

Учреждение образования
«БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Н. Я. Сидельник, В. П. Машковский, П. В. Севрук

ЛЕСНАЯ БИОМЕТРИЯ

ЛАБОРАТОРНЫЙ ПРАКТИКУМ

*Рекомендовано
учебно-методическим объединением по образованию
в области природопользования и лесного хозяйства
в качестве учебно-методического пособия
для студентов учреждений высшего образования
по специальности 1-75 01 01 «Лесное хозяйство»*

Минск 2021

УДК 630*9:519.24(076.5)(075.8)

ББК 43:22.172я73

С34

Рецензенты:

кафедра «Лесоводство, экология и защита леса» Мытищинского филиала Московского государственного технического университета имени Н. Э. Баумана (заведующий кафедрой кандидат биологических наук, доцент *В. А. Липаткин*; доцент кафедры кандидат сельскохозяйственных наук, доцент *П. Г. Мельник*);
начальник отдела дистанционного зондирования и мониторинга лесов РУП «Белгослес»
кандидат сельскохозяйственных наук *М. А. Ильючик*

Все права на данное издание защищены. Воспроизведение всей книги или ее части не может быть осуществлено без разрешения учреждения образования «Белорусский государственный технологический университет».

Сидельник, Н. Я.

С34 Лесная биометрия. Лабораторный практикум : учеб.-метод. пособие для студентов специальности 1-75 01 01 «Лесное хозяйство» / Н. Я. Сидельник, В. П. Машковский, П. В. Севрук. – Минск : БГТУ, 2021. – 120 с.
ISBN 978-985-530-891-2.

В пособии описаны статистические методы анализа массовых данных в случаях применения к лесному хозяйству. Рассмотрены такие вопросы, как группировка экспериментальных данных, вычисление основных статистических показателей, анализ распределения массовых наблюдений, исследование закономерностей связи между параметрами объектов, дисперсионный анализ. В каждой лабораторной работе сформулирована цель, указаны обеспечивающие средства, приведен теоретический материал, касающийся изучаемой темы, даны задания и описан ход выполнения работ с пояснениями и необходимыми для этого рисунками.

Лабораторный практикум предназначен для студентов специальности 1-75 01 01 «Лесное хозяйство», также будет полезен магистрантам, аспирантам, работникам лесохозяйственных учреждений при выполнении обработки массовых данных и статистического анализа их различными методами.

УДК 630*9:519.24(076.5)(075.8)

ББК 43:22.172я73

ISBN 978-985-530-891-2

© УО «Белорусский государственный технологический университет», 2021
© Сидельник Н. Я., Машковский В. П., Севрук П. В., 2021

R – коэффициент корреляции
 R^2 – коэффициент детерминации
 S, σ – среднеквадратическое отклонение (смещенное и несмещенное)
 S^2, σ^2 – эмпирическая дисперсия (смещенная и несмещенная)
 t -критерий – критерий Стьюдента
 t_i – нормированное отклонение
 γ, df, cc – число степеней свободы
 x – значения признака у отдельного объекта (варианта)
 \bar{x} – среднее арифметическое значение выборки
 \bar{x}_2 – среднее квадратическое значение выборки
 x_{\max}, x_{\min} – максимальное и минимальное значения признака в выборке
 χ^2 – критерий Пирсона
 V – коэффициент вариации, %
 Q_1, Q_2, Q_3 – квартили



СТАТИСТИЧЕСКИЕ РЯДЫ

Цель лабораторной работы: приобрести навык проведения первичной обработки данных; составить статистические ряды распределения и выполнить разnosку частот диаметров и высот по классам методом конверта; определить средние значения классов и их накопленные частоты.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Биометрия – это инструмент эмпирического познания живой природы, который призван конкретизировать отображение биологических фактов, придать строгость научным выводам и прогнозам, способствовать целенаправленному исследованию явлений.

Биометрическое исследование всегда основано на *выборке* – множество значений случайной величины, совокупность вариантов, набор чисел. Отдельная варианта – это объект, несущий качественный или числовой признак. *Случайная величина* – численная характеристика, принимающая те или иные заранее точно не известные значения. Термин «выборка» указывает на процесс выбора части из чего-то большего, в данном случае – на процесс получения ограниченного количества значений из генеральной совокупности. *Генеральная совокупность* – это множество всех вариантов определенного типа. Чаще всего получить все возможные значения в принципе невозможно. Поэтому судить о генеральной совокупности приходится, исследуя выборки – по части составлять представление о целом. Варианта качественно или количественно выражает признак данного объекта исследования (полученного при данном уровне фактора внешней среды вполне определенным методом).



это упорядоченное изображение реально существующего распределения особей в группе по величине исследуемого признака. Другими словами, вариационный ряд – это ряд ранжированных значений признака, в котором указана частота их встречаемости в данной совокупности. В результате вариационный ряд представляет собой двойной ряд чисел: 1-й ряд обозначает классы; 2-й – частоты вариант изучаемого признака. Указанный ряд пар чисел составляет статистическое распределение – распределение частот по значениям.

После выбора рекомендуемого числа интервалов определяют величины интервалов будущих вариационных рядов по формуле

$$\lambda = \frac{x_{\max} - x_{\min}}{k}, \quad (1.1)$$

где λ – величина интервала; x_{\max} , x_{\min} – максимальное и минимальное значения признака в выборке; k – выбранное число классов.

Затем определяют границы первого интервала вариационного ряда таким образом, чтобы минимальное значение в выборке попало в его середину:

$$x_{\text{H}}^1 = x_{\min} - \frac{\lambda}{2}; \quad (1.2)$$

$$x_{\text{B}}^1 = x_{\min} + \frac{\lambda}{2} - \varepsilon, \quad (1.3)$$

где x_{H}^1 – нижняя граница первого интервала; x_{B}^1 – верхняя граница первого интервала; ε – величина, характеризующая точность представления исходных данных.

Вычисляют границы остальных интервалов, пользуясь следующими формулами:

$$x_{\text{H}}^{k+1} = x_{\text{H}}^k + \lambda; \quad (1.4)$$

$$x_{\text{B}}^{k+1} = x_{\text{B}}^k + \lambda, \quad (1.5)$$

где x_{H}^k – нижняя граница k -того интервала соответствующего признака; x_{B}^k – верхняя граница k -того интервала соответствующего признака; λ – величина интервала для соответствующего признака.

В качестве значений для сформированных классовых интервалов в дальнейших расчетах используют середины классовых интервалов, которые определяют по формуле

Исходные данные для лабораторных работ по лесной биометрии
 студент ЛХФ 2 курса 1 Группы 2 подгруппы Гладкевич Екатерина Евгеньё

D	H	D	H	D	H	D	H	D	H
27,0	25,0	32,0	25,6	22,0	21,5	26,5	24,3	25,3	21,7
24,0	21,6	45,0	26,5	36,6	26,1	29,9	22,2	22,3	22,6
24,0	24,4	26,5	23,5	36,4	24,4	37,6	27,7	39,5	26,1
50,5	28,1	31,9	24,9	36,0	28,1	42,0	25,0	21,5	27,1
21,5	22,2	35,9	23,9	39,0	27,5	26,6	25,5	26,6	24,4
27,0	23,6	27,6	23,6	27,5	25,9	22,6	22,9	36,6	25,6
19,0	22,0	31,5	24,5	53,4	27,8	37,1	25,6	26,5	23,5
23,6	22,5	30,1	24,5	25,6	24,5	34,6	28,4	39,0	24,0
22,6	24,4	27,5	22,6	20,6	21,7	17,6	18,5	29,5	25,6
47,5	27,3	30,6	22,2	25,1	25,4	28,5	25,4	35,0	24,5
31,6	27,4	38,5	26,5	27,3	25,4	35,1	25,9	41,0	26,4
31,0	24,6	37,0	25,9	31,0	25,1	24,6	23,6	29,3	24,9
25,0	23,6	25,0	24,0	29,6	24,9	28,6	21,1	27,6	29,6
37,5	25,5	44,1	28,0	20,0	22,7	29,2	22,7	26,0	22,5
32,2	24,7	20,6	20,0	28,3	25,3	38,0	27,5	41,2	27,0
34,6	26,0	37,0	23,9	45,0	27,0	37,6	28,8	32,0	25,0
37,6	27,0	24,3	23,3	36,0	28,4	44,0	28,6	43,0	28,0
23,0	22,1	29,5	27,6	53,4	27,6	18,6	20,4	27,1	23,1
33,6	24,0	37,1	27,5	46,4	26,1	21,5	20,0	36,0	26,6
29,6	27,0	27,6	24,6	32,5	25,0	25,0	21,6	21,0	21,6
25,5	23,3	52,3	27,4	26,6	24,5	33,2	26,6	24,0	20,3
28,4	27,0	37,3	25,5	30,9	26,3	31,5	25,9	28,6	24,3
21,5	21,0	25,6	22,2	29,6	21,6	39,9	26,6	41,2	26,4
44,5	28,4	16,0	22,7	23,0	23,6	27,0	24,1	34,6	26,0
27,0	23,6	42,5	27,6	41,0	27,5	30,5	23,1	30,0	25,5
22,0	23,9	31,5	24,6	31,0	24,5	32,1	24,6	38,6	25,1
28,0	25,6	20,5	22,5	27,5	24,0	40,6	28,3	37,0	26,4
45,1	28,2	27,5	25,5	28,3	24,4	34,5	25,6	25,6	23,8
24,6	24,4	32,0	24,6	37,0	23,5	29,0	24,0	26,5	23,4
22,6	21,7	43,0	28,4	26,9	21,3	24,2	21,5	27,5	19,6
29,5	24,0	28,5	24,6	35,0	25,6	25,1	24,5	29,0	23,6
27,0	24,6	35,6	25,1	34,5	24,6	26,7	23,7	23,1	22,1
34,1	25,7	30,5	24,5	26,2	24,6	21,5	22,0	39,0	26,9
28,2	24,6	28,3	24,1	31,1	24,4	44,0	30,3	30,1	25,6
41,7	25,8	43,0	27,0	39,7	23,9	27,0	22,0	31,0	25,5
27,6	25,6	36,6	27,1	60,0	28,5	60,0	27,6	52,1	26,5
16,0	19,6	29,5	24,0	38,3	25,5	20,5	21,8	31,5	25,6
25,8	22,3	29,6	24,6	26,1	22,6	30,7	23,9	28,5	24,0
28,1	22,3	33,5	26,0	35,0	24,6	28,0	25,9	33,0	26,6
37,0	25,2	33,5	25,6	40,6	25,4	35,4	25,4	25,6	25,0

Задание выдал _____

Пример выданного задания

Полученную величину следует округлить таким образом, чтобы точность представления исходных данных и величины интервала была одинаковой (до 0,1), а последняя цифра округления – четной.

Определяем границы первого интервала вариационного ряда по формулам (1.2) и (1.3):

– для диаметров:

$$D_{\text{н}}^1 = D_{\text{мин}} - \frac{\lambda_D}{2} = 16,0 - \frac{4,0}{2} = 14,0;$$

$$D_{\text{в}}^1 = D_{\text{мин}} + \frac{\lambda_D}{2} - \varepsilon = 16,0 + \frac{4,0}{2} - 0,1 = 17,9;$$

– для высот:

$$H_{\text{н}}^1 = H_{\text{мин}} - \frac{\lambda_H}{2} = 18,5 - \frac{1,0}{2} = 18,0;$$

$$H_{\text{в}}^1 = H_{\text{мин}} + \frac{\lambda_H}{2} - \varepsilon = 18,5 + \frac{1,0}{2} - 0,1 = 18,9,$$

где $D_{\text{н}}^1$ – нижняя граница первого интервала для диаметров; $D_{\text{в}}^1$ – верхняя граница первого интервала для диаметров; $H_{\text{н}}^1$ – нижняя граница первого интервала для высот; $H_{\text{в}}^1$ – верхняя граница первого интервала для высот; ε – величина, характеризующая точность представления исходных данных (в нашем случае $\varepsilon = 0,1$).

Вычислим границы остальных интервалов, пользуясь формулами (1.4) и (1.5), и определим количество наблюдений, попавшее в интервалы вариационного ряда, регистрируя наблюдения методом конверта (табл. 1.2). Например, значение диаметра $D = 27,0$ см (рисунок) попадает в границы интервала 26,0–29,9 см, значит ставим в соответствующую строку класса в столбец *Шифр частот* «точку», значение $D = 24,0$ см (рисунок) попадает в границы интервала 22,0–25,9 см – ставим в соответствующую строку класса в столбец *Шифр частот* «точку» и т. д. (табл. 1.3).

В дальнейших расчетах будем использовать не сами диаметры и высоты деревьев, полученные в результате измерений, а середины интервалов и частоты составленных нами вариационных рядов. Для этого рассчитаем их по формуле (1.6), например, для интервала 14,0–17,9: $(14,0 + 17,9 + 0,1) / 2 = 16,0$ см и т. д. Середины интервалов (16, 20, 24 и т. д.), обозначающие границы интервалов (соответственно 14,0–17,9; 18,0–21,9; 22,0–25,9 и т. д.), в практике ведения лесного хозяйства носят название *ступень толщины* (в данном случае, 4-сантиметровые).

Колонка *Частота* представляет собой цифровое обозначение графического шифра (метода конверта).

Таблица 1.3

Распределение наблюдений по интервалам (диаметры)

Классы (интервалы)	Шифр частот	Среднее значение классов (D_i)	Ча- стота (f_i)	Накоп- ленная частота	Относи- тельная частота
14,0–17,9	••	16,0	3	3	1,5
18,0–21,9	☒ ••	20,0	13	16	6,5
22,0–25,9	☒ ☒ ☒	24,0	29	45	14,5
26,0–29,9	☒ ☒ ☒ ☒ ☒ ••	28,0	55	100	27,5
30,0–33,9	☒ ☒ ☒	32,0	30	130	15,0
34,0–37,9	☒ ☒ ☒ ••	36,0	32	162	16,0
38,0–41,9	☒ ☒••	40,0	17	179	8,5
42,0–45,9	☒ ••	44,0	12	191	6,0
46,0–49,9	••	48,0	2	193	1,0
50,0–53,9	••••	52,0	5	198	2,5
54,0–57,9		56,0	0	198	0,0
58,0–61,9	••	60,0	2	200	1,0
<i>Итого</i>	—	—	200	—	100

При анализе массовых данных наряду с частотами вариационных рядов используют накопленные частоты, которые вычисляют как сумму частот текущего и всех предшествующих интервалов, т. е. для расчета столбца *Накопленная частота* надо к частоте данного интервала прибавить сумму частот предыдущих интервалов. Для класса со средним значением 16,0: $3 + 0 = 3$; для класса 20,0: $13 + 3 = 16$; для класса 24,0: $29 + 13 + 3 = 45$ и т. д.

Для того чтобы рассчитать колонку *Относительная частота (частости)*, надо частоту каждого интервала выразить в относительных единицах измерения, т. е. в процентах по отношению к общей сумме наблюдения. Для интервала 16,0: $3 \cdot 100\% / 200 = 1,5\%$; для интервала 20,0: $13 \cdot 100\% / 200 = 6,5\%$; для интервала 24,0: $29 \cdot 100\% / 200 = 14,5\%$ и т. д.

Аналогичные вычисления выполняются и для высот. Например, значение высоты $H = 25,0$ м (рисунок) попадает в границы интервала 25,0–25,9 м, поэтому ставим в соответствующую строку класса в столбец *Шифр частот* точку, значение $H = 21,6$ м

(рисунок) попадает в границы интервала 21,0–21,9 м – ставим в соответствующую строку класса в столбец *Шифр частот* точку и т. д. (табл. 1.4).

Таблица 1.4

Распределение наблюдений по интервалам (высоты)

Классы (интервалы)	Шифр частот	Среднее значение классов (H_i)	Часто- та (f_i)	Накоп- ленная частота	Относи- тельная частота
18,0–18,9	•	18,5	1	1	0,5
19,0–19,9	• •	19,5	2	3	1,0
20,0–20,9	• • • •	20,5	4	7	2,0
21,0–21,9	⊠ • •	21,5	13	20	6,5
22,0–22,9	⊠ ⊠ •	22,5	21	41	10,5
23,0–23,9	⊠ ⊠ • •	23,5	22	63	11,0
24,0–24,9	⊠ ⊠ ⊠ ⊠ • •	24,5	43	106	21,5
25,0–25,9	⊠ ⊠ ⊠ ⊠	25,5	40	146	20,0
26,0–26,9	⊠ ⊠	26,5	18	164	9,0
27,0–27,9	⊠ ⊠ •	27,5	21	185	10,5
28,0–28,9	⊠ • •	28,5	13	198	6,5
29,0–29,9	•	29,5	1	199	0,5
30,0–30,9	•	30,5	1	200	0,5
<i>Итого</i>	–	–	200	–	100

Для того чтобы не запутаться при точковке, можно разбить столбец *Шифр частот* на 5 колонок (по количеству столбцов в задании (рисунок)) и производить разноску частот (точковать) каждой колонки задания в отдельный разбитый столбец таблицы, а потом суммировать полученные результаты в одну. Также для регистрации частот можно использовать программные возможности MS Excel.



ДВУМЕРНАЯ ТАБЛИЦА РАСПРЕДЕЛЕНИЯ

Цель лабораторной работы: составить двумерную таблицу распределения по двум признакам.

Обеспечивающие средства: рабочая тетрадь, ручка, линейка, карандаш, стирка, калькулятор или персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Наряду с простой группировкой данных в интервальные вариационные ряды в некоторых случаях целесообразно сгруппировать данные сразу по двум параметрам. Эта потребность, как правило, возникает в том случае, если необходимо проанализировать связь между двумя случайными величинами. В таких случаях составляют так называемые *таблицы распределения* (корреляционные таблицы, корреляционные решетки). Корреляционной решеткой называется вспомогательная таблица, в которую внесены сгруппированные данные двух признаков с указанием диапазонов классов и частоты встречаемости значений с целью вычисления статистических показателей, например коэффициента корреляции и одновременной проверки правильности составления простых распределений по классам по одному признаку.

В решетке по горизонтали фиксируются значения классов по одному признаку (X) слева направо от меньших значений к большим, а по вертикали – значения классов по другому признаку (Y) снизу-вверх от меньших значений к большим. Такое расположение соответствует построению обычной системы координат, на основе которой строятся линии зависимости признаков. Внутри решетки распределяются варианты сопряженных признаков. По их расположению можно сделать предварительные выводы о характере связи до вычисления.



Порядок выполнения работы

Задание. Составим таблицу распределения 200 деревьев по интервалам диаметра и высоты в чистом сосновом древостое по данным примера (рисунок, лабораторная работа № 1). Для этого воспользуемся интервалами рядов распределения деревьев по диаметрам (табл. 1.3) и высотам (табл. 1.4). Впишем границы интервалов вариационного ряда по диаметрам в первую строку таблицы из лабораторной работы № 2 в порядке возрастания, а границы вариационного ряда по высотам – в первую колонку, но в порядке убывания.

Затем, распределяя наблюдения одновременно по интервалам диаметров и высот и регистрируя их методом конвертов, определим частоты каждой клетки корреляционной решетки (таблица). Например, первая пара значений $D = 27,0$ см и $H = 25,0$ м в примере (рисунок, лабораторная работа № 1) – по диаметру $D = 27,0$ см попадает в столбец интервала 26,0–29,9 см, затем смотрим, куда попадает значение по высотам $H = 25,0$ м – в интервал 25,0–25,9 м. На пересечении данных интервалов в ячейке регистрируем точкой результат (таблица).

Такие же действия повторяем со всеми остальными парами значений для своего варианта (чтобы не запутаться при точковке, можно создать пустые копии данной таблицы для каждого столбца задания и точковать каждый в отдельной копии таблицы, а потом суммировать результат в одну таблицу). Также для регистрации частот можно использовать MS Excel.

В результате сумма частот по диаметрам (нижняя строчка) и по высотам (крайний правый столбец) должны совпасть с частотами из табл. 1.3 и 1.4 соответственно.

Двумерная таблица распределения по интервалам диаметра и высоты

$D \backslash H$	14,0–17,9	18,0–21,9	22,0–25,9	26,0–29,9	30,0–33,9	34,0–37,9	38,0–41,9	42,0–45,9	46,0–49,9	50,0–53,9	54,0–57,9	58,0–61,9	Всего
30,0–30,9				• / 1				• / 1					1
29,0–29,9													1
28,0–28,9						••• / 4	• / 1	••• / 6		• / 1		• / 1	13
27,0–27,9		• / 1		••• / 3	• / 1	••• / 4	••• / 4	••• / 3	• / 1	••• / 3		• / 1	21
26,0–26,9					••• / 4	••• / 5	••• / 6	• / 1	• / 1	• / 1			18
25,0–25,9			••• / 2	••• / 11	••• / 10	••• / 12	••• / 4	• / 1					40
24,0–24,9			••• / 6	••• / 20	••• / 12	••• / 4	• / 1						43
23,0–23,9			••• / 7	••• / 9	••• / 2	••• / 3	• / 1						22
22,0–22,9	• / 1	••• / 5	••• / 7	••• / 7	• / 4								21
21,0–21,9		••• / 4	••• / 6	••• / 3									13
20,0–20,9		••• / 3	• / 1										4
19,0–19,9	• / 1			• / 1									2
18,0–18,9	• / 1												1
<i>Итого</i>	3	13	29	55	30	32	17	12	2	5	0	2	200

СОСТАВЛЕНИЕ СТАТИСТИЧЕСКИХ РЯДОВ И ИХ ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ С ИСПОЛЬЗОВАНИЕМ ПАКЕТА ПРОГРАММ

Цель лабораторной работы: составить вариационные ряды с использованием программных средств на ЭВМ; овладеть навыками графического представления выборки.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office и Statistica 10.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Существует множество пакетов программ, предназначенных для статистического анализа данных. Программное обеспечение анализа данных можно условно разделить на пакеты общего назначения (MS Excel) и специальные программные продукты, которые, в свою очередь, делятся на математические программы (Mathematica, Matlab, Maple, Mathcad), статистические программы (Statistica, StatGraphics, SPSS, Stadia и др.) и пакеты научной графики. Наиболее доступными в вузах и широко применяемыми в научно-исследовательских организациях биологического и экологического профиля являются табличный процессор MS Excel и пакет Statistica. С их помощью можно выполнить группировку исходных данных.

Основными средствами анализа данных в MS Excel выступают статистические и математические функции библиотеки встроенных функций (*Мастер функций*), статистические процедуры надстройки *Пакет анализа* и специальный инструмент для проведения графического анализа – *Мастер диаграмм*.

Пакет Statistica представляет собой интегрированную систему статистического анализа и обработки данных. Система состоит из следующих основных компонентов: многофункциональной систе-

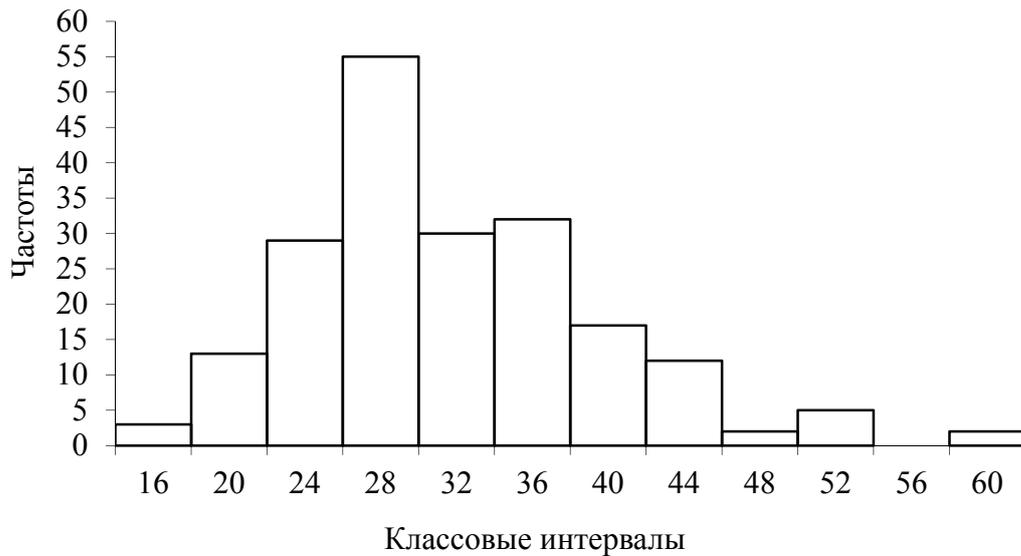


Рис. 3.1. Гистограмма распределения сосновых стволов по диаметрам

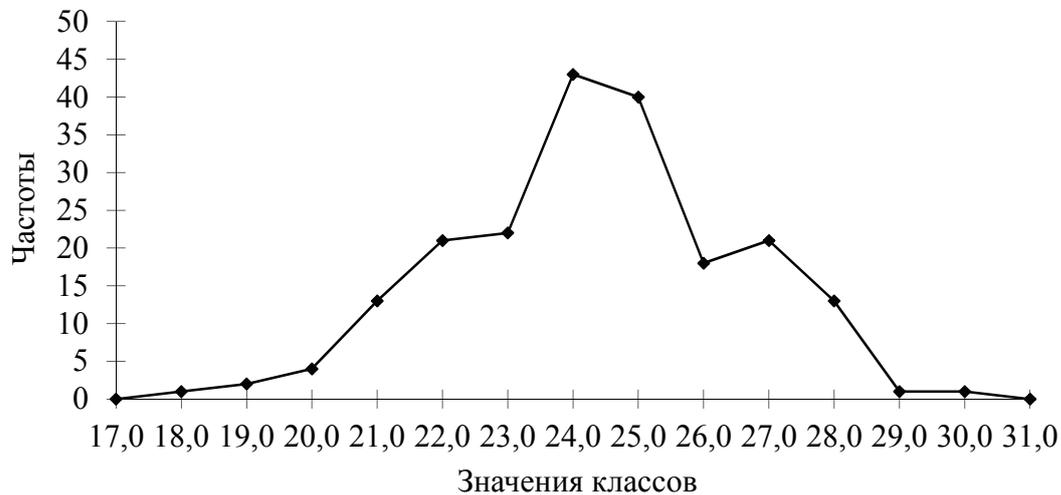


Рис. 3.2. Полигон распределения сосновых стволов по высотам

Кумулята. При построении кумулятивной кривой по оси ординат откладывают накопленные частоты. По оси абсцисс в случае дискретного вариационного ряда откладывают значения классов, а в случае непрерывного – границы интервалов (рис. 3.3).

Огива. Огива строится так же, как и кумулята, однако координаты меняются ролями. На оси абсцисс откладываются частоты, а значения классов или границы интервалов – по оси ординат (рис. 3.4).

Наиболее читаемыми графики получаются, если при их построении соблюдать правило золотого сечения, согласно которому

Порядок выполнения работы

Задание 1. Рассмотрим последовательность действий при выполнении этой работы с помощью пакета Statistica 10.0. Запустим программу Statistica, нажав на одноименный ярлык на *Рабочем столе*. Закроем все окна в программе, если они открылись.

1. Создадим файл для исходных данных, активизировав опцию *Создать...* из пункта меню *Файл*.

