МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВО «Кубанский государственный
аграрный университет имени И. Т. Трубилина»

Ю. Ю. Никифоренко

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ЭКОЛОГИИ И ПРИРОДОПОЛЬЗОВАНИИ

Учебное пособие

Под общей редакцией И. С. Белюченко

Краснодар
КубГАУ
2019

УДК 502.05:311(075.8)
ББК 20.1
Н 62

Р е ц е н з е н т ы :

О. А. Сущенко – ведущий инженер Управления инженерных изысканий АО «НИПИГАЗ», канд. биол. наук;

А. И. Мельченко – профессор кафедры прикладной экологии Кубанского государственного аграрного университета, д-р биол. наук

Никифоренко Ю. Ю.

Н 62 Статистические методы в экологии и природопользовании : учеб. пособие / Ю. Ю. Никифоренко ; под. общ. ред. И. С. Белюченко. – Краснодар : КубГАУ, 2019. – 88 с.

ISBN 978-5-907294-33-2

В учебном пособии рассматриваются вопросы статистической обработки данных экологических исследований. В основу положены технологические приемы обработки экологической информации с использованием электронных таблицах Microsoft Excel и пакета прикладных программ Statistica. Рассматриваются описательные статистики, проверка гипотез о типе распределения, сравнение средних, двухфакторный дисперсионный анализ.

Предназначено для обучающихся по направлению подготовки 05.04.06 Экология и природопользование.

УДК 502.05:311(075.8)
ББК 20.1

ISBN 978-5-907294-33-2

ВВЕДЕНИЕ

Учебное пособие по дисциплине «Статистические методы в экологии и природопользовании» направлено на оказание помощи магистрам, обучающимся по напрвлению 05.04.06 «Экология и природопользование», в процессе обработки и интерпретации количественных и качественных данных. Грамотно подобранные статистические методы позволяют получить достоверную информацию об исследуемом объетке, процессе или явлении. В пособии рассматриваются конкретные примеры и алгоритмы действия по расчету основных статистических характеристик и проведению различных способов анализа данных в программах Microsoft Excel и Statistica.

Особое значение в учебном пособии уделено разбору конкретных примеров, направленных на решение практических задач научной направленности. Приводятся конкретные статистические приемы и методы, помогающие работать с экологической информацией на этапах сбора, подготовки и обобщения данных. В пособии показаны основные возможности работы с электронными таблицами Microsoft Excel и программой Statistica; на простых примерах разобраны их функциональные возможности для целей биологических и экологических исследований.

Статситическая обработка данных, получаемых в ходе научных исследований, позволяет качественно проанализировать результаты и сформулировать достоверные выводы. Умение применять методы статистики является незаменимым навыком при подготовке обучающимися научных статей, курсовых проектов, магистерских диссертаций и т. п. Учебное пособие охватывает основную тематику курсов, изучаемых в рамках направления «Экология и природопользование» и может применяться как для проведения практических и лабораточных занятий, так и для подготовки лекционного материала.

Тема 1. ВВЕДЕНИЕ В СТАТИСТИЧЕСКИЙ АНАЛИЗ В ЭКОЛОГИИ

Использование методов математической статистики в ходе решения различных задач экологической и биологичекой направленности является незаменимой составляющей научного исследования. Значительный объем количественной информации в условиях современных темпов развития науки, требует качественной обработки и интерпретации результатов. Набор определяемых статистических показателей зависит от направленности конкретного исследования и характеристик изучаемых объектов.

Количественный метод является одним из основных в экологии и природопользовании. Интерпретация исследуемых процессов и явлений с использованием статистических расчетов является надежным инструментом для подтверждения или опровержения выдвигаемых гипотез, доказательства теоретических положений, установления причинно-следственных связей и зависимостей, определения влияния факторов среды на свойства живых организмов и экосистем в целом.

Экологическая информация представляет собой набор сведений о концентрации веществ, размерах, весе, возрасте, численности, биомассе, плодовитости организмов, продуктивности экосистем, урожайности сортов, концентрации веществ, соотношении между признаками, дозами факторов, различными количественными показателями и числовыми характеристиками.

