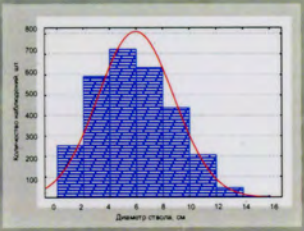
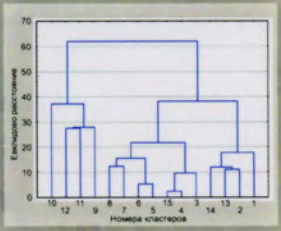
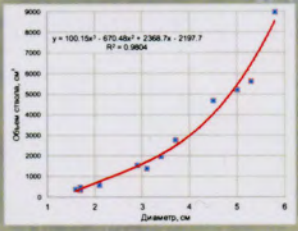
А.С. Бондаренко
А.В. Жигунов

СТАТИСТИЧЕСКАЯ ОБРАБОТКА МАТЕРИАЛОВ ЛЕСОВОДСТВЕННЫХ ИССЛЕДОВАНИЙ

Учебное пособие







САНКТ-ПЕТЕРБУРГ
2016

Министерство образования Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЛЕСОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

САНКТ-ПЕТЕРБУРГСКИЙ НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ
ИНСТИТУТ ЛЕСНОГО ХОЗЯЙСТВА

А.С. Бондаренко
кандидат сельскохозяйственных наук

А.В. Жигунов
доктор сельскохозяйственных наук, профессор

СТАТИСТИЧЕСКАЯ ОБРАБОТКА МАТЕРИАЛОВ ЛЕСОВОДСТВЕННЫХ ИССЛЕДОВАНИЙ

Учебное пособие

Санкт-Петербург
Издательство Политехнического университета
2016

УДК 630*236.4
ББК 65.34
Б81

Рассмотрено и рекомендовано к изданию учебно-методической комиссией Института леса и природопользования Санкт-Петербургского государственного лесотехнического университета 22 сентября 2016 г.

Ответственный редактор
Доктор географических наук, профессор, заведующий кафедрой Санкт-Петербургский Лесотехнический университет *А.С. Алексеев*

Рецензенты:
Доктор сельскохозяйственных наук,
начальник научно-исследовательского отдела
Санкт-Петербургского НИИ лесного хозяйства» *А.Б. Егоров*
Кандидат сельскохозяйственных наук,
директор Ленинградского НИИСХ «Белогорка» *Д.А. Данилов*

Бондаренко А.С. **Статистическая обработка материалов лесоводственных исследований :** Учебное пособие / А.С. Бондаренко, А.В. Жигунов. – СПб: Из-во Политехнического университета, 2016. – 125 с.

Приводятся некоторые общие положения прикладной статистики и наиболее часто используемые методы статистического анализа применительно к материалам лесоводственных исследований: расчёт основных статистик, дисперсионный, регрессионный, корреляционный анализ, многомерные методы статистического анализа. Описана последовательность вычисления статистических показателей и даны примеры обработки данных с помощью калькулятора и с применением программ компьютерного обеспечения MS Excel 2013 и Statistica 12.

Предназначено для инженерно-технических и научных работников лесного комплекса, бакалавров направления 350301 «Лесное дело», магистров направления 350401 «Лесное дело», аспирантов специальности 06.03.01 – Лесные культуры, селекция, семеноводство и 06.03.02 – лесоведение, лесоводство, лесоустройство и лесная таксация.

Табл. 4. Ил. 85. Библиогр.: 23 назв.

Редактор *А. С. Алексеев*
Компьютерная верстка: *А. С. Бондаренко, А. В. Жигунов*

© Бондаренко А.С, Жигунов А.В, 2016
© Санкт-Петербургский государственный лесотехнический университет, 2016
© Санкт-Петербургский политехнический университет Петра Великого, 2016

ISBN 978-5-7422-5495-9

ЦЕЛЬ И ЗАДАЧИ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Лесные сообщества и их структурные элементы являются вероятностными системами с различными уровнями детерминации проходящих в них процессов и явлений. Эффективность ведения лесного хозяйства определяется полнотой научных сведений о естественном формировании лесных фитоценозов и воздействии на них хозяйственных мероприятий.

Основными показателями, которыми оперируют при оценке отдельного дерева, группы деревьев или учетного ряда в лесных культурах, являются: высота и прирост дерева в высоту, диаметр ствола и его радиальный прирост, объем ствола, запас фитомассы и ее фракционный состав, длина хвои, соотношение прироста главного и боковых побегов и др. В результате экспериментальных наблюдений исследователь получает ряд числовых характеристик. Достоверность этих сведений оценивается путем статистической обработки цифрового материала. При этом полученные результаты следует рассматривать в качестве основы для оптимизации управленческих, технологических и экономических процессов.

Статистика - это наука, изучающая методы сбора и интерпретации числовых данных *[Справочник по прикладной статистике, 1989]*. Статистика имеет дело с данными, подверженными разбросу, то есть имеющими в своём составе случайную (стохастическую) компоненту. Такая рассогласованность (разброс) данных может обусловливаться множеством трудноучитываемых факторов (ошибки наблюдения, непредсказуемые флуктуации внешней среды и свойств самого объекта и др.). Чаще всего ситуация слишком сложна, чтобы её можно было изучить на основе полного описания, отражающего все детали. Особенно актуально это по отношению к живым объектам, которые характеризуются высокой степенью сложности. Обычно применяется некоторая математическая модель явления, которая воспроизводит его существенные черты и исключает те, которые предполагаются несущественными. Такая модель включает в себя детерминистские и стохастические (случайные) элементы. Главная задача статистической обработки экспериментальных данных – получение осмысленных заключений из подверженных разбросу (несогласованных) данных.

При написании учебного пособия без прямого цитирования и специальных ссылок были использованы книги и работы ряда авторов *[1-22]*, обратившись к которым можно более подробно познакомиться с теоретической стороной отдельных вопросов.

При иллюстрировании отдельных процедур расчетов статистических показателей используется модуль Excel из пакета MS Office 2013, а также статистическая программа Statistica 12 компании StatSoft.

1. ОБЩИЕ ПРИНЦИПЫ ОБРАБОТКИ ПОЛЕВЫХ ДАННЫХ

В процессе выполнения камеральных работ и обработки данных полевых измерений важную роль играет технология записи данных и перевода их в электронный вид, поскольку от того, в каком виде записываются данные полевых измерений, зависит скорость и качество их дальнейшей обработки. В частности, все современные профессиональные пакеты статистической обработки информации используют один и тот же принцип формирования базы данных, которого рекомендуется придерживаться при записи результатов измерений и последующем переносе их на электронные носители (компьютер). Указанный принцип можно условно сформулировать как «один объект – одна строка, один атрибут – один столбец». При этом важно сразу определить, что является объектом и что – его атрибутами (свойствами). Например, при замерах биометрических показателей в лесных культурах объектом является отдельное дерево, а его отдельными атрибутами – дата наблюдений, лесничество (где располагается соответствующий участок), участковое лесничество, квартал, выдел, номер делянки, варианта, ФИО замерщика, высота, диаметр ствола, наличие повреждений, примечание и т.п. Соответственно, каждый из атрибутов записывается в отдельный столбец. Пример такой записи приведен в табл. 1.

При выполнении полевых измерений данные записываются на бумажные носители простым карандашом, характеризующимся устойчивостью к воде. Возможно использование в полевых условиях электронных систем, таких как ноутбуки, планшетные компьютеры и др., тем не менее, следует учитывать специфику полевых измерений и предпринимать необходимые меры для сохранности вводимых данных (например, использовать периодическое резервное копирование на обособленные портативные носители информации – карточки, флеш-накопители и т.п.) и возможности их однозначной интерпретации при последующей обработке (необходимо исключить какую-либо машинную трансформацию данных при их вводе, сохранении и перезаписи).

Следует отметить, что при записи данных в полевых условиях для экономии времени целесообразно группировать повторяющиеся атрибуты (такие как информация о замерщике, наименование лесничества, лесотаксационные квартал и выдел, номер и площадь повторности и др.) в начале наблюдений, то есть под шапкой таблицы до записей с замерами биомет-

рических показателей и других индивидуальных для каждого дерева данных. Кроме того, для экономии времени имеет смысл не записывать каждый раз один и тот же повторяющийся атрибут (например, номера делянки), а ставить обозначение повтора (- // -).Тем не менее, необходимо помнить, что при переводе данных на электронные носители информации все атрибуты должны быть записаны в программу обработки данных по отдельным столбцам и в полном объёме для каждого растения (без пропусков и посторонних обозначений).

Таблица 1

Пример заполнения журнала полевых измерений в лесных культурах

Пробная площадь	Высота, м	Диаметр, см	Примечание
Дата: 12 сентября 2016г. Замерщик: Петров И.П., Волосовское л-во, кв. 14, выд. 18, Еевр., опытные лесн. к-ры, площадь 5,0 га Повторность 1, площадь 0,13 га			
1	15,1	18	наклон ствола
- // -	16,4	19	
2	18,2	20	
- // -	14,0	15	
- // -	15,0	16	двухверш.
- // -	13,9	14	
Повторность 2, площадь 0,08 га			
2	13,0	15	двухверш.
- // -	14,1	18	
Дата: 15 сентября 2016г. Замерщик: Иванов П.И., Гатчинское л-во, кв. 72, выд. 2, Еевр., опытные лесн. к-ры, площадь 4,5 га Повторность 1, площадь 0,08 га			
1	3,42	6,3	
- // -	2,54	5,8	усыхающее
2	4,36	6,2	
- // -	5,52	8,9	мех. повреждение
- // -	4,28	7,0	

Рассмотрим перевод данных полевых наблюдений на электронные носители информации. Электронная таблица строится по тому же приведённому выше принципу «один объект – одна строка, один атрибут – один столбец», но имеет свою специфику, отличную от методики данных в полевых условиях. В частности, в электронной таблице должны быть запол-

нены все ячейки, которые для удобства были пропущены в полевом журнале. При этом необходимо знать, отсутствует ли значение соответствующего атрибута объекта или оно просто не записано в полевом журнале из соображений экономии времени. При вводе данных на электронные носители информации пропуск в значении атрибута возможен только в том случае, если объект (дерево) не содержит значения для определённого атрибута. Так, например, название лесничества или номер квартала должны быть обязательно внесены в каждую ячейку, поскольку это непременное свойство каждого дерева (все деревья имеют координаты расположения), а такой атрибут как «примечание» может быть заполнен далеко не для всех деревьев (дополнительная информация для некоторых деревьев может отсутствовать), поэтому указанные ячейки для отдельных деревьев могут быть пустыми (ячейки, не содержащие значений для соответствующего атрибута, остаются незаполненными).

Таким образом, приведённая выше таблица в электронном варианте будет иметь следующий вид (табл. 2).

Таблица 2

Пример заполнения электронной таблицы по данным измерений в лесных культурах

Замер-щик	Лес-ничe-ство	Кв	Выд	Повтор-ность	Делянка	Высота, м	Диаметр, см	Примеча-ние
Petrov	VLS	14	18	1	1	15,1	18	наклон ствола
Petrov	VLS	14	18	1	1	16,4	19	
Petrov	VLS	14	18	1	2	18,2	20	
Petrov	VLS	14	18	1	2	14,0	15	
Petrov	VLS	14	18	1	2	15,0	16	двухвер-шинность
Petrov	VLS	14	18	1	2	13,9	14	
Petrov	VLS	14	18	2	2	13,0	15	двухвер-шинность
Petrov	VLS	14	18	2	2	14,1	18	
Ivanov	GTN	72	2	23	1	3,42	6,3	
Ivanov	GTN	72	2	23	1	2,54	5,8	усыхающее
Ivanov	GTN	72	2	12	1	4,36	6,2	
Ivanov	GTN	72	2	12	1	5,52	8,9	мех. по-вреждение
Ivanov	GTN	72	2	12	1	4,28	7,0	

Особое значение при переводе данных на электронные носители информации следует уделять методикам, позволяющим избежать ошибок, а также верификации данных. В частности, при заполнении электронных таблиц необходимо обращать особое внимание на качество записи номинальных переменных, таких, например, как «лесничество», вид повреждения», «форма кроны» и т.п. Поскольку по этим данным в большинстве видов статистического анализа производится группировка, необходимо принимать меры для устранения ошибок при вводе такого типа данных, так как ошибка всего в одном знаке (например, наличие пробела после названия лесничества в одной ячейке и отсутствие этого пробела – в другой) приведет к тому, что программа обработки данных будет оценивать такие значения как различающиеся и, как следствие, такие объекты при группировке будут отнесены к разным совокупностям. В результате возникают ошибки при обработке данных, которые могут иметь решающее значение при формировании выводов. Во избежание такого рода дефектов ввода данных целесообразно использовать некоторые несложные приемы при переводе вводе данных на электронные носители информации:

Прием 1. В отношении номинальных переменных полезно использовать не полные наименования, а его условное сокращенное обозначение. Например, вместо значения «Лодейнопольское» для атрибута «Наименование лесничества» вводить в соответствующую ячейку «LDP», вместо «Гатчинское» - «GTN» и т. д. (таблица 2). Естественно, имеет смысл внести соответствующие пояснения для «обратной расшифровки» в примечания файлов, либо в рабочую документацию (для повышения сохранности данных лучше одновременно и в электронный и в бумажный носитель информации). Следует отметить, что ввод значений номинальных переменных в стандартные статистические пакеты лучше выполнять латиницей, так как англоязычные программы, к которым относится большинство приложений такого рода, зачастую плохо оперируют кирилическими шрифтами.

Прием 2. При внесении данных в электронную таблицу необходимо как можно интенсивнее использовать такой инструмент как «копировать-вставить» (а также инструмент «Автозамена») вместо непосредственной набивки повторяющихся данных с клавиатуры. Такой прием позволяет снизить влияние ослабления внимания исполнителя от усталости, невнимательности и т.п.

Прием 3. Как можно шире использовать автоматическую проверку введенных значений на наличие грубых ошибок. Это можно сделать, например, при помощи модуля «Сводная таблица» в MS Excel, которая

позволяет сгруппировать данные по определённому столбцу таблицы и далее проверить имеются ли в составе этих данных значения, различающиеся из-за ввода ошибочных данных. Вышеприведённая ошибка с пробелом в наименовании лесничества проявится наличием в сгруппированных данных двух похожих лесничеств: одно с пробелом в конце слова, другое – без пробела.

При внесении в электронный вид количественных переменных для этих целей очень удобно использовать для проверки данных функции «максимум» и «минимум» в строке состояния MS Excel в нижней части рабочего окна программы для выявления грубых ошибок. В этом случае при выделении столбца исполнитель видит максимальное (минимальное) значение переменной и имеющийся выброс в данных (например, значение высоты 80 м вместо фактической 8,0 м в лесных культурах) может быть легко диагностирован и уточнен по полевой тетради.

Кроме приведённой простейшей верификации «минимум-максимум» целесообразно использовать более сложные методики выявления выбросов (грубых ошибок), например, по значениям стандартного отклонения и ошибки среднего значения. Так, например, наличие выбивающегося из общего ряда (обычно в большую сторону) значения ошибки среднего того или иного варианта (повторность, делянка) говорит либо о малом числе наблюдений изучаемого варианта, либо о наличии значения переменной, сильно отличающегося от общего массива данных. В этом случае требуется проверка значений биометрического показателя для соответствующего варианта. Примерно такой же смысл имеет и проверка стандартного отклонения с тем отличием, что теоретически стандартное отклонение не зависит от количества измеренных значений для определенного варианта, а зависит от природы изучаемых данных. То есть сильные различия в значениях стандартного отклонения для разных опытных единиц (блоки, делянки) в пределах одного участка говорят либо об ошибках ввода данных, либо о некорректных способах измерения величин и значительном влиянии случайной составляющей изменчивости признака.

Еще одним способом устранения грубых ошибок при вводе количественных переменных является проверка введенных значений на наличие нечисловых значений. Такая проверка легко выполняется средствами любого из статических пакетов либо стандартными функциями проверки свойств и значений MS Excel.

Одним из способов парного сравнения переменных является построение графических зависимостей по двум переменным. Например, при построении графика зависимости высоты от диаметра в MS Excel сразу можно увидеть точки, выбивающиеся из общего вытянутого (в соответ-

ствии с типом зависимости высоты от диаметра) облака точек. Выбивающаяся из общей зависимости точка, так называемый выброс, будет соответствовать дереву с «неправильным» соотношением высоты и диаметра (например, высота 4 м при диаметре дерева 34 см). Проверка значений такого выброса по полевой тетради позволит распознать грубые ошибки, не выявляемые на основе использования описанных выше методик.

2. РАСЧЕТ ОСНОВНЫХ СТАТИСТИК

В практических задачах обычно имеется совокупность наблюдений $x_1, x_2, \ldots, x_n$, на основе которых необходимо сделать те или иные выводы. Для компактного описания данной совокупности используются методы описательной статистики. Показателями, описывающими выборку, являются статистики (или параметры) положения (среднее, мода, медиана, минимум, максимум, верхний и нижний квартили и др.) и разброса (дисперсия, стандартное отклонение, коэффициент вариации, и др.).

Среднее арифметическое значение (mean) показывает "центральное положение" переменной и рассчитывается следующим образом:

$$\overline{X} = \frac{\Sigma x_i}{n} \ ,$$

где $\overline{X}$ - расчётное среднее значение величины;

Σx_i - сумма значений биометрического показателя;

n – количество наблюдений;

Пример. Имеется ведомость замера высот растений в лесных культурах:

Высота, см	46	70	45	10	63	31	64	40	26	23	18	10	18	71	25

Рассчитаем среднее для ряда из 15-ти значений высот растений:

$$\overline{X} = \frac{\Sigma x_i}{n} = \frac{46+70+45+10+63+31+64+40+26+23+18+10+18+71+25}{15} = 37{,}3 \text{ см}$$

Расчёт в MS Excel производится следующим образом: создаем новую книгу (быстрая клавиша Ctrl + N), вводим с клавиатуры данные о значениях биометрических показателей в ячейки листа *MS Excel* (каждое значение – в отдельную ячейку) и устанавливаем курсор в ячейку, где необходимо получить среднее значение. В меню "Формулы", где различные формулы для расчётов сгруппированы в несколько групп: «автосумма», «фи-

нансовые», «логические» и др., выбираем команду "Автосумма". В ниспадающем меню выбираем функцию «Сумма». В выбранной ячейке будет создана формула = СРЗНАЧ(). В круглых скобках функции при помощи клавиатуры (раскладка английская) и манипулятора-мыши выделяем диапазон чисел, нажимаем ОК. В ячейке будет создана формула вида = СРЗНАЧ(A2:A12) со ссылкой на диапазон ячеек (рис. 1). После ввода данных (клавиша "Enter") получаем результат расчета по введенной формуле: 37,3333. Степень округления можно задать при помощи контекстного меню (правая клавиша мыши) «Формат ячеек». Во вкладке «Число» указываем количество знаков после запятой в поле «Число десятичных знаков». В последующих описаниях работы с *MS Excel* для экономии места мы будем приводить описания операций менее подробно, поскольку они в значительной степени однотипны.

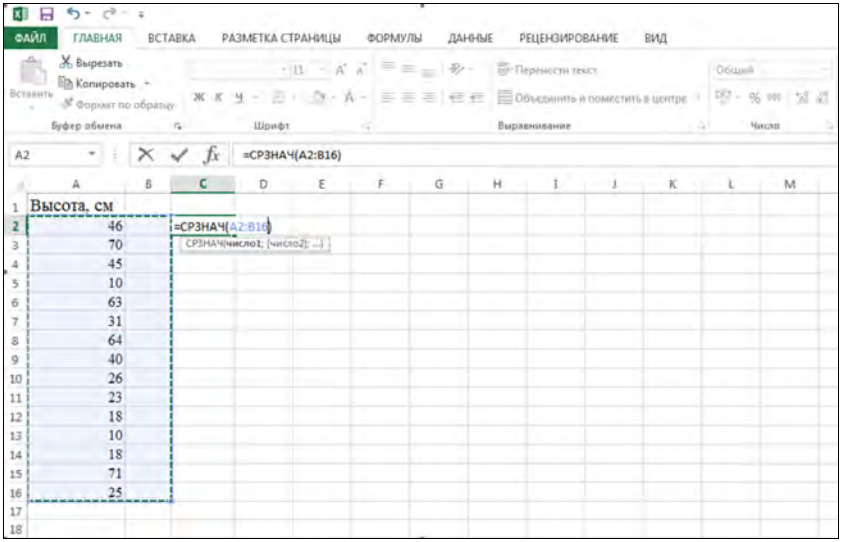


Рис. 1. Ввод формулы в ячейке окна рабочей книги MS Excel

М е д и а н а (median) - это значение, которое разбивает выборку на две равные части. Половина наблюдений лежит ниже медианы и половина - выше ее. Если число наблюдений в выборке нечетно, то медиана вычисляется как среднее двух средних значений. Медиана случайной величины x обозначается обычно как $med\ (x)$.

В *MS Excel* медиана рассчитывается аналогично среднему значению (см. выше) при помощи функции МЕДИАНА().

М о д а (mode) в ы б о р к и - это значение, наиболее часто встречающееся в выборке (максимум функции плотности распределения). Мода случайной величины *x* обозначается как *mod(x)*. В *MS Excel* мода рассчитывается при помощи функции МОДА().

С у м м а к в а д р а т о в о т к л о н е н и й (sum of squares) - это сумма возведённых в квадрат отклонений случайной величины от её среднего значения:

$$SS = \sum(x_i - \bar{x})^2 .$$

Сумма квадратов используется практически во всех разделах статистического анализа, особенно большую роль этот показатель играет в задачах регрессионного и дисперсионного анализа. В *MS Excel* сумма квадратов рассчитывается при помощи функции КВАДРОТКЛ ().

Д и с п е р с и я (variance) - это усреднённая для выборки сумма квадратов отклонений:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} .$$

Более удобной для расчётов является следующая формула:

$$\sigma^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} .$$

Дисперсия является одной из основных характеристик изменчивости изучаемой величины. Важнейшим свойством дисперсии является то, что она может быть расчленена на части, каждая из которых обусловлена каким-либо фактором. Сумма этих частей является полной дисперсией случайной величины.

В *MS Excel* дисперсия выборки рассчитывается при помощи функции ДИСП().

С т а н д а р т н о е (среднеквадратическое) о т к л о н е н и е (standard deviation) также является мерой разнообразия объектов, входящих в выборку. В просторечии стандартное отклонение именуется термином "сигма". Рассчитывается как квадратный корень из дисперсии, следовательно, имеет ту же размерность, что и изучаемый признак:

$$\sigma = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} .$$

Стандартное отклонение показывает, насколько в среднем отклоняется значение признака каждого объекта от среднего арифметического. Чем сильнее разброс значений, тем больше стандартное отклонение. Поскольку стандартное отклонение – величина с определённой единицей измерения (метры, килограммы, штуки и т. д.), то по нему можно сравнивать изменчивость только одних и тех же признаков, выраженных в одних единицах и имеющих сопоставимые средние величины. Это ограничивает применимость данного параметра. Для сравнения уровня изменчивости признаков разной размерности используется коэффициент вариации.

Пример. Поскольку стандартное отклонение является очень важной характеристикой и встречается во всех приложениях статистического анализа, рассмотрим пример расчёта данной величины для диаметров растений:

Диаметр, см	11	12	9	11	10	14	14	16	12	10	15

$$\sigma_1 = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} =$$

$$= \sqrt{\frac{(11^2+12^2+9^2+11^2+10^2+14^2+14^2+16^2+12^2+10^2+15^2) - \frac{(11+12+9+11+10+14+14+16+12+10+15)^2}{11}}{11-1}} =$$

$$= 2{,}27.$$

Расчёт в MS Excel производится при помощи команды СТАНДОТКЛОН(), которую необходимо ввести непосредственно в ячейке после знака равенства $\boxed{=}$ (рис. 2) Знак равенства при вводе в ячейку воспринимается программой как начальный аргумент математической функции, список которых в данной программе очень широк и вызывается вкладкой «Вставить формулу» в меню «Формулы».

К о э ф ф и ц и е н т в а р и а ц и и (coefficient of variation) количественного признака представляет собой отношение стандартного отклонения к среднему арифметическому значению и выражается обычно в процентах:

$$CV = \frac{\sigma}{\overline{X}} \cdot 100\% .$$

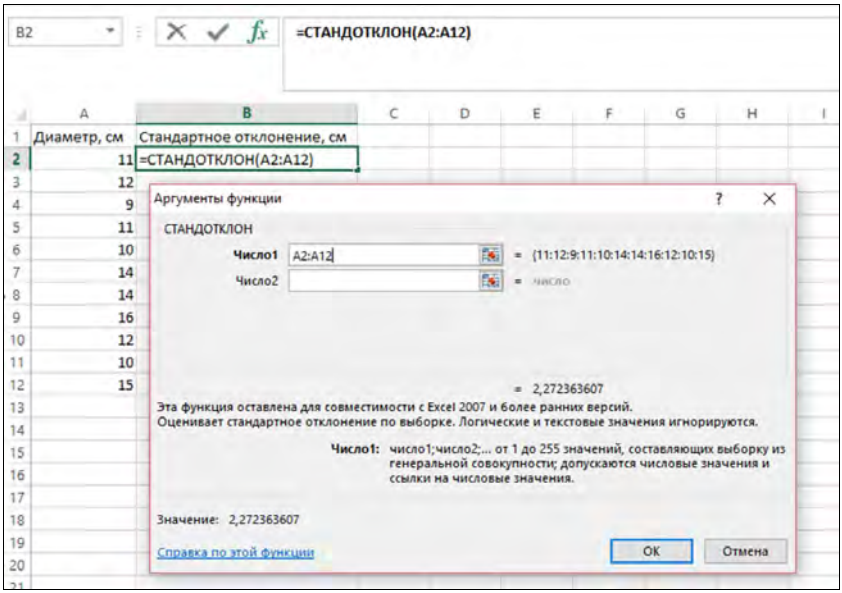


Рис. 2. Диалоговое окно вычисления стандартного отклонения

К о в а р и а ц и я (covariance) является мерой сопряжённой изменчивости двух признаков. Ковариация признаков x и y рассчитывается следующим образом:

$$\sigma_{xy} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{n-1}.$$

Ковариация отражает дисперсию, обусловленную совместным варьированием признаков.

В *MS Excel* ковариация рассчитывается с применением команды КОВАР().

А с и м м е т р и я, или коэффициент косости (skewness), является мерой «скошенности» распределения численностей и для выборки вычисляется по формуле:

$$A = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^3.$$

Для классической функции плотности нормального распределения коэффициент асимметрии равен 0. При левой асимметрии, т. е. когда вер-

шина кривой сдвинута влево, а правая часть кривой растянута – знак положительный, при правой асимметрии – знак отрицательный.

В *MS Excel* асимметрия рассчитывается с применением команды СКОС().

Э к с ц е с с или коэффициент крутости (kurtosis) измеряет "пикообразность" распределения и для выборки вычисляется по формуле:

$$E = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 \right) - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Функция плотности классического нормального распределения имеет эксцесс, равный 0. Если эксцесс значимо превышает 0, то функция плотности является более островершинной, чем при нормальном распределении, если же эксцесс отрицателен - функция плотности плосковершинна (растянута в стороны).

В *MS Excel* асимметрия рассчитывается аналогично вышеприведенному примеру со стандартным отклонением (рис. 2) при помощи команды ЭКСЦЕСС().

Расчёт основных статистик в программе Statistica.

Для того, чтобы начать выполнять расчеты в программе *Statistica* сначала необходимо создать новый документ. Для этого в меню File выбираем строку New либо используем сочетание клавиш CTRL+ N (рис. 3).

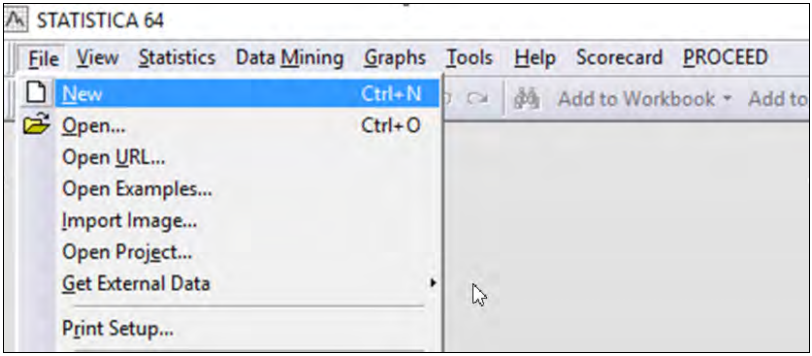


Рис. 3. Фрагмент меню для создания нового документа

Далее в окне «Create New Document» открываем вкладку «Spreadsheet», где выбираем основные параметры электронной таблицы с данными, в частности, количество столбцов (*Number of variables*) и строк (*Number of cases*) файла данных, а также при необходимости - формат отображения вводимых данных (*Display format*). Данное окно приведено на рисунке 4, для примера количество столбцов принято равным двум, а количество строк – 15 (рис. 4).

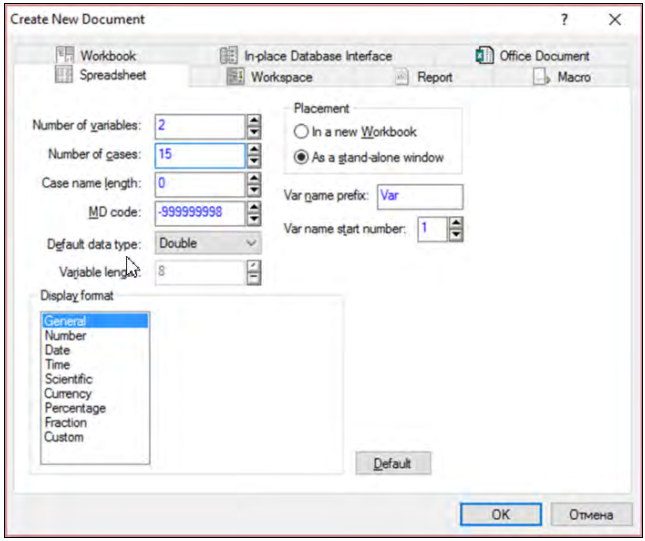


Рис. 4. Диалоговое окно создания новой книги в Statistica

Чтобы поменять название и другие атрибуты переменных, необходимо двойным щелчком нажать на шапку столбца и в открывшемся диалоговом окне набрать необходимое название в окошке «Name» и, при необходимости, тип переменных (рис. 5). Вводим данные (используем приведённые выше данные о высотах деревьев) в ячейке аналогично *MS Excel*. В меню *File* сохраняем по нужному адресу файл данных при помощи меню *Save as* либо при помощи функциональной клавиши F12.

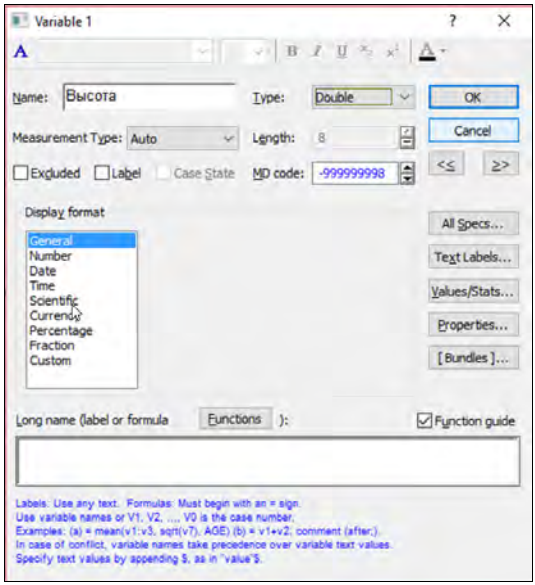


Рис. 5. Диалоговое окно свойств переменной

Далее для расчета основных статистик в меню *Statistics* следует запустить модуль *Basic statistics / Tables* (основные статистики и таблицы). Выбираем необходимый раздел анализа: раздел *Descriptive Statistics* (описательная статистика). Появляется диалоговое окно описательной статистики (рис. 6).

