коэффициентов, подлежащих определению. Поэтому круг выражений, из которых выбирают уравнение регрессии, должен быть ограничен по возможности простыми функциями. Из них особое место занимают те, которые путем алгебраических преобразований могут быть приведены к линейному виду. Это позволяет применить относительно простые, но глубоко разработанные методы статистической оценки. К ним в первую очередь принадлежат показательные, степенные, логарифмические и гиперболические функции. В таблице 16.1 и на рисунках 16.7 и 16.8 приведены типы уравнений, которые, по совершенно справедливому мнению К.Е. Никитина и А.З.Швиденко наиболее часто используют в лесных исследованиях. Эти модели с одной независимой переменной и с двумя коэффициентами, приводимые к линейному виду. В таблице показан вид преобразования и критерии, позволяющие оценить приемлемость данного выражения. Их графическое изображение приведено на рисунке 16.6.

Таблица 16.1 – Некоторые элементарные функции, наиболее часто используемые в лесном деле, одной переменной и преобразование их к линейному виду

Уравнение	Тип преобразования	Уравнение прямой	Уравнение применяется, если постоянной величине равно соотношение	
$y=ax^b$	Логарифмирование	$lgy=lga+blgx$	$\Delta(lgx)/\Delta(lgy)$	16.6
$y=ae^{bx}$	Логарифмирование	$lgy=lga+(blge)x$	$\Delta x/\Delta(lgy)$	16.7
$y=a+bx^n$ (n-известное)	Замена $x'=x^n$, $y'=y$	$y'=a+bx'$	$\Delta(x^n)/\Delta y$	16.8
$y=a+b/x$	Замена $x'=1/x$, $y'=y$	$y'=a+bx'$	$\Delta(1/x)/\Delta lg$	16.9
$1/y=a+bx$	Замена $x'=x$, $y'=1/y$	$y'=a+bx'$	$\Delta x/\Delta(1/y)$	16.10
$x/y=a+bx$ $y=\frac{x}{a+bx}$	Замена $x'=x$, $y'=x/y$	$y'=a+bx'$	$\Delta x/\Delta(x/y)$	16.11

Задачу также можно решить путем предварительного преобразования исходных данных и нанесения их на график в новой системе координат (точки должны располагаться вдоль прямой линии). Первый путь

предпочтительнее, поскольку легко реализуется в программах автоматического выбора модели на компьютере, второй – более нагляден.

В лесном хозяйстве помимо уравнений, приведенных в таблице 16.1 очень широко используются уравнения вида целых полиномов, в основном 2-4 степени: $y=a_0+a_1x+a_2x^2+a_3x^3+\ldots$ К линейному виду они могут быть приведены путем логарифмирования. Их употребление для описания динамики древостоев рекомендуют К.Е. Никитин, В.Ф. Багинский и другие. Этими уравнениями хорошо выражаются зависимости *H=f(D)*. При этом здесь нежелательно использование параболы 2 порядка, которая занижает начальные и завышает конечные данные. А.Г. Мошкалев (1926-1992) использовал для связи *H=f(D)* полиномы 3 степени. В.Ф. Багинский иногда употреблял параболу 4 порядка. Для вычисления коэффициентов в уравнениях вида целых полиномов для ПК имеется хорошее матобеспечение, где дается и оценка уравнения.

Из уравнений, рекомендованных К.Е. Никитиным и А.З. Швиденко, при проведении исследований в лесном хозяйстве наиболее употребительными являются (16.6), (16.9), (16.10) и (16.11). Формы кривых, часто применяемых для моделирования зависимостей в исследованиях по лесному хозяйству, показаны на рисунках 16.7-16.8.

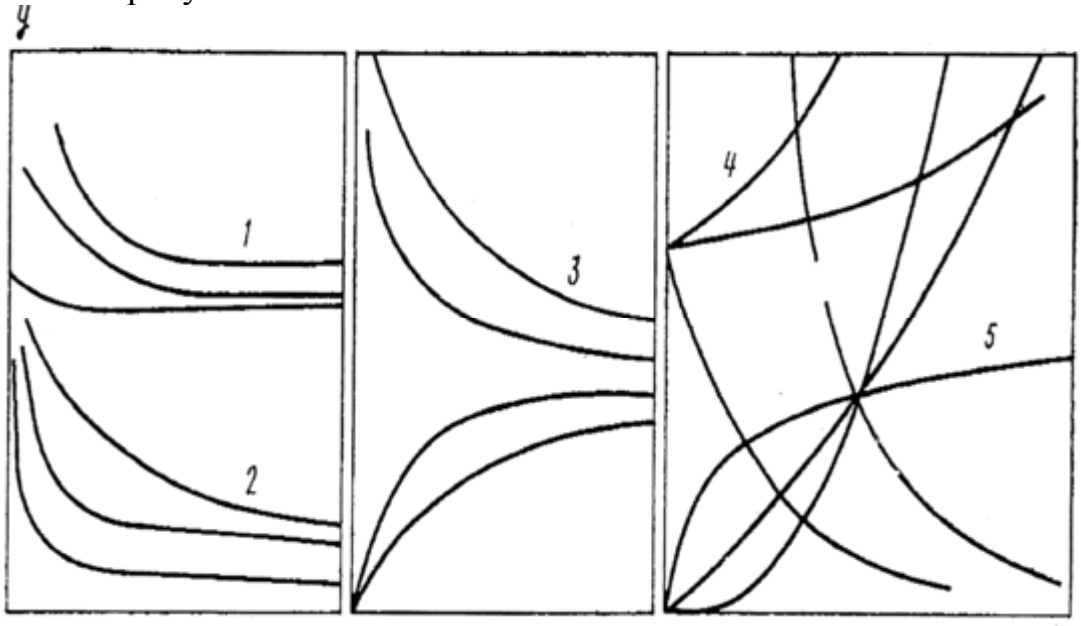


Рисунок 16.7 Графики кривых с двумя параметрами:
1 – $1/y=a+bx$; 2 – $y=a+b/x$; 3 – $y/x=a+b/x$; 4 – $y=ae^{bx}$; 5 – $y=ax^b$

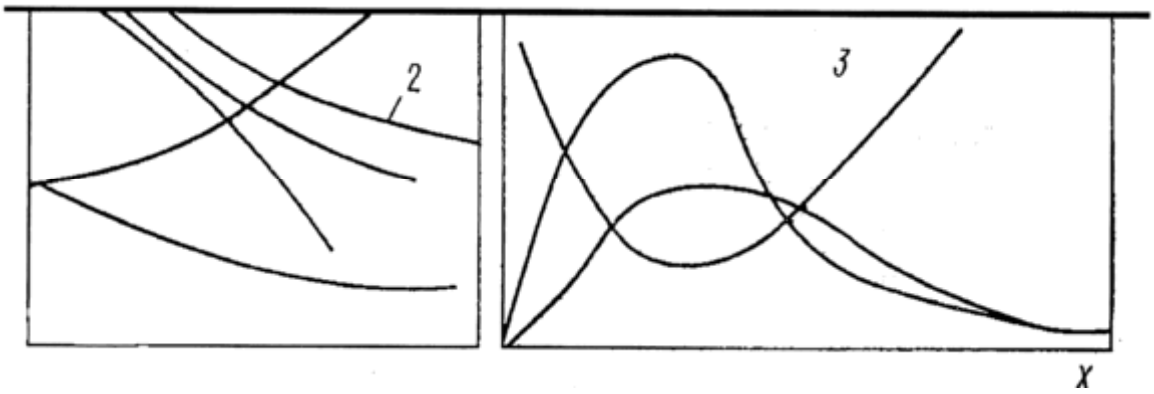


Рисунок 16.8 Графики кривых с тремя параметрами:
$1 - y=ae^{bx}+c$; $2 - y=ax^{b}+c$; $3 - y=ax^{b}e^{cч}$

Анализ основных моделей и выделение наиболее адекватных, применяемых для описания динамики древостоев, строения, связей между запасами насаждений и классами бонитетов, типов леса и описание взаимосвязей между другими таксационными показателями выполнил О.А. Атрощенко в монографии «Моделирование роста леса и лесохозяйственных процессов», которая приведена в списке литературы.

16.3 Верификация регрессионных моделей

Верификация модели - это ее проверка, т.е. подтверждение того, что модель верна (или не верна) и соответствует законам и закономерностям, действующим в генеральной совокупности. Верификация проводится в несколько этапов.

- Проверяется степень сглаживания (выравнивания) опытных данных, т.е. степень совпадения теоретических и экспериментальных значений функции при заданных аргументах. Для этого используют методы, рассмотренные ранее: основную ошибку уравнения регрессии, остаточную дисперсию, отношение общей и остаточной дисперсий, анализ остатков.
- Проверяют значимость коэффициентов уравнения регрессии по t-критерию, где t - должен быть 2 или 3 в зависимости от принятого уровня достоверности.
- Устанавливают степень коррелированности аргументов и принимают решение об их исключении или оставлении.

После выполнения перечисленных процедур приходят к заключению, что уравнение подобрано верно и хорошо описывает закономерность по результатам опыта. Если это не так, то подбирают другую модель.

Но при постановке опыта могут быть ошибки. Количество первичного материала может оказаться недостаточным для отражения всех особенностей изучаемого явления в генеральной совокупности и т.д. Поэтому требуется дополнительная проверка модели, описывающей ту или иную закономерность, но на новом материале, собранном независимо от нашего опыта. Его количество должно быть не меньше, чем использовано для построения модели. При этом сбор материала для проверки должен осуществляться в соответствии с правилами планирования эксперимента, т.е. правилами организации выборочных наблюдений. Иначе выборка будет нерепрезентативной. Здесь нельзя использовать “типичные” места замеров, “типичные” делянки и т.д., т.к. могут возникнуть смещенные оценки, т.е. имеющие систематическую ошибку.

Например, если мы, определяя выход сеянцев в питомнике, взяли более урожайную грядку в качестве “типичной”, то показатели будут завышены и наоборот. Когда нет систематической ошибки, то точность

можно рассчитать (и планировать), но систематическую ошибку исправить практически невозможно, т.к. она не поддается статистической интерпретации. Поэтому при планировании эксперимента надо строго выдерживать статистический метод.

В лесном хозяйстве исследуют обычно достаточно неоднородные совокупности. Например, изучая приживаемость лесных культур, мы можем обнаружить на участке места, где культуры прижились хорошо или где они погибли, скажем, в случае вымочек на понижениях.

Оценивая участки древостоев, пройденные рубками ухода, встречаемся с неравномерностью выборки деревьев на всей площади. Изучая смешанный древостой, видим, что деревья, скажем, сосны и ели, могут располагаться относительно равномерно или биогруппами, но встречаются и относительно большие куртины (до 0,05 га до 0,10 га и больше), состоящие из одной породы.

При проведении исследований (обследований), причем не только в научных, но и практических целях, требуется собирать различную информацию. Поэтому планирование эксперимента при изучении подобных объектов должно строиться на применении выборочных методов, отвергая закладку опытных участков в «типичных» местах.

Сказанное подтвердим примером, когда нам требуется провести исследование в совокупностях различных объектов (таблица 16.2).

Таблица 16.2 – Информация, требуемая для проведения исследования разных объектов

№ п/п	Наименование совокупностей (объектов)	Требуемая информация	Метод сбора
1	Средний состав дубовых древостоев в лесхозе (области, республике)	Состав на выделе в дубовой хозсекции	Маршрутный выборочный учет или анализ таксационных операций
2	Возобновление под пологом леса в ельниках	Средняя численность и возраст возобновления на 1 га	Закладки учетных площадок
3	Состояние 1-2 летних культур сосны	Процент приживаемости	Закладка учетных площадок
4	Средний запас спелых сосновых древостоев	Запас $м^3$/га	Закладка пробных площадей

Чтобы получить статистически значимую информацию, например, для второй из названных совокупностей, т.е. при учете естественного возобновления, надо провести наблюдение на большом числе учетных площадок или проанализировать много выделов. Численность возобновления от выдела к выделу и от площадки к площадке внутри выдела будет сильно варьировать. Коэффициент вариации υ может достигать 100% и выше. Понятно, что в этих условиях ни одна, ни десятки площадок, заложенных по принципу типичной выборки, не дадут надежной инфор-

мации. Так, для учета с точностью 10% надо заложить 100 площадок: $N = \frac{V^2}{P^2} = \frac{100^2}{10^2} = 100$. При 5% точности (с достоверностью 0,68) уже требуется огромный исходный материал: $N = \frac{100^2}{5^2} = \frac{10000}{25} = 400$ шт. Поэтому в практике обычно ограничиваются точностью в 10-15% при достоверности 0,68.

Аналогичная картина будет для первого примера. Состав дубовых насаждений от выдела к выделу и от лесхоза к лесхозу, даже от лесничества к лесничеству будет меняться. Понятно, что «типическая» выборка здесь ничего не дает, т.к. «типичность», т.е. средний состав, еще предстоит определить.

Для обследования лесных культур с целью установления их приживаемости в лесном хозяйстве существуют специальные правила, которые требуют подсчитать количество сеянцев на определенной площади. При этом предполагается выборочным путем, чаще всего методом систематической выборки, закладывать учетные площадки или ленты (ряды).

Для нахождения среднего запаса в древостоях, которые намечаются в главное пользование, необходимо заложить пробные площади, естественно, в спелых сосняках. Но вариация здесь высока из-за различия в условиях произрастания и полноте. Поэтому и в подобных ситуациях планирование эксперимента базируется на методах, разработанных в математической статистике.

Непригодность типического способа отбора выборки в описанных случаях состоит не в том, что способ недостаточно точен, а в том, что он не свободен от субъективизма. Несмотря на этот недостаток, способ типической выборки приходится нередко применять при оценке (таксации) насаждений. Например, при глазомерной таксации средний диаметр деревьев древостоя определяют как среднюю величину из 3-4, взятых “на глаз” средних по толщине деревьев. При детальном обследовании участка леса на зараженность вредителем приходится закладывать пробную площадь в средних (типичных) условиях. Иногда таким же образом решают и задачу оценки возобновления.

По неизбежности, как способ, требующий наименьшего объема наблюдений, метод типической выборки находит широкое применение при решении многих задач в лесном хозяйстве. Следует отметить, что информацию, пригодную для статистической оценки опыта с определенной точностью, этим способом получить невозможно. Можно вычислить среднюю величину признака на основе такого материала. Но ошибку этой средней определять не следует, так как она неправомерна, ибо случайная изменчивость признака (среднее квадратическое отклонение) в опыте не измерялась. В этом случае придется для оценки опыта ограничиться типической выборкой, т.е. только полученной средней величиной признака.

Для статистического истолкования результатов опыта рассмотренных совокупностей (возобновление на вырубке, сосняки лесничества, культуры сосны) с определенной точностью требуется закладка проб по одному из способов случайного отбора, рассматриваемых ниже. При этом

исследователю придется решить два главных вопроса: 1) определить достаточное число наблюдений, 2) правильно отобрать единицы для наблюдений.

При решении первого вопроса можно воспользоваться формулой

$$N = \frac{V^2}{P^2}, \qquad (16.12)$$

где V – коэффициент вариации; P – показатель заданной точности.

Выше показано, что для получения результата с точностью 5% для оценки возобновления, где коэффициент вариации (υ) нами установлен в 100 %, потребовалось бы заложить 400 площадок. При этом наше заключение о том, что полученная выборочная средняя будет отличаться не более чем на 5% от генеральной средней, дается с вероятностью 0,68. Это значит, что в 32 случаях из 100 эта закономерность может и не подтвердиться. Если принять уровень безошибочного суждения 0,05, т.е. делать заключение с вероятностью 0,95, которую следует считать достаточной, то в формулу для определения числа наблюдений нужно ввести множитель t (t - критерий). При вероятности 0,95 t-критерий примерно равен 2 (1,96).

Тогда формула для числа наблюдений будет

$$N = \frac{V^2 \cdot t^2}{P^2} \qquad (16.13)$$

Для нашего объекта число наблюдений составило бы

$$N = (2^2(100^2) / 5^2 = 1600.$$

В других случаях, когда υ по пробной выборке характеризовалось бы меньшим числом, можно было бы планировать меньшее число наблюдений.

Из приведенного примера видно, что, если варьирование значений признака в совокупности велико и полученное N по формуле (16.13) практически недостижимо, исследователь должен пойти на сужение объекта исследований или иногда довольствоваться точностью опыта в 10 и даже 15%.

Правильнее первое решение. Можно взять, например, в опыте не все ельники лесхоза, лесничества, а только какого-то типа леса или типа условий местопроизрастания, желательно наиболее распространенные. В пределах таких объектов варьирование будет значительно меньшим. Вообще, следует придерживаться принципа брать более ограниченные совокупности.

16.4 Применение регрессионных моделей в лесном хозяйстве

В лесоводственных исследованиях регрессионные модели – необходимый инструмент для решения возникающих задач. При этом важнейшим этапом, который надо строго выдерживать и в практической работе, является планирование эксперимента в строгом соответствии с законами биометрии.

Решения вопроса по планированию эксперимента состоит в правильном отборе или размещении единиц наблюдения. Современная статистическая теория рекомендует для этого ряд способов.

Обычно применяют следующие способы: простой случайный отбор, или случайное бесповторное выборочное наблюдение, случайное послойное выборочное наблюдение, систематическое выборочное наблюдение и субвыборочное наблюдение, или двухстадийное наблюдение.

Простой случайный отбор является наиболее распространенным и статистически разработанным методом. Его организуют с помощью какого-либо механизма, обеспечивающего равную возможность для любой единицы попасть в выборку. Обычно для выбора единиц используют таблицу случайных чисел.

Но применить этот способ в чистом виде при исследованиях в лесном хозяйстве часто очень трудно, а иногда и невозможно. Например, закладка учетных площадок в пределах лесхоза потребует большого количества времени на переезды и поиск нужного объекта.