На этот кажущийся первоначально хаотичным набор первичной числовой информации накладываются свойства самих объектов изучения, усиливающие разброс данных, в частности широкая изменчивость живых систем. Современная статистика оказывается столь полезной при обработке численных данных в биологии и экологии именно потому, что она основана на признании этой изменчивости и обладает мощными средствами её учета. В итоге, в кажущемся хаосе полу-

ченных цифр вдруг открываются конкретные закономерности, которые требуют объективной оценки. Подтверждение существования закономерного в видимом хаосе изменчивости достигается посредством использования методов статистического анализа.

Применение прикладных методов статистики к сложным живым системам способствовало появлению нового направления в биологических науках и математике, которое получило название «биометрия». Кроме данного общепризнанного термина, использовались и используются другие – биометрика, вариационная статистика, биологическая статистика, биоматематика, в последнее время компьютерная биометрия. Несмотря на разнообразие понятий, суть данного научного направления остается фактически одной и той же – статистическая обработка результатов наблюдений и экспе-риментов в биологических науках (к коим относится и экология) с целью отделения закономерного от случайного, оценки разнообразных связей и зависимостей между биологическими явлениями, поиска причин, определения влияния фактора и т. д.

В основе биометрии лежат такие разделы математики, как теория вероятностей и математическая статистика. Другой путь называется дедуктивным подходом, при котором на первое место выдвигаются математические модели, основанные на теоретических обобщениях, с последующей проверкой моделей опытом. Этот путь «обслуживается» так называемой математической биологией, исследующей теоретические проблемы с помощью математического моделирования.

Характерной особенностью биометрии является применимость её методов не к единичным фактам, а только к их совокупностям, к массовым явлениям. Именно в сфере массовых случайных явлений обнаруживаются закономерности, не свойственные единичным объектам. В этом плане область приложения статистических методов в биологии и экологии очень значительна, так как многие экологические и биологические явления массовы по своей природе – в них участвуют

не одна клетка, не одна особь, не одна бактерия, не один вид или популяция, а их совокупности, взаимодействующие между собой. Осуществление событий в таких совокупностях может быть оценено вероятностями. Такие проблемы, как изменчивость морфологических, физиологических, экологических признаков животных и растений, возрастная изменчивость органов у человека, установление влияния экологических факторов, количественный учет организмов, классификационные построения в систематике, изучение наследственности в генетике, индивидуальный рост организмов, популяционная динамика численности, особенности сукцессии экосистем, могут изучаться лишь с помощью математических и математико-статистических методов.

В тех областях биологических наук, где исследования проводятся на основе измерений и подсчетов, игнорирование статистической обработки полученного исследователем материала может привести к ошибочным выводам. Напротив, корректное применение биометрических методов увеличивает доказательность сделанных заключений, помогает правильно планировать эксперименты, выявлять скрытые закономерности и правильно их интерпретировать, устанавливать причины наблюдаемых явлений, отделять их от следствий, выделять из множества воздействующих на явление факторов наиболее важные, измерять силу их влияния, дает возможность получить точную количественную характеристику изменчивости исследуемых показателей, оценить достоверность проверяемой гипотезы, определить степень различий между анализируемыми признаками.

Несмотря на ценность применения методов статистики в биологии и экологии, существуют некоторые нюансы, на которые необходимо обращать внимание при использовании методов математической статистики в экологии.

Первая из них – это механическое использование количественных методов анализа в исследованиях, без понимания их сути и приложимости к тем или иным биологическим явле-

ниям и экологическим процессам. Очень важно знать и учитывать особенности и условия применения тех или иных статистических процедур, поскольку любой метод имеет свои ограничения.

Без учета этих ограничений применение соответствующего метода становится математически неправомерно, это приводит к фальсификации результатов и выводов научной работы, к отклонению проверяемой гипотезы там, где на самом деле её нужно было бы принять, к установлению влияния фактора, который в реальности не влияет, к подтверждению не существующих связей между элементами системы. Описаные фальсификации могут возникать при формальном применении биометрических методов с целью создать лишь видимость строгой научности в той или иной исследовательской работе.