2. В открывшемся диалоговом окне *Создать Новый Документ* (рис. 3.5) на вкладке *Таблица* установим необходимое количество переменных – **2** (высота и диаметр); число наблюдений – **200**, формат отображения – **Число** и *Дес. разряды* – **1** (число знаков после запятой у переменной). После этого достаточно нажать кнопку *ОК* в диалоговом окне создания нового документа (рис. 3.5) и файл будет создан.

3. Открыв двойным щелчком левой кнопкой мыши по заголовку переменной *Пер1* диалоговое окно редактирования ее свойств (рис. 3.6), присвоим переменным имена «**D**» и «**H**» соответственно.

4. Введем или скопируем исходные данные (рисунок, лабораторная работа № 1) и сохраним файл с ними (рис. 3.7) с помощью опции *Сохранить* из меню *Файл* на диск *D*, каталог *BIOM*, подкаталог *№группы*, имя файла: *№группы_№подгруппы_Фамилия_Имя* (латинскими буквами).

5. Из пункта меню *Анализ (Статистика)* выберем опцию *Основная статистика / таблицы*, которая откроет диалоговое окно с локальным меню. Из данного меню выберем опцию *Таблицы частот*, которая откроет диалоговое окно *Таблицы частот* (рис. 3.8).

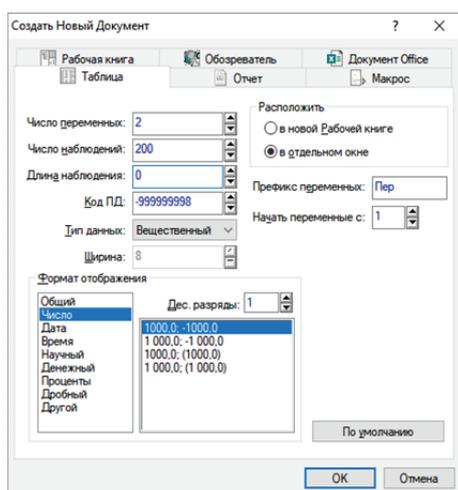


Рис. 3.5. Окно «Создать новый документ» в программе Statistica

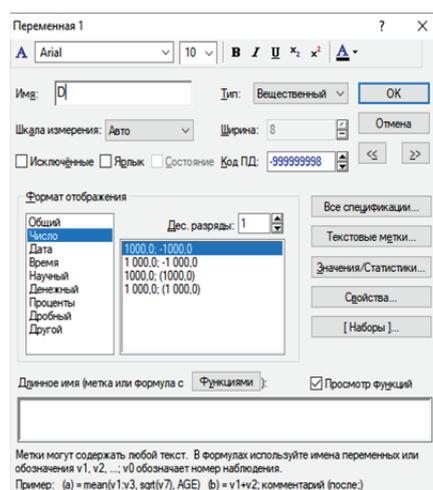


Рис. 3.6. Окно редактирования свойств переменных в программе Statistica

6. В открывшемся диалоговом окне нажмем кнопку *Переменные*. После этого на экране появится еще одно диалоговое окно, содержащее список переменных. Выберем для начала первую переменную, например *D*, щелкнув по ней левой кнопкой мыши. Завершим выбор переменной нажатием кнопки *OK*.

7. В диалоговом окне *Таблицы частот* (рис. 3.8) выберем вкладку *Дополнительно*. Затем с помощью мыши выберем метод группировки данных с явным заданием интервалов *Размер шага*. Введем в поле *Размер шага* величину интервала вариационного ряда по диаметрам $\lambda_D = 4,0$ см (см. лабораторную работу № 1). Щелкнув мышью, уберем отметку у пункта *с мин. знач.* (рис. 3.8) и введем в поле *Начать с* нижнюю границу самого первого интервала (см. лабораторную работу № 1, например, для диаметров $D_n^1 = 14,0$ см).

8. Нажмем кнопку *OK*, и группировка данных будет выполнена. Полученный результат приведен на рис. 3.9.

Рассматриваемая таблица содержит границы интервалов в колонке *От До*. В колонке *Частота* записаны частоты вариационного ряда. Колонка *Кумул. Частота* содержит накопленные частоты, а колонки *Процент* и *Кумул. Процент* – частоты и накопленные частоты полученного вариационного ряда, выраженные в процентах к общему количеству наблюдений.

	1	2
	D	H
1	25,3	21,7
2	22,3	22,6
3	39,5	26,1
4	21,5	27,1
5	26,6	24,4
6	36,6	25,6
7	26,5	23,5
8	39,0	24,0
9	29,5	25,6
10	35,0	24,5
11	41,0	26,4
12	29,3	24,9
13	27,6	29,6
14	26,0	22,5
15	41,2	27,0
16	32,0	25,0
17	43,0	28,0
18	27,1	23,1
19	36,0	26,6
20	21,0	21,6

Рис. 3.7. Созданный файл данных в программе Statistica

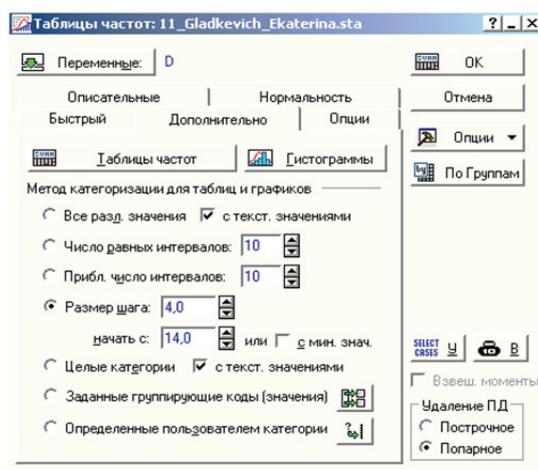


Рис. 3.8. Окно *Таблицы частот* в программе Statistica

Сделаем «снимок экрана», нажав кнопку на клавиатуре *PrtScr*, и сохраним его в личную папку на диск *D*, каталог *BIOM*, подкаталог *№группы*.

Результат группировки в программе Statistica должен совпадать с результатами группировки в табл. 1.3 (для диаметров) и табл. 1.4 (для высот).

Составить вариационные ряды можно также с использованием программного пакета MS Excel. Открыв программу, введем или скопируем исходные значения варианта (200 значений диаметров и высот), создадим таблицу с границами интервалов (табл. 1.3 и 1.4) и используем программные функции MS Excel, например ЧАСТОТА (рис. 3.11), СУММЕСЛИМН, СЧЁТЕСЛИ и др.

Ввод функции в ячейку следующий – ЧАСТОТА (массив данных; массив интервалов). Массив данных – массив или ссылка на множество числовых данных, для которых вычисляются частоты; массив интервалов – массив или ссылка на множество интервалов, в которые группируются значения аргумента «массив данных». Количество элементов в возвращаемом массиве на единицу больше числа элементов в массиве «Массив интервалов». Дополнительный элемент в возвращаемом массиве содержит количество значений, превышающих верхнюю границу интервала, содержащего наибольшие значения.

Важно знать, что функция ЧАСТОТА () вводится как формула массива (рис. 3.11) после выделения диапазона смежных ячеек, в которые требуется вернуть полученный массив распределения (частот), т. е. после ввода формулы необходимо вместо нажатия клавиши *Enter* нажать сочетание клавиш *Ctrl+Shift+Enter*.

	D	H				Среднее значение классов	Частота
1							
2	25,3	21,7	14,0 - 17,9			16,0	3
3	22,3	22,6	18,0 - 21,9			20,0	13
4	39,5	26,1	22,0 - 25,9			24,0	29
5	21,5	27,1	26,0 - 29,9			28,0	55
6	26,6	24,4	30,0 - 33,9			32,0	30
7	36,6	25,6	34,0 - 37,9			36,0	32
8	26,5	23,5	38,0 - 41,9			40,0	17
9	39,0	24,0	42,0 - 45,9			44,0	12
10	29,5	25,6	46,0 - 49,9			48,0	2
11	35,0	24,5	50,0 - 53,9			52,0	5
12	41,0	26,4	54,0 - 57,9			56,0	0
13	29,3	24,9	58,0 - 61,9			60,0	2
14	27,6	29,6					200
15	26,0	22,5					
16	41,2	27,0					
17	32,0	25,0	18,0 - 18,9			18,5	1
18	43,0	28,0	19,0 - 19,9			19,5	2
19	27,1	23,1	20,0 - 20,9			20,5	4
20	36,0	26,6	21,0 - 21,9			21,5	13
21	21,0	21,6	22,0 - 22,9			22,5	21
22	24,0	20,3	23,0 - 23,9			23,5	22
23	28,6	24,3	24,0 - 24,9			24,5	43
24	41,2	26,4	25,0 - 25,9			25,5	40
25	34,6	26,0	26,0 - 26,9			26,5	18
26	30,0	25,5	27,0 - 27,9			27,5	21
27	38,6	25,1	28,0 - 28,9			28,5	13
28	37,0	26,4	29,0 - 29,9			29,5	1
29	25,6	23,8	30,0 - 30,9			30,5	1
30	26,5	23,4					200
31	27,5	19,6					

Рис. 3.11. Результат группировки данных диаметров и высот с помощью функции ЧАСТОТА в MS Excel

Графическое изображение рядов распределения можно выполнить в MS Excel, который позволяет создавать диаграммы и графики различных типов. Различают две категории диаграмм: стандартные и нестандартные. К стандартным относятся гистограммы, линейчатые диаграммы, графики, круговые и точечные диаграммы, диаграммы с областями, а также кольцевые, лепестковые, поверхностные, пузырьковые, конические, цилиндрические и пирамидальные диаграммы. Нестандартные диаграммы обычно формируются на базе стандартных, отличаясь особым оформлением. Для построения диаграмм в MS Excel используется меню *Вставка* и выбирается необходимый тип диаграммы (рис. 3.12).

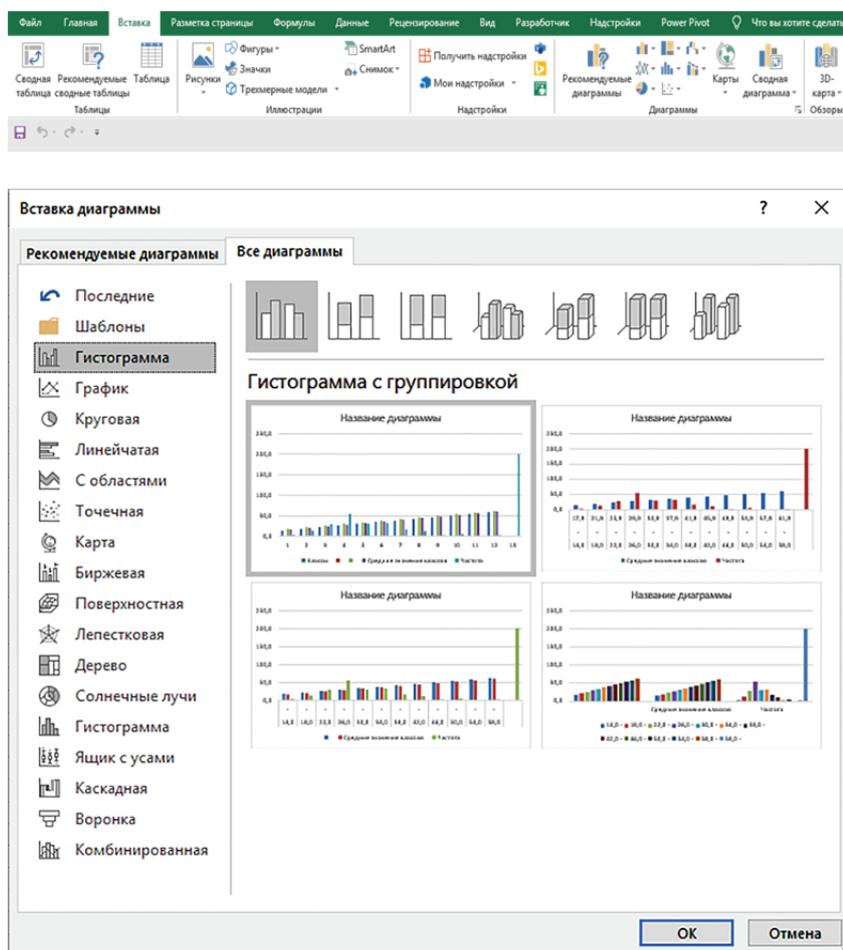


Рис. 3.12. Выбор диаграмм и их типов в MS Excel

Новые возможности построения диаграмм в MS Excel 2019 еще больше упрощают представление цифровой информации в виде понятной графической информации.

ОПРЕДЕЛЕНИЕ ОСНОВНЫХ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Цель лабораторной работы: рассчитать основные статистические показатели, характеризующие выборочную совокупность, и их стандартные ошибки.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 4 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Для того чтобы можно было всесторонне проанализировать исследуемый признак, необходимо вычислить целый ряд статистических показателей, характеризующих объект исследования. При этом статистики делятся на две группы:

- показатели, которые характеризуют центральную тенденцию или уровень ряда (это различные средние величины – мода, медиана, средняя арифметическая, средняя геометрическая и др.);
- показатели, измеряющие степень вариации (размах вариации, дисперсия, среднее квадратическое отклонение, коэффициенты вариации, асимметрии, эксцесса и др.).

Для расчета основных статистических показателей используются *метод моментов* или *непосредственный метод* расчета. В связи с развитием компьютерных технологий последний способ наиболее распространен.

При обработке первичных экспериментальных материалов наиболее важным является расчет средней арифметической величины – основная величина, которая характеризует изучаемую совокупность.

Среднее арифметическое значение. Общая формула для определения величины средней арифметической – это отношение суммы значений всех вариантов (x_i) выборки к их числу (объему вы-

борки n). Средняя арифметическая величина для сгруппированных значений называется *средневзвешенной*:

$$\bar{x} = \frac{\sum_{i=1}^k (x_i \cdot f_i)}{n}, \quad (4.1)$$

где k – количество классов; x_i – значение i -того класса (середины интервала, если ряд интервальный); f_i – частота i -того класса.

Средняя арифметическая величина используется для оценки математического ожидания исследуемой случайной величины.

Среднее квадратическое значение. Данная величина имеет большое значение в лесном хозяйстве для определения одного из основных таксационных показателей – среднего диаметра древостоя, который является среднеквадратической величиной. Это вызвано тем, что, согласно определяющему свойству, при замене всех элементов выборки на среднее квадратическое остается постоянной сумма квадратов элементов выборки. В том случае, если исходные данные сгруппированы в статистический ряд, среднее квадратическое значение определяется по формуле

$$\bar{x}_2 = \sqrt[2]{\frac{\sum_{i=1}^n (x_i^2 \cdot f_i)}{n}}. \quad (4.2)$$

Данная величина позволяет вычислить сумму площадей сечений древостоя (абсолютную полноту древостоя – G), являющуюся важнейшим таксационным показателем:

$$G = \sum_{i=1}^n g_i = \sum_{i=1}^n \frac{\pi \cdot d_i^2}{4} = \frac{\pi}{4} \cdot \sum_{i=1}^n d_i^2 = \frac{\pi}{4} \cdot \sum_{i=1}^n \bar{d}^2,$$

где G – сумма площадей сечения древостоя; g_i – площадь поперечного сечения i -того дерева; d_i – диаметр i -того дерева; $\pi = 3,14$; \bar{d} – среднеквадратический диаметр древостоя.

Определив сумму площадей сечения деревьев в древостое, можно найти его запас по известной формуле лесной таксации: $M = G \cdot HF$, где M – запас древостоя, м³/га; HF – видовая высота, м (находится по специализированной таблице в зависимости от породы и средней высоты древостоя); G – сумма площадей сечения, м²/га.

Средние величины указывают на то значение признака, вокруг которого группируются анализируемые наблюдения, но вокруг

одного и того же значения признака наблюдения могут располагаться совершенно по-разному. Для того чтобы отразить характер расположения наблюдений вокруг среднего, и служат показатели вариации.

Размах вариации. Это наиболее простой показатель, характеризующий распределение вариантов вокруг среднего. Он вычисляется как разность между максимальным и минимальным значениями признака, которые в биометрии называют также *лимитами* (от латинского слова *limes* – предел и обозначают символом *lim*):

$$R = x_{\max} - x_{\min}. \quad (4.3)$$

Если наблюдения плотно группируются вокруг среднего, то лимиты располагаются близко друг к другу и размах вариации оказывается небольшим и наоборот.

Однако размах вариации является ненадежным показателем, так как он вычисляется на основании значений лимитов. Для преодоления отмеченного недостатка необходимо учитывать не только крайние значения признака (лимиты), но и все варианты в выборке, чтобы понять характер распределения всех остальных значений, располагающихся ближе к среднему.

Эмпирическая дисперсия (σ^2 , S^2). Этот показатель получил свое название от латинского слова *dispersio* – рассеяние, он представляет собой средний квадрат отклонений вариант от среднего арифметического:

$$S_x^2 = \frac{\sum_{i=1}^k f_i \cdot (x_i - \bar{x})^2}{n}. \quad (4.4)$$

Выборочная дисперсия, рассчитанная по данной формуле, дает смещенную оценку генеральной дисперсии. Для того чтобы получить несмещенную оценку, в формулу необходимо добавить сомножитель $\frac{n}{n-1}$, называемый *поправкой Бесселя*:

$$\sigma_x^2 = S_x^2 \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^k f_i \cdot (x_i - \bar{x})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^k f_i \cdot (x_i - \bar{x})^2}{n-1}. \quad (4.5)$$

Величина $n-1$ из формулы называется *числом степеней свободы*, которое показывает, сколько в данном случае имеется независимых наблюдений.

В некоторых случаях использование дисперсии оказывается не очень удобным, поскольку в формуле каждое отклонение варианты от среднего значения возводится в квадрат, в итоге дисперсия измеряется в единицах, равных квадрату единицы измерения. Так, например, если высчитывается дисперсия показателя, измеряемого в метрах, то сама дисперсия будет выражаться в квадратных метрах, что само по себе бессмысленно для сравнения. Поэтому часто используется другой, очень близкий к дисперсии показатель вариации.

Среднеквадратическое отклонение. Для избавления от квадратов отклонений прибегают к действию, противоположному возведению в степень, т. е. извлекают квадратный корень из дисперсии. В итоге стандартное отклонение является в ряде случаев более удобной характеристикой вариации признаков, поскольку измеряется в тех же единицах, что и исходные данные:

$$S_x = \sqrt{\frac{\sum_{i=1}^k f_i \cdot (x_i - \bar{x})^2}{n}} \quad \text{– смещенная оценка;} \quad (4.6)$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^k f_i \cdot (x_i - \bar{x})^2}{n-1}} \quad \text{– несмещенная оценка.} \quad (4.7)$$

В связи с этим данный показатель является более естественным и легче поддается анализу. На основе правила «трех сигм» с помощью среднеквадратического отклонения можно производить первоначальную отбраковку значений в выборке. Ведь в интервале $\pm 3\sigma$ от среднеарифметического значения находится 99,7% всех вариантов ряда, в интервале $\pm 2\sigma$ – 95,4% и в интервале $\pm 1\sigma$ – 68,2% (рис. 4.1).

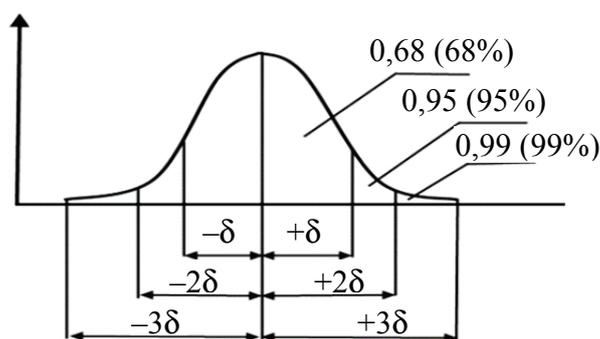


Рис. 4.1. Среднеквадратическое отклонение как мера variability признака

Коэффициент вариации. Дисперсия и среднее квадратическое отклонение довольно полно характеризуют вариацию, однако часто удобнее иметь показатель, оценивающий разброс данных не в абсолютных величинах, а в *относительных*. Таким показателем является *коэффициент вариации*, показывающий, сколько процентов составляет среднее квадратическое отклонение от среднего арифметического:

$$V = \frac{\sigma_x}{x} \cdot 100\%. \quad (4.8)$$

В биометрии этот показатель часто оказывается весьма полезным. Дело в том, что анализу подвергаются, как правило, объекты живой природы, а они с течением времени изменяют свои размеры, растут. В связи с этим часто необходимо анализировать выборки, сделанные для объектов с разным средним возрастом, а следовательно, и с разными средними размерами. Если в таких случаях нужно сравнить степень изменчивости признака в разных выборках, то удобнее оперировать коэффициентом вариации, так как он дает нам величину вариации по отношению к среднему значению.

Варьирование считается слабым при коэффициенте вариации 0–10%, при 10–30% – средним, 30–60% – высоким, 60–100% – очень высоким, при более 100% – аномальным. В лесных биогеоценозах при коэффициенте вариации больше 30–33% распределение можно считать неоднородным (например, древостой будет разновозрастным или в анализ взяты данные второго яруса), т. е. чем меньше значение коэффициента вариации, тем совокупность однороднее. При $V > 50\%$ совокупность вообще неоднородна – сильное разнообразие ряда свидетельствует о малой представительности (типичности) соответствующей средней величины и, следовательно, о нецелесообразности ее использования в практических целях.

В одновозрастных чистых насаждениях, созданных путем посева и посадки и имевших до смыкания крон деревьев одинаковый уход, распределение деревьев по толщине характеризуется симметричной, одновершинной линией, называемой кривой нормального распределения. В этом случае влияние многочисленных факторов, задерживающих рост деревьев или способствующих ему, взаимно уравниваются. Однако довольно часто у кривых распределения появляется асимметрия.

Коэффициент асимметрии. Рассмотренные выше показатели довольно полно характеризуют анализируемые признаки, однако ни один из них не отражает степень симметричности распределе-

Асимметричность кривой распределения есть следствие конкуренции между деревьями. Более крупные деревья, занимающие в насаждении лучшее положение, имеют все преимущества для успешного роста. Поэтому с увеличением возраста правая ветвь кривой распределения, на которой сосредоточены крупные деревья, становится длиннее. Левая ветвь, изображающая отстающие в росте деревья, оказывается более короткой из-за отпада ослабленных деревьев или в результате их вырубki в порядке ухода за лесом. В итоге одновершинная, асимметричная кривая характеризует молодое насаждение или пройденное рубками ухода.

Левовершинное с положительной асимметрией распределение является типичным для распределений числа деревьев по диаметру в одновозрастных древостоях, не пройденных рубками ухода, или пройденных рубками ухода слабой (средней) интенсивности по комбинированному (низовому) способу, или выборочными рубками, при которых вырубаются в основном самые толстые деревья.

Правовершинное с отрицательной асимметрией распределение диаметров деревьев можно получить искусственным путем при интенсивной рубке маломерных деревьев. Это распределение является характерным для распределений высот деревьев в одновозрастном древостое.

По мере увеличения возраста древостоя растет размах ряда распределения деревьев по толщине. Из-за уменьшения числа деревьев кривая становится более плоской, о чем свидетельствует эксцесс.

Эмпирический коэффициент эксцесса. Кроме того, что распределения наблюдений могут отличаться друг от друга по степени асимметричности, они могут иметь разную крутизну. Распределения могут быть *островершинными* и *плосковершинными*. В случае островершинной кривой, когда большое число наблюдений группируется в непосредственной близости от центра распределения, говорят о наличии *положительного* эксцесса. Кривая распределения имеет *отрицательный* эксцесс, если она плосковершинная. Для оценки степени крутизны кривой распределения используется *коэффициент эксцесса* (мера крутости вариационного ряда), который вычисляется по формуле

$$E = \frac{\sum_{i=1}^k f_i \cdot (x_i - \bar{x})^4}{n \cdot \sigma_x^4} - 3. \quad (4.10)$$

(табл. 1.3 и 1.4). Для выполнения вычислений составим вспомогательную табл. 4.1.

Таблица 4.1

Вычисление средних значений (диаметры)

D_i	f_i	$D_i \cdot f_i$	$D_i^2 \cdot f_i$	g_i	$g_i \cdot f_i$
16,0	3	48,0	768,0	0,0201	0,0603
20,0	13	260,0	5 200,0	0,0314	0,4082
24,0	29	696,0	16 704,0	0,0452	1,3108
28,0	55	1 540,0	43 120,0	0,0616	3,3880
32,0	30	960,0	30 720,0	0,0804	2,4120
36,0	32	1 152,0	41 472,0	0,1018	3,2576
40,0	17	680,0	27 200,0	0,1257	2,1369
44,0	12	528,0	23 232,0	0,1521	1,8252
48,0	2	96,0	4 608,0	0,1810	0,3620
52,0	5	260,0	13 520,0	0,2124	1,0620
56,0	0	0,0	0,0	0,2463	0,0000
60,0	2	120,0	7 200,0	0,2827	0,5654
Сумма	200	6 340,0	213 744,0	–	16,7884

Теперь, пользуясь формулой (4.1), вычислим среднее арифметическое значение

$$\bar{D} = \frac{\sum_{i=1}^k (D_i \cdot f_i)}{n} = \frac{6340,0}{200} = 31,70 \text{ см}$$

и по формуле (4.2) среднее квадратическое:

$$\bar{D}_2 = \sqrt[2]{\frac{\sum_{i=1}^k (D_i^2 \cdot f_i)}{200}} = \sqrt[2]{\frac{213\,744,0}{200}} = 32,69 \text{ см.}$$

После на основе полученной средней квадратической вычислим сумму площадей сечений древостоя. Для этого определим площадь сечения, соответствующую среднеквадратическому диаметру:

$$g = \frac{\pi \cdot \bar{D}_2^2}{4} = \frac{3,1416 \cdot 32,69^2}{4} = 839,31 \text{ см}^2 = 0,083\,931 \text{ м}^2.$$

Умножив полученную величину на число стволов, получим сумму площадей сечений древостоя:

$$G = g \cdot n = 0,083\,931 \cdot 200 = 16,79 \text{ м}^2.$$

Сравнивая полученную величину с площадью сечения, вычисленной по данным ряда диаметров (табл. 4.1), видим, что среднее квадратическое значение позволяет очень точно определять важнейший таксационный показатель – сумму площадей сечений древостоя.

Аналогичным образом определим средние значения для ряда высот. Для этого составим сначала вспомогательную табл. 4.2.