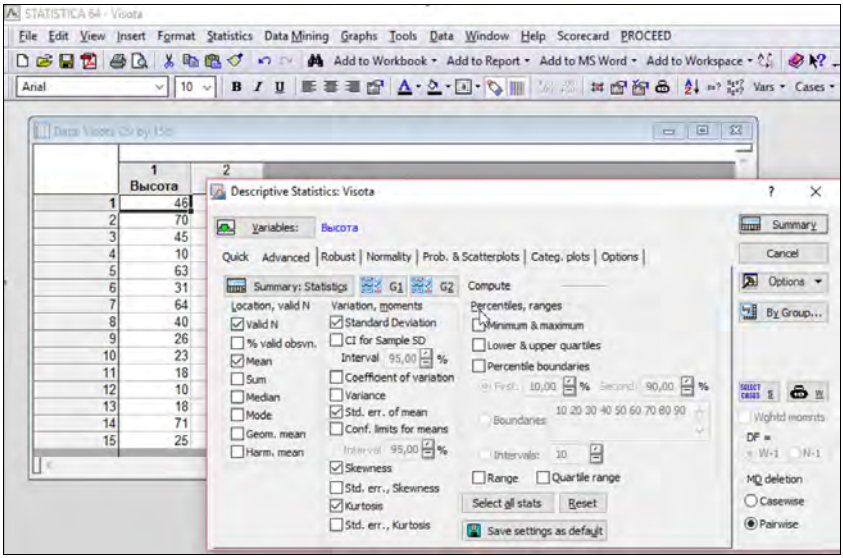


Рис. 6. Диалоговое окно описательных статистик

После нажатия на кнопку *Variables* (переменные) выбираем переменные, для которых необходимо вычислить статистики (можно выбрать несколько переменных). Во вкладке *Advanced* флажками отмечаем необходимые статистики, например, количество наблюдений *(Valid N)*, среднее *(Mean)*, медиану *(Median)*, стандартное отклонение *(Standard deviation)*, ошибку среднего *(Std. err. of mean)*, асимметрию *(Skewness)*, эксцесс *(Kurtosis)*. После нажатия на кнопку *Summary* будет показана таблица с вычисленными значениями (рис. 7).

	Descriptive Statistics (Visota)						
Variable	Valid N	Mean	Median	Std.Dev.	Standard Error	Skewness	Kurtosis
Высота	15	37,33333	31,00000	21,51301	5,554635	0,398868	-1,28022

Рис. 7. Результаты расчета некоторых описательных статистик

При необходимости полученные результаты расчета статистик можно скопировать в *MS Excel* и другие приложения.

3. ТИПЫ ПЕРЕМЕННЫХ ВЕЛИЧИН

В статистическом анализе данных оперируют понятием «переменная» или «случайная» величина. Переменная - это то, что можно измерять, контролировать или что можно изменять в исследованиях.

Типы переменных. В зависимости от количества информации, содержащейся в переменных, они делятся на следующие типы: количественные (интервальные), порядковые (ординальные), качественные (номинальные).

Количественные переменные позволяют измерить в определённой шкале степень проявления изучаемого свойства объекта, численно выразить и сравнить различия между объектами по изучаемому признаку. Такими переменными являются, например, высота и диаметр дерева, возраст, количество семян в шишке, площадь выдела и т. п. При этом можем сказать, что одно дерево по высоте на 2 м превышает другое или количество семян в одной шишке в 3 раза меньше, чем в другой.

Порядковые переменные позволяют ранжировать (упорядочить) объекты, указав, какие из них в большей или меньшей степени обладают качеством, выраженным данной переменной. Однако они не позволяют сказать, "на сколько больше" или "на сколько меньше". Примерами порядковой переменной могут служить классы бонитета (показатель производительности условий произрастания) и балл плодоношения деревьев. Мы можем утверждать, что первый класс бонитета более производительный, чем третий, но не можем сказать насколько в количественном выражении (например, на 10 м^3 по запасу). Точно также мы упорядочиваем деревья по интенсивности плодоношения от минимального (балл – 0) до максимального (балл – 5), но не можем сказать насколько сильно пятый балл плодоношения отличается от нулевого.

В некоторых ситуациях (например, в непараметрических тестах) бывает необходимо перевести количественную шкалу в порядковую. Процедура такого перехода называется *ранжированием.* При этом наблюдения упорядочиваются от меньшего к большему (или наоборот) и присваивают им ранги. Трудности в назначении рангов возникают в том случае, если встречаются совпадающие значения. В этом случае вычисляются средние ранги, которые являются средним арифметическим значением рангов, назначаемых для чисел в том случае, если бы они различались на некоторую небольшую величину. В любом случае в результате расчета наибольший ранг должен быть равен общему количеству элементов.

Для примера рассмотрим выборку 5; 3; 1; 12; 5; 4; 5; 3. Упорядочиваем элементы от меньшего к большему: 1; 3; 3; 4;.5; 5; 5; 12. Присваиваем

ранги этим элементам: 1; 2,5; 2,5; 4; 6; 6; 6; 8. Таким образом, исходный ряд элементов имеет следующую ранжировку: 6; 2,5; 1; 8; 6; 4; 6; 2,5.

Номинальные (качественные) переменные позволяют разбивать исследуемую совокупность на неподдающиеся упорядочиванию однородные по анализируемому свойству классы (номер опытного участка, древесная порода, пол, вариант обработки почвы). Часто номинальные переменные называют категориальными.

Зависимые и независимые переменные. Независимыми называются переменные, которые задаются исследователем, тогда как зависимые - это переменные, которые измеряются или регистрируются: исследователь выбирает некоторую переменную (независимую) и исследует, каким образом изменяется в зависимости от значений этой переменной другая переменная. Первая переменная называется независимой, вторая – зависимой. Термины зависимая и независимая переменная применяются, в основном, в экспериментальном исследовании, где экспериментатор оперирует некоторыми переменными, и в этом смысле они "независимы" от реакций, свойств, намерений и т.д. присущих объектам исследования. Некоторые другие переменные, как предполагается, должны "зависеть" от действий экспериментатора или от экспериментальных условий. Иными словами, зависимость проявляется в ответной реакции исследуемого объекта на посланное на него воздействие. Например, необходимо изучить ход роста по высоте дерева сосны. Исследователь выбирает временной промежуток (допустим, от 10- до 50-летнего возраста) и изучает, как в зависимости от возраста изменяется высота дерева. В этом случае возраст – независимая, высота – зависимая переменная. Другой пример: в эксперименте деревья сосны сравниваются с деревьями ели по содержанию смолистых веществ. В данном случае исследователь выбирает древесную породу и изучает, как в зависимости от вида древесной породы изменяется содержание смолистых веществ. Здесь принадлежность к древесной породе является независимой, а содержание смолистых веществ - зависимой переменной. Отметим, что в первом примере и зависимая, и независимая переменные являются количественными переменными, а во втором независимая переменная относится к номинальным (качественным), а зависимая – к количественным переменным.

4. ОСНОВНЫЕ ПОНЯТИЯ

В ы б о р к а – это некоторое количество объектов (особей), выбранных из данной группировки в целях оценки параметров (среднее, дисперсия и т.д.) этой группировки. Например, деревья в данном насаждении, от которых берутся образцы древесины для определения её плотности, образуют выборку, взятую из данного насаждения. Характерная особенность большинства биологических исследований заключается в том, что выборка содержит лишь малую часть исследуемой группировки, поскольку всю совокупность исследовать невозможно. Выборка является "представителем" той группировки, откуда она взята. То, откуда берётся выборка, в статистике называют генеральной совокупностью. Под ней можно понимать определённую группировку особей. Вообще же в статистике генеральная совокупность понимается шире – как множество каких-либо однородных элементов, из которого по определённому правилу выделяется некоторое подмножество, называемое выборкой. Таким образом, понятия генеральной совокупности и выборки неразрывно связаны. Важнейшим свойством выборки в статистике является её случайность. Принцип случайности означает равную вероятность попадания каждого элемента изучаемой генеральной совокупности в выборку.

Нарушение принципа случайности может привести к серьёзным ошибкам в выводах. Так, например, в некоторых руководствах можно встретить методики отбора "типичных представителей" генеральной совокупности на основе визуальных данных (например, "модельные деревья"). Такая методика сама по себе может быть нужной, например, для подбора особей в экспериментальную группу, где важна выравненность изучаемых объектов по ряду показателей. Тем не менее, для характеристики генеральной совокупности такой отбор неприемлем, поскольку в значительной мере субъективен и не может служить базой для получения статистически обоснованных выводов. Из такой выборки, например, заведомо исключаются нетипичные, но тем не менее, довольно многочисленные особи. Это ведёт к неверным оценкам: занижению показателей разнообразия, смещению межгрупповых различий, искажению вкладов изучаемых факторов в общую изменчивость признака.

Принцип случайности позволяет теоретически описать распределение оцениваемого параметра (например, среднее арифметическое) в генеральной совокупности, состоящей из всех мыслимых случайных выборок данного объёма, которые можно было бы извлечь из данной совокупности. Самое главное то, что наша реальная выборка подчиняется этому распре-

делению лишь тогда, когда она случайна. Это положение является основой статистических методов.

Достоверность (reliability). Достоверность (или надёжность) непосредственно связана с репрезентативностью определенной выборки, на основе которой строятся выводы. Репрезентативность – это показатель того, насколько результаты выборки соответствуют генеральным параметрам. Другими словами, достоверность говорит о том, насколько вероятно, что зависимость, подобная найденной для данной выборки, будет вновь обнаружена (иными словами, подтвердится) на данных другой выборки, извлеченной из той же самой генеральной совокупности. Следует отметить, что конечной целью статистического исследования почти никогда не является изучение данной конкретной выборки. Выборка представляет интерес лишь постольку, поскольку она дает информацию обо всей генеральной совокупности. Если исследование удовлетворяет требованиям, предъявляемым к статистическому исследованию, то достоверность найденных зависимостей между переменными изучаемой выборки можно количественно оценить и представить с помощью стандартной статистической меры (называемой *p*-уровень или статистический уровень значимости, подробнее см. в след. разд.).

Уровень значимости (p-level) представляет собой меру доверия к полученному результату. Ее более высокий уровень соответствует более низкому уровню доверия к найденной в статистическом анализе закономерности. Уровень значимости показывает вероятность ошибки, связанной с распространением наблюдаемого результата на всю генеральную совокупность, т. е вероятность того, что найденная в исследуемой выборке закономерность (связь между переменными) является лишь случайной особенностью данной выборки. Так, например, если уровень значимости *t*-критерия, оценивающего степень различий между двумя выборками, составляет 0,032, то вероятность случайности полученных различий между выборками составит чуть более 3%. Выбор определенного уровня значимости, выше которого результаты отвергаются как ложные, является достаточно произвольным. Обычно во многих областях результат $p \leq 0{,}05$ является приемлемой границей статистической значимости. То есть величина уровня значимости, равная или меньшая, чем 0,05, говорит о достоверности полученного результата. Тем не менее, следует помнить, что этот уровень все еще включает довольно большую вероятность ошибки (5%). Результаты, полученные на уровне $p \leq 0{,}01$, обычно рассматриваются как статистически значимые, а на уровнем $p \leq 0{,}005$ или $p \leq 0{,}001$ - как высоко значимые. Однако следует помнить, что данная классификация уровней значимости достаточно произвольна и является всего лишь неформальным

соглашением, принятым на основе практического опыта в той или иной области исследования.

Ошибка параметра или стандартная ошибка (standard error). Часть практически не дает возможности охарактеризовать целое. Всегда остаётся вероятность того, что выборочные данные недостаточно точно описывают генеральную совокупность, содержат определённую ошибку. Такая ошибка, связанная с переносом результатов, полученных на основе анализа выборки на всю генеральную совокупность, называется ошибкой репрезентативности. Статистическая ошибка показывает, насколько в среднем может отклоняться значение какого-либо параметра (среднее, стандартное отклонение и т. д.), найденное по выборке (так называемая оценка параметра), от "истинного" в генеральной совокупности. Стандартные ошибки основных параметров рассчитываются следующим образом:

ошибка среднего $m_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$;

ошибка стандартного отклонения $m_{\sigma} = \dfrac{\sigma}{\sqrt{2n}}$;

ошибка коэффициента вариации: $m_{cv} = \dfrac{cv}{\sqrt{2n}}$.

Естественно, что чем больше размер выборки, тем меньше значение ошибки (по выборке точнее можно определить "истинное" значение параметра). Это находит отражение в формулах расчёта – значения всех ошибок параметров обратно пропорциональны корню квадратному количества наблюдений. Кроме того, ошибка зависит от степени изменчивости исследуемой величины: чем сильнее случайная величина варьирует, тем больше ошибка (при разбросе показателей "истинное" значение параметра определить сложнее), что также подтверждается формулами: ошибка прямо пропорциональна показателю изменчивости.

Обычно ошибка записывается вместе с исследуемым параметром. Например, средняя высота древостоя составляет 18,3 ± 0,93 м. Это позволяет показать степень точности определения данного параметра. При этом количество значащих цифр ошибки должно быть на одну больше, чем у параметра (грубо говоря, на одну цифру после запятой больше).

На основе значения ошибки рассчитывается значение доверительного интервала параметра.

Доверительный интервал (confidence interval) для какой-либо статистики (например, стандартного отклонения) представляет

собой интервал значений вокруг оценки параметра (вычисленного по выборочным данным), в котором находится "истинное" значение этой статистики. При этом вероятность того, что "истинное" (генеральное) значение не попадает в этот интервал, равно уровню значимости. Например, если среднее выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем значимости p = 0,05 равны 19 и 27 соответственно, то можно заключить с 95%-ной вероятностью, что интервал с границами 19 и 27 "накрывает" среднее популяции. Соответственно, вероятность того, что "истинное" среднее не попало в этот интервал, составляет 5%. Если установить больший уровень доверия (т.е. уровень значимости сделать меньше), то интервал станет шире и вероятность, с которой он "накрывает" неизвестное среднее популяции, возрастает.

Доверительный интервал рассчитывается на основе ошибки параметра и значения t-критерия Стьюдента. Нижняя граница доверительного интервала равна $\bar{x} - tm$, верхняя - $\bar{x} + tm$, где m – ошибка параметра, вычисленная по выборочным данным, t – значение t-критерия для выбранного уровня значимости (количество наблюдений берётся для случая бесконечно большой совокупности). Допустим, вычисленный по выборке средний диаметр культур составляет 7,3 см, стандартная ошибка равна 0,43 (среднее значение записывается как 7,3±0,43 см). Как наиболее распространённый в биологических исследованиях примем уровень значимости, равный 0,05. При этом уровне и бесконечно большой совокупности значение t-критерия составляет 1,96. Следовательно, доверительный интервал для среднего значения составляет $\bar{x} \pm tm$ = 7,3 ± 1,96·0,43 = 7,3±0,84 см. Таким образом, с вероятностью 95% генеральное среднее находится в интервале от 6,46 до 8,14 см (7,3-0,84; 7,3+0,84).

Т о ч н о с т ь (accuracy). Стандартная ошибка является именованной величиной (имеет ту же размерность, что и исследуемый параметр). Поэтому её использование для сравнительной оценки точности, с которой определены выборочные значения параметров (например, средние) по отношению к их генеральным параметрам, неудобно. Для этой цели применяется обычно показатель точности опыта - отношение стандартной ошибки среднего к самому́ среднему значению случайной величины, выраженное в процентах:

$$P = \frac{m_{\bar{x}}}{\bar{x}} \cdot 100 .$$

Эта величина определяет степень надёжности полученных данных, их приближение к генеральной совокупности. В биологических исследованиях удовлетворительной принято считать точность, не превышающую

5%. Исходя из этой точности, определяют и минимальное количество наблюдений (см. п. 4.1).

5. ПЛАНИРОВАНИЕ ВЫБОРКИ

5.1. Установление минимального объёма выборки

Минимальное количество наблюдений можно определить исходя из требуемой точности опыта:

$$n = \frac{t^2 \cdot cv^2}{P^2},$$

где t - критерий Стьюдента для принятого уровня значимости (при обычно принятом уровне значимости 0,05 t-критерий равен 1,96 или, с округлением, 2,0); cv – коэффициент вариации, определяемый по данным прежних наблюдений или на основе пробной выборки; P^2 – точность, планируемая в данном опыте (обычно равна 5-10%).

Пример. Рассмотрим расчёт минимально необходимого количества измерений высоты на постоянном лесосеменном участке сосны. Мы хотим определить среднюю высоту на участке с точностью ±5%. Предыдущие измерения показывают, что в данном возрасте коэффициент вариации высоты на подобных участках составляет около 20%.

Рассчитываем количество наблюдений:

$$n = \frac{t^2 \cdot cv^2}{P^2} = \frac{2^2 \cdot 0{,}20^2}{0{,}05^2} = 64.$$

Таким образом, для того, чтобы узнать на участке среднюю высоту с точностью до 5% нам необходимо измерить высоты 65-70 деревьев (округление производится в большую сторону).

5.2. Способы организации выборки

Основные:

1) *простой случайный отбор* – способ извлечения n объектов из конечной генеральной совокупности N объектов, при котором каждая из возможных выборок имеет равную вероятность быть отобранной. На практике часто нумеруют объекты в генеральной совокупности от 1 до N и затем, используя таблицы случайных чисел или какой-либо другой метод, обеспечивающий равную вероятность выбора объекта, отбирают один за другим n объектов;

2) *регулярный отбор.* Часто применяется вместо случайного отбора. Это наиболее часто применяемый в практике метод, при использовании

которого объекты в выборку отбираются по определённому правилу, не зависящему от свойств объекта. Например, это может быть отбор каждого i-го объекта, отбор через определённые промежутки времени и т. д.;

3) *расслоённый (стратифицированный) отбор.* Заключается в том, что исходная генеральная совокупность объёма N подразделяется на подсовокупности объёма $N_1, N_2, \ldots, N_k$. При этом подсовокупности не содержат общих объектов и вместе исчерпывают всю генеральную совокупность, так что $N_1 + N_2 + \ldots + N_k = N$. Подсовокупности называются стратами, или слоями. Когда слои определены, из каждого извлекается простая случайная выборка объёма соответственно $n_1, n_2, \ldots, n_k$. Во избежание нарушения представленности объектов необходимо, чтобы размер выборок $n_1, n_2, \ldots, n_k$ был пропорционален размеру подсовокупностей $N_1, N_2, \ldots, N_k$. Расслоённый отбор применяется в случаях, когда слои однородны, т.е. входящие в них объекты имеют близкие характеристики.

Следует отметить, что способ сбора исходных данных должен тщательно планироваться и его необходимо полностью описывать в отчётах о выполненной работе.

Примеры*:*

1) *С л у ч а й н ы й о т б о р.* Допустим, у нас есть 320 шишек. Необходимо случайным образом выбрать из них 10 штук.

Присвоим каждой шишке номер - от 1 до 320. Поскольку общее количество объектов – трёхзначное число, необходимо использовать случайные числа с тремя разрядами. Например, это могут быть первые или последние три разряда в таблице случайных чисел *[Митропольский,1969]*. Кроме того, во многие научные калькуляторы встроены генераторы случайных чисел. Поскольку трехзначное случайное число даёт 1000 возможных комбинаций, а нам их необходимо только 320, каждое полученное случайное число необходимо умножить на число а = 320/1000 = 0,32.

Используя таблицу или генератор случайных чисел, получаем 10 чисел (по объёму нужной выборки). Дальнейшие операции представлены в табл. 3. Отметим, что значение $x \cdot a$ округляется в большую сторону. В последней строке таблицы представлены номера шишек, которые необходимо взять для формирования случайной выборки.

Таблица 3

Формирование выборки на основе случайных чисел

Случайное число, *x*	378	685	930	341	374	140	531	391	92	99
x·a	120,8	219,1	297,7	109,0	119,5	44,9	169,8	125,1	29,3	31,6
Номер объекта	121	220	298	109	120	45	170	126	30	32

При использовании компьютерной программы *MS Excel* расчёт случайных чисел сильно упрощается: для этих целей используется функция СЛУЧМЕЖДУ(1;320), которая даёт случайное число, заключенное между 1 и 320. При использовании такого варианта мы сразу получаем номер объекта.

2) *Регулярный отбор.* Применяется, в частности, при исследовании лесных культур. Имеется, например, участок лесных культур с количеством растений 850 шт. Нужна выборка объёмом не менее 100 шт. Таким образом, необходимо исследовать каждое восьмое дерево в ряду (850/100≈8).

3) *Расслоённый отбор.* Данный вид отбора целесообразно использовать, когда внутри совокупности существует неоднородность между объектами. Например, при исследовании плодоношения необходимо учитывать тот факт, что урожайность деревьев в значительной мере зависит от их положения в пологе древостоя, поэтому в сомкнутом насаждении целесообразно разделить деревья на подгруппы, соответствующие классам Крафта. Затем в пределах каждого класса случайным (или регулярным) образом отбираются деревья, на которых и учитывается плодоношение. Количество деревьев, отбираемых в каждом классе, целесообразно взять пропорционально общему количеству деревьев в нем.

6. ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

Конечная цель большинства научных исследований состоит в нахождении связей (зависимостей) между факторами (переменными). Любой закон природы может быть выражен, в конечном счёте, в виде описания характера или структуры взаимосвязей (зависимостей) между изучаемыми переменными величинами.

В зависимости от типа зависимых и независимых переменных различаются методы статистического анализа зависимостей (табл. 4).

Таблица 4

Применение основных методов статистического анализа в зависимости от типа исходных переменных

№ п/п	Цель статистического анализа		Типы переменных		Методы статистического анализа	
			зависимая	**независимая**	**параметрические (распределение признаков приближается к нормальному)**	**непараметрические (распределение признаков существенно отличается от нормального либо неизвестно)**
1	Оценка различий между двумя выборками по значению признака	а) выборки независимые	Количественная	Номинальная	*t*-критерий для независимых выборок	Критерий серий Вальда-Вольфовица, U-критерий Манна-Уитни, двухвыборочный критерий Колмогорова-Смирнова
			Порядковая	Номинальная	—	
			Номинальная	Номинальная	—	Критерий χ^2 (хи-квадрат) Макнемара
		б) выборки зависимые	Количественная	Номинальная	*t*-критерий для зависимых выборок	Критерий знаков, критерий Уилкоксона
			Порядковая	Номинальная	—	
2	Оценка различий между несколькими выборками по значению признака	а) выборки независимые	Количественная	Номинальная	Дисперсионный анализ	Анализ Краскела-Уоллеса, медианный тест
			Порядковая	Номинальная	—	
		б) выборки зависимые	Количественная	Номинальная	Дисперсионный анализ с повторными измерениями	Ранговый дисперсионный анализ Фридмана
			Порядковая	Номинальная	—	
3	Анализ тесноты связи между переменными		Количественная	Количественная	Корреляционный анализ (коэффициент корреляции Пирсона)	—
			Порядковая	Порядковая	—	Анализ ранговых корреляций (коэффициенты корреляции Спирмена и Кендалла)

Продолжение таблицы 4

4	Исследование вида зависимости между количественными переменными		Количественная	Количественная (одна переменная)	Регрессионный анализ а) простая линейная регрессия б) нелинейная регрессия	—
			Количественная	Количественная (несколько переменных)	Регрессионный анализ (множественная линейная регрессия)	—
			Количественная	Номинальная и количественная	Ковариационный анализ	—
5	Оценка различий двух эмпирических распределений, либо отличие эмпирического распределения от теоретического		Все типы переменных	—	Метод χ^2 (хи-квадрат), критерий Колмогорова-Смирнова	
6	Сокращение числа переменных		Количественные		Факторный анализ	
7	Отнесение объектов в заранее известные группы по совокупности количественных характеристик		Номинальная	Количественная	Дискриминантный анализ	
8	Разделение объектов на группы (неизвестные) по совокупности количественных характеристик	а) количество групп известно (задаётся исследователем)	Номинальная	Количественная	Кластерный анализ (раздел *K*-средних)	
		б) количество групп неизвестно (определяется в процессе вычислений)	Номинальная	Количественная	Кластерный анализ (древовидная кластеризация)	

В таблице отдельно приведены методы для признаков как с нормальным или близким к нормальному распределением (количественные переменные), так и с распределением существенно отличающимся от нормального (количественные и порядковые переменные). Методы, применимые к признакам с распределением, существенно отличающимся от нормального, либо для признаков с неизвестным распределением (например, в силу малого количества наблюдений), относят к непараметрическим методам статистического анализа.

Рассмотрим различия между зависимыми и независимыми выборками. Степень различия между средними значениями признака в двух (нескольких) группах зависит от внутригруппового варьирования переменных. Так, например, если средняя высота одного древостоя равна 10 м, а средняя высота другого – 12 м, то разница между ними в 2 м будет важной в случае, когда значения высот первого древостоя изменяются от 9 до 11 м, а второго – от 11 до 13 м (внутригрупповая дисперсия мала). В этом случае различия между древостоями по высоте значительны. Если же в обоих древостоях высота деревьев варьирует от 4 до 16 м (внутригрупповая дисперсия велика), то полученная разница между средними высотами, равная 2 м, будет несущественной и различия между древостоями в силу высокого внутригруппового варьирования выявить не удастся.

Зависимыми являются выборки, в состав которых входят одни и те же объекты, которые измеряются повторно (например, прирост по диаметру одних и тех же деревьев в разные годы наблюдений, количество вредителей на одних и тех же сеянцах до и после опрыскивания препаратом). Применение специальных методов статистического анализа к зависимым выборкам позволяет сравнивать именно фактор, по которым эти выборки различаются (влияние года наблюдения на прирост по диаметру, влияние опрыскивания на количество вредителей и т. п.), без учёта индивидуальных различий объектов, входящих в состав этих выборок (различия в скорости роста по диаметру отдельных деревьев, различия в устойчивости деревьев к вредителю и др.). Такой подход позволяет точнее, чем в случае с независимыми выборками, определить влияние исследуемого фактора на признак.

Независимые выборки состоят из разных объектов (либо из одних и тех же, но параллели между несколькими измерениями одного объекта не рассматриваются) и различаются по значениям интересующего фактора (высота деревьев на площадях с разными вариантами обработки почвы, количество вредителей на площадях с различной концентрацией инсектицида). Количество наблюдений в каждой из независимых выборок может быть различным, в отличие от случая с зависимыми выборками, где произ-

водится попарное сравнение значений признака для каждого объекта и количество наблюдений в рамках сравниваемых выборок, таким образом, оказывается одинаковым.

Рассмотрим подробнее назначение некоторых методов статистической обработки данных, приведенных в табл. 4.

Оценка различий между распределениями. Одной из часто встречающихся задач статистического анализа данных является сравнение распределений. Чаще всего это проверка соответствия эмпирического распределения теоретическому, например, нормальному распределению. Тем не менее, довольно часто исследователь сталкивается с необходимостью сравнить два эмпирических распределения с целью определения того, являются ли выборки взятыми из одной генеральной совокупности либо это разные совокупности и при анализе их нельзя смешивать. Например, мы исследуем зависимость высоты растений от диаметра ствола и имеем измерения этих параметров для двух участков с различной густотой. Для того, чтобы принять решение об объединении этих двух выборок в одну мы можем сравнить два эмпирических распределения высот растений. Если они достоверно не различаются – объединяем их в одну совокупность и рассчитываем интересующие нас зависимости, если же распределения достоверно различаются, то эти зависимости изучаем отдельно для каждого участка. Для оценки различий между распределениями обычно используются метод хи-квадрат (*Chi-Square test,* χ^2) и критерий Колмогорова-Смирнова (*Colmogorov-Smirnov two-sample test*).

Оценка различий между выборками. Сравнение средних значений различных выборок относится к наиболее часто применяемым методам статистического анализа. Если обе выборки подчиняются закону нормального распределения (см. Главу 7), то при сравнении двух выборок используются параметрические методы статистического анализа данных. Если же распределение выборок существенно отличается от нормального (либо зависимые переменные являются порядковыми или номинальными), используются непараметрические тесты.

При сравнении средних значений выборок выделяют четыре различные тестовые ситуации, требующих использования различающегося инструментария статистического анализа данных:

- сравнение двух независимых выборок.
- сравнение двух зависимых (спаренных) выборок
- сравнение более двух независимых выборок
- сравнение более двух зависимых выборок

Охарактеризуем эти ситуации более подробно:

Сравнение двух независимых выборок. При этом сравниваются две группы объектов, различающихся по какому-либо фактору, например, участок культур, на котором применялось удобрение, и участок без удобрений. Если распределение изучаемого признака близко к нормальному (например, высота, диаметр ствола растений на этих участках), то используется параметрический *t*-критерий для независимых выборок (тест Стьюдента). Если же распределение сильно отличается от нормального, либо признак является порядковым (например, балл плодоношения деревьев на участках) или номинальным (цвет семян), то используются непараметрические аналоги *t*-критерия (*U*-критерий Манна-Уитни, критерий Колмогорова-Смирнова и др.).

Сравнение двух зависимых (спаренных) выборок. Сравниваются в разных условиях одни и те же объекты, то есть измерения производятся повторно. Например, производится сравнение параметров одних и тех же растений до и после пересадки на лесокультурную площадь: длина хвои (количественный признак с распределением близким к нормальному), угол крепления ветвей (количественный признак с распределением, существенно отличающимся от нормального), балл плодоношения (порядковая переменная). При сравнении двух зависимых переменных и нормальном распределении признака используется *t*-критерий для зависимых выборок, в иных случаях его непараметрические аналоги (Критерий знаков, критерий Уилкоксона). Следует отметить, что к спаренным выборкам можно применять и соответствующие тесты для независимых выборок. В этом случае точность результатов будет значительно ниже, так как тесты для зависимых выборок позволяют исключить ненужные в данном случае индивидуальные различия между объектами и сосредоточить при анализе все внимание на различиях между градациями фактора (до и после пересадки) без учета других факторов, могущих внести свой вклад в варьирование

Сравнение более двух независимых выборок. Сравнивается несколько групп объектов, различающихся по одному фактору. Используя приведенный выше пример с удобрениями, мы можем иметь три участка: с фосфорным удобрением, с комплексным удобрением и без удобрений. Аналогично случаю для двух выборок выбираем между параметрическими и непараметрическими тестами. Параметрическим вариантом в данном случае является однофакторный дисперсионный анализ (п. 10.1.1). В англоязычной специальной литературе в отношении дисперсионного анализа общепринятой является аббревиатура ANOVA – ANalysis Of VAriance, непараметрическими тестами являются анализ Краскела-Уоллеса и Медианный тест (п. 10.1.2).

Сравнение более двух зависимых выборок. Измерения производятся повторно, но не два, а три и более раз. Например, производится сравнение параметров одних и тех же растений в возрасте 5, 10 и 30 лет, на основании чего можем сделать вывод о влиянии возраста растений на тот или иной признак (длина хвои, цвет коры и др.). В случае нормального распределения признака используется однофакторный дисперсионный анализ с повторными измерениями. Если распределение признака существенно отличается от нормального или он относится к порядковым переменным, используется ранговый дисперсионный анализ Фридмана.

Оценка различий между выборками по нескольким значениям. Если необходимо выполнить сравнение нескольких выборок по нескольким факторам (два и более), то используется многофакторный дисперсионный анализ (в специальной литературе используется аббревиатура MANOVA - Multivariate ANalysis Of VAriance). Допустим, нам необходимо выяснить, как влияет на скорость роста растений применение удобрений (фосфорное, калийное и комплексное) на почвах различного механического состава (песчаная, супесчаная и суглинистая). При этом мы имеем два фактора и одну зависимую переменную. Первый фактор – влияние удобрений с тремя градациями этого фактора: фосфорное, калийное и комплексное удобрение. Второй фактор – влияние механического состава почв с тремя градациями: песчаная, супесчаная и суглинистая. Зависимая переменная – высота растений. Многофакторный дисперсионный анализ (п. 10.1.1) позволяет выявить не только степень влияния каждого из факторов на скорость роста растений (например, выяснится, что влияние механического состава почвы гораздо выше влияния типа применяемого удобрения), но и оценить совместное влияние факторов. Например, если анализ покажет достоверно значимое влияние совместного фактора «удобрение-почва», то мы можем говорить о том, что на почвах разного механического состава удобрения влияют на растения поразному.

Анализ тесноты связи между переменными. Для определения уровня тесноты связи между количественными переменными, например, между высотой и диаметром ствола дерева, используется линейная корреляция Пирсона, которая относится к параметрическим методам. Если одна или несколько переменных являются порядковыми (например, взаимосвязь высоты и балла плодоношения), то используются непараметрические ранговые коэффициенты корреляции Спирмена (*Spearman rank correlation coefficient*) и Кендалла (*Kendall tau rank correlation coefficient*).

Исследование вида зависимости между переменными. Довольно часто исследователю требуется знать вид зависимости какой-либо переменной от одной или нескольких других переменных. Данная задача является предметом изучения регрессионного анализа (л. 13). Конечным итогом такого анализа является определение тесноты связи и получение уравнения зависимости одной переменной (*Y*-переменная) от одной или нескольких других переменных (*X*-переменные). В результате, зная значения *X*-переменных, мы с определенной степенью вероятности можем вычислить значение искомой *Y*-переменной. Например, необходимо по значениям диаметра и высоты ствола дерева найти его объем. Следует отметить, что для получения осмысленных результатов исследователь на основании своих знаний и опыта должен изначально представлять вид регрессионной зависимости, которую имеет смысл опробовать при помощи регрессионного анализа (линейная, полиномиальная, логарифмическая и др.). Регрессионный анализ на основании эмпирических данных позволяет рассчитать только необходимые коэффициенты уравнения, но никогда не позволит определить сам вид зависимости: такие задачи должен решать человек. В приведенном выше примере очевидно, что диаметр и высота связаны с объемом полиномиальной зависимостью типа $V=f\ D^2H$. Тем не менее, если исследователь не имеет представления о типе зависимости между переменными, в качестве крайнего варианта он может подобрать такую зависимость, просто перебирая различные виды зависимости и сравнивая, насколько точно они описывают имеющиеся данные. Но при этом нет никакой гарантии, что найденная зависимость соответствует природе исследуемых данных, а не является хорошей «подгонкой» под них в силу использования в регрессии невообразимого числа математических степеней и коэффициентов. Еще одной из сфер применения регрессионного анализа является определение того, какие из независимых *X*-переменных нужны для описания *Y*-переменной, а какие являются лишними, лишь незначительно «улучшающими» зависимость. Так, например, при изучении зависимости объема ствола от других биометрических показателей, исследователь, скорее всего, придет к выводу, что в структуре регрессионной зависимости диаметр является наиболее важной переменной, а толщина скелетных ветвей – признак, которым можно пренебречь.