Вторая опасность связана с широко распространенным мнением о том, что математическая обработка данных может если не полностью учесть, то свести к минимуму те технические, организационные и методические ошибки, которые возникли при проведении исследования. На это часто надеются недобросовестные исследователи. Данное мнение глубоко ошибочно, статистические методы можно с равным успехом применять как к верным данным, так и к неправильно полученным.

Контрольные вопросы

1. Для каких целей в экологии применяются методы математической статистики?

2. Какие потенциальные задачи решаются в процессе математической обработки экологических данных?

3. Выделите основные проблемы, возникающие на этапе обработки количественной информации.

4. Что поспособствовало появлению в науке нового напрвления под названием «биометрия»?

Тема 2. ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ

Первичная обработка эколого-биологической информации представляет собой набор количественных статистических методов. На первом этапе обработки первичной информации осуществляется упорядочивание информации об объекте и предмете изучения. На этой стадии «сырые» сведения группируются по тем или иным критериям, заносятся в сводные таблицы. Первично обработанные данные, представленные в удобной форме, дают исследователю в первом приближении понятие о характере всей совокупности данных в целом: об их однородности – неоднородности, компактности – разбросанности, четкости – размытости и т. д. Эта информация хорошо считывается с наглядных форм представления данных и дает сведения об их распределении.

В ходе применения первичных методов статистической обработки получаются показатели, непосредственно связанные с производимыми в исследовании измерениями.

2.1 Правила составления сводных таблиц

В большинстве случаев обработку данных экологического мониторинга целесообразно начать с составления таблиц (сводных таблиц) полученных данных.

Как правило, по строкам такой таблицы занесены значения показателей в определенной точке отбора проб, а по столбцам расположены значения каждого заносимого в таблицу признака (измеренного параметра) − в одном столбце находятся значения одного признака по всем точкам отбора проб. Все строки и все столбцы должны быть пронумерованы. Последовательность признаков может быть упорядочена по разным критериям (например, по времени). Пример такой таблицы представлен на рисунке 1.

	A	B	C	D	E	F	G	H
1	ID	Лето 2001	Осень 2001	Среднее	Весна 2002	Лето 2002	Осень 2002	Среднее
2	1.1	3,67	4,24	3,95	5,73	4,44	5,64	5,27
3	1.1.б	4,56	4,12	4,34	5,06	4,88	5,06	5,00
4	1.1.д	3,36	6,46	4,91	5,70	4,65	6,28	5,54
5	1.11	5,33	4,37	4,85	4,27	4,35	5,23	4,62
6	1.11.б	4,12	4,25	4,18	4,26	3,82	6,14	4,74
7	1.11.д	3,42	3,52	3,47	5,57	5,17	3,28	4,67
8	1.13	3,57	4,72	4,14	4,75	4,06	5,34	4,72
9	1.15	5,82	4,26	5,04	5,74	5,17	4,65	5,19
10	1.16.б	3,75	3,63	3,69	4,65	5,54	4,27	4,82
11	1.16.д	3,62	4,70	4,16	4,58	4,36	5,14	4,69
12	1.17	5,22	4,73	4,97		4,92	5,53	5,23
13	1.19	4,15	4,85	4,50	5,32	4,80	5,27	5,13
14	1.21	4,85	3,93	4,39	5,04	4,35	5,38	4,92
15	1.21.б	4,25	4,87	4,56	4,17	3,15	5,63	4,32
16	1.21.д	3,86	5,95	4,90	4,86	4,52	5,17	4,85
17	1.23	3,74	3,85	3,79	5,23	3,27	4,87	4,46
18	1.25	6,47	3,82	5,14	5,25	5,62	5,36	5,41
19	1.3	1,82	4,82	3,32	4,62	5,52	5,47	5,20
20	1.5	5,36	5,23	5,30	5,33	3,27	5,67	4,76
21	1.6.б	5,52	4,03	4,78	4,63	4,57	4,65	4,62
22	1.6.д		4,92	4,92	4,72	5,78	5,42	5,31
23	1.7	3,82	4,12	3,97	6,25	4,46	5,84	5,52
24	1.9	5,64	4,64	5,14	5,07	4,73	4,72	4,84
25	2.1		4,54	4,54	5,08	6,12	7,82	6,34
26	2.1.б	4,16	4,27	4,21	6,12	3,66	5,14	4,97
27	2.1.г	4,37	3,85	4,11	5,04	5,22	5,36	5,21
28	2.11	5,32	4,84	5,08	4,84	6,24	4,92	5,33
29	2.11.б	4,08	6,12	5,10	4,85	3,80	5,72	4,79
30	2.11.г	3,07	4,87	3,97	6,14	4,45	5,46	5,35
31	2.13	2,65	5,86	4,26	4,57	4,65	6,63	5,28
32	2.15	6,24	3,62	4,93	5,77	5,20	5,76	5,58