Таблица 4.2

Вычисление средних значений (высоты)

H_i	f_i	$H_i \cdot f_i$	$H_i^2 \cdot f_i$
18,5	1	18,5	342,3
19,5	2	39,0	760,5
20,5	4	82,0	1 681,0
21,5	13	279,5	6 009,3
22,5	21	472,5	10 631,3
23,5	22	517,0	12 149,5
24,5	43	1 053,5	25 810,8
25,5	40	1 020,0	26 010,0
26,5	18	477,0	12 640,5
27,5	21	577,5	15 881,3
28,5	13	370,5	10 559,3
29,5	1	29,5	870,3
30,5	1	30,5	930,3
Сумма	200	4 967,0	124 276,0

Далее, подставляя полученные суммы в формулы (4.1) и (4.2), вычисляем среднее арифметическое значение

$$\bar{H} = \frac{\sum_{i=1}^k (H_i \cdot f_i)}{n} = \frac{4967,0}{200} = 24,84 \text{ м}$$

и среднее квадратическое:

$$\bar{H}_2 = \sqrt{\frac{\sum_{i=1}^k (H_i^2 \cdot f_i)}{200}} = \sqrt{\frac{124\,276,0}{200}} = 24,93 \text{ м.}$$

Задание 2. Вычислим рассмотренные выше показатели вариации для диаметров и высот деревьев. Проще всего определить размах вариации. Для этого достаточно найти минимальное и максимальное значения:

$$R_D = D_{\max} - D_{\min} = 60,0 - 16,0 = 44,0 \text{ см} - \text{ для диаметров};$$

$$R_H = H_{\max} - H_{\min} = 30,3 - 18,5 = 11,8 \text{ м} - \text{ для высот}.$$

Для того чтобы определить остальные показатели вариации, составим по данным вариационных рядов диаметров и высот вспомогательные табл. 4.3 и 4.4.

Таблица 4.3

Вычисление показателей вариации (диаметры)

D_i	f_i	$D_i - \bar{D}$	$(D_i - \bar{D})^2 \cdot f_i$	$(D_i - \bar{D})^3 \cdot f_i$	$(D_i - \bar{D})^4 \cdot f_i$
16,0	3	-15,70	739,47	-11 609,68	182 271,98
20,0	13	-11,70	1779,57	-20 820,97	243 605,35
24,0	29	-7,70	1 719,41	-13 239,46	101 943,84
28,0	55	-3,70	752,95	-2 785,92	10 307,90
32,0	30	0,30	2,70	0,81	0,24
36,0	32	4,30	591,68	2 544,22	10 940,15
40,0	17	8,30	1 171,13	9 720,38	80 679,15
44,0	12	12,30	1 815,48	22 330,40	274 663,92
48,0	2	16,30	531,38	8 661,49	141 182,29
52,0	5	20,30	2 060,45	41 827,14	849 090,94
56,0	0	24,30	0	0	0
60,0	2	28,30	1 601,78	45 330,37	1 282 849,47
Сумма	200	-	12 766,00	81 958,78	3 177 535,23

Таблица 4.4

Вычисление показателей вариации (высоты)

H_i	f_i	$H_i - \bar{H}$	$(H_i - \bar{H})^2 \cdot f_i$	$(H_i - \bar{H})^3 \cdot f_i$	$(H_i - \bar{H})^4 \cdot f_i$
18,5	1	-6,34	40,2	-254,87	1 615,88
19,5	2	-5,34	57,03	-304,54	1 626,24
20,5	4	-4,34	75,34	-326,98	1 419,09
21,5	13	-3,34	145,02	-484,37	1 617,8
22,5	21	-2,34	114,99	-269,08	629,65
23,5	22	-1,34	39,5	-52,93	70,93
24,5	43	-0,34	4,97	-1,69	0,57
25,5	40	0,66	17,42	11,5	7,59
26,5	18	1,66	49,6	82,34	136,68
27,5	21	2,66	148,59	395,25	1 051,37
28,5	13	3,66	174,14	637,35	2 332,7
29,5	1	4,66	21,72	101,22	471,69
30,5	1	5,66	32,04	181,35	1 026,44
Сумма	200	-	920,56	-285,45	12 006,63

Подставляя значения из этих таблиц в формулы (4.4)–(4.10), получим оценки остальных показателей вариации для диаметров и высот:

– диаметры:

$$S_D^2 = \frac{\sum_{i=1}^k f_i \cdot (D_i - \bar{D})^2}{n} = \frac{12\,766,0}{200} = 63,83;$$

$$\sigma_D^2 = \frac{\sum_{i=1}^k f_i \cdot (D_i - \bar{D})^2}{n-1} = \frac{12\,766,0}{199} = 64,15;$$

$$S_D = \sqrt{\frac{\sum_{i=1}^k f_i \cdot (D_i - \bar{D})^2}{n}} = \sqrt{\frac{12\,766,0}{200}} = 7,989;$$

$$\sigma_D = \sqrt{\frac{\sum_{i=1}^k f_i \cdot (D_i - \bar{D})^2}{n-1}} = \sqrt{\frac{12\,766}{199}} = 8,009;$$

$$V_D = \frac{\sigma_D}{\bar{D}} \cdot 100\% = \frac{7,989}{31,70} \cdot 100\% = 25,33\%;$$

$$A_D = \frac{\sum_{i=1}^k f_i \cdot (D_i - \bar{D})^3}{n \cdot \sigma_D^3} = \frac{81\,958,78}{200 \cdot 8,009^3} = 0,7977;$$

$$E_D = \frac{\sum_{i=1}^k f_i \cdot (D_i - \bar{D})^4}{n \cdot \sigma_D^4} - 3 = \frac{3\,177\,535,23}{200 \cdot 8,009^4} = 3,8614 - 3 = 0,8614;$$

– ВЫСОТЫ:

$$S_H^2 = \frac{\sum_{i=1}^k f_i \cdot (H_i - \bar{H})^2}{n} = \frac{920,56}{200} = 4,60;$$

$$\sigma_H^2 = \frac{\sum_{i=1}^k f_i \cdot (H_i - \bar{H})^2}{n-1} = \frac{920,56}{199} = 4,63;$$

$$S_H = \sqrt{\frac{\sum_{i=1}^k f_i \cdot (H_i - \bar{H})^2}{n}} = \sqrt{\frac{920,56}{200}} = 2,145;$$

$$\sigma_H = \sqrt{\frac{\sum_{i=1}^k f_i \cdot (H_i - \bar{H})^2}{n-1}} = \sqrt{\frac{920,56}{199}} = 2,151;$$

$$V_H = \frac{\sigma_H}{H} \cdot 100\% = \frac{2,151}{24,84} \cdot 100\% = 8,68\%;$$

$$A_H = \frac{\sum_{i=1}^k f_i \cdot (H_i - \bar{H})^3}{n \cdot \sigma_H^3} = \frac{-285,45}{200 \cdot 2,151^3} = -0,1434;$$

$$E_H = \frac{\sum_{i=1}^k f_i \cdot (H_i - \bar{H})^4}{n \cdot \sigma_H^4} - 3 = \frac{12\,006,63}{200 \cdot 2,151^4} = 2,8043 - 3 = -0,1957.$$

Задание 3. Вычислим стандартные ошибки полученных оценок статистических показателей для определения их точности:

диаметры

$$m_{\bar{D}} = \frac{\sigma_D}{\sqrt{n}} = \frac{8,009}{\sqrt{200}} = 0,566;$$

$$m_{\sigma} = \frac{\sigma_D}{\sqrt{2 \cdot n}} = \frac{8,009}{\sqrt{2 \cdot 200}} = 0,400;$$

$$m_V = \frac{V_D}{\sqrt{2 \cdot n}} = \frac{25,33}{\sqrt{2 \cdot 200}} = 1,791;$$

$$m_A = \sqrt{\frac{6}{n}} = \sqrt{\frac{6}{200}} = 0,173;$$

$$m_E = 2 \cdot \sqrt{\frac{6}{n}} = 2 \cdot \sqrt{\frac{6}{200}} = 0,346;$$

высоты

$$m_{\bar{H}} = \frac{\sigma_H}{\sqrt{n}} = \frac{2,151}{\sqrt{200}} = 0,152;$$

$$m_{\sigma} = \frac{\sigma_H}{\sqrt{2 \cdot n}} = \frac{2,151}{\sqrt{2 \cdot 200}} = 0,108;$$

$$m_V = \frac{V_H}{\sqrt{2 \cdot n}} = \frac{8,68}{\sqrt{2 \cdot 200}} = 0,614;$$

$$m_A = \sqrt{\frac{6}{n}} = \sqrt{\frac{6}{200}} = 0,173;$$

$$m_E = 2 \cdot \sqrt{\frac{6}{n}} = 2 \cdot \sqrt{\frac{6}{200}} = 0,346.$$

Вычислим показатель точности для рассматриваемого примера по формуле (4.16):

$$P_D = \frac{m_{\bar{D}}}{D} \cdot 100\% = \frac{0,566}{31,70} \cdot 100\% = 1,79\% \text{ — диаметры;}$$

$$P_H = \frac{m_{\bar{H}}}{H} \cdot 100\% = \frac{0,152}{24,84} \cdot 100\% = 0,61\% \text{ — высоты.}$$

СТРУКТУРНЫЕ ХАРАКТЕРИСТИКИ СТАТИСТИЧЕСКОГО РЯДА

Цель лабораторной работы: рассчитать структурные характеристики вариационных рядов.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Наряду со степенными средними и показателями вариации для характеристики экспериментальных данных используются так называемые структурные характеристики.

Мода. Значение признака, которое наиболее часто встречается в выборке, называется *модой*. Если данные сгруппированы, то класс с максимальным количеством вариантов, называется *модальным*.

Распределение, имеющее один модальный класс, называется *унимодальным*. Если распределение имеет два или более максимума, то оно называется *бимодальным* или *мультимодальным* соответственно.

Если при анализе непрерывно варьирующего признака исходные данные сгруппированы в интервальный ряд, то мода может находиться в любом месте модального интервала. Ее местоположение можно оценить, смоделировав зависимость частоты от величины исследуемого признака в модальном и двух соседних с ним интервалах с помощью параболы второго порядка. Такой подход позволяет получить формулу для приближенной оценки моды:

$$M_o = x_{M_o} + \lambda \cdot \frac{f_{m-1} - f_{m+1}}{2 \cdot (f_{m+1} - 2 \cdot f_m + f_{m-1})}, \quad (5.1)$$

где x_{Mo} – центр модального интервала; λ – величина интервала; f_{m-1} – частота интервала, предшествующего модальному; f_{m+1} – частота интервала, следующего за модальным; f_m – частота модального интервала.

Медиана. Положение экспериментальных данных достаточно хорошо характеризуется различными степенными средними. Однако в случае малой выборки на величину этих статистик могут оказывать довольно значительное влияние крайние варианты, которые являются наименее характерными элементами выборки. Этого недостатка лишена медиана, значение которой определяется наиболее типичными элементами выборки. *Медиана* – это значение признака, которое делит всю выборку на две равные части – половина вариант имеет значения меньшие, чем медиана, а половина – большие.

Проще всего значение медианы определяется в случае несгруппированного набора данных – для этого надо предварительно упорядочить все элементы выборки по возрастанию (*ранжировать*). Например, в выборке из 101 наблюдения медиана будет равна значению 51-го признака, в выборке из 100 наблюдений значение медианы будет находиться между значениями 50-го и 51-го признаков.

В случае, если медиану надо определить для сгруппированного набора данных, начинают с того, что находят, в каком классе она расположено. Проще всего это сделать при наличии накопленных частот вариационного ряда. Класс, в котором находится медиана (медианный интервал), – это первый интервал, у которого накопленная частота окажется больше $n/2$. Предполагая, что внутри медианного интервала наблюдения располагаются равномерно, медиану можно определить по формуле

$$Me = x_{Me} - \frac{\lambda}{2} + \lambda \cdot \left(\frac{\frac{n}{2} - \sum_{i=1}^{j-1} f_i}{f_j} \right), \quad (5.2)$$

где x_{Me} – центр медианного интервала; λ – величина интервала; n – объем выборки; j – номер медианного интервала; $\sum_{i=1}^{j-1} f_i$ – накопленная частота предшествующего медианному интервалу; f_j – частота медианного интервала.

Квантили. Медиана делит вариационный ряд на две равные части. В более общем случае мы можем разделить вариационный ряд на две неравные части в любом соотношении. Статистики, которые отделяют от вариационного ряда определенную часть его членов, называются *квантилями*.

Квантили, отделяющие от вариационного ряда 1, 2, ..., 99 процентов его членов, называются *перцентилями*. С помощью 99 перцентилей P_1, P_2, \dots, P_{99} вариационный ряд делится на 100 равных частей. Девять статистик, которые делят вариационный ряд на десять одинаковых частей, называются *децилями*. *Квартилями* называют три квантиля (Q_1, Q_2 и Q_3), делящие вариационный ряд на четыре равные части. Они соответствуют перцентильям, отделяющим от ранжированного ряда наблюдений 25, 50 и 75% вариант соответственно:

$$Q_1 = P_{25}, \quad Q_2 = P_{50}, \quad Q_3 = P_{75}.$$

Кроме того, квартиль и перцентиль, делящие вариационный ряд на две равные части, соответствуют медиане ряда наблюдений:

$$Q_2 = P_{50} = Me.$$

В практике чаще всего используют перцентили $P_3, P_{10}, P_{25}, P_{50}, P_{75}, P_{90}$ и P_{97} . Определяют квантили аналогично тому, как и медиану. В том случае, если анализируется интервальный вариационный ряд, можно воспользоваться формулой

$$P_L = x_L - \frac{\lambda}{2} + \lambda \cdot \left(\frac{\frac{L \cdot n}{100} - \sum_{i=1}^{j-1} f_i}{f_L} \right), \quad (5.3)$$

где P_L – квантиль, отделяющий от ранжированного ряда L процентов наблюдений; x_L – центр интервала, в который попадает квантиль P_L ; j – номер интервала, в который попадает квантиль P_L ; L – процент наблюдений в выборке, которые меньше, чем квантиль P_L ; $\sum_{i=1}^{j-1} f_i$ – накопленная частота интервала, предшествующего интервалу, в котором находится квантиль P_L . Для определения, в каком интервале находится квантиль, следует воспользоваться накопленными частотами ряда распределения. Первый интервал, у которого накопленная частота окажется больше, чем величина $(L \cdot n / 100)$, и будет таким классом.

Задание. Рассчитать структурные характеристики для вариационного ряда диаметров и высот.

Порядок выполнения работы

Задание. Рассмотрим процесс вычисления структурных характеристик на примере вариационных рядов по диаметру и высоте. Пользуясь формулами (5.1)–(5.3), а также данными табл. 1.3 и 1.4, вычислим структурные характеристики для диаметров и высот:

– диаметры:

$$Mo_D = x_{Mo,D} + \lambda_D \cdot \frac{f_{m-1} - f_{m+1}}{2 \cdot (f_{m+1} - 2 \cdot f_m + f_{m-1})} =$$

$$= 28,0 + 4,0 \cdot \frac{29 - 30}{2 \cdot (30 - 2 \cdot 55 + 29)} = 28,0 + 4,0 \cdot 0,0098 = 28,0;$$

$$Me_D = x_{Me,D} - \frac{\lambda_D}{2} + \lambda_D \cdot \left(\frac{\frac{n}{2} - \sum_{i=1}^{j-1} f_i}{f_j} \right) = 32,0 - \frac{4,0}{2} + 4,0 \cdot \left(\frac{\frac{200}{2} - 100}{30} \right) =$$

$$= 32,0 - 2,0 + 4,0 \cdot 0 = 30,0;$$

$$Q_{1,D} = P_{25,D} = x_{25,D} - \frac{\lambda_D}{2} + \lambda_D \cdot \left(\frac{\frac{25 \cdot n}{100} - \sum_{i=1}^{j-1} f_i}{f_{25}} \right) = 28,0 - \frac{4,0}{2} +$$

$$+ 4,0 \cdot \left(\frac{\frac{25 \cdot 200}{100} - 45}{55} \right) = 28,0 - 2,0 + 4,0 \cdot 0,0909 = 26,4;$$

$$Me_D = Q_{2,D} = P_{50,D} = x_{50,D} - \frac{\lambda_D}{2} + \lambda_D \cdot \left(\frac{\frac{50 \cdot n}{100} - \sum_{i=1}^{j-1} f_i}{f_{50}} \right) = 32,0 - \frac{4,0}{2} +$$

$$+ 4,0 \cdot \left(\frac{\frac{50 \cdot 200}{100} - 100}{30} \right) = 32,0 - 2,0 + 4,0 \cdot 0 = 30,0;$$

$$Q_{3,D} = P_{75,D} = x_{75,D} - \frac{\lambda_D}{2} + \lambda_D \cdot \left(\frac{\frac{75 \cdot n}{100} - \sum_{i=1}^{j-1} f_i}{f_{75}} \right) = 36,0 - \frac{4,0}{2} +$$

$$+ 4,0 \cdot \left(\frac{\frac{75 \cdot 200}{100} - 130}{32} \right) = 36,0 - 2,0 + 4,0 \cdot 0,6250 = 36,5;$$

– ВЫСОТЫ:

$$Mo_H = x_{Mo,H} + \lambda_H \cdot \frac{f_{m-1} - f_{m+1}}{2 \cdot (f_{m+1} - 2 \cdot f_m + f_{m-1})} =$$

$$= 24,5 + 1,0 \cdot \frac{22 - 40}{2 \cdot (40 - 2 \cdot 43 + 22)} = 24,5 + 1,0 \cdot 0,3750 = 24,9;$$

$$Me_H = x_{Me,H} - \frac{\lambda_H}{2} + \lambda_H \cdot \left(\frac{\frac{n}{2} - \sum_{i=1}^{j-1} f_i}{f_j} \right) = 24,5 - \frac{1,0}{2} + 1,0 \cdot \left(\frac{\frac{200}{2} - 63}{43} \right) =$$

$$= 24,5 - 0,5 + 1,0 \cdot 0,8605 = 24,9;$$

$$Q_{1,D} = P_{25,D} = x_{25,D} - \frac{\lambda_D}{2} + \lambda_D \cdot \left(\frac{\frac{25 \cdot n}{100} - \sum_{i=1}^{j-1} f_i}{f_{25}} \right) = 23,5 - \frac{1,0}{2} +$$

$$+ 1,0 \cdot \left(\frac{\frac{25 \cdot 200}{100} - 41}{22} \right) = 23,5 - 0,5 + 1,0 \cdot 0,4091 = 23,4;$$

$$Me_H = Q_{2,H} = P_{50,H} = x_{50,H} - \frac{\lambda_H}{2} + \lambda_H \cdot \left(\frac{\frac{50 \cdot n}{100} - \sum_{i=1}^{j-1} f_i}{f_{50}} \right) = 24,5 - \frac{1,0}{2} +$$

$$+ 1,0 \cdot \left(\frac{\frac{50 \cdot 200}{100} - 63}{43} \right) = 24,5 - 0,5 + 1,0 \cdot 0,8605 = 24,9;$$

$$Q_{3,H} = P_{75,H} = x_{75,H} - \frac{\lambda_H}{2} + \lambda_H \cdot \left(\frac{\frac{75 \cdot n}{100} - \sum_{i=1}^{j-1} f_i}{f_{75}} \right) = 26,5 - \frac{1,0}{2} +$$

$$+ 1,0 \cdot \left(\frac{\frac{75 \cdot 200}{100} - 148}{18} \right) = 26,5 - 0,5 + 1,0 \cdot 0,2222 = 26,2.$$



НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН

Цель лабораторной работы: изучить нормальное распределение случайных величин; научиться рассчитывать теоретические частоты нормального распределения.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Любая случайная величина подчинена какому-либо, как правило, неизвестному закону распределения. *Распределение* – это соотношение между значениями случайной величины и частотой их встречаемости. Одной из задач биометрии и является определение закона распределения анализируемой случайной величины. Обычно решение этой задачи начинается с проверки гипотезы о нормальном распределении данных.

Нормальное распределение (распределение Гаусса) используется для описания распределения непрерывных количественных признаков, на которые действуют множество независимых факторов, при отсутствии доминирующих.

Ее суть заключается в том, что частота отклонения отдельных вариантов от средней арифметической данной совокупности есть функция их величины. Вероятность частоты той или иной варианты в генеральной совокупности и определяется этой функцией. Основными параметрами, характеризующими нормальное распределение, являются *средняя арифметическая* и *стандартное отклонение*.



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 распределенными. 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 счет 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).

Нормальное распределение имеет важное значение в биометрии. На практике очень часто исследуемые случайные величины подчиняются этому закону. Для того чтобы узнать, подчиняется ли случайная величина закону нормального распределения или нет, надо вычислить теоретические частоты вариационного ряда исходя из предположения о нормальном распределении анализируемого параметра и сравнить их с эмпирическими частотами.

Функция нормального распределения $F(x)$ имеет вид

$$F(x) = \int_{-\infty}^x P(z) \cdot d \cdot z = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \int_{-\infty}^x e^{-\frac{(z-m)^2}{2 \cdot \sigma^2}} d \cdot z. \quad (6.1)$$

Интегралы, входящие в это выражение, нельзя выразить через элементарные функции, но их можно вычислить через специальную функцию, которая является интегральной функцией нормального распределения с параметрами $m = 0$ и $\sigma = 1$. Для этого следует перейти к нормированной случайной величине.

Нормированное отклонение – отклонение той или другой варианты (или группы вариантов) от средней арифметической, выра-

женное в сигмах $t_i = \frac{x_i - m}{\sigma}$:

$$F(x) = \frac{1}{\sqrt{2 \cdot \pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} d \cdot t. \quad (6.2)$$

Зная значения параметров m и σ , можно определить теоретические вероятности попадания исследуемой случайной величины в интервалы вариационного ряда (а следовательно, и его теоретические частоты), исходя из предположения, что она подчинена закону нормального распределения. Это позволит изобразить графически кривую нормального распределения (рис. 6.1) и сравнить теоретические и эмпирические частоты вариационного ряда, на основании чего можно будет решить, следует эмпирическое распределение нормальному закону или нет.

Кривая нормального распределения (рис. 6.1) обладает свойствами:

- однозначно определяется двумя параметрами – средним значением и среднеквадратическим отклонением;

- кривая симметрична относительно среднего значения и имеет колоколообразную форму, которая зависит от величины σ , являющейся параметром масштаба, а положение определяется \bar{x} ;

- кривая имеет один максимум, равный $\frac{1}{\sigma\sqrt{2\pi}}$, и две точки перегиба на расстоянии $\pm 1\sigma$ от \bar{x} ;

- ветви кривой асимптотически приближаются к оси абсцисс на расстояние $\pm\infty$.

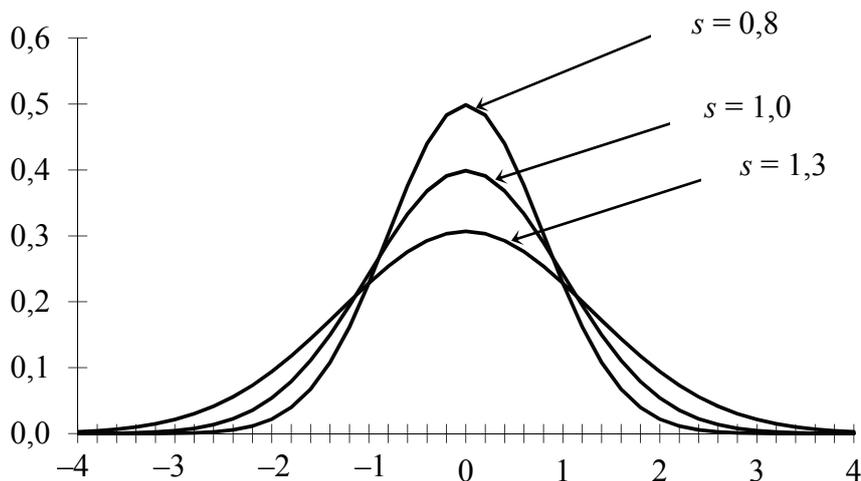


Рис. 6.1. Кривые нормального распределения с разными значениями параметра формы

Практическое значение – основные статистические методы (корреляционный, дисперсионный, регрессионный анализы, t -критерий и др.) основано на предположении о нормальном распределении используемых количественных признаков.

Задание. Рассчитать теоретические частоты нормального распределения для вариационного ряда диаметров и высот и сравнить их с эмпирическими частотами, построить графики сравнения.

Порядок выполнения работы

Задание. Вычислим теоретические частоты для вариационных рядов по диаметру и высоте с помощью данных табл. 1 прил., предполагая, что исследуемая случайная величина распределена по нормальному закону. Для выполнения вычислений составим вспомогательную таблицу (табл. 6.1).

С учетом того, что оценкой параметров нормального распределения методом моментов являются *среднеквадратическое отклонение* и *среднее арифметическое*, вычислим нормированные нижнюю и верхнюю границы интервалов следующим образом:

$$t_i^H = \frac{x_i - \frac{\lambda}{2} - \bar{x}}{\sigma}; \quad (6.3)$$

$$t_i^B = \frac{x_i + \frac{\lambda}{2} - \bar{x}}{\sigma}. \quad (6.4)$$

В отличие от анализируемого вариационного ряда, нормальное распределение определено на интервале от $-\infty$ до $+\infty$. Для того чтобы области определения эмпирического и нормального распределения сделать одинаковыми, добавим **2** дополнительных интервала: *один* перед первым интервалом с границами от $-\infty$ до нижней границы первого интервала, а *второй* – после последнего интервала с границами от верхней границы последнего интервала до $+\infty$. Эмпирические частоты этих дополнительных интервалов будут равны 0 (нулю), так как в исходных данных нет ни одного наблюдения, которое было бы меньше нижней границы первого интервала или больше верхней границы последнего интервала.

Таблица 6.1

**Вычисление теоретических частот
для функции нормального распределения (диаметры)**

D_i	f_i	t_i^H	t_i^B	$\Phi(t_i^H)$	$\Phi(t_i^B)$	P_i	\tilde{f}_i	$\Delta = f_i - \tilde{f}_i$
10,0	0	$-\infty$	-2,46	0,000	0,007	0,007	1,4	-1,4
14,0	3	-2,46	-1,96	0,007	0,025	0,018	3,6	-0,6
18,0	13	-1,96	-1,46	0,025	0,072	0,047	9,4	3,6
22,0	29	-1,46	-0,96	0,072	0,169	0,097	19,4	9,6
26,0	55	-0,96	-0,46	0,169	0,323	0,154	30,8	24,2
30,0	30	-0,46	0,04	0,323	0,516	0,193	38,6	-8,6
34,0	32	0,04	0,54	0,516	0,705	0,189	37,8	-5,8
38,0	17	0,54	1,04	0,705	0,851	0,146	29,2	-12,2
42,0	12	1,04	1,54	0,851	0,938	0,087	17,4	-5,4
46,0	2	1,54	2,04	0,938	0,979	0,041	8,2	-6,2
50,0	5	2,04	2,53	0,979	0,994	0,015	3,0	2,0
54,0	0	2,53	3,03	0,994	0,999	0,005	1,0	-1,0
58,0	2	3,03	3,53	0,999	1,000	0,001	0,2	1,8
62,0	0	3,53	$+\infty$	1,000	1,000	0,000	0,0	0,0
Сумма	200	—	—	—	—	1,000	200,0	0,0

Значения функции нормированного нормального распределения для нижней $\Phi(t_i^H)$ и верхней $\Phi(t_i^B)$ границ интервалов можно найти с помощью табл. 1 прил., используя в качестве аргументов значения t_i^H и t_i^B соответственно (значения функции для $-\infty$ и $+\infty$ будут равны 0 и 1 соответственно). В этой таблице значения функции распределения даны только для положительных аргументов. Если надо найти функцию распределения для отрицательного аргумента, следует воспользоваться соотношением $\Phi(-x) = 1 - \Phi(x)$, которое справедливо, так как нормальное распределение является симметричным.