Многомерные методы. Перечисленные выше методы статистического анализа используют сравнение объектов одновременно по одному-двум признакам. Если же необходимо сравнение объектов одновременно по нескольким признакам (количественным или номинальным), в зависимости от поставленных задач используются методы многомерного анализа данных. Бурное развитие этих методов в последнее время обу-

словлено резко возросшими возможностями вычислительной техники и разработкой соответствующих программных продуктов (*SPSS, SAS, Statistica, NCSS, Statgrafics* и др.). Краткий обзор некоторых многомерных методов приведен в Главе 13.

7. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Нормальное распределение – наиболее распространенный в практической биометрии тип распределения случайной величины. При таком распределении переменная величина распределена непрерывно, причём большинство её значений сосредоточено в точке арифметического среднего. Чем дальше от среднего находится значение исследуемой переменной, тем меньше вероятность её появления. Если по оси X откладывать значения переменной, а по оси Y – количество наблюдений с соответствующим значением переменной, то в случае нормального распределения получим характерную "колоколообразную" кривую. Точная форма классического нормального распределения определяется только двумя параметрами: средним арифметическим значением (обычно обозначается как μ или $\bar{x}$) и стандартным отклонением (σ). При этом характерное свойство нормального распределения состоит в том, что 68% всех его наблюдений лежат в диапазоне ± одно стандартное отклонение от среднего, диапазон ± два стандартных отклонения содержит 95% значений, а диапазон ± три стандартных отклонения – 99,7% значений.

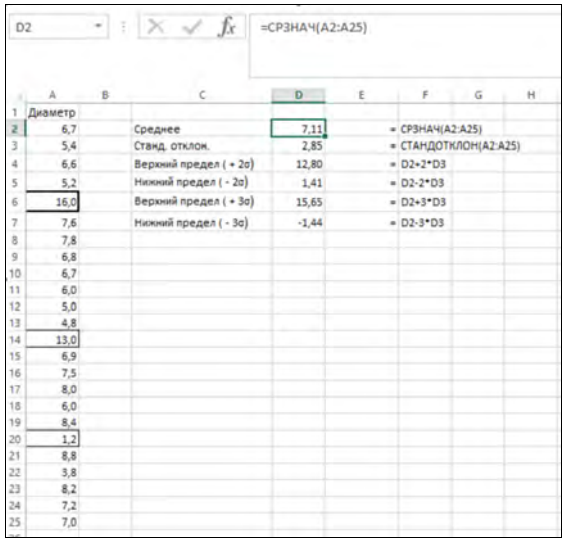
Нормальное распределение является следствием влияния на изучаемый признак множества (в идеале – бесконечного числа) независимых факторов. Например, можно предположить, что рост дерева является результатом воздействия многих независимых факторов, таких как различные генетические предрасположенности, болезни, множество воздействий внешней среды (ветер, свет, почва и т. д.) и множество других трудно учитываемых в полном объеме факторов. Как следствие, размеры деревьев (высота, диаметр) имеют тенденцию быть нормально распределёнными. Если же появляются факторы, действие которых сильно преобладает над другими, то возникают отклонения распределения от нормального. Например, в сомкнутом насаждении ощутимо возрастает влияние фактора конкуренции растений за свет, заставляющий растения тянуться в высоту за счёт снижения прироста по диаметру. В результате появляются отклонения от нормального распределения, в частности в описанном случае распределение признака становится асимметричным (высота приобретает отрицательную, а диаметр – положительную косость).

Применимость нормального распределения. Основные статистические методы (корреляционный, дисперсионный, регрессионный анализы, *t*-критерий и др.) основаны на предположении о нормальном распределении используемых количественных признаков. В случае, когда данные не являются нормально распределенными, можно выбрать один из двух вариантов. Первый: использовать альтернативные, "непараметрические" тесты (так называемые "свободно распределенные критерии", некоторые из которых приведены в п. 9 и 10). Другой вариант – использовать тесты, основанные на изначальном предположении о нормальном распределении, если есть уверенность, что объем выборки достаточно велик. Такая возможность опирается на следующий важный принцип, позволяющий понять популярность тестов, основанных на нормальном распределении признаков: при возрастании объема выборки форма выборочного распределения параметров выборки (например, среднее значение) приближается к нормальному распределению, даже если исходное распределение исследуемых переменных таковым не является.

Отбраковка вариант. Свойства нормального распределения позволяют выявить в данных резко выделяющиеся наблюдения или выбросы *(outliers)* и решать вопрос об их выбраковке (такие отклонения могут возникать в результате неточности измерений, ошибок при записи данных и т. п.).

Мы знаем, что диапазон {среднее ± 2 стандартных отклонения или $\bar{x} \pm 2\sigma$} содержит 95% значений, а диапазон {среднее ± 3 стандартных отклонения или $\bar{x} \pm 3\sigma$} содержит 99,7% значений. Поэтому, если есть уверенность в том, что распределение рассматриваемой выборки – нормальное (например, на основании предыдущего опыта наблюдений за подобными объектами), то можно принять, что варианта, выходящая за пределы $\bar{x} \pm 2\sigma$ является сомнительной и подлежит проверке, а варианта, выходящая за пределы диапазона $\bar{x} \pm 3\sigma$, с большой долей вероятности является выбросом, артефактом (вероятность естественного появления такого значения в рамках нормального распределения составляет десятые доли процента) и подлежит выбраковке.

Пример расчёта в MS Excel. Имеем данные измерений диаметра ствола у 24 деревьев ели. Необходимо определить, нет ли сильно отклоняющихся значений в данной выборке. При помощи формул, описанных выше (п.3), рассчитываем в программе *MS Excel* среднее, стандартное отклонение (сигму), и по их значениям вычисляем верхнюю и нижнюю границы разброса для 2σ и 3σ. Пример набранных данных и формулы для расчетов приведены на Рис. 8.



D2 =СРЗНАЧ(A2:A25)

	A	B	C	D	E	F	G	H
1	Диаметр							
2	6,7		Среднее	7,11		= СРЗНАЧ(A2:A25)		
3	5,4		Станд. отклон.	2,85		= СТАНДОТКЛОН(A2:A25)		
4	6,6		Верхний предел (+ 2σ)	12,80		= D2+2*D3		
5	5,2		Нижний предел (- 2σ)	1,41		= D2-2*D3		
6	16,0		Верхний предел (+ 3σ)	15,65		= D2+3*D3		
7	7,6		Нижний предел (- 3σ)	-1,44		= D2-3*D3		
8	7,8							
9	6,8							
10	6,7							
11	6,0							
12	5,0							
13	4,8							
14	13,0							
15	6,9							
16	7,5							
17	8,0							
18	6,0							
19	8,4							
20	1,2							
21	8,8							
22	3,8							
23	8,2							
24	7,2							
25	7,0							

Рис. 8. Расчет основных статистик

Таким образом, из границ $\bar{x} \pm 2\sigma$ (от 1,41 до 12,80) выбиваются три значения: 16,0; 13,0; 1,2. Правильность этих значений необходимо тщательно проверить по имеющимся полевым данным. Из границ $\bar{x} \pm 3\sigma$ (теоретические рамки от -1,44 до 15,65) выбивается значение 16,0, которое с высокой степенью вероятности является выбросом, поэтому целесообразно его исключить. Следует отметить, что после выбраковки вариант значения основных параметров распределения (среднее, стандартное отклонение и др.) следует пересчитывать заново, так как корректировка исходных данных закономерно изменяет и значения основных параметров распределения. При изменении данных пересчет формул на листе *MS Excel* по умолчанию выполняется автоматически. Тем не менее, полезно учитывать (особенно при работе с генератором случайных чисел по формулам типа СЛЧИС), что для пересчета всех формул листа в *MS Excel* традиционно используется функциональная клавиша F9.

8. ПРОВЕРКА РАСПРЕДЕЛЕНИЯ

Проверка соответствия распределения, полученного в опыте (эмпирическое распределение), предполагаемому (теоретическое распределение) производится на основе применения *критериев согласия (goodness of fit)*. При этом проверяется гипотеза о том, что эмпирическое и теоретическое

распределения согласуются друг с другом, а наблюдаемые расхождения являются несущественными и связаны с влиянием случайных факторов. Обычно критерии согласия используются для того, чтобы оценить, насколько сильно отличается наблюдаемое в опыте распределение от нормального, но данные критерии могут использоваться для сравнения любых других рядов распределения. Например, можно сравнивать не эмпирическое и теоретическое распределения, а два эмпирических ряда распределения с целью выявления взяты ли эти ряды из одной генеральной совокупности (в этом случае различия между ними будут незначимы) или из разных (различия рядов распределения достоверны).

В практике статистического анализа наиболее часто используется критерий согласия χ^2 (хи-квадрат), предложенный К. Пирсоном. Критерий χ^2 рассчитывается следующим образом:

$$\chi^2 = \sum \frac{(n - \tilde{n})^2}{\tilde{n}},$$

где n – экспериментальная частота, $\tilde{n}$ – теоретическая частота.

Пример. Имеются данные измерений диаметров 91 дерева сосны, сгруппированные в 4-сантиметровые ступени толщины:

Ступень толщины, см	20	24	28	32	36	40	44
Наблюдаемая частота (эмпирическая), n	7	8	28	20	18	9	1
Теоретическая частота (нормальное распределение), $\tilde{n}$	6,3	15	23,2	23,4	15,4	6,6	1,8

Необходимо оценить, отличается ли распределение данной выборки от нормального. Частоты нормального распределения (теоретические) приведены выше.

Критерий χ^2 удобно рассчитывать по следующей таблице:

Диаметр, см	Частота		$n - \tilde{n}$	$\frac{(n - \tilde{n})^2}{\tilde{n}}$
	n	$\tilde{n}$		
20	7	6,3	0,7	0,08
24	8	15	-7,0	3,27
28	28	23,2	4,8	0,99
32	20	23,4	-3,4	0,49
36	18	15,4	2,6	0,44
40	9	6,6	2,4	0,87
44	1	1,8	-0,8	0,36
Σ	91	91,7	-0,7	$\chi^2 = 6,50$

Далее оцениваем значимость полученного критерия χ^2 (6,50) на принятом уровне (обычно 0,05) и числе степеней свободы, равном числу классов. Значимость критерия χ^2 можно оценить по стандартным таблицам, имеющимся в руководствах по статистике, либо при помощи функции ХИ2РАСП(), которую можно найти в программе *MS Excel*. Аргументами данной функции являются значение критерия (6,50) и число степеней свободы (6). Результат расчёта функции – значение уровня значимости. В нашем случае он равен 0,37. Поскольку это значение превышает 0,05, можем говорить о том, что различия между эмпирическим (наблюдаемым) и теоретическим (нормальным) распределением недостоверны. Следовательно, распределение нашей выборки диаметров деревьев сосны можно считать нормальным и, соответственно, применять при обработке таких данных стандартные параметрические методы.

Пример расчёта в MS Excel. Сравнение наблюдаемого и ожидаемого распределения в *MS Excel* производится при помощи функции ХИ2ТЕСТ(). Аргументами данной функции являются фактический (наблюдаемый) и ожидаемый (теоретический) интервалы с частотами (рис. 9). Результатом расчёта функции является показатель уровня значимости (0,37).

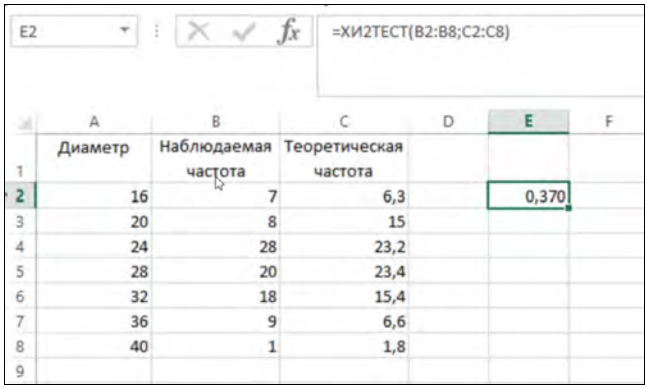


Рис. 9. Сравнение распределений на основе теста Хи-квадрат в MS Excel

Пример расчёта критерия χ^2 в Statistica. Расчёт производится в модуле *Nonparametrics* (непараметрические статистики) меню *Statistics*. Данные выстраиваются по столбцам: отдельно наблюдаемые и теоретические частоты. В стартовой панели *(startup panel)* выбираем *Observed versus expected X?* (наблюдаемые в сравнении с ожидаемыми). Кнопкой *Variables*

выбираем переменные с наблюдаемыми *(observed)* и ожидаемыми *(expected)* частотами (рис. 10) и запускаем процедуру анализа (OK).

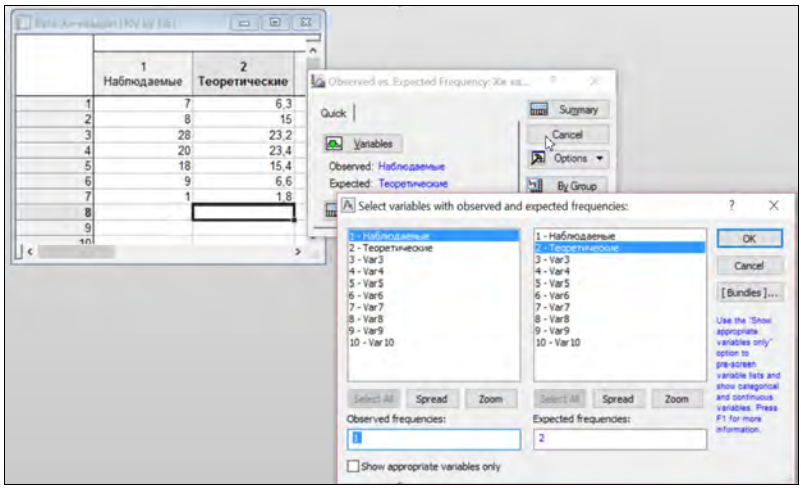


Рис. 10. Сравнение распределений на основе теста хи-квадрат в Statistica

Результатом является таблица (рис. 11), в которой помимо частот, расчёт которых приведён выше, приведено значение критерия χ^2 (*Chi-Square* = 6,49) и его уровень значимости (p = 0,370). Так как уровень значимости превышает 0,05, отличие наблюдаемого распределения от нормального недостоверно.

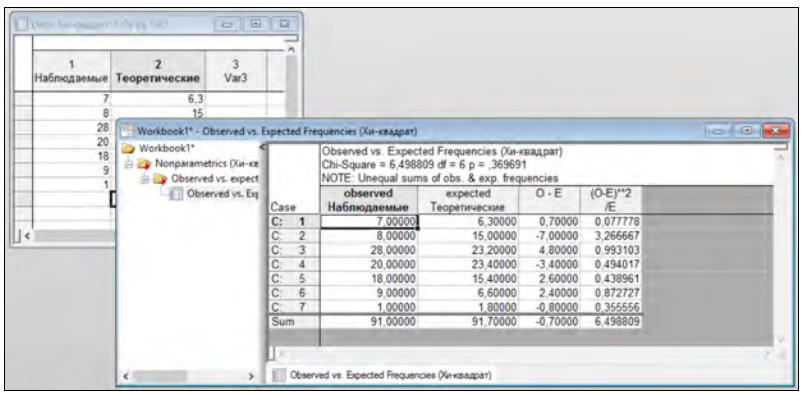


Рис. 11. Результаты сравнения распределений

Проверка распределения в Statistica. Простым и наглядным методом оценки соответствия эмпирического распределения теоретическому нормальному распределению является использование построения графиков пакета *Statistica*. Рассмотрим следующий пример: допустим, мы имеем данные измерений диаметров 244 деревьев сосны. Необходимо оценить соответствие этой выборки модели нормального распределения. Вводим данные в рабочий лист *Statistica* аналогично примеру в главе 2 «*Расчёт основных статистик в программе Statistica*». После ввода данных в меню *Statistics* в модуле *Distribution fitting* (соответствие распределения) *выбираем* требуемое распределение *(в нашем случае это нормальное распределение - Normal*), подтверждаем выбор - OK (рис. 12), и кнопкой *Variable* выделяем переменную, которую проверяем на нормальность распределения (диаметр). Далее нажимаем кнопку *Plot of observed and expected distribution (*график наблюдаемого и теоретического распределения*)*. Результатом расчётов является гистограмма (рис. 13), на которой показано экспериментальное распределение частот (столбики), кривая нормального распределения (линия), а в заголовке – данные, среди которых тип теоретического распределения (*Normal*) значение критерия χ^2 *(Chi-Square)* с числом степеней свободы (df) и уровень значимости данного критерия *(p)*.

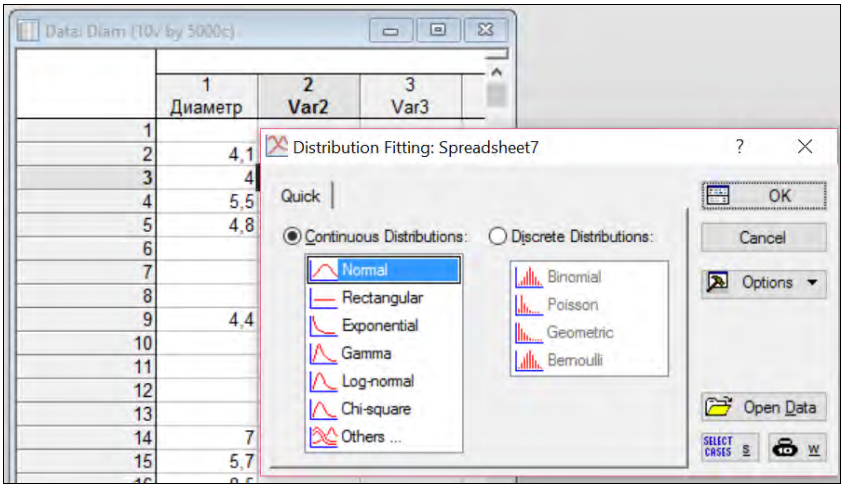


Рис. 12. Проверка соответствия теоретическому распределению

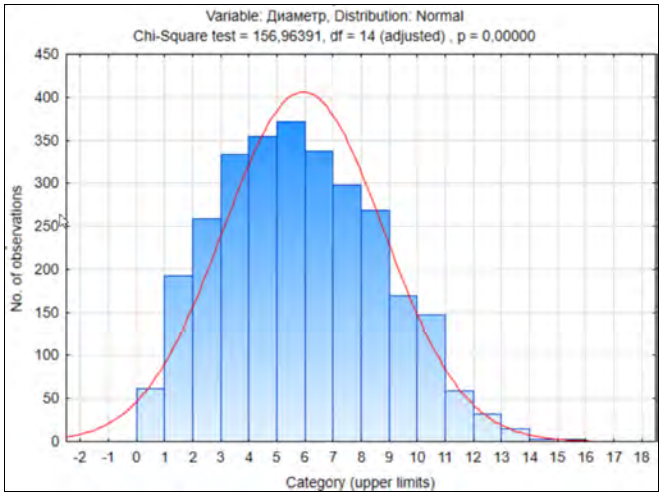


Рис. 13. Гистограмма распределения частот

В связи с тем, что полученное значение уровня значимости (менее 0,001) ниже, чем пороговое значение 0,05, можем говорить о том, что эмпирическое распределение выборки за счет высокого количества наблюдений достоверно отличается от нормального распределения и необходимо выяснять возможную причину такого отклонения. В частности, в приведенном примере мы наблюдаем выраженную положительную асимметрию (смещение ряда влево с пологим правым плечом) и отрицательный эксцесс («плоская» вершина кривой) распределения деревьев по диаметру ствол. Очевидно, источник таких отклонений необходимо устанавливать на основе более подробных дальнейших исследований.

9. СРАВНЕНИЕ ДВУХ ВЫБОРОК

9.1 Независимые выборки

9.1.1. Параметрические тесты

В отношении независимых выборок *t-критерий (t-test for independent samples)* обычно используется при сравнении средних значений двух выборок, в тех случаях, когда не производится попарного сравнения двух объектов как в случае с зависимыми выборками. При таком подходе в подавляющем большинстве случае выборки состоят из разных объектов и различаются по значениям интересующего фактора (высота культур на

площади с уходом и без него, M_{1000} семян в двух лесосеменных районах и т. п.). Фактором (независимой переменной) является признак, по которому различаются выборки (наличие ухода, лесосеменной район), его градациями – значение фактора для каждой из двух выборок (уход есть, ухода нет; западный лесосеменной район, восточный лесосеменной район). В свою очередь зависимой переменной является изучаемый признак (высота, M_{1000} семян). Строго говоря, *t*-критерий пригоден только в том случае, когда распределение признака нормально, хотя, к счастью, довольно устойчив к отклонениям от нормальности.

Для независимых выборок *t*-критерий рассчитывается следующим образом:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{m_1^2 + m_2^2}},$$

где $\bar{x}_1$ и $\bar{x}_2$ - значения сравниваемых средних арифметических;
m_1 и m_2 – соответствующие величины ошибок средних арифметических.

Полученное значение *t*-критерия сравнивается с табличным при числе степеней свободы *n1+n2-2* (*n1* – количество наблюдений в первой выборке; *n2* - количество наблюдений во второй выборке) и принятом уровне значимости (обычно принимается 5%-ный уровень значимости) и определяется достоверность различий средних арифметических двух выборок.

Пример. Рассмотрим расчёт *t*-критерия для независимых выборок на примере диаметров растений. Растения первой выборки выращивались без применения удобрений, растения второй выборки получали подкормку удобрениями:

Диаметр, см вариант 1 (с удобрениями)	11, 12, 9, 11, 10, 14, 14, 16, 12, 10, 15
Диаметр, см вариант 2 (без удобрений)	7, 11, 6, 9, 8, 13, 12, 14, 11, 9, 11, 10, 9

Рассчитываем средние арифметические:

$\bar{x}_1$ = (11+12+9+11+10+14+14+16+12+10+15)/11 = 12,2 см;

$\bar{x}_2$ = (7+11+6+9+8+13+12+14+11+9+11+10+9)/13 = 10,0 см.

Рассчитываем стандартные отклонения (см. п.1):

$$\sigma_1 = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} =$$

$$= \sqrt{\frac{(11^2+12^2+9^2+11^2+10^2+14^2+14^2+16^2+12^2+10^2+15^2) - \frac{(11+12+9+11+10+14+14+16+12+10+15)^2}{11}}{11-1}} =$$

$$= 2{,}27,$$

Таким же образом рассчитываем отклонение для второй выборки

$$\sigma_2 = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}} = 2{,}31$$

Рассчитываем ошибки средних арифметических (см. п.3, ошибка параметра):

$$m_1 = \frac{\sigma_1}{\sqrt{n_1}} = \frac{2{,}27}{\sqrt{11}} = 0{,}6851; \qquad m_2 = \frac{\sigma_2}{\sqrt{n_2}} = \frac{2{,}31}{\sqrt{13}} = 0{,}6407$$

Расчетное значение *t*-критерия составляет:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{m_1^2 + m_2^2}} = \frac{12{,}2 - 10{,}0}{\sqrt{0{,}6851^2 + 0{,}6407^2}} = \frac{2{,}2}{\sqrt{0{,}6851^2 + 0{,}6407^2}} = 2{,}35\,.$$

Теоретическое значение *t*-критерия можно найти в статистических таблицах или воспользоваться компьютерными программами. В частности, в *MS Excel* для этого используется функция СТЬЮДРАСПОБР(), аргументами которой являются вероятность (0,05) и количество степеней свободы (в нашем случае число степеней свободы составляет $n_1+n_2-2 = 22$. Расчет приведенной функции в *MS Excel* дает теоретическое значение *t*-критерия равное 2,07.

В *Statistca* теоретическое значение *t*-критерия можно найти при помощи модуля *Probability calculator* (калькулятор вероятностей) в меню *Statistics* (статистики). В *Probability calculator* выбираем строку *Distributions* (распределения*) и* далее распределение – *t (Student).* В соответствующие поля диалогового окна вводим количество степеней свободы *df* и значение уровня значимости *p* (0,05). Кроме того, отмечаем галочками опции *1-Cumulative p* и *Two-tailed* (рис. 14). После нажатия кнопки *Compute* (вычислить) в окне *t* появляется искомое значение *t*-критерия (рис. 14).

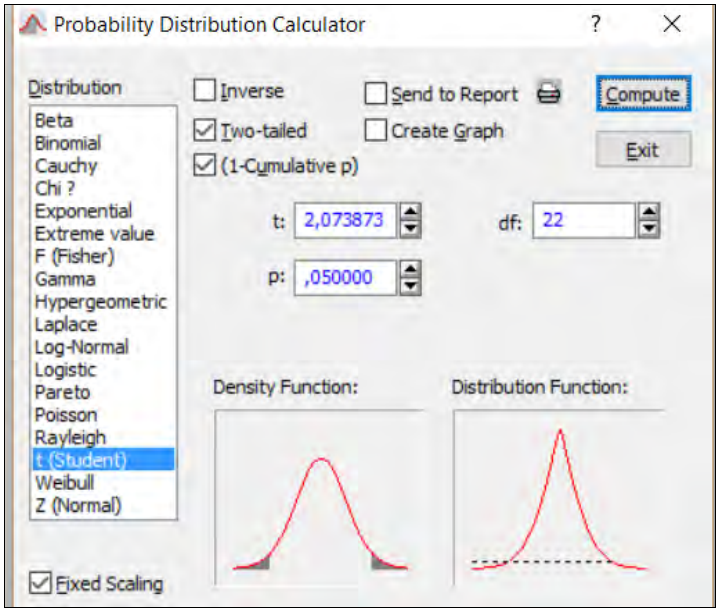


Рис. 14. Диалоговое окно расчета основных параметров распределения

Для нашего примера табличное значение *t*-критерия на 5%-ном уровне значимости при числе степеней свободы $n_1+n_2-2 = 11+13-2 = 22$ составляет 2,07. Рассчитанное значение (2,34) превышает табличное, поэтому различия между выборками достоверны на 5%-ном уровне значимости, другими словами, деревья варианта 1 достоверно отличаются в большую сторону по диаметру от деревьев варианта 2.

Расчёт в MS Excel производится при помощи надстройки "Анализ данных", являющейся специализированным инструментом *MS Excel* для статистической обработки и предоставляющей широкий выбор инструментов анализа данных. Если у Вас в *MS Excel* не установлен данный пакет – Откройте вкладку «Файл» и выберите пункт «Параметры». Далее выберите команду «Надстройки», а затем в поле «Неактивные надстройки приложений» выберите пункт «Пакет анализа» и далее – кнопка «Перейти» (рис. 15).

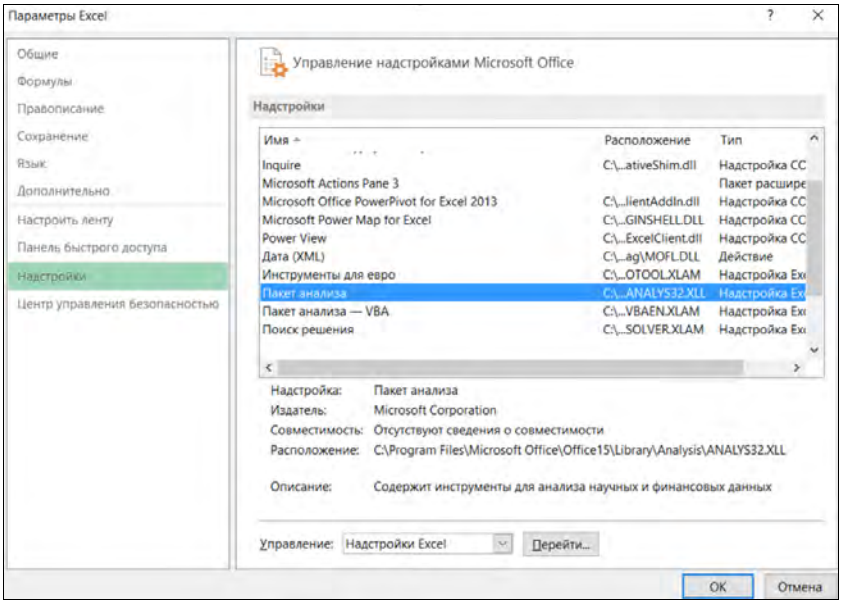


Рис. 15. Меню подключения надстройки "Анализ данных"

Далее в окне «Доступные надстройки» установите флажок «Пакет анализа» и нажмите кнопку ОК. Теперь в основном меню «Данные» появится пункт «Анализ данных» при нажатии на который выводится список инструментов статистического анализа, пригодных для использования при работе с экспериментальными данными (рис. 16). В частности, для нашего случая выбираем строку "Двухвыборочный *t*-тест с различными дисперсиями" и в появившемся диалоговом окне (рис. 17) вводим ссылку на диапазоны с переменными (интервал переменной 1 и интервал переменной 2), а также указываем куда записать результаты анализа (интервал, новый лист или книга).

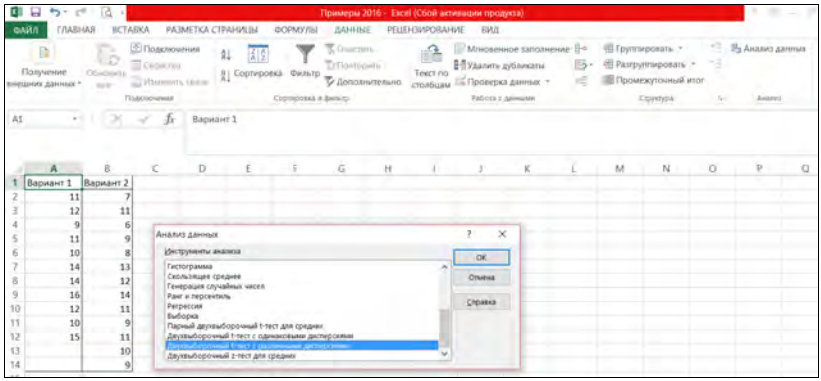


Рис. 16. Диалоговое окно выбора t-критерия с различными дисперсиями

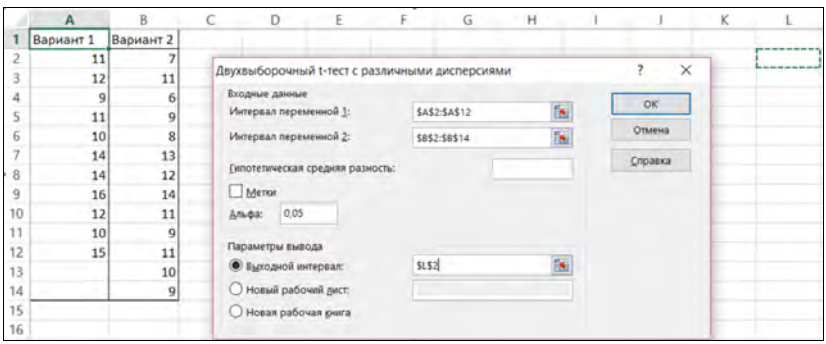


Рис. 17 Диалоговое окно ввода основных параметров расчета t-критерия

Результатом является таблица (рис. 18), в которой указано расчётное значение *t*–критерия (2,33), критическое ("табличное") значение *t*–критерия двухстороннего (2,07), уровень значимости *P* двухстороннего *t*–критерия (0,03). Рассчитанное значение (2,33) превышает критическое, а уровень значимости *P* (0,03) менее 5%, поэтому различия между выборками достоверны на 5%-ном уровне значимости, то есть среднее значение исследуемой переменной варианта 1 достоверно отличается в большую сторону от среднего значения переменной варианта 2. Различия между полученными различными способами значениями *t*–критерия на уровне второго знака после запятой (2,34 и 2,33) объясняются ошибками округления при расчетах и кардинальным образом не влияют на выводы.