Более приемлема и наиболее часто применяется систематическая выборка. По точности и обоснованности она практически не уступает случайной, являясь своеобразным вариантом последней.

Систематическая выборка полностью определяется выбором первого ее члена. Выбирают для обмера или наблюдения, допустим, каждый десятый член, например 10, 20, 30-е и т.д. дерево по перечету или 10, 20, 30-й и т.д. ряд культур. Закладка учетных площадок через определенное расстояние друг от друга представляет также систематическую или механическую выборку. Преимущество такой выборки - легкость ее получения и равномерность распределения по всему объекту. Ее недостатком могут быть случаи, когда совокупность обладает периодической изменчивостью и если интервал между отбираемыми единицами совпадает с длиной волны этого изменения (или кратный ей), то получим выборку со смещением, т.е. с систематической ошибкой.

Допустим, что многорядный агрегат, который применяли при посеве некоторой культуры, имел один из шести или другого числа неисправный захват. Если в последующем учете №№ учетных рядов культур совпадают с рядами, произведенными неисправным захватом, то выборка будет содержать систематическую ошибку. Это всегда следует учитывать при планировании опыта. В условиях равномерной изменчивости признака, можно применять систематическую выборку и без существенной погрешности обрабатывать как случайную. После получения дополнительно-

го материала, собранного с использованием статистического метода, можно рекомендовать ее использование в широкой практике.

Окончательное суждение о верности модели и ее широком применении может быть сделано только после проверки на практике, т.е. модель должна быть опробована в производственных условиях. Только тогда можно учесть все возможные особенности закономерностей, действующих в генеральной совокупности. Если модель подобрана правильно и хорошо проверена, то при ее использовании в практике не обнаруживается существенных недостатков. Но небольшая коррекция все равно может иметь место.

Например, сортиментные таблицы Ф.П. Моисеенко с 1958 по 1968 годы были неоднократно проверены в практике. Но их новый усовершенствованный вариант до издания в 1972 году прошел дополнительную проверку в течении 1970-71 гг. и лишь после внесения небольших уточнений был принят и издан. Зато с 1972 года, т.е. уже более 37 лет эти таблицы применяются в практике без замечаний.

Это положительный пример, хотя есть и такие, когда модель приходилось менять. Например, первоначальную модель для упрощенного отвода лесосек, разработанную в 1986 году В.Ф. Багинским, через год после опытной проверки автор заменил на более совершенную.

Регрессионные модели – основа для составления почти всех нормативов в лесном хозяйстве, и они должны тщательно отбираться и проверяться, чтобы хозяйство в лесу велось на должном научном уровне.

Все нормативы для лесного хозяйства, где встречаются количественные зависимости, разработаны с помощью описанных методов. Так, «Правила рубок в лесах Республики Беларусь» базируются на закономерностях изменения прироста древостоя (z) в зависимости от полноты ($П$). Это регрессионная связь $z=f(A, П, Б)$, где A и $Б$ – возраст и класс бонитета. Выход сортиментов (C) в товарных таблицах определяется по его связи со средними диаметром (D), высотой (H) и классом товарности (K), т.е. $C=f(D, H, K)$. Стоимость древесины (Cm) зависит от породы (P), крупности сортимента (Kp) и его сорта (S), т.е. $Cm=f(P, Kp, S)$. Этот показатель учитывают при расчетах эффективности проводимых лесохозяйственных мероприятий. Спелость леса и возраст рубки находят, используя связи выхода сортиментов, их стоимости и др. от возраста. Эти примеры можно продолжить.

Все сказанное свидетельствует о том, что в лесном хозяйстве регрессионные методы применяются повсеместно.

17. ИЗМЕРЕНИЕ СВЯЗИ МЕЖДУ КАЧЕСТВЕННЫМИ ПРИЗНАКАМИ

17.1 Особенности статистического анализа качественных признаков

В лесном хозяйстве приходится иметь дело не только с объектами, отличающимися количественно, но и качественно. Например, растения и животные одного вида варьируют по окраске, иногда вместо многочисленных замеров величины (веса, размера) семян их можно разделить по таким качественным категориям как цвет (оттенок), листья отличают по форме (эллипсовидные, круглые) и т.д. В этом случае, конечно, можно сделать оцифровку качественных признаков, но это несет определенную условность. Поэтому лучше воспользоваться специальными методами, разработанными для подобных случаев. Такие методы разработаны достаточно полно и хорошо изложены К.Е. Никитиным и А.З. Швиденко, П.Ф. Рокицким, Н.И. Сваловым и другими авторами, материалом которых мы воспользуемся.

Качественные признаки в конкретной задаче рассматривают как постоянные (тогда их обычно относят к основным факторам) или изменчивые. Будем называть качественный признак изменчивым, если он может принимать разные состояния или градации x_i. Например, различные интенсивности окраски хвои, семян, коры и пр.; богатство условий местопроизрастания - боры, субори, судубравы, дубравы; категории защитности лесов и др. Для изменчивых признаков целесообразно вводить понятия, аналогичные статистикам распределения и связи случайной величины - среднее, медиану, моду, корреляции и др.

Простейший случай - когда в опыте отмечают только наличие или отсутствие некоторого признака А, т.е. А имеет две градации: А и “не А”. Тогда, обозначив одни из них 1, а второй 0, изучение статистических задач, связанных с данным качественным признаком, можно свести к закономерностям биномиального распределения. В целом к этому приему прибегают и в более сложных ситуациях: различным качественным градациям ставят в соответствие количественный признак. Например, некоторое количество желудей может быть распределено по градациям интенсивности блеска (блестящие, тусклые, матовые), древостои различаются по типам условий местопроизрастания, где определены их площади или запасы и т.д. Тогда, обозначив значения количественного признака

через Y_i, можно записать качественный признак аналогично ряду распределения случайной величины

$$\begin{matrix} x_1 & x_2 \ldots & x_k \\ y_1 & y_2 \ldots & y_k \end{matrix} \qquad (17.1)$$

Примеры распределения качественных признаков для разных объектов показаны в таблице 17. 1.

Таблица 17.1 – Качественные признаки для разных совокупностей

Наименование совокупностей	Градация	Количество	
		га	%
Семена сосны в партии семян	блестящие	-	75
	тусклые	-	20
	матовые	-	5
	итого	-	100
Змеи (гадюки) в лесах Беларуси	черные	-	85
	серые	-	12
	иные	-	3
	всего	-	100
Распределение сосновых древостоев по группам типов условий произрастания *i* лесхоза.	боры	43260	65,0
	суборы	20880	31,4
	судубравы	1820	2,7
	дубравы	580	0,9
	итого	66540	100

По терминологии К. Джини качественный признак называют упорядоченным, если его градации образуют естественную последовательность, в противном случае он является неупорядоченным. Упорядоченные признаки разделяют на прямолинейные и циклические. Для прямолинейных признаков естественным образом можно выделить две крайние градации, иначе говоря, упорядочить признак по градациям от меньшего к большему или наоборот. В качестве примера приведем интенсивность окраски семян сосны или гадюк, а также увеличение богатства условий местопроизрастания, что показано в таблице 17.1.

Объекты, обладающие качеством прямолинейного типа, можно ранжировать, т.е. сопоставить каждому из них число натурального ряда. Если два объекта (градации) имеют ранги 2 и 5, значит между ними находится 3 других объекта или градации. Ранжировка исходных данных - широко распространенный прием при обработке не только качественных, но и количественных признаков. Можно ранжировать области Беларуси по проценту лесистости, охотничьи угодья по числу диких животных, предприятия лесного хозяйства по количеству работающих или объему заготовливаемой древесины и т.д.

Ранги бывают несвязными, когда градации объектов позволяют осуществить ранжировку полностью, и связными, когда некоторая часть объектов занимает “одинаковое место”. Например: 4 объекта можно равноправно разместить после объекта с рангом 2. Тогда им приписывают ранг $\frac{1}{4}(3+4+5+6) = 4\frac{1}{2}$.

Для циклических признаков две крайние градации нельзя выделить на объективной основе. Примеры циклических признаков - времена года (весна, лето, осень, зима), ориентировка склонов в горной местности по странам света и т.д. Примеры неупорядоченных качественных признаков - типы леса некоторого региона (по любой из существующих классификаций), методы окулировки при прививках и т.д.

Основные статистики для качественных признаков. Для вычисления качественных признаков исходят из условия, что различные градации этих признаков количественно неизмеримы или измеримы, но измерения в силу каких-либо причин проведены не были. Однако для дальнейшего анализа допускаются следующие предположения:

- Имеется возможность установления между градациями некоторой меры различия.
- Различие между соседними градациями одинаково. Такие признаки называют равнопромежуточными.

Если выполняется первое условие, то признак может быть сделан равнопромежуточным введением недостающих, возможно и ненаблюдавшихся, градаций.

Для качественного равнопромежуточного ряда типа (17.1) понятие среднего значения вводится по аналогии с количественными признаками: градации X_i ставится в соответствие некоторое число a, а постоянной разности между X_i и X_{i+1} - некоторое число h. Тогда вместо ряда (17.1) имеем

$$\begin{array}{llll} a, & a+h, & \ldots, & a+kh, \\ y_1, & y_2, & \ldots, & y_k, \end{array} \qquad (17.2)$$

Исследуя ряд (17.2), можно получить некоторые численные придержки относительно качественного признака X. Так, среднее арифметическое получим из следующего уравнения (17.3)

$$\tilde{X}_{\text{кач}} = -\frac{\sum_{i=0}^{k}(a + ih)y_i}{\sum_{i=1}^{k} y_i} \qquad (17.3)$$

Однако формальный перенос операций, введенных для случайных величин, оправдан только в том случае, если вычисленным величинам

можно приписать реальный смысл, и они проясняют какие-либо обстоятельства, связанные с решаемой задачей.

Для примера возьмем распределение сосновых древостоев некоторого (i-того) белорусского лесхоза. Общая площадь сосняков в лесхозе 66540 га. По классам бонитета они распределены следующим образом (таблица 17. 2).

Таблица 17.2 – Распределение сосняков по классам бонитета

Класс бонитета	I^a	I	II	III	IV	V	Va	Итого
Площадь, га	720	2070	41130	17480	2920	1810	410	66540
Площадь, %	1,1	3,1	61,8	26,3	4,4	2,7	0,6	100

Класс бонитета здесь выступает как качественный признак, заменяя показатель условий местопроизрастания. Класс бонитета выражают римскими цифрами а крайние бонитеты имеют индексы «*а*» и «*б*». Для того, чтобы найти средний класс бонитета превратим ряд в таблице 17.2 в равнопромежуточный, зашифровав римские цифры арабскими (таблица 17.3).

Таблица 17.3 – Распределение сосняков по классам бонитета при замене римских цифр арабскими в равнопромежуточном ряду

Класс бонитета	1	2	3	4	5	6	7	Итого
Площадь, га	720	2070	41130	17480	2920	1810	410	66540
Площадь, %	1,1	3,1	61,8	26,3	4,4	2,7	0,6	100

По данным таблицы 17. 3 обычным путем вычислим условное значение класса бонитета $\bar{x} = \frac{\sum x_i m_i}{\sum m_i}$. При этом результат должен быть одинаковым, в пределах точности округления если используем показатели площадей в га или процентах. Для сокращения опустим промежуточные величины.

$$\bar{x} = \frac{226500}{66540} = 3{,}40 \qquad \bar{x} - \frac{340{,}3}{100} - 3{,}4$$

Таким образом, условное среднее равно 3,4. Возвращаясь к таблице 17.2 и заменяя цифры классов бонитета на их обычные обозначения, получим, что средний класс сосновых бонитета сосновых древостоев в нашем лесхозе равен II, 4, что близко к среднереспубликанскому показателю (II, 3) по этой породе.

Подобным образом можно вычислять и другие показатели лесного фонда, чем широко пользуются в практике. Например, нам требуется определить средний фактический запас на 1 га в древостоях, показанных в таблице 17.2.

Для усложнения задачи допустим, что древостоев IV и V класса бонитета у нас нет, а распределение запасов и площадей по классам бонитета выглядит как в таблице 17.4.

Таблица 17.4 – Распределение площадей и средних запасов ($м^3$/га) по классам бонитета для сосновых древостоев

Класс бонитета	I^a	I	II	III	IV	V	V^a	Итого
Шифр класса бонитета	1	2	3	4	5	6	7	-
Площадь, га (П)	720	2070	41130	20400	-	-	2220	66540
Запас $м^3$/га (м)	220	180	150	110	-	-	40	-

Приведенный в таблице ряд не является равнопромежуточным, т. к. отсутствуют значения между III и V^a классами бонитета. Но средний запас здесь легко определяется по формуле

$$M_{ср} = \frac{\sum M_i}{П_i} = \frac{9033300}{66540} \approx 136 \ м^3/га \qquad (17.4)$$

Вычисленная величина имеет реальный смысл, и этот метод широко используется для определения средних показателей в лесном хозяйстве.

Аналогичными приемами можно ввести аналог размаха: широту интервала изменчивости признака, дисперсию и среднее квадратическое отклонение (здесь важна равнопромежуточность ряда), установить медиану, моду и т.д.

17.2. Основные методы анализа качественных признаков

Качественные альтернативы обычно обозначают латинскими буквами, располагая их в виде таблиц. Данные располагают как правило в четырехпольной таблице, показанной на рисунке 17.1, которая является исходным материалом для последующего анализа.

a	b
c	d

Рисунок 17.1 Таблица расположения качественных признаков

Учитывая, что анализ качественных признаков особенно важен в генетике, лесной селекции, молекулярной биологии, то этому вопросу уделено больше внимания в учебниках биометрии для биологов Г.Ф.Лакина, П.В.Рокицкого и др. (они приведены в списке литературы), на которые мы будем опираться в изложении этих вопросов.

Распространенным методом анализа качественных признаков является метод четырех полей. Учитывая обозначения, показанные на рисунке17.1 степень сопряженности, существующей между качественными признаками, или альтернативами, определяется по формуле Д. Юла

$$r = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \qquad (17.5)$$

В данном случае примеры удобно брать из области генетики. Поэтому проведем анализ качественных признаков на излюбленном объекте генетиков – мушке дрозофиле. От скрещивания серого самца, имевшего длинные крылья, с черной самкой, у которой они были рудиментарными, все потомство оказалось одинаковым. А при скрещивании особей из первого поколения между собой во втором поколении гибридов были получены следующие категории мух (таблица 17.5).

Таблица 17.5 – Распределение признаков в потомстве мушки дрозофилы

Признак	Количество, шт	%
Серые длиннокрылые с зачатками крыльев	20	41,7
Черные длиннокрылые	20	8,3
Черные длиннокрылые с зачатками крыльев	100	41,7
Итого	210	100

Теперь сделаем анализ приведенных качественных признаков для установления между ними корреляционных зависимостей. Решение этой задачи сводится к вычислению коэффициента корреляции по методу четырех полей, т.е. по формуле Юла (17.5). Обозначим через Х черную окраску тела, а через У - зачатки крыльев. Затем разместим данные эксперимента в четырехпольной корреляционной таблице (17.5).

Таблица 17.5 – Распределение расщепленных признаков у потомства дрозофилы

У / Х	Незачатковые крылья $У_1$	Зачатковые крылья $У_2$	Сумма
Нечерное тело $Х_1$	a=100	b=20	a+b=120
Черное тело $Х_2$	c=20	d=100	c+d=120
Сумма	a+c=120	b+d=120	a+b+c+d=240

Пользуясь этой таблицей, подставляем в формулу Юла нужные данные и находим величину коэффициента корреляции между указанными признаками:

$$r = \frac{100 \cdot 100 - 20 \cdot 20}{\sqrt{120 \cdot 120 \cdot 120 \cdot 120}} = \frac{9600}{14400} = +0{,}67$$

Оценка достоверности коэффициента корреляции, вычисленного по способу четырех полей, ничем существенно не отличается от обычной оценки значимости этого показателя, которую вычислим по формуле, приведенной ранее (глава 12, раздел 12.2) $m_r = \frac{1-r}{\sqrt{N}}$.

Для нашего примера:

$$m_r = \frac{1 - 0{,}67^2}{\sqrt{240}} = \frac{0{,}551}{15{,}49} = 0{,}035$$

Достоверность коэффициента корреляции определяем по t-критерию Стьюдента $t_\Phi = \frac{r}{m_r}$. Для приведенного примера: $t = \frac{0.67}{0.035} = 19.1$. Эта величина значительно больше, чем ее предельное (табличное) значение, которое находим из приложения Е при 1% уровне значимости , где t=2,58.

Но в зависимости от группировки и качества коррелируемого материала, расположения его по клеткам четырехпольной таблицы, как показали последние исследования, коэффициент корреляции, вычисленный по формуле Юла, может оказаться несколько завышенным. Поэтому в числитель формулы (17.5) рекомендуется вносить поправку следующего вида:

$$r = \frac{(ad - bc) - 0{,}5N}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \qquad (17.6)$$

Для получения информации о сопряженности между альтернативами используется также критерий x^2 Пирсона. Установлено, что между формулой (17.5) Юла и критерием Пирсона “хи-квадрат” существует следующая связь:

$$r = \sqrt{\frac{\chi^2}{n}} \qquad (17.7)$$

Ограничением применения критерия χ^2 является то, что это формула (17.17) применима только к четырехпольной таблице. Так, для взятого нами примера критерий “хи-квадрат” будет следующим

$$\chi^2 = \frac{n(ad-cb)^2}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} = \frac{240 \cdot 9600^2}{120 \cdot 120 \cdot 120 \cdot 120} = 106{,}7,$$

откуда $r = \sqrt{\frac{106{,}7}{240}} = \sqrt{0{,}45} = +0{,}67$. Получается тот же результат, что и при расчете по формуле Юла.