Вероятности для интервалов вариационного ряда вычисляются как разность значений функции распределения для верхней и нижней границ:

$$P_i = \Phi(t_i^B) - \Phi(t_i^H). \quad (6.5)$$

Теперь можно найти теоретические частоты ряда:

$$\tilde{f}_i = n \cdot P_i. \quad (6.6)$$

Аналогичным образом можно вычислить теоретические частоты для вариационного ряда высот (табл. 6.2).

Таблица 6.2

**Вычисление теоретических частот
для функции нормального распределения (высоты)**

H_i	f_i	t_i^H	t_i^B	$\Phi(t_i^H)$	$\Phi(t_i^B)$	P_i	\tilde{f}_i	$\Delta = f_i - \tilde{f}_i$
17,5	0	$-\infty$	-3,18	0,000	0,001	0,001	0,2	-0,2
18,5	1	-3,18	-2,72	0,001	0,003	0,002	0,4	0,6
19,5	2	-2,72	-2,25	0,003	0,012	0,009	1,8	0,2
20,5	4	-2,25	-1,79	0,012	0,037	0,025	5,0	-1,0
21,5	13	-1,79	-1,32	0,037	0,093	0,056	11,2	1,8
22,5	21	-1,32	-0,86	0,093	0,195	0,102	20,4	0,6
23,5	22	-0,86	-0,39	0,195	0,348	0,153	30,6	-8,6
24,5	43	-0,39	0,07	0,348	0,528	0,180	36,0	7,0
25,5	40	0,07	0,54	0,528	0,705	0,177	35,4	4,6
26,5	18	0,54	1,00	0,705	0,841	0,136	27,2	-9,2
27,5	21	1,00	1,47	0,841	0,929	0,088	17,6	3,4
28,5	13	1,47	1,93	0,929	0,973	0,044	8,8	4,2
29,5	1	1,93	2,40	0,973	0,992	0,019	3,8	-2,8
30,5	1	2,40	2,86	0,992	0,998	0,006	1,2	-0,2
31,5	0	2,86	$+\infty$	0,998	1,000	0,002	0,4	-0,4
Сумма	200	-	-	-	-	1,000	200,0	0,0

Последние колонки табл. 6.1 и 6.2, представляющие собой разность между эмпирическими и теоретическими частотами, дают нам информацию о близости теоретического (в данном случае нормального) и эмпирического распределений. Однако по данным отклонениям достаточно трудно принять решение о согласованности эмпирического и теоретического распределений.

Более наглядную картину можно увидеть, изобразив эти распределения графически (рис. 6.2 и 6.3) в виде совмещенного графика гистограммы (эмпирические частоты) и *сглаженной* линии

(теоретические частоты). Однако такие сравнения распределений будут субъективными.

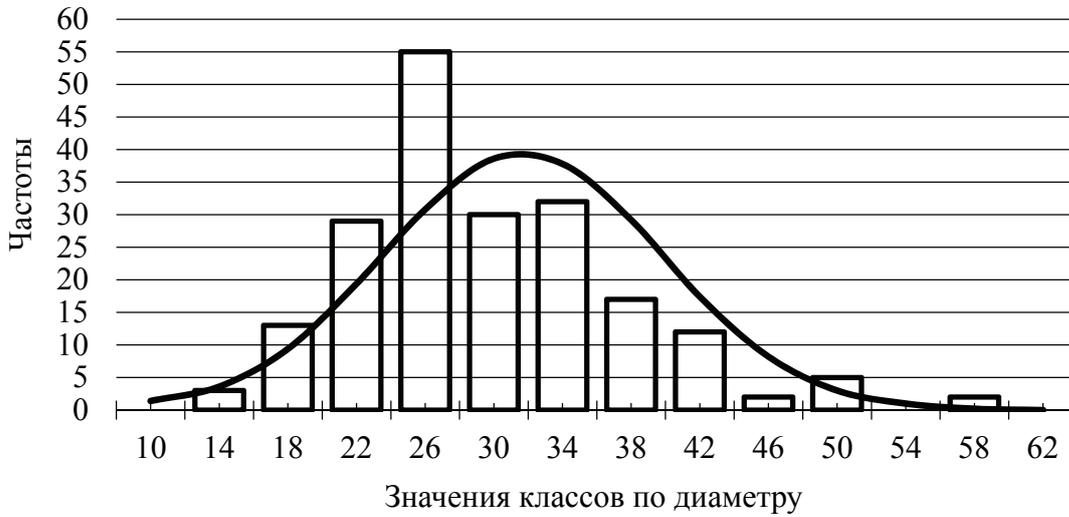


Рис. 6.2. Сравнение эмпирического и нормального распределений сосновых стволов по диаметрам

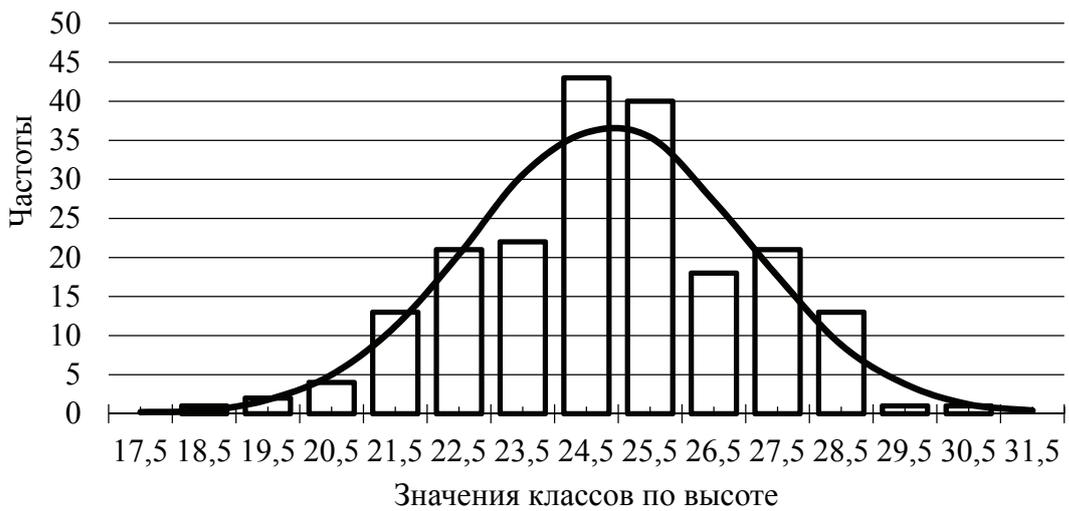


Рис. 6.3. Сравнение эмпирического и нормального распределений сосновых стволов по высотам

Для того чтобы дать объективную оценку согласованности эмпирических и теоретических распределений, необходимо воспользоваться специальными методиками проверки статистических гипотез, например, с помощью программы Statistica или MS Excel.

СТАТИСТИЧЕСКАЯ ПРОВЕРКА НЕПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ

Цель лабораторной работы: определить, подчиняются ли экспериментальные данные закону о нормальном распределении с помощью критериев согласия Пирсона и Колмогорова.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Для статистической проверки гипотез, т. е. отклонения или принятия ее, применяются так называемые критерии достоверности – специальные статистики, позволяющие судить о надежности выводов относительно принятой гипотезы, ожидаемого результата и т. д.

Все критерии достоверности условно подразделены на две группы: *параметрические* (критерии, определяющие достоверность различий двух выборок по определенным параметрам этой совокупности (средние, дисперсии и т. д.) называются *критериями различия*) и *непараметрические* (критерии, определяющие степень соответствия эмпирического распределения известному теоретическому и не использующие параметры этой совокупности, называются *критериями согласия*).

Критерии согласия Пирсона (χ^2) и Колмогорова – Смирнова (λ) относятся к непараметрическим критериям. При сравнении фактического и теоретического значений этих показателей делаются аналогичные предыдущим выводы о данном распределении. *Недостаток* данных критериев в том, что они дают более грубую оценку различий выборок по сравнению с параметрическими.

Для того чтобы выяснить, подчиняются ли экспериментальные данные какому-либо закону распределения, надо сформулировать статистическую гипотезу в отношении распределения анализируемой случайной величины и затем проверить ее.

Рассмотрим процесс проверки непараметрической гипотезы с помощью одного из критериев согласия – критерия согласия Пирсона.

На первом этапе следует выдвинуть нулевую гипотезу, состоящую в том, что анализируемый признак подчиняется какому-либо закону распределения (рисунок). Далее, исходя из предположения о том, что нулевая гипотеза справедлива, следует вычислить статистику χ^2 :

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - \tilde{f}_i)^2}{\tilde{f}_i}. \quad (7.1)$$

Полученное значение сравнивается с квантилем распределения Пирсона χ^2 , который в качестве параметров распределения использует уровень значимости (обычно $\alpha = 0,05$) и число степеней свободы:

$$\gamma = k - p - 1, \quad (7.2)$$

где k – общее число степеней свободы, равное числу слагаемых в формуле (7.1); p – число параметров теоретической функции распределения, которые оценивались по анализируемым данным (для нормального распределения – 2).



Схема принятия решения о нулевой гипотезе
с помощью критериев согласия

Уровень значимости – это тот теоретический процент значений нормального распределения, который можно отбросить, не

учитывать, дабы с меньшими усилиями получить основную информацию об изучаемом явлении. Уровень значимости, равный 0,05 (5%), можно интерпретировать так: имеется всего 5% шансов, что полученная величина не будет соответствовать изучаемой совокупности.

При вычислении критерия Пирсона следует иметь в виду, что теоретические частоты не должны быть меньше пяти. В том случае, если теоретические частоты оказываются недостаточно большими, нужно объединять маленькие классы в большие.

Для оценки достоверности различий эмпирического распределения от теоретического следует сравнить вычисленную величину χ^2 с табличной. Если вычисленная величина χ^2 равна или превышает табличную $\chi^2(\alpha, \gamma)$, то эмпирическое распределение от теоретического отличается достоверно. Тем самым гипотеза об отсутствии этих различий будет опровергнута. Если же $\chi^2 < \chi^2(\alpha, \gamma)$, нулевая гипотеза остается в силе.

Для оценки достоверности различий между фактическим и теоретическим распределениями наряду с критерием Пирсона (χ^2) может применяться также и критерий λ_K (лямбда), предложенный А. Н. Колмогоровым и Н. В. Смирновым:

$$\lambda_K = D_{\max} \cdot \sqrt{n}. \quad (7.3)$$

Условием применения критерия служит достаточное число (не менее 100) наблюдений. Для проверки непараметрической гипотезы о виде распределения случайной величины необходимо вычисленную статистику λ сравнить с квантилем распределения Колмогорова λ_α . В том случае, если вычисленная статистика окажется больше, чем табличное значение $\lambda > \lambda_\alpha$, мы отвергаем нулевую гипотезу о виде распределения, в противном случае гипотеза принимается. Если при проверке непараметрических гипотез используются оценки параметров распределения, полученные по материалам анализируемой выборки, то критерий Колмогорова показывает лучшее согласие теоретических распределений с экспериментальными данными, чем критерий согласия Пирсона χ^2 .

Задание 1. Рассчитать критерий согласия Пирсона и определить, подчиняются ли экспериментальные данные закону о нормальном распределении.

Задание 2. Определить, подчиняются ли экспериментальные данные закону о нормальном распределении с помощью критерия согласия Колмогорова.

Порядок выполнения работы

Задание 1. Рассмотрим процесс проверки непараметрической гипотезы с помощью критерия Пирсона на примере распределений диаметров и высот деревьев в сосновом древостое. Нулевая гипотеза будет заключаться в предположении, что анализируемые случайные величины подчиняются закону нормального распределения. Исходя из такого предположения, вычислим статистику χ^2 для вариационного ряда по диаметрам. Для этого составим вспомогательную табл. 7.1. При составлении ее эмпирические и теоретические частоты вариационного ряда диаметров можно взять из табл. 6.1. Далее следует объединить интервалы таким образом, чтобы теоретические частоты в укрупненных классах были не меньше *пяти*. Дальнейшие расчеты (три последние колонки табл. 7.1) выполняем, используя эмпирические и теоретические частоты, полученные после укрупнения классов. Сумма, полученная в последней колонке таблицы, и будет статистикой Пирсона χ^2 .

Таблица 7.1

Вычисление критерия согласия Пирсона χ^2 (диаметры)

D_i	f_i		\tilde{f}_i		$f_i - \tilde{f}_i$	$\frac{(f_i - \tilde{f}_i)^2}{\tilde{f}_i}$
	до укрупнения	после укрупнения	до укрупнения	после укрупнения		
10,0	0	3	1,4	5,0	-2,0	0,80
14,0	3		3,6			
18,0	13	13	9,4	9,4	3,6	1,38
22,0	29	29	19,4	19,4	9,6	4,75
26,0	55	55	30,8	30,8	24,2	19,01
30,0	30	30	38,6	38,6	-8,6	1,92
34,0	32	32	37,8	37,8	-5,8	0,89
38,0	17	17	29,2	29,2	-12,2	5,10
42,0	12	12	17,4	17,4	-5,4	1,68
46,0	2	9	8,2	12,4	-3,4	0,93
50,0	5		3,0			
54,0	0		1,0			
58,0	2		0,2			
62,0	0		0,0			
Сумма	200	200	200,0	200,0	0,0	36,46

Пользуясь табл. 2 прил., найдем соответствующий квантиль распределения Пирсона χ^2 , чтобы, сравнивая его с вычисленной статистикой χ^2 , проверить нулевую гипотезу. Уровень значимости (вероятность отклонения правильной нулевой гипотезы) примем $\alpha = 0,05$. С учетом объединения интервалов и того, что мы оценили два параметра нормального распределения (σ и m) по материалам наших экспериментальных данных, вычислим число степеней свободы, пользуясь формулой (7.2):

$$\gamma = k - p - 1 = 9 - 2 - 1 = 6.$$

Определив необходимые параметры, найдем квантиль распределения Пирсона $\chi^2_{0,05;6} = 12,592$ по табл. 2 прил. Так как вычисленная статистика Пирсона $\chi^2 = 36,46$ превышает табличное значение 12,592, то мы отклоняем нулевую гипотезу, т. е. распределение диаметров деревьев в древостое не подчиняется закону нормального распределения.

Аналогичным образом вычислим критерий Пирсона для распределения высот деревьев в древостое (табл. 7.2).

Таблица 7.2

Вычисление критерия согласия Пирсона χ^2 (высоты)

H_i	f_i		\tilde{f}_i		$f_i - \tilde{f}_i$	$\frac{(f_i - \tilde{f}_i)^2}{\tilde{f}_i}$
	до укрупнения	после укрупнения	до укрупнения	после укрупнения		
17,5	0	7	0,2	7,4	-0,4	0,02
18,5	1		0,4			
19,5	2		1,8			
20,5	4		5,0			
21,5	13	13	11,2	11,2	1,8	0,29
22,5	21	21	20,4	20,4	0,6	0,02
23,5	22	22	30,6	30,6	-8,6	2,42
24,5	43	43	36,0	36,0	7,0	1,36
25,5	40	40	35,4	35,4	4,6	0,60
26,5	18	18	27,2	27,2	-9,2	3,11
27,5	21	21	17,6	17,6	3,4	0,66
28,5	13	13	8,8	8,8	4,2	2,00
29,5	1	2	3,8	5,4	-3,4	2,14
30,5	1		1,2			
31,5	0		0,4			
Сумма	200	200	200,0	200,0	0,0	12,62

верхним границам интервалов. Абсолютные величины разностей между эмпирической и теоретической функциями распределения приведены в последней колонке вспомогательной таблицы (по модулю). Максимальное из этих значений равно 0,177.

Таблица 7.3

Вычисление критерия согласия Колмогорова λ (диаметры)

D_i	f_i	Накопленные частоты	Эмпирическая функция распределения (F_n)	t_i^B	$\Phi(t_i^B)$	$ F_n - \Phi(t_i^B) $
14,0	3	3	0,015	-1,96	0,025	0,010
18,0	13	16	0,080	-1,46	0,072	0,008
22,0	29	45	0,225	-0,96	0,169	0,056
26,0	55	100	0,500	-0,46	0,323	0,177
30,0	30	130	0,650	0,04	0,516	0,134
34,0	32	162	0,810	0,54	0,705	0,105
38,0	17	179	0,895	1,04	0,851	0,044
42,0	12	191	0,955	1,54	0,938	0,017
46,0	2	193	0,965	2,04	0,979	0,014
50,0	5	198	0,990	2,53	0,994	0,004
54,0	0	198	0,990	3,03	0,999	0,009
58,0	2	200	1,000	3,53	1,000	0,000
Сумма	200	—	—	—	—	—

Далее с помощью формулы (7.3) вычислим критерий Колмогорова:

$$\lambda = D_{\max} \cdot \sqrt{n} = 0,177 \cdot \sqrt{200} = 2,50.$$

В табл. 3 прил. для уровня значимости $\alpha = 0,05$ найдем квантиль распределения Колмогорова $\lambda_\alpha = 1,36$. Так как вычисленное значение критерия Колмогорова $\lambda = 2,50$ больше, чем табличное $\lambda_\alpha = 1,36$, то гипотезу о нормальном распределении диаметров деревьев в древостое мы отклоняем.

Попробуем с помощью критерия Колмогорова проверить нулевую гипотезу о нормальном распределении высот в сосновых древостоях по материалам, представленным в табл. 1.4 и 6.2. Для этого выполним аналогичные вспомогательные расчеты для распределения по высотам как для распределения стволов по диаметрам в табл. 7.4.

Таблица 7.4

Вычисление критерия согласия Колмогорова λ (высоты)

H_i	f_i	Накопленные частоты	Эмпирическая функция распределения (F_n)	t_i^B	$\Phi(t_i^B)$	$ F_n - \Phi(t_i^B) $
18,5	1	1	0,005	-2,72	0,003	0,002
19,5	2	3	0,015	-2,25	0,012	0,003
20,5	4	7	0,035	-1,79	0,037	0,002
21,5	13	20	0,100	-1,32	0,093	0,007
22,5	21	41	0,205	-0,86	0,195	0,010
23,5	22	63	0,315	-0,39	0,348	0,033
24,5	43	106	0,530	0,07	0,528	0,002
25,5	40	146	0,730	0,54	0,705	0,025
26,5	18	164	0,820	1,00	0,841	0,021
27,5	21	185	0,925	1,47	0,929	0,004
28,5	13	198	0,990	1,93	0,973	0,017
29,5	1	199	0,995	2,40	0,992	0,003
30,5	1	200	1,000	2,86	0,998	0,002
Сумма	200	—	—	—	—	—

Максимальная разность между эмпирической и теоретической функциями распределения для высот равна 0,033. Используя это значение, с помощью формулы (7.3) вычислим критерий Колмогорова:

$$\lambda = D_{\max} \cdot \sqrt{n} = 0,033 \cdot \sqrt{200} = 0,47.$$

Теперь в табл. 3 прил. для уровня значимости $\alpha = 0,05$ найдем квантиль распределения Колмогорова $\lambda_\alpha = 1,36$. Так как вычисленное значение критерия Колмогорова $\lambda = 0,47$ меньше, чем табличное $\lambda_\alpha = 1,36$, то в данном случае мы принимаем гипотезу о нормальном распределении высот деревьев в древостое.

ВЫЧИСЛЕНИЕ ОСНОВНЫХ СТАТИСТИК И АНАЛИЗ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН С ИСПОЛЬЗОВАНИЕМ ПАКЕТА ПРОГРАММ

Цель лабораторной работы: рассчитать основные показатели вариации с использованием специализированных пакетов статистических программ; выполнить анализ распределений случайных величин в программе Statistica.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office и Statistica 10.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

В результате первичной обработки исходных данных исследователь получает простой набор обобщающих количественных характеристик, с помощью которых можно сжато описать выборки любого объема. Расчет всех этих характеристик по формулам вручную или на калькуляторе требует значительных затрат времени. Естественно, что подобный подход в современных исследованиях редко используется, поэтому возникает необходимость овладения навыками подобных расчетов на персональном компьютере, что повышает качество анализа и сокращает затраты времени.

В программе Statistica показатели вариации выборки можно вычислять с помощью встроенных функций. Программной реализацией расчета статистических характеристик выборочной совокупности является модуль *Описательная статистика (Descriptive statistics)*, включенный в большинство пакетов для статистической обработки данных. Работая в данном модуле, достаточно сделать ссылку на массив данных выборки, выбрать необходимые

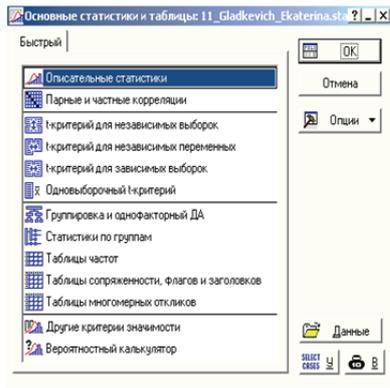


Рис. 8.1. Диалоговое окно модуля *Основные статистики и таблицы*

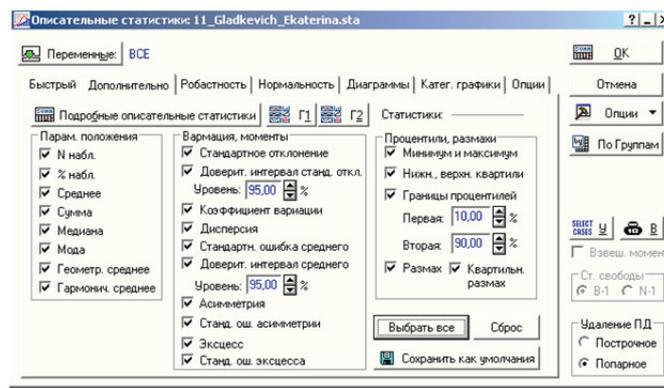


Рис. 8.2. Диалоговое окно модуля *Описательные статистики (Descriptive statistics)* пакета Statistica

6. Для выполнения вычислений нажмем кнопку *OK* (рис. 8.2). На экране появится таблица, содержащая вычисленные основные статистики для выбранных переменных (рис. 8.3).

Рабочая книга1* - Описательные статистики (11_Gladkevich_Ekaterina.sta)										
Описательные статистики (11_Gladkevich_Ekaterina.sta)										
Переменная	N набл.	Среднее	Медиана	Мода	Частота моды	Сумма	Минимум	Максим.	Нижняя Квартиль	Е К
D	200	31,65900	29,95000	27,00000	6	6331,800	16,00000	60,00000	26,50000	:
H	200	24,76450	24,60000	24,60000	12	4952,900	18,50000	30,30000	23,60000	:

Рис. 8.3. Результаты вычисления основных статистик в программе Statistica

7. Полученные результаты сведем в табл. 8.1, в которой, кроме статистик, приведены названия вычисленных показателей, используемые в программе Statistica.

Основные статистические показатели в MS Excel вычисляются с помощью встроенных функций *МАКС* (массив); *МИН* (массив); *СРЗНАЧ* (массив); *ДИСП* (массив); *СТАНДОТКЛОН* (массив); *МОДА* (массив); *МЕДИАНА* (массив); *КВАРТИЛЬ* (массив; часть); *ПЕРСЕНТИЛЬ* (массив; номер); *ЭКСЦЕСС* (массив); *СКОС* (массив) – асимметрия и других или с помощью *Пакета анализа*. Для этого выберем пункт меню *Данные*, опцию *Анализ данных* (рис. 8.4) и выберем *Описательная статистика*. В диалоговом окне *Описательная статистика* (рис. 8.5) выберем массив исходных данных и укажем дополнительные опции (рис. 8.5).

Нажмем кнопку *OK* (рис. 8.5), в результате будет выполнен расчет основных статистик (рис. 8.6) в программе MS Excel.

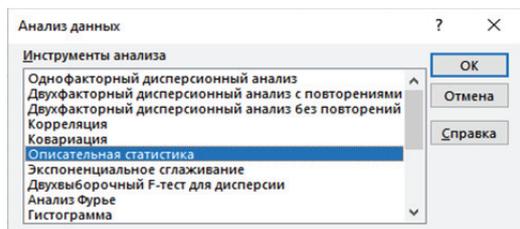


Рис. 8.4. Диалоговое окно *Анализ данных* в MS Excel

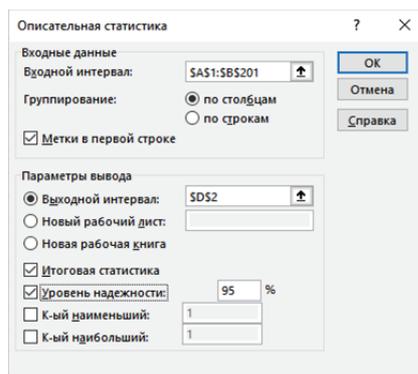


Рис. 8.5. Диалоговое окно *Описательная статистика* в MS Excel

Таблица 8.1

Основные статистики для диаметров и высот деревьев

Название показателя	Название показателя, используемое в программе Statistica	Значение показателя	
		для ряда диаметров	для ряда высот
Число наблюдений	N набл.	200	200
Среднее арифметическое	Среднее	31,66	24,76
Нижний доверительный предел	Доверит. -95,000%	30,547	24,765
Верхний доверительный предел	Доверит. +95,000%	32,771	25,062
Среднее геометрическое	Геометр. среднее	30,723	24,671
Среднее гармоническое	Гармонич. среднее	29,833	24,576
Медиана	Медиана	29,95	27,00
Мода	Мода	27,0	24,6
Частота модальной варианты	Частота моды	6	12
Сумма	Сумма	6331,8	4952,9
Минимальное значение	Минимум	16,0	18,5
Максимальное значение	Максимум	60,0	30,3
Нижний квартиль (первый)	Нижняя квартиль	26,5	23,6
Верхний квартиль (третий)	Верхняя квартиль	36,8	26,1
Перцентиль 10-й	Перцентиль 10,00000	22,6	21,9
Перцентиль 90-й	Перцентиль 90,00000	42,25	27,6
Размах вариации	Размах	44,0	11,8
Дисперсия	Дисперсия	63,547	4,556
Стандартное отклонение	Станд. откл.	7,972	2,134
Коэффициент вариации	Коэф. вар.	25,18	8,62
Стандартная ошибка среднего	Станд. ошибка	0,564	0,151
Коэффициент асимметрии	Асимметрия	0,871	-0,187
Стандартная ошибка коэффициента асимметрии	Стданд. ошибка асимметрия	0,172	0,172
Коэффициент эксцесса	Эксцесс	1,035	-0,100
Стандартная ошибка коэффициента эксцесса	Стданд. ошибка эксцесс	0,342	0,342

	D	H	
Среднее	31,659	Среднее	24,7645
Стандартная ошибка	0,5636792	Стандартная ошибка	0,150932
Медиана	29,95	Медиана	24,6
Мода	27	Мода	24,6
Стандартное отклонение	7,971628	Стандартное отклонение	2,134495
Дисперсия выборки	63,546853	Дисперсия выборки	4,55607
Экцесс	1,0351952	Экцесс	-0,09987
Асимметричность	0,870845	Асимметричность	-0,18651
Интервал	44	Интервал	11,8
Минимум	16	Минимум	18,5
Максимум	60	Максимум	30,3
Сумма	6331,8	Сумма	4952,9
Счет	200	Счет	200
Уровень надежности(95,0%)	1,1115509	Уровень надежности(95,0%)	0,297631

Рис. 8.6. Результаты вычисления основных статистик в программе MS Excel

Задание 2. Анализ распределения случайных величин с помощью пакета Statistica 10.0 можно выполнить следующим образом.