Двухвыборочный t-тест с различными дисперсиями		
	Переменная 1	*Переменная 2*
Среднее	12,18181818	10
Дисперсия	5,163636364	5,333333333
Наблюдения	11	13
Гипотетическая разность средних	0	
df	21	
t-статистика	2,326251769	
P(T<=t) одностороннее	0,015048693	
t критическое одностороннее	1,720742903	
P(T<=t) двухстороннее	0,030097387	
t критическое двухстороннее	2,079613845	

Рис. 18. Результаты расчета двухстороннего t–критерия

Расчёт в Statistica. В отличие от *MS Excel*, данные в статистических пакетах, в том числе в *Statistica*, принято выстраивать следующим образом: первая переменная – независимая (группирующая), то есть номер выборки (1 или 2), вторая – зависимая переменная, т.е. диаметры (рис. 19). В меню *Statistics в* модуле *Basic statistics / Tables* (основные статистики / таблицы) выбираем раздел *t-test, independent, by groups (*t-критерий, независимые выборки, по группам*)*. Кнопкой *Variables* выбираем переменные: *grouping* (группирующая) – номер выборки, *dependent* (зависимая) – диаметр (рис. 20).

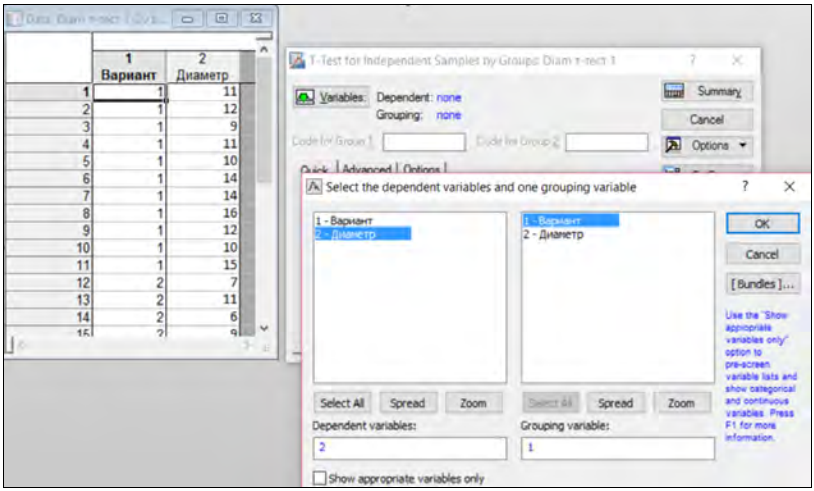


Рис. 19. Фрагмент рабочего окна при расчете t-критерия

Результатом анализа является таблица (рис. 20), в которой наиболее важным показателем является значение уровня значимости *(p)* равное 0,0298. Поскольку это значение меньше 0,05, то расчётное значение *t*-критерия (2,32) достоверно на 5%-ном уровне значимости, что подтверждает также и автоматическая заливка табличных данных красным цветом.

Variable	T-tests; Grouping: Вариант (Diam т-тест 1) Group 1: 1 Group 2: 2				
	Mean 1	Mean 2	t-value	df	p
Диаметр	12,18182	10,00000	2,322978	22	0,029821

Рис. 20. Результаты расчета t- критерия

Следует отметить, что если мы в качестве порога достоверности возьмём уровень значимости 0,01, то приведенные различия будут недостоверными (p>0,01). Таким образом, вывод по результатам анализа зависит от принятого уровня значимости. Так как в биологических исследованиях принят уровень значимости 0,05, то делаем вывод о том, что средние рассматриваемых выборок достоверно (на 5%-ном уровне значимости) различаются.

9.1.2. Непараметрические тесты

При неизвестном распределении признаков либо распределении, существенно отличающемся от нормального, целесообразно использовать непараметрические методы (методы, свободные от распределения). Непараметрические критерии можно использовать как для количественных, так и для порядковых (ранговых) переменных.

Критерий Манна-Уитни *(Mann-Whitney test)* является наиболее мощным непараметрическим аналогом *t*-критерия для независимых выборок. При этом зависимый признак относится порядковым переменным (ранги). Можно использовать и количественные признаки при условии малой выборки и (или) сильно отклоняющемся от нормального распределении. В этом случае при выполнении анализа количественный признак предварительно пересчитывается программой в порядковый (см. п. 2). Следует помнить, что при переводе количественной шкалы измерения в ранговую теряется часть информации, заключённой в переменных, поэтому при имеющейся возможности (распределение признака близко к нормальному) по отношению к количественным признакам целесообразно все-таки использовать стандартные параметрические методы (*t*-критерий).

U-критерий Манна-Уитни рассчитывается следующим образом:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1,$$

или

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2,$$

где R_1 – сумма рангов, относящихся к группе с объёмом выборки n_1;
R_2 - сумма рангов, относящихся к группе с объёмом выборки n_2.

Приведённые формулы связаны соотношением: $U=n_1n_2-U'$. Это соотношение можно использовать для проверки правильности расчётов. Полученное значение *U*-критерия сравнивается с критическим при выбранном уровне значимости. Если вычисленное значение больше или меньше табличных значений, то различия между выборками считаются достоверными.

Пример расчёта в Statistica. Рассмотрим влияние ухода на сохранность культур ели: вариант 1 – культуры без ухода, вариант 2 – культуры с уходом. В таблице приведены результаты измерений сохранности по отдельным учетным площадкам, сгруппированным в эти варианты:

Сохранность, % вариант 1 (без ухода)	17, 11, 34, 28, 24, 13, 46, 20, 25
Сохранность, % вариант 2 (с уходом)	25, 79, 61, 82, 40, 16, 72, 90, 58, 65, 58

Выстраиваем данные аналогично предыдущему примеру для t-критерия, то есть первая переменная – независимая, где указываются варианты обработки - номер выборки (1 или 2), вторая – зависимая переменная, т.е. сохранность растений по отдельным пробам (рис. 21). В меню *Statistics в* модуле *Nonparametrics* (непараметрические статистики) выбираем раздел *Comparing two independent samples (groups)* - Сравнение двух независимых выборок (по группам). После подтверждения выбора (OK) кнопкой *Variables* выбираем переменные: *grouping* (группирующая) – номер выборки (в нашем случае вариант ухода), *dependent* (зависимая) – сохранность (рис. 22).

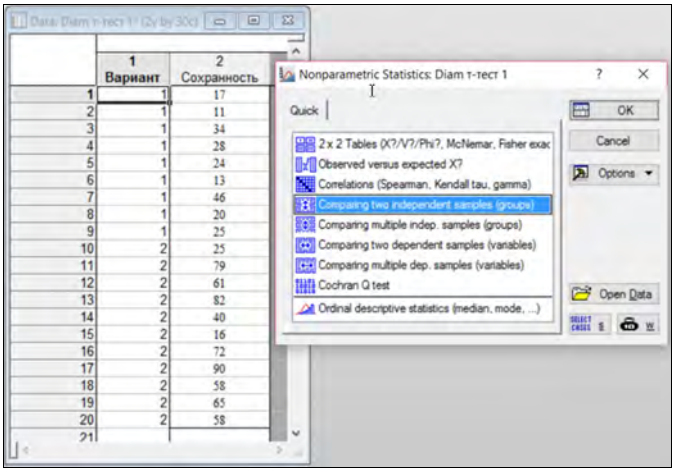


Рис. 21. Выбор меню для сравнения двух независимых выборок

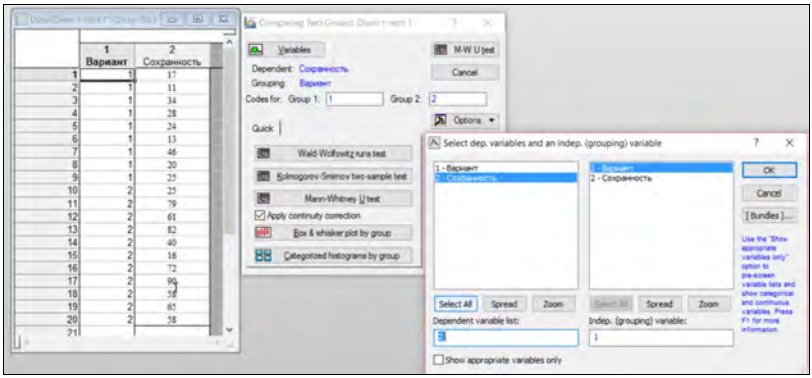


Рис. 22. Выбор переменных для выполнения сравнения двух независимых выборок

Далее в меню анализа выбираем метод обработки данных: *Mann-Whitney U-test*. Результатом является таблица (рис. 23), в которой, прежде всего, обращаем внимание на значение уровня значимости *p-value* (0,004).

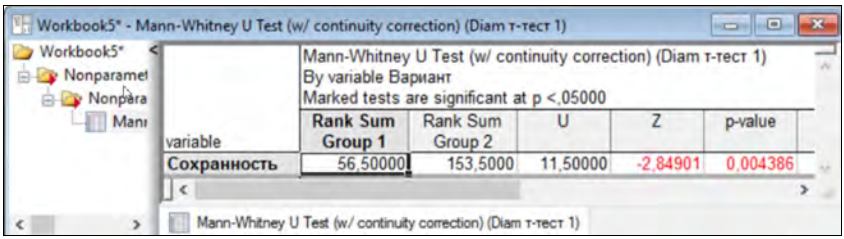


Рис. 23. Результаты расчета теста Манна-Уитни

Поскольку полученное значение *p-value* не превышает 0,05, то можем говорить о том, что значение *U*-критерия (11,5) достоверно на 5%-ном уровне значимости (а также на уровне значимости 1%), следовательно, сохранность культур ели, в которых проводился уход (вариант 2), достоверно превышает сохранность в культурах ели без ухода (вариант 1), то есть влияние ухода оказалось статистически достоверным на принятом уровне значимости 0,05.

9.2. Зависимые выборки

9.2.1. Параметрические тесты

Для зависимых выборок *(t-test for dependent samples)* *t*-критерий предназначен для определения степени различий между средними значениями двух *зависимых* выборок при распределении изучаемого показателя, близком к нормальному. Использование зависимых выборок полезно в тех случаях, когда важный источник внутригрупповой вариации может быть определен и исключен из анализа. Это относится к экспериментам, в которых две сравниваемые группы состоят из одних и тех же объектов, которые тестировались дважды. Например, на одних и те же деревьях сосны в разные годы производится измерение годичных приростов по высоте. Необходимо оценить, различаются ли разные годы по интенсивности семеношения деревьев, т.е. сравнивается две выборки (группы): средняя интенсивность семеношения в один год (первая выборка) и средняя интенсивность - в другой год (первая выборка). В данном случае значительная часть изменчивости внутри каждой группы может быть объяснена индивидуальными различиями деревьев. Если та же самая выборка тестируется дважды, то можно легко исключить эту часть изменчивости. Вместо исследования каждой группы отдельно и анализа исходных значений, можно рассматривать разности между двумя измерениями для каждого дерева. Вычитая первые значения из вторых (семеношение дерева в первый год минус семеношение этого же дерева во второй год) и анализируя затем только эти "чистые (парные) разности", мы исключаем ту часть вариации, которая является результатом различия в исходных уровнях значений признака у индивидуумов. Именно так и проводятся вычисления в *t*-критерии для зависимых выборок. В сравнении с *t*-критерием для независимых выборок такой подход дает более точный результат (критерий становится более чувствительным). Фактор (зависимая переменная), его градации и независимая переменная определяются по тому же принципу, что и в описанном выше примере по использованию *t*-критерия для независимых выборок.

Для зависимых выборок *t*-критерий рассчитывается в следующей последовательности:

1) рассчитываются попарные различия между значениями переменных исследуемых выборок:

$$z_i = x_i - y_i,$$

где x_i – значение переменной для i-го объекта первой выборки, y_i – соответствующее значение переменной для i- го объекта второй выборки;

2) *t*-критерий равен

$$t = \frac{\bar{z}}{m_z},$$

где $\bar{z}$ - среднее значение переменной *z*, m_z – ошибка среднего для *z*.

Полученное значение *t*-критерия сравнивается с табличным при числе степеней свободы *n*-1. Если полученное значение превышает табличное, различие между средними значениями зависимых выборок считается достоверным на выбранном уровне значимости (в биологии обычно используется уровень значимости 0,05).

Пример.

Рассчитаем, различаются ли два разных года наблюдения по величине годичных приростов в высоту деревьев ели. Поскольку приросты измеряются у одних и тех же деревьев и сравниваются попарно, используем *t*-критерий для зависимых выборок. В таблице исходных данных сразу рассчитываем различия между значениями переменных исследуемых выборок (*z*).

Номер дерева	1	2	3	4	5	6	7	8	9	10	11	12	Среднее
Прирост в возрасте 10 лет, x_i	26	38	19	28	28	17	22	30	24	38	21	10	25,1
Прирост в возрасте 11 лет, y_i	20	38	18	25	42	20	14	26	20	29	15	5	22,7
z_i	6	0	1	3	-14	-3	8	4	4	9	6	5	2,4

Рассчитываем стандартное отклонение (см. п.3, основные статистики) для параметра *z* составляет:

$$\sigma_Z = \sqrt{\frac{\sum z_i^2 - \frac{(\sum z_i)^2}{n}}{n-1}} =$$

$$= \sqrt{\frac{(6^2+0^2+1^2+3^2+(-14)^2+(-3)^2+8^2+4^2+4^2+9^2+6^2+5^2) - \frac{(6+0+1+3-14-3+8+4+4+9+6+5)^2}{12}}{12-1}} =$$

$$= 6{,}171.$$

Ошибка *z* составляет: $m_Z = \frac{\sigma_Z}{\sqrt{n}} = \frac{6{,}171}{\sqrt{12}} = 1{,}78$.

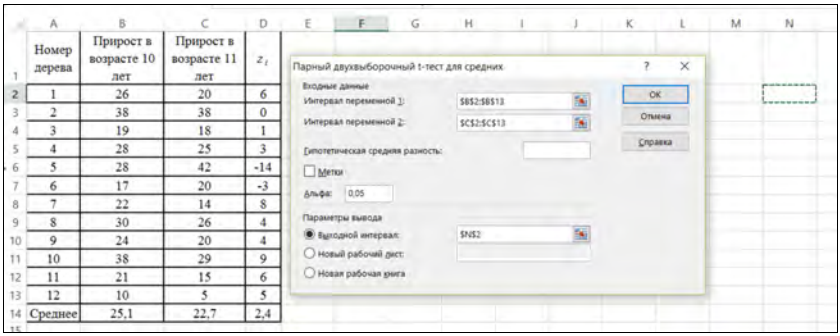
Значение *t*-критерия для зависимых выборок равно:

$$t = \frac{\bar{z}}{m_z} = \frac{2{,}4}{1{,}78} = 1{,}35.$$

Табличное значение *t*-критерия при уровне значимости 0,05 и числе степеней свободы *n*-1 = 11 составляет 2,20. Расчётное значение *t*-критерия (1,35) меньше табличного, следовательно, между выборками нет достовер-

ных различий на принятом уровне значимости. То есть значения приростов в высоту деревьев ели в возрасте 10 и 11 лет достоверно не различаются.

Расчёт в MS Excel производится при помощи описанного выше пакета "Анализ данных", запускаемого из меню "Данные". В списке "Инструменты анализа" выбираем строку "Парный двухвыборочный *t*-тест для средних" и в появившемся диалоговом окне, также как в *t*-тесте для независимых выборок, вводим ссылку на диапазоны с переменными и параметры вывода результатов (рис. 24).



Номер дерева	Прирост в возрасте 10 лет	Прирост в возрасте 11 лет	z_i
1	26	20	6
2	38	38	0
3	19	18	1
4	28	25	3
5	28	42	-14
6	17	20	-3
7	22	14	8
8	30	26	4
9	24	20	4
10	38	29	9
11	21	15	6
12	10	5	5
Среднее	25,1	22,7	2,4

Рис. 24. Окно выбора параметров расчета t-критерия для зависимых выборок

В результатах анализа (рис. 25) обращаем внимание на значение двухстороннего уровня значимости *P (P(T<=t) двухстороннее), равное* 0,20. Поскольку это значение превышает 0,05 (5%), приведенное в табличке значение *t*-критерия (1,36) недостоверно и различия между средними двух выборок считаются недостоверными на 5%-ном уровне значимости.

Парный двухвыборочный t-тест для средних		
	Переменная 1	*Переменная 2*
Среднее	25,08333	22,66667
Дисперсия	66,62879	104,9697
Наблюдения	12	12
Корреляция Пирсона	0,798248	
Гипотетическая разность средних	0	
df	11	
t-статистика	1,356563	
P(T<=t) одностороннее	0,101054	
t критическое одностороннее	1,795885	
P(T<=t) двухстороннее	0,202108	
t критическое двухстороннее	2,200985	

Рис. 25. Результаты расчета t-критерия для зависимых выборок

Расчёт в Statistica. Данные, в отличие от ранее рассмотренных вариантов с независимыми выборками выстраиваются в два столбца, каждая выборка - в отдельный столбец. Это делается для получения возможности однозначно соотнести разные измерения (различные столбцы) для одного объекта (один объект – одна строка) – рис. 26.

Процедуру расчета запускаем при помощи меню *Statistics*, далее выбираем модуль *Basic statistics / Tables* и запускаем строку *t-test, dependent samples* (рис. 26).

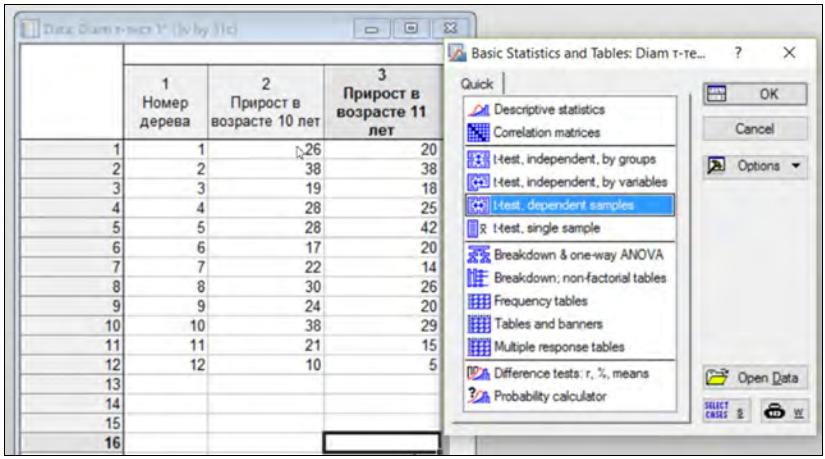


Рис. 26. Выбор меню для сравнения двух зависимых выборок

После подтверждения выбора (OK) кнопкой *Variables* определяем зависимые выборки: First variable list и Second variable list. Соответственно, выбираем переменные 2 и 3 (рис. 27) и запускаем анализ кнопкой Summary.

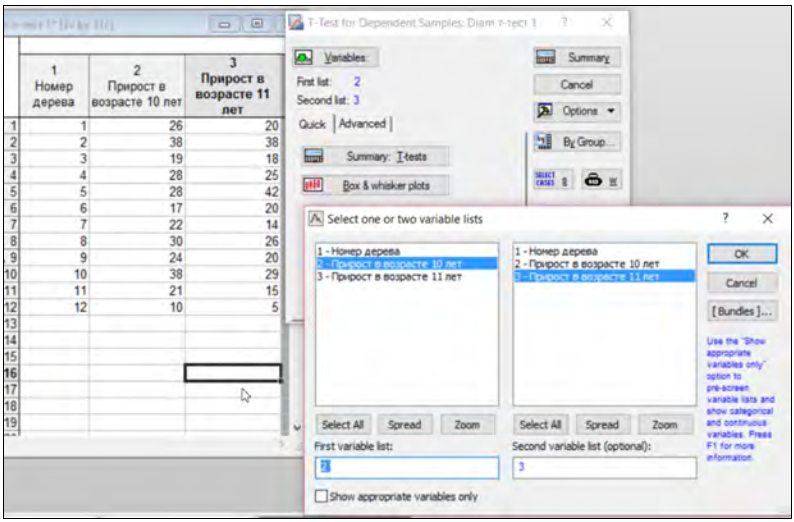


Рис. 27. Выбор переменных для выполнения сравнения двух зависимых выборок

Результатом является таблица (рис. 28), в которой уровень значимости *(p)* составляет 0,20, что превышает 0,05, следовательно, значение полученного *t*-критерия (1,356) является недостоверным на 5%-ном уровне значимости.

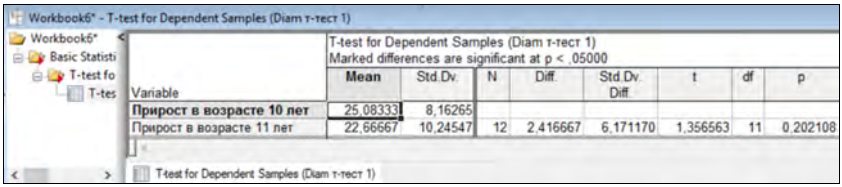


Рис. 28. Результаты расчета t-критерия для зависимых выборок

Таким образом, результаты анализа говорят о том, что значения приростов деревьев ели в возрасте 10 и 11 лет не различаются на 5%-ном уровне значимости.

9.2.2. Непараметрические тесты

Критерий Уилкоксона *(Wilcoxon test)* является непараметрическим аналогом *t*-критерия для зависимых выборок. Зависимый признак при этом выражается в количественной или порядковой (ранги) шкалах.

Пример расчёта в Statistica. Имеются данные о среднем количестве семян в шишке для деревьев сосны в разные годы:

Номер дерева	1	2	3	4	5	6	7	8	9	10	11
Среднее кол-во семян в шишке, шт (1-й год наблюдения)	10	14	22	28	13	17	32	27	30	25	10
Среднее кол-во семян в шишке, шт (2-й год наблюдения)	13	26	28	30	10	35	27	33	35	26	15

Поскольку нет данных о том, что переменные распределены по закону нормального распределения, используем непараметрический тест Уилкоксона, для чего запускаем в меню *Statistics* модуль *Nonparametrics* (непараметрические статистики) и выбираем раздел *Comparing two dependent samples (variables)* - Сравнение двух зависимых выборок (по переменным).

После подтверждения выбора (OK) кнопкой *Variables* определяем зависимые выборки: First variable list и Second variable list. Соответственно, выбираем переменные 2 и 3 (рис. 29) и запускаем анализ кнопкой *Wilcoxon matched pairs test (*парный тест Уилкоксона*)*.

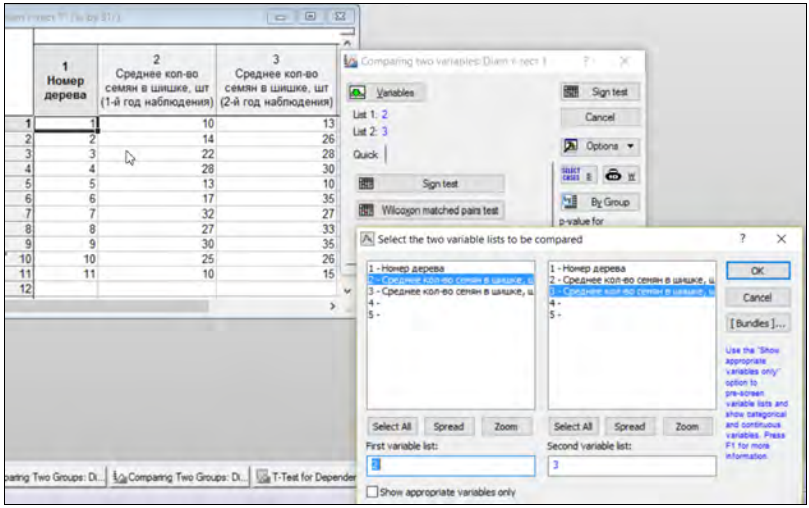


Рис. 29. Выбор переменных для выполнения непараметрического критерия Уилкоксона (сравнение двух зависимых выборок)

В таблице результата (рис. 30), сравниваем значение уровня значимости (p-value) с принятым (5%).

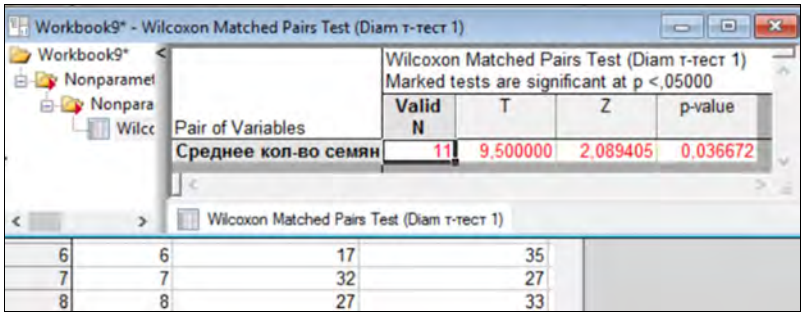


Рис. 30. Результаты расчета непараметрического критерия Уилкоксона

Так как полученный уровень значимости (0,037) меньше, чем принятый (0,05), можно говорить о том, что деревья в первый и второй годы наблюдений достоверно различаются по среднему количеству семян в шишках. Достоверность различий также, как и в предыдущих примерах подтверждается красным цветом шрифта результатов анализа данных.

9. СРАВНЕНИЕ НЕСКОЛЬКИХ ВЫБОРОК

10.1. Независимые выборки

10.1.1. Параметрические тесты

Дисперсионный анализ (ANOVA – ANalysis Of VAriances). Сравнение средних на основе *t*-критерия является удобным методом в том случае, когда имеются две выборки, различающиеся по градациям одного фактора. Если же градаций фактора много (сравнивается более двух выборок), то эффективным инструментом является дисперсионный анализ. Сущность его заключается в сравнении варьирования (дисперсии) переменной, которое обусловлено влиянием изучаемого фактора, с варьированием признака, обусловленным случайными факторами. Влияние фактора оценивается по величине межгрупповой вариации - изменчивости средних значений признака для нескольких градаций фактора. Сила влияния разнообразных случайных факторов оценивается по величине изменчивости в пределах каждой из градации – внутригрупповая вариация. Таким образом, при дисперсионном анализе общая дисперсия признака (отсюда название) расчленяется на дисперсию, обусловленную влиянием изучаемого фактора, и случайную дисперсию. Если первая достоверно превышает вторую, то влияние фактора считается значимым. Отношение межгрупповой (факториальной) дисперсии к внутригрупповой (случайной) называется *F*-критерием Фишера (по имени исследователя предложившего в 1920 г. концепцию дисперсионного анализа). По величине *F*-критерия в дисперсионном анализе судят о достоверности влияния фактора на изучаемый показатель.

Примеры внутригруппового и межгруппового варьирования приведены на графике, называемом диаграммой размаха (рис. 31).

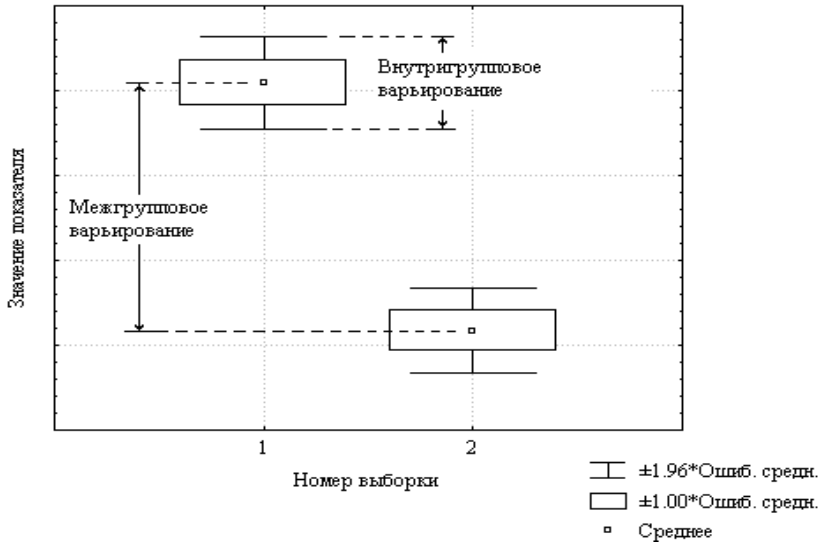


Рис. 31. Пример диаграмм размаха

Чем больше межгрупповое варьирование и меньше внутригрупповое, тем более значимыми являются различия между средними значениями групп.

Дисперсионный анализ помимо определения влияния одного фактора на значение признака (однофакторный дисперсионный анализ) позволяет определять влияние нескольких факторов (многофакторный дисперсионный анализ). При этом оценивается как отдельное влияние факторов на признак, так и их взаимодействие.

Однофакторный дисперсионный анализ. Произведем расчёт однофакторного дисперсионного комплекса на следующем примере. На лесосеменной плантации измеряется длина пыльников у отдельных деревьев сосны. Необходимо оценить, различаются ли отдельные деревья по длине пыльников, другими словами, как влияют индивидуальные особенности деревьев на длину пыльников. Получаем следующие исходные данные:

Номер дерева	Длина пыльников, мм					
1	40	24	46	44	35	
2	21	31	17	29	39	29
3	27	20	39	30		
4	17	21	23	24	21	

В данном случае фактором являются индивидуальные особенности деревьев, градации фактора – отдельные деревья (всего имеется четыре градации по числу деревьев). Фактор – независимый признак. Зависимый признак – длина пыльников.

Расчёт производится в следующей последовательности.

1. Оцениваем общую сумму квадратов *(SS total),* которая рассчитывается как сумма квадратов отклонений (см. п.1) для всех наблюдаемых значений (в нашем случае 20). Для расчёта можем воспользоваться функцией КВАДРОТКЛ в *MS Excel* или рассчитать вручную:

$$SS_T = \sum(x_i - \bar{x})^2 = 1526{,}55.$$

2. Оцениваем случайную сумму квадратов отклонений *(SS error).* Для этого оцениваем сумму квадратов для каждой градации фактора и суммируем их. В нашем примере градации фактора – отдельные деревья:

1 дерево: $SS_1 = \sum(x_i - \bar{x})^2 = 308{,}80$,

2 дерево: $SS_2 = 301{,}33$,

3 дерево: $SS_3 = 186{,}00$,

4 дерево: $SS_4 = 28{,}80$.

Объединяя, получаем случайную сумму квадратов:

$$SS_E = SS_1 + SS_2 + SS_3 + SS_4 = 308{,}80 + 301{,}33 + 186{,}00 + 28{,}80 = 824{,}93\ .$$

3. Рассчитываем факториальную сумму квадратов (*SS factor* или *SS effect*), для чего из общей суммы квадратов вычитаем случайную сумму квадратов:

$$SS_F = SS_T - SS_E = 1526{,}55 - 824{,}93 = 701{,}62\ .$$

4. Оцениваем количество степеней свободы *(df)* для каждой суммы квадратов. Для общей суммы квадратов *df* равно общему количеству наблюдений минус один:

$$df_T = N - 1 = 20 - 1 = 19$$

Факториальное *df* рассчитывается по тому же принципу: количество градаций фактора минус один:

$$df_F = n - 1 = 4 - 1 = 3$$

df для случайной суммы квадратов равна *df* общая минус *df* факториальная:

$$df_E = df_T - df_F = 19 - 3 = 16$$

5. Рассчитываем дисперсии, которые в дисперсионном анализе ещё называют средними квадратами *(MS)*. Последние рассчитываются делением суммы квадратов *(SS)* на количество степеней свободы *(df)*.

Общий средний квадрат: $MS_T = SS_T / df_T = 1526{,}55 / 19 = 80{,}34$.

Факториальный средний квадрат: $MS_F = SS_F / df_F = 701{,}62 / 3 = 233{,}87$.

Случайный средний квадрат: $MS_E = SS_E / df_E = 824{,}93 / 16 = 51{,}56$.

6. Определяем значение *F*-критерия и его значимость.

$$F = \frac{MS_F}{MS_E} = \frac{233{,}87}{51{,}56} = 4{,}53.$$

Значимость *F*-критерия определяется по стандартным таблицам. Если расчётное значение *F*-критерия превышает табличное на выбранном уровне значимости, влияние фактора считается достоверным.

Можно рассчитать значимость *F*-критерия при помощи статистических компьютерных программ. В частности, в *Statistica* это делается в меню *Statistics* при помощи модуля *Probability calculator* (калькулятор вероятностей), строка Distributions. В открывшемся окне колонке распределений *Distribution* выбираем распределение *F(Fisher),* в поля вводим значения *F*-критерия и количеств степеней свободы для MS_F и MS_E (3 и 16 соответственно), а также отмечаем галочкой опцию *1-Cumulative p* (рис. 32). После нажатия кнопки *Compute* (вычислить) в окне уровня значимости *(p)* появляется искомое значение. Поскольку полученное значение уровня значимости (0,0176) менее 0,05, можно говорить о том, что значение *F*-критерия достоверно на 5%-ном уровне значимости. Следовательно, деревья достоверно различаются по длине пыльников, что подтверждает влияние такого фактора как индивидуальные свойства дерева на значение переменной «длина пыльников».