Коэффициент ассоциации. Английский ученый Дж. Юл предложил для оценки "тесноты" сопряженности между двумя парами альтернатив относить значение, стоящее в числителе формулы (17.5), не к корню квадратному из произведений суммарных данных по столбцам и строкам четырехпольной таблицы, а к сумме двух произведений ad+bc. Сконструированный таким образом показатель (обозначаемый К) получил название коэффициента ассоциации Юла. Его формула следующая:

$$K = \frac{ad-bc}{ad+bc} \qquad (17.8)$$

Для нашего примера этот показатель

$$K = \frac{100 \cdot 100 - 20 \cdot 20}{100 \cdot 100 + 20 \cdot 20} = +0{,}92$$

Коэффициент ассоциации всегда больше коэффициента корреляции. Коэффициент ассоциации свидетельствует о наличии параллелизма между числовыми значениями признаков, без учета их вариабельности, а следовательно, и не дает точного представления о существующей между ними связи. В этом и заключается причина того, что коэффициент ассоциации не получил широкого применения в практике.

Коэффициент контингенции. Существует еще один оригинальный показатель взаимной сопряженности, называемый коэффициентом контингенции и обозначаемый С. Он вычисляется по формуле Пирсона (17.9):

$$C = \sqrt{\frac{\Phi^2}{1+\Phi^2}}, \qquad (17.9)$$

где величина $1+\Phi^2 = \left(\frac{p_{xy}^2}{p_x} : p_y\right)$. Здесь p_{xy} - частоты, заключенные в клетках корреляционной решетки; p_x - число случаев в каждом вертикальном, а p_y - число случаев в каждом горизонтальном ряду корреляционной таблицы.

Покажем вычисление коэффициента контингенции на том же примере расщепления признаков в гибридном потомстве дрозофилы. Ход вычисления представлен в таблице (17.6).

Таблица 17.6 – Вычисление коэффициента контингенции

X_i \ $У_i$		Незачатковые крылья, Y_1	Зачатковые крылья, Y_2	P_x	$\sum\left(\frac{P_{xy}^2}{P_x}:P_y\right)$
1	P_{xy}	100	20	120	86,67/120=0,722
	P_{xy}^2	10000	400	-	
	$P_{xy}^2 : P_x$	83,33	3,33	86,67	
2	P_{xy}	20	100	120	0,722
	P_{xy}^2	400	10000	-	
	$P_{xy}^2 : P_x$	3,33	83,33	86,67	
	P_y	120	120	240	1,444

Из этой таблицы следует, что $1+\Phi^2=1{,}444$, откуда $\Phi^2=0{,}444$. Подставляя найденные значения в формулу (17.9), получим:

$$C=\sqrt{\frac{0{,}444}{1{,}444}}=\sqrt{0{,}307}=+0{,}55$$

Коэффициент контингенции меньше коэффициента корреляции, и по разности этих показателей можно судить о прямолинейной или криволинейной зависимости между признаками.

Преимущество коэффициента контингенции перед коэффициентом ассоциации заключается в том, что он позволяет измерять сопряженность не только между двумя, но и большим числом коррелируемых признаков.

В исследованиях по лесному хозяйству для определения тесноты связи между качественными признаками применяют методы, основанные на корреляции рангов. Такую корреляции с вычислением критерия Спирмена мы описали в главе 12, когда исследовали корреляцию между количественными признаками (x) и (y). Этот коэффициент можно использовать и для оценки ранговой корреляции качественных признаков.

Для установления корреляции качественных признаков существует также ранговый коэффициент корреляции Кенделла. Ранговые коэффициенты корреляции, описывающие качественные признаки, изменяются как и обычные коэффициенты корреляции от -1 до +1 в том же значении, т. е. как показатели прямой и обратной связи. Они меняют знак на обратный, если одну из

последовательностей рангов заменить на сопряженную, т.е. на место первого элемента поставить последний, за ним предпоследний и т.д.

Коэффициент ранговой корреляции Кендела определяется по формуле

$$\tau = \frac{2\sum x_j n \cdot y_i n}{N(N-1)} \qquad (17.10)$$

При замене $\Sigma x_j n \cdot y_i n$ на S получим $r = \frac{2S}{N(N-1)}$ (17.10а)

Формулы (17.10) и (17.10а) применяются в последовательности $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$. Каждой паре рангов (x_i-x_j, y_i-y_j) ставят в соответствие функцию. Последняя соответственно равна следующим величинам

x_{in} или y_{in} { +1 при $x_i, y_i < x_j, y_i$; 0 при $x_i, y_i = x_j, y_i$; -1 при $x_i > x_j, y_j$.

Вычисления, как правило, выполняют на компьютере по специальной программе, входящей в сертифицированное матобеспечение.

Для относительно простых случаев, т.е. при небольшой величине N, можно провести ручной счет, воспользовавшись формулой

$$\tau = \frac{4P}{N(N-1)} - 1 \qquad (17.11)$$

где P – количество пар, образующих прямую последовательность, имеют прямой порядок (j<i).

Основная ошибка τ определяется по формуле

$$m_\tau = \sqrt{\frac{2(2N+5)}{9N(N-1)}} \; . \qquad (17.12)$$

В качестве примера рассмотрим есть ли связь между светолюбием древесной породы и ее продуктивностью. Возьмем 6 древесных пород, растущих в одинаковых условиях произрастания – $Д_2$, т.е. в кисличном типе леса. Первый ряд (x_i) характеризует ранжирование по светолюбию, второй – по продуктивности (запас на 1 га).

Породы	С	Е	Л	Д	Б	Ос
x_i	2	6	1	4	3	5
y_j	3	2	1	5	6	4

Перепишем теперь ряд *x* в порядке возрастания рангов.

Породы	Л	С	Б	Д	Ос	Е
x_i	1	2	3	4	5	6
y_j	1	3	6	5	4	2

Теперь подсчитаем, сколько пар в ряду у образуют прямой порядок. Для Л таких пар 5(С, Б, Д, Ос, Е), для породы С таких пар 3 (Б, Д, Ос). соответственно для Б – 0, Д – 0, Ос – 0, Е – 0.

Тогда Р=5+3+0+0+0+0=8.

По формуле 17.11 определим τ и его основную ошибку (17.12)

$$\tau = \frac{4 \cdot 8}{6(6-1)} - 1 = \frac{32}{30} - 1 = 1{,}07 - 1 = 0{,}07$$

$$m_\tau = \sqrt{\frac{2(2 \cdot 6 + 5)}{9 \cdot 6(6-1)}} = \sqrt{\frac{34}{270}} = \sqrt{0{,}015} = \pm 0{,}12 .$$

Отношение τ к его основной ошибке равно $t = \frac{0{,}07}{0{,}12} = 0{,}58$. Так как m_i<3 и даже <2, то искомая зависимость недостоверна. Достоверность коэффициента ранговой корреляции Кендела t можно установить и по таблице критических значений t-критерия Стьюдента, приведенной в приложении Е.

17.3 Метод индексов

Наряду с методом корреляции и регрессионным анализом в практической работе лесоводов применяется метод индексов. Сущность его заключается в том, что величина одного признака, обычно меньшая, выражается в долях единицы или в процентах от величины другого признака:

$$I = \frac{Y}{X} \cdot 100\% , \qquad (17.13)$$

где I - индекс, выраженный в процентах;

Y и X - величины соответствующих признаков.

Метод индексов давно нашел широкое применение в области морфологии. Например, им пользуются, при оценке телосложения животных и человека. Достоинство этого метода заключается в его простоте и общедоступности, а также и в том, что он позволяет более полно, чем отдельно взятые измерения, характеризовать особенности телосложения и их динамику в онтогенезе. Поэтому наибольшее количество примеров описано именно биологами. Приведем те из них, которые описал Г.Ф. Лакин. Сделаем анализ двигательной активности мышей разных видов (таблица17.7).

Таблица 17.7 – Анализ двигательной активности мышей разных видом

Виды грызунов	Вес тела, г	Вес мозга, мг	Весовой индекс мозга	Двигательная активность животных
Мышь желтогорлая	24,2	726,0	21,8	Очень быстрый зверек, совершает дальние экскурсии
Мышь	23,5	569,9	13,6	Менее подвижная, чем мышь

полевая				желтогорлая
Полевка рыжая	21,3	541,3	13,8	Подвижная, но слабо связана с норой
Полевка серая	23,8	462,0	8,9	Больше держится около норы, слабо подвижная

Усиление двигательной активности животных приводит, очевидно, к относительно более быстрому увеличению веса мозга по сравнению с весом тела. Поэтому у зверьков, ведущих более активный образ жизни, обнаруживаются и более высокие весовые показатели мозга.

Чтобы установить филогенетическое положение вида, характеризовать высоту его прогрессивного развития, М.Ф. Никитенко использовал другой индекс, названный им коэффициентом теленцефализации и представляющий вес переднего мозга, выраженный в процентах к общему весу головного мозга.

В таблице 17.4 показано значение этого индекса у некоторых видов млекопитающих и у человека.

Таблица 17.8 – Значение индекса у некоторых млекопитающих и человека

Виды	Вес, г		Коэффициент теленцефализации, %
	головного мозга	переднего мозга	
Еж	2,85	1,36	47,8
Крот	0,85	0,44	51,2
Кролик	22,05	11,31	51,3
Белка	5,89	3,13	53,3
Бобр	41,40	26,60	64,3
Собака	71,10	48,49	68,2
Волк	218,80	153,20	70,3
Кабан	173,50	123,35	71,1
Дельфин	612,00	464,00	74,6
Макак	93,20	74,83	80,3
Человек	1482,70	1260,3 - 1289,9	85,0 - 87,0

Видно, что по мере движения вверх по лестнице эволюции индекс постепенно увеличивается: у насекомоядных он равен 47%, а у приматов достигает 80%, т.е. обнаруживает определенную закономерность в филогенетическом ряду животных.

Чтобы составить представление о возрастных изменениях пропорций тела у павианов-гамадрилов и макаков-резус в первые девять лет их жизни, были использованы коэффициенты роста, т.е. индексы, представляющие отношения между конечными и начальными размерами головы и туловища.

Результаты показаны в таблице 17.9, из которой видно, что индексы линейного роста разных частей тела у гамадрилов и макаков неодинако-

вы. Это следствие гетеродинамии, т.е. непропорционального увеличения размеров данных признаков в постэмбриональном развитии животных.

Индексы - показатели довольно неустойчивые, так как на их величине сильно сказываются ошибки, допускаемые при измерениях животных. Чтобы устранить этот недостаток, прибегают к предварительной обработке результатов измерений путем вычисления средних арифметических отдельных измерений или выравниванию динамических рядов, а иногда находят средние арифметические из отдельных индексов.

Таблица 17.9 – Индексы линейного роста разных частей тела у гамадрилов и резусов

Размеры головы и туловища	Гамадрилы		Резусы	
	самцы	самки	самцы	самки
Общая длина головы	2,1	1,9	1,8	1,7
Поперечный диаметр головы	1,8	1,6	1,5	1,5
Передне-задний диаметр головы	1,6	1,5	1,5	1,4
Скуловой диаметр	2,3	1,9	2,0	1,9
Глубина груди	3,3	2,7	2,9	2,6
Ширина груди	3,2	2,6	2,7	2,3
Окружность груди	3,5	2,8	3,0	2,7
Передняя длина туловища	3,1	2,7	3,2	3,1
Ширина таза	3,3	3,1	2.9	2,9

Например, прежде чем выяснить возрастные изменения пропорций тела у самок павианов-гамадрилов, результаты их измерений были подвергнуты предварительной обработке: по каждому промеру вычислялись средние арифметические, которые затем выравнивались по способу скользящей средней. В итоге получились данные, собранные в таблице 17.10.

Таблица 17.10 – Соотношение различных частей тела у гамадрилов

Возраст обезьян (мес.)	Число измерений	Длина корпуса (туловище+ +голова), см	Длина туловища, см	Окружность груди, см	Живой вес, кг
-	-	1	2	3	4
0-6	11	24,3	16,0	21,9	1,27
6-12	16	31,2	22,5	27,2	2,49
12-18	23	35,3	24,9	30,8	3,54
18-24	21	38,4	26,8	33,0	4,26
24-36	8	42,8	30,8	37,9	5,53
36-60	13	46,7	33,4	43,4	8,09
60 и >	9	50,4	38,0	48,4	12,90

На основании данных таблицы 17.10 вычислены индексы, приведенные в таблице 17.11, которые показывают, каким образом изменяются с возрастом животных соотношения между взятыми признаками.

Таблица 17.11 – Изменение индексов размера тела у гамадрилов с возрастом

Возраст обезьян, мес.	Индексы					
	1	2	3	4	5	6
	1:4	2:4	3:4	1:2	1:3	2:3
0 - 6	19,1	12,3	16,9	1,5	1,1	0,7
6 - 12	12,5	9,0	10,9	1,4	1,2	0,8
12 - 18	10,1	7,1	8,7	1,5	1,1	0,8
18 - 24	8,9	6,2	7,7	1,5	1,2	0,8
24 - 36	7,8	5,6	6,9	1,4	1,1	0,8
36 - 60	5,8	3,8	5,4	1,4	1,1	0,8
60 и >	3,9	3,1	3,9	1,3	1,1	0,8

Если значения индексов нанести на график, как это показано на рисунке 17.2, нетрудно убедиться в довольно строгой зависимости между ними.

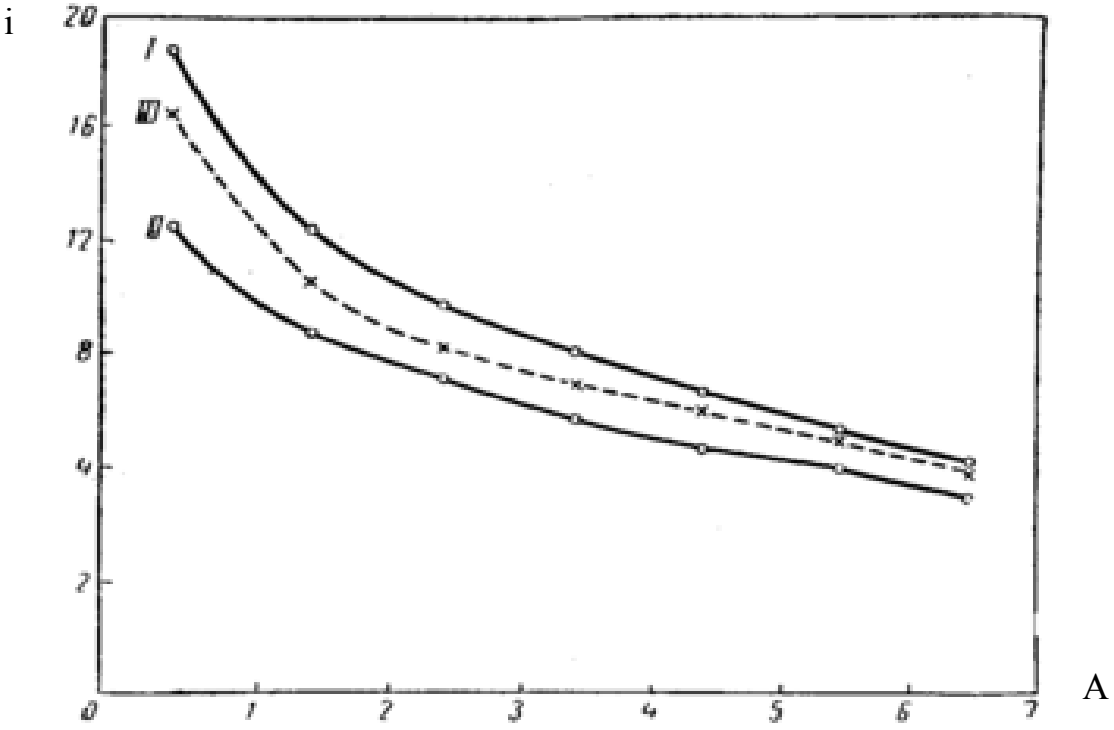


Рисунок 17.2 Возрастные изменения индексов между основными размерами тела и живым весом у самок павианов - гамадрилов:

на оси абсцисс - возраст в годах, на оси ординат - значения индексов

Она выражается уравнением общего вида

$$y = \left(\frac{x}{a} - b\right)^2,$$

где y - живой вес обезьяны в кг;
x - длина туловища в см.;
a и b - параметры уравнения, a = 10, b = 0,5.

Отсюда определить живой вес (y) растущей самки павиана-гамадрила по длине ее туловища (x), выраженной в см, можно по следующей формуле:

$$y = \left(\frac{x}{a} - 0{,}5\right)^2$$

Индексы могут быть одноименные, когда сопоставляемые признаки выражены в одних и тех же величинах измерения, и разноименные, когда признаки выражаются разными единицами меры, как, например, вес и линейные размеры тела и т.п. Различают также индексы ростовые, объемные, весо-ростовые и другие, а также индексы туловища, головы и т.д.

Недостатки метода индексов

За более чем столетнюю историю этого метода было предложено немало различных индексов как зоотехниками, так и антропологами. Однако несмотря на простоту, общедоступность и широкую популярность метода, он имеет весьма существенные недостатки. Дело в том, что индексы - величины относительные. Они не учитывают ни вариабельность признаков, ни степень сопряженности между ними, как это свойственно другим биометрическим показателям - коэффициентам регрессии и корреляции. Известно также, что биологические признаки не только варьируют, но обнаруживают неодинаковую скорость развития и роста (гетеродинамия), что не может не отразиться на величине индексов. Постоянство индексов сохраняется только при относительно одинаковых скоростях развития частей тела в онтогенезе, т.е. в случаях гомодинамии, когда отношения средних арифметических сопоставляемых признаков X и Y равняются коэффициентам регрессии, т.е. при условии, что

$$\frac{\overline{X}_x}{\overline{X}_y} = R_{x/y} \quad \text{и} \quad \frac{\overline{X}_y}{\overline{X}_x} = R_{y/x} \qquad (17.14)$$

Но такое равенство в соотносительной изменчивости различных частей организма, как правило, не наблюдается. Индексы и коэффициенты регрессии обычно полностью не совпадают друг с другом. Иллюстраци-

ей этого может служить таблица 17.12, в которой приводятся сравнительные данные определения нормального веса человека по длине его тела на основании использования индекса Брока.

$$i = P/TL, \text{ где}$$

P – вес тела человека; T – обхват (периметр) грудной клетки; L – рост человека.

