Городничев Руслан Михайлович, к.б.н., зав. лаб. БиоМ, доцент ЭГО ИЕН СВФУ

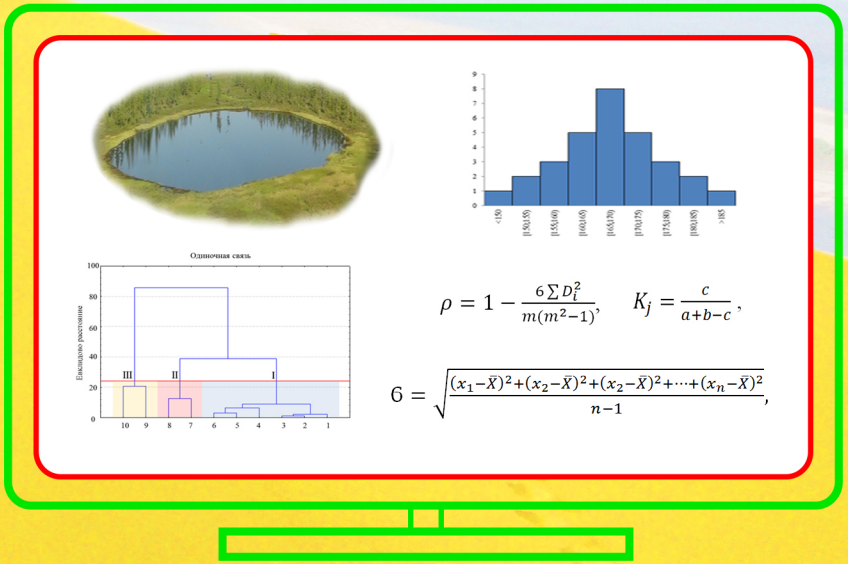
Пестрякова Людмила Агафьевна, д.г.н., г.н.с. ЭГО ИЕН СВФУ, руководитель магистратуры по профилю «Геоэкология»

Ушницкая Лена Алексеевна, н.с. лаб. БиоМ ЭГО ИЕН СВФУ

Левина Сардана Николаевна, м.н.с., БиоМ ЭГО ИЕН СВФУ

Давыдова Парасковья Васильевна, вед. инженер БиоМ ЭГО ИЕН СВФУ

Методы экологических исследований
ОСНОВЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ



Якутск 2019

Министерство науки и высшего образования Российской Федерации
Северо-Восточный федеральный университет имени М.К. Аммосова
Институт естественных наук
Эколого-географическое отделение

Методы экологических исследований

ОСНОВЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

Учебно-методическое пособие

Якутск
2019

УДК 31:574(075.8)
ББК 20.080я73

Утверждено учебно-методическим советом СВФУ

Авторы:
Р.М. Городничев, Л.А. Пестрякова, Л.А. Ушницкая,
С.Н. Левина, П.В. Давыдова

Рецензенты:
Ю.И. Трофимцев, д.т.н., ИМИ СВФУ им. М.К. Аммосова (г. Якутск),
М.М. Черосов, д.б.н., ИБПК СО РАН (г. Якутск)

Методы экологических исследований. Основы статистической обработки данных: учебно-методическое пособие / [Р.М. Городничев и др.]. – Якутск : Издательский дом СВФУ, 2019. – 94 с.
ISBN 978-5-7513-2737-8

В учебно-методическом пособии приведены описание, алгоритмы и примеры расчетов широко применяемых методов статистической обработки данных, в том числе основных статистических характеристик, индексов биоразнообразия и таксономического сходства, коэффициентов корреляции, критериев оценки статистической значимости, различий выборок объектов исследований, а также иерархического кластерного анализа.

Предназначено для студентов бакалавриата и магистратуры направлений «Экология и природопользование», «География» и «Биология».

УДК 31:574(075.8)
ББК 20.080я73

ISBN 978-5-7513-2737-8

ОГЛАВЛЕНИЕ

Предисловие

Статистическая обработка данных является универсальным инструментов получения информации о различных процессах и явлениях живой и неживой природы. Различные методы обработки данных, применение персональных компьютеров и специализированного программного обеспечения являются важным условием проведения современных исследований. В настоящее время, применяя статистические методы, можно установить наличие связей характеристик, процессов и явлений, разбивать большие совокупности данных на группы и классы, используя объективные (основанные на математических процедурах) критерии, устанавливать существенность различий признаков групп объектов исследования и многое др. Статистические методы позволяют анализировать большие массивы информации, находя смыслы, скрытые от человеческих глаз. Использование современных методов статистики и компьютерного оборудования обеспечивает высокую скорость и безошибочность вычислительных процедур, позволяет получить результаты в наглядном виде. Широкий спектр методов обработки информации требует понимания базовых вычислительных процедур и основ статистической обработки данных, на достижение которых направлено данное учебно-методическое пособие.

Цель учебно-методического пособия – произвести описание основных методов статистической обработки данных, применяемых в экологии.

Основные задачи:

– произвести теоретический обзор наиболее эффективных и широко применяемых в экологических исследованиях методов статистической обработки данных;

– привести примеры расчетов численных характеристик и алгоритмов действий, лежащих в основе рассматриваемых методов статистической обработки данных;

– привести примеры использования методов статистической обработки данных на практике;

– выработать навыки применения методов статистической обработки данных у обучающихся путем выполнения заданий самостоятельной работы по пройденным темам.

Все методы статистической обработки данных, приведенные в данной книге, основаны на вычислительных процедурах, которые могут быть реализованы вручную без использования специального оборудования и программного обеспечения. В книге освещены основные элементарные статистические характеристики, индексы биоразнообразия и таксономического сходства, произведено описание простейших процедур проверки данных на соответствие закону нормального распределения, приведены описания расчетов параметрических и непараметрических коэффициентов корреляции и критериев, характеризующих значимость различий характеристик выборок данных. Отдельно рассмотрены процедуры иерархического кластерного анализа, основанные на различных алгоритмах кластеризации.

Учебно-методическое пособие предназначено для освоения таких дисциплин как: «Методы экологических исследований», «Эколого-аналитическая оценка природных сред» и «Компьютерные технологии и статистические методы в экологии и природопользовании». Методы и алгоритмы, описываемые в книге, необходимы при написании (отвечающих требованиям современных образовательных стандартов) курсовых и выпускных квалификационных работ в области наук о Жизни, наук о Земле и других направлений, связанных с обработкой численной информации об объектах окружающей среды.

Данное учебно-методическое пособие предназначено для обучающихся бакалавриата и магистратуры эколого-географического отделения Института естественных наук Северо-Восточного федерального университета имени М.К. Аммосова, а также может быть использовано студентами и специалистами естественных и гуманитарных направлений других подразделений и организаций. Книга предназначена для получения начальных знаний в области статистической обработки данных.

1. Элементарные статистические характеристики

Данный раздел посвящен наиболее часто используемым при оперировании с численными характеристиками параметрам, которые применяются в качестве исходных элементов большинства статистических анализов.

1. **Среднее арифметическое значение.** Наиболее широко известный и часто употребляемый параметр «среднее арифметическое значение». Применяется в самых разнообразных случаях, когда нужно получить обобщенное представление о значении какой-либо характеристики. Например, исследователь занимается изучением успеваемости студентов разных групп по математике. Для упрощения процесса сравнения он может вычислить среднее арифметическое значение оценки студентов в каждой из этих групп, и получившиеся значения сопоставить, сделав общий вывод, о том, что студенты одной из групп более успешны.

Вычисляется среднее арифметическое по формуле (1.1):

$$\bar{X}=\frac{x_1+x_2+x_3+\cdots+x_n}{n}, \quad (1.1)$$

где $\bar{X}$ – среднее арифметическое значение;
x_i – значения рассматриваемой характеристики;
n – количество объектов исследования.

Пример: Масса 5 яблок соответственно равна 250, 220, 200, 223 и 230 г. Среднее арифметическое массы яблок будет рассчитано следующим образом: $\bar{X}=\frac{250+220+200+223+230}{5} \approx 224{,}6$ (г).

2. **Медиана (*Me*).** Другая разновидность среднего значения группы объектов исследования – это медиана. Медиана – это такое значение рассматриваемой характеристики, которое расположено посередине вариационного ряда. То есть половина объектов исследования меньше значений медианы, а другая половина – больше. Лучше всего понять, что такое медиана, используя примеры.

Пример: Вернувшись к предыдущему примеру с яблоками, укажем медиану. Итак, для начала нужно построить вариационный ряд, то есть упорядочить данные в порядке возрастания или убывания. Масса 5 яблок упорядоченная по возрастанию: 200, 220, 223, 230 и 250 г. Здесь значение параметра (масса яблок), расположенное посередине ряда и будет медианой. То есть медиана в данном случае равна 223.

В случае с нечетным количеством объектов исследования медиана будет найдена как в отмеченном выше примере. Если объектов будет четное количество, то расчет немного усложнится. Предположим, что выборка объектов исследования насчитывает 6 яблок, имеющих массу 250, 220, 200, 223, 230 и 218 г. Для вычисления медианы строим вариационный ряд. Он будет иметь следующий вид: 200, 218, 220, 223, 230 и 250. Далее находим значения расположенные в средней части ряда. В нашем случае это 220 и 223. Вычислим среднее арифметическое этих чисел, оно и будет медианой: $Me = \frac{220+223}{2} \approx$ $\approx 221{,}5$ (г).

3. **Мода (*Мо*).** Еще одним типом среднего значения является мода. Мода – это наиболее часто повторяющееся значение данных. Запомнить смысл данной характеристики очень просто. В жизни модным называют какой-либо часто используемый или «широко распространенный» предмет. Например: модные штаны, модная шляпа и др.

Пример. Приведен рост нескольких студентов: 175, **155**, 163, **155**, 162, 174 см. В данном случае чаще всех повторяется значение 155. Оно и будет являться модой.

4. **Линейное отклонение.** В повседневной жизни часто приходится определять положение одного объекта относительно другого. Например, стол стоит в 1 м от стены, или школа расположена в 2 км к югу от дома. Как происходит определение этого относительного положения? Мы отмеряем физическое расстояние (линейное отклонение) между конкретными объектами.

На рисунке 1.1 изображены рекламные щиты, расположенные на разном расстоянии от дома. Для того, чтобы определить расстояние (относительное расположение) щита I от щита II, необходимо будет

провести следующую манипуляцию: $13 - 7 = 6$ (м). В данном случае мы определили простое отклонение первого объекта от второго.

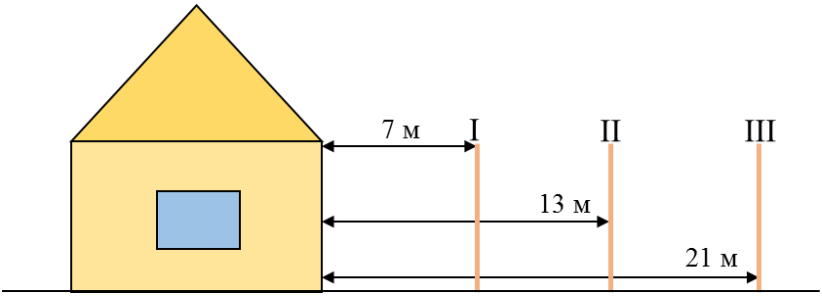


Рисунок 1.1. Дом и расположенные на различном удалении от него рекламные щиты

На практике часто приходится измерять, как сильно значения параметра группы объектов исследования отклонены от среднего арифметического значения. Это необходимо для определения того насколько данные однородны.

Пример: Определим среднее расстояние рекламных щитов до дома (рисунок 1.1): $\bar{X}=\frac{7+13+21}{3} \approx 13{,}67$ (м).

Отметим на вышеприведенном графическом примере высчитанное среднее расстояние щитов от дома (рисунок 1.2).

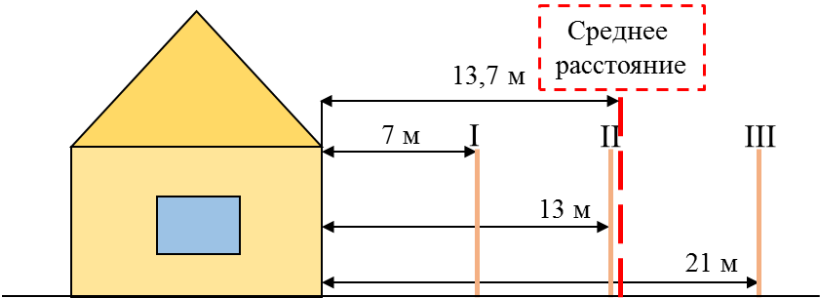


Рисунок 1.2. Среднее расстояние рекламных щитов от дома

Теперь определим, насколько же I, II и III рекламные щиты «отклонены» от среднего. Для этого проведем арифметическую манипуляцию аналогичную предыдущей.

Расстояние (отклонение) I, II и III рекламных щитов от значения среднего расстояния будут определены соответственно: $7 - 13{,}67 = -6{,}67$; $13 - 13{,}67 = -0{,}67$ и $21 - 13{,}67 = 7{,}33$.

Теперь высчитаем среднее отклонение щитов от значения среднего расстояния. Для этого применим формулу (1.1) для расчета среднего арифметического значения: $\bar{X}=\frac{-6{,}67+(-0{,}67)+7{,}33}{3} = 0$.

Ответ 0. Положительные и отрицательные значения при суммировании как бы «компенсируются». В этом случае не получится установить «разброс значений» параметра в большую и меньшую сторону от среднего арифметического. Поэтому для получения значений отклонения от среднего отличных от нуля было принято решение избавиться от отрицательных значений данных. Для этого применены 2 подхода: использование модуля (в формуле среднего линейного отклонения) и возведение в квадрат (в формулах стандартного отклонения и дисперсии).

5. **Среднее линейное отклонение (*а*)** (1.2).

$$a = \frac{|x_1-\bar{X}|+|x_2-\bar{X}|+|x_3-\bar{X}|+\cdots+|x_n-\bar{X}|}{n}, \quad (1.2)$$

где $\bar{X}$ – среднее арифметическое значение;
x_i – значения рассматриваемой характеристики;
n – количество объектов исследования.

6. **Стандартное отклонение (или среднее квадратическое отклонение) (б)** – также характеризует разброс значений параметров вокруг среднего арифметического и вычисляется по формуле (1.3), если количество объектов исследования n>30, или по формуле (1.4), если количество объектов исследования n<30.

$$б = \sqrt{\frac{(x_1-\bar{X})^2+(x_2-\bar{X})^2+(x_2-\bar{X})^2+\cdots+(x_n-\bar{X})^2}{n}}, \quad (1.3)$$

где $\bar{X}$ – среднее арифметическое значение;
x_i – значения рассматриваемой характеристики;
n – количество объектов исследования.

$$6 = \sqrt{\frac{(x_1-\bar{X})^2+(x_2-\bar{X})^2+(x_2-\bar{X})^2+\cdots+(x_n-\bar{X})^2}{n-1}}, \quad (1.4)$$

Стандартное отклонение – наиболее часто используемая мера отклонения значений исследуемой характеристики от среднего арифметического значения. Стандартное отклонение выступает исходным параметров для расчета достаточно большого количества статистических параметров.

7. **Дисперсия.** Стандартное отклонение, возведенное в квадрат (или стандартное отклонение без извлечения квадратного корня) получило название дисперсия ($б^2$).

При n>30 6^2 рассчитывается по формуле (1.5):

$$6^2 = \frac{(x_1-\bar{X})^2+(x_2-\bar{X})^2+(x_2-\bar{X})^2+\cdots+(x_n-\bar{X})^2}{n}, \quad (1.5)$$

При n<30 6^2 рассчитывается по формуле (1.6):

$$6^2 = \frac{(x_1-\bar{X})^2+(x_2-\bar{X})^2+(x_2-\bar{X})^2+\cdots+(x_n-\bar{X})^2}{n-1}, \quad (1.6)$$

Приведем примеры расчетов среднего линейного, стандартного отклонений и дисперсии, используя данные таблицы 1.1.

Таблица 1.1

Значения массы тела взрослых самцов популяции волка

№	Значения массы тела, кг
1	42
2	43
3	42,5
4	42,5
5	43,3

6	42
7	44,3
8	44

Итак, для начала рассчитаем среднее значение массы тела $\bar{X}$, применив формулу (1.1): $\bar{X} = \frac{42+43+42,5+42,5+43,3+42+44,3+44}{8} = 43$.

Далее по формуле (1.2) вычислим среднее линейное отклонение:

$$a = \frac{|42-43|+|43-43|+|42,5-43|+|42,5-43|+|43,3-43|+|42-43|+ +|44,3-43|+|44-43|}{8} = 0,7.$$

Используя формулу (1.4) вычислим стандартное отклонение:

$$6 = \sqrt{\frac{(42-43)^2+(43-43)^2+(42,5-43)^2+(42,5-43)^2+(43,3-43)^2+ +(42-43)^2+(44,3-43)^2+(44-43)^2}{8-1}} = 0,9.$$

Применив формулу (1.6) или возведя в квадрат стандартное отклонение, вычислим дисперсию: $6^2 = 0,9^2 = 0,81$.

8. **Коэффициент вариации (*V*).** Параметр, который показывает какую долю от среднего арифметического значения, выраженную в процентах, составляет стандартное отклонение (1.7).

$$V = \frac{6}{\bar{X}}100\%, \quad (1.7)$$

где Б – стандартное отклонение;

$\bar{X}$ – среднее арифметическое значение.

Рассчитаем стандартное отклонение для примера из таблицы 1.1, подставив в формулу (1.7) известные значения Б и $\bar{X}$.

$V = \frac{0,9}{43}100\% = 2,1\%$.

9. **Стандартная ошибка среднего (SD_x)** показывает насколько может отклоняться среднее значение выборки, состоящей из определенного количества объектов исследования (n), извлекаемых из генеральной совокупности (теоретически представленной всем множеством таких объектов) от среднего значения генеральной совокупности (то есть высчитанного по всему множеству объектов). Вычисляется по формуле (1.8).

$$SD_x = \frac{6}{\sqrt{n}}, \quad (1.8)$$

где Б – стандартное отклонение;
n – количество объектов исследования.

Обычно стандартная ошибка среднего записывается сразу после среднего значения отделяясь от него знаком ±. Запись имеет вид: $\bar{X} \pm$.

Рассчитаем стандартную ошибку среднего для примера из таблицы 1.1, подставив в формулу (1.8) известные значения Б и n.

$SD_x = \frac{0,9}{\sqrt{8}} = 0,3$.

Записываем ответ следующим образом: 43±0,3.

Задания к разделу 1 для самостоятельного выполнения

1. Произведите расчет среднего арифметического для ряда данных: 1,5; 2,7; 1,9; 2,3; 4,1; 3,2.
2. Рассчитайте среднее линейное отклонение для ряда данных: 1,5; 2,7; 1,9; 2,3; 4,1; 3,2).
3. Укажите моду в ряде данных: 1; 2; 3; 4; 5; 6; 7; 8 ; 9; 2.
4. Чему равна медиана ряда данных: 1; 2; 3; 4; 5; 6; 7; 8 ; 9?
5. Рассчитайте стандартное отклонение для ряда данных: 1,5; 2,7; 1,9; 2,3; 4,1; 3,2.
6. Рассчитайте коэффициент вариации для ряда данных: 1,5; 2,7; 1,9; 2,3; 4,1; 3,2.
7. Рассчитайте ошибку среднего для ряда данных: 1,5; 2,7; 1,9; 2,3; 4,1; 3,2.

2. Индексы биоразнообразия

Биоразнообразие (биологическое разнообразие) – одно из ключевых понятий современной экологии. Биоразнообразие подразумевает разнородность и многообразие всех форм жизни на разных уровнях ее организации и оценивается путем вычисления разнообразных индексов. Ниже приведены формулы и описание наиболее популярных индексов разнообразия.

Индекс Шеннона-Уивера *H* [Shannon, 1948; Shannon, Weaver, 1949] является одним из наиболее часто используемых индексов разнообразия и вычисляется по формуле (2.1):

$$H = -\sum \frac{n_i}{N} \ln \frac{n_i}{N}, \quad (2.1)$$

где n_i – общая численность вида или внутривидовой разновидности;
N – общая численность отмеченных особей.

Следует отметить, что в формуле данного индекса применяют логарифмы с различным основанием ($\log_2 X$; lg и др.).