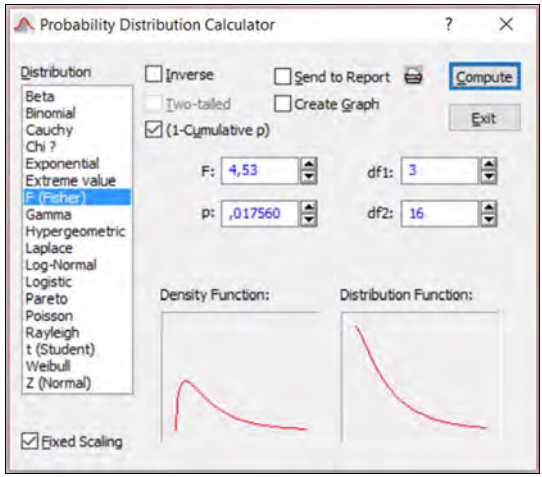


Рис. 32. Диалоговое окно расчета F-критерия

Расчёт в MS Excel по схеме однофакторного дисперсионного анализа выполняется, как обычно, при помощи команды "Анализ данных" в меню "Данные". В списке "Инструменты анализа" выбираем строку "Однофакторный дисперсионный анализ" и в появившемся диалоговом окне (рис. 33) вводим ссылку на диапазоны с переменными (каждый столбец – отдельное дерево, если в выделяемый диапазон включены номера деревьев – необходимо обязательно отметить опцию «Метки в первой строке»). Кроме того, указываем параметры вывода (куда записывать результаты).

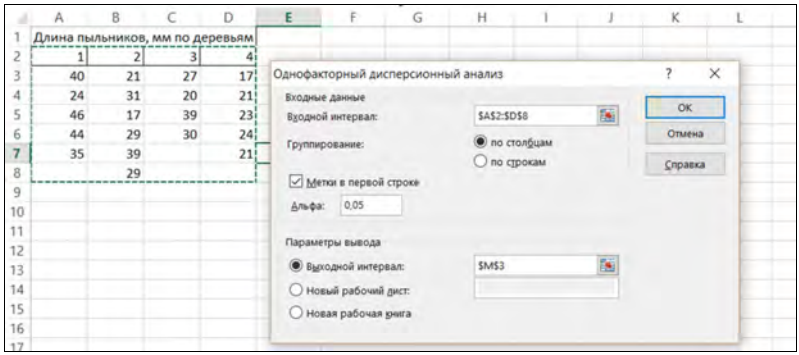


Рис. 33. Диалоговое окно ввода параметров однофакторного дисперсионного анализа

Результатом является таблица (рис. 34), в которой, в частности, указано расчётное значение *F*-критерия (4,536), критическое ("табличное") значение *F*-критерия двухстороннего (3,239), уровень значимости *P* (0,017). Рассчитанное значение превышает критическое, а уровень значимости *P* (0,017) менее 5%, поэтому влияние фактора достоверно на 5%-ном уровне значимости (различия между деревьями по длине пыльников достоверны).

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
1	5	189	37,8	77,2
2	6	166	27,66667	60,26667
3	4	116	29	62
4	5	106	21,2	7,2

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	701,6167	3	233,8722	4,53607	0,01747917	3,238871517
Внутри групп	824,9333	16	51,55833			
Итого	1526,55	19				

Рис. 34. Результаты расчета однофакторного дисперсионного анализа

Пример расчёта в Statistica. В пакете *Statistica* данные для однофакторного дисперсионного анализа (как и для большинства других видов статистического анализа) выстраиваются, как это было показано выше, по переменным: каждая переменная (фактор) – отдельный столбец (рис. 35). Однофакторный анализ в *Statistica* удобнее всего выполнять в модуле *Basic statistics / Tables* (основные статистики и таблицы), запускаемого из меню *Statistics,* далее - строка *Breakdown & one-way ANOVA* (классификация и однофакторный дисперсионный анализ).

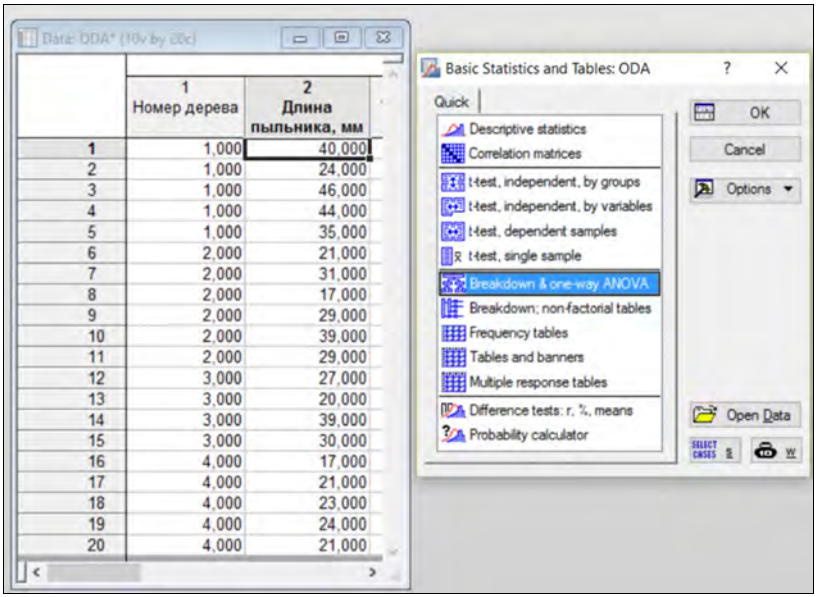


Рис. 35. Выбор меню однофакторного дисперсионного анализа

Кнопкой *Variables* выбираем фактор *(Grouping variables)* – номер дерева, а также зависимую переменную *(Dependent variables)* – длину пыльников (рис. 35) и запускаем процесс вычислений кнопкой OK. На вкладке *Quick* непосредственно результаты дисперсионного анализа можем запустить кнопкой *Analysis of Variance*, выбор необходимых статистик (среднее, количество наблюдений, ошибка среднего, стандартное отклонение и др.) и запуск процесса их расчета кнопкой *Summary*, а также графики размахов варьирования - на вкладке *Descriptives* с выбором соответствующих опций (рис. 36).

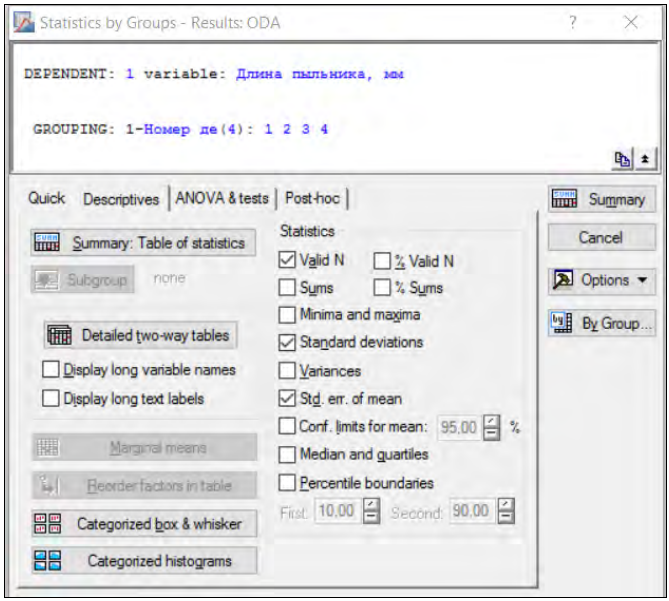


Рис. 36. Выбор опций однофакторного дисперсионного анализа

При запуске дисперсионного анализа из вкладки *Quick* кнопкой *Analysis of Variance* получаем таблицу результатов анализа (рис. 37).

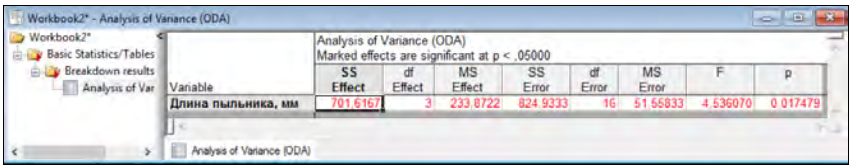


Рис. 37. Результаты однофакторного дисперсионного анализа

Так как влияние фактора достоверно на уровне значимости 0,05 (p<0,05), строка результатов выделена красным цветом. Таким образом, отдельные деревья достоверно различаются по длине пыльников (влияние индивидуальных свойств дерева на формирование пыльников достоверно). В *Statistica* можно вывести диаграммы размаха для различных градаций изучаемого фактора, подобные описанной выше (рис. 27). Для этого на той же вкладке *Descriptives* необходимо запустить процесс построения графика кнопкой *Categorized box & whisker* и далее в диалоговом окне *Box-Whisker*

Type выбрать в каких единицах измерять диаграмму размаха. Нагляднее всего, если варьирование переменной выражается в единицах ошибки среднего *Mean/SE/1.96*SE*, что позволяет оценить степень различий между изучаемыми группами (рис. 38). В частности, можем наблюдать значительные и достоверные различия («усы» не пересекаются) по длине пыльников между первым и четвертым деревом.

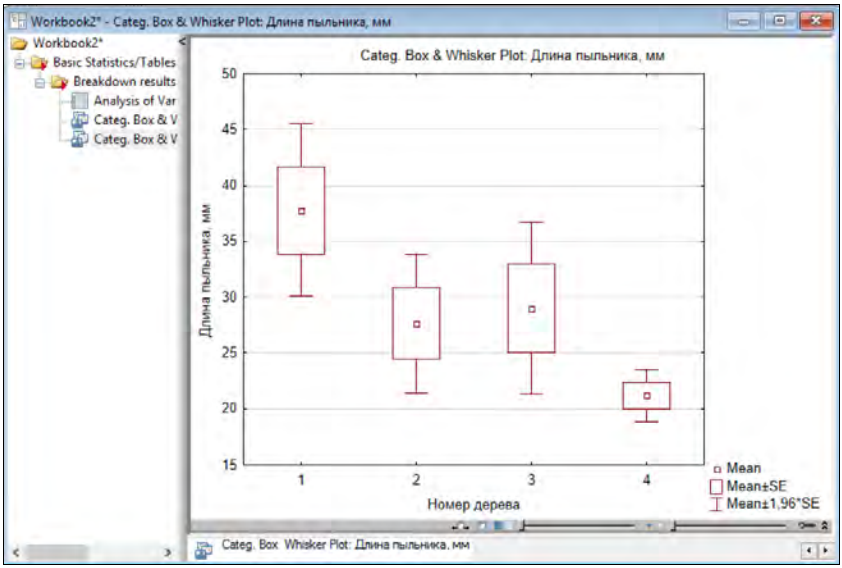


Рис. 38. Гистограмма с основными параметрами варьирования выборок

Двухфакторный дисперсионный анализ.

Используется в том случае, когда необходимо оценить влияние на какой-либо признак не одного, как это имеет место в однофакторном анализе, а двух факторов (если факторов несколько, то дисперсионный анализ называется многофакторным).

Пример. Рассмотрим следующий случай: необходимо оценить влияние типа почвы и удобрений на рост деревьев сосны. Допустим, что известны диаметры растений на четырёх делянках, различающихся по типу почвы и составу применяемых удобрений:

Варианты опыта		Почва	
		песчаная	суглинистая
Удобрения	Вариант 1	8 12 13	7 3
	Вариант 2	3 4 5	4 8 5

Двухфакторный дисперсионный анализ позволяет ответить на следующие вопросы:

1) различаются ли песчаная и суглинистая почвы по влиянию на скорость роста растений (фактор 1);

2) различаются ли удобрения вариантов 1 и 2 по влиянию на скорость роста растений (фактор 2);

3) влияют ли удобрения двух вариантов одинаково на разных типах почв (эффект взаимодействия факторов 1 и 2).

Пример расчёта в Statistica. Ввод данных – стандартный для статистических пакетов (рис. 39). Расчеты дисперсионного комплекса для многофакторного (в том числе, двухфакторного) дисперсионного анализа производится в модуле *ANOVA* (меню *Statistics*). Для нашего случая (классические факторы, представляющие собой номинальные переменные) выбираем тип анализа Factorial ANOVA (факторный дисперсионный анализ) и подтверждаем выбор кнопкой OK (рис. 39). Далее в диалоговом окне, вызываемом кнопкой *Variables*, выбираем зависимые переменные (*Dependent*) и факторы (*Categorical factors*). Соответственно, в данном дисперсионном комплексе одна зависимая переменная – диаметр деревьев и два фактора – механический состав почвы и наличие (отсутствие) удобрений (рис. 40), далее – ввод (*OK*).

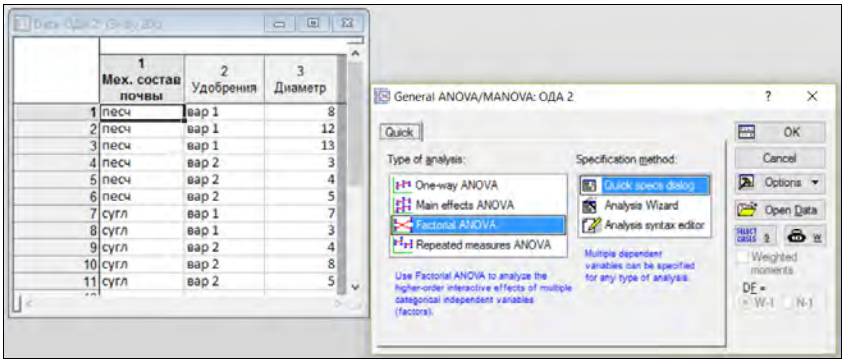


Рис. 39. Начальное меню многофакторного дисперсионного анализа

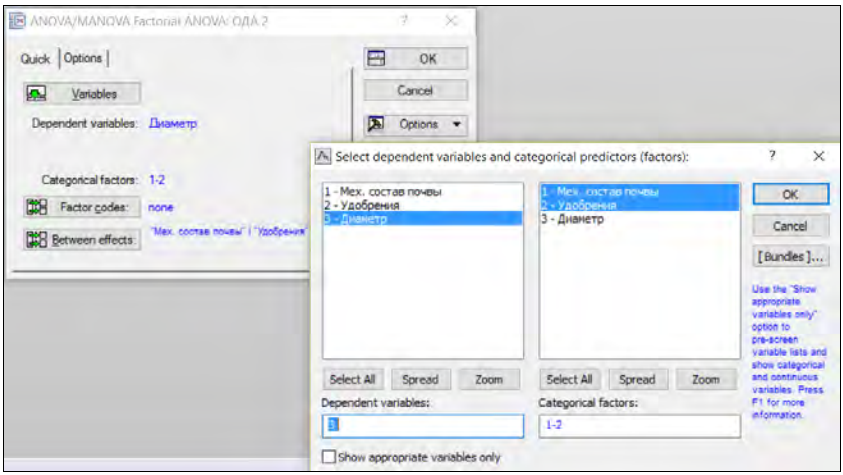


Рис. 40. Диалоговое окно выбора переменных дисперсионного анализа

В появившемся диалоговом окне запускаем таблицу результатов анализа кнопкой *All effects*. Результат анализа для нашего примера приведен на рис. 41 (красным цветом выделены строки, в которых фактор оказывает достоверное влияние на изучаемую переменную). Первой строкой можем видеть свободный член Intercept. В значительной степени упрощая ситуацию, можем сказать, что он характеризует отличие изучаемой зависимой переменной от нулевого значения (нулевая гипотеза состоит в том, что генеральное среднее изучаемого признака равна нулю) и в нашем слу-

чае может не рассматриваться. Далее оценивается влияние на зависимую переменную изучаемых факторов.

Effect	Univariate Tests of Significance for Диаметр (ОДА 2) Sigma-restricted parameterization Effective hypothesis decomposition				
	SS	Degr. of Freedom	MS	F	p
Intercept	439,1852	1	439,1852	94,11111	0,000026
Мех. состав почвы	12,5185	1	12,5185	2,68254	0,145465
Удобрения	26,7407	1	26,7407	5,73016	0,047902
Мех. состав почвы*Удобрения	39,1852	1	39,1852	8,39683	0,023061
Error	32,6667	7	4,6667		

Рис. 41. Результаты многофакторного дисперсионного анализа

Таблица результатов показывает, что механический состав почвы достоверно не влияет на скорость роста растений (уровень значимости $p = 0,14$, то есть превышает пороговое значение 0,05); два варианта применения удобрений (фактор «Удобрения») достоверно различаются по влиянию на скорость роста растений (уровень значимости $p<0,05$); взаимодействие типа почвы и варианта применения удобрений (обозначено в таблице «Мех. состав почвы*Удобрения») достоверно влияет на рост растений ($p<0,05$), т. е. удобрения влияют по-разному на разных почвах. Таким образом, при помощи двухфакторного дисперсионного анализа мы ответили на три вопроса, поставленные ранее.

10.1.2. Непараметрические тесты

Дисперсионный анализ Краскела-Уоллеса *(Kruskal-Wallis test)* является непараметрическим аналогом однофакторного дисперсионного анализа для независимых выборок и служит для определения достоверности различий между несколькими выборками по значению изучаемого показателя. При этом распределение признака сильно отличается от нормального, иначе использование непараметрического теста не имеет смысла, поскольку параметрические тесты являются более мощным инструментом статистического анализа. В качестве независимых переменных в непараметрическом дисперсионном анализе Краскела-Уоллеса могут выступать количественные и порядковые переменные.

Пример расчёта в Statistica. Рассмотрим влияние влажности почвы на сохранность культур сосны. Участки культур сгруппированы по трём градациям влажности почвы.

Сохранность, % вариант 1 (сухие почвы)	25, 38, 60, 29, 43, 18, 36, 24, 52.
Сохранность, % вариант 2 (свежие почвы)	89, 79, 61, 82, 40, 20, 72, 90, 58, 65, 58.
Сохранность, % вариант 3 (влажные почвы)	17, 11, 29, 13, 44, 13, 28, 6, 25, 12.

Так как показатель сохранности, скорее всего, не является нормально распределённым признаком, а группировка включает более двух вариантов, применяем для анализа непараметрический дисперсионный анализ Краскела-Уоллеса. Данные выстраиваются так же, как и в предыдущем случае (рис. 42): первая переменная – группирующая (номер варианта), вторая – зависимая (процент сохранности). Расчёт производится в модуле *Nonparametrics* (непараметрические статистики) из меню *Statistics*. В стартовой панели непараметрических статистик выбираем подходящий для наших условий вариант обработки: *Comparing multiple indep. samples (groups)* – Сравнение нескольких независимых выборок (групп) и запускаем меню дальнейшего выбора вариантов (рис. 42).

Далее в меню кнопкой Variables выбираем зависимую (сохранность растений) и независимую группирующую (влажность почвы) переменные и кнопкой Summary: *Kruskal-Wallis ANOVA & Median test* (дисперсионный анализ Краскела-Уоллеса и медианный тест), после чего запускаем процесс вывода результатов анализа кнопкой Summary (рис. 43)

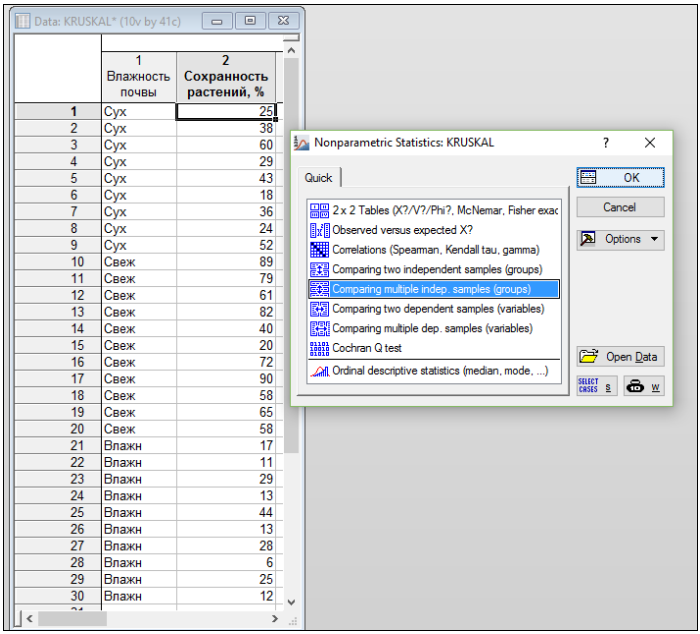


Рис. 42. Исходное меню непараметрического дисперсионного анализа

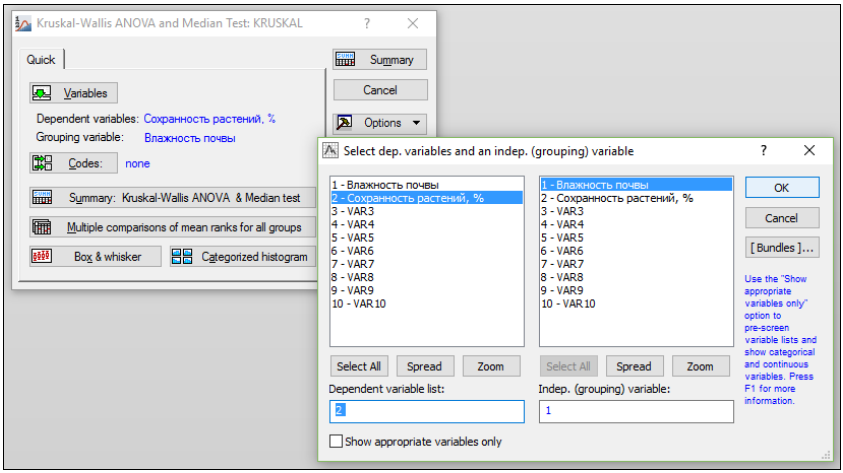


Рис. 43. Диалоговое окно выбора переменных

Результатом являются две таблицы, первая – результат анализа Краскела-Уоллеса (рис. 44), вторая – результат медианного теста, который является его аналогом. В обеих таблицах, прежде всего, обращаем внимание на значение уровня значимости *p* в заголовке таблицы.

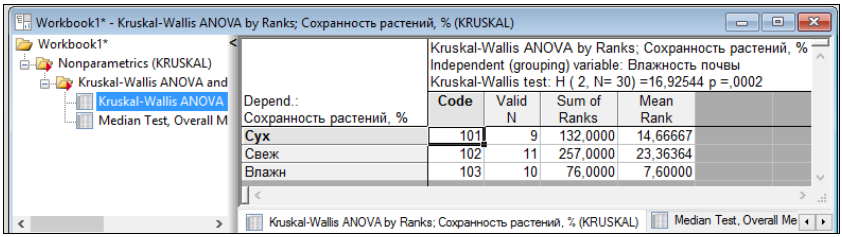


Рис. 44. Результаты непараметрического дисперсионного анализа

Поскольку и в результатах анализа Краскела-Уоллеса (*p*=0,0002), и в медианном тесте (*p*=0,010) значение уровня значимости не превышает принятого нами порога 0,05, можем говорить о том, что рассматриваемые варианты почвы достоверно (на уровне значимости 0,05) различаются по сохранности культур сосны. Другими словами, влияние влажности почвы на сохранность культур сосны достоверно на 5%-ном уровне значимости.

10.2. Зависимые выборки

10.2.1. Параметрические тесты

Как мы уже говорили выше (п. 5) дисперсионный анализ с повторными измерениями целесообразно проводить в тех случаях, когда переменная измеряется *для одних и тех же объектов* повторно (при разных условиях или в разное время). В этом случае говорят о факторе повторных измерений (называемого также *внутригрупповым фактором,* поскольку для оценки его значимости вычисляется внутригрупповая сумма квадратов). Если сравниваются разные группы объектов (например, четыре клона на лесосеменной плантации), то разница между группами описывается *межгрупповым фактором.* Способы вычисления критериев значимости для двух описанных типов факторов различны, но общая их логика и интерпретации совпадают.

Во многих случаях в эксперименте требуется включить в план дисперсионного анализа и межгрупповой фактор, и фактор повторных измерений. Например, измеряется длина хвои деревьев для трёх разных семей

сосны («семья» - межгрупповой фактор) до и после пересадки на лесокультурную площадь («влияние пересадки»- внутригрупповой фактор). Два измерения («до» и «после») каждого дерева образуют внутригрупповой фактор (или фактор с повторными измерениями). Расчёт и интерпретация главных эффектов и взаимодействий для межгрупповых факторов и факторов повторных измерений аналогичны приведённым выше для независимых выборок. Оба типа факторов могут, очевидно, взаимодействовать между собой (например, после пересадки у одной семьи длина хвои увеличивается, у другой – уменьшается).

Пример расчёта в Statistica. Для иллюстрирования описанной статистической процедуры рассмотрим случай, сходный с приведенным для двух выборок (см. п. 9.1.1) – различия между годами наблюдения по величине годичных приростов в высоту деревьев ели. Отличие здесь в том, что в данном случае используются данные не по двум, а по четырём годам наблюдения:

Номер дерева	1	2	3	4	5	6	7	8	9	10	11	12
Прирост в возрасте 10 лет	26	38	19	28	28	17	22	30	24	38	21	10
Прирост в возрасте 11 лет	20	38	18	25	42	20	14	26	20	29	15	5
Прирост в возрасте 12 лет	27	26	22	25	37	21	20	34	25	28	23	14
Прирост в возрасте 13 лет	30	34	30	24	46	30	18	40	26	37	29	15

Необходимо отметить, что при анализе зависимых выборок важно соответствие между строками таблицы, так как строка соответствует одному отдельному объекту – дереву, которому соответствует несколько наблюдений по разным столбцам таблицы (значение прироста в разные годы). В связи с этим, данные для зависимых выборок выстраиваются по столбцам: каждая градация фактора (в нашем случае это годы наблюдений) – отдельный столбец (рис. 41).

Расчёт производится в модуле *ANOVA (*запускается из меню *Statistics).* В появившемся диалоговом окне *General ANOVA/MANOVA: ANOVA* выбираем пункт *Repeated measures ANOVA* (дисперсионный анализ с повторными измерениями) и запускаем процедуру кнопкой ОК (рис. 45).

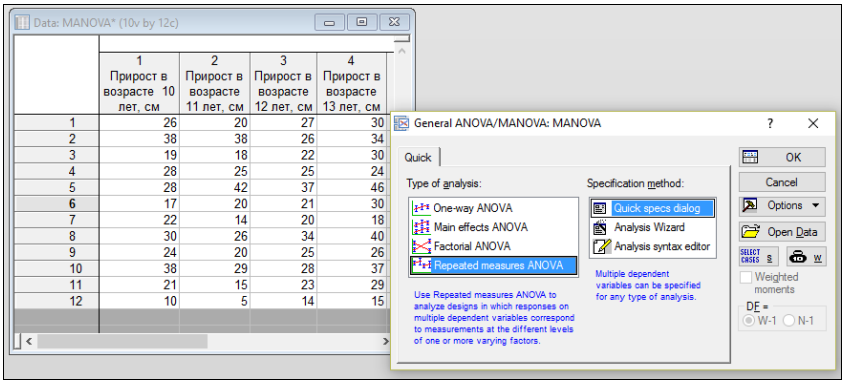


Рис. 45. Исходное меню дисперсионного анализа с повторными измерениями

Далее кнопкой *Variables* выбираем список переменных, при этом отмечаем все сопряжённые переменные (годы наблюдений) в поле *Dependent variable*, поле независимых переменных *(Categorical factors)* оставляем пустым и подтверждаем выбор кнопкой OK (рис. 46).

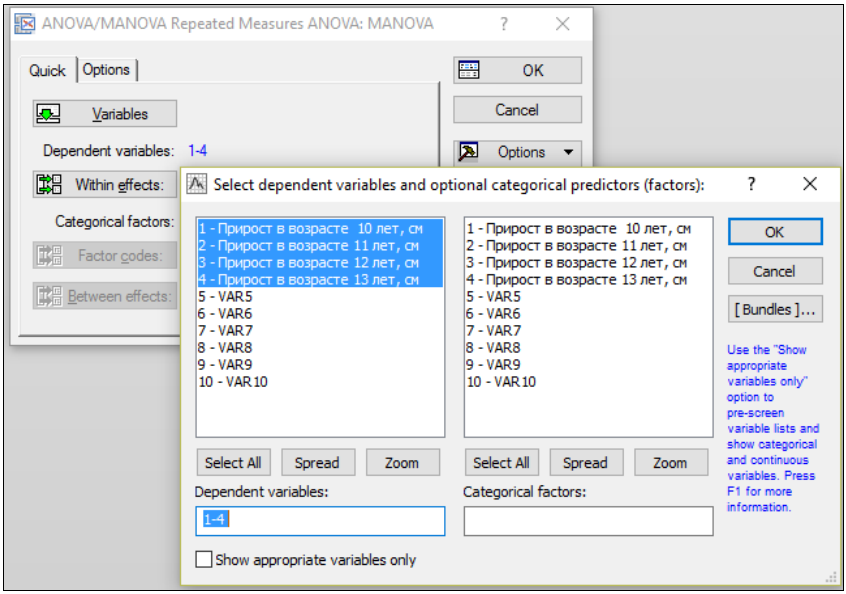


Рис. 46. Выбор переменных в дисперсионном анализе с повторными измерениями

Далее в том же окне *ANOVA/MANOVA Repeated Mesures ANOVA* (рис. 45) кнопкой *Within effects* запускаем окно вводим количество градаций фактора *No. of levels* (у нас их четыре – по числу годов измерений прироста) и название фактора (назовём его, например, *ГОД*) – Рис. 47.

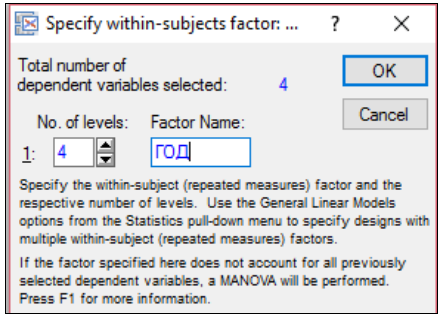


Рис. 47. Диалоговое окно выбора фактора и количества его градаций

После запуска процедуры анализа *(OK)* в появившемся диалоговом окне запускаем таблицу дисперсионного анализа *All effects* (все эффекты). Результатом будет таблица (рис. 48), в которой указано расчётное значение *F*-критерия для нашего фактора ГОД (6,120), а также уровень значимости *p-level* (0,00199).

Effect	Repeated Measures Analysis of Variance (MANOVA) Sigma-restricted parameterization Effective hypothesis decomposition				
	SS	Degr. of Freedom	MS	F	p
Intercept	31724,08	1	31724,08	136,9609	0,000000
Error	2547,92	11	231,63		
ГОД	331,75	3	110,58	6,1203	0,001988
Error	596,25	33	18,07		

Рис. 48. Результаты дисперсионного анализа с повторными измерениями

Так как показатель уровня значимости ниже, чем пороговое значение 0,05, можем говорить о том, что величина F-критерия дисперсионного анализа достоверно высока (о чём свидетельствует красный цвет строки). Следовательно, приросты деревьев ели разных лет наблюдения достоверно различаются и, таким образом, влияние фактора «год наблюдений» на значение переменной «годичный прирост» достоверно на уровне значимости 0,05.

10.2.2. Непараметрические тесты

Дисперсионный анализ Фридмана *(Friedman ANOVA)* является непараметрическим тестом для определения различий между средними значениями признака нескольких зависимых выборок (например, значения признака за ряд лет наблюдений). При этом распределение изучаемого признака неизвестно, либо существенно отличается от нормального. В противном случае целесообразно использовать дисперсионный анализ, так как он более полно использует имеющуюся в данных информацию.

Пример расчёта в Statistica. Имеются данные о среднем количестве семян в шишке у деревьев сосны обыкновенной в разные годы:

Номер дерева	1	2	3	4	5	6	7	8	9	10	11
Среднее кол-во семян в шишке, шт (2000г.)	10	14	22	28	13	17	32	27	30	25	10
Среднее кол-во семян в шишке, шт (2001г.)	13	26	28	30	10	35	27	33	35	26	15
Среднее кол-во семян в шишке, шт (2002г.)	19	18	20	16	10	10	23	35	24	16	18

Необходимо выяснить существуют ли различия между разными годами наблюдения по урожайности деревьев (количеству шишек). Так как неизвестно, соответствуют ли переменные нормальному распределению, используем непараметрический дисперсионный анализ Фридмана. Сначала в рабочем листе *Statistica* выстраиваем данные аналогично приведенному выше примеру для параметрического теста с зависимыми выборками: каждый отдельный год наблюдения – отдельная переменная (столбец), а измерения каждого отдельного дерева – в индивидуальной строке, т.е. отдельная строка соответствует отдельному объекту наблюдения (рис. 49).