Рисунок 1 – Пример заполнения данных для статистической обработки

На рисунке представлены данные по содержанию тяжелых металлов в почвах на территории определенного хозяйства. По строкам расположены точки отбора проб, а по столбцам содержание подвижной формы меди в различные сезоны года.

2.2 Проверка данных

После создания таблицы на бумаге или в электронных таблицах на компьютере необходимо проверить качество полученных данных. Для этого часто достаточно внимательно осмотреть массив данных. Начать проверку следует с выявления ошибок (описок), которые заключаются в том, что неправильно написан порядок числа. Например, 100 написано вместо 10; 9,4 вместо 94 и т. п. При внимательном просмотре по столбцам это легко обнаружить, поскольку сравнительно ред-

ко встречаются параметры, которые сильно варьируют. Чаще всего значения одного параметра имеют один порядок или ближайшие порядки. При наборе данных на компьютере важно соблюдать требования к формату данных в используемой статистической программе. Прежде всего, это относится к знаку, который должен отделять в десятичном числе целую часть от дробной (точка или запятая).

Затем массив данных надо проверить на наличие «*выскакивающих» вариант* − выделяющихся значений, которые могли быть получены в результате неточных измерений, ошибок в записях, отвлечения внимания испытуемого и т.д. Если обнаружены «подозрительные» значения, то нужно принять обоснованное решение об их выбраковке. Его можно принять, используя достаточно мощный параметрический критерий t *(критерий выпада).* Он рассчитывается по следующей формуле:

$$t = \frac{V - M}{\sigma} \geq t_{кр} \ , \qquad (1)$$

где t − критерий выпада; V − выпадающее значение признака; σ – стандартное отклонение; M − средняя величина признака для всей группы; $t_{кр}$ − стандартные значения критерия выпадов.

Стандартное значение критерия выпадов ($t_{кр}$) определяется для трех уровней доверительной вероятности по специальной таблице «Значения критерия t для обработки выпадающих вариант при разных уровнях значимости (p)».

Смысл критерия в том, чтобы определить, находится ли данная варианта в интервале, характерном для большинства членов выборки, или же вне его.

Допустим, нами принят уровень значимости p = 0,05 (доверительная вероятность 0,95), а значение критерия t составило 1,5. Поскольку 95 % вариант лежат в пределах M ± 1,96 σ

(1,5 меньше 1,96), следовательно, данная варианта лежит в указанном интервале. Если же значение критерия больше, например, 2,4, то это означает, что данное значение не относится к анализируемой совокупности (выборки), включающей 95 % вариант, а есть проявление иных закономерностей, ошибок и поэтому должно быть исключено из рассмотрения.

После исключения выпадающих значений первичные статистические параметры вычисляются заново.

Контрольные вопросы

1. Введение в статистический анализ в экологии. Цели, круг потенциально решаемых задач, примеры конкретных приложений, компьютерные программы.

2. Этапы технологического процесса автоматизированной обработки экологической информации.

3. Средства автоматизации обработки данных. Базы данных дистрибутивной информации.

4. Ошибки в данных, их природа и устранение.

5. Обзор современных пакетов математической и статистической обработки данных.

6. Использование некоторых пакетов для обработки экологической информации на ПК.