Индекс выравненности Пиелу *E* [Pielou, 1975] вычисляется на основании индекса Шеннона-Уивера по формуле (2.2):

$$E = \frac{H}{\ln S}, \quad (2.2)$$

где H –индекс Шеннона-Уивера;
S – число отмеченных в водном объекте видов.

При вычислении индеса Пиелу применяется логарифм с тем же основанием, что был применен при вычислении индекса Шеннона-Уивера.

Индекс (Мера доминирования *C*) Симпсона [Simpson, 1949] вычисляется по формуле (2.3):

$$C = \sum p_i^2 = \sum \left(\frac{n_i}{N}\right)^2, \quad (2.3)$$

где C – концентрация доминирования (Мера доминирования Симпсона);
p_i – относительная значимость (доля вида);
n_i – общая численность особей вида или внутривидовой разновидности;
N – общая численность отмеченных особей.

Индекс разнообразия Симпсона *D* вычисляется по формуле (2.4):

$$D = \frac{1}{C}, \quad (2.4)$$

где C – мера доминирования (индекс Симпсона).

Индекс Маргалефа *d* [Margalef, 1958] рассчитывается по формуле (2.5):

$$d = \frac{(S-1)}{\ln N}, \quad (2.5)$$

где d – индекс видового богатства;
S – количество видов и внутривидовых таксонов;
N – общее количество зафиксированных особей.

Индекс видового разнообразия Менхиника D_{mn} [Menhinick, 1964] вычисляется по формуле (2.6):

$$D_{mn} = \frac{S}{\sqrt{N}}, \quad (2.6)$$

где S – количество видов и внутривидовых таксонов;
N – общее количество зафиксированных особей.

Индекс Животовского μ [Животовский, 1980] рассчитывается по формуле (2.7):

$$\mu = \left(\sum \sqrt{p_i}\right)^2, \quad (2.7)$$

где p_i – относительная значимость (доля вида).

Доля редких видов *h* рассчитывается по формуле (2.8):

$$h = 1 - \frac{\mu}{S}, \quad (2.8)$$

где μ – индекс Животовского;
S – количество видов и внутривидовых таксонов.

Все указанные выше индексы характеризуются своими особенностями. Например, такой показатель как индекс Шеннона-Уивера позволяет учитывать одновременно и видовое богатство, и количественные различия между видами. Индекс Пиелу (E) указывает, насколько относительная численность особей при данном количестве видов распределена в сообществе равномерно. Низкие значения показателя свидетельствуют о дисбалансе, демонстрирующем наличие таксонов, резко отличающихся по количеству особей. Мера доминирования Симпсона (индекс Симпсона) позволяет оценить, насколько равномерно распределены доли отдельных таксонов в сообществе. Высокие значения параметра указывают на дисбаланс в пользу численности небольшого количества видов. Мера доминирования принимает большие значения в экосистемах с ярко выраженными доминантами (то есть при наличии видов с большим количеством особей). Индекс Маргалефа характеризуется особенностью: его значения тем выше, чем выше количество видов и ниже количество особей. Низкие значения индекса, свидетельствуют об относительно малом количестве видов на фоне относительно большого количества особей. Особенностями характеризуются и другие индексы разнообразия. Для лучшего понимания того, как рассчитываются индексы, и как они соотносятся друг с другом, приведем пример их расчета.

Пример расчета индексов разнообразия. Имеются сведения о количестве особей различных видов диатомовых водорослей (таблица 2.1) в пробе воды озерной экосистемы (для простоты расчета количество видов и количество особей сокращено). Необходимо произвести расчет всех индексов разнообразия.

Таблица 2.1

Виды водорослей озера Радуга

№	Название вида	Озеро Радуга
1	*Achnanthidium minutissimum* (Kütz.) Czarnecki	13
2	*Staurosira venter* (Ehr.) Cleve & Möller	5
3	*Staurosirella pinnata* Ehr	9
4	*Tabellaria fenestrata* (Lungb.) Kutz.	12
5	*Tabellaria flocculosa* (Roth.) Kutz.	8

1) Рассчитаем индекса Шеннона-Уивера.

Сначала произведем расчет *N*: $N = 13 + 5 + 9 + 12 + 8 = 47$.

Подставим имеющиеся данные в формулу (2.1):

$H = -\left(\frac{13}{47}\ln\frac{13}{47} + \frac{5}{47}\ln\frac{5}{47} + \frac{9}{47}\ln\frac{9}{47} + \frac{12}{47}\ln\frac{12}{47} + \frac{8}{47}\ln\frac{8}{47}\right)$=1,6.

2) Произведем расчет индекса выравненности Пиелу по формуле (2.2): $E = \frac{1,6}{\ln 5}$=0,97.

3) Рассчитаем меру доминирования Симпсона по формуле (2.3):

C=$\left(\frac{13}{47}\right)^2 + \left(\frac{5}{47}\right)^2 + \left(\frac{9}{47}\right)^2 + \left(\frac{12}{47}\right)^2 + \left(\frac{8}{47}\right)^2 = 0,22$.

4) Произведем вычисление индекса разнообразия Симпсона по формуле (2.4): $D = \frac{1}{0,22}$=4,6.

5) Рассчитаем индекс Маргалефа по формуле (2.5): $d = \frac{(5-1)}{\ln 47} = 1,04$.

6) Произведем вычисление индекс видового разнообразия Менхиника по формуле (2.6): $D_{mn} = \frac{5}{\sqrt{47}}$ = 0,73.

7) Рассчитаем индекс Животовского по формуле (2.7):

μ=$\left(\sqrt{\frac{13}{47}} + \sqrt{\frac{5}{47}} + \sqrt{\frac{9}{47}} + \sqrt{\frac{12}{47}} + \sqrt{\frac{8}{47}}\right)^2$ =4,9.

8) Определим долю редких видов по формуле (2.8):

$h = 1 - \frac{4,9}{5} = 0,03$.

Задания к разделу 2 для самостоятельного выполнения

По исходным количественным данным, приведенным в таблице 2.2, произвести расчет всех индексов разнообразия (указанных в разделе 2) для любых 3 озер.

Таблица 2.2

Таксономический состав диатомовых водорослей различных озер

Название вида	Озеро 1	Озеро 2	Озеро 3	Озеро 4	Озеро 5	Озеро 6	Озеро 7
Achnanthidium minutissimum (Kütz.) Czarnecki	53	12	48	5	3	45	
Staurosira venter (Ehr.) Cleve & Möller	48			2			
Staurosirella pinnata Ehr	101				45		
Tabellaria fenestrata (Lungb.) Kutz.	56	35					
Tabellaria flocculosa (Roth.) Kutz.	50		23			212	
Navicula cryptocephala Kützing	12			15		56	
Pinnularia major (Kützing) Rabenhorst	18		23				5
Gomphonema acuminatum Ehrenberg	23	12		5	45		
Eunotia praerupta Ehrenberg	78				45		55
Ellerbeckia arenaria (Moore ex Ralfs) R.M.Crawford	56		35	21			

3. Сходство таксономического состава экосистем

На практике часто возникает необходимость оценить насколько похожи различные экосистемы по таксономическому составу организмов. Например, исследователь имеет списки видов встречающихся в различных экосистемах и задается вопросом, насколько сильно таксономический состав данных экосистем схож. В

этом случае используются коэффициенты таксономического сходства. Существуют различные коэффициенты сходства. Наиболее часто используются несколько из них: коэффициент Жаккара, коэффициент Съеренсена и коэффициент Брея-Кертиса. Все указанные коэффициенты являются парными, то есть рассчитываются по таксономическим спискам пар (2-х) экосистем.

Коэффициент Жаккара (K_j) вычисляется по формуле (3.1) [Jaccard, 1901].

$$K_j = \frac{c}{a+b-c}, \quad (3.1)$$

где a – количество (учтенных) видов первой экосистемы (или пробной площадки, территории, пробы и др.);
b – количество видов второй экосистемы;
c – количество общих для 1-ой и 2-ой экосистем видов.

Коэффициент Серенсена (K_s) рассчитывается по формуле (3.2) [Sørensen, 1948].

$$K_s = \frac{2c}{a+b}, \quad (3.2)$$

где обозначения в формуле соответствуют таковым, указанным для коэффициента Жаккара.

Коэффициент Брея-Кертиса (K_{B-C}) рассчитывается по формуле (3.3) [Bray, Curtis, 1957]:

$$K_{B-C} = \frac{2\sum N_{min}}{N_a+N_b}, \quad (3.3)$$

где N_a – общая сумма количественных показателей (например, общее количество учтенных особей) первой экосистемы;
N_b – общая сумма количественных показателей второй экосистемы;
$\sum N_{min}$ – сумма наименьших значений количественных показателей (для каждого таксона, встреченного в обеих экосистемах, выбирается наименьший количественный показатель, то есть показатель только

одной экосистемы. Далее такие минимальные значения всех видов суммируются).

Указанные коэффициенты сходства принимают значения от 0 до 1, где 0 – это полное отсутствие таксономического сходства, 1 – это полное сходство. Коэффициенты Жаккара и Серенсена учитывают только факт встречаемости таксонов в экосистемах без учета количественных соотношений между ними. Коэффициент Брея-Кертиса позволяет учитывать не только наличие общих таксонов экосистем, но и количественные соотношения между ними.

Для того, чтобы лучше понять, для чего применяются отмеченные выше коэффициенты, как они рассчитываются и чем отличаются, приведем пример.

Рассчитаем таксономическое сходство 2-х озерных экосистем (таблица 3.2).

Таблица 3.2

Таксономических состав диатомовых водорослей 2-х озерных экосистем

№	Название таксона (вида)	Озеро 1	Озеро 2
1	*Achnanthidium minutissimum* (Kütz.) Czarnecki	53	12
2	*Staurosira venter* (Ehr.) Cleve & Möller	48	
3	*Staurosirella pinnata* Ehr	101	
4	*Tabellaria fenestrata* (Lungb.) Kutz.	56	35
5	*Tabellaria flocculosa* (Roth.) Kutz.	50	
6	*Navicula cryptocephala* Kützing	12	5
7	*Pinnularia major* (Kützing) Rabenhorst	18	
8	*Gomphonema acuminatum* Ehrenberg	23	24
9	*Eunotia praerupta* Ehrenberg	78	
10	*Ellerbeckia arenaria* (Moore ex Ralfs) R.M.Crawford	56	3

**В столбцах 3 и 4 указаны количества отмеченных в пробах озер особей*

Пример расчета коэффициента Жаккара для таблицы 3.2. Из 10 отмеченных в 1-ой экосистеме видов, во второй имеется 5, то есть количество общих видов равно 5. Следовательно расчеты принимают следующий вид: $K_j = \frac{5}{10+5-5} = 0{,}5$.

Пример расчета коэффициент Серенсена для таблицы 2:

$K_s = \frac{2*5}{10+5} = 0{,}67.$

Пример расчета коэффициента Брея-Кертиса.

$N_a = 53 + 48 + 101 + 56 + 50 + 12 + 18 + 23 + 78 + 56 = 495;$

$N_b = 12 + 35 + 5 + 24 + 3 = 79;$

$\sum N_{min} = 12 + 35 + 5 + 23 + 3 = 78;$

$K_{B-C} = \frac{2*78}{495+79} = 0{,}27.$

Если перед исследователем стоит задача рассчитать таксономическое сходство большого количества экосистем, то для удобства представления информации результаты вычислений заносятся в специальную таблицу «Матрицу таксономического сходства». В первом столбце и первой строчке указанной таблицы записываются названия экосистем (условные обозначения, номера, имена собственные и др.), в остальных ячейках – вычисленные для пар экосистем коэффициенты таксономического сходства. Для лучшего усвоения материала построим условную таблицу таксономического сходства (таблица 3.3).

Таблица 3.3

Матрица таксономического сходства 5 условных экосистем

	Экосистема 1	Экосистема 2	Экосистема 3	Экосистема 4	Экосистема 5
Экосистема 1	1	0,3	0,2	0,15	0,55
Экосистема 2	0,3	1	0,25	0,2	0,4
Экосистема 3	0,2	0,25	1	0,6	0,2
Экосистема 4	0,15	0,2	0,6	1	0,2
Экосистема 5	0,55	0,4	0,2	0,2	1

Таксономическое сходство пар экосистем записано в ячейках, расположенных на пересечении воображаемых перпендикуляров, опускаемых от соответствующих названий. Для одной из экосистем выбирается название из первого столбца матрицы, для второй из первой строчки. Легко обратить внимание, что на пересечении перпендикуляров, проводимых от названий одной и той же экосистемы (например, от «Экосистема 1» из первых столбца и стороки), расположены цифры «1», что свидетельствует о полном сходстве таксономического состава. Таким образом, сходство таксономического

состава экосистемы 1 и экосистемы 2, определяемое по таблице 3.3, составляет 0,3; экосистемы 3 и экосистемы 5 – 0,2; экосистемы 4 и экосистемы 2 – 0,2 и т.д.

Задания к разделу 3 для самостоятельного выполнения

По исходным количественным данным, приведенным в таблице 3.4, произвести расчет всех индексов сходства для 5 учетных площадок. Ориентироваться на пример расчетов, приведенных в разделе 3. Результаты представить в виде матрицы таксономического сходства (смотреть таблицу 3.3).

Таблица 3.4

Количество учтенных особей грызунов различных учетных площадок

Название вида	Учетная площадка 1	Учетная площадка 2	Учетная площадка 3	Учетная площадка 4	Учетная площадка 5
Обыкновенная белка	12		25	7	8
Жёлтый суслик		25	12		
Малый суслик	32			12	12
Сурок-байбак			5		
Соня-полчок	12		12	14	15
Степная мышовка		20			
Лесная мышовка	22		5	7	9
Большой тушканчик	12	22			11
Ондатра			12	12	
Обыкновенная полёвка	17	31			7
Мышь-малютка	22		8	9	
Полуденная песчанка	14		11		10

**В столбцах 3-5 указаны количества отмеченных особей*

4. Проверка данных на соответствие закону нормального распределения

В данном разделе приведена краткая обобщенная информация о нормальном распределении данных и наиболее простых способах проверки соответствия данных закону нормального распределения (в идеализированном случае). Раздел позволяет получить только наиболее общие представления «о нормальности». Более подробно с законом нормального распределения можно ознакомиться, применив специализированную учебную литературу [Полякова, Шаброва, 2015; Елисеева, 2011; Калинина, 2013; Гмурман, 2013; Лысенко, Дмитриева, 2013 и др.].

В природе все находится в гармонии. Те или иные признаки каких-либо объектов природы, как правило, проявляются согласно следующему закону: типичными (средними) значениями признака характеризуется большая часть объектов, по мере отклонения от среднего значения признака – количество объектов, характеризующихся такими значениями признаками, постепенно сокращается. Чем больше характеристика объектов отклонена от нормы (от среднего) тем более он уникален. Тем меньше вероятность среди всего многообразия найти данный объект.

Простой пример: распределение роста у человека. Большинство людей обладают средним ростом. В то время как люди небольшого роста и высокие люди встречаются значительно реже.

Подобным образом обстоит дело и с большинством других признаков у объектов живой и неживой природы. Такое «типичное» распределение данных получило название «нормального».

Множество статистических методов основаны на использовании особенностей нормального распределения данных. Такие методы принято называть «параметрическими». К этой группе статистических методов из рассматриваемых в данном пособии относятся: расчет t-критериев Стьюдента и определение значений коэффициента линейной корреляции Пирсона. Для того, чтобы использовать данные методы сначала нужно проверить распределение значений исходных

характеристик. В случае, если закон распределения не близок «нормальному», то параметрические методы не могут быть использованы.

Методы, для которых соблюдение закона нормального распределения является необязательным, получили название «непараметрические». Непараметрическими методами из рассматриваемых в учебнике являются: t-критерий Уилкоксона; U-критерий Манна-Уитни; коэффициент ранговой корреляции Спирмена и методы кластерного анализа.

Как же осуществить проверку данных на соответствие (близость) закону нормального распределения? Существуют различные методы. В целом они могут быть подразделены на 2 группы: графические и расчетные. С использованием современного программного обеспечения и стандартных офисных компьютеров проверка нормальности распределения не составляет большого труда и занимает считанные секунды. В Excel, Statistica, SPSS и других программах существует широкий инструментарий для проверки распределения исследуемых данных: построение графиков распределения (гистограммы, нормальные вероятностные графики, ящичные диаграммы с усами), вычисление формальных критериев (Колмогорова-Смирнова, Шапиро-Уилка) и прочих характеристик распределения. В данном разделе будут рассмотрены наиболее простые методы проверки на нормальность, которые можно применить без использования специализированного программного обеспечения.

Графический способ. Среди всего множества методов наиболее распространенным и простым является метод (графический) построения диаграммы распределения данных. Суть метода заключается в том, чтобы имеющиеся значения исследуемых характеристик разбить на ряд равных диапазонов. Далее нужно посчитать количество объектов исследования вошедших в эти диапазоны и построить гистограмму, где по оси OX следует отложить указанные диапазоны значений (в порядке возрастания), а по оси OY количество объектов исследования. Если указанная диаграмма будет иметь колоколообразный вид и будет симметричной (или близкой к

таковой), то считают, что распределение близко нормальному. При этом значение среднего арифметического и медианы должны быть также близкими (и моды в идеальных случаях). Такой графический способ применим в тех случаях, когда выборка представлена достаточно большим количеством объектов исследования (от нескольких десятков и больше).

Приведем пример графического метода определения соответствия распределения данных закону нормального распределения. Пусть имеются сведения о росте группы людей, состоящей из 30 человек (таблица 4.1).

Таблица 4.1

Рост группы людей

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Рост, см	147	151	152	156	157	157	161	162	162	163	164	165	166	166	167	167	167	168	169	171	171	172	172	173	176	177	178	182	183	191

Рост людей группы варьирует в диапазоне от 147 до 191 см. Произведем разбивку значений роста на удобное количество диапазонов. Указанных диапазонов должно быть достаточное количество, в противном случае диаграммы не дадут представления о характере распределения данных. Оптимальным количеством (на наш взгляд) является 9-15 диапазонов. В случае с таблицей 4.1 удобно использовать диапазон по 5 см. Таким образом, получится 9 диапазонов равной ширины (по 5 см).

Производим подсчет количеств объектов исследования (людей) вовлеченных в каждый диапазон. Отобразим полученный результат в виде таблицы 4.2.

Таблица 4.2

Группы объектов исследования по значениям роста

Группы	Частота (количество людей)
<150	1
[150;155)	2
[155;160)	3

[160;165)	5
[165;170)	8
[170;175)	5
[175;180)	3
[180;185)	2
≥185	1

Используя данные таблицы 4.2 построим гистограмму (рисунок 4.1).

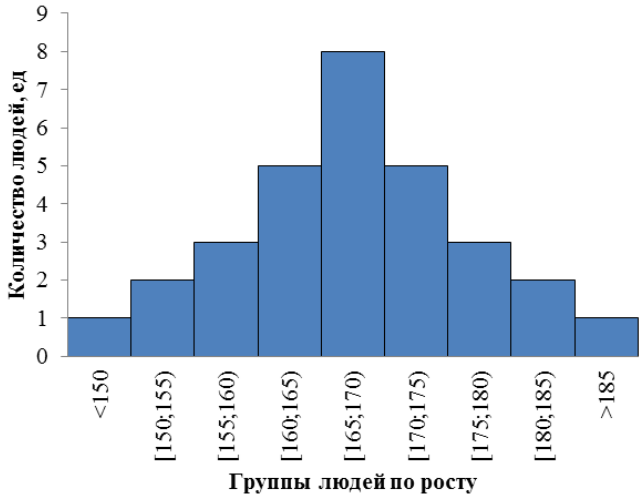


Рисунок 4.1. Гистограмма распределения объектов исследования

Гистограмма распределения объектов исследования близка колоколообразной форме, симметрична. Таким образом, распределение близко нормальному. Далее необходимо вычислить среднее арифметическое, медиану и моду.

Расчет среднего ($\bar{X}$), медианы (*Me*) и моды (*Mo*):

$$\bar{X} = \frac{(147+151+152+156+157+161+162+162+163+164+165+166+166+ +167+167+168+169+167+171+171+172+173+176+177+178+ +182+183+191)}{30} = 167{,}1.$$

$$Me = \frac{167+167}{2} = 167.$$

Mo =167 (наиболее часто повторяющееся число).

Из расчетов видно, среднее арифметическое, медиана и мода практически равны, что также является одним из косвенных признаков нормального распределения данных.