Таблица 17.11 – Сравнительные данные определения нормального веса человека по длине его тела

Длина тела, см	Вес тела, кг		
	по индексу Брока	по уравнению регрессии	
		мужчины	женщины
150	50	54	51,5
155	55	57	55,0
160	60	60	58,5
165	65	63,5	62,0
170	70	66,5	65,5
175	75	70	69,0
180	80	73	72,5
185	85	76	76,0

Из таблицы 17.8 видно, что оценка веса тела по его длине на основании индекса Брока лишь приблизительно совпадает с величинами, вычисленными по уравнению регрессии, да и то лишь при среднем росте человека 160 - 165 см.

Антропологи считают, что метод индексов совершенно непригоден для оценки физического развития организма и должен быть оставлен как пройденный этап в развитии науки. В то же время зоотехники и многие врачи продолжают пользоваться индексами в своей работе, применяя их там, где можно ограничиться приближенными данными и где не требуется достаточно точных показателей физического развития организма. В настоящее время этот недостаток в науке прошлого века полностью преодолен; стали применяться и более совершенные методы исследования биологических явлений, и это тоже вполне закономерное явление. Несомненно, что наиболее точные методы для оценки физического развития человека и животных - это методы регрессии и корреляции. Только благодаря этим методам оценка физического развития человека была поставлена на строго научную основу - она производится в настоящее время с помощью номограмм по трем шкалам измерений - длине тела, обхвату груди и живому весу человека.

Однако, по-видимому, всякий метод имеет присущие ему положительные черты и недостатки, и различные по своей конструкции биометрические показатели не исключают, а дополняют друг друга. Поэтому в практической деятельности лесоводу приходится выбирать то одни,

то другие методы, сообразуясь с целым рядом обстоятельств, от которых зависит этот выбор. Все дело в том, чтобы понимать конструкцию данного показателя, видеть его положительные стороны и недостатки. В умелых руках метод индексов может оказать лесоводам неплохую услугу. Индексы в равной мере характеризуют и прямолинейные, и криволинейные и какие угодно соотношения, существующие между биологическими признаками, чего нельзя сказать о других показателях связи. Поэтому там, где этот метод приложим и дает удовлетворительные результаты, его можно использовать хотя бы как вспомогательное средство в комплексной оценке различных признаков в биологии, сельском и лесном хозяйстве.

17.4. Применение статистических методов исследования качественных признаков в лесном хозяйстве

Методы исследования качественных признаков в лесном хозяйстве применяются достаточно широко, хотя по объему использования эти способы уступают количественному анализу явлений. Применение методов исследования качественных признаков значительно расширилось за последние 10-15 лет. Лесоводы, как и ученые других специальностей, изучают лесных зверей и птиц, где необходимо выделять ряд качественных признаков. Качественные признаки применяют для оценки доброкачественности лесных семян, спелости лесных ягод и т.д. Здесь можно назвать работы А.З. Швиденко, В.Б. Гедых, Г.Г. Гончаренко, В.Е. Падутова и других. Для изучения лесных зверей и птиц, где необходимо выделить ряд качественных признаков (окрас, наличие или отсутствие рогов и т.д.), для анализа лесных семян (цвет, блеск и др.), формы листьев (округлые, эллипсовидные и пр.) используют различные качественные показатели: ранговой корреляции Кендалла, метод индексов и т.д. Находят применение и другие статистические методы оценки качественных признаков.

Особое место в исследованиях по лесному хозяйству занимает метод индексов. Его широко применяют лесоводы в своей практике, находя его весьма удобным и часто незаменимым. Наибольшие применение метод индексов находит при сопоставлении величин, имеющих разные единицы измерения. Так, если требуется учесть биомассу живого напочвенного покрова, скажем черники в совокупности, с запасом древостоя, для вычисления показателей комплексной продуктивности разных типов леса в переводе на 1 га, то метод индексов нас устроит.

Когда О.В. Лапицкой необходимо было определить возраст эколого-экономической спелости древостоев наших лесов, то потребовалось совместить экономические показатели, которые выражались величиной наивысшей рентабельности в некотором возрасте, и экологические, выражавшиеся в тоннах связанного CO_2 на 1 га. Найти возраст оптимального сочетания этих показателей, оцениваемых по разным критериям, оказалось возможным, применив метод индексов. Возраст древостоев, в котором сумма индексов

выражающих экономические и экологические показатели, была максимальной, характеризовал эколого-экономическую спелость.

Метод индексов широко используется в лесной таксации. Так В.В.Загреев (1932-1993) разработал специальный метод исследования динамики древостоев и составления таблиц хода роста с помощью метода индексов.

Он установил, что среди большого числа кривых хода роста нормальных насаждений по любому таксационному показателю существует много сходных. На основании анализа свыше 400 таблиц хода роста нормальных насаждений различных природных районов СССР и ряда зарубежных стран этот автор выявил возможность систематизации, классификации и стандартизации таких таблиц. Для сосняков им разработаны таблицы типов роста насаждений, представляющие собой усредненные и систематизированные с определенной градацией относительные (индексные) ряды хода по отдельным таксационным показателям. Например, для характеристики с точностью ±3% большого количества существующих таблиц хода роста сосновых насаждений имеющих большое многообразие линий хода роста в высоту, оказалось достаточным иметь одну таблицу, включающую лишь 13 типовых рядов. Такие индексные таблицы (графики) служат для сравнительной качественной оценки и группировки таблиц по степени сходства и различия в характере (типе кривых) хода роста. На рисунке 17.3 для примера приведен график типовых линий хода роста сосновых насаждений по сумме площадей сечений.

Типовые графики и таблицы имеют общее значение, так как заменяют все встречающиеся в природе линии хода роста насаждений по отдельным таксационным показателям. Типизация хода роста насаждений позволяет привести все таблицы хода роста нормальных насаждений в строгую систему, устранить существующую путаницу в их практическом применении, оценить существующие таблицы хода роста; выявить общие закономерности и географические различия и особенности роста насаждений.

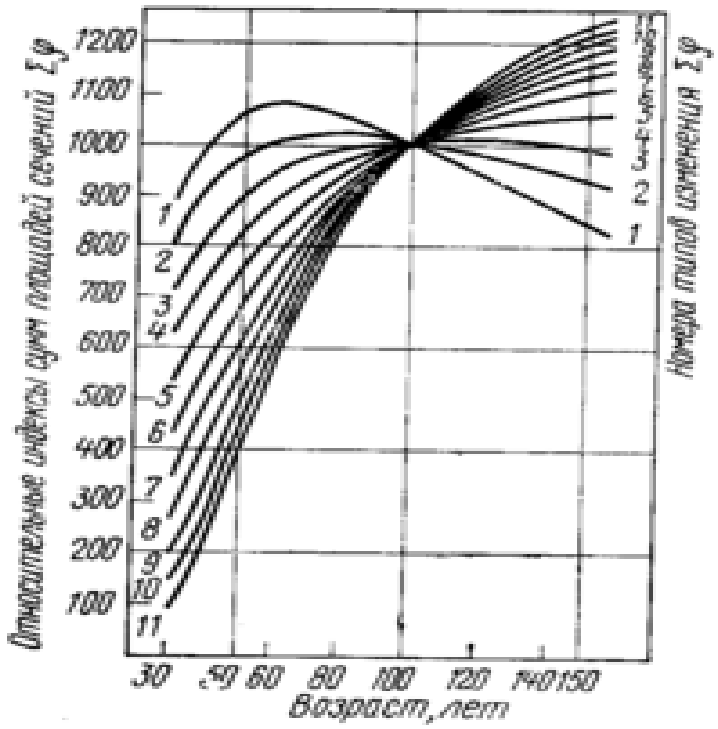


Рисунок 17.3 Типовые линии изменения индексов сумм площадей сечений

Но метод индексов имеет и недостатки, которые проявились и в данном случае. Типовые линии В.В. Загреева имеют достоверность 0,68, т. е. в 32% случаев точность типовых линий выходит за установленные пределы. Потому типовые линии правомерно применить лишь для большого количества объектов, не менее крупного региона.

В то же время метод индексов требует очень тщательного и высокопрофессионального предварительного анализа изучаемого явления. Без соблюдения этого условия, как и в случаях, описанных выше, он приводит к ложным выводам. Примером некорректного использования метода индексов могут служить исследования, проведенные в 70-е годы прошлого века В.И. Зерновым по изучению влияния широкомасштабной гидротехнической мелиорации сельхозземель на прирост прилегающих лесных массивов. Этот случай описан в литературе В.Ф. Багинским. В.И. Зернов использовал относительный (индексный) показатель $K_i = \frac{Z_a}{M_a}$, где Z_a - текущий прирост в $м^3$ в возрасте a; M_a - запас в этом же возрасте. Но в силу того, что примененные индексы K_i имеют сходимость рядов, то спустя 10-20 лет после наблюдения различия искусственно нивелируются, хотя на самом деле существенная разница сохранялась. Полученная величина оказалась непригодной для анализа указанного явления, т.к. искажала его суть и привела к ложным выводам. В результате неверного методического подхода к использованию метода индексов 10-летняя работа целой лаборатории БелНИИЛХа оказалась выполненной впустую. Предотвратить негативные последствия этих ложных выводов стало возможным лишь благодаря корректному ана-

лизу названной работы при ее последующем рецензировании и проверке до практического использования.

Из приведенных примеров следует вывод, который проходит красной нитью через все изучение биометрии: прежде чем приступать к манипулированию цифрами, надо изучить суть явления и понять его основные законы и закономерности на общем (профессиональном) уровне. Для этого нужен предварительный анализ, а уж затем корректно следует смоделировать явление или процесс с помощью приемов биометрии.

18. ДИСПЕРСИОННЫЙ АНАЛИЗ

18.1 Понятие о дисперсионном анализе

При проведении исследований часто надо определить, насколько существенно влияние одного или нескольких факторов на конечный результат. Не всегда здесь можно применить регрессионный анализ с получением разного вида моделей. Наиболее часто такие проблемы возникают в генетике, лесной селекции, лесовосстановлении, лесной энтомологии и в других областях лесной науки. В этом случае необходимо проводить статистический анализ результатов наблюдений, зависящих от разных одновременно действующих факторов, делать выбор этих факторов и оценку “силы” их влияния.

Основой решения перечисленных вопросов является изучение стабильности и однородности дисперсий изучаемого признака и разложение ее на составляющие, порожденные действием рассматриваемых факторов. Некоторые из факторов, меняющиеся в эксперименте или наблюдении (например, порода деревьев, тип леса, цвет желудей и пр.), могут быть качественными, другие количественными. Количественными величинами обычно выражают параметры деревьев и древостоев: высота, диаметр, запас древостоя и т. д. В то же время некоторые из этих показателей могут выражаться и качественными признаками, например деревья крупные, мелкие или деревья первой, второй, третьей величины и т. д.

В зависимости от соотношения между количественными и качественными факторами применяют один из трех достаточно близких по идеям и математическому аппарату методов анализа: регрессионный, дисперсионный и ковариационный. В регрессионном анализе подход количественный, в дисперсионном все факторы рассматривают как качественные; в ковариационном анализе часть факторов изучают как количественные, другую часть - как качественные.

Сказанное не означает, что в дисперсионном анализе факторами не могут быть количественные переменные - применение дисперсионного анализа к количественным факторам при обработке лесоводственной информации встречается повседневно. Однако заключение о влиянии факторов делают на качественной основе: проверяют гипотезу о влиянии данного фактора при выбранном уровне значимости. Если изучают влияние одного меняющегося фактора, то дисперсионный анализ называют однофакторным, двух – соответственно двухфакторным и т.д.

Для характеристики основных типов моделей, встречающихся в дисперсионном анализе, рассмотрим два реальных примера.

В первом примере допустим, что в вегетационных опытах испытывали влияние двух уровней радиоактивного загрязнения почвы цезием-137 на рост и развитие сеянцев сосны, ели, дуба и березы. Методами дисперсионного анализа требуется оценить влияет ли порода и уровень загрязнения цезием-137 на интенсивность роста.

Во втором случае из совокупности спелых сосновых одновозрастных древостоев Гомельской области отобрано случайным образом 50 насаждений, для которых определены средняя высота и средний коэффициент формы q_2. Требуется установить, влияет ли средняя высота древостоя на величину q_2.

При схожести в общей постановке вопроса ситуация в примерах существенно различается. В первом примере факторы представляют собой фиксированные, постоянные величины. Исходные данные для анализа можно здесь рассматривать как выборку из бесчисленного множества опытов над конкретной породой с данными плотностями радиоактивного загрязнения. Во втором - изучаемый фактор (средняя высота) сам является случайной величиной, а выбранные для исследования насаждения могут рассматриваться как случайная выборка из всех сосновых древостоев не только данной области, но и всей Беларуси.

Аналогично приведенным примерам возможны три основные модели дисперсионного анализа: с фиксированными (постоянными) факторами, со случайными и со смешанными, т.е. когда часть факторов постоянна, а другая - случайная. Общая теория разработана для модели, названной первой. Для двух остальных в сложных задачах не всегда можно найти приемлемую теоретическую схему. Правда, в большинстве практических случаев для всех трех моделей можно применять одинаковые вычислительные схемы и оценки. К тому же во многих задачах, возникающих в лесном хозяйстве, можно ограничиться первой моделью. Пример по оценке влияния высоты на q_2 целесообразно использовать в модели с постоянными факторами: сначала зафиксировать некоторые значения (уровни) высот, а затем по ним выбирать насаждения.

Сущность дисперсионного анализа. Дисперсионный, или вариансный, анализ (analysis of variance) представляет собой в настоящее время самостоятельную и очень важную главу биологической статистики. Сущность его заключается в установлении роли отдельных факторов в изменчивости того или иного признака.

Дело в том, что влияние тех или других факторов на изучаемый признак (или признаки) никогда не может быть выделено в чистом виде. Хотя при проведении опытов и стараются сохранить условия максимально однородными, все же различные опыты дают несколько неодинаковые результаты. Объясняется это тем, что на них влияют многочисленные случайные обстоятельства, и другие факторы, меняющиеся от опыта к опыту и не поддающиеся контролю. Тем более велика роль таких дополнительных неконтролируемых факторов при проведении анализа не в экспериментальных условиях, а при изменениях непосредственно в лесу.

Поэтому возникает задача разложения общей изменчивости признака на составные части: с одной стороны, определяемые изучаемыми конкретными факторами, а с другой — вызываемые случайными, неконтролируемыми причинами. Дисперсионный анализ позволяет оценивать значимость влияния отдельных факторов, а также их относительную роль в общей изменчивости.

Методы дисперсионного анализа были разработаны английским математиком и биологом Р.Фишером и применялись первоначально главным образом для анализа результатов опытов в растениеводстве и в животноводстве. Для различных схем опытов были разработаны соответствующие схемы дисперсионного анализа. Однако в дальнейшем выявилась полная возможность использования дисперсионного анализа как при изучении биологического материала, взятого из природы, так и любых экспериментальных данных, в том числе и в лесном хозяйстве. Поэтому описание дисперсионного анализа приводится во всех учебниках по биометрии. Наиболее полными здесь являются работы К.Е. Никитина и А.З. Швиденко, П.Ф. Рокицкого, Н.Н. Свалова, на которые мы будем опираться в изложении настоящей главы.

Общие предпосылки. Представим себе, что мы анализируем отклонение некоторой случайной величины (или нескольких случайных величин) от средней арифметической. Исследуемым объектом пусть будет популяция деревьев сосны. При этом считаем, что отклонение от среднего значения ($\overline{X}$) в некоторой степени связано с действием на данную величину какого-то определенного фактора, например географического, т.е. его влияние может быть выражено в принадлежности к некоторому типу роста. Тогда

$$x - \overline{X} = A + C \qquad (18.1)$$

где $\overline{X}$ — средняя арифметическая популяции;

x — конкретное значение переменной (варианта);

A—доля отклонения переменной, связанная с влиянием данного конкретного фактора;

C — остаточная часть отклонения, не объяснимая влиянием данного фактора. Это смесь всех неконтролируемых и неопределенных факторов, иначе говоря, результат случайных отклонений.

Очень важно, что в фактическом отклонении варианты (переменной) от средней фигурируют 2 компонента:

а) та часть отклонения, которая зависит именно от данного фактора, т. е. по нашей символике - A

б) остаточная часть, не зависящая от данного фактора - C

В таком случае можно сравнить значения A и C.

При достоверном влиянии изучаемого фактора значение A будет в достаточной степени превышать значение C. По степени превышения A над C можно судить о том, насколько достоверно влияние фактора, A.

Приведенную общую схему, относящуюся к отдельному отклонению, можно перенести на вариацию многих вариант, т. е. выразить степень вариации дисперсий через величину

$$\sigma_0^2 = \sigma_A^2 + \sigma_C^2 \qquad (18.2)$$

т. е. общая дисперсия равна сумме 2 дисперсий, а именно определяемой вариацией фактора А, и дисперсии, определяемой другими, неконтролируемыми (случайными) причинами – С.

Более сложный случай — отклонение переменной х от средней арифметической популяции $\overline{X}$ под влиянием 2 причин: влияния факторов А и В. Например, фактором А для наших деревьев сосны может быть географический район, т. е. влияние местности, а фактором В — класс роста.