7. Основные принципы записи информации для электронных таблиц, статистических пакетов и баз данных.

8. Правила составления сводных таблиц. Проверка данных.

Тема 3. ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Описательная статистика – это техника сбора и суммирования количественных данных, которая используется для превращения массы цифровых данных в форму, удобную для восприятия и обсуждения. Методы описательной статистики позволяют оценить точность и достоверность полученных результатов и избежать ошибочных выводов.

Цель описательной статистики – обобщение первичных результатов, полученных в результате наблюдений и экспериментов и представление набора данных для последующего анализа. Важнейшим моментов при этом является сравнение различных объектов (например, ключевых участков мониторинга, различающихся степенью антропогенной нагрузки). При этом очень важно уметь доказать, что обнаруженное различие действительно существует, а не обусловлено статистической погрешностью оценки. На этом этапе анализа необходимо полученные результаты сравнить с литературными данными, с ПДК, фоновыми значениями и т. д.

3.1 Расчет описательных статистик при помощи электронных таблиц Microsoft Excel

Несмотря на то, что Excel существенно уступает специализированным статистическим пакетам обработки данных, тем не менее, этот раздел математики представлен здесь наиболее полно и позволяет проводить необходимый статистический анализ экологических данных.

При рассмотрении применения методов обработки статистических данных в данной лабораторной работе ограничимся только простейшими и наиболее часто встречаемыми описательными статистиками, реализованными в мастере функций Excel.

Функция СРЗНАЧ вычисляет среднее арифметическое из нескольких массивов (аргументов) чисел. *Среднее значение* случайной величины представляет собой наиболее типичное,

наиболее вероятное ее значение, своеобразный центр, вокруг которого разбросаны все значения признака.

Функция ДИСП позволяет оценить дисперсию по выборочным данным. *Дисперсия* является мерой изменчивости, вариации признака и представляет собой средний квадрат отклонений каждого из значений признака от среднего значения по всей совокупности. В отличие от других показателей вариации дисперсия может быть разложена на составные части, что позволяет тем самым оценить влияние различных факторов на вариацию признака.

Функция СТАНДОТКЛОН вычисляет стандартное отклонение. *Стандартное отклонение* (или среднее квадратическое отклонение) является мерой изменчивости (вариации) признака. Оно показывает, на какую величину в среднем отклоняются данные от среднего значения признака.

Если совокупность неоднородна, следует исключить из нее самые «аномальные» наблюдения, поскольку они, скорее всего, нетипичны для данного исследования и могут повлиять на дальнейшие результаты. Для устранения аномальных наблюдений используется правило «трех сигм»: наблюдение признается аномальным и отбрасывается, если его отклонение от выборочной средней более чем в 3 раза превышает среднеквадратическое отклонение выборки, т. е. не удовлетворяют неравенству: $\left|\bar{x} - x_i\right| \le 3\sigma$.

Стандартная ошибка среднего. Стандартная ошибка среднего это величина, на которую отличается среднее значение выборки от среднего значения генеральной совокупности при условии, что распределение близко к нормальному.

95%-й доверительный интервал для среднего. Интервал, в который с вероятностью 0,95 попадает среднее значение признака генеральной совокупности.

К сожалению, пакет Microsoft Excel не рассчитывает такие часто применяемые статистики, как коэффициент вариации и относительная ошибка среднего значения (точность опыта). Но их определение не представляет большого труда. Коэффициент

вариации (%) – это отношение стандартного отклонения к среднему значению, умноженное на 100 %:

$$V = \frac{S_x}{\bar{x}} \cdot 100\,\% \qquad (2)$$

Коэффициент вариации, как дисперсия и стандартное отклонение, является показателем изменчивости признака. Коэффициент вариации не зависит от единиц измерения, поэтому удобен для сравнительной оценки различных статистических совокупностей. При величине коэффициента вариации составляющего до 10 % изменчивость оценивается как слабая, 11 – 25 % – средняя, более 25 % – сильная (Лакин, 1990).