Расчет асимметрии и эксцесса. Для проверки распределения на нормальность также используют метод расчета асимметрии и эксцесса. Для нормального распределения асимметрия и эксцесс равны 0. В реальности имеются некоторые отклонения от 0. Если это отклонение невелико, то можно считать, что распределении близко нормальному.

Рассчитывается асимметрия (A) по формуле (4.1):

$$A = \frac{\sum(x_i - \bar{X})^3}{n6^3}, \quad (4.1)$$

где А – асимметрия;

x_i – значения рассматриваемой характеристики в каждом конкретном случае;

$\bar{X}$ – среднее арифметическое значение;

σ – стандартное отклонение;

n – количество объектов исследования.

Эксцесс (E) рассчитывается по формуле (4.2):

$$E = \frac{\sum(x_i - \bar{X})^4}{n6^4} - 3. \quad (4.2)$$

После расчетов асимметрии и эксцесса полученные их значения сравниваются с показателями, именуемыми ошибками репрезентативности асимметрии (m_A) и эксцесса (m_E), которые вычисляются согласно формулам (4.3) и (4.4).

$$m_A = \sqrt{\frac{6}{n}}. \quad (4.3)$$

$$m_E = \sqrt{\frac{6}{n}}2. \quad (4.4)$$

Если сравниваемые значения асимметрии и эксцесса по модулю (без учета знаков ±) меньше трехкратного значения соответствующих ошибок репрезентативности, то распределение близко нормальному. Если нет – то распределение не соответствует нормальному.

Произведем расчеты асимметрии и эксцесса для рассмотренного выше примера. Для использования формул асимметрии и эксцесса необходимо помимо среднего, которое уже рассчитано, вычислить стандартное отклонение (σ).

Формулы стандартного отклонения, асимметрии и эксцесса содержат общую часть $(x_i - \bar{X})$. Для того, чтобы облегчить и упорядочить вычисления всех этих параметров воспользуемся табличной формой записи результатов вычислительных процедур (таблица 4.3).

Таблица 4.3

Упорядоченное представление расчетов

№	x_i	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$	$(x_i - \bar{X})^3$	$(x_i - \bar{X})^4$
1	2	3	4	5	6
1	147	$(147 - 167{,}1) = -20{,}1$	404,0	–8120,6	163224,1
2	151	$(151 - 167{,}1) = -16{,}1$	259,2	–4173,3	67189,8
3	152	$(152 - 167{,}1) = -15{,}1$	228,0	–3443,0	51988,6
4	156	$(156 - 167{,}1) = -11{,}1$	123,2	–1367,6	15180,7
5	157	$(157 - 167{,}1) = -10{,}1$	102,0	–1030,3	10406,0
6	157	$(157 - 167{,}1) = -10{,}1$	102,0	–1030,3	10406,0
7	161	$(161 - 167{,}1) = -6{,}1$	37,2	–227,0	1384,6
8	162	$(162 - 167{,}1) = -5{,}1$	26,0	–132,7	676,5
9	162	$(162 - 167{,}1) = -5{,}1$	26,0	–132,7	676,5
10	163	$(163 - 167{,}1) = -4{,}1$	16,8	–68,9	282,6
11	164	$(164 - 167{,}1) = -3{,}1$	9,6	–29,8	92,4
12	165	$(165 - 167{,}1) = -2{,}1$	4,4	–9,3	19,4
13	166	$(166 - 167{,}1) = -1{,}1$	1,2	–1,3	1,5
14	166	$(166 - 167{,}1) = -1{,}1$	1,2	–1,3	1,5
15	167	$(167 - 167{,}1) = -0{,}1$	0,01	–0,001	0,0001
16	167	$(167 - 167{,}1) = -0{,}1$	0,01	–0,001	0,0001
17	167	$(167 - 167{,}1) = -0{,}1$	0,01	–0,001	0,0001
18	168	$(168 - 167{,}1) = 0{,}9$	0,8	0,7	0,7

19	169	$(169-167{,}1)=1{,}9$	3,6	6,9	13,0
20	171	$(171-167{,}1)=3{,}9$	15,2	59,3	231,3
21	171	$(171-167{,}1)=3{,}9$	15,2	59,3	231,3
22	172	$(172-167{,}1)=4{,}9$	24,0	117,6	576,5
23	172	$(172-167{,}1)=4{,}9$	24,0	117,6	576,5
24	173	$(173-167{,}1)=5{,}9$	34,8	205,4	1211,7
25	176	$(176-167{,}1)=8{,}9$	79,2	705,0	6274,2
26	177	$(177-167{,}1)=9{,}9$	98,0	970,3	9606,0
27	178	$(178-167{,}1)=10{,}9$	118,8	1295,0	14115,8
28	182	$(182-167{,}1)=14{,}9$	222,0	3307,9	49288,4
29	183	$(183-167{,}1)=15{,}9$	252,8	4019,7	63912,9
30	191	$(191-167{,}1)=23{,}9$	571,2	13651,9	326280,9

Далее находим значения $\sum(x_i-\bar{X})^2$; $\sum(x_i-\bar{X})^3$ и $\sum(x_i-\bar{X})^4$, которые будут равны соответственно суммам всех значений столбиков 4, 5 и 6 таблицы 4.3.

$\sum(x_i-\bar{X})^2=404+259{,}2+228+123{,}2+102+\cdots+222+252{,}8+$
$+571{,}2=2800{,}7.$

$\sum(x_i-\bar{X})^3=-8120{,}6+(-4173{,}3)+(-3443)+\cdots+4019{,}7+$
$+13651{,}9=4748{,}8.$

$\sum(x_i-\bar{X})^4=163224{,}1+67189{,}8+51988{,}6+\cdots+63912{,}9+$
$+326280{,}9=793849{,}5.$

Далее определяем значение стандартного отклонения (σ):

$6=\sqrt{\frac{2800{,}7}{30-1}}$=9,8.

Рассчитываем асимметрию (A) подставив все известные значения в формулу (4.1): $A=\frac{4748{,}8}{30*9{,}8^3}$=0,17.

Рассчитаем эксцесс (E) подставив все известные значения в формулу (4.2): $E=\frac{793849{,}5}{30*9{,}8^4}-3=-0{,}16.$

Далее вычислим ошибки репрезентативности асимметрии (m_A) и эксцесса (m_E) по формулам (4.3) и (4.4) соответственно:

$m_A=\sqrt{\frac{6}{30}}=0{,}45.$

$m_E = \sqrt{\frac{6}{30}} * 2 = 0{,}9.$

Так как $|A| < 3 * m_A$ и $|E| < 3 * m_E$, то распределение близко нормальному.

В целом, проверка на соответствие закону нормального распределения требует достаточно большого количества вычислений, особенно если выборка состоит из десятков объектов исследования. Существующие компьютерные программы, о которых упоминалось ранее, позволяют проверить распределение данных за несколько минут. Поэтому проверка данных на соответствие закону нормального распределения должна осуществляться с большей надежностью не по одному критерию (графическому или расчетному), а по их совокупности.

Задания к разделу 4 для самостоятельного выполнения

Произвести расчет асимметрии и эксцесса для данных, представленных в таблице 4.4. По результатам расчетов сделать вывод о соответствии распределения данных нормальному.

Таблица 4.4

Концентрация свинца в пробах почв, различных экосистем

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Pb, мг/кг	9,3	9,0	9,2	8,5	3,4	2,5	3,3	4,1	6,7	7,2	5,3	5,4	8,1	5,8	9,1	1,5	4,7	6,3	5,2	7,9	4,9	5,8	2,8	3,2	5,3	7,8	7,2	7,3	6,5	5,9

5. Взаимосвязь характеристик объектов исследования

Очень часто на практике возникает необходимость узнать связаны ли характеристики каких-либо предметов, объектов или явлений между собой. Для этих целей используется корреляционный анализ, то есть процедура вычисления коэффициентов корреляции.

В данном разделе будет рассмотрено 2 наиболее широко применяемых коэффициента корреляции: коэффициент линейной корреляции Пирсона [Pearson, 1895] и коэффициент ранговой корреляции Спирмена [Spearman,1904]. В независимости от того, какой коэффициент корреляции используется (так как отличаются только формулы расчета) необходимо руководствоваться следующим порядком действий:

А. Расчет значения коэффициента корреляции по формуле (5.2) (коэффициент Пирсона) или по формуле (5.3) (коэффициент Спирмена).

Б. Сравнение расчетного значения с табличным (таблица 5.1 для коэффициента Пирсона и таблица 5.2 для коэффициента Спирмена).

Г. Если значение расчетного коэффициента по модулю (без учета знака) равно или выше значений табличного показателя – то рассчитанный коэффициент корреляции является статистически значимым, если наоборот – то рассчитанный коэффициент корреляции не является статистически значимым.

Д. В зависимости от знака коэффициента корреляции определяют направленность связи (положительная или отрицательная) между характеристиками, в зависимости от значения – силу связи (связь сильная, средняя или слабая).

Коэффициенты корреляции могут принимать значения от -1 до 1. Значения коэффициента корреляции указывают на характер (знак коэффициента) и силу связи (числовое значение) между переменными.

Сила связи переменных. При значениях коэффициента корреляции по модулю (т.е. без учета знака) равных значениям меньшим 0,5 связь считают слабой (незначительной). В таких случаях, как правило, считают, что связь между характеристиками не выражена

или отсутствует. Если коэффициент корреляции равен или превышает 0,5, но меньше или равен 0,7 – то связь средняя. При значениях больших 0,7 (до 1,0) – связь сильная. При установлении силы связи на знак (+ или –) коэффициента внимания не обращают.

Знак коэффициента корреляции указывает на характер взаимосвязи. Если коэффициент корреляции положительный, то увеличение одной характеристики сопровождается увеличением второй характеристики. Если знак отрицательный, то увеличение одного из пары параметров сопровождается уменьшением значений второго.

Проверка статистической значимости коэффициента корреляции. Важным условием при вычислении любого коэффициента корреляции является проверка его статистической значимости. В случае если коэффициент корреляции не окажется статистически значимым, его использование для дальнейших работ недопустимо. Для расчета статистической значимости коэффициентов корреляции необходимо использовать специальные стандартные таблицы «Критические значения коэффициента корреляции (Пирсона или Спирмена)». Данные таблицы можно найти в Интернете или в специализированной литературе (таблицы 5.1 и 5.2).

Таблица 5.1

Фрагмент таблицы «Критические значения коэффициента корреляции Пирсона» [Table of…, 2019]

Число степеней свободы, $k=n-2$	при p=0,05
5	0,75
6	0,71
7	0,67
8	0,63
9	0,6

Для использования таблицы 5.1 необходимо рассчитать k – число степеней свободы, значения данного параметра вписаны в столбец 1 таблицы 5.1. k рассчитывается по формуле (5.1).

$$k = n - 2, \quad (5.1)$$

k – число степеней свободы;
n – количество объектов исследования.

Таблица 5.2

Фрагмент таблицы «Критические значения коэффициента корреляции Спирмена» [Суходольский, 1988]

m	при p=0,05
5	0,94
6	0,85
7	0,78
8	0,72
9	0,68

В таблице 5.2 m – это количество объектов исследования. Во втором столбце таблиц 5.1 и 5.2 обозначен уровень статистической значимости «p». В большинстве экологических исследований принят уровень статистической значимости равный 5 % (или 0,05). Существуют и более строгие уровни: 1 % (0,01), 0,1 % (0,001) и др. Расчетный коэффициент корреляции считается статистически значимым при соответствующем уровне статистической значимости (например, при p=0,05), если он равен значениям или превышает значения коэффициента из таблицы критических значений.

Коэффициент линейной корреляции Пирсона. Вычисляется по формуле (5.2).

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}, \quad (5.2)$$

где x_i – конкретные значения переменной X;
$\bar{X}$ – среднее арифметическое значение переменной X;
y_i – конкретные значения переменной Y;
$\bar{Y}$ – среднее арифметическое значение переменной Y.

Коэффициент корреляции Спирмена схож с коэффициентом Пирсона, как в интерпретации, так и по диапазону значений, которые он принимает. Различия кроются в том, что коэффициент Спирмена рассчитывается не по исходным данных, а по рангам (целым натуральным числам: 1, 2, 3 и т.д.), которые присваиваются исходным значениям в порядке возрастания или убывания. Плюс коэффициента Спирмена заключается еще и в том, что ранги можно присваивать и нечисловым параметрам, характеризующимся изменением степени своей интенсивности (цвету, успеваемости студентов, типам населенных пунктов по размеру и др.). Например, у исследователя стоит задача проранжировать концентрацию растворенного в воде различных озер кальция, которая соответствует следующим исходным значениям: 7,8 (первое озеро); 13,2; 14,1; 14,5 и 15,7 мг/л (пятое озеро). Ранги целесообразно присвоить исходным значениям следующим образом: 1 (первое озеро), 2, 3, 4 и 5 (пятое озеро). Расчет коэффициента Спирмена производится по формуле (5.3).

$$\rho = 1 - \frac{6\sum D_i^2}{m(m^2-1)}, \quad (5.3)$$

где D_i^2 – квадрат разности рангов пар наблюдений (характеристик);
m – количество пар рангов наблюдений.

Для лучшего усвоения материала произведем расчет коэффициентов корреляции на примере (таблица 5.3). Необходимо установить влияет ли удаление от трасы на концентрацию свинца в почве. Пример дан заведомо простой с целью научиться рассчитывать коэффициенты корреляции (числа в таблицах подобраны для удобства расчета «вручную»).

Таблица 5.3

Концентрация свинца в почве на различном удалении от автострады

№	Расстояние точки опробования от автострады, м	Концентрация свинца в почве, мг/кг
1	5	7,8
2	11	7,3
3	16	6,0

4	21	5,1
5	24	5,2
6	29	3,8
7	32	2,8
8	36	2,9

Расчет коэффициента линейной корреляции Пирсона. Итак, произведем расчет коэффициента Пирсона, применив формулу (5.2). Сначала необходимо вычислить средние значения характеристик для второго и третьего столбцов таблицы 5.3.

$\bar{X} = \frac{5+11+16+21+24+29+32+36}{8} = 21{,}8.$

$\bar{X}$ – среднее арифметическое расстояния точки опробования от автострады.

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

$\bar{Y}$ – среднее арифметическое концентрации свинца.

Далее найдем все разности $(x_i - \bar{X})$ и $(y_i - \bar{Y})$. Вычисления для удобства запишем в виде таблицы (таблица 5.4).

Таблица 5.4

Расчет разностей $(x_i - \bar{X})$ и $(y_i - \bar{Y})$

№	$(x_i - \bar{X})$	$(y_i - \bar{Y})$
1	$(5 - 21{,}8) = -16{,}8$	$(7{,}8 - 5{,}1) = 2{,}7$
2	$(11 - 21{,}8) = -10{,}8$	$(7{,}3 - 5{,}1) = 2{,}2$
3	$(16 - 21{,}8) = -5{,}8$	$(6{,}0 - 5{,}1) = 0{,}9$
4	$(21 - 21{,}8) = -0{,}8$	$(5{,}1 - 5{,}1) = 0{,}0$
5	$(24 - 21{,}8) = 2{,}2$	$(5{,}2 - 5{,}1) = 0{,}1$
6	$(29 - 21{,}8) = 7{,}2$	$(3{,}8 - 5{,}1) = -1{,}3$
7	$(32 - 21{,}8) = 10{,}2$	$(2{,}8 - 5{,}1) = -2{,}3$
8	$(36 - 21{,}8) = 14{,}2$	$(2{,}9 - 5{,}1) = -2{,}2$

Далее найдем произведение: $(x_i - \bar{X})(y_i - \bar{Y})$. Вычисления для удобства также будем использовать табличную запись (таблица 5.5).

Таблица 5.5

Расчет произведения $(x_i - \bar{X})(y_i - \bar{Y})$

№	$(x_i - \bar{X})(y_i - \bar{Y})$
1	$-16{,}8 * 2{,}7 = -45{,}4$
2	$-10{,}8 * 2{,}2 = -23{,}8$
3	$-5{,}8 * 0{,}9 = -5{,}2$
4	$-0{,}8 * 0{,}0 = 0{,}0$
5	$2{,}2 * 0{,}1 = 0{,}2$
6	$7{,}2 * (-1{,}3) = -9{,}4$
7	$10{,}2 * (-2{,}3) = -23{,}5$
8	$14{,}2 * (-2{,}2) = -31{,}2$

Найдем сумму $\sum(x_i - \bar{X})(y_i - \bar{Y})$:

$\sum(x_i - \bar{X})(y_i - \bar{Y}) = -45{,}4 + (-23{,}8) + (-5{,}2) + 0{,}0 + 0{,}2 + (-9{,}4) + +(-23{,}5) + (-31{,}2) = -138{,}2.$

Следующим шагом является нахождение $(x_i - \bar{X})^2$ и $(y_i - \bar{Y})^2$. Вычисления запишем в виде таблицы (таблица 5.6).

Таблица 5.6

Расчет $(x_i - \bar{X})^2$ и $(y_i - \bar{Y})^2$

№	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})^2$
1	$(-16{,}8)^2 = 282{,}2$	$2{,}7^2 = 7{,}3$
2	$(-10{,}8)^2 = 116{,}6$	$2{,}2^2 = 4{,}8$
3	$(-5{,}8)^2 = 33{,}6$	$0{,}9^2 = 0{,}8$
4	$(-0{,}8)^2 = 0{,}6$	$0{,}0^2 = 0{,}0$
5	$2{,}2^2 = 4{,}8$	$0{,}1^2 = 0{,}01$
6	$7{,}2^2 = 51{,}8$	$(-1{,}3)^2 = 1{,}7$
7	$10{,}2^2 = 104{,}0$	$(-2{,}3)^2 = 5{,}3$
8	$14{,}2^2 = 201{,}6$	$(-2{,}2)^2 = 4{,}8$

Далее найдем $\sum(x_i - \bar{X})^2$ и $\sum(y_i - \bar{Y})^2$:

$\sum(x_i - \bar{X})^2 = 282{,}2 + 116{,}6 + 33{,}6 + 0{,}6 + 4{,}8 + 51{,}8 + 104{,}0 + +201{,}6 = 795{,}5.$

$\sum(y_i - \bar{Y})^2 = 7{,}3 + 4{,}8 + 0{,}8 + 0{,}0 + 0{,}01 + 1{,}7 + 5{,}3 + 4{,}8 = 24{,}8.$

Следующий шаг – подставляем все вычисленные значения в формулу (5.2): $r = \frac{-138,2}{\sqrt{795,5*24,8}} = -0,98$.

Так как $|-0,98| > 0,7$, то корреляционная связь сильная (отрицательная).

Проверяем статистическую значимость, сравнивая расчетное значение коэффициента корреляции с табличным, используя таблицу «Критические значения коэффициента корреляции Пирсона» (таблица 5.1).

Для этого сначала вычислим k – число степеней свободы по формуле (5.1): $k = 8 - 2 = 6$.

Далее найдем табличное значение коэффициента Пирсона при k=6 и уровне статистической значимости p=0,05. Табличный коэффициент Пирсона равен 0,71.

Так как $|-0,98| > 0,71$, то связь является статистически значимой при p=0,05.

Таким образом, концентрация свинца в почве (в нашем конкретном случае) связана с расстоянием до автотрассы. По мере удаления от трассы концентрация свинца снижается, на что указывает сильная отрицательная статистически значимая связь.

Пример расчета коэффициента Спирмена. Вычислим для указанного выше примера коэффициент ранговой корреляции Спирмена, используя формулу (5.3). Сначала необходимо всем значениям характеристики 1 (расстояние точки опробования от автострады) и характеристики 2 (концентрация свинца в почве) присвоить ранги. Для удобства представления данных будем использовать табличную запись (таблица 5.7). Ранги (это натуральные числа от 1 по возрастающей) в каждом случае присваиваются значениям от меньшего к большему.

Дальнейшие расчеты осуществляются с применением присвоенных рангов. Далее найдем разность пар рангов D_i и квадрат разности пар рангов D_i^2 (таблица 5.8).