Тогда

$$x - \overline{X} = A + B + AB + C \qquad (18.3)$$

Здесь А — доля отклонения, связанная с влиянием фактора А;

В — доля отклонения, связанная с влиянием фактора В;

АВ — доля отклонения, связанная с влиянием не отдельных факторов А и В, а их взаимодействия;

С—остаточная, случайная часть отклонения.

В значениях дисперсий общая дисперсия σ_0^2 может быть представлена как

$$\sigma_0^2 = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_e^2 \qquad (18.4)$$

При большом числе факторов схему можно усложнять и дальше. Так, при 3 факторах

$$x - \overline{X} = A + B + C + AB + BC + AC + ABC + C \qquad (18.5)$$

A, B, C—главные факторы;

AB, BC и AC — взаимодействие первого порядка;

ABC—взаимодействие второго порядка.

Аналогично можно выразить изменчивость вариант в дисперсиях (σ^2) и среднеквадратических отклонениях - σ.

Нетрудно заметить, что сказанное выше непосредственно связано с тем, что изложено в разделе, когда описывали дисперсии.

Градации факторов и их характер. Обычно каждый изучаемый в эксперименте фактор А имеет не одно, а несколько значении, которые называют градациями или уровнями фактора А. В пределах же каждого уровня отдельные переменные (варианты) принимают разные значения, т.е. наблюдается случайная вариация. То же относится и к более сложным случаям, когда в общей изменчивости участвует несколько факторов, каждый из которых может иметь свои уровни. Проводя дисперсионный анализ влияния различ-

ных факторов, следует иметь в виду различный характер уровней факторов. В одних случаях эти уровни фактически точно установлены. Например, изучая влияние сезонов года, выделяют зиму, весну, лето, осень. Внешние условия этих сезонов года строго фиксированы.

С другой стороны, могут быть такие факторы, уровни которых не являются точно фиксированными или которые имеют вообще все возможные случайные градации. Так, например, среди факторов, влияющих на плодоношение дубовых деревьев является возраст насаждения, его полнота, условия погоды (особенно поздние весенние заморозки) и многое другое. Но каждому из этих признаков свойственна своя вариация и достаточно большая Такие факторы называют случайными, понимая под этим только то, что случайными могут быть разные их уровни. Впрочем, надо иметь в виду, что случайные уровни некоторых из них тоже можно сделать фиксированными.

Отсюда следует, что возможны очень разные схемы или модели дисперсионного анализа. Они могут различаться по числу анализируемых факторов (одно-, двух-, трехфакторные и т. д.), по характеру градаций внутри факторов: с фиксированными факторами, со случайными или смешанные схемы.

Есть так называемые иерархические модели, которые широко используются в лесоводстве. В этом случае уровни одного фактора не располагаются случайно среди уровней других факторов, но связаны с ними иерархически. Так, в лесоводстве, изучая леса, мы выделяем роды, виды, подвиды древесных растений. Но в пределах некоторых географических районов деревья даже одного вида могут иметь разный рост в зависимости от количества тепла, осадков и преобладающих почв. Подрбные схемы будут рассмотрены ниже.

При наличии единых общих принципов конкретные методы дисперсионного анализа будут зависеть от того, с какой схемой расположения материала приходится иметь дело.

Таким образом, весь изучаемый материал может быть разбит на ряд групп, различающихся как по отдельным факторам, так и по их градациям. Изучение методами дисперсионного анализа вариации внутри этих групп, между группами и, наконец, вариации всего материала в целом дает возможность установить, влияют ли данные факторы на изменчивость или нет и какие из них имеют больший удельный вес в общей изменчивости.

Нулевая гипотеза. Как и в других случаях статистического анализа, при дисперсионном анализе следует исходить из первоначально принимаемой нулевой гипотезы, а именно: что данный фактор А (или В, или С и т.д.) не влияет. Если правильна нулевая гипотеза, σ_A^2 должна быть равна нулю (то же относится к σ_B^2, σ_C^2 и т.д.), т. е. вся вариация сводится только к случайной.

Для того чтобы отбросить нулевую гипотезу, нужно доказать, что σ_A^2 достоверно (т. е. с вероятностью не меньшей чем 0,95, или с уровнем значимости 0,05) отличается от нуля. Достоверность значения σ_A^2 может быть

установлена, как это обычно делают по отношению к любому статистическому показателю, т. е. путем деления его на его ошибку, т. е. $\sigma_A = \frac{A}{m_A}$, где σ_A показывает уровень достоверности.

Простейшая схема варьирования при различии по одному фактору. Для того чтобы понимать смысл расчетов при дисперсионном анализе, очень важно с самого начала ясно представлять возможную вариацию в тех группах, на которые разбивается фактический материал.

Разберем простейшую схему, когда анализируется влияние только одного фактора, могущего принимать разные градации, или количественные уровни:

1, 2, . . . , i, . . . , a (18.6)

Отдельные наблюдения (варианты) разбиваются на группы согласно этим градациям фактора, изучаемого в опыте или при наблюдениях в природе. Важно, что изучаемый фактор только один, например: вид удобрения, вносимого в питомнике, или принадлежность к разным древесным видам, или влияние радиоактивного загрязнения, или роль способов обработки почвы и т. д. При наличии двух или нескольких факторов потребуются более сложные схемы.

Распределение вариант при различии по одному фактору представлено в таблице 18.1.

Таблица 18.1 – Схема варьирования при различии групп по одному фактору

Группы по одному фактору	Отдельные варианты (наблюдения) x_{ij}							Суммы по группам	Средние по группам $\bar{x}_i$
	1	2	3	..	j	..	n		
1	x_{11}	x_{12}	x_{13}		x_{1j}		x_{1n}	$\sum x_1=T_1$	$\bar{x}_1$
2	x_{21}	x_{22}	x_{23}		x_{2j}		x_{2n}	$\sum x_2=T_2$	$\bar{x}_2$
:									
i	x_{i1}	x_{i2}	x_{i3}		x_{ij}		x_{in}	$\sum x_i=T_i$	$\bar{x}_i$
:									
a	x_{a1}	x_{a2}	x_{a3}		x_{aj}		x_{an}	$\sum x_a=T_a$	$\bar{x}_a$
								$\sum xij=T$	$\bar{x}$

Число наблюдений (вариант) в каждой группе n, но равное число в группах не обязательно. При неравном числе можно исходить из среднего числа n_i.

$$N = an\ (=an_i) \qquad (18.7)$$

Обычно разные уровни принято обозначать буквой i, а отдельные варианты (наблюдения) — буквой j. Поэтому каждую варианту, независимо от того, где она находится, можно обозначать в общем виде как x_{ij}. В пределах каждого уровня (группы) отдельные варианты принимают случайные значения:

$$x_{.i}, x_{.2}, x_{.3}, \ldots, x_{.j}, \ldots, x_{.n}$$

Суммы вариант по каждой группе (в графе «суммы по группам») обозначены буквами T_1, T_2,... , T_i, ..., T_a. В общем виде их можно обозначить T_i. Общая же сумма всех вариант $\sum xij=T$. В последней графе даны средние по группам: $\bar{x}_1$, $\bar{x}_2$, . . ., $\bar{x}_i$, . . . , $\bar{x}_a$. В общем виде групповые средние можно обозначить через $\bar{x}_i$. Общую же среднюю для всех вариант всех групп — через $\bar{x}$.

Разное варьирование вариант и его характеристика. После введения всех этих обозначений можно приступить к разбору варьирования данных, представленных в таблице 18.1.

Можно выделить 3 типа, или направления, варьирования:

а) общее варьирование всех вариант (x_{ij}), независимо от того, в какой группе они находятся, вокруг общей средней $\bar{x}$;

б) варьирование групповых средних $\bar{x}_i$, или, иначе, средних каждого уровня данного изучаемого фактора, вокруг общей средней $\bar{x}$;

в) варьирование вариант x_{ij} внутри каждой группы вокруг каждой групповой средней $\bar{x}_i$.

Для характеристики этих варьирований при проведении дисперсионного анализа используются уже известные из прежних разделов настоящего учебного пособия величины:

а) суммы квадратов отклонений от средней арифметической;

б) средние квадраты отклонений, т.е. суммы квадратов, деленные на количество степеней свободы. Это дисперсии σ^2.

Суммы квадратов. Для всех 3 типов варьирования можно вычислить суммы квадратов. Слово «отклонений» для краткости будем отбрасывать. В общем виде они будут следующими:

1. Общая сумма квадратов

$$\sum_{ij} (x_{ij} - \bar{x})^2 \qquad (18.8)$$

Значок ij около знака суммы обозначает, что суммирование производится по всем вариантам всех групп.

2. Сумма квадратов для групповых средних

$$\sum_{i} n_i (\bar{x}_i - \bar{x})^2 \qquad (18.9)$$

Чтобы эта величина была того же порядка, что и первая, введен множитель n_i, т.е. среднее число вариант в каждой группе. Если число вариант во всех группах одинаково, то просто n.

3. Сумма квадратов отклонений вариант от групповых средних внутри каждой группы, иначе говоря, для случайной вариации внутри групп

$$\sum_i \left[\sum_j (x_{ij} - \bar{x}_i)^2 \right] \qquad (18.10)$$

Два знака сумм указывают, что суммирование производится дважды: внутри каждой группы, т.е. по отдельным j (от 1 до n), а затем по всем уровням i – от 1 до a.

Степени свободы. Чтобы вычислить средние квадраты (дисперсии), надо разделить каждую сумму квадратов на соответствующие им числа степеней свободы, которые будут следующими:

для общей дисперсии

$$df = N - 1 \quad (N = an); \qquad (18.11)$$

для дисперсии групповых средних

$$df = a - 1; \qquad (18.12)$$

для случайной вариации вариант внутри групп

$$df = (n - 1)\, a = na - a = N - a. \qquad (18.13)$$

Нетрудно заметить, что сумма чисел степеней свободы для групповых средних и для вариации внутри групп должна равняться числу степеней свободы для общей дисперсии:

$$(N - a) + (a - 1) = N - 1 \qquad (18.14)$$

Общая схема дисперсионного анализа при одном факторе. Общая схема дисперсионного анализа приведена в таблице 18.2.

Из таблицы 18.2 видно, что общая вариация разлагается на 2 компонента: один из них — это вариация групповых средних (по градациям фактора A) вокруг общей средней $\bar{x}$; другой — вариация отдельных вариант внутри групп. Последнюю вариацию можно рассматривать как случайную в том смысле, что она создается многими неконтролируемыми факторами кроме учитываемого фактора A. При делении сумм квадратов, обозначаемых ss, на число степеней свободы получаются средние квадраты (дисперсии) — ms,

непосредственно измеряющие суммарную вариацию (формула (18.15), и 2 ее компонента (формулы (18.16) и (18.17)).

Таблица 18.2 – Схема дисперсионного анализа (анализа дисперсии) при одном факторе

Источник варьирования	Сумма квадратов ss	Число степеней свободы df	Средний квадрат ms	Номер формулы для ms
Общее (все варианты)	$\sum_{ij}(x_{ij}-\bar{x})^2$	N-1	$\frac{1}{N-1}\sum_{ij}(x_{ij}-\bar{x})^2$	(18.15)
Групповые средние (фактор А)	$\sum_{i} n_i(\bar{x}_i-\bar{x})^2$	a-1	$\frac{1}{a-1}\sum_{i} n_i(\bar{x}_i-\bar{x})^2$	(18.16)
Варианты внутри групп (случайные отклонения)	$\sum_{i}\left[\sum_{j}(x_{ij}-\bar{x}_i)^2\right]$	N-a	$\frac{1}{N-a}\sum_{i}\left[\sum_{j}(x_{ij}-\bar{x}_i)^2\right]$	(18.17)

В дальнейшем мы увидим, что весь этот анализ понадобится для того, чтобы сравнить 2 средних квадрата — второй и третий, пользуясь критерием

$$F=\left(\frac{\sigma_1^2}{\sigma_2^2}\right) \qquad (18.18)$$

Рабочие формулы для вычисления сумм квадратов. Вычисление сумм квадратов отклонений непосредственно по исходным данным вполне возможно, но требует много труда. Его редко используют даже при компьютерной обработке. Поэтому лучше воспользоваться рабочими формулами, основанными на одной из формул для суммы квадратов отклонений, а именно той, где сумма квадратов отклонений вычисляется по значениям вариант:

$$\sum x_i^2-\frac{(\sum x_i)^2}{n} \qquad (18.19)$$

Второй член является как бы поправкой к первому. В литературе он обозначают буквой С, т.е. $C=\left(\sum x_i\right)^2/n$.

Если далее использовать приведенные выше обозначения $\sum x_i$ для каждой группы (уровня фактора А) через T_i (T_1, T_2, . . .,T_i..., T_a), суммы всех вариант—Т, число наблюдений в каждой группе обозначать n_i, общее число вариант —N, то рабочие формулы будут выглядеть довольно просто:

общая сумма квадратов

$$\sum_{ij} x_{ij}^2-\frac{T^2}{N} \qquad (18.20)$$

сумма квадратов для групповых средних

$$\sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N} \qquad (18.21)$$

сумма квадратов для вариант внутри групп, т.е. для случайных отклонений

$$\sum_{ij} x_{ij}^2 - \sum_i \frac{T_i^2}{n_i} \qquad (18.22)$$

Практически совсем не обязательно вычислять все 3 суммы квадратов, достаточно вычислить только 2, например, первую и вторую. Третья может быть получена путем вычитания второй из первой.

При делении сумм квадратов на числа степеней свободы получаются средние квадраты (вариансы). Таким образом, рабочие формулы для них будут следующими:

для общего варьирования

$$ms = \sigma^2 = \frac{1}{N-1}\left(\sum_{ij} x_{ij}^2 - \frac{T^2}{N}\right) \qquad (18.23)$$

ля групповых средних

$$ms = \sigma^2 = \frac{1}{a-1}\left(\sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}\right) \qquad (18.24)$$

для случайных отклонений

$$ms = \sigma^2 = \frac{1}{N-a}\left(\sum_{ij} x_{ij}^2 - \sum_i \frac{T_i^2}{n_i}\right) \qquad (18.25)$$

Ниже на примерах однофакторной и двухфакторной модели рассмотрены основные методы дисперсионного анализа и иллюстрирующие их вычислительные схемы.

18.2 Однофакторный дисперсионный анализ

Для понимания сути однофакторного дисперсионного анализа допустим, что изучается зависимость случайной величины (x) от меняющегося фактора A, градации (или уровни) которого обозначены A_i. Тогда каждое

наблюдение можно обозначить через x_{ij}, где i - уровень фактора A, j -номер наблюдения. Исходные данные удобно представить в виде таблице 18.3.

Таблица 18.3 – Исходные данные для однофакторного дисперсионного анализа

Уровни фактора A_i	Результаты измерений	Среднее по факторам
A_1	$x_{11}\ x_{12}\ .\ .\ .\ x_{1j}\ .\ .\ .x_{1m1}$	$\tilde{X}_1$
A_2	$x_{21}\ x_{22}\ .\ .\ .\ x_{2j}\ .\ .\ .x_{2m1}$	$\tilde{X}_2$
. . .	. . . . .	. . .
A_i	$x_{i1}\ x_{i2}\ .\ .\ .\ x_{ij}\ .\ .\ .x_{im1}$	$\tilde{X}_i$
. . .	. . . . .	. . .
A_k	$x_{k1}\ x_{k2}\ .\ .\ .\ x_{kj}\ .\ .\ .x_{km1}$	$\tilde{X}_k$

В таблице 18.3 k строк (или групп) - по числу уровней фактора A, в каждой группе m_i наблюдений (число наблюдений неодинаково); при равном числе наблюдений подход не меняется, но приводимые ниже формулы несколько упрощаются, так как все m_i равны m. Для подготовки данных к анализу образуем суммы квадратов, где знаком S^2 обозначены дисперсии:

$$s_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{m_i}(x_{ij} - \tilde{X})^2$$ - общую сумму квадратов всех наблюдений от общего среднего $\tilde{X}$;

$$s_M^2 = \sum_{i=1}^{k} m_i(\tilde{X}_i - \tilde{X})^2$$ - сумму квадратов отклонений групповых средних $\tilde{X}$ i от общего среднего $\tilde{X}$, взвешенную через число наблюдений по группам:

$$s_b^2 = \sum_{i=1}^{k}\sum_{j=1}^{m_i}(x_{ij} - \tilde{X})^2$$ - сумму квадратов отклонений внутри групп (от групповых средних).

Простые преобразования позволяют разложить первую сумму на две другие, что аналогично (18.2)

$$s_0^2 = s_M^2 + s_b^2 \qquad (18.2a)$$

Эта уже известная нам формула (18.2) является основой дисперсионного анализа. Рассмотрим оценки дисперсий, связанных с введенными суммами.

Сумма S_0^2 связана с оценкой общей дисперсии изучаемого признака, если ее разделить на число степеней свободы n-1=k$\sum_i$ (m_i -1). По сумме S_M^2 можно оценить дисперсию между уровнями факторов A_i - межгрупповую дисперсию. Число степеней свободы k-1. Наконец, S_b^2 позволяет оценить дисперсию внутри групп (или остаточную). Так как оценка дисперсии каждой из групп связана с m_i - 1 степенью свободы, то общее число степеней свободы k$\sum$ (m_i -1) = N-k, где N — число наблюдений.