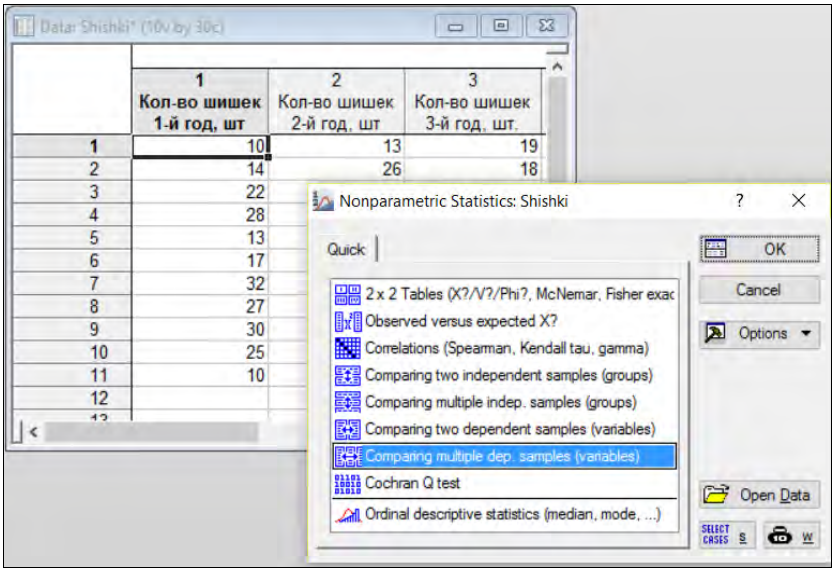


Рис. 49. Начальный выбор непараметрического дисперсионного анализа

Расчёт производится в модуле *Nonparametrics* (непараметрические статистики) из меню *Statistics*. В стартовой панели непараметрических статистик выбираем необходимый вариант обработки: *Comparing multiple dep. samples (variables)* – Сравнение нескольких зависимых выборок (по переменным) и запускаем меню дальнейшего выбора вариантов кнопкой OK (рис. 44). В окне *Friedman ANOVA by Ranks* кнопкой *Variables* отмечаем наши переменные и запускаем процедуру анализа кнопкой Summary: Friedman ANOVA & Kendall's concordance (рис. 50).

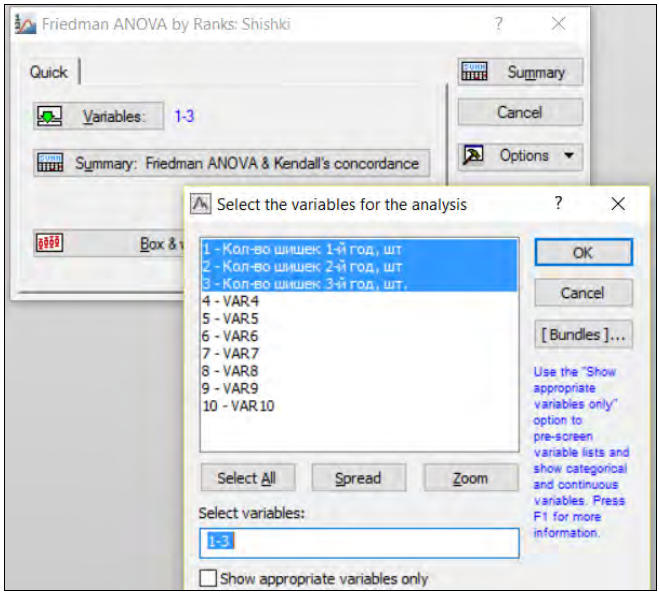


Рис. 50. Выбор переменных непараметрического дисперсионного анализа

В таблице результата (рис. 51) в верхней строке приведено значение критерия хи-квадрат *(Chi Sqr.* = 4,326) и его достоверность (уровень значимости p = 0,115).

Variable	Friedman ANOVA and Kendall Coeff. of Concordance (S ANOVA Chi Sqr. (N = 11, df = 2) = 4,325581 p = ,11501 Coeff. of Concordance = ,19662 Aver. rank r = ,11628			
	Average Rank	Sum of Ranks	Mean	Std.Dev.
Кол-во шишек 1-й год, шт	1,818182	20,00000	20,72727	8,210858
Кол-во шишек 2-й год, шт	2,500000	27,50000	25,27273	8,787388
Кол-во шишек 3-й год, шт,	1,681818	18,50000	19,00000	6,928203

Рис. 51. Результаты непараметрического дисперсионного анализа

Так как значение уровня значимости превышает 0,05, мы не можем с уверенностью утверждать, что влияние изучаемого фактора достоверно. Следовательно, достоверность различий между средними значениями количества семян в шишке деревьев сосны по годам наблюдения (общие средние по годам приведены в столбце Mean) не доказана.

11. АНАЛИЗ ТЕСНОТЫ СВЯЗИ МЕЖДУ ПЕРЕМЕННЫМИ

Коэффициент корреляции (correlation coefficient) используются для определения тесноты линейной связи между количественными переменными. Он может изменяться от -1 до +1. Значение -1 говорит о полностью отрицательной корреляции (при возрастании значений одной переменной значения другой пропорционально убывают), значение +1 означает полностью положительную корреляцию (с возрастанием значений одной переменной значения другой пропорционально возрастают). Значение коэффициента, равное 0, означает отсутствие корреляции (переменные не связаны друг с другом линейной зависимостью).

Чаще всего для определения тесноты связи между количественными переменными используется коэффициент корреляции Пирсона *(Pearson correlation coefficient r),* который рассчитывается по следующей формуле:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

где r – коэффициент корреляции между признаками x и y;

σ_{xy} – ковариация признаков x и y;

σ_x и σ_y – стандартное отклонение для признаков x и y, соответственно.

Для выборочных данных эту формулу можно записать в следующем виде:

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

Ошибка коэффициента корреляции рассчитывается по формуле:

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

Доверительный интервал и значимость коэффициента корреляции оценивается по t-критерию Стьюдента (см. п. 1).

Пример. Рассчитаем коэффициент корреляции между высотой и диаметром 12 деревьев сосны.

Высота, м, x	Диаметр, см y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
8,1	16	1,07	-0,20	-0,21	1,14	0,04
7,3	16,4	0,27	0,20	0,05	0,07	0,04
5,8	14	-1,23	-2,20	2,71	1,52	4,84
8,2	18,3	1,17	2,10	2,46	1,36	4,41
7,8	17,5	0,77	1,30	1,00	0,59	1,69
6,7	16,4	-0,33	0,20	-0,07	0,11	0,04
6,6	19,1	-0,43	2,90	-1,25	0,19	8,41
7,2	18,1	0,17	1,90	0,32	0,03	3,61
6,2	14,6	-0,83	-1,60	1,33	0,69	2,56
5,6	11,9	-1,43	-4,30	6,15	2,05	18,49
8,9	19,7	1,87	3,50	6,55	3,49	12,25
6	12,4	-1,03	-3,80	3,91	1,07	14,44
	Σ	0	0	22,95	12,31	70,82

$$\bar{x} = \frac{8{,}1+7{,}3+5{,}8+8{,}2+7{,}8+6{,}7+6{,}6+7{,}2+6{,}2+5{,}6+8{,}9+6{,}0}{12} = 7{,}03\,.$$

$$\bar{y} = \frac{16+16{,}4+14+18{,}3+17{,}5+16{,}4+19{,}1+18{,}1+14{,}6+11{,}9+19{,}7+12{,}4}{12} = 16{,}2\,.$$

Рассчитываем коэффициент корреляции:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{22{,}95}{\sqrt{12{,}31\cdot 70{,}82}} = 0{,}78\,.$$

Ошибка коэффициента корреляции:

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0{,}78^2}{12-2}} = 0{,}198\,.$$

Таким образом, коэффициент корреляции между высотой и диаметром составляет:

$$r = 0{,}78 \pm 0{,}198$$

Доверительный интервал для коэффициента корреляции рассчитывается на основе ошибки коэффициента корреляции с помощью t-критерия Стьюдента (см. п.3, доверительный интервал) и, например, для 5%-ного уровня значимости составляет $r \pm t\cdot m_r = 0{,}78 \pm 1{,}96\cdot 0{,}198 = 0{,}78 \pm 0{,}388$. Таким образом, с вероятностью 95% генеральное ("истинное") значение коэффициента корреляции находится в интервале от 0,39 до 1,00 (нижняя

граница 0,78-0,388 = 0,39; верхняя граница 0,78+0,388 = 1,17 и поскольку значение более1,00 не имеет смысла – верхняя граница принимается равной 1,00). Поскольку 0 не входит в доверительный интервал, можно говорить о том, что между высотой и диаметром деревьев существует достоверная (на 5%-ном уровне значимости) положительная связь.

В *MS Excel* коэффициент корреляции между двумя массивами данных рассчитывается при помощи функции КОРРЕЛ().

Пример расчёта в Statistica. Коэффициент корреляции между двумя или несколькими массивами данных рассчитывается в модуле *Basic statistics / Tables* (основные статистики и таблицы), запускаемом из меню *Statistics*. Данные располагаются по столбцам: каждая переменная (высота, диаметр, объем ствола) – отдельный столбец. В модуле *Basic statistics / Tables* (основные статистики и таблицы) запускаем *Correlation matrices* и кнопкой *One variable list* (единый список переменных) выбираем переменные для анализа. При этом можно выбрать несколько переменных (рис. 52).

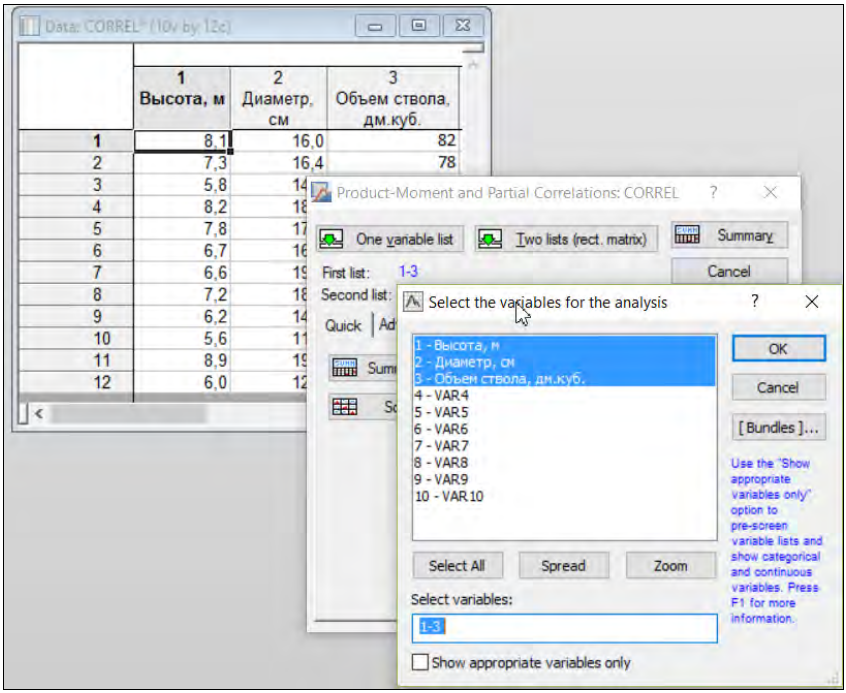


Рис. 52. Выбор переменных для расчета коэффициентов корреляции

Запускаем процедуру анализа (*Summary*). Результатом является корреляционная матрица (рис. 53), в которой коэффициенты, достоверно отличающиеся от нуля, отмечены красным цветом. Помимо этого, в таблице приведены средние значения и стандартные отклонения изучаемых показателей.

Variable	Correlations (CORREL) Marked correlations are significant at p < ,05000 N=12 (Casewise deletion of missing data)				
	Means	Std.Dev.	Высота, м	Диаметр, см	Объем ствола, дм.куб.
Высота, м	7,03333	1,05773	1,000000	0,777382	0,905506
Диаметр, см	16,20000	2,53736	0,777382	1,000000	0,957730
Объем ствола, дм.куб.	76,75000	31,36044	0,905506	0,957730	1,000000

Рис. 53. Результаты расчета коэффициентов корреляции

Ранговый коэффициент корреляции Спирмена *(correlation coefficient Spearman R)* используется обычно для расчета корреляции между переменными, выраженными в порядковой шкале. Так же как и линейный, коэффициент корреляции Пирсона может изменяться от -1 до +1.

Коэффициент корреляции Спирмена рассчитывается по формуле:

$$R = 1 - \frac{6\sum d^2}{n^3 - n},$$

где d –разность между рангами признаков x и y; n – численность выборки.

Ошибка коэффициента корреляции Спирмена рассчитывается по формуле:

$$m_R = \sqrt{\frac{1}{n-1}}\ .$$

Доверительный интервал и значимость коэффициента корреляции Спирмена, оценивается сходным с коэффициентом корреляции Пирсона образом (см. выше).

Пример. Рассчитаем, изменяется ли величина прироста в высоту с увеличением возраста у отдельного дерева сосны (существует ли корреляция между возрастом и величиной прироста). Поскольку возраст в данном случае измеряется целыми годами, он является порядковой переменной. В связи с этим, целесообразно использовать ранговую корреляцию, а именно коэффициента корреляции Спирмена.

Расчёт выполняется по следующей таблице:

Возраст, лет, x	Прирост, см, y	Ранг, x	Ранг, y	Разность рангов, d	d^2
8	27	1	1	0	0
9	50	2	6	-4	16
10	42	3	4	-1	1
11	55	4	9	-5	25
12	44	5	5	0	0
13	41	6	3	3	9
14	54	7	8	-1	1
15	52	8	7	1	1
16	60	9	10	-1	1
17	36	10	2	8	64
	Σ	55	55	0	118

Определение рангов для значений переменных – см. п.2, переменные. Правильность определения рангов проверяется следующим образом: суммы рангов переменных должны быть равны $n(n+1)/2$ и соответствовать друг другу. В нашем случае $n(n+1)/2 = 10\cdot(10+1)/2 = 55$ (правильно).

Ранговый коэффициент корреляции равен:

$$R = 1 - \frac{6\sum d^2}{n^3 - n} = 1 - \frac{6 \cdot 118}{10^3 - 10} = 1 - \frac{708}{990} = 0{,}28.$$

Ошибка коэффициента корреляции:

$$m_R = \sqrt{\frac{1}{n-1}} = \sqrt{\frac{1}{10-1}} = 0{,}302.$$

Таким образом, коэффициент корреляции между возрастом и приростом составляет:

$$R = 0{,}28 \pm 0{,}302$$

Доверительный интервал для коэффициента корреляции (см. выше) на 5%-ном уровне значимости составляет $R \pm t\cdot m_R = 0{,}28 \pm 1{,}96\cdot 0{,}302 = 0{,}28 \pm 0{,}592$. Таким образом, с вероятностью 95% генеральное ("истинное") значение коэффициента корреляции находится в интервале от -0,31 до 0,87 (0,28-0,592 = -0,31; 0,28+0,592 = 0,87). Поскольку 0 входит в доверительный интервал, можно говорить о том, что на 5%-ном уровне значимости полученный коэффициент корреляции достоверно не отличается от нуля (между возрастом и приростом в высоту дерева сосны не обнаружено достоверной связи).

Пример расчёта в Statistica. Ввод данных (рис. 54) осуществляется аналогично вышеприведенному примеру с параметрическим коэффициентом корреляции Пирсона. Коэффициент корреляции Спирмена рассчитывается в модуле *Nonparametrics* (непараметрические статистики), запускаемом из меню *Statistics.* Далее выбираем раздел *Correlations (Spearman, Kendall tau, gamma).*

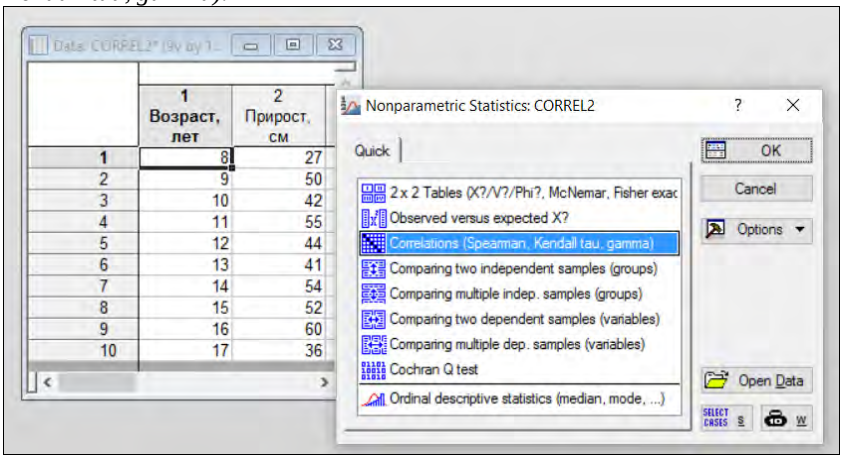


Рис. 54. Исходное меню выбора непараметрических коэффициентов корреляции

Кнопкой variables выбираем список переменных (рис. 55), нажимаем ОК, в окошке *Compute* выбираем тип выводимых данных (*Square matrix* – квадратная матрица) и кнопками, расположенными ниже, запускаем процесс расчета нужного нам показателя (Spearman rank R – ранговый коэффициент корреляции Спирмена)

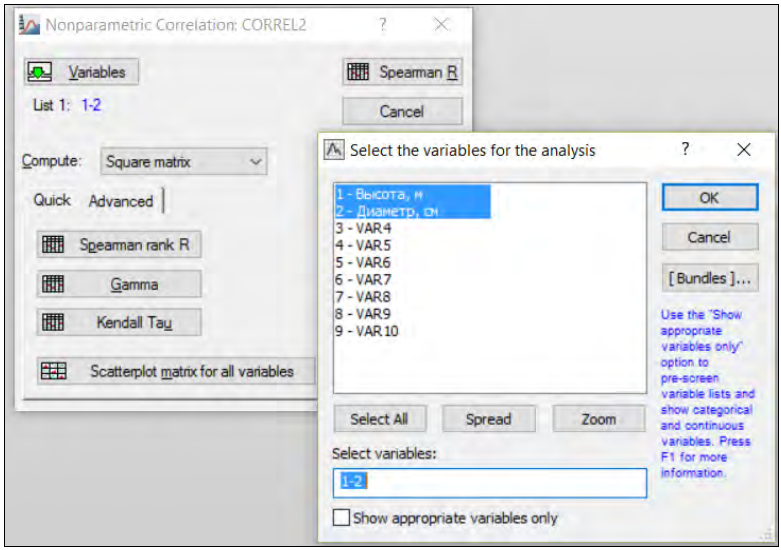


Рис. 55. Выбор переменных для расчета коэффициента корреляции Спирмена

Получаем таблицу результатов (рис. 56).

Variable	Spearman Rank Order Correlations (CORREL2) MD pairwise deleted Marked correlations are significant at p <,05000	
	Высота, м	Диаметр, см
Высота, м	1,000000	0,284848
Диаметр, см	0,284848	1,000000

Рис. 56. Результаты расчета коэффициента корреляции Спирмена

из полученных данных следует вывод о том, что ранговая корреляция изучаемых показателей недостоверна (этот вывод мы уже сделали выше на основе расчета границ доверительного интервала, который включает в себя нулевое значение), поскольку значения не отмечены красным шрифтом как достоверные на уровне значимости 5%

12. РЕГРЕССИОННЫЙ АНАЛИЗ

Задачей регрессионного анализа является оценка связи между количественными переменными. По сравнению с корреляционным анализом, который позволяет определить тесноту взаимосвязи между количественными переменными и её направленность (положительная или отрицательная корреляция), регрессионный анализ позволяет сделать более детальный анализ взаимосвязи переменных. Прежде всего, наряду с оценкой тесноты связи, он позволяет найти уравнение зависимости между переменными и, следовательно, при необходимости определить значения одной переменной (зависимой) по значениям другой (независимой). При помощи регрессионного анализа мы можем найти конкретные значения коэффициентов такого уравнения, получить оценку их значимости, точности уравнения и определить эффективность модели при изучении изменчивости признаков.

12.1. Простая линейная регрессия

Изучение метода регрессионного анализа целесообразно начать с простой линейной регрессии. При этом уравнение взаимосвязи между двумя переменными (уравнение регрессии) в этом случае имеет следующий вид:

$$Y = a + bX,$$

где X – независимая переменная; Y – зависимая переменная; a и b – коэффициенты уравнения.

Задачей регрессионного анализа является определение конкретных значений коэффициентов a и b при известных из опыта значениях переменных X и Y, которые измерены попарно. Кроме того, на основании регрессионного анализа определяются показатели эффективности полученного уравнения, то есть насколько точно оно описывает имеющиеся данные.

Например, нам необходимо определить зависимость высоты от диаметра дерева сосны на основе линейной регрессии. Имеем следующие данные (для удобства запишем их по возрастанию переменной X):

Диаметр, см, (X)	1,7	1,8	2,0	2,1	2,9	3,1	3,4	3,7	4,5	5,0	5,0	5,8
Высота, м, (Y),	2,5	2,2	2,4	2,0	2,8	2,2	2,6	3,1	3,6	3,5	3,2	4,1

Представив эти данные в системе координат (рис. 57, а), можно видеть, что точки образуют узкое поле вокруг воображаемой наклонной прямой линии. При помощи регрессионного анализа мы можем найти уравнение этой линии (уравнение регрессии). Производится это при помощи метода наименьших квадратов, который подбирает линию так, чтобы сумма квадратов отклонений точек от этой линии была наименьшей. На Рис. 57, б приведена такая линия.

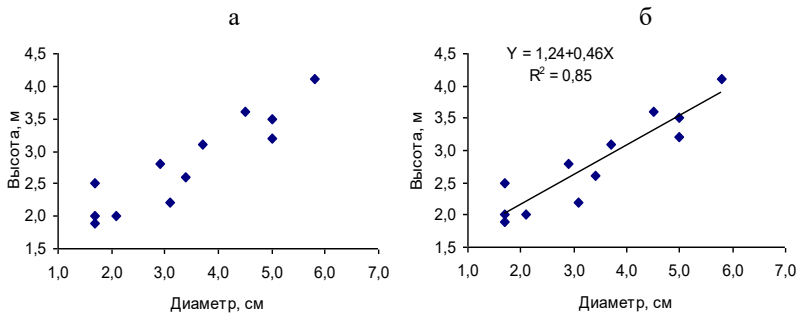


Рис. 57. Построение регрессии по эмпирическим данным

Уравнение этой линии Y=1,24+0,46X. Для того, чтобы определить насколько точно точки ложатся на полученную прямую, используется коэффициент детерминации R^2, который показывает, какая часть варьирования точек объясняется найденной линией регрессии. Измеряется R^2 в долях единицы или в процентах (в нашем случае 0,85 или 85%). Максимальное значение R^2 равное 1 (или 100%), означает, что все точки лежат на линии регрессии, т. е. линия регрессии на 100% объясняет изменчивость зависимой переменной.

Кроме того, важными показателями, получаемыми в процессе регрессионного анализа, являются показатели значимости как коэффициентов регрессионного уравнения, так и регрессии в целом (т. е. имеет она смысл или нет, случайна ли найденная зависимость). Значимость регрессии (регрессионного уравнения) определяется обычно на основе F-критерия, для которого вычисляется значение уровня значимости *(p)* и по которому судят о значимости регрессионного уравнения в целом. Чтобы определить значимость отдельных коэффициентов регрессионного уравнения (a и b), используется обычно t-критерий, для которого также рассчитываются уровни значимости, по которым определяем достоверность этих коэффициентов.

Пример расчёта в MS Excel. Воспользуемся данными о высоте и диаметре деревьев сосны, приведенными выше. Линейный регрессионный анализ в *MS Excel* производится при помощи команды "Анализ данных" (активация надстройки "Анализ данных" в *MS Excel* - см. п. 8.1.1) в меню "Данные". В списке "Инструменты анализа" выбираем строку "Регрессия" и в появившемся диалоговом окне (рис. 58) вводим ссылку на диапазоны с зависимой (*Y*) и независимой (*X*) переменными, а также параметры вывода (интервал, новый лист или книга) и некоторые опции, относящиеся к параметрам выводимых данных (по необходимости).

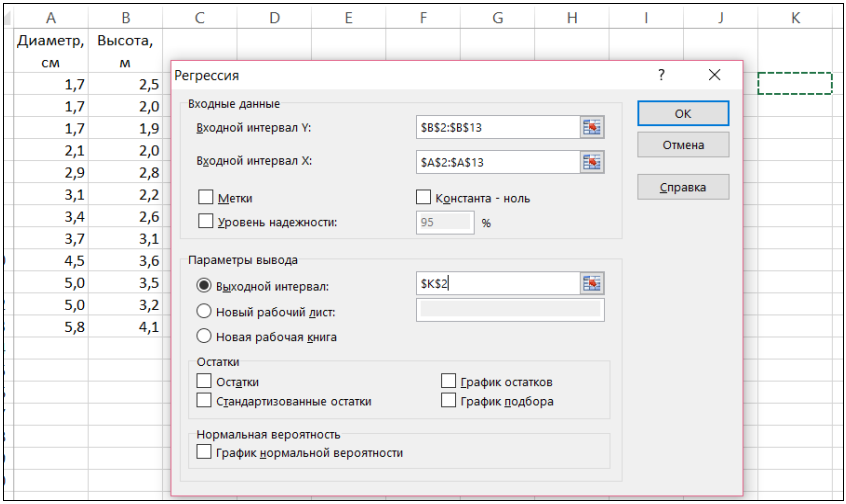


Рис. 58. Диалоговое окно основных параметров регрессионного анализа

Результатом является таблица (рис. 59), в которой следует обратить внимание на следующие показатели:

1) Коэффициент детерминации (*R*-квадрат), составляющий 0,847, говорит о том, что варьирование *Y* на 85% описывается регрессионной линией, что подтверждает достаточно высокую степень линейной зависимости.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,92085018
R-квадрат	0,84796505
Нормированный R-квадрат	0,83276156
Стандартная ошибка	0,29306497
Наблюдения	12

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	1	4,79029591	4,79029591	55,77435	2,13959E-05
Остаток	10	0,858870756	0,085887076		
Итого	11	5,649166667			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значени	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	1,24345767	0,223904273	5,553523644	0,000243	0,744567863	1,742347	0,744568	1,742347
Переменная X 1	0,45759872	0,061272773	7,468222754	2,14E-05	0,321074472	0,594123	0,321074	0,594123

Рис. 59. Результаты регрессионного анализа

2) Значения коэффициентов *a* (коэффициент *Y*-пересечения, равен 1,243) и *b* (коэффициент переменной *X1*, равен 0,458). Таким образом, уравнение линейной регрессии имеет вид $Y = 1{,}243 + 0{,}458X$.

3) Уровни значимости t-статистик для коэффициентов *a* и *b (P*-значение*)*. Для *a* уровень значимости составляет 0,00024, для *b* – 0,00002 ($2{,}14*10^{-5}$). Так как оба уровня значимости существенно меньше 0,05, коэффициенты *a* и *b* являются достоверными на 5%-ном уровне значимости (важны для вычисленной зависимости).

4) Уровень значимости *F*-критерия дисперсионного анализа, оценивающего достоверность регрессионного уравнения в целом. Значимость *F* в нашем случае составляет 0,00002 (меньше 0,05). Это говорит о том, что уравнение в целом хорошо описывает зависимость *Y* от *X* (т. е. зависимость с высокой степенью достоверности не является случайной).

Пример расчёта в Statistica. Линейный регрессионный анализ выполняется в модуле *Multiple Regression*, запускаемом из меню *Statistics*. Кнопкой *Variables* выбираем зависимую *(Dependent var.)* и независимую *(Independent variable list)* переменные (рис. 60). После запуска процедуры анализа (*OK*) в появившемся диалоговом окне кнопкой *Summary: Regression results* (итоги: результаты регрессии) выводим таблицу результатов (рис. 61), где приведены общие результаты регрессии (*Multiple Regression Results*) и, кроме того, отдельными кнопками можем вывести более подробные результаты регрессионного анализа. Так, например во вкладке *Advanced* (расширенные) можем вывести отдельными окнами основные результаты регрессионного анализа (*Summary: Regression results*), дисперсионный анализ (ANOVA) и многие другие.

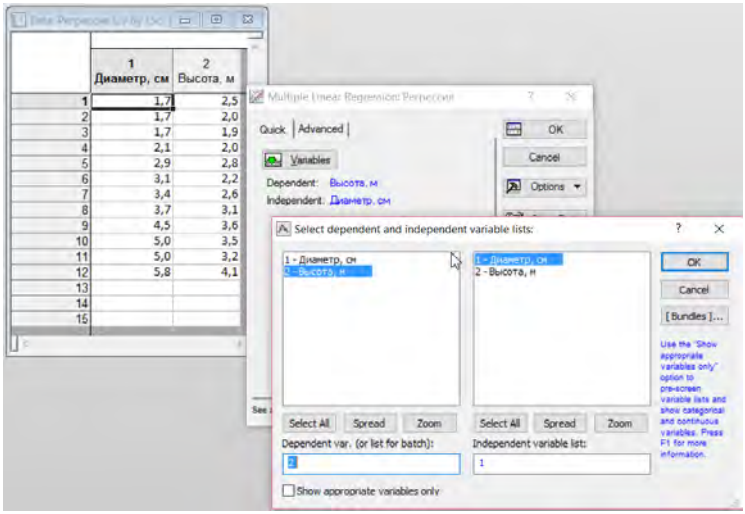


Рис. 60. Выбор переменных регрессионного анализа

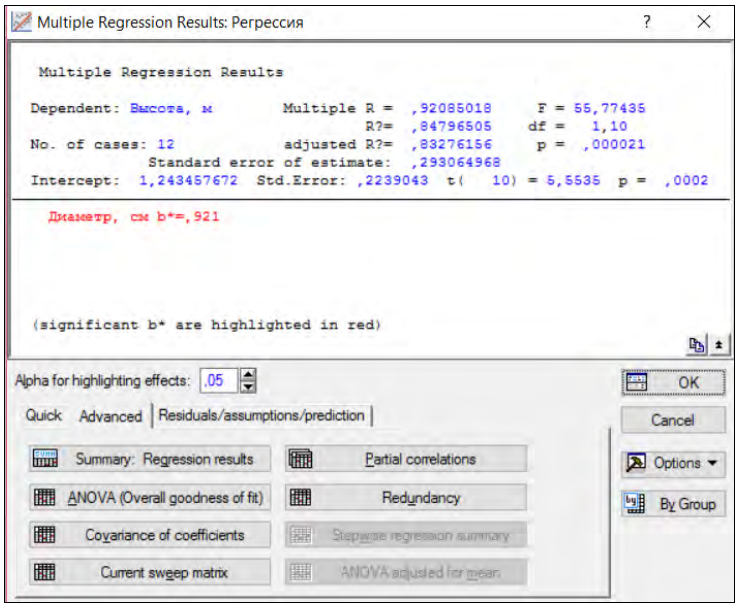


Рис. 61. Промежуточное диалоговое окно регрессионного анализа

В частности, в результатах регрессионного анализа (рис. 62), выводимых кнопкой *Summary: Regression results* вкладки *Quick (быстрые результаты)* обращаем внимание на следующие показатели:

1) Коэффициент детерминации (*R?*), который составляет 0,8479 (точки хорошо ложатся на прямую).

2) Значения коэффициентов *a* (*B* для *Intercpt*, равен 1,243) и *b* (*B* для переменной «Диаметр» равен 0,458). Следовательно, уравнение имеет вид $Y = 1{,}243 + 0{,}458X$.

3) Уровни значимости *t*-критерия для коэффициентов *a* и *b* (*p-level*). Для *a* рассчитанный уровень значимости составляет 0,000243, для *b* – 0,000021. Оба уровня значимости менее 0,05, следовательно коэффициенты *a* и *b* – достоверны на 5% уровне значимости, что подтверждается и красным цветом шрифта соответствующих строк таблицы.

4) Уровень значимости *F*-критерия (в третьей строке после значения *F*-критерия, $p<0{,}0002$), оценивающего достоверность регрессионного уравнения в целом. Значимость *F*-критерия, составляющая менее 0,00002 говорит о высокой степени достоверности полученного уравнения зависимости *Y* от *X*.

N=12	Regression Summary for Dependent Variable: Высота, м (Регрессия) R= ,92085018 R?= ,84796505 Adjusted R?= ,83276156 F(1,10)=55,774 p<,00002 Std.Error of estimate: ,29306					
	b*	Std.Err. of b*	b	Std.Err. of b	t(10)	p-value
Intercept			1,243458	0,223904	5,553524	0,000243
Диаметр, см	0,920850	0,123302	0,457599	0,061273	7,468223	0,000021

Рис. 62. Результаты регрессионного анализа

Для того, чтобы построить по исходным данным график зависимости, необходимо в меню Graphs (графики) основной программы запустить строку Scatterplots (точечный), в открывшемся окне кнопкой Variables выбрать переменные X и Y для построения графика (рис. 63) и, при необходимости, изменить тип графика на множественный (*Graph type – Multiple*), задать линии, показывающие границы разброса (опции группы regression bands) и др. Во вкладке *Advanced* представлен более широкий выбор инструментов регрессионного анализа, в частности, в ней можем выбрать отличные от линейной типы зависимости (полиномиальная, логарифмическая и др.), отметить опции вывода коэффициента линейной корреляции, коэффициента детерминации и др. Так, при использовании опций, приведенных на Рис. 64 после запуска процедуры получаем график линейной регрессии со всей необходимой информацией (рис. 65).