Таблица 5.7

Ранги, присвоенные исследуемым характеристикам

№	Характеристика 1	Ранги для характеристики 1	Характеристика 2	Ранги для характеристики 2
	Расстояние точки опробования от автострады, м		Концентрация свинца в почве, мг/кг	
1	5	1	7,8	8
2	11	2	7,3	7
3	16	3	6,0	6
4	21	4	5,1	4
5	24	5	5,2	5
6	29	6	3,8	3
7	32	7	2,8	1
8	36	8	2,9	2

Таблица 5.8

Разность рангов

№	Ранги для характеристики 1	Ранги для характеристики 2	D_i	D_i^2
1	1	8	$1-8=-7$	49
2	2	7	$2-7=-5$	25
3	3	6	$3-6=-3$	9
4	4	4	$4-4=0$	0
5	5	5	$5-5=0$	0
6	6	3	$6-3=3$	9
7	7	1	$7-1=6$	36
8	8	2	$8-2=6$	36

Найдем сумму квадратов разности пар рангов $\sum D_i^2$:

$\sum D_i^2 = 49+25+9+0+0+9+36+36 = 164.$

Далее подставляем все известные сведения в формулу (5.3):

$\rho = 1-\frac{6*164}{8(8^2-1)} = -0{,}95.$

Так как $|-0{,}95| > 0{,}7$, то корреляционная связь сильная (отрицательная).

Проверяем статистическую значимость сравнивая расчетное значение коэффициента корреляции с табличным, используя таблицу «Критические значения коэффициента корреляции Спирмена» (таблица 5.2).

При m=8 в и уровне статистической значимости p=0,05 табличный коэффициент Спирмена равен 0,72.

Так как $|-0{,}95| > 0{,}72$, то связь является статистически значимой при p=0,05.

Таким образом, концентрация свинца в почве (в нашем конкретном случае) связана с расстоянием до автотрассы. По мере удаления от трассы концентрация свинца снижается, на что указывает сильная отрицательная статистически значимая связь.

Задания к разделу 5 для самостоятельного выполнения

По исходным данным (таблица 5.9) рассчитать коэффициент линейной корреляции Пирсона и коэффициент ранговой корреляции Спирмена (при уровне статистической значимости p=0,05), сформулировать вывод о характере и силе связи между изучаемыми характеристиками. Сравнить полученные значения коэффициентов Пирсона и Спирмена.

Таблица 5.9

Концентрация меди в почве на различном удалении от предприятия-загрязнителя

№	Расстояние точки опробования от предприятия, м	Cu, мг/кг
1	20	9
2	35	7,5
3	40	7,7
4	50	6,1
5	55	5,6
6	65	4,5
7	70	4
8	75	3,5

6. Определение значимости различий выборок объектов исследования

На практике часто возникает необходимость определить, различаются ли характеристики объектов исследования друг от друга. И значимо ли это различие. Одним из наиболее широко используемых расчетных показателей в этой области выступает t-критерий Стьюдента [Student, 1908]. Различают 2 t-критерия Стьюдента: критерий для зависимых данных и критерий для независимых данных. Оба этих критерия являются параметрическими, то есть при их вычислении обязательным условием является распределение данных близкое нормальному (смотреть раздел 4).

t-критерий Стьюдента для зависимых данных необходим для того, чтобы определить значительно ли изменились характеристики объектов исследования после наступления какого-либо события. То есть в данном случае производится сравнение объектов исследования с самими собой. Осуществляется сопоставление их характеристик в начальный момент времени и после проведения над объектами исследования каких-либо манипуляций. Например: можно, таким образом, сравнить средний бал группы студентов до проведения обучающего семинара и после его проведения. Можно сопоставить выбросы предприятий до внедрения очистных установок и выбросы после внедрения таких установок и др.

Т-критерий Стьюдента для зависимых данных рассчитывается по формуле (6.1):

$$t = \frac{|M_d|\sqrt{n}}{6_d}, \quad (6.1)$$

где t – t-критерий Стьюдента;
Md – среднее арифметическое значение разностей показателей (измерений до наступления какого-либо события и измерений после наступления события);
$б_d$ – стандартное отклонение разностей показателей;

n – количество объектов исследования в группе (должно быть одинаковым до наступления события и после его наступления).

После расчета t-критерия по формуле (6.1) необходимо произвести проверку его статистической значимости, сопоставив расчетное значение с табличным. Далее приведен фрагмент таблицы «Критических значения t-критерия Стьюдента» (таблица 6.1). Это стандартная таблица, которую можно найти в статистических справочниках или в Интернете. Для определения табличного значения t-критерия (с которым будем сопоставлять расчетное значение) необходимо дополнительно знать еще 2 параметра: число степеней свободы (f) и уровень статистической значимости (p).

Число степеней свободы для зависимых данных рассчитывается по формуле (6.2):

$$f = n - 1, \quad (6.2)$$

где f – число степеней свободы;
n – количество объектов исследования.

Уровень статистической значимости задается заранее. Его не нужно отдельно рассчитывать. Для большинства экологических исследований используют p=0,05.

Вычислив значение f, находим табличный (таблица 6.1, второй столбик) t-критерий Стьюдента при соответствующем уровне значимости (в нашем случае p=0,05).

Таблица 6.1

Критические значения t-критерия Стьюдента

[Values of…, 2019]

Число степеней свободы, f	Значение t-критерия Стьюдента при p=0,05
1	12,706
2	4,303
3	3,182
4	2,776

5	2,571
6	2,447
7	2,365
8	2,306
9	2,262
10	2,228
11	2,201
12	2,179
13	2,160
14	2,145
15	2,131
16	2,120
17	2,110

Если значение расчетного (рассчитанного по формуле (6.1)) t-критерия Стьюдента равны или превышают значения табличные (таблица 6.1) – то различия значений объектов исследования значительны (значимы), если наоборот – то различия значений характеристик объектов исследования (до манипуляций над ними и после манипуляций) незначительны.

Пример расчета t-критерия Стьюдента для зависимых данных. Допустим, имеется группа предприятий, на которых была внедрена система очистки выбросов. Данные по выбросам до внедрения фильтров и после приведены в таблице 6.2.

Таблица 6.2

Выбросы предприятий до и после внедрения системы очистки

№	Выбросы предприятия до внедрения очистных установок, т/год	Выбросы предприятия после внедрения очистных установок, т/год
1	2,3	1,5
2	2,8	2,2
3	4,2	3,5
4	6,2	6,5

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

Для начала найдем значения разностей показателей, указанных в столбцах 2 и 3 таблицы 6.2 (таблица 6.3).

Таблица 6.3

Вычисление разности показателей

№	Выбросы предприятия до внедрения очистных установок, т/год	Выбросы предприятия после внедрения очистных установок, т/год	Разность показателей
1	2,3	1,5	$2,3 - 1,5 = 0,8$
2	2,8	2,2	$2,8 - 2,2 = 0,6$
3	4,2	3,5	$4,2 - 3,5 = 0,7$
4	6,2	6,5	$6,2 - 6,5 = -0,3$
5	3,1	2,5	$3,1 - 2,5 = 0,6$
6	5,3	4,5	$5,3 - 4,5 = 0,8$
7	5,1	4	$5,1 - 4 = 1,1$
8	2,3	1,4	$2,3 - 1,4 = 0,9$
9	3,1	2,3	$3,1 - 2,3 = 0,8$
10	2,8	2,8	$2,8 - 2,8 = 0$

Далее найдем M_d и $б_d$:

$$M_d = \frac{0,8+0,6+0,7+(-0,3)+0,6+0,8+1,1+0,9+0,8+0}{10} = 0,6;$$

$$б_d = \sqrt{\frac{(0,8-0,6)^2+(0,6-0,6)^2+(0,7-0,6)^2+(-0,3-0,6)^2+(0,6-0,6)^2+ +(0,8-0,6)^2+(1,1-0,6)^2+(0,9-0,6)^2+(0,8-0,6)^2+(0-0,6)^2}{10-1}} = 0,43;$$

Подставим все необходимые показатели в формулу (6.1):

$$t = \frac{|0,6|\sqrt{10}}{0,43} = 4,41.$$

Далее найдем значение *f*.

$f = 10 - 1 = 9$.

По известному f, пользуясь таблицей критических значений, определим табличное значение t-критерия Стьюдента при p=0,05.

Сравним рассчитанный и табличный t-критерий. Так как расчетный критерий больше табличного (4,41>2,262), то различия значимы при p=0,05. То есть установки по очистке воздуха действительно повлияли на количество выбросов.

Т-критерий для независимых выборок – это модификация t-критерия, которая позволяет сравнивать группы разных объектов исследования. Например, озерные экосистемы различных населенных пунктов, популяции зайца различных территорий, популяции разных видов грызунов и др. При этом в сравниваемых группах (популяциях) может быть различное количество объектов исследования.

Логика, этапы и последовательность расчета t-критерия Стьюдента для независимых данных совпадает с таковыми при расчете критерия для зависимых данных. Отличия заключаются лишь в формулах для расчета t и f.

При расчете t-критерия Стьюдента для независимых данных используют формулу (6.3):

$$t = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}}, \quad (6.3)$$

где t – t-критерий Стьюдента для независимых данных;

M_1 – среднее арифметическое значение первой группы объектов исследования;

M_2 – среднее арифметическое значение второй группы объектов исследования;

m_1 – стандартная ошибка среднего первой группы объектов исследования;

m_2 – стандартная ошибка среднего второй группы объектов исследования.

Подробнее о расчете стандартной ошибки среднего можно узнать в разделе 1.

Расчет числа степеней свободы (f) для независимых данных осуществляют по формуле (6.4):

$$f = (n_1 + n_2) - 2, \quad (6.4)$$

где n_1 – количество объектов в первой группе;
n_2 – количество объектов во второй группе.

Пример расчета t-критерия для независимых данных. Необходимо выяснить существенны ли различия массы тела самцов двух групп волков (таблица 6.4).

Таблица 6.4

Значения массы тела взрослых самцов двух популяций волка

№	Значения массы тела 1-ой популяции, кг	Значения массы тела 2-ой популяции, кг
1	53,6	42
2	46,3	43
3	44,2	42,5
4	49,1	42,5
5	48,1	43,3
6	47	42
7	46	44,3
8	49,2	44
9	46,5	
1	48,3	

Сначала рассчитаем M_1 и M_2:

$$M_1 = \frac{53{,}6+46{,}3+44{,}2+49{,}1+48{,}1+47+46+49{,}2+46{,}5+48{,}3}{10} = 47{,}8;$$

$$M_2 = \frac{42+43+42{,}5+42{,}5+43{,}3+42+44{,}3+44}{8} = 43.$$

Для расчета m_1 и m_2 необходимо сначала рассчитать соответствующие стандартные отклонения:

$$6_1 = \sqrt{\frac{(53{,}6-47{,}8)^2+(46{,}3-47{,}8)^2+(44{,}2-47{,}8)^2+(49{,}1-47{,}8)^2+ +(48{,}1-47{,}8)^2+(47-47{,}8)^2+(46-47{,}8)^2+(49{,}2-47{,}8)^2+ +(46{,}5-47{,}8)^2+(48{,}3-47{,}8)^2}{10-1}} = 2{,}55;$$

$$6_2 = \sqrt{\frac{(42-43)^2+(43-43)^2+(42{,}5-43)^2+(42{,}5-43)^2+ +(43{,}3-43)^2+(42-43)^2+(44{,}3-43)^2+(44-43)^2}{8-1}} = 0{,}87.$$

Далее рассчитаем m_1 и m_2.

$$m_1 = \frac{6_1}{\sqrt{n_1}} = \frac{2{,}55}{\sqrt{10}} = 0{,}81;$$

$$m_2 = \frac{6_2}{\sqrt{n_2}} = \frac{0{,}87}{\sqrt{8}} = 0{,}31.$$

Подставляем полученные значения в формулу (6.3):

$$t = \frac{47{,}8-43}{\sqrt{0{,}81^2+0{,}31^2}} = 5{,}53.$$

Далее найдем значение *f*.

$$f = (10+8)-2 = 16.$$

По известному *f*, пользуясь таблицей критических значений (таблица 6.1), определим табличное значение t-критерия Стьюдента при p=0,05. Сравним рассчитанный и табличный t-критерий. Так как расчетный критерий больше табличного (5,53>2,120), то различия групп объектов исследования значимы (при p=0,05). То есть массы самцов двух групп волка действительно значительно отличаются. Волки первой группы, как правило, обладают большей массой.

Непараметрические критерии сравнения значимости различия выборок. В случае если выборка не подчиняется закону нормального распределения данных, вместо t-критерия Стьюдента используются непараметрические критерии. Они могут использовать данные, имеющие нормальное (близкое нормальному) распределение и отклоняющиеся от закона нормального распределения. Обычно если объектов исследования мало, то предпочтительнее использовать непараметрические критерии. Еще одним плюсом становится относительная простота их расчета. В разделе будет рассмотрено 2 основных непараметрических критерия: Т-критерий для зависимых данных Уилкоксона (иногда Вилкоксона или W-критерий Уилкоксона) [Wilcoxon, 1945] и U-критерий Манна-Уитни для независимых данных [Mann, Whitney, 1947].

Т-критерий Уилкоксона для зависимых данных применяется для тех же целей, что и t-критерий Стьюдента для зависимых данных. То есть в случаях, когда нужно сравнить значения какого-либо параметра

объектов исследования измеренного до какого-то события и после. То есть когда дело имеют, как правило, с одними и теми же объектами исследования. Например: с пациентами до использования лекарства и после его использования; с компонентами экосистем до загрязнения и после загрязнения и т.д.

Рассчитывается критерий Уилкоксона по следующему алгоритму:

1. Составить таблицу, где в двух параллельных столбцах указываются значения характеристик объектов исследования «до» (1 столбец) и «после» (2 столбец).

2. Вычислить разность между соответствующими значениями «после» и «до» (от значений «после» отнимаем значения «до»). Записать значения этой разности в отдельном столбце таблицы. Определить, какое количество разностей получилось со знаком «+», а какое со знаком «–». Значения того знака, который преобладают на этом этапе называют «типичный сдвиг». Оставшиеся данные – «сдвиг нетипичный».

3. Переписать полученный столбец разностей показателей, выраженный в абсолютных значениях, то есть без указания «±».

4. Произвести сортировку столбца разностей, полученного на предыдущем этапе, от меньшего значения к большему.

5. В новом столбце таблицы произвести нумерацию данных от меньшего к большему, используя порядковые числа (1, 2, 3 и т.д.). Присвоить значениям разностей ранги от меньшего значения к большему, так, что наименьшему значению присваивают ранг 1, всем последующим в порядке возрастания 2, 3, 4 и т.д. В случае наличия равных разностей (при совпадении значений разности) ранг для этих равных разностей будет вычислен как среднее арифметическое значение их порядковых номеров.

То есть по формуле (6.5):

$$r = \frac{n_1 + n_2 + \cdots + n_i}{N}, \quad (6.5)$$

где n_i – порядковые номера равных разностей;
N – количество равных разностей.

6. Произвести контроль правильности присвоения рангов сопоставив общую сумму рангов и контрольную сумму рангов, вычислив контрольную сумму рангов ($\sum x_{ij}$) по формуле (6.6):

$$\sum x_{ij} = \frac{(1+n)n}{2}, \quad (6.6)$$

где n – количество объектов исследования.

Если сумма и контрольная сумма совпадают, то ранжирование выполненно правильно.

7. Обозначить все ранги (используя любой графический прием: выделив их подчеркиванием, цветом и др.), присвоенные «нетипичным сдвигам». Посчитать сумму рангов «нетипичных сдвигов» (обозначается буквой «*T*»).

8. Используя таблицу критических значений Т-критерия Уилкоксона (таблица 6.5), найти критическое значения T ($T_{кр}$) при заданном уровне значимости (например, при $p<0,05$) и количестве объектов исследования (n).

9. Если расчетный Т-критерий Уилкоксона ($T_{эмп}$) меньше или равен $T_{кр}$, то сдвиг в «типичном направлении» статистически достоверно преобладает.

Таблица 6.5

Критические значения Т-критерия Уилкоксона (фрагмент оригинала)

[Т-критерий Вилкоксона, 2006-2019]

n	p<0,05	p<0,01
5	0	—
6	2	—
7	3	0
8	5	1
9	8	3
10	10	5
11	13	7
12	17	9
13	21	12
14	25	15

15	30	19
16	35	23
17	41	27
18	47	32
19	53	37
20	60	43

Пример расчета Т-критерия Уилкоксона. Для лучшего усвоения материала произведем расчет Т-критерия Уилкоксона на примере (таблица 6.6). Дана концентрация меди в воде 10 водоемов, на которых расположены предприятия, оказывающие негативное воздействие на качество воды. На всех 10 предприятиях внедрили новую систему управления технологическими процессами, которая должна снизить негативное воздействие на водоемы. Задача: проверить снизилась ли статистически значимо (при $p<0,05$) концентрация меди в воде водоемов после внедрения системы управления.

Таблица 6.6

Концентрация меди в воде водоемов

№	Концентрация меди в воде водоемов до внедрения системы управления, мг/л	Концентрация меди в воде водоемов после внедрения системы управления, мг/л
1	2,7	2,3
2	2,0	1,5
3	1,3	1,0
4	1,5	1,1
5	1,1	1,3
6	2,2	1,5
7	1,4	1,1
8	2,1	1,1
9	2,3	2,4
10	2,5	2,0

Итак, применив Т-критерий Уилкоксона, проверим, значимо ли изменилась концентрация меди в воде водоемов. Воспользуемся приведенным ранее алгоритмом.

Так как таблица со значениями показателя «до» и «после» уже составлена, произведем расчет разностей показателей «после» и «до». Для удобства указанные манипуляции будем пошагово отображать в табличном виде (таблица 6.7).

Таблица 6.7

Нахождение разности значений показателя «после» и «до»

№	Концентрация меди в воде водоемов до внедрения системы управления, мг/л	Концентрация меди в воде водоемов после внедрения системы управления, мг/л	Разность [после – до]
1	2,7	2,3	$2{,}3 - 2{,}7 = -0{,}4$
2	2,0	1,5	$1{,}5 - 2{,}0 = -0{,}5$
3	1,3	1,0	$1{,}0 - 1{,}3 = -0{,}3$
4	1,5	1,1	$1{,}1 - 1{,}5 = -0{,}4$
5	1,1	1,3	$1{,}3 - 1{,}1 = 0{,}2$
6	2,2	1,5	$1{,}5 - 2{,}2 = -0{,}7$
7	1,4	1,1	$1{,}1 - 1{,}4 = -0{,}3$
8	2,1	1,1	$1{,}1 - 2{,}1 = -1{,}0$
9	2,3	2,4	$2{,}4 - 2{,}3 = 0{,}1$
10	2,5	2,0	$2{,}0 - 2{,}5 = -0{,}5$

Далее определяем количество, каких знаков (+ или –) преобладает среди значений разности. Так как преобладают отрицательные значения, то типичным сдвигом будет считаться сдвиг в отрицательную сторону.

Вписываем в новый столбец таблицы (таблица 6.8) значения разностей показателей «после» и «до» в абсолютных значениях (то есть без указания знаков «±»).

Таблица 6.8

Абсолютные значения разности показателя «после» и «до»

№	Концентрация меди в воде водоемов до внедрения системы управления, мг/л	Концентрация меди в воде водоемов после внедрения системы управления, мг/л	Разность [после – до]	Абсолютные значения разности

1	2,7	2,3	$2,3 - 2,7 = -0,4$	0,4
2	2,0	1,5	$1,5 - 2,0 = -0,5$	0,5
3	1,3	1,0	$1,0 - 1,3 = -0,3$	0,3
4	1,5	1,1	$1,1 - 1,5 = -0,4$	0,4
5	1,1	1,3	$1,3 - 1,1 = 0,2$	0,2
6	2,2	1,5	$1,5 - 2,2 = -0,7$	0,7
7	1,4	1,1	$1,1 - 1,4 = -0,3$	0,3
8	2,1	1,1	$1,1 - 2,1 = -1,0$	1,0
9	2,3	2,4	$2,4 - 2,3 = 0,1$	0,1
10	2,5	2,0	$2,0 - 2,5 = -0,5$	0,5

Далее произведем сортировку абсолютных значений разности от меньшего значения к большему (таблица 6.9).

Таблица 6.9

Сортировка абсолютных значений разности показателя «после» и «до»

№	Концентрация меди в воде водоемов до внедрения системы управления, мг/л	Концентрация меди в воде водоемов после внедрения системы управления, мг/л	Разность [после – до]	Абсолютные значения разности
9	2,3	2,4	$2,4 - 2,3 = 0,1$	0,1
5	1,1	1,3	$1,3 - 1,1 = 0,2$	0,2
3	1,3	1	$1,0 - 1,3 = -0,3$	0,3
7	1,4	1,1	$1,1 - 1,4 = -0,3$	0,3
1	2,7	2,3	$2,3 - 2,7 = -0,4$	0,4
4	1,5	1,1	$1,1 - 1,5 = -0,4$	0,4
2	2	1,5	$1,5 - 2,0 = -0,5$	0,5
10	2,5	2	$2,0 - 2,5 = -0,5$	0,5
6	2,2	1,5	$1,5 - 2,2 = -0,7$	0,7
8	2,1	1,1	$1,1 - 2,1 = -1,0$	1,0

В новом столбце таблицы произведем нумерацию абсолютных значений разности данных от меньшего к большему, используя порядковые числа (1, 2, 3 и т.д.) (таблица 6.10).