1. Запустим программу Statistica, если не запущена, и откроем файл с данными, как описано в задании 1.

2. Выбрав опцию *Подгонка распределений* из меню *Анализ* (рис. 8.7), откроем диалоговое окно *Подгонка распределений*, содержащее список распределений (рис. 8.8), для которых можно выполнить вычисления. Из данного списка выберем распределение *Нормальное* и нажмем кнопку *OK*.

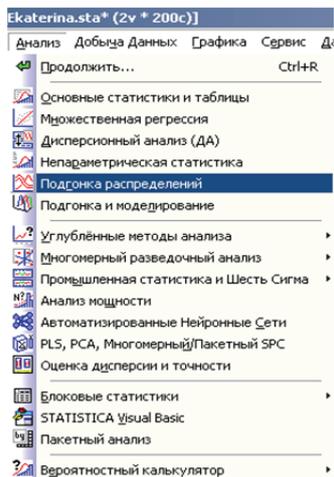


Рис. 8.7. Пункт меню *Подгонка распределений*

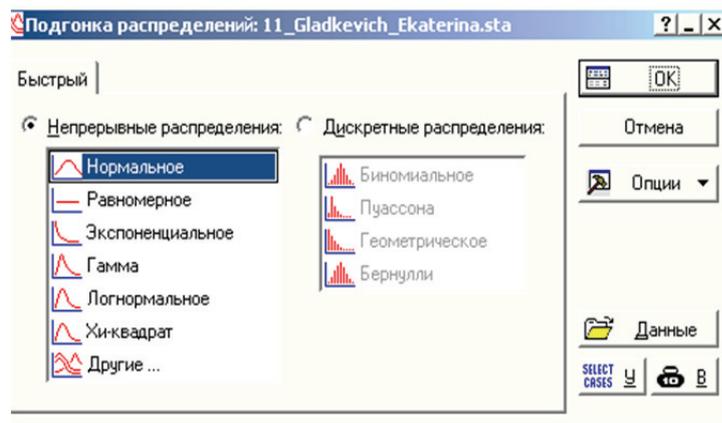


Рис. 8.8. Диалоговое окно модуля *Подгонка распределений* пакета Statistica

3. В открывшемся диалоговом окне *Подгонка к непрерывным распределениям* (рис. 8.9) нажмем кнопку *Переменные*. После того как на экране появится еще одно диалоговое окно, содержащее список переменных, выберем какую-нибудь переменную (для диаметров – *D*) и нажимаем кнопку *ОК* для возврата в окно настройки распределений.

4. Далее во вкладке *Параметры* (рис. 8.10) при необходимости редактируем такие поля, как *Число групп* – число интервалов при группировке данных (для диаметров – 12), *Нижняя граница* – нижняя граница первого интервала минус 0,1 ($\epsilon = 0,1$) (для диаметров – $14,0 - 0,1 = 13,9$), *Верхняя граница* – верхняя граница последнего интервала (для диаметров – 61,9), а также поля, содержащие параметры распределения (можно оставить незаполненными).

5. Для выполнения вычислений нажмем кнопку *ОК* или на вкладке *Быстрый* (рис. 8.9) *Наблюдаемые и ожидаемые частоты*. На экране появится таблица, содержащая эмпирические и теоретические частоты (рис. 8.11).

6. В данной таблице (рис. 8.11) в колонке *Верхняя граница* приведены верхние границы интервалов; в колонке *Наблюд. Частота* – эмпирические частоты; в колонке *Кумул. Наблюд.* – накопленные эмпирические частоты; в колонке *Процент. Наблюд.* – эмпирические частоты, выраженные в процентах от общего количества наблюдений; *Кумул. % Наблюд.* – накопленные эмпирические частоты, выраженные в процентах; *Ожидаем. Частота* – теоретические частоты; *Кумул. Ожидаем.* – накопленные теоретические частоты; *Процент. Ожидаем.* – теоретические частоты, выраженные в процентах; *Кумул. % Ожидаем.* – накопленные теоретические частоты, выраженные в процентах; *Наблюд. – Ожидаем.* – отклонение эмпирических частот от теоретических.

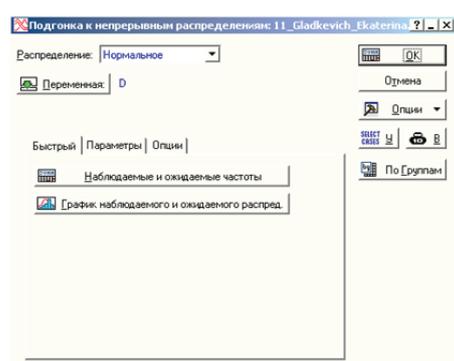


Рис. 8.9. Диалоговое окно *Подгонка к непрерывным распределениям* пакета Statistica

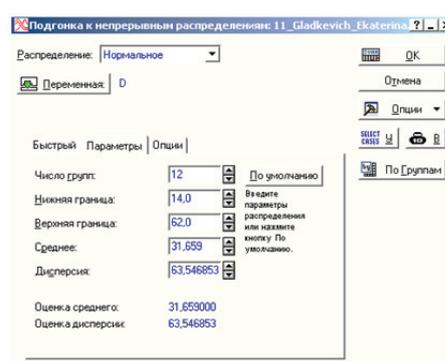


Рис. 8.10. Вкладка *Параметры* пакета Statistica

Рабочая книга1* - Перемен.: D, Распред.: Нормальное (11_Gladkevich_Ekaterina.sta)

Перемен.: D, Распред.: Нормальное (11_Gladkevich_Ekaterina.sta)
 Хи-квадрат = 19,52701, сс = 6 (скорр.), p = 0,00336

Верхняя Граница	Наблюд. Частота	Кумул. Наблюд.	Процент Наблюд.	Кумул. % Наблюд.	Ожидаем. Частота	Кумул. Ожидаем.	Процент Ожидаем.	Кумул. % Ожидаем.	Наблюд. - Ожидаем.
<= 17,90000	3	3	1,50000	1,5000	8,43481	8,4348	4,21740	4,2174	-5,43481
21,90000	13	16	6,50000	8,0000	13,65224	22,0870	6,82612	11,0435	-0,65224
25,90000	29	45	14,50000	22,5000	24,91553	47,0026	12,45776	23,5013	4,08447
29,90000	55	100	27,50000	50,0000	35,53333	82,5359	17,76667	41,2680	19,46667
33,90000	30	130	15,00000	65,0000	39,60241	122,1383	19,80120	61,0692	-9,60241
37,90000	32	162	16,00000	81,0000	34,49319	156,6315	17,24660	78,3158	-2,49319
41,90000	17	179	8,50000	89,5000	23,47812	180,1096	11,73906	90,0548	-6,47812
45,90000	12	191	6,00000	95,5000	12,48790	192,5975	6,24395	96,2988	-0,48790
49,90000	2	193	1,00000	96,5000	5,19012	197,7876	2,59506	98,8938	-3,19012
53,90000	5	198	2,50000	99,0000	1,68531	199,4730	0,84265	99,7365	3,31469
57,90000	0	198	0,00000	99,0000	0,42750	199,9005	0,21375	99,9502	-0,42750
< бесконеч.	2	200	1,00000	100,0000	0,09955	200,0000	0,04977	100,0000	1,90045

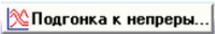
Рис. 8.11. Результаты вычислений для нормального распределения для ряда диаметров, выполненных с помощью программы Statistica

7. Кроме того, в окне результатов (рис. 8.11) приводится информация, необходимая для проверки нулевой гипотезы о согласованности эмпирического и теоретического распределений. Это статистика Пирсона χ^2 – величина *Хи-квадрат (Chi-Square)*, число степеней свободы – *сс (df)* и вероятность совершения ошибки первого рода (вероятность отклонения справедливой нулевой гипотезы) – *p*. Эти данные занесем в таблицу для анализа распределений (табл. 8.2) в строку *Нормальное* распределение.

Таблица 8.2

Анализ распределения диаметров деревьев в древостое

Наименование распределения	Критерий Пирсона χ^2 (<i>Хи-квадрат</i>)	Число степеней свободы (<i>сс</i>)	Вероятность совершения ошибки 1-го рода (<i>p</i>)	Оценка
Нормальное	19,52701	6	0,00336	–
Равномерное	154,61615	9	0,00000	–
Экспоненциальное	462,17025	9	0,00000	–
Гамма	8,67503	5	0,12275	+
Лог-нормальное	7,06830	5	0,21561	+
Хи-квадрат	9,41717	6	0,15144	+

8. Вернемся в диалоговое окно *Подгонка к непрерывным распределениям* (рис. 8.9). Для этого над кнопкой *Пуск* нажмем на одноименную кнопку в окне программы Statistica  *Подгонка к непрерывным распределениям*. Далее, выбирая по очереди следующие теоретические распределения с помощью поля с

выпадающим списком *Распределение* в диалоговом окне *Подгонка к непрерывным распределениям* (рис. 8.9), выполним расчеты: **критерия Пирсона, числа степеней свободы и вероятности совершения ошибки 1-го рода** по очереди для каждого из распределений: равномерное, экспоненциальное, гамма, лог-нормальное, хи-квадрат. Полученные результаты сведем в табл. 8.2 в соответствующие строки.

9. Выполним анализ полученных результатов, который в данном случае показывает, что в отношении трех распределений (гамма, лог-нормальное, хи-квадрат) гипотезу можно принять, так как вероятность совершения ошибки 1-го рода для них выше уровня значимости ($\alpha = 0,05$). Гипотезы, выдвинутые в отношении других распределений, следует отвергнуть. Из трех распределений, которые хорошо подходят к экспериментальным данным, лучшим следует считать лог-нормальное распределение, так как вероятность совершения ошибки 1-го рода для него максимальная.

10. Аналогичным образом выполним расчеты для вариационного ряда высот. Для этого повторно выполним действия 2–9 задания 2, но введем требуемые данные для ряда высот во кладке *Параметры* для *нормального* распределения. Полученные результаты соберем в табл. 8.3.

Таблица 8.3

Анализ распределения высот деревьев в древостое

Наименование распределения	Критерий Пирсона χ^2 (<i>Хи-квадрат</i>)	Число степеней свободы (<i>сс</i>)	Вероятность совершения ошибки 1-го рода (<i>p</i>)	Оценка
Нормальное	13,44788	7	0,08676	+
Равномерное	214,67700	10	0,00000	–
Экспоненциальное	1478,03542	5	0,00000	–
Гамма	13,89280	7	0,05312	+
Лог-нормальное	15,03286	7	0,03558	–
Хи-квадрат	301,57619	11	0,00000	–

11. Выполним анализ полученных результатов, который в данном случае показывает, что гипотезы в отношении всех распределений, кроме нормального и гамма, следует отвергнуть, так как для них вероятность сделать при этом ошибку ниже, чем уровень значимости $\alpha = 0,05$. Лучшим распределением следует счи-

тать нормальное, так как вероятность совершения ошибки 1-го рода для него максимальная.

12. Для лучших распределений построим графики сравнения эмпирического и теоретического распределений. Для этого в диалоговом окне *Подгонка к непрерывным распределениям* (рис. 8.9) выберем лучшее распределение и на вкладке *Быстрый* нажмем кнопку *График наблюдаемого и ожидаемого распред.*, на экране появится график, на котором будут сравниваться эмпирическое (изображено в виде гистограммы) и теоретическое распределения (рис. 8.12).

На этом же графике (рис. 8.12) приведена информация для проверки гипотезы о согласованности эмпирического и теоретического распределений с помощью критерия Пирсона.

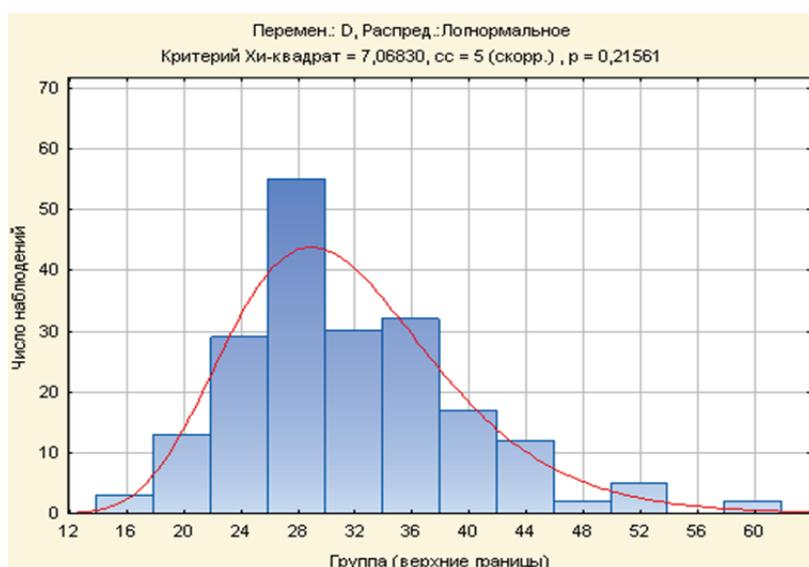


Рис. 8.12. Сравнение эмпирического и теоретического (логнормального) распределений для ряда диаметров в программе Statistica

В рассматриваемых примерах с помощью программы Statistica могут быть получены несколько иные значения, чем ранее, когда подробно рассматривался процесс вычисления теоретических частот и проверки гипотезы о нормальном распределении диаметров и высот (см. лабораторную работу № 7). Это вызвано тем, что в программе Statistica параметры нормального распределения оцениваются на основании исходных данных без выполнения группировки, а в рассматриваемом ранее примере вычисления выполнялись по данным интервальных вариационных рядов, что всегда связано с потерей точности. Кроме того, при вы-

числениях функции нормального распределения программа Statistica использует аппроксимирующие алгоритмы, а в первом варианте расчета применялись табличные значения функции нормального распределения. Принимая во внимание сказанное, целесообразно выбрать результаты, полученные с помощью программы Statistica, как более достоверные при наличии отличий в результатах расчета.

В MS Excel можно выполнить построение встроенных функций распределений, например нормального – НОРМ.РАСП (X; среднее; стандартное отклонение; интегральная); экспоненциального – ЭКСП.РАСП (X; лямбда; интегральная); гамма – ГАММА.РАСП (x; альфа; бета; интегральная); бета-распределения – БЕТА.РАСП (x; альфа; бета; интегральная; [A]; [B]); лог-нормального ЛОГНОРМ.РАСП (x; среднее; стандартное отклонение; интегральная); распределения Пуассона – ПУАССОН (x; среднее; интегральная); биномиального – БИНОМРАСП (число успехов; число испытаний; вероятность успеха; интегральная); распределения Стьюдента – СТЬЮДРАСП (x; степени свободы; хвосты); распределения Пирсона (хи-квадрат) ХИ2РАСП (x; степени свободы); распределения Фишера – ФРАСП (x; степени свободы₁; степени свободы₂) и др.

Для проверки согласованности теоретического и эмпирического распределений в MS Excel можно использовать тест критерия согласия (хи-квадрат) Пирсона ХИ2.ТЕСТ (фактический интервал; ожидаемый интервал).



КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Цель лабораторной работы: выполнить корреляционный анализ; рассчитать коэффициент корреляции, корреляционное отношение и их стандартные ошибки, степень криволинейности.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

В предыдущих лабораторных работах высоты и диаметры анализировались по отдельности, однако в природе многие случайные величины в той или иной степени связаны друг с другом и в большинстве исследований требуется выяснить наличие связи между ними. Зависимости и связи в природе, имеющие общие методы их статистического измерения, называют корреляцией, связью или зависимостью.

В природе могут быть обнаружены связи 2 типов: функциональные и корреляционные. *Функциональная* связь – это такая связь между признаками (показателями), при которой с изменением одного признака (показателя) на определенную величину, другой признак (показатель) изменяется тоже на определенную величину.

Корреляционная связь – это такая связь, при которой изменение одного признака у ряда особей на определенную величину сопровождается изменениями другого признака на различные (варьирующие) значения.

По математическим особенностям корреляционная связь различных признаков может быть: *прямой* (положительной) и *обратной* (отрицательной); *прямолинейной* и *криволинейной*; *простой* и *множественной*; между *количественными* признаками и *качественными*.



и к оценке достоверности выборочных показателей корреляции. Для этого производится разноска полученных численных значений в одну общую двумерную таблицу, называемую *корреляционной решеткой* (см. лабораторную работу № 2). Коэффициенты связи не объясняют причинности связи между изучаемыми показателями. Причинность взаимосвязи между различными признаками и факторами может быть определена только физиологическими, биохимическими и другими методами. С помощью коэффициентов связи можно определить величину связи, направление и ее тип.

Для того чтобы оценить тесноту связи между случайными величинами, в случае линейных зависимостей (или типа, близкого к прямолинейному) удобно использовать коэффициент корреляции:

$$R = \frac{\sum_{i=1}^k f_i \cdot (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot S_x \cdot S_y}. \quad (9.1)$$

По результатам расчетов делается вывод о характере связи. Коэффициент корреляции R может принимать значения от -1 до $+1$. Чем ближе значение R к 1 , тем сильнее связь между данными признаками, при $R = 1,0$ связь из корреляционной переходит в функциональную; $0,91-0,999$ – связь очень сильная; $0,71-0,90$ – сильная; $0,51-0,70$ – средняя; $0,31-0,50$ – слабая; $0,1-0,30$ – очень слабая; $R = 0$ указывает на отсутствие прямолинейной связи (но при этом может быть криволинейная связь).

Коэффициент корреляции, взятый в квадрат (R^2), называется коэффициентом *детерминации* и показывает долю изменчивости признака, обусловленную влиянием изменчивости воздействующего фактора.

Для оценки криволинейных связей коэффициентом корреляции R пользоваться не рекомендуется, так как он не дает точной ее оценки или вообще не может ее определить. Для нелинейных зависимостей лучше использовать показатель, предложенный Пирсоном, который называется корреляционным отношением (η):

$$\eta = \sqrt{\frac{\sum_{i=1}^k f_{x,i} \cdot (\bar{y}_{x,i} - \bar{y})^2}{n \cdot S_y^2}}. \quad (9.2)$$

Величина корреляционного отношения всегда положительна, изменяется от 0 до 1. Когда групповые средние одинаковы (не варьируют), то $\eta = 0$ или связь отсутствует. В случае строго прямолинейной связи (все точки лежат на одной линии) $\eta = R = 1$, в остальных случаях $\eta > R$.

По соотношению величины коэффициента корреляции и корреляционного отношения можно сделать вывод о характере связи: прямолинейна она или криволинейна. Чем значительнее корреляционное отношение превышает коэффициент корреляции, тем более криволинейной является эта связь. Для оценки степени криволинейности связи вычисляют меру криволинейности как разницу между квадратами корреляционного отношения и коэффициента корреляции:

$$K = \eta^2 - R^2. \quad (9.3)$$

Стандартные ошибки коэффициента корреляции и корреляционного отношения можно оценить с помощью следующих выражений:

$$S_R = \frac{\sqrt{1 - R^2}}{\sqrt{n - 2}} \quad (9.4)$$

и

$$S_\eta = \frac{\sqrt{1 - \eta^2}}{\sqrt{n - 2}}. \quad (9.5)$$

Задание. Рассчитать коэффициент корреляции, корреляционное отношение, стандартные ошибки коэффициента корреляции и корреляционного отношения, степени криволинейности, определить тип связи.

Порядок выполнения работы

Задание. Вычислим показатели связи для пары случайных величин – диаметров и высот деревьев в древостое. Для того чтобы выполнить вычисления, составим вспомогательную таблицу.

Для определения групповых средних (\bar{H}_D) нужно частоты в ячейках по строчкам таблицы умножить на соответствующие средние значения классов, затем сложить их и всю сумму разделить на f_D того же столбца.

Вспомогательная таблица для вычисления коэффициента корреляции и корреляционного отношения

$\frac{H}{D}$	16,0	20,0	24,0	28,0	32,0	36,0	40,0	44,0	48,0	52,0	56,0	60,0	f_H	$H_i - \bar{H}$
30,5								1					1	5,66
29,5				1									1	4,66
28,5						4	1	6		1		1	13	3,66
27,5		1		3	1	4	4	3	1	3		1	21	2,66
26,5					4	5	6	1	1	1			18	1,66
25,5			2	11	10	12	4	1					40	0,66
24,5			6	20	12	4	1						43	-0,34
23,5			7	9	2	3	1						22	-1,34
22,5	1	5	7	7	1								21	-2,34
21,5		4	6	3									13	-3,34
20,5		3	1										4	-4,34
19,5	1			1									2	-5,34
18,5	1												1	-6,34
f_D	3	13	29	55	30	32	17	12	2	5	0	2	200	
\bar{H}_D	20,17	22,12	23,09	24,28	25,07	25,97	26,32	28	27	27,50	-	28	-	
$\bar{H}_D - \bar{H}$	-4,67	-2,72	-1,75	-0,56	0,23	1,13	1,48	3,16	2,16	2,66	-	3,16	-	
$f_D(\bar{H}_D - \bar{H})^2$	65,43	96,18	88,81	17,25	1,59	40,86	37,24	119,83	9,33	35,38	-	19,97	531,87	-
$D_i - \bar{D}$	-15,70	-11,70	-7,70	-3,70	0,30	4,30	8,30	12,30	16,30	20,30	-	28,30	-	
$f_{Dit} \cdot (H_i - \bar{H}) \times (D_i - \bar{D})$	220,11	414,41	391,62	113,59	2,04	155,32	209,33	466,42	70,42	269,99	-	178,86	2492,11	

Например, для столбца 16,0:

$$(1 \times 18,5 + 1 \times 19,5 + 1 \times 22,5) / 3 = 20,17 \text{ м.}$$

Чтобы рассчитать $f_{DH} \cdot (H_i - \bar{H}) \cdot (D_i - \bar{D})$ для столбца 16,0:

$$(1 \times (-6,34) + 1 \times (-5,34) + 1 \times (-2,34)) \times (-15,70) = 220,11.$$

Подставляя значения сумм из данной таблицы в формулы (9.1) и (9.2), получим:

$$R = \frac{\sum_{i=1}^k f_i \cdot (D_i - \bar{D}) \cdot (H_i - \bar{H})}{n \cdot S_D \cdot S_H} = \frac{2492,11}{200 \cdot 7,989 \cdot 2,145} = 0,727;$$

$$\eta^2 = \frac{\sum_{i=1}^k f_{D,i} \cdot (\bar{H}_{d,i} - \bar{H})^2}{n \cdot S_H^2} = \frac{531,87}{200 \cdot 5,164} = 0,578$$

или

$$\eta = \sqrt{0,578} = 0,760.$$

Пользуясь выражениями (9.4) и (9.5), вычислим стандартные ошибки коэффициента корреляции и корреляционного отношения:

$$S_R = \frac{\sqrt{1 - R^2}}{\sqrt{n - 2}} = \frac{\sqrt{1 - 0,727^2}}{\sqrt{200 - 2}} = \frac{0,6866}{14,07} = 0,049;$$

$$S_\eta = \frac{\sqrt{1 - \eta^2}}{\sqrt{n - 2}} = \frac{\sqrt{1 - 0,760^2}}{\sqrt{200 - 2}} = \frac{0,6499}{14,07} = 0,046.$$

Полученные результаты говорят о том, что между диаметрами и высотами деревьев в древостое существует сильная корреляционная связь ($R = 0,727$), а тот факт, что корреляционное отношение значительно превышает коэффициент корреляции, показывает нам, что эта зависимость скорее криволинейная, чем прямолинейная. Вычислим меру криволинейности по формуле (9.3):

$$K = \eta^2 - R^2 = 0,760^2 - 0,727^2 = 0,5776 - 0,5285 = 0,049.$$



РЕГРЕССИОННЫЙ АНАЛИЗ

Цель лабораторной работы: выполнить регрессионный анализ; рассчитать коэффициенты регрессий и определить стандартные ошибки регрессий; построить графики зависимости высоты от диаметра.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 4 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

С помощью корреляционного метода можно определить степень тесноты связи между признаками, но выяснить, как *количественно* меняется результативный признак при изменении другого, нельзя.

В этом случае следует воспользоваться регрессионным анализом, который, в отличие от корреляционного, изучает *эффект влияния одного признака на другой*, зависимость признака от фактора, характер влияния фактора на признак. Термин «регрессия» ввел Ф. Гальтон. Смысл его заключался в том, что коррелирующие пары в биологических объектах, обнаруживающие отклонения от средней линии, определяющей корреляцию признаков совокупности, имеют тенденцию возврата к этой средней, если только действуют одни случайные причины.

Регрессионный анализ возможен при наличии всего лишь нескольких пар сопряженных наблюдений, но при условии сильных связей между признаками. Подобно корреляции, регрессия может быть парной (простой) и множественной, по форме связи – линейной и нелинейной, по зависимости – односторонней (изменяется лишь один признак под влиянием другого) и двусторонней (изменяются оба признака под воздействием друг друга).



уравнения регрессии заданного вида таким образом, чтобы сумма квадратов отклонений эмпирических значений зависимой переменной от теоретических значений (рис. 10.2) была наименьшей

$$\sum_{i=1}^n (y_i - \tilde{y}_i)^2 \Rightarrow 0.$$

Метод наименьших квадратов применяется чаще всего для решения различных задач, связанных с обработкой результатов опыта. Наиболее важным приложением этого метода является решение задачи сглаживания экспериментальной зависимости, т. е. изображения опытной функциональной зависимости аналитической формулой. При этом метод наименьших квадратов не решает вопрос о выборе общего вида аналитической функции, а дает возможность при заданном типе аналитической функции $y = f(x)$ подобрать наиболее вероятные значения для параметров этой функции. Для того чтобы получить оценку коэффициентов уравнения методом наименьших квадратов, следует решить систему нормальных уравнений.

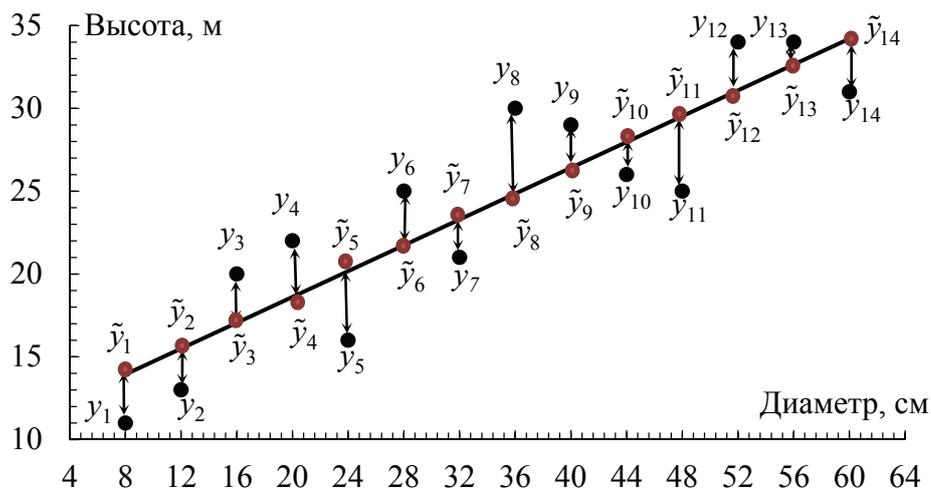


Рис. 10.2. Сущность метода наименьших квадратов (на примере графика высот)

Линии регрессии дают наглядное представление о форме и тесноте корреляционной связи между признаками. Но форма связи у биологических признаков разнообразна, поэтому задача анализа состоит в том, чтобы любую форму такой связи выразить уравнением определенной функции (линейной, параболической, гиперболической и др.).