Дальнейший анализ зависит от типа рассматриваемой модели. Для модели с фиксированными факторами ответ на основной вопрос дисперсионного анализа сводится к проверке гипотезы $H_0: \overline{X}_1 = \overline{X}_2 = \ldots = \overline{X}_k$, т.е. утверждения, что все групповые средние не зависят от влияния фактора A.. Тогда, если верна H_0, межгрупповая дисперсия (в генеральной совокупности) должна быть равна внутригрупповой, т. е. сформулированная гипотеза может быть заменена эквивалентной $H_0: \sigma_M^2 = \sigma_b^2$. Допустим, что x_{ij} - независимые наблюдения над случайной величиной $\overline{X}$, распределенной нормально со средним $\overline{X}$ и дисперсией σ^2. Тогда отношение

$$F(k, n-k) = \frac{s_M^2 / (k-1)}{s_b^2 / (n-k)} \qquad (18.26)$$

используют в качестве статистической характеристики критерия:
если вычисленное значение F меньше табличного при уровне значимости a, то гипотезу об отсутствии влияния фактора A не отклоняют. Если же факторы случайны, то проверка гипотезы о равенстве групповых средних представляет небольшой интерес (уровни фактора A — сами случайные величины) и проверяют гипотезу о том, что межгрупповая дисперсия в генеральной совокупности равна нулю $H_0: \sigma_M^2 = 0$.

В статистической теории показано, что в качестве статистической характеристики критерия можно применять величину (18.26). В дальнейшем мы не будем обсуждать различия между моделями со случайными и фиксированными факторами. Полная схема однофакторного дисперсионного анализа приведена в таблице 18.4.

Таблица 18.4 – Схема однофакторного анализа

Тип дисперсии	Число степеней свободы	Сумма квадратов	Оценка дисперсии	Гипотеза при факторах		Статистический критерий
				фиксированных	случайных	
Межфакторная (между группами)	k-1	$s_M^2 = \sum_{i=1}^{k} m_i(\tilde{X}_i - \tilde{X})^2$	$\sigma_M^2/(k-1)$	$\overline{X}_1 = \overline{X}_2 = \ldots = \overline{X}_k,$	$\sigma_M^2 = 0$	$\frac{s_M^2/(k-1)}{s_b^2/(n-k)}$
Внутрифакторная (внутри групп)	$k\sum(m_i-1) = n-k$	$s_b^2 = \sum_i \sum_j (x_{ij} - \tilde{X}_i)^2$	$\sigma_b^2/(n-k)$			
Общая	n-1	$s_o^2 = \sum_i \sum_j (x_{ij} - \tilde{X})^2$				

Влияние фактора A можно оценить иным путем. Тогда корреляционное отношение $\eta_{yx}^2 = \sigma_M^2 / \sigma_o^2$ и гипотеза о равенстве групповых средних в генеральной совокупности сводится к виду $H_0 : \eta^2 = 0$ при альтернативной $H_a : \eta^2 > 0$. Величина η^2 связана с F простым соотношением и подчиняется β-распределению со степенями свободы k_1. Есть таблицы критических значений корреляционного отношения η^2 при a= 0,05 (приложение Т); если фактическое значение η^2 больше табличного, то гипотеза о равенстве групповых средних отклоняется. Использование η^2 в качестве оценки влияния иногда более удобно, чем (18.26).

Для расчета мощности F-критерия (если H_0 не отклонена) необходимо предположить, что верна альтернативная гипотеза $H_a : \eta^2 > 0$. Тогда F-критерий подчиняется нецентральному F-распределению, зависящему от числа степеней свободы k_a и k_b и параметра нецентральности

$$\delta^2 = n\eta^2 \qquad (18.27)$$

Обычно при расчете мощности, т.е. при работе с нецентральным F-распределением, пользуются графиками, построенными аналогично графику F-распределения. На рисунке 18.1 в качестве примера приведен график мощности F-критерия при a = 0,05 для числа степеней свободы числителя k_i = 7; при k_1 = 1 можно пользоваться графиком мощности одностороннего t-критерия. Для других k_i графики мощности приведены в специальных изданиях.

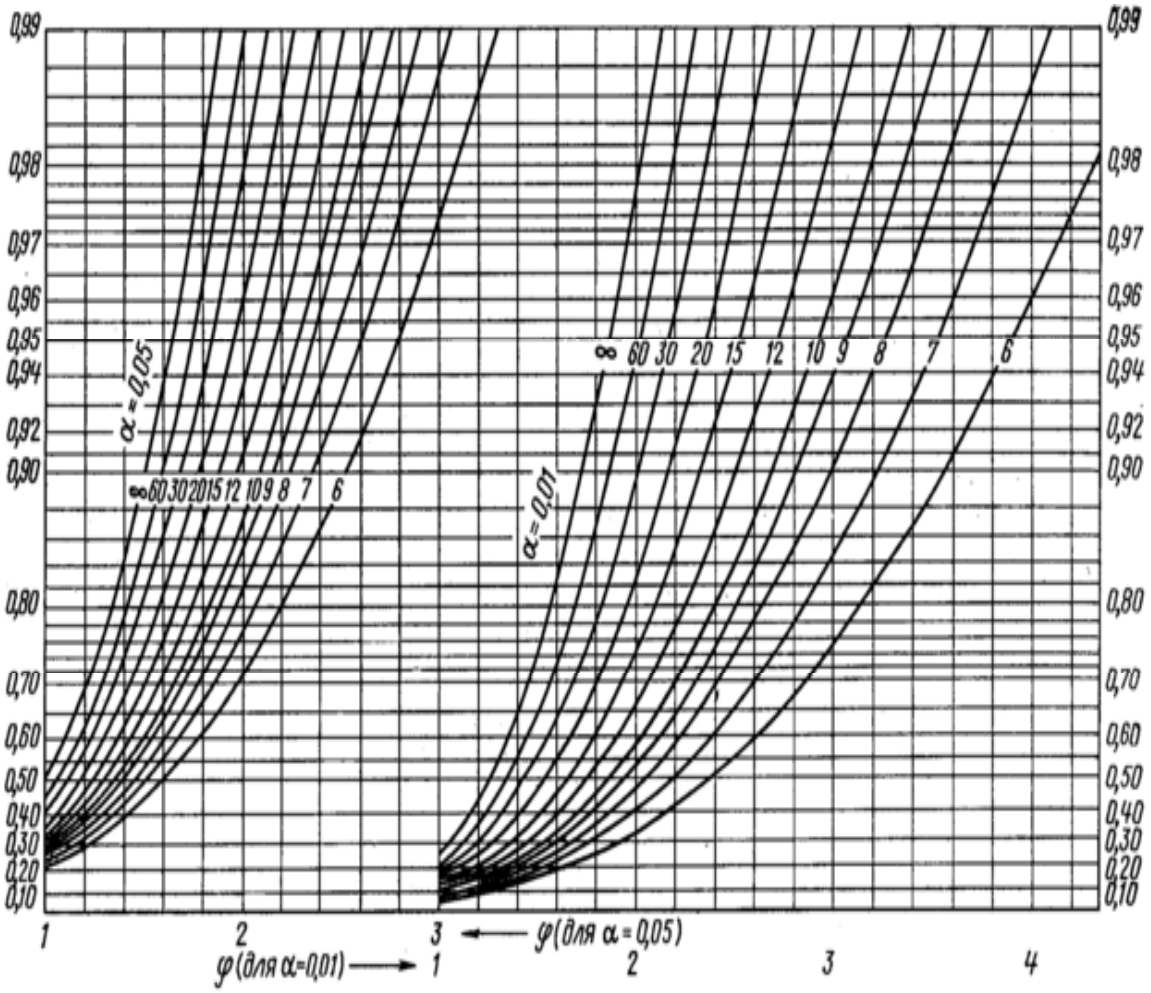


Рисунок 18.1 График функции мощности F-критерия для $k_1=7$ (по К.Е. Никитину и А.З. Швиденко)

В качестве статистической характеристики H_a на рисунке 18.1 использована величина

$$\varphi = \sqrt{n\eta_a^2 / k_1} \qquad (18.28)$$

где η_a^2 — альтернативное корреляционное отношение.

При расчетах мощности F-критерия следует иметь в виду, что альтернативная гипотеза $H_a : \eta^2 > 0$ сложна и для расчетов необходимо выбрать некоторое конкретное значение η_a^2, при котором теснота связи в генеральной совокупности имеет в рамках решаемой задачи практическое значение, т. е. проверить H_o против простой $H_a : \eta^2 = \eta_a^2$. Техника расчета мощности F-критерия рассматривается в табл. 18.4. Заметим, что при помощи (18.28) определяют η^2, которое может быть еще признано существенным при данных $k_м$, k_b, и $1 - \beta$

$$\eta^2 = \varphi^2 k_a / n \ , \qquad (18.29)$$

а также число наблюдений, обеспечивающее признание существенности наличия связи с данным η^2 и при заданных а и β.

Обычно в дисперсионном анализе реальных задач мало отвергнуть гипотезу H_0. Если признается, что фактор влияет на изучаемый признак, то для выяснения соотношения между групповыми средними можно применить t-критерий Стьюдента для попарного сравнения $x_{i.}$.

Определенный интерес представляет метод множественного сравнения, позволяющий оценить сравнения любых сочетаний групповых средних. Допустим, например, что в таблице 18.4 исходные данные представляют результаты измерений для составления некоторых лесотаксационных таблиц и необходимо выбрать целесообразное сочетание уровней A_i, т.е. решить вопрос о составлении двух отдельных таблиц для A_1+A_2 и $A_3+...+A_k$ или трех таблиц $A_1+A_2+A_3$, $A_4+A_5+. . .+A_k$ и т.д. Применяют два метода множественного сравнения: S-метод Шеффе и T-метод Тьюки. Рассмотрим применение первого из них (анализ — однофакторный).

Пусть представляет интерес множественное сравнение

$$\varphi=c_1\bar{x}_1+ c_2\bar{x}_2+ . . . c_k\bar{x}_k, \quad \sum c_i = 0 \qquad (18.30)$$

для которого несмещенной оценкой служит

$$\hat{\varphi}=c_1\tilde{X}_1+ c_2\tilde{X}_2+ . . . c_k\tilde{X}_k, \qquad (18.31)$$

Для статистики $\hat{\varphi}$ оценку дисперсии вычисляют по формуле

$$\hat{\sigma}_\varphi^2 = \frac{s_b^2}{k\sum(m_i - 1)}\sum_i \frac{c_i^2}{m_i^2} \qquad (18.32)$$

$$\hat{\sigma}_\varphi^2 = \frac{s_b^2}{n - k}\sum_i \frac{c_i^2}{m_i^2} \qquad (18.33)$$

где σ_b^2—сумма квадратов внутригрупповых отклонений.

Тогда для сравнения φ доверительный интервал при доверительной вероятности 1 - а имеет вид

$$\hat{\varphi} - s\,\hat{\sigma}_\varphi \le \varphi \le \hat{\varphi} + s\,\hat{\sigma}_\varphi \qquad (18.34)$$

где постоянная s определяется по формуле

$$s^2=(k-1)\,F_a(k-1;\ n-k), \qquad (18.35)$$

где $F_a(k-1; n-k)$ - квантиль F-распределения с числом степеней свободы (k - 1), (n-k).

В качестве примера проведем дисперсионный анализ для определения влияния средней высоты (H) на видовое число (f). Исходные данные для расчета показаны в таблице 18.5.

Таблица 18.5 – Исходные данные для дисперсионного анализа влияния средней высоты на видовое число

H_i, м	Видовые числа, f_{ij} •1000	$\sum f_{ij}$	n_i	$\tilde{f}_i$
1	2	3	4	5
22	455, 436, 466, 467, 446, 483	2753	6	458,8
24	467, 446, 502, 448, 429	2292	5	458,4
26	465, 466, 417, 510, 480	2238	5	467,6
28	502, 489, 442, 530, 467, 501	2931	6	488,5
30	452, 467, 456, 433, 467, 456	2731	6	455,2
32	503, 483, 458, 451, 469	2364	5	472,8
34	446, 427, 430	1303	3	434,3
36	468, 434, 407, 370	1679	4	419,8
		Σ18391	Σ40	$\tilde{f}$=460

Здесь изучается влияние средней высоты древостоя на величину среднего видового числа условно одновозрастных спелых ельников. При расчетах на компьютерах суммы s_0^2, s_M^2 и s_B^2 удобнее вычислять по формулам

$$s_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{m_i} x_{ij}^2 - n\tilde{X}^2 \qquad (18.36)$$

$$s_M^2 = \sum_{i=1}^{k} m_i \tilde{X}_i^2 - n\tilde{X}^2 \qquad (18.37)$$

$$s_B^2 = \sum_{i=1}^{k}\sum_{j=1}^{m_i} x_{ij}^2 - \sum_{i=1}^{k} m_i \tilde{X}_i^2 , \qquad (18.38)$$

а вместо исходных данных использовать их отклонения от некоторого начального значения, например, от общего среднего $\tilde{X}$, что упрощает расчеты. Групповые средние приведены в колонке 5. Число групп k=8, общее число наблюдений n = 40. Общее среднее $\tilde{f}$=460. Перейдем к отклонениям от среднего (таблица 18.6) и вычислим показатели, необходимые для применения формул (18.36 - 18.38).

Из таблицы 18.5 s_0^2 = 36195, s_M^2 = 14587, s_B^2 = 36195 - 14587 = 21608. Результаты вычислений запишем в таблицу 18.6, учитывая, что число степе-

ней свободы для групповой дисперсии равно k-1=8-1=7, для общей N -1=40-1=39, а для внутригрупповой N -k =40-8=32.

Таблица 18.6 – Вычисление сумм квадратов в однофакторном дисперсионном анализе

H_i, м	$f_{ij} - \tilde{f} = x_{ij}$	$\sum x_{ij}$	m_i	$\tilde{X}_i$	$\sum x_{ij}^2$	$\sum m_i \tilde{x}_i^2$
22	-5, -24, +6, +7, -14, +23	7	6	-1,2	1411	8,6
24	+7, -14, +42, -12, -31	-8	5	-1,6	3114	12,8
26	+5, +6, -43, +50, +20	+38	5	+7,6	4810	288,8
28	+42,+29, -18,+70, +7, +41	+171	6	+28,5	9559	4873,5
30	-8, +7, -4, -4, -27, +7	-29	6	-4,8	923	138,2
32	+43, +23, -2, -9, +9	+64	5	+12,8	2544	819,2
34	-14, -33, -30,	-77	3	-25,7	2185	1981,5
36	+8, -26, -53, -90	-161	4	-40,2	11649	6464,2
		-9	40	-0,23	36195	14587

Таблица 18.7 – Итоги однофакторного дисперсионного анализа

Тип дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Межгрупповая	14587	7	2084
Внутригрупповая	21608	32	675
Общая	36195	39	928

Статистическая характеристика ($F_{выч}$), полученная из (18.26), равна $F_{выч}$ = 2084/675 =3,09. При a=0,05 табличное значение F , взятое из приложения Ж, при ν=7 и N -k=32 будет равно 2,3. Так как $F_{выч} > F_{табл.}$, то гипотезу об отсутствии влияния высоты на среднее видовое число древостоя отклоняют: средние $\tilde{f}$ в генеральной совокупности не все равны между собой, а зависят от значения средней высоты.

Для расчета методом множественного сравнения предположим, что необходимо выяснить, для каких значений высот можно составить единые таблицы, использующие средние видовые числа. Испытаем, например, возможность объединения высот 22, 24, 26 в одну группу, остальных — во вторую. Функция сравнения по (18.30)

$$\hat{\varphi} = \frac{1}{3}(458,8+458,4+467,6) - \frac{1}{5}(488,5+455,2+472,8+434,3+419,8) = 7,5$$

а оценка дисперсии по (18.32)

$$\hat{\sigma}_{\varphi} = \left\{675\left[\frac{1}{9}\left(\frac{1}{36}+\frac{1}{25}+\frac{1}{25}\right)+\frac{1}{25}\left(\frac{1}{36}+\frac{1}{36}+\frac{1}{25}+\frac{1}{9}+\frac{1}{16}\right)\right]\right\}^{1/2} = 3,85$$

Постоянную s находим из (18.35)

$s=[(8-1)\ F_{0,05}\ (7,32)]^{1/2} = (7 \cdot 2,3)^{1/2} \approx 4$, т. е. для вероятности 0,95 имеем доверительный интервал $7,5 - 4 \bullet 3,85 \le \varphi \le 7,5+4 \bullet 3,85$ или —7,9 $\le \varphi \le$ 22,9.

Так как доверительный интервал содержит ноль, нет оснований объединять материал в указанные группы в зависимости от значении высоты.

В данном примере изменена постановка задачи: вместо однофакторного применен двухфакторный анализ (включен дополнительно средний диаметр), после чего удалось удовлетворительным образом сгруппировать материал. Далее используем корреляционное отношение и расчет мощности критерия. Проверим H_0 несколько иначе. Вычислим корреляционное отношение η^2, равное отношению межгрупповой дисперсии к сумме квадратов (таблица 18.7): $\hat{\eta}^2 = 14587 / 36195 = 0,403$. Гипотеза $H_0 : \eta^2 = 0$ равносильна $H_0 : \overline{X}_1 = \overline{X}_2 \ldots = \overline{X}_k$. Проверим $H_0 : \eta^2 = 0$ при альтернативной $H_a : \eta^2 > 0$. Из приложения Т находим критические значения $\eta^2 = 0,387$ при $k_1 = 7$, $k_2 = 32$, т.е. гипотезу об отсутствии влияния высоты на видовое число при a =0,05 отклоняем.

18.3 Двухфакторный дисперсионный анализ

Выше уже указывалось, что при участии в общей вариации 2 факторов А и В анализ осложняется наличием взаимодействия между этими факторами. Поэтому общая сумма квадратов при двухфакторной схеме разлагается на 4 компонента: а) вариация под влиянием фактора А; б) вариация под влиянием фактора В; в) вариация под совместным влиянием А и В, т.е. взаимодействия А и В, и г) случайные отклонения.

Кроме того, надо помнить, что при двухфакторной схеме каждый уровень одного фактора должен сочетаться с любым уровнем второго фактора. Так, если изучаются какие-то данные за 3 года о животных из 3 различных местообитаний, то необходимо, чтобы по каждому месту были данные всех трех лет. Если же этого нет, то нужно применять другую схему анализа.