Таблица 6.10

Присвоение номеров абсолютным значения разности показателя «после» и «до»

№	Концентрация меди в воде водоемов до внедрения системы управления, мг/л	Концентрация меди в воде водоемов после внедрения системы управления, мг/л	Разность [после – до]	Абсолютные значения разности	Номера
9	2,3	2,4	$2{,}4 - 2{,}3 = 0{,}1$	0,1	1
5	1,1	1,3	$1{,}3 - 1{,}1 = 0{,}2$	0,2	2
3	1,3	1	$1{,}0 - 1{,}3 = -0{,}3$	0,3	3
7	1,4	1,1	$1{,}1 - 1{,}4 = -0{,}3$	0,3	4
1	2,7	2,3	$2{,}3 - 2{,}7 = -0{,}4$	0,4	5
4	1,5	1,1	$1{,}1 - 1{,}5 = -0{,}4$	0,4	6
2	2	1,5	$1{,}5 - 2{,}0 = -0{,}5$	0,5	7
1	2,5	2	$2{,}0 - 2{,}5 = -0{,}5$	0,5	8
6	2,2	1,5	$1{,}5 - 2{,}2 = -0{,}7$	0,7	9
8	2,1	1,1	$1{,}1 - 2{,}1 = -1{,}0$	1	10

Далее необходимо присвоить значениям разностей ранги от меньшего значения к большему, так, что наименьшему значению присваивают ранг 1, всем последующим в порядке возрастания 2, 3, 4 и т.д. В нашем случае имеются одинаковые значения разности (по 2 раза повторяются значения 0,3; 0,4 и 0,5). Ранг для этих равных разностей будет вычислен как среднее арифметическое значение их порядковых номеров. Для 0,3 порядковые номера (см. таблицу 6.10) 3 и 4; для 0,4 – 5 и 6; для 0,5 – 7 и 8.

То есть для разности 0,1 ранг равен 1, для разности 0,2 ранг равен 2, для разностей 0,3; 0,4 и 0,5 рассчитывается по формуле (6.5):

$r_{0,3} = \frac{3+4}{2} = 3{,}5.$

$r_{0,4} = \frac{5+6}{2} = 5{,}5.$

$r_{0,5} = \frac{7+8}{2} = 7{,}5.$

Для оставшихся значений (0,7 и 1,0) ранги будут соответствовать их порядковым номерам.

Запишем все полученные ранги в новый столбец таблицы (таблица 6.11).

Таблица 6.11

Ранги абсолютных значений разности показателя «после» и «до»

№	Концентрация меди в воде водоемов до внедрения системы управления, мг/л	Концентрация меди в воде водоемов после внедрения системы управления, мг/л	Разность [после – до]	Абсолютные значения разности	Номера	Ранги разности
9	2,3	2,4	$2{,}4 - 2{,}3 = 0{,}1$	0,1	1	1
5	1,1	1,3	$1{,}3 - 1{,}1 = 0{,}2$	0,2	2	2
3	1,3	1	$1{,}0 - 1{,}3 = -0{,}3$	0,3	3	3,5
7	1,4	1,1	$1{,}1 - 1{,}4 = -0{,}3$	0,3	4	3,5
1	2,7	2,3	$2{,}3 - 2{,}7 = -0{,}4$	0,4	5	5,5
4	1,5	1,1	$1{,}1 - 1{,}5 = -0{,}4$	0,4	6	5,5
2	2	1,5	$1{,}5 - 2{,}0 = -0{,}5$	0,5	7	7,5
10	2,5	2	$2{,}0 - 2{,}5 = -0{,}5$	0,5	8	7,5
6	2,2	1,5	$1{,}5 - 2{,}2 = -0{,}7$	0,7	9	9
8	2,1	1,1	$1{,}1 - 2{,}1 = -1{,}0$	1	10	10

Произведем контроль правильности присвоения рангов сопоставив общую сумму рангов и контрольную сумму рангов.

Общую сумму рангов ($\sum r$) получаем простым суммированием всех рангов, присвоенных разностям:

$\sum r = 1 + 2 + 3{,}5 + 3{,}5 + 5{,}5 + 5{,}5 + 7{,}5 + 7{,}5 + 9 + 10 = 55.$

Контрольную сумму рангов ($\sum x_{ij}$) вычисляем по формуле (6.6).

$\sum x_{ij} = \frac{(1+10)10}{2} = 55.$

Так как общая сумма рангов и контрольная сумма совпадают, то ранги присвоены правильно. Далее, выделим все ранги, присвоенные «нетипичному сдвигу» (в нашем случае положительным разностям), используя окрашивание соответствующих ячеек таблицы (таблица 6.12).

Таблица 6.12

Выделение рангов «нетипичного сдвига»

№	Концентрация меди в воде водоемов до внедрения системы управления, мг/л	Концентрация меди в воде водоемов после внедрения системы управления, мг/л	Разность [после – до]	Абсолютные значения разности	Номера	Ранги разности
9	2,3	2,4	$2{,}4 - 2{,}3 = 0{,}1$	0,1	1	1
5	1,1	1,3	$1{,}3 - 1{,}1 = 0{,}2$	0,2	2	2
3	1,3	1	$1{,}0 - 1{,}3 = -0{,}3$	0,3	3	3,5
7	1,4	1,1	$1{,}1 - 1{,}4 = -0{,}3$	0,3	4	3,5
1	2,7	2,3	$2{,}3 - 2{,}7 = -0{,}4$	0,4	5	5,5
4	1,5	1,1	$1{,}1 - 1{,}5 = -0{,}4$	0,4	6	5,5
2	2	1,5	$1{,}5 - 2{,}0 = -0{,}5$	0,5	7	7,5
10	2,5	2	$2{,}0 - 2{,}5 = -0{,}5$	0,5	8	7,5
6	2,2	1,5	$1{,}5 - 2{,}2 = -0{,}7$	0,7	9	9
8	2,1	1,1	$1{,}1 - 2{,}1 = -1{,}0$	1	10	10

Далее найдем сумму рангов «нетипичных сдвигов» ($T_{эмп}$):
$T_{\text{эмп}} = 1 + 2 = 3$.

Используя таблицу критических значений Т-критерия Уилкоксона (таблица 6.1), найдем критическое значения $T_{\text{кр}}$ при, при $p<0{,}05$ и количестве объектов исследования n=10.

Так как расчетный Т-критерий Уилкоксона ($T_{эмп}$) меньше $T_{кр}$, то сдвиг в «типичном направлении» статистически достоверно преобладает. То есть концентрация меди в воде указанных водоемов действительно снизилась статистически значимо (при $p<0{,}05$) после внедрения системы управления на предприятиях.

U-критерий Манна-Уитни для независимых данных является широко используемым непараметрическим аналогом соответствующего t-критерия Стьюдента. То есть он применяется в тех случаях, когда нужно определить, существенно ли отличаются значения параметра двух выборок, представленных разными

объектами (2 популяции волка, 2 группы озер, 2 группы предприятий и т.д.). При этом эти 2 выборки могут состоять из различного количества объектов.

Алгоритм для расчета U-критерия следующий:

1. Из двух сравниваемых выборок анализируемых характеристик составляется единый ряд данных.

2. Единый ряд сортируется по возрастанию значений признака (от меньшего к большему).

3. После чего происходит ранжирование единого ряда с присвоением рангов от меньшего значения к большему. Наименьшему значению данных присваивается ранг 1, последующим ранги 2; 3 и т.д. В случае если имеются одинаковые значения в отсортированном ряде, то их ранг будет вычислен как среднее арифметическое порядковых номеров данных значений в отсортированном ряде.

Например, имеется отсортированный ряд данных, состоящий из 6 чисел: 0,1; 0,2; 0,3; 0,4; 0,4 и 0,5. Ранг 1 присвоим значению 0,1, ранг 2 – значению 0,2, ранг 3 – значению 0,3. Так как 0,4 повторяется 2 раза и стоит на 4-м и 5-м месте в ряде данных, то для него ранг (r_x) будет рассчитан следующим образом:

$r_{0,4} = \frac{4+5}{2}$=4,5.

Для значения 0,5 ранг будет равен 6 (по порядковому номеру в отсортированном ряду). Таким образом, получим следующую последовательность рангов: 1; 2; 3; 4,5; 4,5 и 6.

4. Далее высчитываем сумму рангов внутри первой выборки и сумму рангов внутри второй выборки.

5. Используем большую из ранговых сумм (T_x) для вычисления U-критерия (U) по формуле (6.7):

$$U = n_1 n_2 + \frac{n_x(n_x+1)}{2} - T_x\,, \quad (6.7)$$

где n_1 – количество объектов первой выборки;
n_2 – количество объектов второй выборки;

n_x – количество объектов в выборке, обладающей большей ранговой суммой T_x.

6. Сопоставляем вычисленное значение U-критерия (U) с его табличным критическим значением ($U_{кр}$) при соответствующем уровне статистической значимости (например, при p<0,05) (таблица 6.13). Если расчетное значение меньше или равно табличному, то различия выборок статистически значимы.

Как пользоваться таблицей критических значений U-критерия Манна-Уитни? Критические значения U-критерия определяются исходя из количеств объектов исследования в первой (n_1) и второй выборках (n_2), указанных соответственно в первом столбце и второй строке таблицы 6.13. Критическое значение U-критерия находится в ячейках, расположенных в месте пересечения перпендикуляров, опускаемых от значений соответствующих количеств объектов первой и второй выборки. Например, если первая выборка содержит 7 значений, а вторая 10, то критическое значение U-критерия равно 14.

Таблица 6.13

Критические значения U-критерия Манна-Уитни при p=0,05

[Billiet, 2003]

n_1	n_2											
	7	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
3	1	2	2	3	3	4	4	5	5	6	6	7
4	3	4	4	5	6	7	8	9	10	11	11	12
5	5	6	7	8	9	11	12	13	14	15	17	18
6	6	8	10	11	13	14	16	17	19	21	22	24
7	8	10	12	14	16	18	20	22	24	26	28	30
8	10	13	15	17	19	22	24	26	29	31	34	36
9	12	15	17	20	23	26	28	30	34	37	39	42
10	14	17	20	23	26	29	33	36	39	42	45	48
11	16	19	23	26	30	33	37	40	44	48	51	55
12	18	22	26	29	33	37	41	45	49	53	57	61
13	20	24	28	33	37	41	45	50	54	59	63	67
14	22	26	31	36	40	45	50	55	59	64	67	74
15	24	29	34	39	44	49	54	59	64	70	75	80
16	26	31	37	42	47	53	59	64	70	75	81	86
17	28	34	39	45	51	57	63	67	75	81	87	93
18	30	36	42	48	55	61	67	74	80	86	93	99
19	32	38	45	52	58	65	72	78	85	92	99	106
20	34	41	48	55	62	69	76	83	90	98	105	112

Пример вычисления U-критерия Манна-Уитни. Даны значения длины стволов двух групп одновозрастных берез (таблица 6.14). Первая группа включает 10 деревьев, произрастающих у предприятия, осуществляющего, вероятно, негативное воздействие на рост деревьев. Вторая группа, расположенная вдали от негативного воздействия предприятия, включает 12 берез. Существует гипотеза, что березы второй группы выше. Необходимо это проверить, применив U-критерий Манна-Уитни.

Таблица 6.14

Значения длины ствола двух групп сосен

№	Значения длины ствола деревьев, расположенных близ предприятия, м	Значение длины ствола на фоновом участке, м
1	5	6
2	5,5	6,1
3	6	7
4	4,5	7,1
5	6,1	6,2
6	4,7	6,5
7	4,8	6,4
8	5,1	6,0
9	4,4	6,2
10	4,4	6,7
11		6,1
12		6,3

Перед началом всех манипуляций произведем выделения объектов выборки 1 (березы у предприятия) и объектов выборки 2 (березы на фоновом участке) графически, для того, чтобы знать какой выборке принадлежат объекты. Для этого можно воспользоваться различными цветами, например, окрасить значения выборки 1 в зеленый цвет, а значения выборки 2 в красный (таблица 6.15).

Таблица 6.15

Графическое выделение объектов выборки 1 и выборки 2

№	Значения длины ствола деревьев, расположенных близ предприятия, м	Значение длины ствола на фоновом участке, м
1	5	6
2	5,5	6,1

3	6	7
4	4,5	7,1
5	6,1	6,2
6	4,7	6,5
7	4,8	6,4
8	5,1	6,0
9	4,4	6,2
10	4,4	6,7
11		6,1
12		6,3

Объединим обе выборки в единый ряд данных и произведем сортировку значений от меньшего к большему (таблица 6.16).

Таблица 6.16

Единый ряд отсортированных в порядке возрастания данных

№	Единый ряд данных
1	4,4
2	4,4
3	4,5
4	4,7
5	4,8
6	5
7	5,1
8	5,5
9	6
10	6
11	6
12	6,1
13	6,1
14	6,1
15	6,2
16	6,2
17	6,3
18	6,4
19	6,5
20	6,7
21	7
22	7,1

Далее, производим ранжирование единого ряда с присвоением рангов от меньшего значения к большему. Ранги указываем в отдельном столбике таблицы согласно руководству, приведенному при описании процедуры вычисления U-критерия. Для всех повторяющихся значений единого ряда данных (4,4; 6; 6,1 и 6,2) ранг (r_x) будет вычислен как среднее арифметическое порядковых номеров данных значений в отсортированном ряде. Порядковые номера отражены в 1-ом столбце таблицы 6.16.

$r_{4,4} = \frac{1+2}{2} = 1{,}5.$

$r_{6,0} = \frac{9+10+11}{3} = 10.$

$r_{6,1} = \frac{12+13+14}{3} = 13.$

$r_{6,2} = \frac{15+16}{2} = 15{,}5.$

Для всех остальных значений ранг будет соответствовать их порядковому номеру в отсортированном ряде (таблица 6.17).

Таблица 6.17

Ранжированный единый ряд отсортированных в порядке возрастания данных

№	Единый ряд данных	Ранг
1	4,4	1,5
2	4,4	1,5
3	4,5	3
4	4,7	4
5	4,8	5
6	5	6
7	5,1	7
8	5,5	8
9	6	10
10	6	10
11	6	10
12	6,1	13
13	6,1	13
14	6,1	13
15	6,2	15,5
16	6,2	15,5
17	6,3	17

18	6,4	18
19	6,5	19
20	6,7	20
21	7	21
22	7,1	22

Далее высчитываем сумму рангов, принадлежащих первой выборке (выделены зеленым цветом) T_1, и сумму рангов внутри второй выборки (выделены красным цветом) T_2.

$T_1 = 1{,}5 + 1{,}5 + 3 + 4 + 5 + 6 + 7 + 8 + 10 + 13 = 59.$

$T_2 = 13 + 13 + 15{,}5 + 15{,}5 + 17 + 18 + 19 + 20 + 21 + 22 = 194.$

Большую из ранговых сумм (194) используем для вычисления U-критерия (U) по формуле (6.7):

$$U = 10 * 12 + \frac{12(12+1)}{2} - 194 = 4.$$

Сопоставляем вычисленное значение U-критерия (U=4) с его табличным критическим значением ($U_{кр}$) при уровне статистической значимости p<0,05. Критическое значение равно 29. Так как расчетное значение меньше табличного (т.е. 4<29), то различия выборок статистически значимы. Соответственно деревья фонового участка действительно выше таковых, произрастающих близ предприятия.

Задания к разделу 6 для самостоятельного выполнения

1. Используя критерии для зависимых данных (t-критерий Стьюдента и Т-критерия Уилкоксона), определить значимость различий характеристик, указанных в таблице 6.18

Таблица 6.18.

Значения сбросов водопользователей до внедрения водоочистных установок и после внедрения

№	Сбросы предприятия до внедрения водоочистных установок, т/год	Выбросы предприятия после внедрения водоочистных установок, т/год
1	2,4	1,6
2	2,8	2,2

3	4,2	3,5
4	6,2	6,5
5	3,1	2,5
6	5,3	4,8
7	5,2	4
8	2,3	1,5
9	3,1	2,9
10	2,8	2,7

2. Используя критерии для независимых данных (t-критерий Стьюдента и U-критерий Манна-Уитни), определить значимость различий характеристик, указанных в таблице 6.19

Таблица 6.19

Значения массы тела взрослых самцов двух популяций волка

№	Значения массы тела 1-ой популяции, кг	Значения массы тела 2-ой популяции, кг
1	53,6	42
2	46,3	43
3	44,2	42,5
4	49,1	42,5
5	48,1	43,3
6	47	42
7	46	44,3
8	49,2	44
9	46,5	43,1
10	48,3	42
11	49,3	42,9
12	46	
13	50,1	
14	49,2	
15	48	

7. Группировка объектов исследования с применением процедур иерархического кластерного анализа

В случаях, когда большое количество объектов исследования с заданными характеристиками нужно подразделить на отдельные группы применяют кластерный анализ. Кластерный анализ – это целая группа методов группировки объектов или признаков (характеристик) объектов. В данном разделе будут рассмотрены общие принципы кластерного анализа (иерархического кластерного анализа) и основные широко употребляемые методы.

Принципы кластерного анализа основаны на том, что между исследуемыми объектами (точками), обладающими конкретными значениями признаков (координатами), можно установить расстояние. Объекты, расположенные на небольшом удалении друг от друга, образуют сходную группу (кластер), объекты, расположенные на большом удалении, представляют разные группы (кластеры).

Для лучшего понимания темы рассмотрим простой графический пример. Дано 10 предприятий с различной численностью персонала и количеством производимых отходов (таблица 7.1).

Таблица 7.1

Количество отходов, производимое предприятием с различной численностью работников

Предприятия	Количество работников, ед.	Количество отходов, ц/год
1	5	6
2	7	7
3	8	6,5
4	15	12
5	21	14
6	24	13
7	56	35
8	67	41
9	100	120
10	120	125

По значениям таблицы 7.1 построим двумерный график распределения предприятий по значениям, указанным в столбиках 2 и 3 (рисунок 7.1).

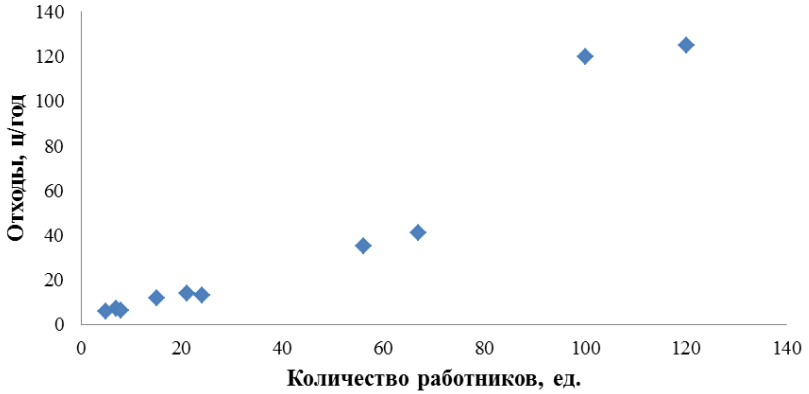


Рисунок 7.1. Количество отходов, производимое предприятием с различной численностью работников

На данном рисунке можно выделить 3 отдельные группы точек (предприятий) (рисунок 7.2). Группа 1 (кластер 1) предприятия с небольшой численностью персонала и с относительно небольшим количеством производимых отходов, группа 2 – предприятия со средним количеством работников и производимых отходов и группа 3 – предприятия с большой численностью персонала и большим количеством производимых отходов.

В указанном примере произведено определение расстояний между точками, без каких бы то ни было вычислений. На графике и так видно, какие точки расположены близко, а какие на значительном расстоянии. В указанном примере мы визуально оценили геометрическое (евклидово) расстояние между объектами, рассчитываемое по формуле (7.1).

$$d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}, \quad (7.1)$$

где $d(A,B)$ – евклидово расстояние между точками A и B;
x_i – координата x соответствующей точки;
y_i – координата y соответствующей точки.