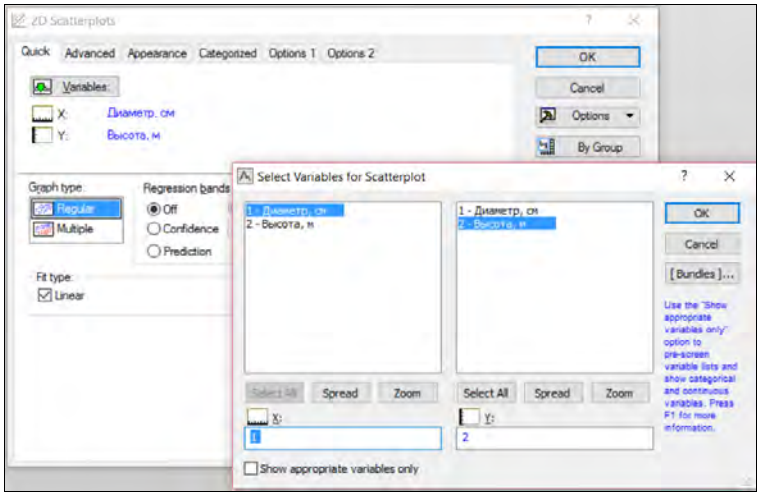


Рис. 63. Выбор переменных для построения графика

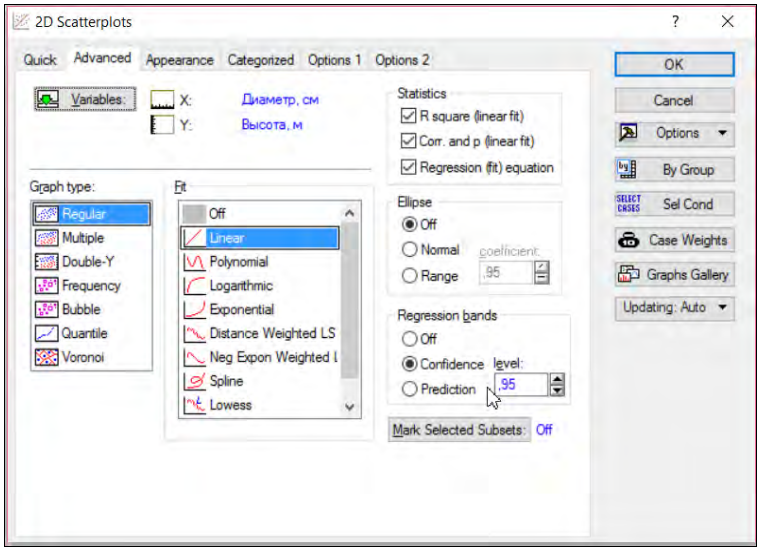


Рис. 64. Диалоговое окно параметров графика

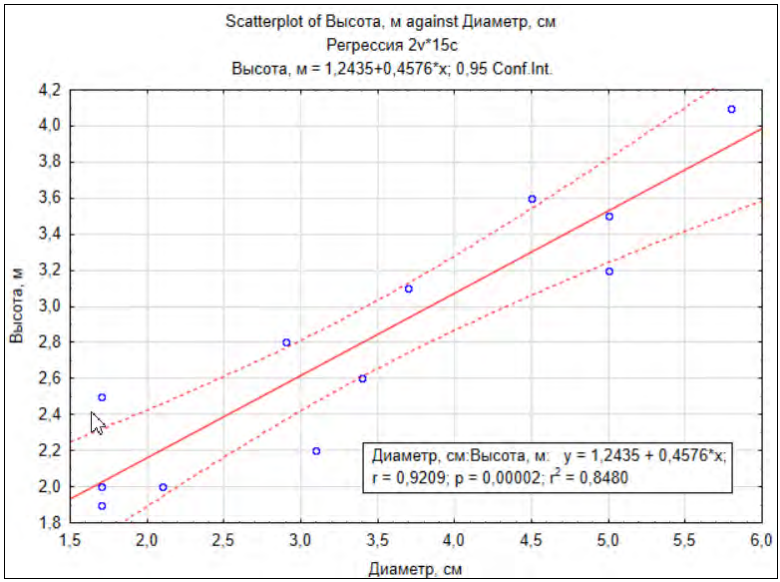


Рис. 65. График зависимости между переменными

На полученном графике приведены уравнение регрессии, линейный коэффициент корреляции между переменными (0,92), коэффициент детерминации (0,85) а пунктирными линиями показана область (доверительный интервал), в которой с вероятностью 95% находится «истинная» линия регрессии.

11.2. Множественная линейная регрессия

В отличие от простой линейной регрессии, где используется одна независимая переменная *X*, множественная линейная регрессия рассматривает зависимость *Y* от нескольких переменных *X*. Следовательно, уравнение множественной регрессии можем записать следующим образом:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n,$$

где *Y* – зависимая переменная; X_1; X_2… X_n – независимые переменные; a, b_1; b_2…b_n– коэффициенты уравнения.

После расчёта значений коэффициентов уравнения мы можем определить, какие из коэффициентов достоверны, а какие – нет и делать заключение относительно того, какие из независимых переменных *X* наиболее

тесно связаны с результирующей переменной *Y*, и какие из них можно из уравнения исключить.

Пример расчёта в Statistica. Допустим, необходимо определить, насколько сильно связана высота дерева сосны со следующими переменными: диаметр ствола, прирост в высоту за последний год, диаметр кроны:

Высота, м *(Y)*	3,1	3,6	2,5	2,0	4,1	3,5	3,2	2,8	2,2	2,6	1,9	2,0	2,3	3,4
Диаметр, см *(X_1)*	3,7	4,5	1,7	1,7	5,8	5,0	5,0	2,9	3,1	3,4	1,7	2,1	3,2	4,0
Прирост, см *(X_2)*	50	32	24	40	53	56	17	56	39	56	40	36	41	46
Диаметр кроны, м *(X_3)*	1,4	1,3	1,8	0,7	1,7	1,5	1,4	1,3	1,5	1,4	1,2	1,3	1,2	1,5

Множественный линейный регрессионный анализ выполняется аналогично вышеописанной простой линейной регрессии в модуле *Multiple Regression*. Кнопкой *Variables* выбираем зависимую *(Dependent var.)* и несколько независимых *(Independent variable list)* переменных (рис. 66).

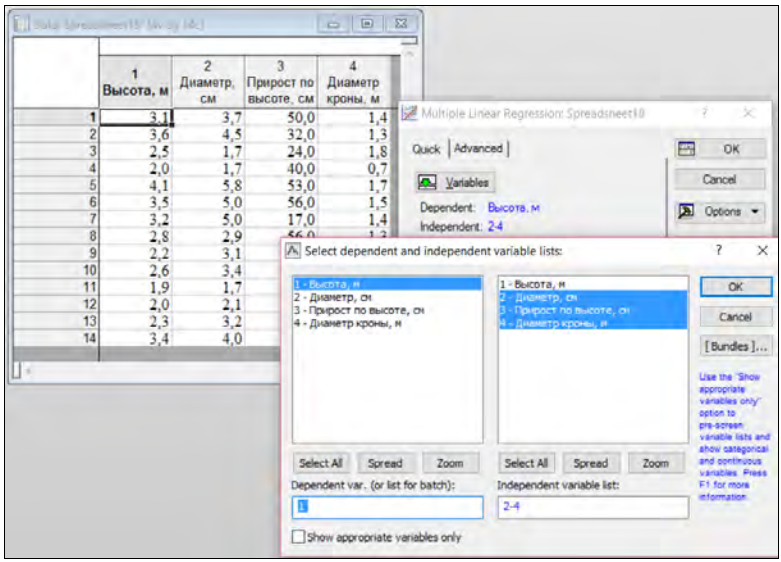


Рис. 66. Выбор переменных множественной регрессии

После запуска процедуры анализа (см. п. 11.1 Простая линейная регрессия) выводим таблицу результатов анализа. В таблице результатов (рис. 67), обращаем внимание, прежде всего, на коэффициенты уравнения и их значимость.

N=14	Regression Summary for Dependent Variable: Высота, м (Spreadsheet18) R= ,91804726 R?= ,84281078 Adjusted R?= ,79565401 F(3,10)=17,873 p<,00024 Std.Error of estimate: ,31369					
	b*	Std.Err. of b*	b	Std.Err. of b	t(10)	p-value
Intercept			0,605738	0,560668	1,080387	0,305337
Диаметр, см	0,826831	0,141507	0,428784	0,073384	5,843018	0,000163
Прирост по высоте, см	0,056360	0,128609	0,003242	0,007398	0,438225	0,670538
Диаметр кроны, м	0,161488	0,138657	0,433536	0,372242	1,164662	0,271188

Рис. 67. Результаты расчета множественной регрессии

В соответствии со значениями коэффициентов (столбец b), уравнение регрессии имеет следующий вид:

$$Y = 0{,}606 + 0{,}429 \cdot X_1 + 0{,}003 \cdot X_2 + 0{,}433 \cdot X_3$$

Достоверность значений коэффициентов определяется по столбцу уровня значимости (*p-level*). Можем видеть, что достоверным на уровне значимости 0,05 является только коэффициент при диаметре ствола (значение менее 0,05), что подтверждается красным шрифтом текста данной строки. Остальные коэффициенты недостоверны. Таким образом, прирост по высоте и диаметр кроны мы можем убрать из уравнения регрессии, несмотря на высокое значение коэффициента детерминации *(R?),* равное 84%, и высокую достоверность всего уравнения регрессии (p<0,00024). Такие высокие значения определяются, по-видимому, только высокой связью диаметра с высотой.

Для проверки этого предположения оставляем в уравнении только те переменные, коэффициенты которых достоверны, т. е. диаметр ствола. Результатом такого анализа является таблица (рис. 68), из которой следует, что по сравнению с предыдущим случаем коэффициент детерминации уменьшился незначительно (82%), а уровень значимости всего уравнения даже возрос (p<0.00001).

N=14	Regression Summary for Dependent Variable: Высота, м (Spreadsheet18) R= ,90545769 R?= ,81985362 Adjusted R?= ,80484142 F(1,12)=54,612 p<,00001 Std.Error of estimate: ,30656					
	b*	Std.Err. of b*	b	Std.Err. of b	t(12)	p-value
Intercept			1,196790	0,231898	5,160852	0,000237
Диаметр, см	0,905458	0,122524	0,469559	0,063540	7,390027	0,000008

Рис. 68. Результаты расчета множественной регрессии

Таким образом, в окончательное уравнение регрессии из трёх исследованных переменных целесообразно включать только диаметр ствола, остальные переменные для регрессионной зависимости не представляют интереса. Также, как и в отношении простой линейной регрессии мы можем выполнить все более подробные анализы и построить необходимые графики (см. п. 11.1).

12.3. Нелинейная регрессия

Довольно часто при выполнении регрессионного анализа линейная зависимость недостаточно хорошо описывает зависимость между переменными. В этом случае целесообразно использовать нелинейную регрессию.

Пример расчёта в Statistica. Необходимо найти уравнение взаимосвязи между диаметром и объёмом ствола дерева сосны:

Диаметр ствола (D), см	3,7	4,5	1,7	1,6	5,8	5,3	5,0	2,9	3,1	3,4	1,7	2,1
Объём ствола (V), см^3	2759	4673	470	376	8965	5606	5200	1531	1374	1954	357	573

Поскольку очевидно, что между диаметром и объемом ствола связь нелинейная, а параболическая целесообразно использовать в данном случае нелинейную зависимость, а уравнение вида:

$$Y = a + bX + cX^2 + dX^3,$$

где X – независимая переменная; Y – зависимая переменная; a, b, c, d – коэффициенты уравнения. В данном случае использование полинома третьей степени связано тем, что объем является кубической функцией линейных размеров.

В пакете *Statistica* расчёт нелинейной регрессии запускаем при помощи строки *Advanced Linear/Nonlinear Models* в меню *Statistics*. Далее запускаем модуль *Nonlinear Estimation* (нелинейное оценивание) – Рис. 69.

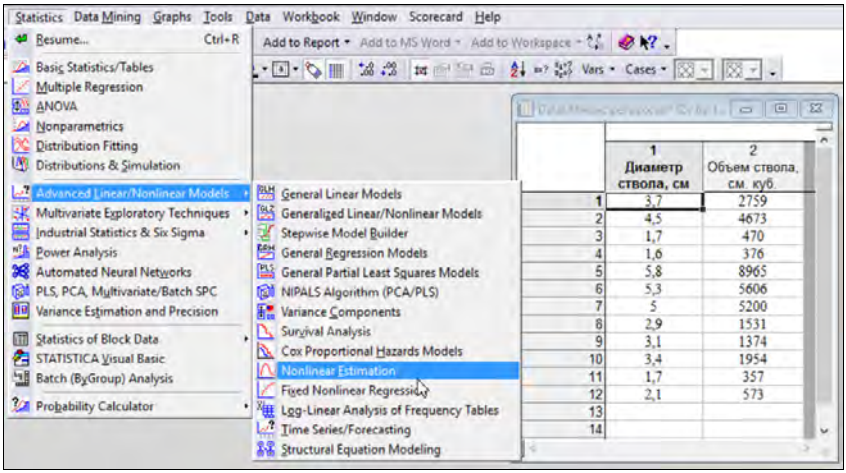


Рис. 69. Меню выбора нелинейной регрессии

Выбираем *User-specified regression, least squares* (регрессия, определяемая пользователем, метод наименьших квадратов) и кнопкой *Function to be estimated* (оцениваемая функция) запускаем диалоговое окно. После ввода (OK) в поле *Estimated function* набираем необходимую функцию (рис. 70). В данной формуле v1 и v2 – используемые переменные (сокращенное от *variable 1* и *variable 2*). Транскрипцию набора необходимой функции, в частности знаки умножения (*), степенной функции (^) и другие можно увидеть ниже в этом же окне (рис. 70).

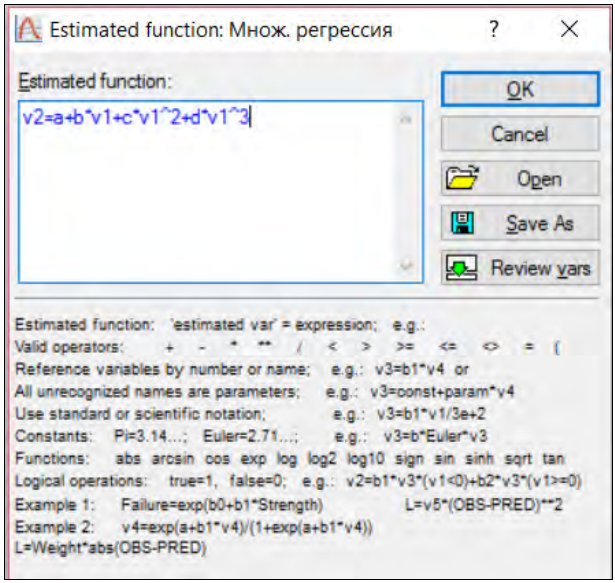


Рис. 70. Окно ввода заданной пользователем функции регрессии

После подтверждения модели (*OK*) запускаем процесс оценки. В диалоговом окне *Results* кнопкой Summary: *Parameter estimates* выводим таблицу результатов анализа (рис. 71), а кнопкой *Fitted 2D function & observed vals* – график (рис. 72). Коэффициенты уравнения приведены в первом столбце с заголовком *Estimate*. В других столбцах приведены стандартные ошибки коэффициентов (Standard error), достоверность коэффициентов (*p-level*), а также верхний и нижний доверительные интервалы (*Conf Limit*).

	Model is: v2=a+b*v1+c*v1^2+d*v1^3 (Множ. регрессия) Dep. Var. : Объем ствола, см. куб. Level of confidence: 95.0% (alpha=0.050)					
	Estimate	Standard error	t-value df = 8	p-value	Lo. Conf Limit	Up. Conf Limit
a	-2197,68	2910,273	-0,755147	0,471805	-8908,79	4513,419
b	2368,68	2884,964	0,821044	0,435403	-4284,06	9021,424
c	-670,48	860,955	-0,778761	0,458534	-2655,84	1314,887
d	100,15	79,108	1,266056	0,241115	-82,27	282,577

Рис. 71. Результаты расчета заданной пользователем функции регрессии

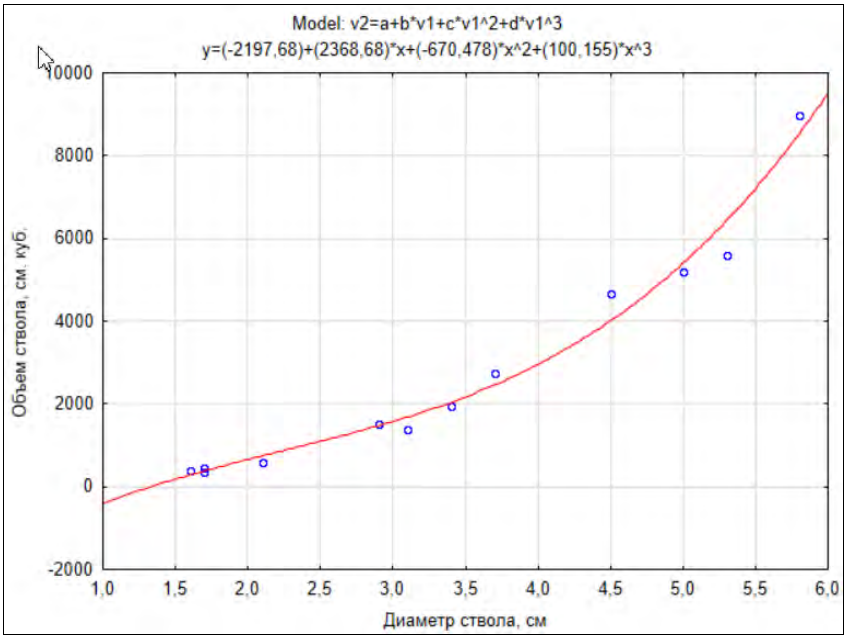


Рис. 72. График заданной пользователем функции регрессии

В соответствии с результатами анализа (см. Рис. 71) запрошенное уравнение регрессии с округлением коэффициентов до целых значений имеет следующий вид:

$$Y = -2797 + 2369 \cdot X - 670 \cdot X^2 + 100 \cdot X^3$$

Построение регрессионных кривых в MS Excel. Некоторые виды регрессионных зависимостей достаточно подробно можно проанализировать

в *MS Excel*. Для этого сначала необходимо построить график зависимости одной переменной от другой. После ввода данных в книгу *MS Excel* и выделения нужных данных (необходимо для упрощения процесса построения диаграммы), в меню "Вставка" на вкладке «Диаграммы» выбираем нужный тип диаграммы. В данном случае удобнее всего воспользоваться точечной диаграммой (рис. 73).

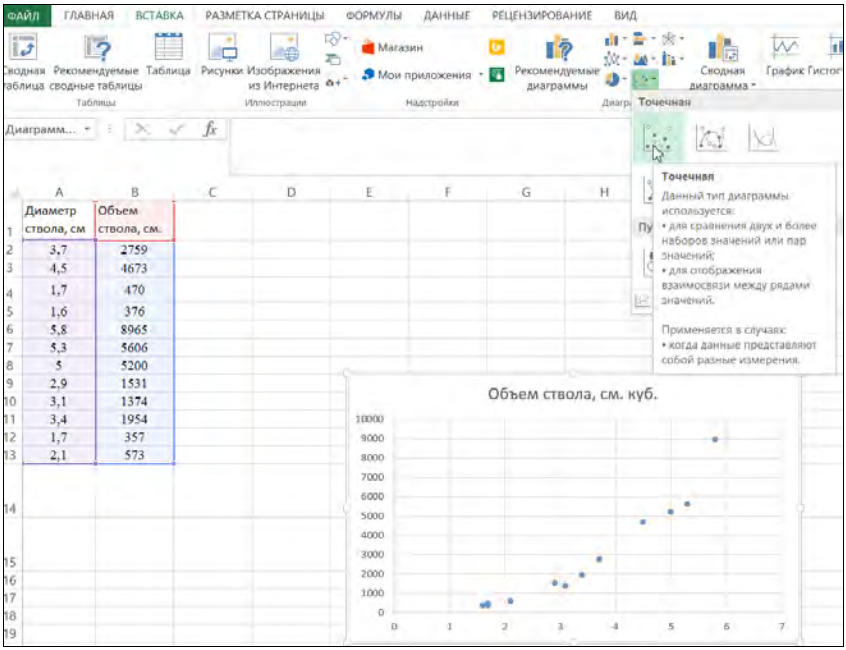


Рис. 73. Построение диаграммы в MS Excel

В итоге будет построен первоначальный черновой вариант диаграмма, после чего во вкладке *«Добавить элемент диаграммы»* можем последовательно дополнить все необходимые элементы диаграммы: название диаграммы, подписи и параметры осей, линию тренда и др. (рис. 74).

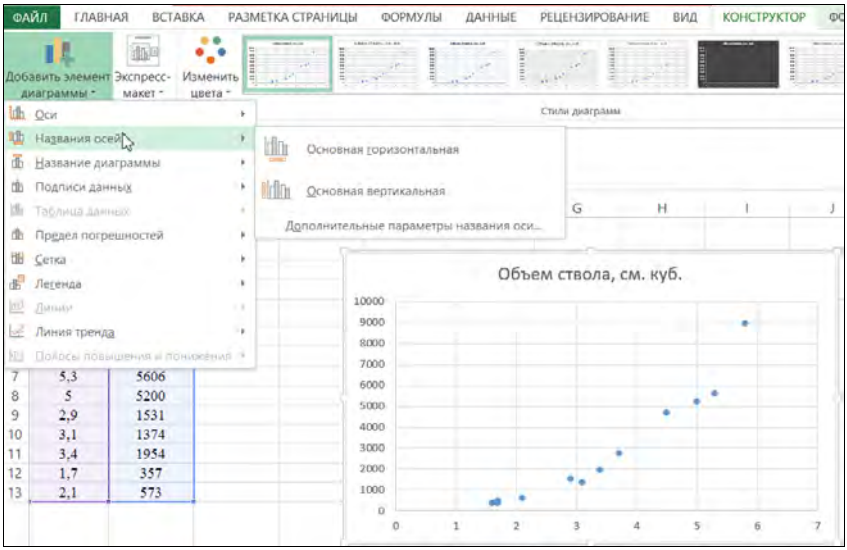


Рис. 74. Редактирование параметров диаграммы в MS Excel

В результате строим диаграмму со всеми необходимыми нам атрибутами: подписи и диапазон значений осей, линия тренда, коэффициент детерминации и др. (рис. 75). При помощи линии тренда *MS Excel* позволяет подбирать необходимую исследователю аппроксимацию: линейная $(a+bX)$, логарифмическая $(a \cdot ln(X)+b)$, полиномиальная $(a + bX + cX^2 + .. + nX^n)$ вплоть до полинома 6-й степени, степенная (aX^b), экспоненциальная $(a \cdot e^{bX})$, линейная фильтрация (скользящее среднее по нескольким точкам) и др.

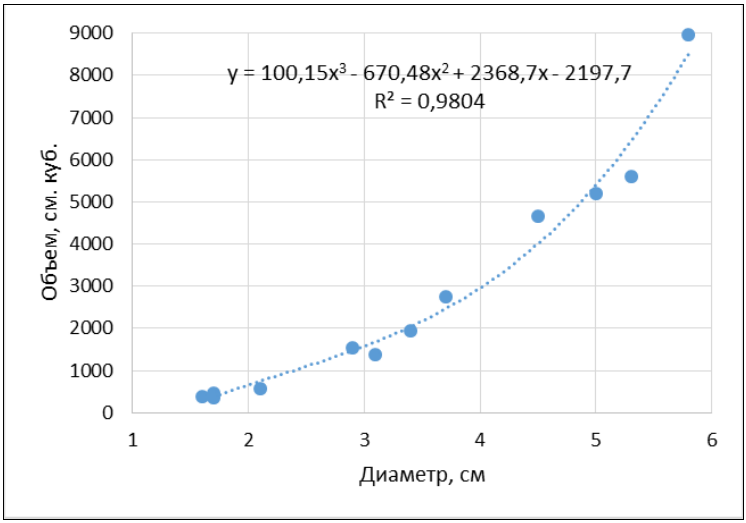


Рис. 75. Результат построения диаграммы с нелинейной зависимостью переменных

Можем отметить, что приведенная регрессионная зависимость характеризуется коэффициентом детерминации R^2=0,98, то есть полученная линия регрессии на 98% объясняет варьирование зависимой переменной, то есть зависимость практически функциональная.

12. НЕКОТОРЫЕ МЕТОДЫ МНОГОМЕРНОГО АНАЛИЗА ДАННЫХ

В предыдущих разделах рассматривались методы, в которых анализируется один или два признака объекта. Например, в регрессионном анализе мы рассматривали зависимость высоты дерева от его диаметра, т. е. использовали два признака, две характеристики для каждого объекта (дерева). Тем не менее, довольно часто возникает необходимость изучения объектов с одновременным учетом значительного числа их характеристик. Раздел математической статистики, исследующий эксперименты, в которых каждый объект характеризуется несколькими показателями, называется многомерным статистическим анализом. Многомерные методы в настоящее время привлекают пристальное внимание исследователей и находят широкое применение. Так, например, в лесоводстве многомерные классификации широко применяются при изучении популяционной структуры

видов, в лесной типологии, ресурсоведении, генетике, фенетике и других разделах лесной науки и практики. В зависимости от решаемых задач (п. 5) используют различные виды многомерного анализа: дискриминантный, факторный, ковариационный, кластерный анализы, многомерное шкалирование, анализ таблиц сопряжённости. Кратко рассмотрим применение некоторых из этих методов.

13.1. Кластерный анализ

Кластерный анализ позволяет разбивать изучаемую совокупность на группы сходных объектов одновременно по нескольким признакам. При этом в качестве меры разделения объектов по группам является расстояние между этими объектами в многомерном пространстве изучаемых признаков. По степени сходства-различия между объектами они и объединяются в группы однородных объектов, называемые кластерами.

Существует несколько методов кластерного анализа. Наиболее распространены в практике лесного хозяйства метод К-средних (K-means) и иерархическая кластеризация (Hierarchical).

Кластеризация К-средних разбивает все множество имеющихся объектов на заданное пользователем число кластеров таким образом, чтобы средние значения по каждой из переменных для кластеров максимально различались, то есть кластеры (группы) были как можно сильнее удалены друг от друга по изучаемому комплексу признаков.

Иерархическая кластеризация является наиболее гибким из методов кластерного анализа, поэтому нашла наиболее широкое применение в лесоводственных исследованиях. Она позволяет исследователю детально изучить структуру различий между объектами и выбрать оптимальное число кластеров. Рассмотрим данный метод кластеризации более подробно.

Основой кластерного анализа является расчёт "расстояний" (меры расстояния) между объектами по совокупности их количественных характеристик и группировка этих объектов таким образом, чтобы внутри групп расстояние между объектами было минимальным, а между группами – максимальным. Любой количественный признак можно рассматривать как меру расстояния между объектами. Так, например, разница по высоте между двумя деревьями является мерой расстояния между ними по признаку «высота». Если же имеются измерения нескольких признаков для каждого объекта (высота, диаметр ствола, длина хвои и т. д.), кластерный анализ позволяет свести это множество признаков (расстояний) к одной обобщённой мере. В том случае, когда используется очень ограниченное количество переменных (не более трёх), то расстояние между объектами

можно представить наглядно. Допустим, у нас есть данные о высоте и диаметре двух деревьев. Изобразим расстояния в пространстве признаков графически. Отложим высоту по оси *X*, а диаметр – по оси *Y*. Отметим на этом графике два дерева точками (рис. 76). Расстояние между этими точками может служить обобщённой мерой расстояния между объектами в многомерном (в данном случае двухмерном) пространстве признаков.

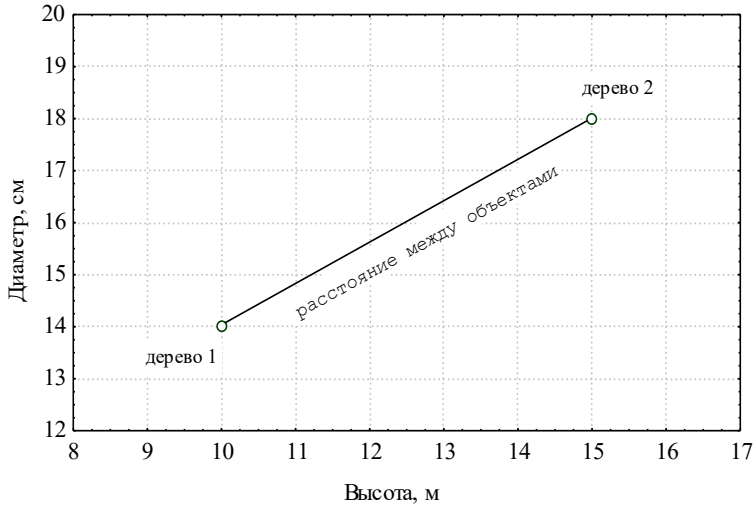


Рис. 76. Графическое изображение расстояния между объектами

Если признаков больше трёх, то наглядно представить расстояние между ними затруднительно, тем не менее это расстояние можно вычислить математическими методами.

В кластерном анализе используются различные меры расстояния между объектами (правильнее было бы назвать их мерами несходства).

Евклидово расстояние (Euclidean distance). Наиболее общий тип расстояния, представляющий собой аналог многомерный аналог двухмерного геометрического расстояния между точками на рис. 76. Евклидово расстояние между объектами *x* и *y* вычисляется следующим образом:

$$d_{xy} = \sqrt{\sum (x_i - y_i)^2} \ ,$$

где d_{xy} – расстояние между объектами *x* и *y*; x_i – значение *i*-го признака для объекта *x*; y_i– значение *i*-го признака для объекта *y*.

Чтобы избежать применения квадратного корня, а также для придания большего веса более отдаленным друг от друга объектам, часто величина расстояния возводится в квадрат *(d^2xy)* и называется «квадратичным евклидовым расстоянием».

Манхэттенское расстояние (расстояние городских кварталов – city block distance). Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как расстояние Евклида. Однако следует отметить, что для Манхэттенского расстояния влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат). Манхэттенское расстояние вычисляется по формуле:

$$d_{xy} = \sum |x_i - y_i| .$$

Расстояние Чебышева (Chebychev distance). Это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате (каким-либо одним измерением). Расстояние Чебышева вычисляется по формуле:

$$d_{xy} = max(x_i - y_i) .$$

Расстояние Махаланобиса (Mahalanobis distance). Это расстояние связано с корреляциями переменных. Когда корреляция между переменными равна нулю, расстояние Махаланобиса эквивалентно квадратичному евклидову расстоянию.

Существуют и другие меры расстояний между объектами в многомерном пространстве (напр. метрики Минковского, степенное расстояние и др.).

Процедура кластерного анализа рассчитывает расстояния между всеми парами объектов, на основе чего составляется матрица расстояний между изучаемыми объектами.

Следующим этапом анализа является собственно кластеризация – последовательное объединение объектов в кластеры, т. е. в группы, где сходство между объектами выше (расстояние между ними меньше), чем сходство с другими объектами или кластерами. При этом методы кластеризации довольно разнообразны, результаты могут различаться в зависимости от применяемого метода, что зачастую приводит к неоднозначности интерпретации результатов. Поэтому кластерный анализ служит прежде всего для выявления скрытой структуры имеющегося материала. В дальнейшем к полученным кластерам могут применяться теоретически более

обоснованные статистические методы (корреляционный, дисперсионный, регрессионный анализ).

В результате кластерного анализа можем получить дендрограмму сходства объектов. На Рис. 77 приведён пример такой дендрограммы. По оси *X* откладываются объекты исследования (опытные участки), по оси *Y* – расстояние между объектами (в данном случае это Евклидово расстояние), вычисленное по комплексу показателей этих объектов (влажность почвы, содержание гумуса, процентное соотношение различных видов травянистых растений и т. п.).