Термином линейный регрессионный анализ обозначают такое прогнозирование, которое описывается линейной взаимосвязью между исследуемыми переменными: $y = b_0 + b_1x$.

В случае криволинейных зависимостей применяются математические функции следующего вида: гиперболическая $y = b_0 + b_1 / x$; показательная $y = b_0 + b_1^x$; степенная $y = b_0x^{b_1}$; параболическая $y = b_0 + b_1x + b_2x^2$; полиномиальная 3-й степени $y = b_0 + b_1x + b_2x^2 + b_3x^3$; логарифмическая $y = b_0 + b_1 \lg x$; экспоненциальная $y = b_0 \exp(b_1x)$ и др.

Решение математических уравнений связи предполагает вычисление по исходным данным их параметров (свободного члена b_0 и коэффициентов регрессии b_1, b_2, \dots).

Стандартная ошибка регрессионного уравнения (т. е. ошибка, с которой любое измеренное значение (y) предсказывается для данного значения (x) по регрессионному уравнению) для сгруппированных данных может быть вычислена по формуле

$$s = \sqrt{\frac{\sum f_{i,j} \cdot (y_j - \tilde{y}_i)^2}{n - m - 1}}. \quad (10.1)$$

Чем меньше данная стандартная ошибка, тем более точно регрессия описывает опытные данные.

Задание 1. Получить регрессионную модель линейной зависимости высоты от диаметра деревьев в сосновом древостое методом наименьших квадратов, рассчитать стандартную ошибку регрессии уравнения и отобразить графически зависимость между диаметрами и высотами.

Задание 2. Получить регрессионное уравнение параболы второго порядка, описывающее зависимость высоты от диаметра в чистом сосновом древостое методом наименьших квадратов, рассчитать стандартную ошибку регрессии и отобразить графически зависимость между диаметрами и высотами.

Порядок выполнения работы

Задание 1. Линейная взаимосвязь между исследуемыми переменными выражается уравнением общего вида:

$$y = b_0 + b_1x. \quad (10.2)$$

Для построения регрессионной модели данной линейной зависимости высоты от диаметра деревьев в сосновом древостое нужно выполнить оценку его коэффициентов b_0 и b_1 уравнения прямой линии методом наименьших квадратов, для чего решим систему нормальных уравнений:

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum_{i=1}^k f_i \cdot D_i = \sum_{i=1}^k f_i \cdot H_i, \\ b_0 \cdot \sum_{i=1}^k f_i \cdot D_i + b_1 \cdot \sum_{i=1}^k f_i \cdot D_i^2 = \sum_{i=1}^k f_i \cdot H_i \cdot D_i. \end{cases} \quad (10.3)$$

Рассмотрим процесс вычисления коэффициентов уравнения прямой, моделирующей зависимость между высотами и диаметрами. Для этого на основе корреляционной решетки (таблица, лабораторная работа № 2) составим вспомогательную таблицу для вычисления всех необходимых для расчетов сумм (таблица). В данной таблице суммы вычисляются сначала по интервалам, а затем складываются.

Чтобы рассчитать строку $\sum f_{i,j} \cdot H_j \cdot D_i$, надо для столбца 16,0: $(1 \times 18,5 + 1 \times 19,5 + 1 \times 22,5) \times 16,0 = 968,0$; для столбца 20,0: $(3 \times 20,5 + 4 \times 21,5 + 5 \times 22,5 + 1 \times 27,5) \times 20,0 = 5750,0$.

Подставив значения сумм в систему нормальных уравнений (10.3), получим

$$\begin{cases} b_0 \cdot 200 + b_1 \cdot 6340,0 = 4967,0, \\ b_0 \cdot 6340,0 + b_1 \cdot 213\,744,0 = 159\,946,0. \end{cases} \quad (10.4)$$

Решим полученную систему уравнений. Для этого разделим каждое из уравнений системы (10.3) на коэффициенты при параметре b_0 :

$$\begin{cases} b_0 + b_1 \cdot 31,70 = 24,835, \\ b_0 + b_1 \cdot 33,714 = 25,228. \end{cases} \quad (10.5)$$

Теперь вычтем первое уравнение системы (10.5) из второго

$$b_1 \cdot 2,014 = 0,393 \quad (10.6)$$

и выразим из полученного уравнения (10.6) коэффициент b_1 :

$$b_1 = \frac{0,393}{2,014} = 0,195. \quad (10.7)$$

Вспомогательная таблица для вычисления коэффициентов регрессий

H	D	16,0	20,0	24,0	28,0	32,0	36,0	40,0	44,0	48,0	52,0	56,0	60,0	f_{H_i}	$\sum f_i \cdot H_i$
$30,5$									1					2	30,5
$29,5$					1				6				1	9	29,5
$28,5$							4	1	6		1			17	370,5
$27,5$			1		3	1	4	4	3	1	3		1	29	577,5
$26,5$						4	5	6	1	1	1			44	477,0
$25,5$				2	11	10	12	4	1					36	1020,0
$24,5$				6	20	12	4	1						21	1053,5
$23,5$				7	9	2	3	1						22	517,0
$22,5$			5	7	7	1								7	472,5
$21,5$			4	6	3									6	279,5
$20,5$			3	1										2	82,0
$19,5$					1									2	39,0
$18,5$														3	18,5
f_D		3	13	29	55	30	32	17	12	2	5	0	2	200	4 967,0
$\sum f_i \cdot D_i$		48,0	260,0	696,0	1 540,0	960,0	1 152,0	680,0	528,0	96,0	260,0	-	120,0		6 340,0
$\sum f_i \cdot D_i^2$		768,0	5 200,0	16 704,0	43 120,0	30 720,0	41 472,0	27 200,0	23 232,0	4 608,0	13 520,0	-	7 200,0		213 744,0
$\sum f_i \cdot D_i^3$		122 880,0	104 000,0	400 896,0	1 207 360	983 040,0	1 492 992	1 088 000,0	1 022 208,0	221 184,0	703 040,0	-	43 200,0		7 667 008,0
$\sum f_i \cdot D_i^4$		196 608,0	2 080 000,0	9 621 504	33 806 080	31 457 280	53 747 712	43 520 000	44 977 152	10 616 832	36 558 080	-	25 920 000		292 501 248,0
$\sum f_i \cdot H_i \cdot D_i$		968,0	5750,0	16068,0	37394,0	24064,0	29916,0	17900,0	14 784,0	2 592,0	7 150,0	-	3 360,0		159 946,0
$\sum f_i \cdot H_i \cdot D_i^2$		15 488,0	115 000,0	385 632,0	1 047 032	770 048,0	1 076 976	716 000,0	650 496,0	124 416,0	371 800,0	-	201 600,0		5 474 488,0
\tilde{H}_i		21,8	22,6	23,3	24,1	24,9	25,7	26,5	27,2	28,0	28,8	29,6	30,4		-
$\sum f_{i,j} \cdot (H_j - \tilde{H}_i)^2$		16,67	42,13	50,36	151,20	32,20	68,28	25,00	24,68	2,50	10,45	-	12,02		435,5
\tilde{H}_i		20,6	21,9	23,1	24,2	25,2	26,0	26,7	27,2	27,6	27,9	28,0	28,0		-
$\sum f_{i,j} \cdot (H_j - \tilde{H}_i)^2$		9,23	39,68	49,04	149,75	31,90	66,00	26,88	24,68	1,22	2,80	-	0,50		401,7

Уравнение прямой

Уравнение параболы

Подставляя вычисленное значение коэффициента b_1 в первое уравнение системы (10.5) и, выразив из него коэффициент b_0 , получим

$$b_0 = 24,835 - b_1 \cdot 31,70 = 24,835 - 0,195 \cdot 31,70 = 18,65. \quad (10.8)$$

Таким образом, у нас получилась регрессионная модель зависимости высоты от диаметра деревьев в сосновом древостое следующего вида:

$$\tilde{y} = 18,65 + 0,195 \cdot x, \quad (10.9)$$

или, используя другие обозначения:

$$\tilde{H} = 18,65 + 0,195 \cdot D. \quad (10.10)$$

Затем вычисляется по столбцам: для столбца 16,0: $18,65 + 0,195 \times 16,0 = \mathbf{21,8}$ м; для столбца 20,0: $18,65 + 0,195 \times 20,0 = \mathbf{22,6}$ м.

Пользуясь полученным регрессионным уравнением прямой линии, определим теоретические высоты \tilde{H}_i и сумму квадратов отклонений эмпирических высот от теоретических $\sum f_{i,j} \cdot (H_j - \tilde{H}_i)^2$ (таблица).

Для этого надо для столбца 16,0: $((18,5 - 21,8)^2 \times 1) + ((19,5 - 21,8)^2 \times 1) + ((22,5 - 21,8)^2 \times 1) = \mathbf{16,67}$ м; для столбца 20,0: $((22,5 - 22,6)^2 \times 3) + ((21,5 - 22,6)^2 \times 4) + ((22,5 - 22,6)^2 \times 5) + ((27,5 - 22,6)^2 \times 1) = \mathbf{42,13}$ м.

Полученное значение суммы квадратов отклонений 435,5 мы можем использовать для вычисления стандартной ошибки регрессионного уравнения прямой:

$$m_{1D} = \sqrt{\frac{\sum f_{i,j} \cdot (H_j - \tilde{H}_i)^2}{n - 2}} = \sqrt{\frac{435,5}{200 - 2}} = 1,48. \quad (10.11)$$

На рис. 10.3 изображено полученное регрессионное уравнение прямой линии.

Задание 2. Взаимосвязь между исследуемыми переменными в виде параболы второго порядка выражается уравнением общего вида:

$$y = b_0 + b_1x + b_2x^2. \quad (10.12)$$

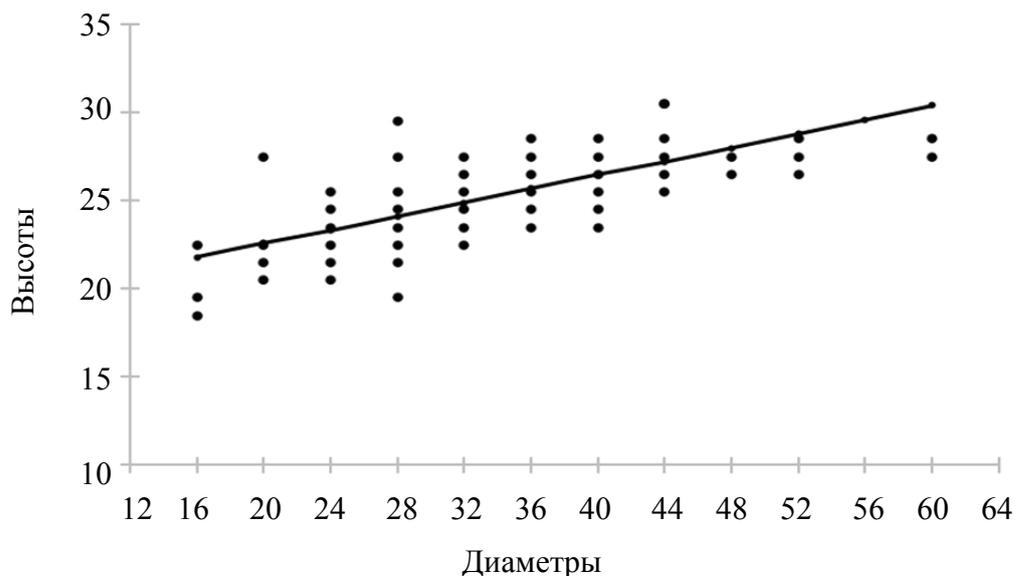


Рис. 10.3. Зависимость между высотами и диаметрами деревьев в древостое (прямая)

Для построения данной регрессионной модели зависимости высоты от диаметра деревьев выполним оценку коэффициентов параболы второго порядка b_0 , b_1 и b_2 методом наименьших квадратов, для чего решаем систему нормальных уравнений:

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum_{i=1}^k f_i \cdot D_i + b_2 \cdot \sum_{i=1}^k f_i \cdot D_i^2 = \sum_{i=1}^k f_i \cdot H_i, \\ b_0 \cdot \sum_{i=1}^k f_i \cdot D_i + b_1 \cdot \sum_{i=1}^k f_i \cdot D_i^2 + b_2 \cdot \sum_{i=1}^k f_i \cdot D_i^3 = \sum_{i=1}^k f_i \cdot H_i \cdot D_i, \\ b_0 \cdot \sum_{i=1}^k f_i \cdot D_i^2 + b_1 \cdot \sum_{i=1}^k f_i \cdot D_i^3 + b_2 \cdot \sum_{i=1}^k f_i \cdot D_i^4 = \sum_{i=1}^k f_i \cdot H_i \cdot D_i^2. \end{cases} \quad (10.13)$$

Вычислим коэффициенты уравнения параболы второго порядка, описывающей связь высот и диаметров деревьев в древостое. Для выполнения вычислений используем значения вспомогательной таблицы. Подставив найденные значения в систему нормальных уравнений (10.13), получим

$$\begin{cases} b_0 \cdot 200 + b_1 \cdot 6\,340,0 + b_2 \cdot 213\,744,0 = 4\,967,0, \\ b_0 \cdot 6\,340,0 + b_1 \cdot 213\,744,0 + b_2 \cdot 7\,667\,008,0 = 159\,946,0, \\ b_0 \cdot 213\,744,0 + b_1 \cdot 7\,667\,008,0 + \\ + b_2 \cdot 292\,501\,248,0 = 5\,474\,488,0. \end{cases} \quad (10.14)$$

$$\tilde{H} = 13,855 + 0,4880 \cdot D - 0,0042 \cdot D^2. \quad (10.23)$$

С помощью полученного уравнения регрессии определим теоретические высоты \tilde{H}_i (подставив в уравнение (10.23) значения D_i из таблицы), а также сумму квадратов отклонений (таблица) эмпирических высот от теоретических (аналогично вычислениям с уравнением прямой, только берутся значения теоретических высот для уравнения параболы второго порядка).

Используя сумму квадратов отклонений 401,7, вычислим стандартную ошибку регрессионного уравнения параболы второго порядка:

$$m_{2D} = \sqrt{\frac{\sum f_{i,j} \cdot (H_j - \tilde{H}_i)^2}{n-3}} = \sqrt{\frac{401,7}{200-3}} = 1,43. \quad (10.24)$$

На рис. 10.4 изображено полученное регрессионное уравнение параболы второго порядка.

Стандартная ошибка регрессии параболы второго порядка (1,43 м) меньше, чем стандартная ошибка регрессионного уравнения прямой (1,48 м). Это говорит о том, что парабола в нашем случае более точно описывает опытные данные, чем прямая.

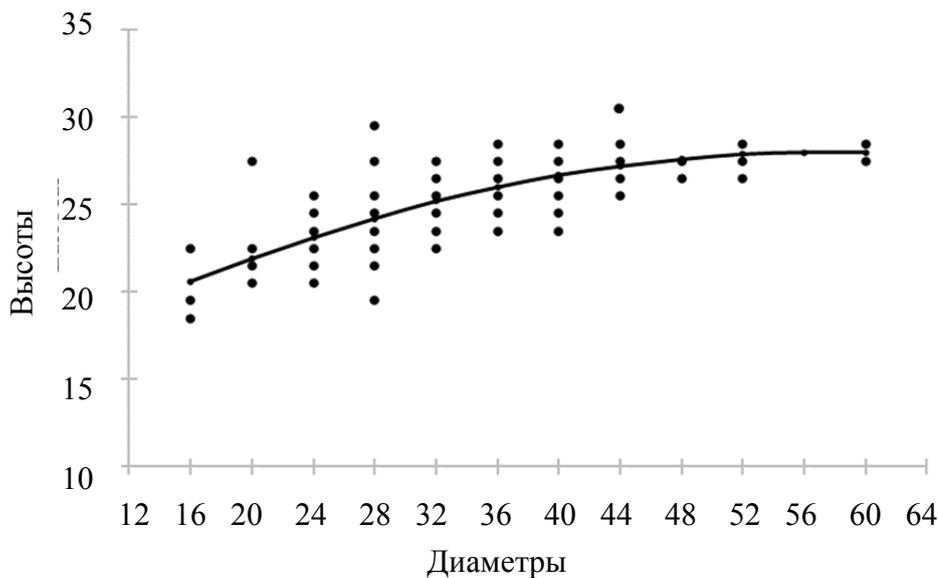


Рис. 10.4. Зависимость между высотами и диаметрами деревьев в древостое (парабола второго порядка)

Наличие модели, позволяющей оценивать значения высот деревьев в древостое исходя из их диаметра, может оказать большую практическую пользу, так как трудоемкость измерения высоты растущего дерева значительно выше, чем трудоемкость измерения его диаметра.



РЕГРЕССИОННЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ПАКЕТА ПРОГРАММ

Цель лабораторной работы: научиться выполнять регрессионный анализ зависимости диаметров и высот в чистом сосновом древостое с использованием пакета программ Statistica или MS Excel.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office и Statistica 10.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Регрессионный анализ является одним из самых востребованных методов статистического исследования. С его помощью можно установить степень влияния независимых величин на зависимую переменную. Регрессия позволяет прогнозировать зависимую переменную на основании значений признака. В функционале Microsoft Excel имеются инструменты, предназначенные для проведения подобного вида анализа, множество функций, которые оценивают не только наклон и сдвиг линии регрессии, характеризующей линейную взаимосвязь между факторами, но и регрессионную статистику. MS Excel позволяет осуществить данную процедуру с помощью функции *Анализ данных – Регрессия*. Результаты выполнения функции аналогичны результатам программы Statistica, но пользователь может выбрать место их размещения: на используемом или на новом листе рабочей книги.

В программе Statistica оценка коэффициентов однофакторной и многофакторной линейной регрессии осуществляется в отдельном модуле *Множественная регрессия (Multiple regression)*. Результаты просматриваются в отдельном окне, в котором есть коэффициенты, оцененные методом наименьших квадратов, коэффициент детерминации, статистика Фишера оценки значимости регрессии, статистики Стьюдента оценки значимости коэффициентов,



3. Для выполнения расчетов в файле исходных данных создадим **четыре новые переменные** (D^2 , D^3 , $1/D$, $\ln(D)$). Для этого в меню *Вставка* выберем *Добавить переменные...*. В открывшемся диалоговом окне *Добавить переменные в таблицу* (рис. 11.1) в поле *Число* внесем число переменных – **4**, которые следует добавить, и нажмем кнопку *ОК*. Далее дадим название добавленным переменным и вычислим их значения. Для этого выполним двойной щелчок левой кнопкой мыши на заголовке новой переменной (например, на *НовПер1*, которая откроет диалоговое окно свойств переменных (рис. 11.2).

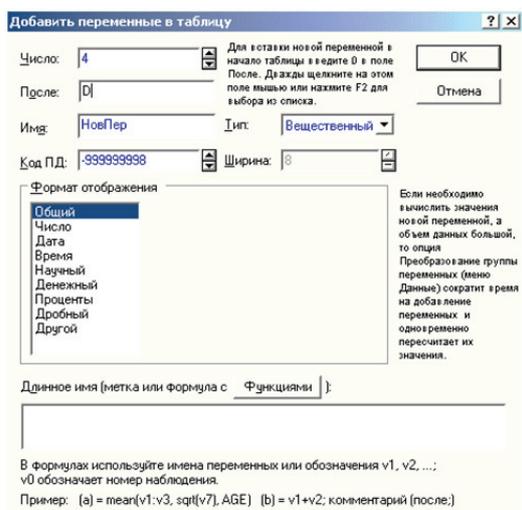


Рис. 11.1. Диалоговое окно добавления переменных в таблицу

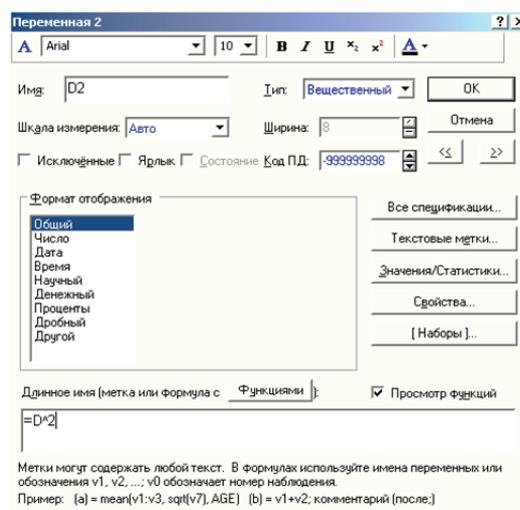


Рис. 11.2. Диалоговое окно изменения свойств переменных

4. В окне (рис. 11.2) изменим имя в поле *Имя* (например, для D^2 ввести D^2), а в поле *Длинное имя*, расположенном в нижней части диалогового окна, введем формулу для вычисления значений переменной (например, возвести диаметр в квадрат, введя $=D^2$). После нажатия кнопки *ОК* все изменения будут выполнены, а если вводилась формула, то после подтверждения (нажать кнопку *Да*) будут вычислены и значения переменной. Повторим данные действия по созданию и расчету для остальных новых переменных (D^2 , D^3 , $1/D$, $\ln(D)$), введя соответственно формулы D^2 , D^3 , $1/D$, $\text{LOG}(D)$.

5. В меню *Анализ* (рис. 8.7) выберем *Множественная регрессия*, при этом откроется диалоговое окно *Множественная регрессия* (рис. 11.3).

6. В открывшемся диалоговом окне нажмем кнопку *Переменные* (рис. 11.1). После этого в диалоговом окне *Списки зависимых*

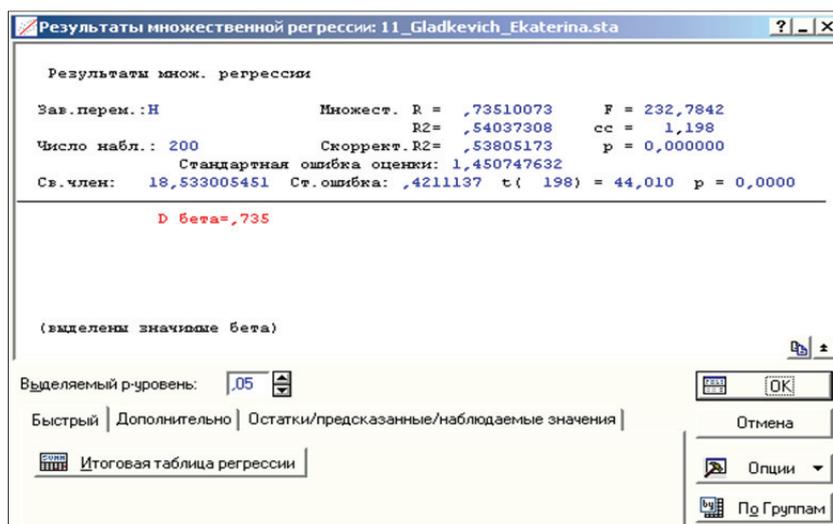


Рис. 11.5. Окно результатов регрессионного анализа

8. После заполнения строки для уравнения прямой нажмем кнопку *Отмена* и в появившемся окне (рис. 11.4) укажем новые независимые переменные для уравнения параболы (D , $D2$, зажав *Ctrl*), зависимая переменная (H) остается, запускаем расчеты кнопками *OK* (рис. 11.4) и (рис. 11.3) и заносим полученные данные (рис. 11.5) в табл. 11.1 в строку уравнения параболы.

9. Нажмем кнопку *Отмена* и по очереди выполним аналогичные действия (№ 6–8) для оставшихся регрессионных моделей – кубического уравнения (D , $D2$, $D3$), гиперболы ($1/D$), логарифмического уравнения ($\text{LOG}(D)$). Такие операции продолжают-ся до тех пор, пока все графы таблицы не будут заполнены.

Статистические показатели, характеризующие уравнения регрессии

Уравнение	Коэффициент корреляции (Множеств. R)	Коэффициент детерминации (R^2)	Критерий Фишера (F)	Стандартная ошибка оценки	Вывод
$H = b_0 + b_1 \cdot D$	0,735101	0,540373	232,784	1,450748	Дост.
$H = b_0 + b_1 \cdot D + b_2 \cdot D^2$	0,763811	0,583408	137,942	1,384663	Дост./ +
$H = b_0 + b_1 \cdot D + b_2 \cdot D^2 + b_3 \cdot D^3$	0,763806	0,583400	91,520	1,388077	Не дост.
$H = b_0 + b_1 / D$	0,757999	0,574564	267,405	1,395745	Дост.
$H = b_0 + b_1 \ln(D)$	0,758844	0,575845	268,810	1,393643	Дост.

бранного уровня значимости и соответствующего числа степеней свободы (указано в скобках в заголовке колонки после символа t), гипотеза отвергается. Колонка p -знач. содержит вероятность того, что гипотеза о равенстве коэффициента регрессии нулю верна.

Итоги регрессии для зависимой переменной: H (11_Gladkevich_Ekaterina.sta)						
R= ,76381134 R2= ,58340776 Скоррект. R2= ,57917840						
F(2,197)=137,94 p<0,0000 Станд. ошибка оценки: 1,3847						
N=200	БЕТА	Ст.Ош. БЕТА	В	Ст.Ош. В	t(197)	p-знач.
Св.член			13,12559	1,264272	10,38193	0,000000
D	1,96062	0,275530	0,52498	0,073776	7,11582	0,000000
D2	-1,24296	0,275530	-0,00468	0,001036	-4,51115	0,000011

Рис. 11.6. Таблица с параметрами регрессионного уравнения

Для параболы второго порядка (рис. 11.6) получились высокие значения t -критерия Стьюдента и значения вероятностей (p -знач.), которые меньше уровня значимости $\alpha = 0,05$. Они свидетельствуют о том, что параметрическую гипотезу о равенстве коэффициентов регрессии нулю следует отвергнуть.

Это значит, что все независимые переменные вносят существенный вклад в уравнение регрессии. Полученное нами уравнение параболы второго порядка выглядит следующим образом:

$$H = 13,12\ 559 + 0,52\ 498 \cdot D - 0,00\ 468 \cdot D^2.$$

Значения t -критерия Стьюдента для регрессионных коэффициентов параболы третьего порядка получились следующие:

$$t_{b_0} = 10,38\ 193; t_{b_1} = 7,11\ 582; t_{b_2} = -4,51\ 115.$$

Также обращает на себя внимание логарифмическое уравнение, для которого характерен самый большой критерий Фишера и достаточно большие значения коэффициента корреляции и коэффициентов детерминации, а также не слишком высокая стандартная ошибка оценки. Самый большой критерий Фишера говорит нам о том, что по соотношению точность – сложность логарифмическое уравнение является самым удачным уравнением, т. е. оно с довольно высокой точностью описывает зависимость высот от диаметров, являясь значительно более простым уравнением, чем парабола второго порядка (в логарифмическом уравнении используется только одна независимая переменная – $\ln(D)$), а в параболе

второго порядка – две). Логарифмическое уравнение следует предпочесть в том случае, если простота уравнения для исследователя является главным фактором подбора.