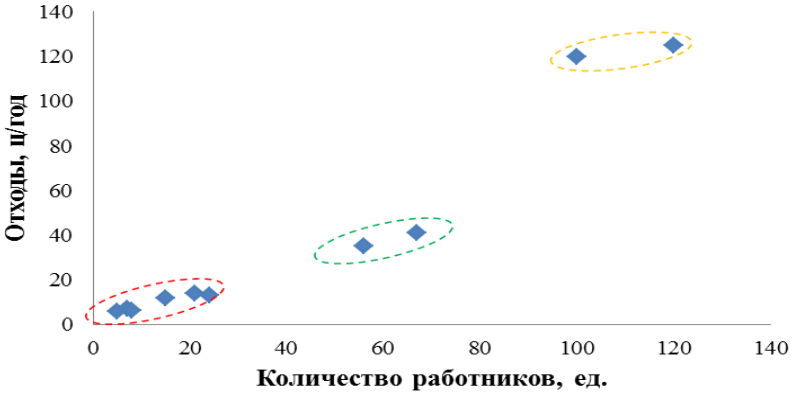


Рисунок 7.2. Выделенные кластеры

В случае если объекты исследования характеризуются 3 признаками, то они будут обладать 3 координатами. В этом случае формула для евклидова расстояния принимает следующий вид (7.2):

$$d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}, \quad (7.2)$$

где $d(A,B)$ – евклидово расстояние между точками A и B;
x_i – координата x соответствующей точки;
y_i – координата y соответствующей точки;
z_i – координата z соответствующей точки.

В этом случае также можно построить трехмерный график функции, на котором будут изображены точки. Но что делать, если характеристик объектов исследования будет больше 3? Как представить себе эту картину в пространстве? В этом случае наглядный график построить не представляется возможным, однако расстояние между объектами и в этом случае можно определить. В

формулу евклидова расстояния только лишь нужно будет добавить новые координаты. Координат будет столько, сколько параметров, характеризующих объекты исследования. В этом случае говорят о евклидовом расстоянии в многомерном пространстве (7.3):

$$d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2 + \cdots + (n_A - n_B)^2}, \quad (7.3)$$

где $d(A,B)$ – евклидово расстояние между точками А и В;
x_i – координата x соответствующей точки;
y_i – координата y соответствующей точки;
z_i – координата z соответствующей точки;
n_i – координата n соответствующей точки (количество совпадает с количеством характеристик).

Пример расчета евклидова расстояния. Допустим, имеются сведения о характеристиках 2-х организаций (таблица 7.2).

Таблица 7.2

Сведения об организациях

Организация	Количество работников, ед.	Количество отходов, ц/год	Количество автотранспорта, ед.	Ежегодный расход бумаги, ящ./год
Офис	5	6	1	3
Типография	7	7	1	3,5

Произведем расчет евклидова расстояния, применив формулу (7.3) и сведения из таблицы 7.2:

$d(A,B) = \sqrt{(5-7)^2 + (6-7)^2 + (1-1)^2 + (3-3{,}5)^2}$=2,3.

Евлидово расстояние не единственная мера близости объектов, однако, в кластерном анализе оно используется наиболее часто, поэтому для объяснения процедур кластерного анализа далее будет использовано именно оно. Итак, что же нужно сделать после вычисления расстояния между объектами? Для удобства изложения ниже приведены все основные этапы кластерного анализа.

Этапы кластерного анализа. Для проведения кластерного анализа необходимо произвести следующие действия:

1. Рассчитать расстояние между всеми объектами (как было указано ранее);

2. Произвести объединение наиболее близких точек. Далее, согласно одному из алгоритмов объединения (метод одиночной связи, метод полной связи или метод невзвешенного попарного среднего арифметического) произвести последовательное (на каждом шаге присоединяется 1 объект) объединение всех исследуемых объектов в кластеры (группы);

3. Построить график расстояний объединения объектов (дендрограмму);

4. Определить на дендрограмме количество кластеров, которые следует выделить для указанных объектов исследования.

1. Расчет расстояния между всеми объектами (точками). Расстояние должно быть рассчитано между всеми объектами исследования. Для простоты расчетов расстояний возьмем таблицу 7.1. Расстояние между объектами таблицы 7.2 рассчитывалось бы точно также, но координат было бы больше.

Итак, применяя формулу (7.3) рассчитаем расстояние между первой точкой и всеми остальными, далее между второй точкой и всеми оставшимися, между третьей точкой и всеми объектами и т.д., пока не будет вычислено расстояние между всеми точками.

Сначала произведем расчет евклидова расстояния между первым объектом и всеми остальными:

$d(1,2) = \sqrt{(5-7)^2 + (6-7)^2} = 2{,}2;$

$d(1,3) = \sqrt{(5-8)^2 + (6-6{,}5)^2} = 3{,}0;$

$d(1,4) = \sqrt{(5-15)^2 + (6-12)^2} = 11{,}7;$

$d(1,5) = \sqrt{(5-21)^2 + (6-14)^2} = 17{,}9;$

$d(1,6) = \sqrt{(5-24)^2 + (6-13)^2} = 20{,}3;$

$d(1,7) = \sqrt{(5-56)^2 + (6-35)^2} = 57{,}8;$

$d(1,8) = \sqrt{(5-67)^2 + (6-41)^2} = 71{,}2;$

$d(1,9) = \sqrt{(5-100)^2 + (6-120)^2} = 148{,}4;$

$d(1,10) = \sqrt{(5-120)^2 + (6-125)^2} = 165{,}5.$

Далее произведем расчет евклидова расстояния между вторым объектом и всеми остальными, кроме первого. Расстояние от второго до первого (d(2,1)) уже рассчитано, оно соответствует дистанции d(1,2). Нет разницы, как измерять расстояние: от первого объекта до второго или от второго до первого. Оно будет равным.

$d(2,3) = \sqrt{(7-8)^2 + (7-6{,}5)^2} = 1{,}1;$

$d(2,4) = \sqrt{(7-15)^2 + (7-12)^2} = 9{,}4;$

$d(2,5) = \sqrt{(7-21)^2 + (7-14)^2} = 15{,}7;$

$d(2,6) = \sqrt{(7-24)^2 + (7-13)^2} = 18{,}0;$

$d(2,7) = \sqrt{(7-56)^2 + (7-35)^2} = 56{,}4;$

$d(2,8) = \sqrt{(7-67)^2 + (7-41)^2} = 69{,}0;$

$d(2,9) = \sqrt{(7-100)^2 + (7-120)^2} = 146{,}3;$

$d(2,10) = \sqrt{(7-120)^2 + (7-125)^2} = 163{,}4.$

Произведем расчет евклидова расстояния между третьим объектом и всеми остальными, кроме первого и второго (эти расстояния уже рассчитаны).

$d(3,4) = \sqrt{(8-15)^2 + (6{,}5-12)^2} = 8{,}9;$

$d(3,5) = \sqrt{(8-21)^2 + (6{,}5-14)^2} = 15{,}0;$

$d(3,6) = \sqrt{(8-24)^2 + (6{,}5-13)^2} = 17{,}3;$

$d(3,7) = \sqrt{(8-56)^2 + (6{,}5-35)^2} = 55{,}8;$

$d(3,8) = \sqrt{(8-67)^2 + (6{,}5-41)^2} = 68{,}3;$

$d(3,9) = \sqrt{(8-100)^2 + (6{,}5-120)^2} = 146{,}1;$

$d(3,10) = \sqrt{(8-120)^2 + (6{,}5-125)^2} = 163{,}1.$

Произведем расчет евклидова расстояния между четвертым объектом и всеми оставшимися, кроме первого, второго и третьего.

$d(4,5) = \sqrt{(15-21)^2 + (12-14)^2} = 6{,}3;$

$d(4,6) = \sqrt{(15-24)^2 + (12-13)^2} = 9{,}1;$

$d(4,7) = \sqrt{(15-56)^2 + (12-35)^2} = 47{,}0;$

$d(4,8) = \sqrt{(15-67)^2 + (12-41)^2} = 59{,}5;$

$d(4,9) = \sqrt{(15-100)^2 + (12-120)^2} = 137{,}4;$

$d(4,10) = \sqrt{(15-120)^2 + (12-125)^2} = 154{,}3.$

Произведем расчет евклидова расстояния между пятым объектом и оставшимися, кроме первого, второго, третьего и четвертого.

$d(5,6) = \sqrt{(21-24)^2 + (14-13)^2} = 3{,}2;$

$d(5,7) = \sqrt{(21-56)^2 + (14-35)^2} = 40{,}8;$

$d(5,8) = \sqrt{(21-67)^2 + (14-41)^2} = 53{,}3;$

$d(5,9) = \sqrt{(21-100)^2 + (14-120)^2} = 132{,}2;$

$d(5,10) = \sqrt{(21-120)^2 + (14-125)^2} = 148{,}7.$

Произведем расчет евклидова расстояния между шестым объектом и всеми остальными, кроме первого, второго, третьего, четвертого и пятого.

$d(6,7) = \sqrt{(24-56)^2 + (13-35)^2} = 38{,}8;$

$d(6,8) = \sqrt{(24-67)^2 + (13-41)^2} = 51{,}3;$

$d(6,9) = \sqrt{(24-100)^2 + (13-120)^2} = 131{,}2;$

$d(6,10) = \sqrt{(24-120)^2 + (13-125)^2} = 147{,}5.$

Произведем расчет евклидова расстояния между седьмым объектом и всеми остальными, кроме первого, второго, третьего, четвертого, пятого и шестого.

$d(7,8) = \sqrt{(56-67)^2 + (35-41)^2} = 12{,}5;$

$d(7,9) = \sqrt{(56-100)^2 + (35-120)^2} = 95{,}7;$

$d(7,10) = \sqrt{(56-120)^2 + (35-125)^2} = 110{,}4.$

Произведем расчет евклидова расстояния между восьмым объектом и всеми остальными, кроме первого, второго, третьего, четвертого, пятого, шестого и седьмого.

$d(8,9) = \sqrt{(67-100)^2 + (41-120)^2} = 85{,}6;$

$d(8,10) = \sqrt{(67-120)^2 + (41-125)^2} = 99{,}3.$

Произведем расчет евклидова расстояния между девятым объектом и оставшимися, кроме первого, второго, третьего, четвертого, пятого, шестого, седьмого и восьмого.

$d(9,10) = \sqrt{(100-120)^2 + (120-125)^2} = 20{,}6.$

Таким образом, получены расстояния между всеми имеющимися точками. Далее для более удобного восприятия информации результаты вычислений записывают в виде матрицы расстояний, то есть в виде таблицы с указанием расстояний между объектами (точками) (таблица 7.3).

Таблица 7.3

Матрица расстояний между точками

Точки	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	2,2	3,0	11,7	17,9	20,3	57,8	71,2	148,4	165,5
2	2,2	0	1,1	9,4	15,7	18,0	56,4	69,0	146,3	163,4
3	3,0	1,1	0	8,9	15,0	17,3	55,8	68,3	146,1	163,1
4	11,7	9,4	8,9	0	6,3	9,1	47,0	59,5	137,4	154,3
5	17,9	15,7	15,0	6,3	0	3,2	40,8	53,3	132,2	148,7
6	20,3	18,0	17,3	9,1	3,2	0	38,8	51,3	131,2	147,5
7	57,8	56,4	55,8	47,0	40,8	38,8	0	12,5	95,7	110,4
8	71,2	69,0	68,3	59,5	53,3	51,3	12,5	0	85,6	99,3
9	148,4	146,3	146,1	137,4	132,2	131,2	95,7	85,6	0	20,6
10	165,5	163,4	163,1	154,3	148,7	147,5	110,4	99,3	20,6	0

Первый столбец и первая строка таблицы 7.3 содержат название объектов исследования, наименования точек между которыми произведено определение расстояния. Евклидово расстояние указано в ячейках, расположенных на пересечении номеров соответствующих объектов. Для одного объекта номер берем из первого столбца, для второго из первой строки. Так, на пересечении строк и столбцов с одинаковыми номерами указано число 0, что обусловлено тем, что расстояние объекта «до самого себя» равно 0. Таким образом, в матрице расстояний имеется «диагональ» из нолей, относительно которой, все числа зеркально отражены. В результате чего матрицей удобно пользоваться: легко можно отсчитывать расстояние, как от строки, так и от столбца. Результаты будут одинаковыми.

2. Объединение объектов (точек), с применением различных алгоритмов кластерного анализа. В наиболее обобщенном виде процедура объединения объектов в кластеры (группы) выглядит следующим образом: после построения матрицы расстояний между объектами исследования (таблица 7.3) необходимо произвести объединение двух наиболее близкорасположенных объектов. После их объединения производится объединение следующей пары наиболее близкорасположенных объектов. Данная процедура производится последовательно «шаг за шагом» (по одной точке), пока не будут соединены все объекты. С каждым шагом происходит объединение все более удаленных объектов.

Как объединять объекты? Как правило, первый шаг любого алгоритма всегда одинаковый: объединяются два наиболее близких объекта. Далее в зависимости от алгоритма существуют различия. В учебнике рассмотрим 3 основных метода (алгоритма) кластерного анализа [Estivill-Castro, 2002]: метод одиночной связи (метод ближнего соседа) [Florek et al., 1951; McQuitty, 1957; Sneath; 1957]; метод полной связи (метод дальнего соседа) [Sørensen, 1948] и метод невзвешенного попарного среднего арифметического [Sokal, Michener, 1958].

Метод одиночной связи (метод ближнего соседа). После шага один производится построение новой матрицы расстояний, теперь в этой матрице объединенные на первом шаге объекты будут представлять единый кластер с указанием расстояний до оставшихся объектов (точек). Суть метода ближнего соседа в том, что в новых ячейках матрицы будут записаны минимальные расстояния от образованного кластера, до оставшихся необъединенных в кластеры объектов. Чтобы понять изложенное, произведем описанные манипуляции на примере (таблица 7.3).

Наименьшее расстояние в таблице 7.3 зафиксировано между объектами 2 и 3, оно составляет 1,1. На первом шаге производим объединение этих объектов в единый кластер. Осуществим перестроение матрицы расстояний таким образом, чтобы внутри этой матрицы объекты 2 и 3 были представлены единым кластером «2;3», а расстояние внутри матрицы было пересчитано в соответствии с

алгоритмом «одиночной связи»: то есть указываем кратчайшее расстояние от одного из объектов кластера «2;3» до всех оставшихся объектов матрицы. Например, расстояние от точки 2 до точки 1 составляет 2,2, а расстояние от точки 3 до точки 1 составляет 3,0. Так как дистанция 2,2 короче, то ее и следует записать в новую матрицу расстояний (таблица 7.4).

Таблица 7.4

Перерасчет матрицы расстояний после первого шага кластеризации

Точки	**1**	**2;3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	2,2	11,7	17,9	20,3	57,8	71,2	148,4	165,5
2;3	2,2	0	8,9	15,0	17,3	55,8	68,3	146,1	163,1
4	11,7	8,9	0	6,3	9,1	47,0	59,5	137,4	154,3
5	17,9	15,0	6,3	0	3,2	40,8	53,3	132,2	148,7
6	20,3	17,3	9,1	3,2	0	38,8	51,3	131,2	147,5
7	57,8	55,8	47,0	40,8	38,8	0	12,5	95,7	110,4
8	71,2	68,3	59,5	53,3	51,3	12,5	0	85,6	99,3
9	148,4	146,1	137,4	132,2	131,2	95,7	85,6	0	20,6
10	165,5	163,1	154,3	148,7	147,5	110,4	99,3	20,6	0

В итоге полученная пересчитанная матрица расстояний стала короче первоначальной на 1 столбец и 1 строку. В новой таблице снова производится нахождение самой короткой дистанции. Наименьшее расстояние (2,2) отмечено между кластером «2;3» и объектом 1. Производим объединение объектов в новый кластер «2;3;1». Снова производим перестроение матрицы расстояний согласно алгоритму «ближнего соседа» основываясь на значениях из таблицы 7.4 (таблица 7.5). Вписываем в новую матрицу наиболее краткие расстояния для пары объектов «2;3» и 1. Так расстояние от точки 1 до объекта 4 равно 11,7, а от кластера «2;3» – 8,9, значит новый кластер «2;3;1» будет находиться на расстоянии 8,9 от кластера 4. Подобным образом вычисляются расстояния от кластера «2;3;1» до всех объектов (5, 6, 7, 8, 9 и 10).

Таблица 7.5

Перерасчет матрицы расстояний после второго шага кластеризации

Точки	2;3;1	4	5	6	7	8	9	10
2;3;1	0	8,9	15,0	17,3	55,8	68,3	146,1	163,1
4	8,9	0	6,3	9,1	47,0	59,5	137,4	154,3
5	15,0	6,3	0	3,2	40,8	53,3	132,2	148,7
6	17,3	9,1	3,2	0	38,8	51,3	131,2	147,5
7	55,8	47,0	40,8	38,8	0	12,5	95,7	110,4
8	68,3	59,5	53,3	51,3	12,5	0	85,6	99,3
9	146,1	137,4	132,2	131,2	95,7	85,6	0	20,6
10	163,1	154,3	148,7	147,5	110,4	99,3	20,6	0

Новая матрица расстояний также короче предыдущей на 1 столбик и на 1 строчку. В ней снова находим наименьшее расстояние между объектами, объединяем наиболее близкие объекты в кластер и производим перерасчет матрицы расстояний по алгоритму «метода одиночной связи», как было осуществлено на первом и втором шаге кластеризации. Подобные манипуляции проводят до тех пор, пока в таблице не будут объединены все объекты (кластеры). На последнем шаге кластеризации матрица расстояний будет состоять из 2 столбцов и 2 строчек. Ниже последовательно приведены матрицы всех последующих шагов кластеризации (Таблицы 7.6-7.11).

Таблица 7.6

Перерасчет матрицы расстояний после третьего шага кластеризации

Точки	2;3;1	4	5;6	7	8	9	10
2;3;1	0	8,9	15,0	55,8	68,3	146,1	163,1
4	8,9	0	6,3	47,0	59,5	137,4	154,3
5;6	15,0	6,3	0	38,8	51,3	131,2	147,5
7	55,8	47,0	38,8	0	12,5	95,7	110,4
8	68,3	59,5	51,3	12,5	0	85,6	99,3
9	146,1	137,4	131,2	95,7	85,6	0	20,6
10	163,1	154,3	147,5	110,4	99,3	20,6	0

Таблица 7.7

Перерасчет матрицы расстояний после четвертого шага кластеризации

Точки	2;3;1	5;6;4	7	8	9	10
2;3;1	0	8,9	55,8	68,3	146,1	163,1
5;6;4	8,9	0	38,8	51,3	131,2	147,5
7	55,8	38,8	0	12,5	95,7	110,4
8	68,3	51,3	12,5	0	85,6	99,3
9	146,1	131,2	95,7	85,6	0	20,6
10	163,1	147,5	110,4	99,3	20,6	0

Таблица 7.8

Перерасчет матрицы расстояний после пятого шага кластеризации

Точки	5;6;4;2;3;1	7	8	9	10
5;6;4;2;3;1	0	38,8	51,3	131,2	147,5
7	38,8	0	12,5	95,7	110,4
8	51,3	12,5	0	85,6	99,3
9	131,2	95,7	85,6	0	20,6
10	147,5	110,4	99,3	20,6	0

Таблица 7.9

Перерасчет матрицы расстояний после шестого шага кластеризации

Точки	5;6;4;2;3;1	7;8	9	10
5;6;4;2;3;1	0	38,8	131,2	147,5
7;8	38,8	0	85,6	99,3
9	131,2	85,6	0	20,6
10	147,5	99,3	20,6	0

Таблица 7.10

Перерасчет матрицы расстояний после седьмого шага кластеризации

Точки	5;6;4;2;3;1	7;8	9;10
5;6;4;2;3;1	0	38,8	131,2
7;8	38,8	0	85,6
9;10	131,2	85,6	0

Таблица 7.11

Перерасчет матрицы расстояний после восьмого шага кластеризации

Точки	5;6;4;2;3;1;7;8	9;10
5;6;4;2;3;1;7;8	0	85,6
9;10	85,6	0

Таким образом, на последнем шаге происходит объединение 2-х оставшихся кластеров на расстоянии 85,6 (таблица 7.11). После построения всех матриц расстояний и проведения всех шагов кластеризации наступает следующий этап кластерного анализа. Для удобства восприятия информации необходимо построить график объединения объектов исследования в кластеры (рисунок 7.3), где по горизонтальной оси откладываются название объектов (номера), а по вертикальной расстояние на котором они объединены.