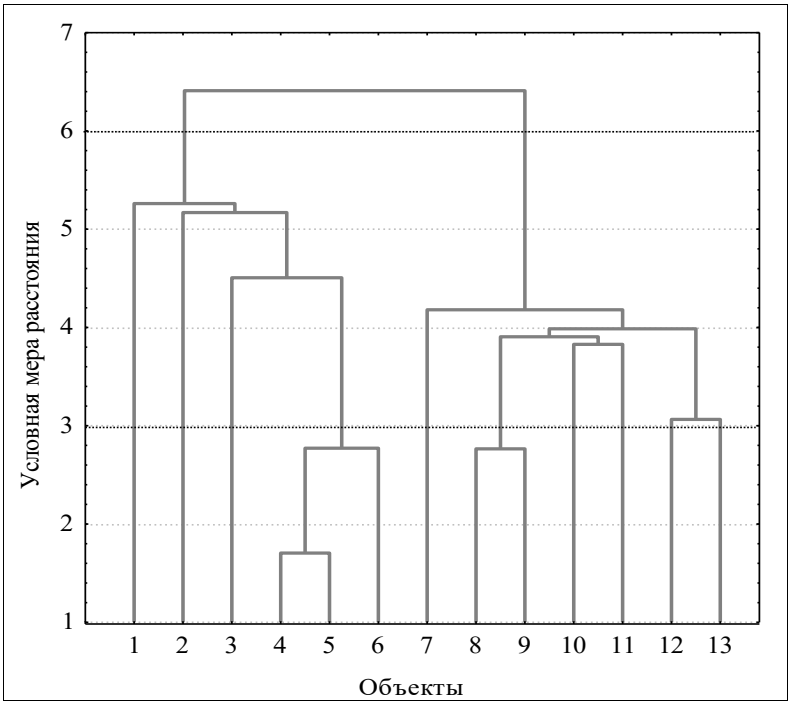


Рис. 77. Дендрограмма сходства объектов

При анализе дендрограммы удобнее всего идти снизу вверх. Можем констатировать, что в нижней части каждый объект является отдельным кластером (каждый объект сам по себе является отдельным кластером). Теперь представим, что постепенно мы "ослабляем" критерий того, какие объекты являются уникальными, а какие нет. Другими словами, увеличи-

ваем расстояние, в пределах которого два или более объекта объединяются в один кластер. В результате, мы связываем вместе всё большее и большее число объектов и, соответственно, кластеров. Момент объединения тех или иных объектов отмечается горизонтальными линиями. Так, например, когда мы достигаем условного расстояния, равного трём (пунктирная линия), у нас объединёнными оказываются объекты 4, 5, 6 (один кластер), а также 8 и 9 (второй кластер). Все остальные объекты являются пока отдельными кластерами. Таким образом, всего мы имеем 10 кластеров (количество пересечений с пунктирной линией). При условном расстоянии, равном шести, мы имеем два кластера: первый состоит из объектов 1-6, второй – из 7-13. При расстоянии, равном семи, все объекты объединены в один общий кластер.

Исследователь на основании собственного опыта и знания изучаемых объектов, а также используя визуальный анализ такой дендрограммы решает, какое количество кластеров (групп) целесообразно выделить в изучаемой совокупности и как их интерпретировать (например, участки группируются по возможным градациям лесорастительного эффекта).

Пример построения дендрограммы в Statistica. В географических культурах измерены основные биометрические показатели потомств 15-ти различных происхождений и получены средние значения по признакам:

Номер потомства	Диаметр, см	Высота, см	Сумма приростов в высоту за последние 3 года, см	Площадь кроны, $м^2$
1	4,7	340,7	95,1	3,0
2	4,1	324,5	102,2	2,8
3	3,9	287,0	94,8	2,1
4	3,4	277,5	93,7	2,5
5	3,3	263,7	74,6	1,3
6	4,1	260,3	78,6	2,0
7	2,9	246,9	70,9	1,6
8	2,6	241,1	81,8	2,2
9	7,2	375,5	151,8	5,1
10	11,9	463,5	132,4	10,9
11	8,3	427,0	135,6	5,4
12	10,1	403,0	148,8	7,9
13	6,1	323,5	112,9	4,2
14	6,0	320,0	124,3	3,3
15	5,4	277,0	92,4	2,9

Необходимо построить дендрограмму сходства для этих потомств и попытаться выявить какие из географических происхождений близки по комплексу признаков, а какие значительно различаются. Выстраиваем данные стандартным для *Statistica* образом: отдельный столбец – отдельный признак, отдельная строка – отдельный объект (в нашем случае объектом является потомство географического происхождения).

В пакете *Statistica* многомерные методы собраны в строке *Multivariate Exploratory Technique* (многомерные методы исследования) меню *Statistics*. Далее выбираем *Cluster Analisys* – кластерный анализ (рис. 78).

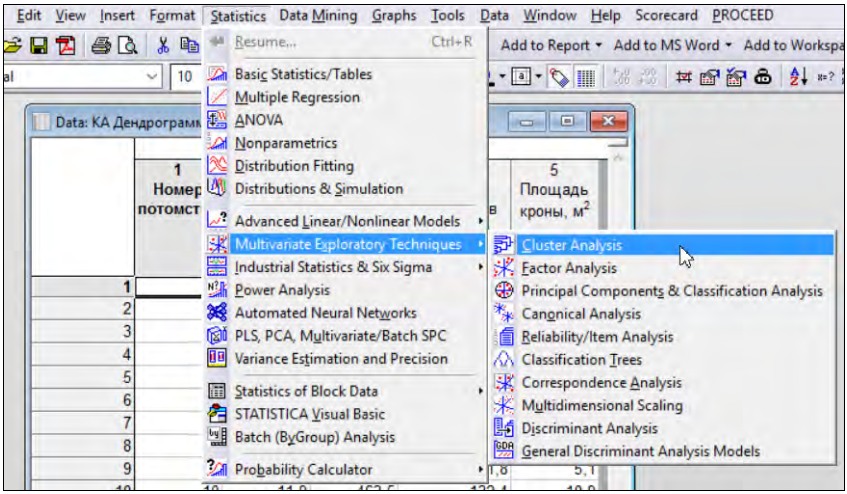


Рис. 78. Выбор кластерного анализа в основном меню

Из меню *Cluster Analisis* (кластерный анализ) запускаем необходимый нам метод кластеризации *Joining (tree clustering)* – объединение (дерево кластеризации). В появившемся диалоговом окне во вкладке *Advanced* выбираем необходимые параметры анализа. В частности, в окне Input file выбираем строки или столбцы нам необходимо разбить по кластерам (в нашем случае для кластеризации, потомства разных географических происхождений у нас расположены по строкам – Cases (rows). В окне Amalgamation (linkage) rule (правило для объединения) – оставляем Single Linkage (метод «одиночной связи»). В окне Distance mesure (мера расстояния) выбираем упомянутое выше Евклидово расстояние - Euclidean distance, а кнопкой *Variables* выбираем для анализа все биометрические показатели потомств (рис. 79). Далее запускаем процесс кнопкой OK.

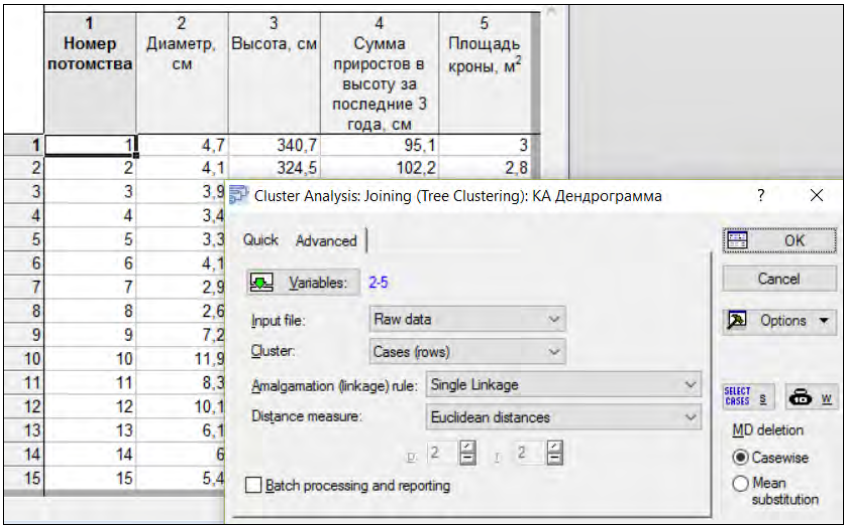


Рис. 79. Выбор основных параметров кластерного анализа

После запуска процедуры расчёта *(OK)* появляется панель с результатами анализа (рис. 80), в которой на вкладке *Advanced* мы можем выбрать ориентацию дендрограммы: горизонтальную *(Horizontal hierarchical tree plot)* или вертикальную *(Vertical icicle plot)*, матрицу расстояний между объектами *(Distance matrix)* и другие данные. Вертикальная дендрограмма для наших данных приведена на Рис. 81.

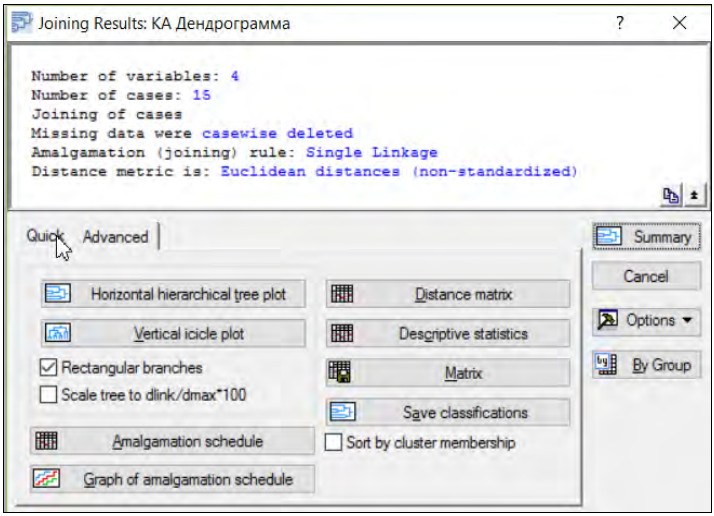


Рис. 80. Промежуточные результаты кластерного анализа

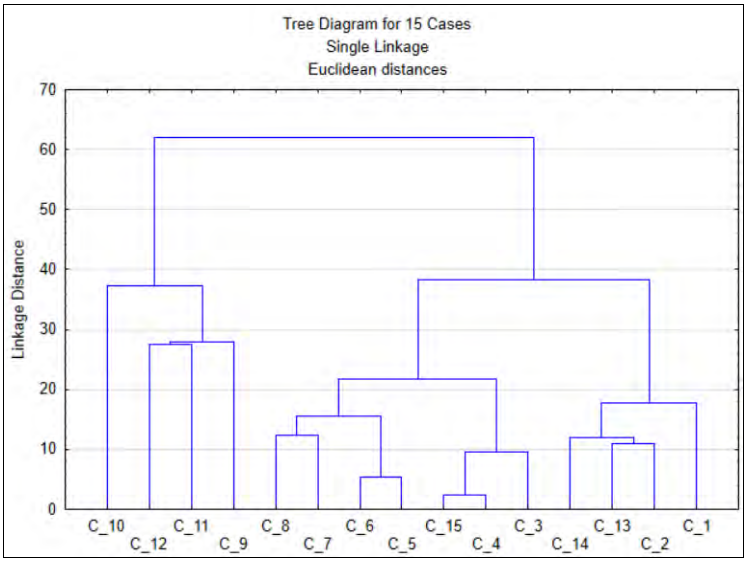


Рис. 81. Вертикальная дендрограмма сходства объектов

Можем отметить, что потомства различных географических происхождений группируются в два крупных, достаточно удалённых друг от друга кластера на дистанции 60 по оси Y. Далее, приблизительно на расстоянии равном 35 каждая из этих наиболее крупных аггрегаций в свою очередь разделяется еще на две группы. Таким образом на условном расстоянии 30 мы получаем четыре кластера:

кластер №1: отдельное от всех потомство 10;

кластер №2: потомства 9, 11, 12;

кластер №3: потомства 3, 4, 5, 6, 7, 8, 15;

кластер №4: потомства 1, 2, 13, 14.

При изменении условного расстояния (например, если берем эту меру равной 15) – получаем большее количество кластеров. Потомства внутри полученных кластеров по комплексу изученных биометрических показателей ближе друг к другу, чем потомства разных кластеров. Различия могут быть обусловлены генетическими, климатическими и др. факторами, поэтому такой анализ является разведочным для более подробных дальнейших исследований.

Пример расчёта кластеризации "К-средних" в Statistica. При проведении многих лесоводственных исследований часто появляется необходимость определить величину различий между деревьями в популяции по их размерам (высоте, диаметру, объёму ствола). Обычно всю совокупность деревьев исследуемой популяции разделяют на строго определенное количество групп. Довольно часто в лесоводстве используется разделение деревьев по размерам на пять групп (например, классификация Крафта). При этом исследователи разделяют совокупность деревьев на пять равных групп по высоте или диаметру в какой-либо один год наблюдения (такие группы называются ранговыми). При этом ранговое положение дерева в популяции со временем может меняться, что усложняет отнесение его к какой-либо ранговой группе. В связи с этим желательно учитывать данные о высотах деревьев на некотором изучаемом промежутке времени (за ряд лет), а не только за какой – либо отдельный год изучения. Если у нас есть данные о размерах деревьев за несколько лет наблюдения можно использовать эту информацию для деления деревьев на группы. Удобно сделать это при помощи метода кластерного анализа "*К*-средних".

Допустим, у нас есть данные о росте 25-ти деревьев сосны:

Номер дерева	Высота в возрасте, лет									
	5	6	7	8	9	10	11	12	13	14
1	128	142	180	230	274	314	340	368	405	450
2	73	85	100	131	167	181	199	232	274	315
3	90	134	176	226	257	282	317	357	402	450
4	28	45	85	132	178	238	313	340	380	430
5	77	102	122	150	191	221	249	295	345	385
6	45	60	95	113	134	161	179	210	248	280
7	46	50	74	104	129	157	174	182	194	210
8	69	93	120	156	189	230	264	301	354	400
9	117	151	198	248	272	290	312	347	397	450
10	49	62	82	118	156	201	239	274	321	370
11	66	83	90	110	134	174	201	248	305	360
12	136	156	198	243	289	338	364	386	420	450
13	44	71	105	150	158	175	194	226	286	340
14	132	158	194	211	224	248	269	315	365	410
15	48	61	86	115	149	164	193	232	277	320
16	37	50	72	83	103	118	136	148	166	195
17	36	43	59	76	93	123	140	160	192	235
18	26	29	37	47	75	117	149	187	235	285
19	47	52	56	77	87	107	147	165	190	200
20	23	30	47	60	74	94	110	136	171	215
21	17	30	37	45	64	74	86	107	152	190
22	18	29	35	44	58	68	83	109	127	150
23	21	35	39	46	71	92	102	118	136	160
24	31	50	67	79	87	101	116	128	140	175
25	107	119	133	147	188	216	245	279	314	350

Необходимо разделить эту совокупность на пять групп в соответствии с их высотами в диапазоне возрастов 5-14 лет. При этом полученные группы (назовём их группами роста) могут содержать разное количество деревьев. Основной критерий разделения: наибольшее расстояние между группами по комплексу измерений высоты растений за ряд лет.

Для выполнения процедуры анализа в строке *Multivariate Exploratory Technique* (многомерные методы исследования) меню *Statistics* аналогично примеру с иерархической кластеризацией (см. выше) выбираем *Cluster Analisis* (кластерный анализ), в открывшемся окне выбираем метод K-means clustering (кластеризация «К-средних») и запускаем процедуру анализа (OK).

В открывшемся диалоговом окне *Cluster Analisis* во вкладке *Advanced* выбираем необходимые параметры анализа: в окне Input file выбираем объекты, которые группируются в строки *(Cases)* или столбцы *(Variables)*. В нашем случае группируются деревья, расположенные по строкам, поэтому выбираем *Cases (rows)*. Кроме того, в окошке *Number of Clusters* (количество кластеров) вводим количество групп, на которые необходимо разделить объекты, в нашем случае – пять штук (рис. 82).

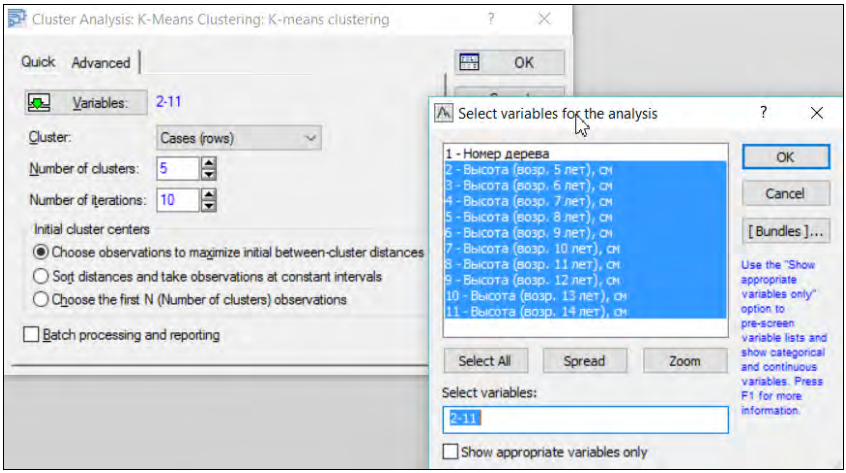


Рис. 82. Выбор комплекса переменных кластерного анализа «К-средних»

Начальное количество кластеров выбираем по принципу максимизации начального расстояния между кластерами (опция *Choose observations to maximize initial between-cluster distances*).

Кнопкой *Variables* выбираем переменные, все десять высот (рис. 82). После запуска процедуры анализа *(OK)* выводится панель с результатами анализа, в рамках которой во вкладке *Advanced* (рис. 83) можем вывести значимость различий между полученными группами на основе дисперсионного анализа *(Analysis of variance)*, средние значения переменных (высот) для каждого кластера и расстояния между кластерами по Евклидову расстоянию *(Summary: Cluster means & Euclidean distances)*, график средних значений для выявленных групп *(Graph of means)*, описательную статистику (среднее, дисперсия и стандартное отклонение) для каждой группы *(Descriptive statistics for each cluster)*, а также объекты, входящие в каждую группу и расстояние от этих объектов до центра кластера *(Members of each cluster & distances)*. Кнопкой *Save classifications and distances* можем вывести в отдельное окно анализируемые количественные переменные, к каким кластерам они принадлежат и расстояние отдельных объектов от центров соответствующих кластеров (рис. 84).

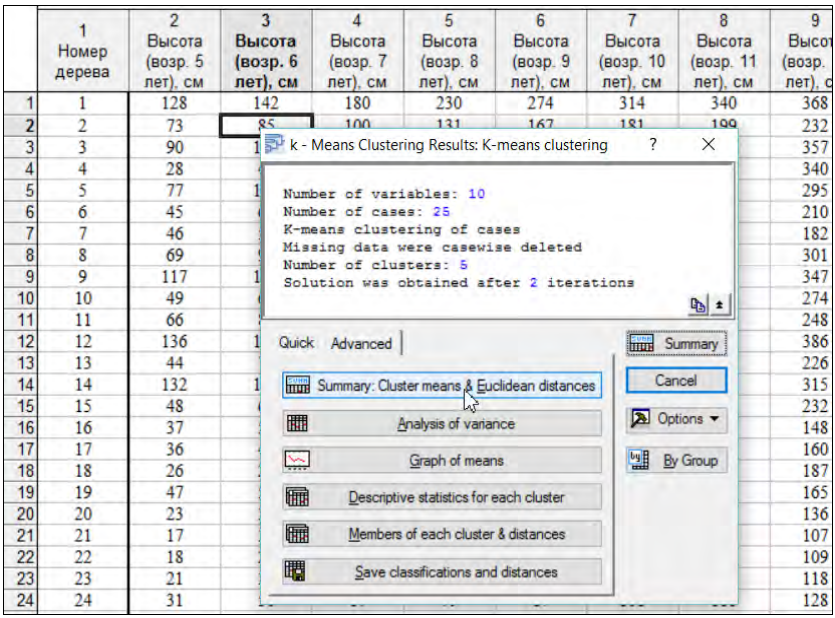


Рис. 83. Промежуточная панель кластерного анализа «К-средних»

Data: Spreadsheet9 (4v by 25c)

K-means clustering

	1 Номер дерева	2 CASE_NO	3 CLUSTER	4 DISTANCE
C_1	1	1	1	11,67
C_2	2	2	2	10,61
C_3	3	3	1	12,79
C_4	4	4	4	33,92
C_5	5	5	4	10,05
C_6	6	6	2	20,02
C_7	7	7	3	20,22
C_8	8	8	4	9,58
C_9	9	9	1	8,11
C_10	10	10	4	22,03
C_11	11	11	2	17,80
C_12	12	12	1	25,53
C_13	13	13	2	11,76
C_14	14	14	1	32,67
C_15	15	15	2	6,33
C_16	16	16	3	16,25
C_17	17	17	3	5,47
C_18	18	18	3	28,12
C_19	19	19	3	11,05
C_20	20	20	5	15,96
C_21	21	21	5	9,60
C_22	22	22	5	14,98
C_23	23	23	5	7,56
C_24	24	24	5	15,02
C_25	25	25	4	25,42

Рис. 84. Принадлежность объектов к кластерам и расстояния от центра кластера

При выводе графика средних значений по кластерам кнопкой *Graph of means* (график по средним значениям) получаем соответствующий график средних значений по рассчитанным кластерам, где каждый из кластеров в максимальной степени удален от других (рис. 85).

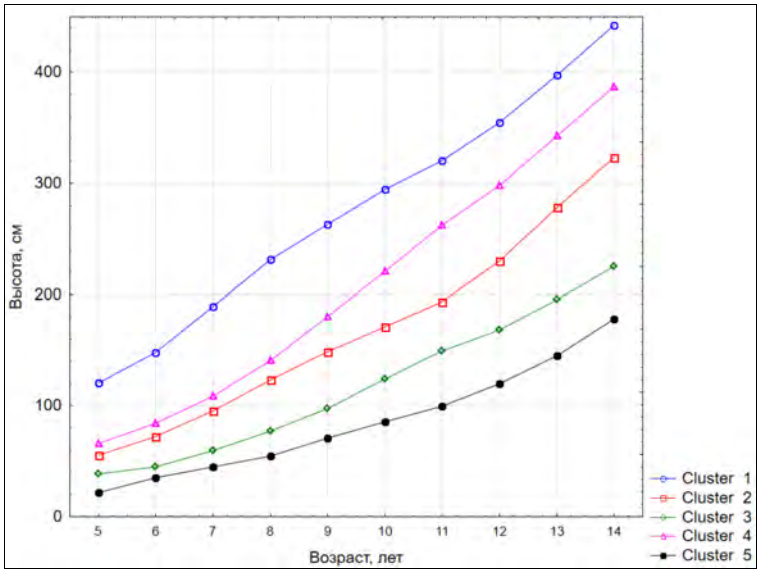


Рис. 85. График средних значений признака по кластерам

В соответствии с полученным графиком, самой быстрорастущей группой является кластер №1 (можно интерпретировать как первую группу роста - лидеры), далее следуют кластеры №4, 2, 3 (вторая, третья и четвёртая группы роста) и наиболее медленнорастущий кластер №5 (пятая группа роста или деревья-аутсайдеры). Так как при кластеризации деревьев используется информация за несколько лет наблюдения, то полученные группы роста являются интегральной оценкой скорости роста дерева. А поскольку группы чётко отграничиваются друг от друга на всем изучаемом промежутке времени (рис. 80), можем утверждать, что деревья разных групп роста достоверно отличаются друг от друга по скорости роста в высоту, что при желании можно подтвердить при помощи дисперсионного анализа (кнопка *Analysis of variance*), показывающего, что различия между кластерами достоверны на уровне значимости 0,05.

13.2. Дискриминантный анализ

Дискриминантный анализ позволяет изучать различия между группами объектов одновременно по нескольким показателям, причем, в отличие от кластерного анализа группы объектов при выполнении дискриминантного анализа уже заранее известны. Для каждой из групп известны

измерения нескольких количественных характеристик на основе дискриминантного анализа мы можем найти способ, при помощи которого на основе этих количественных характеристик можем определить принадлежность объекта к той или иной заранее известной группе. Это позволяет для новых объектов на основе их количественных характеристик предсказывать группы, к которым эти объекты относятся, что и является задачей дискриминантного анализа.

В какой-то мере дискриминантный анализ является аналогом множественного регрессионного анализа с той разницей, что на основе набора количественных характеристик объекта в регрессионном анализе мы рассчитываем результирующую количественную переменную, а в кластерном анализе результатом является номинальная переменная. То есть, в регрессионном анализе зависимая переменная – качественная, а в дискриминантном анализе – номинальная (см. табл.4). При этом дискриминантный анализ решает, по сути, те же задачи, что и множественный регрессионный анализ: предсказание зависимой переменной (категорий номинального признака) и определение того, какие независимые переменные лучше всего подходят для такого предсказания.

13.3. Факторный анализ

При исследовании таких сложных систем, какими являются биологические объекты, мы часто не можем непосредственно измерить факторы, определяющие свойства этих объектов, а зачастую нам неизвестно даже число и содержательный смысл этих факторов. Для непосредственных измерений могут быть доступны величины, зависящие от этих факторов. Когда влияние неизвестного фактора проявляется в нескольких измеряемых признаках, эти признаки могут иметь тесную связь друг с другом, и общее количество влияющих факторов может быть гораздо меньшим, чем число измеряемых переменных. Факторный анализ позволяет перейти от этих переменных к факторам, влияющим на них. При этом, во-первых, можно уменьшить количество переменных, используемых в анализе (уменьшение избыточности данных), а во-вторых, возможно найти содержательное толкование факторов, определяющих свойства объектов, и, в-третьих, определить, какие признаки объекта наиболее представительны, т. е., по каким показателям можно наиболее точно определить степень влияния того или иного нерегистрируемого фактора.

ЗАКЛЮЧЕНИЕ

В данном учебном пособии мы придерживались классических подходов в статистической обработке полевых данных. Тем не менее, пристального внимания заслуживают современные подходы к обработке экспериментальных данных, такие, например, как байесовские методы оценки вероятностей. В силу ограниченного объема учебного пособия и соответствующей учебной дисциплины в него не вошли многие методы статистической обработки полевых данных. Это, прежде всего, относится к многомерным методам, таким как дискриминантный и факторный анализы, полезным при исследовании большого числа зависимых переменных. Недостаточно полно представлены такие методы изучения зависимостей как регрессионный и значительное количество непараметрических методов исследования. Отсутствует информация по таким перспективным направлениям как многомерное шкалирование, анализ таблиц сопряженности, нейросетевых моделей, марковские процессы принятия решений, теории игр и др. Кроме того, в контексте лесоводственного направления исследований гораздо большего внимания, на наш взгляд, заслуживают вопросы методики планирования эксперимента. Эти и другие методы могут быть успешно использованы в лесном деле. Опираясь на предложенные в данном учебном пособии материалы, можно самостоятельно освоить новые методы и грамотно применять их в своей практической работе.

В данном учебном пособии основное внимание при подборе методов статистической обработки данных уделялось наиболее широко применяющимся стандартным статистическим методам, имеющим очевидную интерпретацию и рациональную последовательность применения. В результате была описана пошаговая последовательность вычисления статистических показателей на основе наглядных примеров из лесоводственной практики с обработкой данных на основе применения современных программ компьютерного обеспечения MS Excel 2013 и Statistica 12.

Основываясь на стандартной общепринятой логике статистического исследования, экспериментатор может провести самостоятельное исследование зависимостей и сделать обоснованные выводы и заключения относительно объектов, характеризующихся значительной изменчивостью. Выявленные в результате таких исследований закономерности могут быть положены в основу практического применения и оптимизации технологических, экологических и экономических решений по ведению лесного хозяйства как при естественном формировании лесных фитоценозов, так и под воздействием хозяйственных мероприятий. Именно практическая направленность является одной из особенностей данного учебного пособия. Более

подробно познакомится с теоретической стороной вопроса можно, обратившись к специальной литературе, часть которой приведена в библиографическом списке.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Исследование зависимостей. - М.: Финансы и статистика, 1983. - 487с.
2. Бондаренко А.С. Применение кластерного анализа для оценки скорости роста древесных пород // Восстановление лесов и ресурсо- и энергосберегающие технологии лесного комплекса. – Воронеж: Воронеж. гос. лесотехн. акад., 2000. – С. 40-42.
3. Боровиков В. П. Statistica. Искусство анализа данных на компьютере: для профессионалов. – СПб, 2003. – 260с.
4. Герасимов Ю.Ю., Хлюстов В.К. Математические методы и модели в расчётах на ЭВМ: применение в лесоуправлении и экологии: Учебник для лесных вузов. – М.: МГУЛ, 2001. – 260с.
5. Доспехов Б.А. Методика полевого опыта (с основами статистической обработки результатов исследований). – 5-е изд., доп. и перераб. – М.: Агропромиздат, 1985. – 351с.
6. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. - М.: Статистика, 1986. – 366с.
7. Жигунов А.В., Маркова И.А., Бондаренко А.С. Статистическая обработка материалов лесокультурных исследований: Учеб. пособие. – СПб: ЛТА, 2002. – 86с.
8. Животовский Л. А. Популяционная биометрия. – М.: Наука, 1991. – 271с.
9. Ивантер Э. В., Коросов А. В. Введение в количественную биологию: Учеб. пособие. – Петрозаводск: Изд-во ПетрГУ, 2011. – 302с.
10. Карманова И.В. Математические методы роста и продуктивности растений. - М.: Наука, 1976. - 223 с.
11. Кендалл М. Ранговые корреляции. - М.: Статистика, 1975. - 216 с.
12. Ким О. Дж., Мьюллер Ч.У., Клекка У.Р. Факторный, дискриминантный и кластерный анализ: под ред. И. С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с
13. Котов М. М., Лебедева Э. П. Применение биометрических методов в лесной селекции: Учеб. пособие. – Горький, 1977. – 120с.
14. Кравцов Б. А., Милютин Л. И. Возможности применения многомерной классификации при изучении популяций древесных растений // Пространственно-временная структура лесных биогеоценозов. – Новосибирск: Наука, 1981. – С. 47-66.
15. Маслаков Е. Л., Покрывайло В. Д., Товкач Л. Н. Математические методы при изучении роста лесных культур: Метод. указания. – Л.: ЛенНИИЛХ, 1983. – 60 с.

16. Митропольский А.К. Техника статистических вычислений. – М.: Наука, 1971. – 576с.

17. Никитин К.Е., Швиденко А.З. Методы и техника обработки лесоводственной информации. – М.: Лесн. пром–сть, 1978. – 272с.

18. Рунион Р. Справочник по непараметрической статистике: Современный подход. – М.: Финансы и статистика, 1982. – 198с.

19. Свалов Н.Н. Вариационная статистика. - М.: Лесн. пром-сть, 1977.- 176 с.

20. Справочник по прикладной статистике. В 2-х т. / Под. ред. Э. Ллойда, У. Ледермана. – М., 1989, 1990. – 510 с., 526 с.

21. Степин В.В. Применение математических моделей в лесном хозяйстве: Бюлл. ЦБНТИ. – М. – 1975. – 47 с.

22. Тюрин Ю. Н., Макаров А. А. Анализ данных на компьютере. – М.: ИНФРА-М, 2002. – 528 с.

23. Шеффе Г. Дисперсионный анализ. -– М.: Наука, 1980. – 512 с.

ОГЛАВЛЕНИЕ

Цель и задачи статистической обработки экспериментальных данных …….……... **3**
1. Общие принципы обработки полевых данных …….…… **4**
2. Расчет основных статистик ……..…… **9**
3. Типы переменных величин ……..…… **18**
4. Основные понятия ……. **20**
5. Планирование выборки ……... **24**
5.1. Установление минимального объема выборки ….……. **-**
5.2. Способы организации выборки …….…….. **-**
6. Исследование зависимостей …….…… **26**
7. Нормальное распределение ……..…… **35**
8. Проверка распределения …….……. **37**
9. Сравнение двух выборок …….……. **43**
9.1. Независимые выборки ……..……. **-**
9.1.1. Параметрические тесты ……..……. **-**
9.1.2. Непараметрические тесты ……..……. **50**
9.2. Зависимые выборки ……..……. **52**
9.2.1. Параметрические тесты ……..……. **-**
9.2.2. Непараметрические тесты ……..……. **42**
10. Сравнение нескольких выборок …….…... **60**
10.1. Независимые выборки …….…… **-**
10.1.1. Параметрические тесты ……..……... **-**
10.1.2. Непараметрические тесты …….…… **71**
10.2. Зависимые выборки …….…… **74**
10.2.1. Параметрические тесты ……..……. **-**
10.2.2. Непараметрические тесты ……..……. **78**
11. Анализ тесноты связи между переменными ……... **81**
12. Регрессионный анализ ……..……... **88**
12.1. Простая линейная регрессия ……..……... **-**
12.2. Множественная линейная регрессия ……..…… **95**
12.3. Нелинейная регрессия ……..…….. **98**
13. Некоторые методы многомерного анализа данных …… **104**
13.1. Кластерный анализ ……..……... **105**
13.2. Дискриминантный анализ ……..…… **118**
13.3. Факторный анализ ……..…… **119**
13. Заключение……..……..… **120**
13. Рекомендуемая литература ……..…… **122**

Бондаренко Александр Сергеевич
Жигунов Анатолий Васильевич

СТАТИСТИЧЕСКАЯ ОБРАБОТКА МАТЕРИАЛОВ ЛЕСОВОДСТВЕННЫХ ИССЛЕДОВАНИЙ

Учебное пособие

Налоговая льгота — Общероссийский классификатор продукции
ОК 005-93, т. 2; 95 3005 — учебная литература

Подписано в печать 26.10.2016. Формат 60 x 84/16. Печать цифровая.
Усл. печ. л. 7,75. Тираж 80. Заказ 2900.

Отпечатано с готового оригинал-макета, предоставленного авторами,
в Издательско-полиграфическом центре
Политехнического университета.
195251, Санкт-Петербург, Политехническая ул., 29.
Тел.: (812) 552-77-17; 550-40-14.