Распределение вариант при варьировании по 2 факторам показано в таблице 18.8. В графах «вар.» помещены варианты, в графах «пок.» — показатели Т и x.

Символом r обозначается количество групп (уровней) по фактору А, т.е. количество горизонтальных рядов (1, 2, 3, ..., i, ..., r}; c —количество групп (уровней) по фактору В, т.е. количество вертикальных столбцов, или колонок {1,2,3,...,j, ..., c); n — число наблюдений в каждой клетке таблицы. В данном случае n равно 3, но не обязательно, чтобы оно было одинаковым во всех клетках. Все же для простоты расчетов выгоднее последнее, тогда nrb = N, т.е. общему числу всех наблюдений.

Так как варьирование групп по фактору А и по фактору В всегда сравнивают со случайными отклонениями вариант в пределах каждой группы как мерилом случайной вариации (σ_e^2), то последняя должна быть измерена на достаточном материале. Это значит, что в каждой клетке надо иметь не менее 2 наблюдений, а еще лучше, если их будет больше.

Каждая варианта (наблюдение) может быть обозначена в общем виде как x_{ijk}, т.е. как k-ое наблюдение в ряду i и в вертикальном столбце j. Конкретная же варианта x имеет 3 значка. Первый обозначает номер группы по фактору А, т.е. номер горизонтальной строчки, второй — номер группы по фактору В, т.е. номер вертикального столбца, третий — номер в данной клетке. В каждой клетке даны сводные показатели: сумма вариант клетки (T..) и средняя арифметическая их ($\bar{x}$..). Значки при них указывают номера горизонтальной строчки и вертикального столбца. В общем виде показатели для каждой клеточки T_{ij} и $\bar{x}_{ij}$.

Показатели для горизонтальных строчек, то есть для градаций фактора А, даны справа в вертикальных столбцах: T_1, T_2, . . ., Ti., . . . , T_r и соответственно $\bar{x}_{1.}$, $\bar{x}_{2.}$ и т.д. В общем виде их будем обозначать $T_{i.}$ и $\bar{x}_{i.}$ или просто T_i и $\bar{x}_i$.

Для вертикальных столбцов (градаций по фактору В) показатели представлены в нижней части таблицы 18.8. Это суммы $T_{.1}$, $T_{.2}$, . . ., $T_{.j}$, . . . , $T_{.c}$ и средние $\bar{x}_{.1}$, $\bar{x}_{.2}$, . . ., $\bar{x}_{.j}$, . . . , $\bar{x}_{.c}$ В общем виде они будут обозначаться как $T_{.j}$, и $\bar{x}_{.j}$ или просто T_j и $\bar{x}_j$.

Общая сумма всех вариант всех клеточек обозначается T, а общая средняя арифметическая — $\bar{x}$.

Вычисление сумм квадратов и средних квадратов. После введения всех этих обозначении можно перейти к построению общих формул сумм квадратов, необходимых для проведения дисперсионного анализа при 2 факторах.

Они будут следующими:

Общая сумма квадратов

$$\sum_{ijk}(x_{ijk} - \bar{x})^2 \qquad (18.39)$$

т. е. простая сумма квадратов отклонений всех наблюдении от общей средней арифметической.

Сумма квадратов для варьирования по фактору А

$$nc\sum_{i}(\bar{x}_i - \bar{x})^2 \qquad (18.40)$$

т. е. помноженная на nc сумма квадратов отклонений всех значений $\bar{x}_i$ от общей средней арифметической.

Таблица 18.8

Схема варьирования при различии групп по 2 факторам

Группы (уровни) по фактору A	Группы (уровни) по фактору B и отдельные наблюдения x_{ijk} внутри них												Сумма по группам фактора A T_i	Средние по группам фактора A $\bar{x}_i$
	1		2		3		...	j		...	c			
	вар.	пок.	вар.	пок.	вар.	пок.	...	вар.	пок.	...	вар.	пок.		
1	x_{111} x_{112} x_{113}	T_{11} $\bar{x}_{11}$	x_{121} x_{122} x_{123}	T_{12} $\bar{x}_{12}$	x_{131} x_{132} x_{133}	T_{13} $\bar{x}_{13}$		x_{1j1} x_{1j2} x_{1j3}	T_{1j} $\bar{x}_{1j}$		x_{1c1} x_{1c2} x_{1c3}	T_{1c} $\bar{x}_{1c}$	$T_{1\cdot}$	$\bar{x}_{1\cdot}$
2	x_{211} x_{212} x_{213}	T_{21} $\bar{x}_{21}$	x_{221} x_{222} x_{223}	T_{22} $\bar{x}_{22}$	x_{231} x_{232} x_{233}	T_{23} $\bar{x}_{23}$		x_{2j1} x_{2j2} x_{2j3}	T_{2j} $\bar{x}_{2j}$		x_{2c1} x_{2c2} x_{2c3}	T_{2c} $\bar{x}_{2c}$	$T_{2\cdot}$	$\bar{x}_{2\cdot}$
3	x_{311} x_{312} x_{313}	T_{31} $\bar{x}_{31}$	x_{321} x_{322} x_{323}	T_{32} $\bar{x}_{32}$	x_{331} x_{332} x_{333}	T_{33} $\bar{x}_{33}$		x_{3j1} x_{3j2} x_{3j3}	T_{3j} $\bar{x}_{3j}$		x_{3c1} x_{3c2} x_{3c3}	T_{3c} $\bar{x}_{3c}$	$T_{3\cdot}$	$\bar{x}_{3\cdot}$
...														
i	x_{i11} x_{i12} x_{i13}	T_{i1} $\bar{x}_{i1}$	x_{i21} x_{i22} x_{i23}	T_{i2} $\bar{x}_{i2}$	x_{i31} x_{i32} x_{i33}	T_{i3} $\bar{x}_{i3}$		x_{ij1} x_{ij2} x_{ij3}	T_{ij} $\bar{x}_{ij}$		x_{ic1} x_{ic2} x_{ic3}	T_{ic} $\bar{x}_{ic}$	$T_{i\cdot}$	$\bar{x}_{i\cdot}$
...														
r	x_{r11} x_{r12} x_{r13}	T_{r1} $\bar{x}_{r1}$	x_{r21} x_{r22} x_{r23}	T_{r2} $\bar{x}_{r2}$	x_{r31} x_{r32} x_{r33}	T_{r3} $\bar{x}_{r3}$		x_{rj1} x_{rj2} x_{rj3}	T_{rj} $\bar{x}_{rj}$		x_{rc1} x_{rc2} x_{rc3}	T_{rc} $\bar{x}_{rc}$	$T_{r\cdot}$	$\bar{x}_{r\cdot}$
Суммы по группам фактора B T_j		$T_{\cdot 1}$		$T_{\cdot 2}$		$T_{\cdot 3}$			$T_{\cdot j}$			$T_{\cdot c}$	T	—
Средние по группам фактора B $\bar{x}_j$		$\bar{x}_{\cdot 1}$		$\bar{x}_{\cdot 2}$		$\bar{x}_{\cdot 3}$			$\bar{x}_{\cdot j}$			$\bar{x}_{\cdot c}$	—	$\bar{x}$

Сумма квадратов для варьирования по фактору B

$$nr\sum_{j}(\bar{x}_j-\bar{x})^2 \qquad (18.41)$$

т. е. помноженная на nr сумма квадратов отклонений всех значений $\bar{x}_j$ от общей средней арифметической.

Сумма квадратов для взаимодействия A и B

$$n\sum_{ij}\left[\bar{x}_{ij}-(\bar{x}_i-\bar{x})-(\bar{x}_j-\bar{x})-\bar{x}\right]^2=n\sum_{ij}\left[\bar{x}_{ij}-\bar{x}_i-\bar{x}_j+\bar{x}\right]^2 \qquad (18.42)$$

Наконец, сумма квадратов для случайных отклонений

$$\sum_{ijk}(x_{ijk}-\bar{x}_{ij})^2 \qquad (18.43)$$

т. е. сумма квадратов отклонений вариант от средних отдельных клеток таблицы.

Числа степеней свободы df таковы: для общего варьирования rcn-1, для варьирования по фактору A r-1, для варьирования по фактору B c-1, для взаимодействия A и B (r-1)(c-1), для случайных отклонений rc(n-1).

Общая схема разложения вариации при двухфакторной схеме дисперсионного анализа представлена в табл. 18.8. В ней же даны и общие формулы для средних факторов, которые получаются путем деления сумм квадратов на число степеней свободы.

Рабочие формулы при двухфакторном анализе. Для упрощения расчетов лучше применить рабочие формулы для сумм квадратов, а именно:

для общего варьирования

$$\sum_{ijk}x_{ijk}^2-\frac{T^2}{nrc} \qquad (18.44)$$

для варьирования по фактору A

$$\frac{1}{nc}\sum_{i}T_i^2-\frac{T^2}{nrc} \qquad (18.45)$$

для варьирования по фактору B

$$\frac{1}{nr}\sum_{j}T_j^2-\frac{T^2}{nrc} \qquad (18.46)$$

для варьирования, характеризующего взаимодействие A и B,

$$\frac{1}{n}\sum_{ij} T_{ij}^2 - \frac{1}{nc}\sum T_i^2 - \frac{1}{nr}\sum T_j^2 + \frac{T^2}{nrc} \qquad (18.47)$$

для варьирования случайных отклонений (внутри всех групп)

$$\sum_{ijk} x_{ijk}^2 - \frac{1}{n}\sum_{ij} T_{ij}^2 \qquad (18.48)$$

В этих формулах n — число вариант в каждой клеточке;
c — число вертикальных столбцов, т.е. групп по фактору B;
r — число горизонтальных строчек, т.е. групп по фактору A.

Величина $\sum x_{ijk}^2$ означает уже фигурировавшую ранее сумму квадратов всех вариант. Далее приходится иметь дело с различными суммами вариант:

T_{ij} — сумма вариант по отдельным клеткам (как рядов, так и столбцов);

T_i — сумма вариант для i-рядов, т.е. рядов по уровням (группам) фактора A;

T_j — сумма вариант для j-столбцов, т.е. колонок по уровням (группам) фактора B;

T — общая сумма всех вариант.

Применение этих формул для сумм квадратов дает возможность пользоваться не средними, имеющимися в таблице 18.8, а только суммами вариант, кроме только того, что понадобится сумма квадратов всех вариант $\sum x_{ijk}^2$. Поэтому в схеме варьирования таблицы 18.8 можно не записывать средних в отдельных клетках.

Коэффициенты при отдельных суммах ($\frac{1}{nc}, \frac{1}{nr}$ и т.д.) служат для приведения всех величин к одному порядку. Число степеней свободы для всех сумм квадратов приведено в таблице 18.9.

Таблица 18.9 – Схема дисперсионного анализа при 2 факторах

Источник варьирования	Сумма квадратов ss	Число степеней свободы df	Средний квадрат ms	
Общее	$\sum_{ijk}(\bar{x}_{ijk}-\bar{x})^2$	rcn-1	$\frac{1}{rcn-1}\sum_{ijk}(\bar{x}_{ijk}-\bar{x})^2$	18.49
Фактор А (групповые средние по фактору А)	$nc\sum_{i}(\bar{x}_i-\bar{x})^2$	r-1	$\frac{nc}{r-1}\sum_{i}(\bar{x}_i-\bar{x})^2$	18.50
Фактор В (групповые средние по фактору В)	$nr\sum_{j}(\bar{x}_j-\bar{x})^2$	c-1	$\frac{nr}{c-1}\sum_{j}(\bar{x}_j-\bar{x})^2$	18.51
Взаимодействие А и В	$n\sum_{ij}(\bar{x}_{ij}-\bar{x}_i-\bar{x}_j+\bar{x})^2$	(r-1)(c-1)	$\frac{n}{(r-1)(c-1)}\sum_{ij}(\bar{x}_{ij}-\bar{x}_i-\bar{x}_j+\bar{x})^2$	18.52
Случайные отклонения	$\sum_{ijk}(\bar{x}_{ijk}-\bar{x}_{ij})^2$	rc(n-1)	$\frac{1}{rc(n-1)}\bullet\sum_{ijk}(\bar{x}_{ijk}-\bar{x}_{ij})^2$	18.53

Поэтому можно записать следующие рабочие формулы для средних квадратов, получающиеся путем деления сумм квадратов на соответствующие числа степеней свободы:

для общего варьирования

$$ms=\frac{1}{rcn-1}\left(\sum_{ijk}x_{ijk}^2-\frac{T^2}{nrc}\right) \qquad (18.54)$$

для варьирования по фактору А

$$ms=\frac{1}{r-1}\left(\frac{1}{nc}\sum_{i}T_i^2-\frac{T^2}{nrc}\right) \qquad (18.55)$$

для варьирования по фактору В

$$ms=\frac{1}{c-1}\left(\frac{1}{nr}\sum_{j}T_j^2-\frac{T^2}{nrc}\right) \qquad (18.56)$$

для варьирования А и В

$$ms=\frac{1}{(r-1)(c-1)}\left(\frac{1}{n}\sum_{ij}T_{ij}^2-\frac{1}{nc}\sum T_i^2-\frac{1}{nr}\sum T_j^2+\frac{T^2}{nrc}\right) \quad (18.57)$$

для случайных отклонений

$$ms=\frac{1}{rc(n-1)}\left(\sum_{ijk}x_{ijk}^2-\frac{1}{n}\sum_{ij}T_{ij}^2\right) \quad (18.58)$$

Примеры дисперсионного анализа при двухфакторной схеме.

Приведем пример, описанный П.Ф. Рокицким. Допустим, что проводились опыты по удобрению карповых прудов известью (600 кг/га негашеной извести), суперфосфатом (72,8 кг/га P_2O_5) и известью и суперфосфатом вместе (с трехкратной повторностью). Четвертый пруд в каждом блоке не удобрялся. Окончательные данные о продуктивности прудов (в переводе на 300 рыб в каждом пруду) представлены в таблице 18.10.

Таблица 18.10 – Продуктивность карповых прудов с применением удобрений

Группы по фактору А (кальциевые удобрения)	Группы по фактору В (фосфорные удобрения)				T_i	$\bar{x}_i$
	О		Р			
	варианты x_{ijk}	T_{ij}	варианты x_{ijk}	T_{ij}		
О	58 84 39	T_{11}=181	72 72 64	T_{12}=208	$T_{1.}$=389	$\bar{x}_{1.}$=64,83
Са	49 55 48	T_{21}=152	74 74 85	T_{22}=233	$T_{2.}$=385	$\bar{x}_{2.}$=64,17
T_j		$T_{.1}$=333		$T_{.2}$=441	T=774	
$\bar{x}_j$		$\bar{x}_{.1}$=55,5		$\bar{x}_{.2}$=73,5		$\bar{x}$=64,5

$$\sum_{ijk}x_{ijk}^2=52312$$

Таким образом, n=3, r=2, c=2, N=12. Применение рабочих формул позволит вычислить значения сумм квадратов.

Общая сумма квадратов

$$\sum_{ijk} x_{ijk}^2 - \frac{T^2}{rnc} = 58^2 + 84^2 + \cdots 74^2 + 85^2 - \frac{599076}{12} =$$

= 52312 - 49923 = 2389.

Сумма квадратов для варьирования по фактору А (кальций)

$$\frac{1}{nc}\sum_i T_i^2 - \frac{T^2}{nrc} = \frac{1}{6}(389^2+385^2)\text{-}49923=49924\text{-}49923=1$$

Сумма квадратов для варьирования по фактору В (фосфор)

$$\frac{1}{nr}\sum_j T_j^2 - \frac{T^2}{nrc} = \frac{1}{6}(333^2+441^2)\text{-}49923 = 50895\text{-}49923 = 972.$$

Сумма квадратов для взаимодействия А и В,

$$\frac{1}{n}\sum_{ij} T_{ij}^2 - \frac{1}{nc}\sum T_i^2 - \frac{1}{nr}\sum T_j^2 + \frac{T^2}{nrc} = \frac{1}{3}(181^2+208^2+152^2+233^2)\text{-}$$

- 49924-50895+49923=243.

Сумма квадратов для случайных отклонений

$$\sum_{ijk} x_{ijk}^2 - \frac{1}{n}\sum_{ij} T_{ij}^2 = 52312\text{-}51139 = 1173.$$

Сводка результатов дисперсионного анализа дана в таблице 18.11

Таблица 18.11 – Дисперсионный анализ данных о влиянии удобрений Ca, P и Ca+P на продуктивность карповых прудов

Источник варьирования	ss	df	ms	F фактическое	F табличное	
					при P=0,05	при P=0,01
Общее	2389	11	-	-	-	-
Ca	1	1	1	1/147=0,007	-	-
P	972	1	972	972/147=6,6	5,32	11,26
Ca+P	243	1	243	243/147=1,7	5,32	11,26
Случайные отклонения	1173	8	147	-	-	-

С помощью критерия F проверяется достоверность средних квадратов для источников варьирования: Ca, P и Ca+P. Роль Ca оцениваем как F=1/147 = 0,007, т.е. роль Ca не доказана.

Для влияния P F = 972/147 = 6,6. Табличные значения F, взятые из приложения Ж, при df = 1 и df=8: для P=0,05-5,32 и для P = 0,01-11,26. Таким образом, эффект фосфора можно принять доказанным (нулевая гипотеза отвергается). Все же полученное значение F удовлетворяет только уров-

ню значимости P=0,05. При более жестких требованиях следовало бы воздержаться от окончательного вывода до проведения новых, более полных исследований. Дело в том, что опыты были поставлены только на трех повторностях, поэтому число степеней свободы для случайных колебаний (df=8) ниже минимально допустимого числа 10, о чем говорилось выше. F для взаимодействия Ca+ очень мало (= 1,7), поэтому в данном случае влияние взаимодействия не доказано. Нулевая гипотеза остается в силе.