Имея график объединения объектов исследования в кластеры (дендрограмму), можно проследить, как объекты объединялись в группы, на каком расстоянии друг от друга и в какой последовательности. В зависимости от целей и задач исследования полученный древовидный график, состоящий из отдельных веток (кластеров), соединяющих объекты исследования пошагово, можно подразделить на нужное исследователям количество групп объектов исследования (не большее, чем количество объектов исследования). Универсальных методов выделения нужного количества кластеров на дендрограмме не существует. В разных случая применяются разные подходы. Широко используется, например, метод согласно которому количество кластеров определяется как разность между количеством объектов выборки (в нашем случае 10) и шагом кластеризации, после которого наблюдается скачкообразное увеличение расстояния объединения. В нашем случае такое скачкообразное увеличение расстояния наблюдается после 6-го шага. То есть согласно данному подходу можно выделить 4 кластера.

Также количество кластеров можно определить «на глаз» после того, как расстояние объединения начинает значительно возрастать

(рисунок 7.4). Для удобства можно провести линию разграничения кластеров, разделяющую отдельные группы объектов исследования.

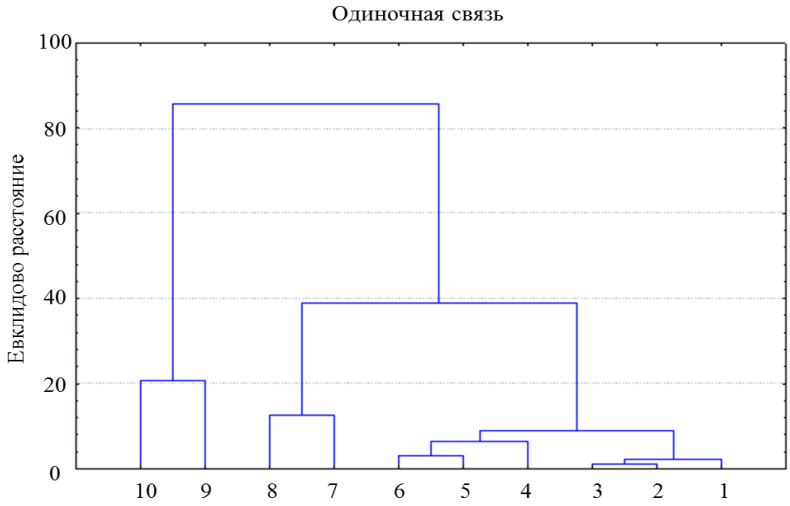


Рисунок 7.3. Дендрограмма кластеризации методом одиночной связи

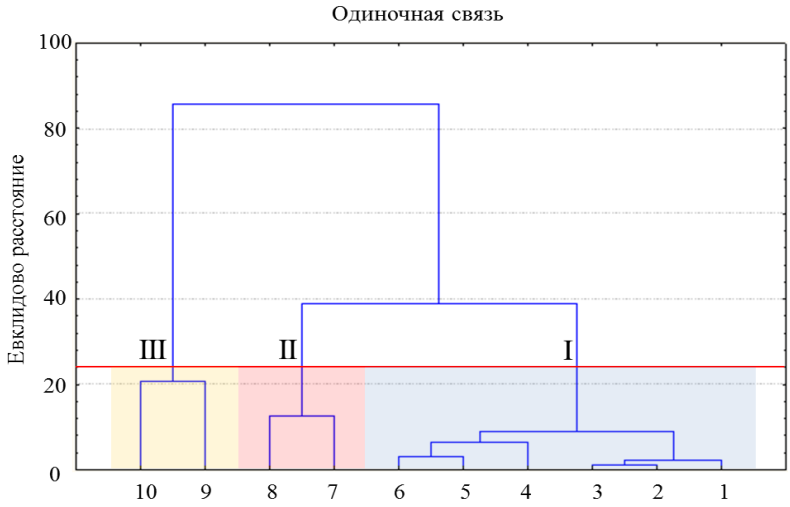


Рисунок 7.4. Один из способов выделения кластеров на дендрограмме

Таким образом, на дендрограмме (рисунок 7.4) произведено выделение трех кластеров. Первый кластер объединяет объекты 1,2,3,4,5,6, второй – 7,8 и третий 9 и 10. Далее производят подробное изучение каждой отдельной группы объектов исследования, сравнение групп между собой и прочие манипуляции, зависящие от целей и задач исследования. Отдельные ветки, из которых состоят крупные кластеры, нередко именуют «субкластерами». Например, первый кластер состоит из двух субкластеров: субкластера «5;6;4» и субкластера «2; 3;1».

Метод полной связи (метод дальнего соседа). Метод полной связи практически полностью совпадает с методом одиночной связи. Как и метод одиночной связи, алгоритм полной связи начинается после построения матрицы расстояний между объектами исследования. На первом шаге метода (как и для метода одиночной связи) производится объединение наиболее близкорасположенных объектов. Далее необходимо построить (как и для метода одиночной связи) новую матрицу расстояний, теперь в этой матрице объединенные на первом шаге объекты будут представлять единый кластер с указанными расстояниями до оставшихся объектов (точек). Суть метода полной связи в том, что в новых ячейках матрицы будут записаны не минимальные расстояния (как в методе одиночной связи), а максимальные расстояния от объектов, слагающих данный кластер, до оставшихся необъединенных в кластеры объектов. Чтобы понять изложенное произведем описанные манипуляции на примере (таблица 7.3).

Наименьшее расстояние в таблице 7.3 зафиксировано между объектами 2 и 3, оно составляет 1,1. На первом шаге производим объединение этих объектов в единый кластер. Осуществим перестроение матрицы расстояний таким образом, чтобы внутри этой матрицы объекты 2 и 3 были представлены единым кластером «2;3», а расстояние внутри матрицы было пересчитано в соответствии с алгоритмом «полной связи»: то есть указываем максимальное расстояние от одного из объектов кластера «2;3» до всех оставшихся объектов матрицы. Например, расстояние от точки 2 до точки 1 составляет 2,2, а расстояние от точки 3 до точки 1 составляет 3,0. Так

как дистанция 3,0 больше, то ее и следует записать в новую матрицу расстояний (таблица 7.12). Дистанция от точки 2 до точки 4 составляет 9,4, а от точки 3 до точки 4 – 8,9. Так как 9,4 больше 8,9, то в новую матрицу расстояний записываем 9,4. Подобным образом устанавливаются расстояние от нового кластера до всех оставшихся точек. Расстояния между всеми остальными точками (не вошедшими в кластер) записываются без изменений.

Таблица 7.12

Перерасчет матрицы расстояний после первого шага кластеризации методом полной связи

Точки	**1**	**2;3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	3,0	11,7	17,9	20,3	57,8	71,2	148,4	165,5
2;3	3,0	0	9,4	15,7	18,0	56,4	69,0	146,3	163,4
4	11,7	9,4	0	6,3	9,1	47,0	59,5	137,4	154,3
5	17,9	15,7	6,3	0	3,2	40,8	53,3	132,2	148,7
6	20,3	18,0	9,1	3,2	0	38,8	51,3	131,2	147,5
7	57,8	56,4	47,0	40,8	38,8	0	12,5	95,7	110,4
8	71,2	69,0	59,5	53,3	51,3	12,5	0	85,6	99,3
9	148,4	146,3	137,4	132,2	131,2	95,7	85,6	0	20,6
10	165,5	163,4	154,3	148,7	147,5	110,4	99,3	20,6	0

В итоге полученная пересчитанная матрица расстояний стала короче первоначальной на 1 столбец и 1 строку. В новой таблице снова производится нахождение самой короткой дистанции. Наименьшее расстояние (3,0) отмечено между кластером «2;3» и объектом 1. Производим «шаг 2» – объединение объектов в новый кластер «2;3;1». Снова пересчитаем матрицу расстояний согласно алгоритму «полной связи» основываясь на значениях из таблицы 7.12. Вписываем в новую матрицу (таблица 7.13) наибольшие расстояния для пары объектов «2;3» и 1. Так расстояние от точки 1 до объекта 4 равно 11,7, а от кластера «2;3» – 9,4, значит новый кластер «2;3;1» будет находиться на расстоянии 11,7 от кластера 4. Расстояние от точки 1 до объекта 5 равно 17,9, а от кластера «2;3» – 15,7. В новую матрицу записываем максимальное значение (17,9). Подобным образом находим расстояние от нового кластера до всех оставшихся точек.

Таблица 7.13

Перерасчет матрицы расстояний после второго шага кластеризации методом полной связи

Точки	2;3;1	4	5	6	7	8	9	10
2;3;1	0	11,7	17,9	20,3	57,8	71,2	148,4	165,5
4	11,7	0	6,3	9,1	47,0	59,5	137,4	154,3
5	17,9	6,3	0	3,2	40,8	53,3	132,2	148,7
6	20,3	9,1	3,2	0	38,8	51,3	131,2	147,5
7	57,8	47,0	40,8	38,8	0	12,5	95,7	110,4
8	71,2	59,5	53,3	51,3	12,5	0	85,6	99,3
9	148,4	137,4	132,2	131,2	95,7	85,6	0	20,6
10	165,5	154,3	148,7	147,5	110,4	99,3	20,6	0

Новая матрица расстояний снова короче предыдущей на 1 столбик и на 1 строчку. В ней также находим наименьшее расстояние, производим объединение наиболее близких объектов, далее осуществляем перерасчет матрицы расстояний по алгоритму «метода полной связи», как было осуществлено на первом и втором шаге кластеризации. Подобные манипуляции проводят пошагово до тех пор, пока в таблице не будут объединены все объекты (кластеры). На последнем шаге кластеризации матрица расстояний будет состоять из 2 столбцов и 2 строчек. Ниже последовательно приведены матрицы всех последующих шагов кластеризации (Таблицы 7.14-7.19).

Таблица 7.14

Перерасчет матрицы расстояний после третьего шага кластеризации методом полной связи

Точки	2;3;1	4	5;6	7	8	9	10
2;3;1	0	11,7	20,3	57,8	71,2	148,4	165,5
4	11,7	0	9,1	47,0	59,5	137,4	154,3
5;6	20,3	9,1	0	40,8	53,3	132,2	148,7
7	57,8	47,0	40,8	0	12,5	95,7	110,4
8	71,2	59,5	53,3	12,5	0	85,6	99,3
9	148,4	137,4	132,2	95,7	85,6	0	20,6
10	165,5	154,3	148,7	110,4	99,3	20,6	0

Таблица 7.15

Перерасчет матрицы расстояний после четвертого шага кластеризации методом полной связи

Точки	2;3;1	5;6;4	7	8	9	10
2;3;1	0	20,3	57,8	71,2	148,4	165,5
5;6;4	20,3	0	47,0	59,5	137,4	154,3
7	57,8	47,0	0	12,5	95,7	110,4
8	71,2	59,5	12,5	0	85,6	99,3
9	148,4	137,4	95,7	85,6	0	20,6
10	165,5	154,3	110,4	99,3	20,6	0

Таблица 7.16

Перерасчет матрицы расстояний после пятого шага кластеризации методом полной связи

Точки	2;3;1	5;6;4	7;8	9	10
2;3;1	0	20,3	71,2	148,4	165,5
5;6;4	20,3	0	59,5	137,4	154,3
7;8	71,2	59,5	0	95,7	110,4
9	148,4	137,4	95,7	0	20,6
10	165,5	154,3	110,4	20,6	0

Таблица 7.17

Перерасчет матрицы расстояний после шестого шага кластеризации методом полной связи

Точки	2;3;1;5;6;4	7;8	9	10
2;3;1;5;6;4	0	71,2	148,4	165,5
7;8	71,2	0	95,7	110,4
9	148,4	95,7	0	20,6
10	165,5	110,4	20,6	0

Таблица 7.18

Перерасчет матрицы расстояний после седьмого шага кластеризации методом полной связи

Точки	2;3;1;5;6;4	7;8	9;10
2;3;1;5;6;4	0	71,2	165,5
7;8	71,2	0	110,4
9;10	165,5	110,4	0

Таблица 7.19

Перерасчет матрицы расстояний после восьмого шага кластеризации методом полной связи

Точки	2;3;1;5;6;4;7;8	9;10
2;3;1;5;6;4;7;8	0	165,5
9;10	165,5	0

Таким образом, на последнем шаге происходит объединение 2-х оставшихся кластеров на расстоянии 165,5. После построения всех матриц расстояний и проведения всех шагов кластеризации, как и в случае применения алгоритма одиночной связи, необходимо построить дендрограмму (рисунок 7.5).

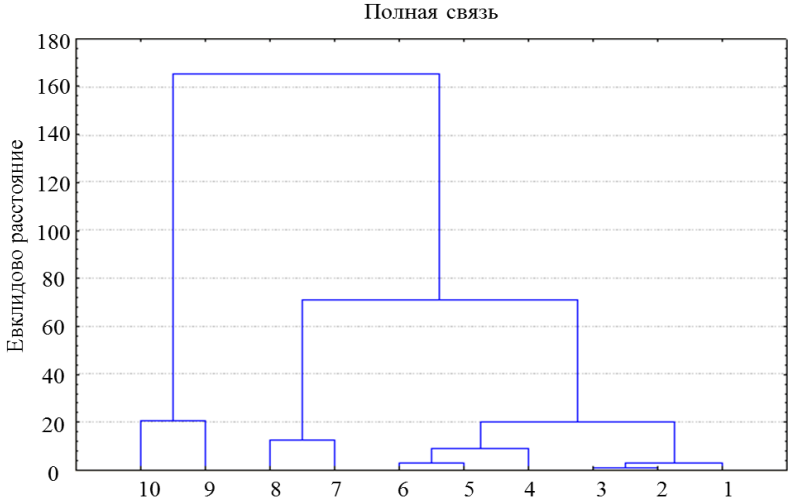


Рисунок 7.5. Дендрограмма кластеризации методом полной связи

Далее на дендрограмме выделяют количество групп объектов исследования в зависимости от целей и задач, стоящих перед исследователями. Одним из способов выделения кластеров является проведение линии разделения кластеров в месте резкого увеличения расстояния объединения кластеров (рисунок 7.6). На рисунке 7.6

показан один из вариантов разделения дендрограммы на отдельные группы.

Таким образом, на дендрограмме (рисунок 7.8) произведено выделение трех кластеров. Первый кластер объединяет объекты 1,2,3,4,5,6, второй – 7,8 и третий 9 и 10. Далее производят подробное изучение каждой отдельной группы объектов исследования, сравнение групп между собой и прочие манипуляции, зависящие от целей и задач исследования. Также можно рассмотреть группы объектов внутри кластеров – субкластеры.

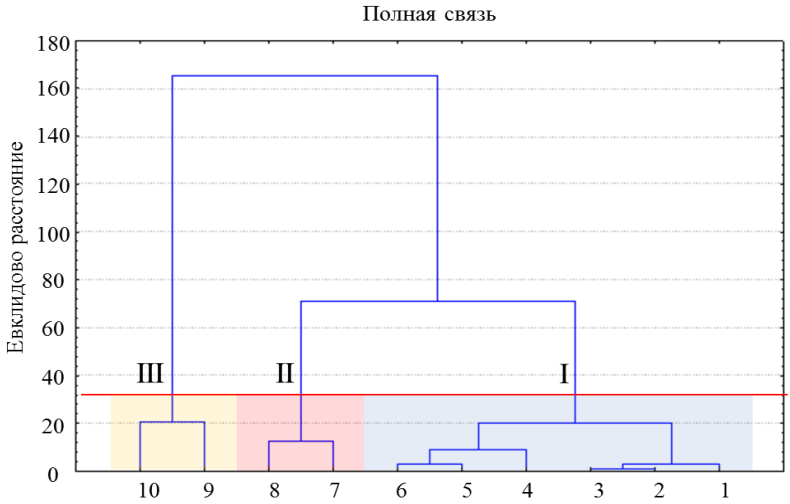


Рисунок 7.6. Один из способов выделения кластеров на дендрограмме

Метод невзвешенного попарного среднего арифметического (*unweighted pair-group method using arithmetic averages*, UPGMA). Этот метод также схож с описанными выше алгоритмами кластеризации. Также как и в предыдущих методах на каждом шаге происходит объединение наиболее близких объектов в кластеры. Различия метода заключаются в особенностях перерасчета расстояний от образованного кластера и остальных объектов.

Перерасчет расстояния будет осуществляться по формуле (7.4):

$$d(AB,C) = \frac{N_A d_{(A,C)} + N_B d_{(B,C)}}{N_A + N_B}, \quad (7.4)$$

где *d(AB,C)* – это расстояние от кластера, образованного путем соединения объектов (кластеров) A и B, до объекта (кластера) C;
N_A – количество объектов в кластере A;
N_B – количество объектов в кластере B;
d(A,C) – расстояние от кластера A до кластера C;
d(B,C) – расстояние от кластера B до кластера C.

На примере таблицы 7.3 по аналогии с предыдущими алгоритмами кластеризации проведем кластерный анализ методом невзвешенного попарного среднего арифметического. Итак, имея матрицу расстояний (таблица 7.3), на первом шаге кластеризации проведем объединение двух наиболее близких объектов: 2 и 3 (расстояние 1,1). Далее произведем перерасчет расстояний между образованным кластером «2;3» и оставшимися объектами по формуле (7.4).

Расчет расстояния от кластера «2;3» до объекта 1:

$$d(23,1) = \frac{1*2,2+1*3,0}{1+1} = 2,6.$$

Расчет расстояния от кластера «2;3» до объекта 4:

$$d(23,4) = \frac{1*9,4+1*8,9}{1+1} = 9,2.$$

Расчет расстояния от кластера «2;3» до объекта 5:

$$d(23,5) = \frac{1*15,7+1*15,0}{1+1} = 15,4.$$

Расчет расстояния от кластера «2;3» до объекта 6:

$$d(23,6) = \frac{1*18,0+1*17,3}{1+1} = 17,7.$$

Расчет расстояния от кластера «2;3» до объекта 7:

$$d(23,7) = \frac{1*56,4+1*55,8}{1+1} = 56,1.$$

Расчет расстояния от кластера «2;3» до объекта 8:

$$d(23,8) = \frac{1*69,0+1*68,3}{1+1} = 68,7.$$

Расчет расстояния от кластера «2;3» до объекта 9:

$$d(23,9) = \frac{1*146,3+1*146,1}{1+1} = 146,2.$$

Расчет расстояния от кластера «2;3» до объекта 9:

$d(23,9) = \frac{1*163,4+1*163,1}{1+1} = 163,3.$

После того, как все расстояния будут рассчитаны, произведем построение новой матрицы расстояний (таблица 7.20).

Таблица 7.20

Перерасчет матрицы расстояний после первого шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	1	2;3	4	5	6	7	8	9	10
1	0	2,6	11,7	17,9	20,3	57,8	71,2	148,4	165,5
2;3	2,6	0	9,2	15,4	17,7	56,1	68,7	146,2	163,3
4	11,7	9,2	0	6,3	9,1	47,0	59,5	137,4	154,3
5	17,9	15,4	6,3	0	3,2	40,8	53,3	132,2	148,7
6	20,3	17,7	9,1	3,2	0	38,8	51,3	131,2	147,5
7	57,8	56,1	47,0	40,8	38,8	0	12,5	95,7	110,4
8	71,2	68,7	59,5	53,3	51,3	12,5	0	85,6	99,3
9	148,4	146,2	137,4	132,2	131,2	95,7	85,6	0	20,6
10	165,5	163,3	154,3	148,7	147,5	110,4	99,3	20,6	0

Находим в новой матрице расстояний наименьшее расстояние. На втором шаге кластеризации объединяются кластер «2;3» и объект 1 (расстояние 2,6). Производим пересчет расстояний от нового кластера «2;3;1» до всех оставшихся объектов.

Расчет расстояния от кластера «2;3;1» до остальных объектов будет отличаться от того, как оно рассчитывалось на предыдущем шаге. Здесь стоит принять во внимание количество объектов, входящих в кластер «2;3». Он состоит из двух объектов: объекта 2 и объекта 3, – что должно быть отражено в расчетах далее.