Относительная ошибка среднего значения (%) – отношение стандартной ошибки среднего к среднему значению, умноженное на 100 %.

$$C_S = \frac{S_{\bar{x}}}{\bar{x}} \cdot 100\%. \qquad (3)$$

Процент расхождения между генеральной и выборочной средней показывает, на сколько процентов можно ошибиться, если утверждать, что генеральная средняя равна выборочной средней. Если относительная ошибка не превышает 5 %, то точность исследований (точность опыта) оценивается как хорошая, до 10 % – удовлетворительная. Точность 3–5 % при вероятности 0,95, а в некоторых случаях и при вероятности 0,68, является вполне достаточной для большинства задач экологического мониторинга.

В пакете Microsoft Excel помимо мастера функций имеется набор более мощных инструментов для работы с несколькими выборками и углубленного анализа данных, называемый ***Пакет анализа***, который может быть использован для решения задач статистической обработки выборочных данных.

Рассмотрим пример расчета основных статистических характеристик для данных экологического исследования почв на территории 1-го отделения учебного хозяйства «Кубань» Кубанского ГАУ (таблица 1).

Таблица 1 – Данные экологического мониторинга почв на территории 1-го отделения учхоза «Кубань» Кубанского ГАУ

№ п/п	№ пробы	Орг. вещ-во, %	Микроорганизмы, $*10^9$ экз./г	NO_3, мг/кг	P_2O_5, мг/кг	Мезофауна, экз./10кг	Физ. глина, %
1	1.4	3,7	5,1	23,0	16,3	0,2	72,4
2	1.10	3,5	0,8	17,0	28,6	0,2	70,3
3	2.4	3,3	1,7	78,0	24,4	3,8	69,8
4	2.9	3,2	1,8	38,0	18,3	3,2	69,7
5	3.6	3,1	3,3	7,0	32,7	0,2	68,7
6	3.7	3,1	0,9	21,0	26,8	0,6	68,7
7	4.3	4,0	1,0	11,0	92,6	0,8	73,9
8	4.6	3,5	1,5	32,0	27,9	0,6	72,5
9	5.1	3,7	1,1	27,0	26,4	4,4	70,8
10	5.6	3,0	2,8	62,0	56,1	0,6	67,9
11	5.4	2,7	1,0	7,0	37,9	1,8	66,9
12	6.4	2,9	0,3	46,0	72,5	1,0	68,1
13	7.0	3,5	0,8	30,0	56,3	1,8	71,1
14	7.1	3,6	1,7	21,0	22,7	1,8	71,8
15	7.2	3,8	4,5	18,0	32,4	1,2	74,2

Для установки пакета Анализ данных в Microsoft Excel сделайте следующее:

– в меню Сервис выберите команду Надстройки;

– в списке установите флажок Пакет анализа.

Для использования статистического Пакета анализа данных необходимо:

– указать курсором мыши на пункт меню Сервис и щелкнуть левой кнопкой мыши;

– в раскрывающемся списке выбрать команду Анализ данных (если команда Анализ данных отсутствует в меню Сервис, то необходимо установить в Microsoft Excel пакет анализа данных);

– выбрать строку Описательная статистика и нажать ОК;

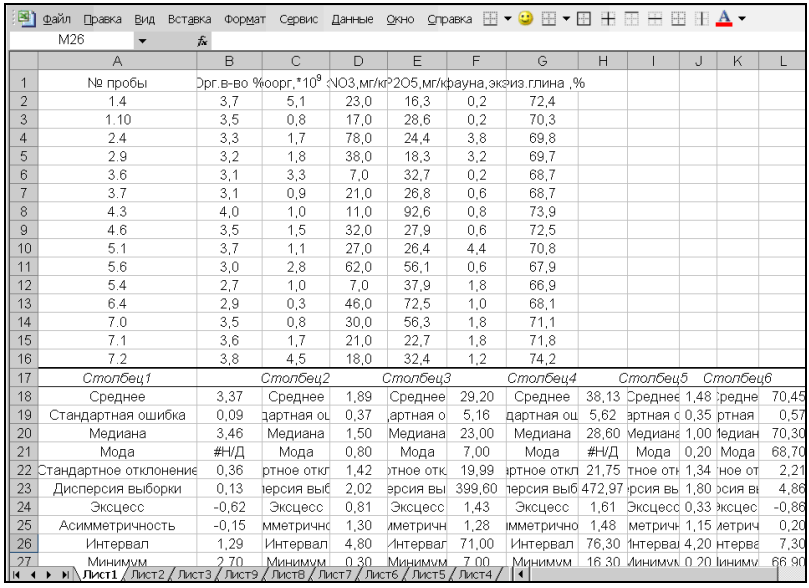
– в появившемся диалоговом окне указать входной интервал, то есть ввести ссылки на ячейки, содержащие анализируемые данные;