Оцениваемые параметры при двухфакторном дисперсионном анализе. При двухфакторном дисперсионном анализе, как и при анализе по одному фактору, значения средних квадратов оценивают определенные параметры вариации: при фиксированных уровнях факторов - условные, которые можно обозначит буквой χ^2, и при случайных уровнях - отражающие действительную случайную вариацию и поэтому обозначающиеся обычными σ^2.

Оцениваемые параметры для второго случая при двухфакторной схеме будут следующими (средние квадраты отмечены только номерами):

Источник варьирования	ms	Оцениваемые параметры
Фактор A	1	$\sigma_e^2+n\sigma_{AB}^2+nc\sigma_A^2$
Фактор B	2	$\sigma_e^2+n\sigma_{AB}^2+nr\sigma_B^2$
Взаимодействие A и B	3	$\sigma_e^2+n\sigma_{AB}^2$
Случайные отклонения	4	σ_e^2

Если сравнить оцениваемые параметры для 4 источников варьирования, то можно сделать вывод о возможности оценки влияния факторов A и B путем деления их средних квадратов не на средний квадрат случайных отклонении (ms_4), а на средний квадрат взаимодействия (ms_3). Этот способ получения F для A и B имеет смысл только при достоверном наличии взаимодействия, ибо в этом случае параметры, оцениваемые ms_1 и ms_3, отличаются на $nc\sigma_A^2$, а оцениваемые ms_2 и ms_3 — на $nr\sigma_B^2$. Однако некоторые авторы считают более правильным во всех случаях брать знаменателем для F ms_4, т.е. σ_e^2. Для доказательства же влияния взаимодействия, очевидно, возможен только один способ — деление ms_3 на ms_4.

Определение точных величин σ_A^2, σ_B^2 и σ_{AB}^2, совершенно необходимое в ряде селекционно-генетических исследований, может быть сделано путем ряда последовательных вычитаний средних квадратов и делений разностей на коэффициенты согласно формулам оцениваемых параметров.

Так,

$$\sigma_A^2 = \frac{ms_1 - ms_3}{nc}; \qquad \sigma_B^2 = \frac{ms_2 - ms_3}{nr}; \qquad \sigma_{AB}^2 = \frac{ms_3 - ms_4}{n}$$

18.4 Многофакторный дисперсионный анализ. Использование дисперсионного анализа в лесном хозяйстве

Дисперсионный анализ при трехфакторной схеме. При структуре материала, различающегося по 3 факторам, применяется принципиально та же схема анализа, что и при различиях по 2 факторам, но она более сложна и поэтому требует большого внимания при расчетах.

Общая сумма квадратов разлагается на 8 компонентов:
1. Эффект фактора A.
2. Эффект фактора B.
3. Эффект фактора C.
4. Взаимодействие A и B.
5. Взаимодействие A и C.
6. Взаимодействие B и C.
7. Взаимодействие A, B и C вместе (взаимодействие второго порядка).
8. Случайные отклонения.

Каждая отдельная варианта обозначается 4 значками, а именно x_{ijkl}. Соответствующие средние: $\bar{x}$ - средняя арифметическая всех наблюдений; $\bar{x}_i$, $\bar{x}_j$ и $\bar{x}_k$ - средние для уровней по фактору A, по фактору B и по фактору C отдельно: $\bar{x}_{ij}$, $\bar{x}_{ik}$ и $\bar{x}_{jk}$ - средние для всех уровней по 2 факторам без учета третьего; $\bar{x}_{ijk}$ - средние всех клеток решетки.

Чтобы не спутать буквы, можно обозначить число групп по факторам A, B и C одной буквой r со значками 1, 2, 3. Тогда общая схема анализа может быть представлена в таблице 18.12.

Средний квадрат, как обычно, получают делением суммы квадратов на число степеней свободы, поэтому для экономии места его можно не включать в таблицу. Общая схема анализа в сущности та же, которая была изложена выше для дисперсионного анализа по 2 факторам. В частности, таков же расчет сумм квадратов и степеней свободы для взаимодействия по двум факторам. Наряду с учетом взаимодействия A и B добавляется учет взаимодействий A и C и B и C. Новым является учет взаимодействия всех 3 факторов A, B и C.

Пользование квадратами отклонений различных средних от общей средней в случае анализа по 3 факторам еще более осложнило бы технику расчетов, поэтому и здесь для подсчета сумм квадратов целесообразно пользоваться рабочими формулами, в которых фигурируют квадраты вариант и суммы вариант по группам.

Таблица 18.12 – Схема дисперсионного анализа при 3 факторах

Источник вариации	ss	df
Общее	$\sum_{ijkl} (x_{ijkl}-\bar{x})^2$	$nr_1r_2r_3-1$
Фактор A	$nr_2r_3 \sum_{i} (\bar{x}_i-\bar{x})^2$	r_1-1
Фактор B	$nr_1r_3 \sum_{j} (\bar{x}_j-\bar{x})^2$	r_2-1
Фактор C	$nr_1r_2 \sum_{k} (\bar{x}_k-\bar{x})^2$	r_3-1
Взаимодействие A и B	$nr_3 \sum_{ij} (\bar{x}_{ij}-\bar{x}_i-\bar{x}_j+\bar{x})^2$	$(r_1-1)(r_2-1)$
Взаимодействие B и C	$nr_1 \sum_{jk} (\bar{x}_{jk}-\bar{x}_j-\bar{x}_k+\bar{x})^2$	$(r_2-1)(r_3-1)$
Взаимодействие A и C	$nr_2 \sum_{ik} (\bar{x}_{ik}-\bar{x}_i-\bar{x}_k+\bar{x})^2$	$(r_1-1)(r_3-1)$
Взаимодействие A,B и C	$n \sum_{k} (\bar{x}_{ijk}-\bar{x}_{ij}-\bar{x}_{ik}-\bar{x}_{jk}+\bar{x}_i+\bar{x}_j+x_k-\bar{x})^2$	$(r_1-1)(r_2-1)(r_3-1)$
Случайные отклонения	$\sum_{ijkl} (x_{ijkl}-\bar{x}_{ijk})^2$	$r_1r_2r_3(n-1)$

Они будут следующими:
общая изменчивость

$$\sum_{ijkl} x_{ijkl}^2 - \frac{T^2}{nr_1r_2r_3}$$

эффект A

$$\frac{1}{nr_2r_3}\sum_{i} T_i^2 - \frac{T^2}{nr_1r_2r_3}$$

эффект B

$$\frac{1}{nr_1r_3}\sum_{j} T_j^2 - \frac{T^2}{nr_1r_2r_3}$$

эффект C

$$\frac{1}{nr_1r_2}\sum_{k} T_k^2 - \frac{T^2}{nr_1r_2r_3}$$

взаимодействие A и B

$$\frac{1}{nr_3}\sum_{ij} T_{ij}^2 - \frac{1}{nr_2r_3}\sum_{i} T_i^2 - \frac{1}{nr_1r_3}\sum_{i} T_j^2 + \frac{T^2}{nr_1r_2r_3}$$

взаимодействие B и C

$$\frac{1}{nr_1}\sum_{jk}T_{jk}^2-\frac{1}{nr_1r_3}\sum_{j}T_j^2-\frac{1}{nr_1r_2}\sum_{k}T_k^2+\frac{T^2}{nr_1r_2r_3}$$

взаимодействие A и C

$$\frac{1}{nr_2}\sum_{ik}T_{ik}^2-\frac{1}{nr_2r_3}\sum_{i}T_i^2-\frac{1}{nr_1r_2}\sum_{k}T_k^2+\frac{T^2}{nr_1r_2r_3}$$

взаимодействие A, B и C

$$\frac{1}{n}\sum_{ikj}T_{ijk}^2-\frac{1}{nr_3}\sum_{ij}T_{ij}^2-\frac{1}{nr_1}\sum_{jk}T_{jk}^2-\frac{1}{nr_2}\sum_{ik}T_{ik}^2+$$
$$+\frac{1}{nr_2r_3}\sum_{i}T_i^2+\frac{1}{nr_1r_3}\sum_{j}T_j^2-\frac{1}{nr_1r_2}\sum_{k}T_k^2-\frac{T^2}{nr_1r_2r_3}$$

случайные отклонения

$$\sum_{ijkl}x_{ijkl}^2-\frac{1}{n}\sum_{ijk}T_{ijk}^2$$

Во всех этих формулах поправка одна и та же - $\frac{T^2}{nr_1r_2r_3}$, т.е. квадрат суммы всех вариант, деленный на общее их количество. При вычислении первой части рабочих формул важно не спутать, какие конкретно суммы надо возводить в квадрат. Чтобы не загромождать текст, ограничимся только этими формулами для сумм квадратов в буквенной символике без окончательных формул для средних квадратов и без приведения конкретных примеров. Оцениваемые средними квадратами параметры при 3 факторах принципиально не отличаются от указанных выше для случая двухфакторного дисперсионного анализа. Они приведены в таблице 18.13. для случая, когда уровни по всем факторам будут случайными. Если же различия между уровнями по одному из факторов являются не случайными, а фиксированными (смешанная схема), соответствующий компонент вариации надо обозначать не σ^2, а каким-либо иным значком, например x^2, как указывалось при разборе однофакторной схемы, или просто K со значком, обозначающим данный фактор A, B, C и т. д.

Таблица 18.13 – Оцениваемые параметры при трехфакторном дисперсионном анализе

Источник варьирования	ms	Оцениваемые параметры
Фактор A	1	$\sigma_e^2 + n\sigma_{ABC}^2 + nr_3\sigma_{AB}^2 + nr_2\sigma_{AC}^2 + nr_2r_3\sigma_A^2$
Фактор B	2	$\sigma_e^2 + n\sigma_{ABC}^2 + nr_3\sigma_{AB}^2 + nr_1\sigma_{BC}^2 + nr_1r_3\sigma_B^2$
Фактор C	3	$\sigma_e^2 + n\sigma_{ABC}^2 + nr_2\sigma_{AC}^2 + nr_1\sigma_{BC}^2 + nr_1r_2\sigma_C^2$
Взаимодействие A и B	4	$\sigma_e^2 + n\sigma_{ABC}^2 + nr_3\sigma_{AB}^2$
Взаимодействие B и C	5	$\sigma_e^2 + n\sigma_{ABC}^2 + nr_1\sigma_{BC}^2$
Взаимодействие A и C	6	$\sigma_e^2 + n\sigma_{ABC}^2 + nr_2\sigma_{AC}^2$
Взаимодействие A, B и C	7	$\sigma_e^2 + n\sigma_{ABC}^2$
Случайные отклонения	8	σ_e^2

Разбор параметров, оцениваемых различными средними квадратами, показывает, что в данном случае оценить с помощью критерия F достоверность влияния отдельных факторов и их взаимодействия значительно сложнее, чем в предыдущих случаях дисперсионного анализа. Поэтому мы ограничимся сказанным**,** отослав читателя, которому понадобятся эти методы, к специальной литературе.

Иерархическая схема дисперсионного анализа. Все предыдущие схемы были факторными. В них предусматривалось, что уровни одного фактора сочетаются с любыми уровнями всех других факторов. Таким образом создаются группы вариант, на которые действуют любые сочетания всех изучаемых факторов. Обычные факторные схемы чаще всего применяются в опытах, план которых строится экспериментатором заранее. Очевидно, что такой план должен предусматривать наличие всех сочетаний градаций разных факторов (или почти всех, что иногда возможно при так называемых «выпавших» группах опыта).

Однако при анализе материала, взятого из природы или из хозяйства, обычные факторные схемы могут быть неосуществимыми, так как внутри градаций (групп) фактора A возможны различные, отличающиеся друг от друга градации (группы) факторов B, C и т.д.

Так, например, при изучении данных об удоях коров-дочерей, происходящих от разных родителей и относящихся к разным породам, обнаружится определенная связь между группами и влияющими на них факторами (рисунок 18.2).

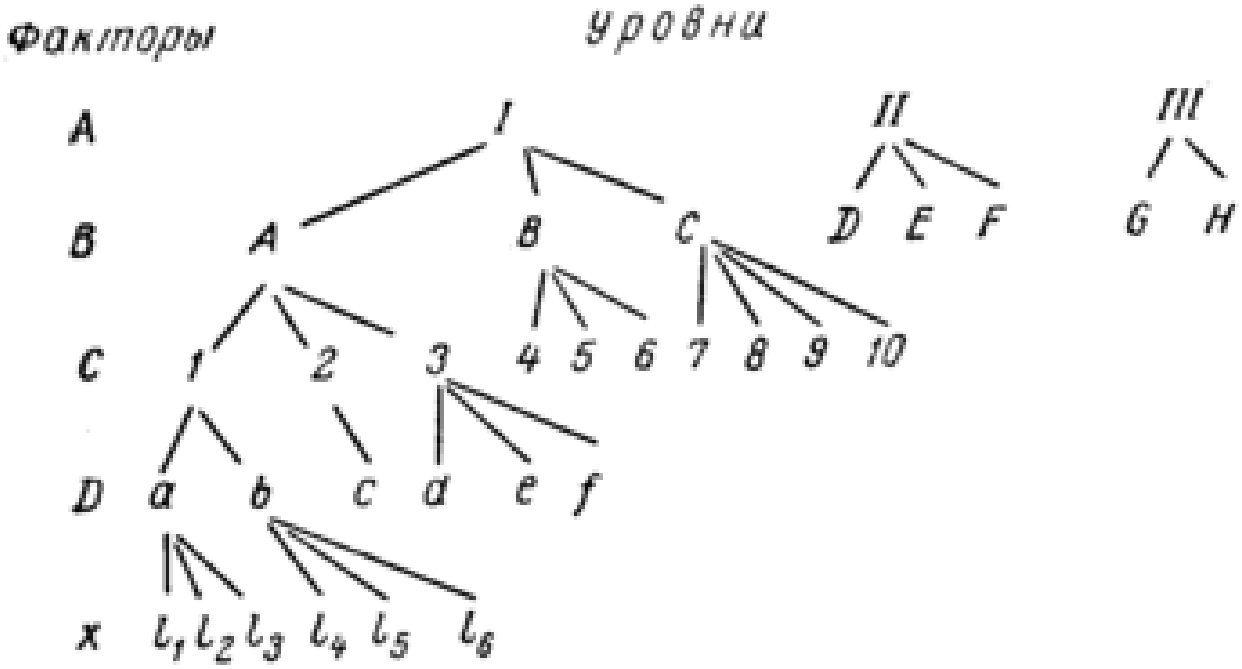


Рисунок 18.2 Схема иерархических связей между факторами и их уровнями (по П.Ф. Рокицкому). Уровни низшего порядка располагаются только внутри определенных уровней высшего порядка. Факторы: A — породы; B — быки; C — покрытые ими коровы; D — дочери; x — варианты, т.е. удои коров-дочерей по отдельным лактациям.

К породе I относятся только быки A, B, C. Остальные быки других пород (II и III). Бык A покрыл коров 1, 2, 3; бык B — коров 4, 5 и 6; бык C — опять иных коров 7, 8, 9, 10 и т.д. Корова 1 дала дочерей a и b; корова 2— дочь c; корова 3 — дочерей d, e и f и т.д. Наконец, от каждой дочери было изучено по нескольку лактаций l_1, l_2 и т.д.

Варьирующие по отдельным лактациям удои коров зависят от 4 факторов: породы, быки, матери, дочери, но связь между ними осуществляется по иерархической лестнице — от более общих факторов к более частным, или от факторов высшего порядка к факторам низшего порядка. Поэтому такие схемы, или модели, получили название иерархических.

Подобным же образом может быть сгруппирован зоологический или ботанический материал по факторам: виды, подвиды, экотипы, места обитания, выборки, отдельные экземпляры. Для иерархических схем характерно отсутствие свободных сочетаний между градациями факторов A, B, C и т.д., в этом их отличие от рассмотренных выше факторных схем. Так, коровы-дочери могут быть только от определенных матерей, а не от любых коров популяции. Одни коровы-матери покрыты одними быками, а другие — другими. Определенные экотипы входят в состав одних подвидов, другие—в состав других и т.д.

Иерархические лестницы могут быть короче или длиннее в зависимости от количества учитываемых факторов.

Применение дисперсионного анализа в лесном хозяйстве. В лесном хозяйстве (в исследованиях) дисперсионный анализ применяется широко. Он является основой при доказательстве сходства и различий у потомства при проведении работ по селекции лесных древесных видов. Им пользуются для определения эффективности проводимых мероприятий в лесном хозяйстве, в частности, при применении удобрений, гербицидов, даже при проведении рубок ухода, хотя в последнем случае чаще используют регрессионный анализ. Вычисления в настоящее время проводят только на компьютерах, по сертифицированным стандартным программам, которые имеются во всех системах матобеспечения для ПК.

ЗАКЛЮЧЕНИЕ

Обобщая вышеизложенное, приходим к выводу, что настоящее учебное пособие позволяет студентам в полной мере освоить курс биометрии, который читается в унмверситете по специальности «Лесное хозяйство».

В то же время пособие обеспечивает потребности магистрантов, аспирантов и научных работников в понимании сути исследуемых ими явлений при анализе различных биометрических показателей, которые вычисляют по стандартным компьютерным программам, входящим в состав сертифицированного метематического обеспечения для ПК.

В процессе пользования настоящим пособием у читателей могут появиться различные замечания и пожелания. Дальнейшее совершенствование всех изданий подобного рода должно идти по пути учета этих предложений. Поэтому через некоторое время – 5-7 лет – пособие должно переиздаваться с учетом новых материалов в области математической статистики и пожеланий пользователей.

При чтении курса лесной биометрии преподаватель должен акцентировать внимание студентов на те моменты, которые обязательны к усвоению и на вопросы, желательные к расширенному (факультативному) изучению. Это раскрывает творческие возможности преподавателей биометрии и студентов и отвечает современным требованиям к методике чтения лекций.