Расчет расстояния от кластера «2;3;1» до объекта 4:

$d(231,4) = \frac{2*9,2+1*11,7}{2+1} = 10.$

Расчет расстояния от кластера «2;3;1» до объекта 5:

$d(231,5) = \frac{2*15,4+1*17,9}{2+1} = 16,2.$

Расчет расстояния от кластера «2;3;1» до объекта 6:

$d(231,6) = \frac{2*17,7+1*20,3}{2+1} = 18,6.$

Расчет расстояния от кластера «2;3;1» до объекта 7:

$$d(231,7) = \frac{2*56,1+1*57,8}{2+1} = 56,7.$$

Расчет расстояния от кластера «2;3;1» до объекта 8:

$$d(231,8) = \frac{2*68,7+1*71,2}{2+1} = 69,5.$$

Расчет расстояния от кластера «2;3;1» до объекта 9:

$$d(231,9) = \frac{2*146,2+1*148,4}{2+1} = 146,9.$$

Расчет расстояния от кластера «2;3;1» до объекта 10:

$$d(231,9) = \frac{2*163,3+1*165,5}{2+1} = 164.$$

После вычислений произведем составление новой матрицы расстояний (таблица 7.21), где также объединим в кластер наиболее близкорасположенные объекты.

Таблица 7.21

Перерасчет матрицы расстояний после второго шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	**2;3;1**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2;3;1	0	10,0	16,2	18,6	56,7	69,5	146,9	164
4	10,0	0	6,3	9,1	47,0	59,5	137,4	154,3
5	16,2	6,3	0	3,2	40,8	53,3	132,2	148,7
6	18,6	9,1	3,2	0	38,8	51,3	131,2	147,5
7	56,7	47,0	40,8	38,8	0	12,5	95,7	110,4
8	69,5	59,5	53,3	51,3	12,5	0	85,6	99,3
9	146,9	137,4	132,2	131,2	95,7	85,6	0	20,6
10	164	154,3	148,7	147,5	110,4	99,3	20,6	0

На третьем шаге кластеризации объединяются объекты 5 и 6 (расстояние 3,2).

Произведем расчет расстояний от кластера «5;6» до остальных объектов.

Расчет расстояния от кластера «5;6» до объекта 2;3;1:

$$d(56,231) = \frac{1*16,2+1*18,6}{1+1} = 17,4.$$

Расчет расстояния от кластера «5;6» до объекта 4:

$d(56,4) = \frac{1*6,3+1*9,1}{1+1} = 7,7.$

Расчет расстояния от кластера «5;6» до объекта 7:

$d(56,7) = \frac{1*40,8+1*38,8}{1+1} = 39,8.$

Расчет расстояния от кластера «5;6» до объекта 8:

$d(56,8) = \frac{1*53,3+1*51,3}{1+1} = 52,3.$

Расчет расстояния от кластера «5;6» до объекта 9:

$d(56,9) = \frac{1*132,2+1*131,2}{1+1} = 131,7.$

Расчет расстояния от кластера «5;6» до объекта 10:

$d(56,10) = \frac{1*148,7+1*147,5}{1+1} = 148,1.$

Рассчитанные значения вписываем в новую матрицу расстояний (таблица 7.22).

Таблица 7.22

Перерасчет матрицы расстояний после третьего шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	**2;3;1**	**4**	**5;6**	**7**	**8**	**9**	**10**
2;3;1	0	10,0	17,4	56,7	69,5	146,9	164
4	10,0	0	7,7	47,0	59,5	137,4	154,3
5;6	17,4	7,7	0	39,8	52,3	131,7	148,1
7	56,7	47,0	39,8	0	12,5	95,7	110,4
8	69,5	59,5	52,3	12,5	0	85,6	99,3
9	146,9	137,4	131,7	95,7	85,6	0	20,6
10	164	154,3	148,1	110,4	99,3	20,6	0

На четвертом шаге кластеризации снова объединяем наиболее близкие объекты (кластер 5;6 и объект 4). Производим расчет расстояний от нового кластера «5;6;4» до всех объектов.

Расчет расстояния от кластера «5;6;4;» до объекта 2;3;1:

$d(564,231) = \frac{2*17,4+1*10}{2+1} = 14,9.$

Расчет расстояния от кластера «5;6;5;» до объекта 7:

$d(564,7) = \frac{2*39,8+1*47}{2+1} = 42,2.$

Расчет расстояния от кластера «5;6;5;» до объекта 8:

$d(564,8) = \frac{2*52,3+1*59,5}{2+1} = 54,7.$

Расчет расстояния от кластера «5;6;5;» до объекта 9:

$d(564,9) = \frac{2*131,7+1*137,4}{2+1} = 133,6.$

Расчет расстояния от кластера «5;6;5;» до объекта 10:

$d(564,10) = \frac{2*148,1+1*154,3}{2+1} = 150,2.$

Рассчитанные значения вписываем в новую матрицу расстояний (таблица 7.23).

Таблица 7.23

Перерасчет матрицы расстояний после четвертого шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	**2;3;1**	**5;6;4**	**7**	**8**	**9**	**10**
2;3;1	0	14,9	56,7	69,5	146,9	164
5;6;4	14,9	0	42,2	54,7	133,6	150,2
7	56,7	42,2	0	12,5	95,7	110,4
8	69,5	54,7	12,5	0	85,6	99,3
9	146,9	133,6	95,7	85,6	0	20,6
10	164	150,2	110,4	99,3	20,6	0

На пятом шаге кластеризации снова объединяем наиболее близкие объекты (кластеры 7 и 8). Производим расчет расстояний от нового кластера «7;8» до всех объектов из таблицы (таблица 7.23).

Расчет расстояния от кластера «7;8;» до объекта 2;3;1:

$d(78,231) = \frac{1*56,7+1*69,5}{1+1} = 63,1.$

Расчет расстояния от кластера «7;8;» до объекта 5;6;4:

$d(78,564) = \frac{1*42,2+1*54,7}{1+1} = 48,5.$

Расчет расстояния от кластера «7;8;» до объекта 9:

$d(78,9) = \frac{1*95,7+1*85,6}{1+1} = 90,7.$

Расчет расстояния от кластера «7;8;» до объекта 10:

$d(78,10) = \frac{1*110,4+1*99,3}{1+1} = 104,9.$

Рассчитанные значения вписываем в новую матрицу расстояний (таблица 7.24).

Таблица 7.24

Перерасчет матрицы расстояний после пятого шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	2;3;1	5;6;4	7;8	9	10
2;3;1	0	14,9	63,1	146,9	164
5;6;4	14,9	0	48,5	133,6	150,2
7;8	63,1	48,5	0	90,7	104,9
9	146,9	133,6	90,7	0	20,6
10	164	150,2	104,9	20,6	0

На шестом шаге кластеризации объединяются кластеры 2;3;1 и 5;6;4 (расстояние 14,9).

Произведем расчет расстояний от кластера «2;3;1;5;6;4» до остальных объектов.

Расчет расстояния от кластера «2;3;1;5;6;4» до объекта 7;8. Так как кластер «2;3;1;5;6;4» состоит из двух субкластеров по 3 объекта, то расчет расстояния будет осуществляться следующим образом:

$d(231564{,}78) = \frac{3*63{,}1+3*48{,}5}{3+3} = 55{,}8.$

Расчет расстояния от кластера «2;3;1;5;6;4» до объекта 9:

$d(231564{,}9) = \frac{3*146{,}9+3*133{,}6}{3+3} = 140{,}3.$

Расчет расстояния от кластера «2;3;1;5;6;4» до объекта 10:

$d(231564{,}10) = \frac{3*164+3*150{,}2}{3+3} = 157{,}1.$

Рассчитанные значения вписываем в новую матрицу расстояний (таблица 7.25).

Таблица 7.25

Перерасчет матрицы расстояний после шестого шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	2;3;1;5;6;4	7;8	9	10
2;3;1;5;6;4	0	55,8	140,3	157,1
7;8	55,8	0	90,7	104,9
9	140,3	90,7	0	20,6
10	157,1	104,9	20,6	0

На седьмом шаге кластеризации объединяются кластеры 9 и 10 (расстояние 20,6).

Далее необходимо снова произвести расчет расстояний от кластера «9;10» до остальных объектов.

Расчет расстояния от кластера «9;10;» до кластера «2;3;1;5;6;4»:

$$d(910,231564) = \frac{1*140,3+1*157,1}{1+1} = 148,7.$$

Расчет расстояния от кластера «9;10;» до кластера «7;8»:

$$d(910,78) = \frac{1*90,7+1*104,9}{1+1} = 97,8.$$

Рассчитанные значения вписываем в новую матрицу расстояний (таблица 7.26).

Таблица 7.26

Перерасчет матрицы расстояний после седьмого шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	**2;3;1;5;6;4**	**7;8**	**9;10**
2;3;1;5;6;4	0	55,8	148,7
7;8	55,8	0	97,8
9;10	148,7	97,8	0

На пятом шаге кластеризации снова объединяем наиболее близкие объекты: кластеры «2;3;1;5;6;4» и «7;8». Произведем расчет расстояний от нового кластера «2;3;1;5;6;4;7;8» до кластера «9;10», учитывая, что первый субкластер «2;3;1;5;6;4» состоит из 6 объектов, а второй «7;8» – из 2-х.

$$d(23156478,910) = \frac{6*148,7+2*97,8}{6+2} = 136.$$

Рассчитанное значение впишем в последнюю матрицу расстояний (таблица 7.27).

Таблица 7.27

Перерасчет матрицы расстояний после восьмого шага кластеризации методом невзвешенного попарного среднего арифметического

Точки	**2;3;1;5;6;4;7;8**	**9;10**
2;3;1;5;6;4;7;8	0	136
9;10	136	0

На девятом шаге кластеризации соединяются оставшиеся кластеры: «2;3;1;5;6;4;7;8» и «9;10». Процесс кластеризации завершен. Теперь снова необходимо построить дендрограмму – график, демонстрирующий порядок объединения объектов в кластеры и расстояния на котором объекты объединились (рисунок 7.7). Далее на дендрограмме выделяют количество групп объектов исследования в зависимости от целей и задач, стоящих перед исследователями. Одним из способов выделения кластеров является проведение линии разделения кластеров в месте резкого увеличения расстояния объединения кластеров (рисунок 7.8). На рисунке 7.8 показан один из вариантов разделения дендрограммы на отдельные группы.

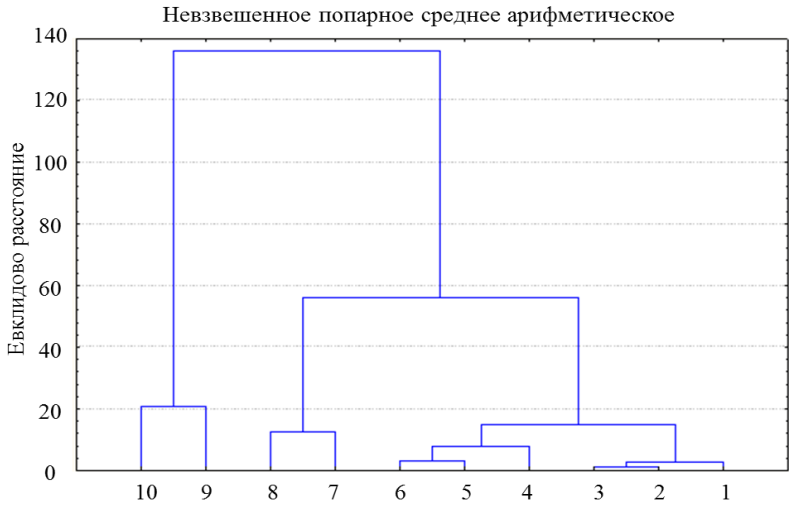


Рисунок 7.7. Дендрограмма кластеризации методом невзвешенного попарного среднего арифметического

Таким образом, на дендрограмме (рисунок 7.8) произведено выделение трех кластеров. Первый кластер объединяет объекты 1,2,3,4,5,6, второй – 7,8 и третий 9 и 10. Далее производят подробное изучение каждой отдельной группы объектов исследования, сравнение

групп между собой и прочие манипуляции, зависящие от целей и задач исследования.

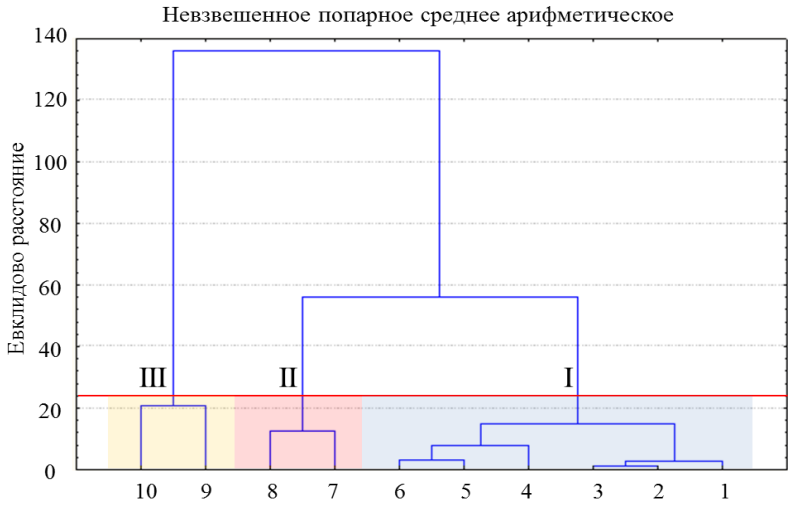


Рисунок 7.8. Дендрограмма кластеризации методом невзвешенного попарного среднего арифметического с выделением групп объектов исследования

Задания к разделу 7 для самостоятельного выполнения

1. Используя данные таблицы 7.28 произвести кластерный анализ методом одиночной связи с построением дендрограммы и выделением кластеров.

Таблица 7.28

Выбросы 5 предприятий, т/год

№	SO_2	N_xO_y	CO_z
1	7,5	3	15
2	4,5	4,3	12,2
3	2,4	7,1	10
4	12,5	9,9	13
5	5,5	8,6	14,2

2. Используя данные таблицы 7.29 произвести кластерный анализ методом полной связи с построением дендрограммы и выделением кластеров.

Таблица 7.29

Химический состав воды 5 озер зоны тундры

№	pH	Минерализация, мг/л	Кислород, мг/л
1	7,5	50	12
2	4,5	45	7
3	5	47	6,5
4	9,5	35	5,5
5	10	30	5

3. Используя данные таблицы 7.30 произвести кластерный анализ методом невзвешенного попарного среднего арифметического с построением дендрограммы и выделением кластеров.

Таблица 7.30

Характеристики населенных 5 населенных пунктов

№	Численность насления, тыс. чел.	Площадь, км^2	Площадь зеленых насаждений, км^2
1	7,5	8	9
2	4,5	3	1,5
3	5	5	2
4	15	12	5,5
5	12	10	5

Литература

Гмурман, В. Е. Теория вероятностей и математическая статистика: учебное пособие для бакалавров / В. Е. Гмурман. – Москва: Юрайт, 2013. – 479 с.

Елисеева, И. И. Статистика: учебник для вузов / И. И. Елисеева. – Москва: Издательсво Юрайт, 2011. – 565 с.

Животовский, Л. А. Показатели внутрипопуляционного разнообразия / Л. А. Животовский // Журнал общей биологии. – 1980. – Т. 41, № 6. – С. 828-836.

Калинина, В. Н. Теория вероятностей и математическая статистика: учебник для бакалавров / В. Н. Калинина. – Москва: Юрайт, 2013. – 472 с.

Лысенко, С. Н. Общая теория статистики: учебное пособие / С. Н. Лысенко, И. А. Дмитриева. – Москва: ИД ФОРУМ, НИЦ ИНФРА-М, 2013. – 208 с.

Полякова, В. В. Основы теории статистики / В. В. Полякова, Н. В. Шаброва. – Екатеринбург: Изд-во Уральского университета, 2015. – 148 с.

Суходольский, Г. В. Основы математической статистики для психологов: Учебник / Г. В. Суходольский. – СПб.: Изд-во С.-Петербургского университета, 1998. – 464 с.

Т-критерий Вилкоксона [Электронный ресурс] // Новый семестр. – 2006-2019. – Режим доступа: https://math.semestr.ru/group/wilcoxon.php. – (Дата обращения: 08.04.2019).

Billiet, P. Critical Values for the Mann-Whitney U-Test [Electronic resource] / P. Billiet // The Open Door Web Site. – 2003. – Uniform Resource Locator: http://www.saburchill.com/IBbiology/downloads/002.pdf. – (Access date: 10.04. 2019).

Bray, J. R. An ordination of upland forest communities of southern Wisconsin / J. R. Bray, J. T. Curtis // Ecological Monographs. – 1957. – Vol. 27, № 4. – P. 325-349.

Estivill-Castro, V. Why so many clustering algorithms: a position paper / V. Estivill-Castro // ACM SIGKDD Explorations Newsletter. – 2002. – Vol. 4 (1). – P. 65–75.

Jaccard, P. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines / P. Jaccard // Bulletin de la Société Vaudoise des Sciences Naturelles. – 1901. – Vol. 37. – P. 241-272.

Mann, H. B., Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other / H. B. Mann, D. R. Whitney // Annals of Mathematical Statistics. – 1947. – Vol. 18 (1). – P. 50-60.

Margalef, R. Information theory in ecology / R. Margalef // International Journal of General Systems. – 1958. – Vol. 3. – P. 36-71.

McQuitty, L. L. Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies / L. L. McQuitty // Educational and Psychological Measurement. – 1957. – Vol. 17. – P. 207-229.

Menhinick, E. F. A Comparison of some species-individuals diversity indices applied to samples of field insects / E. F. Menhinick // Ecology. – 1964. – Vol. 45, № 4. – P. 859-861.

Pearson, K. Notes on Regression and Inheritance in the Case of Two Parents / K. Pearson // Proceedings of the Royal Society of London, – 1895. – Vol. 58. – P. 240-242.

Pielou, E. C. Ecological diversity / E. C. Pielou. – New York: Gordon and Breach Science Publisher, 1975. – 165 p.

Shannon, C. E. A mathematical theory of communication / C. E. Shannon // The Bell System Technical Journal. – 1948. – Vol. 27. – P. 379-423.

Shannon, C. E. The mathematical theory of communication / C. E. Shannon, W. Weaver. – Illinois: University of Illinois, 1949. – 125 p.

Simpson, E. H. Measurement of diversity / E. H. Simpson // Nature. – 1949. – Vol. 163. – P. 688.

Sneath, P. H. A. The Applications of Computers to Taxonomy / P. H. A. Sneath // Journal of General Microbiology. – 1957. – Vol. 17. – P. 201-206.

Sokal, R, Michener, C. A statistical method for evaluating systematic relationships / R. Sokal, C. Michener // University of Kansas Science Bulletin. – 1958. – Vol. 38. P. 1409–1438.

Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter / T. Sørensen // Kongelige Danske Videnskabernes Selskab. – 1948. – Vol. 5, № 4. – P. 1-34.

Spearman, C. The proof and measurement of association between two things / C. Spearman // The American Journal of Psychology. – 1904. – Vol. 15, № 1. – P. 72-101.

Student. The probable error of a mean / Student // Biometrika. – 1908. – Vol. 6 (1). – P. 1-25.

Sur la liaison et la division des points d'un ensemble fini / J. Florek, J. Lukaszewicz, H. Steinhaus, S. Zybrzycki // Colloquia Mathematicae. – 1951. – Vol. 2, № 3-4. – P. 282-285.

Table of Critical Values: Pearson Correlation [Electronic resource] // Statistics Solutions . – 2019. – Uniform Resource Locator: https://www.statisticssolutions.com/ table-of-critical-values-pearson-correlation. – (Access date: 14.05.2019).

Values of the t-distribution (two-tailed) [Electronic resource] // MedCalc Software bvba. – 2019. – Uniform Resource Locator: https://www.medcalc.org/manual/t-distribution.php. – (Access date: 14.05.2019).

Wilcoxon, F. Individual Comparisons by Ranking Methods / F. Wilcoxon // Biometrics Bulletin. – 1945. – Vol. 1, № 6. – P. 80-83.

Учебное издание

***Городничев** Р.М., **Пестрякова** Л.А., **Ушницкая** Л.А. и др.*

Методы экологических исследований.
Основы статистической обработки данных

Учебно-методическое пособие

Печатается в авторской редакции
Дизайн обложки: *Р.М. Городничев*

Подписано в печать 18.06.19. Формат 60х84/16.
Печать цифровая. Печ.л. 4,25. Уч.-изд. 5,3. Тираж 50 экз. Заказ № 231.
Издательский дом Северо-Восточного федерального университета,
677891, г. Якутск, ул. Петровского, 5

Отпечатано с готового оригинал-макета в типографии Издательского дома СВФУ