В пакете Statistica расчет нелинейной регрессии можно также выполнить при помощи модуля *Нелинейное оценивание*, который запускает из меню *Анализ*, подменю *Углубленные методы анализа* команда *Нелинейное оценивание* (рис. 11.7).

Выберем *Регрессия пользователя – метод наим. квадратов МНК*, (метод наименьших квадратов), в появившемся окне *Функция пользователя, оценки МНК* выберем и нажмем кнопку *Оцениваемая функция*, которая откроет одноименное диалоговое окно (рис. 11.8). Здесь наберем требуемую функцию в виде условных обозначений (рис. 11.8).

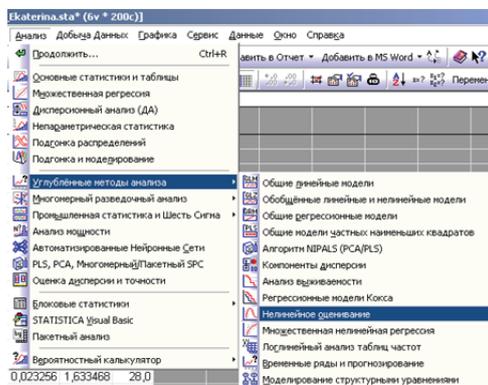


Рис. 11.7. Меню выбора нелинейной регрессии

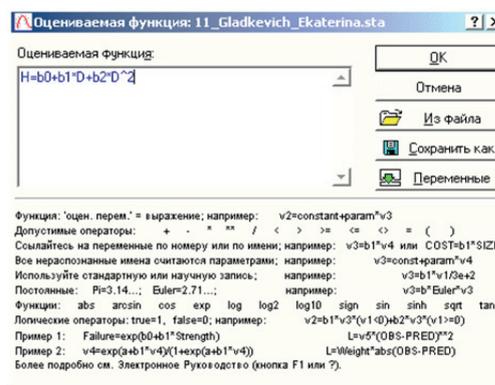


Рис. 11.8. Окно ввода заданной пользователем функции регрессии

В данной формуле используемые переменные H и D (пример для параболы второго порядка приведен на рис 11.8), а транскрипцию набора необходимой функции, в частности знаки умножения (*), степенной функции (^) и другие можно увидеть внизу этого же окна (рис. 11.8).

После подтверждения набранной модели кнопкой *OK* запустим процесс оценки. В появившемся окне *Оценка нелинейной модели МНК* нажмем повторно *OK*, после чего откроется диалоговое окно *Результаты*. В данном окне нажав также кнопку *OK* или на вкладке *Быстрый* кнопку *Оценки параметров модели*, выводится таблица результатов анализа (рис. 11.9), а нажав кнопку *Подогнанная функция и наблюдаемые значения (2M)* на вкладке *Быстрый* – график (рис. 11.10). Коэффициенты уравнения приведены в первом столбце с заголовком *Оценка*. В других столбцах даны

стандартные ошибки коэффициентов (*Стандарт ошиб.*), достоверность коэффициентов (*p-знач*), а также верхний и нижний доверительные интервалы (*Ниж. Дов Предел* и *Вер. Дов Предел*).

Модель: $H=b_0+b_1*D+b_2*D^2$ (11_Gladkevich_Ekaterina.sta)						
Зав. Пер. : H						
Уров. значимости: 95.0% (альфа=0.050)						
	Оценка	Стандарт ошиб.	t-знач. сс = 197	p-знач.	Ниж. Дов Предел	Вер. Дов Предел
b0	13,12559	1,264272	10,38193	0,000000	10,63234	15,61883
b1	0,52498	0,073776	7,11582	0,000000	0,37949	0,67047
b2	-0,00468	0,001036	-4,51115	0,000011	-0,00672	-0,00263

Рис. 11.9. Результаты расчета заданной пользователем функции регрессии

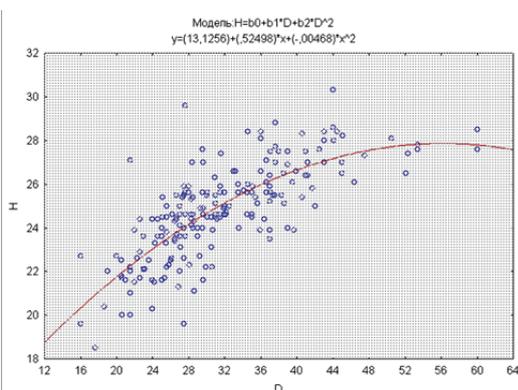


Рис. 11.10. График заданной пользователем функции регрессии

В соответствии с результатами анализа (рис. 11.9) заданное для оценки уравнение регрессии с округлением коэффициентов до целых значений имеет следующий вид:

$$H = 13,12\ 559 + 0,52498D - 0,0468D^2.$$

Некоторые виды регрессионных зависимостей достаточно подробно можно проанализировать в MS Excel. Для этого сначала построим график зависимости одной переменной от другой. После ввода данных в книгу MS Excel и выделения введенных данных (для упрощения процесса построения диаграммы) в меню *Вставка* на вкладке *Диаграммы* (рис. 11.11) выберем нужный тип диаграммы (точечная).

На графике щелкнем правой кнопкой по любой точке диаграммы и выберем опцию *Добавить линию тренда*. С правой стороны (рис. 11.12) отобразятся *типы линии тренда* (типа аппроксимации): экспоненциальная, линейная, логарифмическая, полиномиальная, степенная и скользящее среднее. Выберем, например, *полином 2-й степени* (парабола) и отметим *Показывать уравнение на диаграмме* и *Поместить на диаграмму величину достоверности аппроксимации (R^2)*.

Уравнение регрессии и квадрат корреляционного отношения (коэффициент детерминации) находятся в правом нижнем углу диаграммы (рис. 11.13).

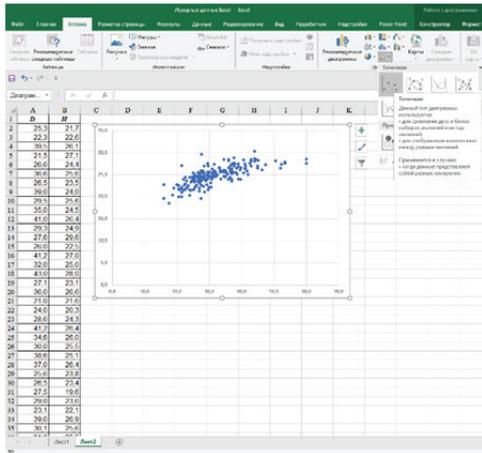


Рис. 11.11. Построение точечной диаграммы в MS Excel

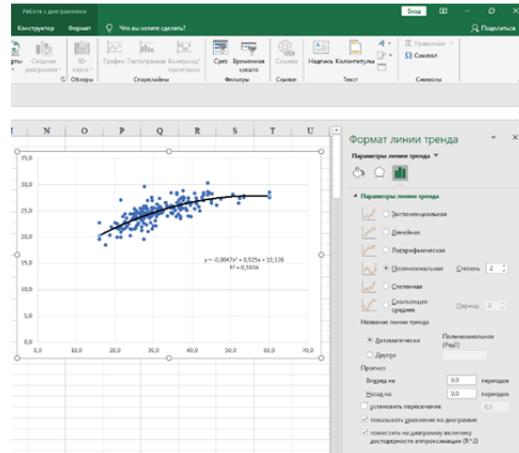


Рис. 11.12. Тип аппроксимации и параметры кривой в MS Excel

Приведенная регрессионная параболическая зависимость характеризуется коэффициентом детерминации $R^2 = 0,5834$, т. е. полученная линия регрессии на 58,34% объясняет варьирование зависимой переменной.

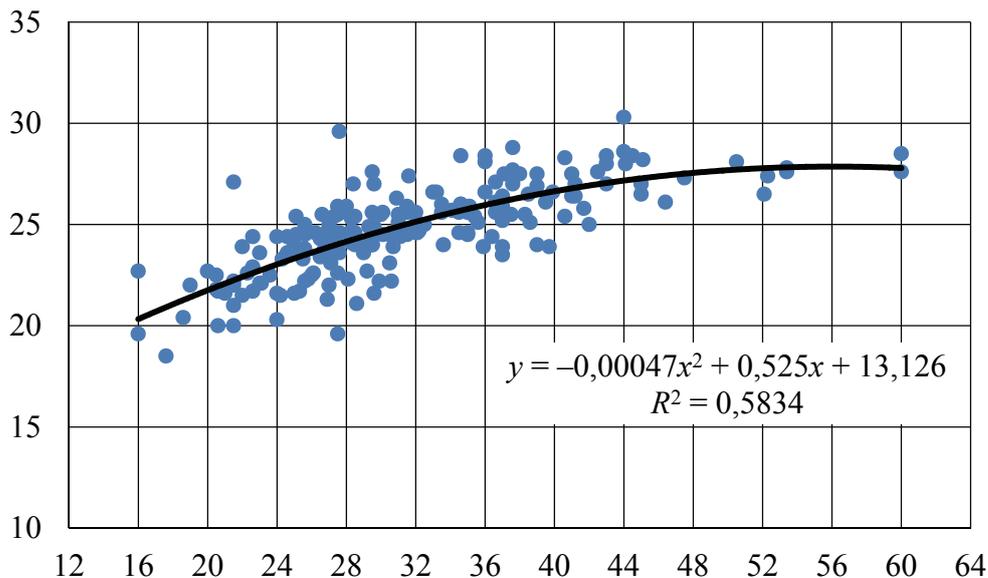


Рис. 11.13. Результат построения диаграммы с нелинейной зависимостью

Для проведения регрессионного анализа и получения коэффициентов уравнения удобен (особенно для множественной линейной регрессии) инструмент *Регрессия* из пакета *Анализ данных* (рис. 11.14). Для реализации выберем в MS Excel пункт меню

Данные команду *Анализ данных*, в появившемся диалоговом окне (рис. 11.14) в списке выберем *Регрессия*. В появившемся диалоговом окне *Регрессия* (рис. 11.15) зададим *Входной интервал Y* (ввести ссылку на диапазон анализируемых зависимых данных, содержащий один столбец данных, например, *H*), укажем *Входной интервал X* (ввести ссылку на диапазон независимых данных, содержащий до 16 столбцов, в нашем случае *D*).

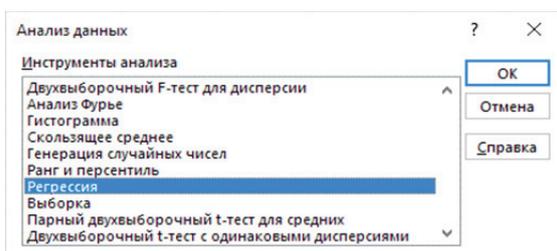


Рис. 11.14. Пакет *Анализ данных* в MS Excel

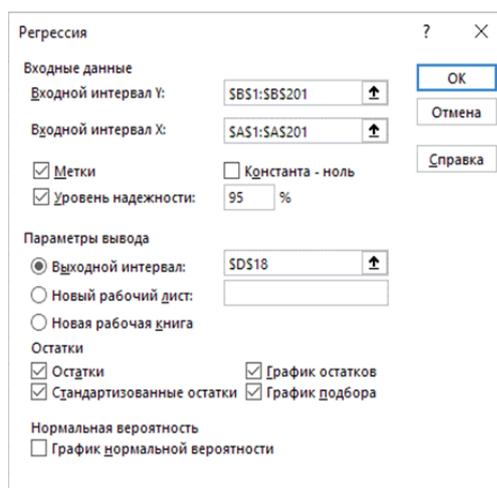


Рис. 11.15. Заполнение в диалоговом окне *Регрессия* в MS Excel

Если необходимо указать *Метки* (учитывает первую строку данных как названия столбцов), *Константа – ноль* (указывает на наличие или отсутствие свободного члена в уравнении регрессии), зададим *Выходной интервал* (достаточно указать левую верхнюю ячейку для будущего диапазона) или *Новый рабочий лист* (можно создать произвольное имя нового листа). Если необходимо получить информацию об остатках и графики остатков, поставим соответствующие флажки. Если необходимо визуально проверить отличие экспериментальных точек от предсказанных по регрессионной модели, установим флажок в поле *График подбора*, нажмем кнопку *OK*. В результате выводится ряд таблиц, данные которых свидетельствуют о качестве полученной регрессионной модели, значениях и значимости коэффициентов линейного уравнения (рис. 11.16).

Выходной диапазон будет включать в себя результаты регрессионного, корреляционного и дисперсионного анализа (рис. 11.16): *множественный R* (коэффициент множественной корреляции), *R-квадрат* (коэффициент детерминации), *нормированный R-квадрат*

(скорректированный коэффициент детерминации для множественной регрессии), *стандартная ошибка* (стандартная ошибка уравнения регрессии), *df* (число степеней свободы вариации), *SS* (сумма квадратов отклонений), *MS* (дисперсия на одну степень свободы вариации), *F* (фактическое значение *F*-критерия), *значимость F* (уровень значимости *F*-критерия), *коэффициенты* (параметры регрессии), *стандартная ошибка* (стандартные ошибки параметров регрессии), *t-статистика* (фактические значения *t*-критерия Стьюдента), *p-значение* (уровень значимости *t*-критерия Стьюдента), *нижние 95%* и *верхние 95%* (доверительные интервалы параметров регрессии).

Если надо выполнить регрессионный анализ для полинома 2-й степени (параболы), то добавим данные независимой переменной, возведенной в квадрат (в нашем случае – D^2), и включим эти данные во *Входной интервал X* для проведения анализа (рис. 11.17).

Значения коэффициентов регрессии для полинома 2-й степени (рис. 11.17) находятся в столбце *Коэффициенты* и соответствуют: *Y-пересечение* – b_0 ; переменная $D1$ – b_1 , переменная $D2$ – b_2 и т. д.

По результатам анализа (рис. 11.16 и 11.17) заданные для оценки уравнения регрессий имеют следующий вид:

– *линейная регрессия*: $H = 18,533 + 0,1968 \cdot D$;

– *параболическая регрессия*: $H = 13,1256 + 0,5250 \cdot D - 0,0047 \cdot D^2$.

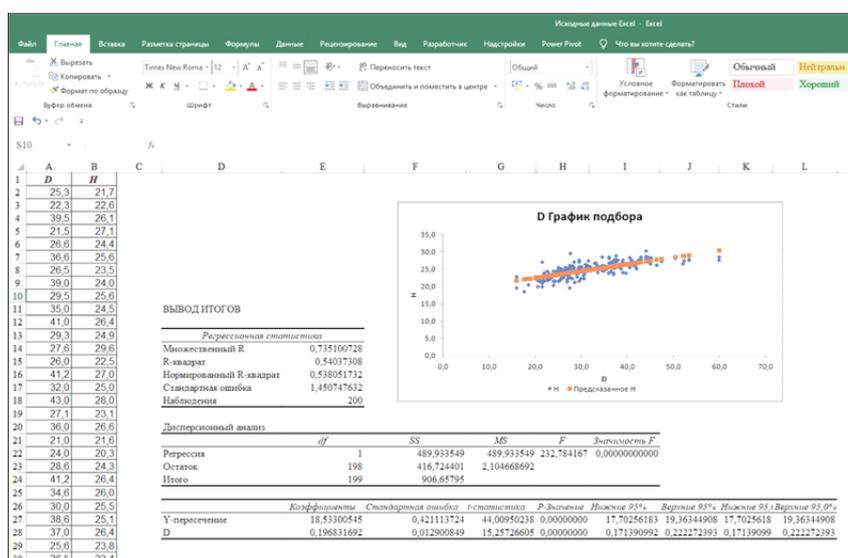


Рис. 11.16. Результаты регрессионного анализа и график соответствия экспериментальных точек и предсказанных по линейной регрессионной модели

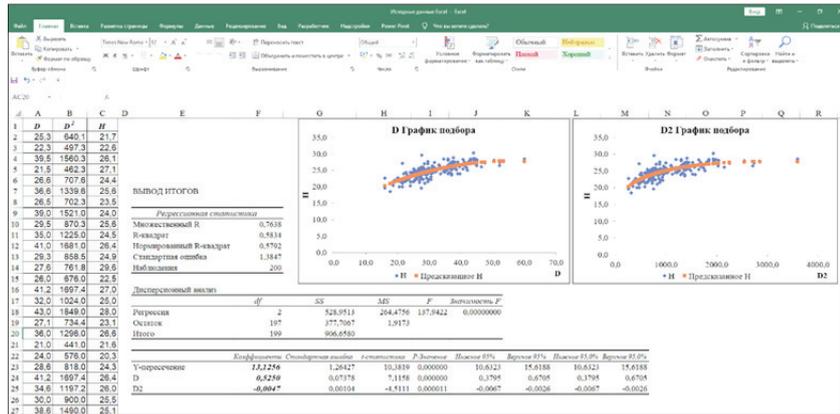


Рис. 11.17. Результаты регрессионного анализа и график соответствия экспериментальных точек и предсказанных по регрессионной модели параболы

В столбце *P-значение* приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда $P > 0,05$, коэффициент может считаться нулевым (это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную). Приводимое значение *R-квадрат* (коэффициент детерминации) показывает, с какой степенью точности и какой процент исходных данных аппроксимирует полученное регрессионное уравнение.

Для проведения регрессионного анализа в MS Excel имеются встроенные статистические функции: ОТРЕЗОК (для расчета коэффициента b_0 в парной линейной регрессии, определяющего отрезок, отсекаемый линией регрессии по оси Y); НАКЛОН (для расчета коэффициента b_1 в парной линейной регрессии, определяющего наклон линии регрессии); ЛИНЕЙН (для расчета множественной линейной регрессии), ТЕНДЕНЦИЯ (для прогноза по множественной линейной регрессии); ПРЕДСКАЗ (для прогноза по парной линейной регрессии); ЛГРФПРИБЛ (для расчета экспоненциальной регрессии), РОСТ (для прогноза по экспоненциальной регрессии) и др.

Для проверки результата дисперсионного анализа в MS Excel (рис. 11.16 и 11.17) могут быть использованы функция ФТЕСТ (массив1; массив2) и процедура *Пакета анализа – Двухвыборочный F-тест для дисперсий* (рис. 11.14).



АНАЛИЗ РАЗЛИЧИЙ ДВУХ ВЫБОРОК

Цель лабораторной работы: научиться выполнять оценку различий выборок и проводить однофакторный дисперсионный анализ.

Обеспечивающие средства: рабочая тетрадь, ручка, калькулятор, линейка, карандаш, стирка, персональный компьютер с установленным пакетом MS Office.

Продолжительность работы: 2 ч.

Общие положения, основные термины и вопросы для проработки лекционного материала и подготовки к лабораторной работе

Обнаружение достоверных отличий статистических параметров – первый шаг к познанию новых биологических закономерностей, причем количественно доказанных. Достоверность различий между генеральными совокупностями может быть определена с помощью критерия Стьюдента (t) и Фишера (F), наименьшей существенной разности (НСР) и др. Данные статистики по результатам расчета сравниваются с табличными при выбранном уровне значимости (обычно 0,05) и числе степеней свободы (объемы выборок без числа ограничений).

Отличия между средними могут иметь два противоположных источника: а) обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности; б) выборки взяты из разных генеральных совокупностей, отличие средних вызвано в основном действием разных доминирующих факторов (а также и случайно).

Сравнение выборочных совокупностей по критерию Стьюдента (t) позволяет утверждать с некоторой долей уверенности сходство или различие между средними выборок по разнице между ними:

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}}. \quad (12.1)$$

Следует помнить, что разность средних берется по модулю, т. е. без учета знака. Полученное этим способом значение критерия Стьюдента сравнивают с табличным при выбранном уровне значимости (обычно для $\alpha = 0,05$) и числе степеней свободы (*объемы выборки без числа ограничений*, $df = n_1 + n_2 - 2$). Если полученное значение (величина) критерия больше табличного, значит различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны.

Изменчивость признаков зависит от многих внешних и внутренних факторов, что привело к необходимости разработки соответствующих математических методов, с помощью которых можно было бы определять влияние отдельных факторов и оценивать их относительную роль в общей изменчивости признаков. Одним из таких методов является дисперсионный анализ.

Дисперсионный анализ (вариансный анализ) – это статистический метод измерения связи изучаемого признака с факторами, оказывающими влияние на этот признак.

Сущность дисперсионного анализа заключается в изучении статистического влияния одного или нескольких факторов на результативный признак. Для этого общий размах вариации распределяют по источникам и оценивают достоверности влияния факторов. Степень влияния факторов на признак оценивают по удельному весу соответствующей факторной дисперсии в общей дисперсии признака.

Дисперсионный анализ используют для установления достоверности и силы влияния факторов на признак, а также для определения относительной доли одного или нескольких факторов в общей изменчивости признака.

Результативный признак (y) – элементарное качество или свойство объектов, изучаемое как результат влияния факторов.

Фактор – любое влияние (воздействие или состояние), разнообразие которых отражается на разнообразии результативного признака.

Градация фактора – степень действия фактора или состояние объектов изучения, которая включает те объекты (с их вариантами), которые подверглись одной степени действия фактора или находились в одном из изучаемых состояний. Факторы, уровни которых не являются точно фиксированными или которые имеют

вообще все возможные случайные градации, называют случайными.

Существуют разнообразные схемы дисперсионного анализа, различаемые 1) *по числу анализируемых факторов* – различают одно-, двух- и трехфакторный дисперсионный анализ (если факторов больше – многофакторным); 2) *характеру градаций внутри фактора* – различают с фиксированными градациями, со случайными и с иерархическими градациями.

В зависимости от распределения вариантов по градациям фактора методы дисперсионного анализа бывают: *равномерные* – с одинаковой повторностью по вариантам; *пропорциональные* – с повышенной повторностью; *неравномерные* – разным числом повторностей с «выпавшими» данными.

В основе дисперсионного анализа лежит закон сложения дисперсий (вариаций), в соответствии с которым *общая дисперсия* (D_0) результативного признака при сгруппированных данных равна сумме *межгрупповой* (факторной) ($D_{гр}$) и *внутригрупповой* (остаточной) дисперсий ($D_{ост}$):

$$D_0 = D_{гр} + D_{ост}; \quad (12.2)$$

$$D_0 = \sum (x_i - \bar{x})^2; \quad (12.3)$$

$$D_{гр} = \sum (x_j - \bar{x})^2 \cdot n_j, \quad (12.4)$$

где x_j – групповые средние; n_j – численность групп.

$$D_{ост} = \sum (x_i - \bar{x}_j)^2, \quad (12.5)$$

где \bar{x}_j – среднее для всех групп.

$D_{ост}$ рассчитывается по каждой группе, после чего находится их сумма, которая и является остаточной вариацией всего комплекса.

Общая дисперсия характеризует общую изменчивость результативного признака. *Межгрупповая* (факторная) дисперсия показывает вариацию результативного признака под влиянием факторного признака, положенного в основу группировки. *Остаточная дисперсия* показывает вариацию результативного признака, обусловленную действием случайных (неучтенных) факторов.

Если исследуется влияние нескольких факторных признаков на результативный, то рассчитываются объемы вариации, обусловленные влиянием каждого фактора и совместным влиянием анализируемых факторов. Например, при изучении влияния факторов A и B общий объем вариации равен

$$D_0 = D_A + D_B + D_{AB} + D_{\text{ост}}, \quad (12.6)$$

где D_A, D_B, D_{AB} – объемы вариации, обусловленные факторами A и B , а также их взаимодействием AB .

Рассчитываются дисперсии на одну степень свободы. Факторная (групповая) дисперсия:

$$D_{\text{гр}} = \frac{W_{\text{гр}}}{m-1}, \quad (12.7)$$

где $W_{\text{гр}}$ – сумма квадратов отклонений групповых средних от общей средней; m – число групп. Остаточная (случайная) дисперсия:

$$D_{\text{ост}} = \frac{W_{\text{ост}}}{n-m}, \quad (12.8)$$

где $W_{\text{ост}}$ – сумма квадратов отклонений остаточных средних от общей средней; n – число наблюдений.

Остаточная вариация обычно вычисляется из формулы соотношений дисперсий:

$$D_{\text{ост}} = D_0 - D_{\text{гр}}. \quad (12.9)$$

Результаты вычислений, выполняемых в ходе дисперсионного анализа, обычно представляют в виде табл. 12.1.

Отношение факторной дисперсии к случайной, рассчитанных на одну степень свободы (табл. 12.1), носит название критерия Фишера (F), который позволяет определить достоверность вывода, сделанного по выборочному обследованию. Фактическое значение критерия Фишера сравнивается с теоретическим с заданной доверительной вероятностью (обычно 0,95), которое берется из специальных таблиц.

Таблица 12.1

Схема однофакторного дисперсионного анализа

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средний квадрат
Межгрупповая	$D_A = \sum_{i=1}^r n_i \cdot (\bar{x}_{A_i} - \bar{x})^2$	$\gamma_A = r - 1$	$S_A^2 = \frac{D_A}{\gamma_A}$
Внутригрупповая	$D_e = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{A_i})^2$	$\gamma_e = n - r$	$S_e^2 = \frac{D_e}{\gamma_e}$
Полная (общая)	$D_y = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2$	$\gamma_y = n - 1$	$S_y^2 = \frac{D_y}{\gamma_y}$

Если $F_{\text{факт}} > F_{\text{табл}}$, то делается вывод о неслучайном характере вариации, о существенности влияния фактора на зависимую переменную. Если $F_{\text{факт}} \leq F_{\text{табл}}$, то можно утверждать, что различия между дисперсиями (факторной и остаточной) носят случайный характер.

Дисперсионный анализ осуществляется в следующей последовательности: расчет общей, факторной и остаточных сумм квадратов отклонений; вычислений дисперсий на одну степень свободы; расчеты F -критерия и его сравнение с табличным для установления достоверности влияния каждого фактора на варьирующий признак; вычисление корреляционных отношений, показывающих степень влияния факторов на зависимую переменную.

Задание 1. Выполнить сравнение выборочных совокупностей для диаметров и для высот по критерию Стьюдента.

Задание 2. Выполнить однофакторный дисперсионный анализ для двух выборочных совокупностей сосны обыкновенной в разных условиях местопроизрастания (типах леса).

Порядок выполнения работы

Задание 1. Выполним сравнение двух выборочных совокупностей (одна из варианта в методическом указании, а вторая – полученная по заданию) по критерию Стьюдента (t). Данные для анализа приведены в табл. 12.2.

Таблица 12.2

Исходные данные для сравнения

Показатель	По методическим указаниям		По заданию	
	D , см	H , м	D , см	H , м
Среднее значение	31,70	24,84	32,60	24,67
Ошибка среднего значения	0,566	0,152	0,603	0,156

По формуле (12.1) определяем критерий Стьюдента:

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} = \frac{|31,70 - 32,60|}{\sqrt{0,566^2 + 0,603^2}} = 1,09 \text{ – для диаметров;}$$

$$t = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} = \frac{|24,84 - 24,67|}{\sqrt{0,152^2 + 0,156^2}} = 0,78 \text{ – для высот.}$$

Сопоставляем табличные значения критерия Стьюдента $t_{0,05;200} = 1,97$ (табл. 4 прил.) при $\alpha = 0,05$ для объема выборки 200 наблюдений с расчетным. Поскольку $t_{0,05; 200} > t_{\Phi}$ и для диаметров, и для высот, то разность между средними признается несущественной (недостовой). Следовательно, данные двух выборочных совокупностей, как для диаметров, так и для высот принадлежат одной общей генеральной совокупности и их можно объединить.

Задание 2. Рассмотрим пример однофакторного дисперсионного анализа с одинаковым количеством наблюдений в группах на примере варианта из данных методических указаний и данных, полученных в выданном задании.

Проанализируем, зависит ли диаметр древостоя от древесной породы. Для этого воспользуемся сгруппированными данными измерений высот у 200 деревьев сосны обыкновенной, произрастающей в кисличном типе леса (методические указания) и 200 деревьев сосны обыкновенной, произрастающей в мшистом типе леса (задание). Исходные данные и основные расчеты для данного примера приведены в табл. 12.3.

С помощью формулы (4.1) вычислим групповые средние дисперсионного комплекса:

– для сосняка кисличного:

$$\bar{x}_1 = \frac{\sum_{j=1}^{n_1} x_{1,j} \cdot f_i}{n_1} = \frac{6340,0}{200} = 31,70 \text{ см.}$$

– для сосняка мшистого:

$$\bar{x}_2 = \frac{\sum_{j=1}^{n_2} x_{2,j} \cdot f_i}{n_2} = \frac{5730,4}{200} = 28,65 \text{ см.}$$

Рассчитаем общую среднюю всего дисперсионного комплекса. Для этого воспользуемся количеством деревьев и средними для каждой породы из второй колонки табл. 12.3 и вычислим с помощью формулы:

$$\bar{x} = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} x_{i,j}}{\sum_{i=1}^r n_i} = \frac{\sum_{i=1}^r n_i \cdot \bar{x}_{A_i}}{\sum_{i=1}^r n_i} = \frac{200 \cdot 31,70 + 200 \cdot 28,65}{200 + 200} = 30,176 \text{ см.}$$

Таблица 12.3

Диаметр стволов в древостое

$D_i(x_{i,j})$	f_i	$x_{i,j} - \bar{x}$	$(x_{i,j} - \bar{x})^2 \cdot f_i$	$x_{i,j} - \bar{x}_{A_i}$	$(x_{i,j} - \bar{x}_{A_i})^2 \cdot f_i$
Сосняк кисличный					
16,0	3	-14,176	602,877	-15,70	739,470
20,0	13	-10,176	1 346,163	-11,70	1 779,570
24,0	29	-6,176	1 106,146	-7,70	1 719,410
28,0	55	-2,176	260,424	-3,70	752,950
32,0	30	1,824	99,809	0,30	2,700
36,0	32	5,824	1 085,407	4,30	591,680
40,0	17	9,824	1 640,687	8,30	1 171,130
44,0	12	13,824	2 293,236	12,30	1 815,480
48,0	2	17,824	635,390	16,30	531,380
52,0	5	21,824	2 381,435	20,30	2 060,450
56,0	0	25,824	0	24,30	0
60,0	2	29,824	1 778,942	28,30	1 601,780
Сумма	200	–	13 230,515	–	12 766,000
Среднее	31,70	–	–	–	–
Сосняк мшистый					
14,7	3	-15,476	718,520	-13,95	583,975
17,9	11	-12,276	1 657,702	-10,75	1 271,661
21,1	29	-9,076	2 388,840	-7,55	1 653,948
24,3	39	-5,876	1 346,568	-4,35	738,656
27,5	32	-2,676	229,151	-1,15	42,467
30,7	33	0,524	9,061	2,05	138,412
33,9	23	3,724	318,968	5,25	633,455
37,1	10	6,924	479,418	8,45	713,687
40,3	9	10,124	922,458	11,65	1 221,083
43,5	3	13,324	532,587	14,85	661,389
46,7	4	16,524	1 092,170	18,05	1 302,921
49,9	2	19,724	778,072	21,25	902,955
53,1	2	22,924	1 051,020	24,45	1 195,409
Сумма	200	–	11 524,534	–	11 060,019
Среднее	28,65	–	–	–	–

Теперь можно приступить к определению сумм квадратов отклонений. Для вычисления общей суммы квадратов отклонений необходимо вычислить отклонения наблюдений от общей средней и возвести их в квадрат (3 и 4 колонки табл. 12.3) и умножить на количество деревьев в данном классе. Сумма этих значений и даст нам искомое значение:

$$\bar{D}_y = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2 \cdot f_i = 13\,230,515 + 11\,524,534 = 24\,755,050.$$

Сумма квадратов отклонений, обусловленная межгрупповой дисперсией, может быть вычислена с использованием полученных ранее средних по формуле:

$$D_x = \sum_{i=1}^r n_i \cdot (\bar{x}_{A_i} - \bar{x})^2 \cdot f_i = 200 \cdot (31,70 - 30,176)^2 + 200 \cdot (28,65 - 30,176)^2 = 929,0304.$$

Теперь найдем сумму квадратов отклонений наблюдений от групповых средних (случайные отклонения). Для этого воспользуемся вычисленными в табл. 12.3 квадратами отклонений наблюдений от групповых средних (6 колонка). Искомое значение будет равняться сумме этих величин:

$$D_e = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_{A_i})^2 \cdot f_i = 1766,0 + 11\,060,019 = 23\,826,019.$$

Эта величина вместе с суммой квадратов отклонений, обусловленной межгрупповой дисперсией $D_x = 929,0304$, должна дать общую сумму квадратов отклонений $D_y = 24\,755,050$:

$$D_e + D_x = 929,0304 + 23\,826,019 = 24\,755,050 = D_y.$$

Такую особенность можно использовать либо для проверки правильности выполненных расчетов, либо для вычисления одной из трех сумм квадратов отклонений по известным двум другим. Далее определим число степеней свободы для каждого источника вариации:

– для общей дисперсии:

$$\gamma_y = n - 1 = 200 + 200 - 1 = 400 - 1 = 399;$$

– для межгрупповой дисперсии:

$$\gamma_x = r - 1 = 2 - 1 = 1;$$

– для остаточной дисперсии:

$$\gamma_e = n - r = 400 - 2 = 398.$$

Вместе с числом степеней свободы для межгрупповой дисперсии степени свободы для остаточной дисперсии должны дать общее число степеней свободы всего комплекса:

$$\gamma_x + \gamma_e = 1 + 398 = 399 = \gamma_y.$$

Делением сумм квадратов отклонений на соответствующие числа степеней свободы определим выборочные дисперсии:

– общая дисперсия всего комплекса:

$$S_y^2 = \frac{D_y}{\gamma_y} = \frac{24\,755,050}{399} = 62,0427;$$

– межгрупповая дисперсия:

$$S_x^2 = \frac{D_x}{\gamma_x} = \frac{929,0304}{1} = 929,0304;$$

– остаточная дисперсия:

$$S_e^2 = \frac{D_e}{\gamma_e} = \frac{23\,826,019}{398} = 59,8644.$$

Определив все необходимые дисперсии, можно приступить к проверке гипотезы о равенстве между собой остаточной и факториальных дисперсий. Для этого вычислим F -статистику Фишера:

$$F_x = \frac{S_x^2}{S_e^2} = \frac{929,0304}{59,8644} = 15,52.$$

Результаты вычислений, выполняемых в ходе дисперсионного анализа, представлены в виде табл. 12.4.

Таблица 12.4

Результаты однофакторного дисперсионного анализа

Компонент дисперсии	Сумма квадратов	Число степеней свободы	Средний квадрат	F фактическое	F табличное
Межгрупповая	929,0304	1	929,0304	15,52	3,98
Остаточная	23 826,019	398	59,8644	–	–
Полная (общая)	24 755,050	399	62,0427	–	–

Далее найдем в табл. 5 прил. квантиль распределения Фишера с 1 и 398 степенями свободы для уровня значимости $\alpha = 0,05$. Это значение равно $F_{0,05;2;82} = 3,98$. Так как вычисленная статистика Фишера (15,52) превышает табличное значение (3,98), мы отклоняем гипотезу о равенстве между собой межгрупповой и остаточной дисперсий. Следовательно, мы обнаружили статистически достоверное влияние древесного вида на диаметр деревьев в древостое.

СПИСОК ВОПРОСОВ ДЛЯ САМОКОНТРОЛЯ

Лабораторная работа № 1

1. Что называется выборочной совокупностью?
2. Как определяется рекомендуемое число классов?
3. Для чего находят максимальные и минимальные значения?
4. Какой графический шифр используется в лесном хозяйстве? Его сущность.
5. Сколько наблюдений должно быть в спелом сосновом древостое?
6. Что такое ступень толщины в практике ведения лесного хозяйства?

Лабораторная работа № 2

1. Что такое корреляционная решетка (двумерная таблица распределения)?
2. Порядок составления и регистрации наблюдений в двумерной таблице.
3. Для чего нужно составлять таблицу распределения, и где она используется?
4. Какие предварительные выводы можно сделать по расположению частот?

Лабораторная работа № 3

1. Какие пакеты программ можно использовать для статистического анализа?
2. Порядок построения графиков гистограммы и полигона.
3. Расскажите, как строятся графики куммуляты и огивы.
4. Порядок построения статистического ряда с использованием пакетов программ Statistica или MS Excel.
5. Какие данные вводятся в программу для получения результата и что получается в итоге?

Лабораторная работа № 4

1. Назовите основные статистические показатели.
2. Какая формула используется для определения среднего диаметра древостоя?
3. Что означает средневзвешенная величина?
4. В каких единицах измеряются диаметр, высота, сумма площадей сечений?
5. Что такое площадь сечения, как она выглядит (фигура)?
6. Чем отличаются смещенная и несмещенная дисперсии?

7. Чем отличаются дисперсия от среднеквадратического отклонения?
8. Коэффициент асимметрии (формула, разновидности, графическое отображение). Что означает левосторонняя и правосторонняя асимметрия? (Нарисовать график.)
9. Коэффициент эксцесса (формула, разновидности, графическое отображение). Что означает островершинный и плосковершинный эксцесс? (Нарисовать график.)
10. Коэффициент вариации, точность оценки средней величины и их практическое значение.

Лабораторная работа № 5

1. Что такое мода и как вычисляется?
2. Как называется распределение, которое имеет один, два или более максимума?
3. Что такое медиана и как вычисляется?
4. Чем отличаются квантили и перцентили от медианы?

Лабораторная работа № 6

1. Нормальное распределение и его практическое значение.
2. Свойства кривой нормального распределения.
3. Вычисление теоретических частот для функции нормального распределения.
4. Графическое отображение сравнения эмпирического и нормального распределений.

Лабораторная работа № 7

1. Критерии достоверности и их классификация.
2. Критерии согласия Пирсона и Колмогорова и их практическое значение.
3. Как определить, подчиняется ли данное распределение закону нормального распределения?
4. Вычисление критерия согласия Пирсона и его табличного значения, сравнение и вывод.
5. В зависимости от чего находят в таблице квантиль распределения Пирсона?
6. Вычисление критерия согласия Колмогорова и его табличного значения, сравнение и вывод.

Лабораторная работа № 8

1. Порядок работ по вычислению основных статистик с использованием пакетов программ Statistica или MS Excel.
2. Порядок работ по анализу распределения случайной величины с использованием пакетов программ Statistica или MS Excel.

ПРИЛОЖЕНИЕ

Таблица 1

Нормальное распределение

<i>x</i>	0	1	2	3	4	5	6	7	8	9
0,0	0,500	0,504	0,506	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	0,540	0,544	0,548	0,552	0,556	0,560	0,564	0,567	0,571	0,575
0,2	0,579	0,583	0,587	0,591	0,595	0,599	0,603	0,606	0,610	0,614
0,3	0,618	0,622	0,626	0,629	0,633	0,637	0,641	0,644	0,648	0,652
0,4	0,655	0,659	0,663	0,666	0,670	0,674	0,677	0,681	0,684	0,688
0,5	0,691	0,695	0,698	0,702	0,705	0,709	0,712	0,716	0,719	0,722
0,6	0,728	0,729	0,732	0,736	0,739	0,742	0,745	0,749	0,752	0,755
0,7	0,758	0,761	0,764	0,767	0,770	0,773	0,776	0,779	0,782	0,785
0,8	0,788	0,791	0,794	0,797	0,800	0,802	0,805	0,808	0,811	0,813
0,9	0,816	0,819	0,821	0,824	0,826	0,829	0,831	0,834	0,836	0,839
1,0	0,841	0,844	0,846	0,848	0,851	0,853	0,855	0,858	0,860	0,862
1,1	0,864	0,866	0,869	0,871	0,873	0,875	0,877	0,879	0,881	0,883
1,2	0,885	0,887	0,889	0,891	0,893	0,894	0,896	0,898	0,900	0,901
1,3	0,903	0,905	0,907	0,908	0,910	0,911	0,913	0,915	0,916	0,918
1,4	0,919	0,921	0,922	0,924	0,925	0,926	0,928	0,929	0,931	0,932
1,5	0,933	0,934	0,936	0,937	0,938	0,939	0,941	0,942	0,943	0,944
1,9	0,971	0,972	0,973	0,973	0,974	0,974	0,975	0,976	0,976	0,977
2,0	0,977	0,978	0,978	0,979	0,979	0,980	0,980	0,981	0,981	0,982
2,1	0,982	0,983	0,983	0,983	0,984	0,984	0,985	0,985	0,985	0,986
2,2	0,986	0,986	0,987	0,987	0,987	0,988	0,988	0,988	0,989	0,989
2,3	0,989	0,990	0,990	0,990	0,990	0,991	0,991	0,991	0,991	0,992
2,4	0,992	0,992	0,992	0,992	0,993	0,993	0,993	0,993	0,993	0,994
2,5	0,994	0,994	0,994	0,994	0,994	0,995	0,995	0,995	0,995	0,995
2,6	0,995	0,995	0,996	0,996	0,996	0,996	0,996	0,996	0,996	0,996
2,7	0,996	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997	0,997
2,8	0,997	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,998
2,9	0,998	0,998	0,998	0,998	0,998	0,998	0,998	0,999	0,999	0,999
3,0	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,1	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999
3,2	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	0,999	1,000
3,3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
3,4	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
3,5	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
1,6	0,945	0,946	0,947	0,948	0,950	0,951	0,952	0,953	0,954	0,954
1,7	0,955	0,956	0,957	0,958	0,959	0,960	0,961	0,962	0,961	0,963
1,8	0,964	0,965	0,966	0,966	0,967	0,968	0,969	0,969	0,970	0,971

Таблица 2

Критические значения χ^2 распределения

$\gamma \backslash \alpha$	0,99	0,95	0,90	0,80	0,70	0,50	0,30	0,10	0,05	0,01	0,001
1	0,00016	0,00393	0,0158	0,0642	0,148	0,455	1,074	2,706	3,841	6,635	10,827
2	0,0201	0,103	0,211	0,446	0,713	1,386	2,408	4,605	5,991	9,210	13,8915
3	0,115	0,352	0,584	1,005	1,424	2,366	3,665	6,251	7,815	11,345	16,266
4	0,297	0,711	1,064	1,649	2,195	3,357	4,878	7,779	9,488	13,277	18,467
5	0,554	1,145	1,610	2,343	3,000	4,351	6,064	9,236	11,070	15,086	20,515
6	0,872	1,635	2,204	3,070	3,828	5,348	7,231	10,645	12,592	16,812	22,457
7	1,239	2,167	2,833	3,822	4,671	6,346	8,363	12,017	14,067	18,475	24,322
8	1,646	2,733	3,490	4,594	5,527	7,344	9,524	13,362	15,507	20,090	26,125
9	2,088	3,325	4,168	5,380	6,393	8,343	10,656	14,684	16,919	21,366	27,877
10	2,558	3,940	4,865	6,179	7,267	9,342	11,781	15,987	18,307	23,209	29,588
11	3,053	4,575	5,578	6,989	8,148	10,341	12,899	17,275	19,675	24,725	31,264
12	3,571	5,226	6,304	7,807	9,034	11,340	14,011	18,519	21,026	26,207	32,909
13	4,107	5,892	7,042	8,634	9,926	12,340	15,119	19,812	22,362	27,688	34,528
14	4,660	6,571	7,790	9,467	10,821	13,339	16,222	21,064	23,685	29,141	36,123
15	5,229	7,261	8,547	10,307	11,721	14,339	17,322	22,307	24,996	30,578	37,697
16	5,812	7,962	9,312	11,152	12,624	15,338	18,418	23,542	26,296	32,000	39,252
17	6,408	8,672	10,085	12,002	13,531	16,338	19,511	24,769	27,587	33,409	40,790
18	7,015	9,390	10,865	12,857	14,440	17,338	20,601	25,989	28,869	34,805	42,312
19	7,633	10,117	11,651	13,716	15,352	18,338	21,689	27,204	30,144	36,191	43,820
20	8,260	10,851	12,443	14,578	16,266	19,337	22,775	28,412	31,410	37,566	45,315
21	8,897	11,591	13,240	15,445	17,182	20,337	23,853	29,615	32,671	38,932	46,797
22	9,542	12,338	14,041	16,310	18,101	21,337	24,939	30,813	33,924	40,289	48,268
23	10,196	13,091	14,848	17,187	19,021	22,337	26,018	32,007	35,172	41,638	49,728
24	10,856	13,848	15,659	18,062	19,943	23,337	27,096	33,196	36,415	42,980	51,179
25	11,524	14,611	16,473	18,940	20,867	24,337	28,172	34,382	37,652	44,314	52,620
26	12,198	15,379	17,292	19,820	21,792	25,336	29,246	35,563	38,885	45,642	54,052
27	12,879	16,151	18,114	20,703	22,719	26,336	30,319	36,741	40,113	46,963	55,476
28	13,565	16,928	18,939	21,588	23,647	27,336	31,391	37,916	41,337	48,278	56,893
29	14,256	17,708	19,768	22,475	24,577	28,336	32,461	39,087	42,557	49,588	58,302
30	14,953	18,493	20,599	23,364	25,508	29,336	33,530	40,256	43,773	50,892	59,703

В таблице приведены значения квантилей $\chi^2_{\alpha, \gamma}$ в зависимости от числа степеней свободы γ и вероятности α такими, что $P(\chi^2 \geq \chi^2_{\alpha, \gamma}) = \alpha$.

Таблица 3

Квантили распределения Колмогорова λ_α

α	λ_α	α	λ_α	α	λ_α
0,99	0,44	0,50	0,83	0,15	1,14
0,90	0,57	0,40	0,89	0,10	1,22
0,80	0,64	0,30	0,97	0,05	1,36
0,70	0,71	0,25	1,02	0,02	1,53
0,60	0,77	0,20	1,07	0,01	1,63

В таблице приведены значения квантилей λ_α в зависимости от вероятности α такой, что $P(\lambda \geq \lambda_\alpha) = \alpha$.

Таблица 4

Значения критерия Стьюдента при различных уровнях значимости

γ	Уровень значимости			γ	Уровень значимости		
	0,05	0,01	0,001		0,05	0,01	0,001
2	4,30	9,93	31,60	21	2,08	2,83	3,82
3	3,18	5,84	12,94	22	2,07	2,82	3,79
4	2,78	4,60	8,61	23	2,07	2,81	3,77
5	2,57	4,03	6,86	24	2,06	2,80	3,75
6	2,45	3,71	5,96	25	2,06	2,79	3,73
7	2,37	3,50	5,41	26	2,06	2,78	3,71
8	2,31	3,36	5,04	27	2,05	2,77	3,69
9	2,26	3,25	4,78	28	2,05	2,76	3,67
10	2,23	3,17	4,49	29	2,04	2,76	3,66
11	2,20	3,11	4,44	30	2,04	2,75	3,65
12	2,18	3,06	4,32	40	2,02	2,70	3,55
13	2,16	3,01	4,22	50	2,01	2,68	3,50
14	2,15	2,98	4,14	60	2,00	2,66	3,46
15	2,13	2,95	4,07	80	1,99	2,64	3,42
16	2,12	2,92	4,02	100	1,98	2,63	3,39
17	2,11	2,90	3,97	120	1,98	2,63	3,37
18	2,10	2,88	3,92	200	1,97	2,60	3,34
19	2,09	2,86	3,88	500	1,96	2,59	3,31
20	2,09	2,85	3,85	∞	1,96	2,58	3,29

Таблица 5

F-распределение Фишера (в таблице приведены верхние 5% – точки $F(\gamma_6, \gamma_3, 0,95)$)

γ_6	γ_3																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,45	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,43	4,28	4,21	4,15	4,10	4,06	4,02	3,97	3,91	3,84	3,81	3,77	3,74	3,70	3,67	3,63
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,47	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,82	1,79	1,74	1,69	1,64	1,58	1,51
50	4,03	3,18	2,79	2,56	2,4	2,29	2,2	2,13	2,07	2,03	1,95	1,87	1,78	1,74	1,69	1,63	1,58	1,51	1,44
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,29	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

ЛИТЕРАТУРА

1. Атрощенко, О. А. Лесная биометрия: учеб. пособие для студентов специальности «Лесное хозяйство» / О. А. Атрощенко, В. П. Машковский. – Минск: БГТУ, 2009. – 340 с.
2. Бондаренко, А. С. Статистическая обработка материалов лесоводственных исследований: учеб. пособие / А. С. Бондаренко, А. В. Жигунов. – СПб.: Из-во Политехнического университета, 2016. – 125 с.
3. Герасимов, Ю. Ю. Математические методы и модели в расчетах на ЭВМ: применение в лесоуправлении и экологии: учеб. для лесных вузов / Ю. Ю. Герасимов, В. К. Хлюстов. – М.: МГУЛ, 2001. – 260 с.
4. Ивантер, Э. В. Введение в количественную биологию: учеб. пособие / Э. В. Ивантер, А. В. Коросов. – Петрозаводск: Изд-во ПетрГУ, 2011. – 302 с.
5. Смольянов, А. Н. Математические методы в лесном хозяйстве: тексты лекций / А. Н. Смольянов, А. В. Мироненко. – Воронеж: М-во образования и науки РФ, ФГБОУ ВПО «ВГЛТА». 2013. – 143 с.
6. Гефан, Г. Д. Статистический метод и основы его применения: учеб. пособие / Г. Д. Гефан. – Иркутск: ИрГУПС, 2003. – 208 с.
7. Борздова, Т. В. Основы статистического анализа и обработка данных с применением Microsoft Excel: учеб. пособие / Т. В. Борздова. – Минск: ГИУСТ БГУ, 2011. – 75 с.
8. Башмакова, И. Б. Математическая статистика: учеб. пособие / И. Б. Башмакова, И. И. Кораблева, С. С. Прасникова. – СПб.: СПбГАСУ, 2017. – 68 с.
9. Белокуров, С. Г. Математические методы в биологии: учеб.-метод. пособие / С. Г. Белокуров, М. С. Трескин. – Караваево: Костромская ГСХА, 2015. – 50 с.
10. Сушко, Г. Г. Биометрия: методические указания для проведения лабораторных работ. В 2-х ч. / Г. Г. Сушко, И. А. Литвенкова. – Витебск: ВГУ имени П. М. Машерова, 2019. – Ч. 2. – 47 с.
11. Сиделев, С. И. Математические методы в биологии и экологии: введение в элементарную биометрию: учеб. пособие / С. И. Сиделев; Яросл. гос. ун-т имени П. Г. Демидова. – Ярославль: ЯрГУ, 2012. – 140 с.
12. Бараз, В. Р. Использование MS Excel для анализа статистических данных: учеб. пособие / В. Р. Бараз, В. Ф. Пегашкин. – Нижний Тагил: НТИ (филиал) УрФУ, 2014. – 181 с.
13. Справочник по прикладной статистике. В 2-х т. / под. ред. Э. Ллойда, У. Ледермана. – М., 1989. – Т. 1. – 510 с.; 1990. – Т. 2. – 526 с.

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ	3
ПЕРЕЧЕНЬ ОСНОВНЫХ УСЛОВНЫХ ОБОЗНАЧЕНИЙ	4
Лабораторная работа № 1. СТАТИСТИЧЕСКИЕ РЯДЫ	6
Лабораторная работа № 2. ДВУМЕРНАЯ ТАБЛИЦА РАСПРЕДЕЛЕНИЯ	14
Лабораторная работа № 3. СОСТАВЛЕНИЕ СТАТИСТИЧЕСКИХ РЯДОВ И ИХ ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ С ИСПОЛЬЗОВАНИЕМ ПАКЕТА ПРОГРАММ	18
Лабораторная работа № 4. ОПРЕДЕЛЕНИЕ ОСНОВНЫХ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ	28
Лабораторная работа № 5. СТРУКТУРНЫЕ ХАРАКТЕРИСТИКИ СТАТИСТИЧЕСКОГО РЯДА	42
Лабораторная работа № 6. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН	48
Лабораторная работа № 7. СТАТИСТИЧЕСКАЯ ПРОВЕРКА НЕПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ	55
Лабораторная работа № 8. ВЫЧИСЛЕНИЕ ОСНОВНЫХ СТАТИСТИК И АНАЛИЗ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН С ИСПОЛЬЗОВАНИЕМ ПАКЕТА ПРОГРАММ	63
Лабораторная работа № 9. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ...	73
Лабораторная работа № 10. РЕГРЕССИОННЫЙ АНАЛИЗ	79
Лабораторная работа № 11. РЕГРЕССИОННЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ ПАКЕТА ПРОГРАММ	89
Лабораторная работа № 12. АНАЛИЗ РАЗЛИЧИЙ ДВУХ ВЫБОРОК	102
СПИСОК ВОПРОСОВ ДЛЯ САМОКОНТРОЛЯ	111
ПРИЛОЖЕНИЕ	114
ЛИТЕРАТУРА	118

Учебное издание

Сидельник Николай Ярославович
Машковский Владимир Петрович
Севрук Павел Владимирович

ЛЕСНАЯ БИОМЕТРИЯ

ЛАБОРАТОРНЫЙ ПРАКТИКУМ

Учебно-методическое пособие

Редактор *Е. И. Гоман*
Компьютерная верстка *А. А. Селиванова*
Дизайн обложки *П. П. Падалец*
Корректор *Е. И. Гоман*

Подписано в печать 28.05. 2021. Формат 60×84¹/₁₆.
Бумага офсетная. Гарнитура Таймс. Печать ризографическая.
Усл. печ. л. 7,0. Уч.-изд. л. 7,2.
Тираж 170 экз. Заказ .

Издатель и полиграфическое исполнение:
УО «Белорусский государственный технологический университет».
Свидетельство о государственной регистрации
издателя, изготовителя, распространителя печатных изданий
№ 1/227 от 20.03.2014.
Ул. Свердлова, 13а, 220006, г. Минск.