– указать выходной интервал, то есть ввести ссылку на ячейку, в которую будут выведены результаты анализа;

– в разделе Группирование переключатель установить в положение по столбцам или по строкам;

– установить флажок в поле Итоговая статистика и нажать ОК.

На рисунке 2 приводится пример расчета статистических характеристик, выполненного с помощью Пакета анализа Microsoft Excel для данных экологического мониторинга почв на территории 1-го отделения учебного хозяйства «Кубань» Кубанского ГАУ.



	A	B	C	D	E	F	G	H	I	J	K	L
1	№ пробы	Орг.в-во	%оорг,*10	NO3,мг/кг	P2O5,мг/кг	фауна,экз	из.глина ,%					
2	1.4	3,7	5,1	23,0	16,3	0,2	72,4					
3	1.10	3,5	0,8	17,0	28,6	0,2	70,3					
4	2.4	3,3	1,7	78,0	24,4	3,8	69,8					
5	2.9	3,2	1,8	38,0	18,3	3,2	69,7					
6	3.6	3,1	3,3	7,0	32,7	0,2	68,7					
7	3.7	3,1	0,9	21,0	26,8	0,6	68,7					
8	4.3	4,0	1,0	11,0	92,6	0,8	73,9					
9	4.6	3,5	1,5	32,0	27,9	0,6	72,5					
10	5.1	3,7	1,1	27,0	26,4	4,4	70,8					
11	5.6	3,0	2,8	62,0	56,1	0,6	67,9					
12	5.4	2,7	1,0	7,0	37,9	1,8	66,9					
13	6.4	2,9	0,3	46,0	72,5	1,0	68,1					
14	7.0	3,5	0,8	30,0	56,3	1,8	71,1					
15	7.1	3,6	1,7	21,0	22,7	1,8	71,8					
16	7.2	3,8	4,5	18,0	32,4	1,2	74,2					
17	*Столбец1*		*Столбец2*		*Столбец3*		*Столбец4*		*Столбец5*		*Столбец6*	
18	Среднее	3,37	Среднее	1,89	Среднее	29,20	Среднее	38,13	Среднее	1,48	Средне	70,45
19	Стандартная ошибка	0,09	дартная ош	0,37	артная о	5,16	дартная ош	5,62	артная о	0,35	ртная	0,57
20	Медиана	3,46	Медиана	1,50	Медиана	23,00	Медиана	28,60	Медиана	1,00	Медиан	70,30
21	Мода	#Н/Д	Мода	0,80	Мода	7,00	Мода	#Н/Д	Мода	0,20	Мода	68,70
22	Стандартное отклонение	0,36	ртное откл	1,42	отное отк	19,99	артное откл	21,75	тное отк	1,34	ное от	2,21
23	Дисперсия выборки	0,13	ерсия выб	2,02	ерсия вы	399,60	ерсия выб	472,97	рсия вы	1,80	сия вы	4,86
24	Эксцесс	-0,62	Эксцесс	0,81	Эксцесс	1,43	Эксцесс	1,61	Эксцесс	0,33	ксцесс	-0,86
25	Асимметричность	-0,15	имметрично	1,30	имметрично	1,28	имметрично	1,48	метрично	1,15	метрич	0,20
26	Интервал	1,29	Интервал	4,80	Интервал	71,00	Интервал	76,30	Интервал	4,20	нтерва	7,30
27	Минимум	2,70	Минимум	0,30	Минимум	7,00	Минимум	16,30	Минимум	0,20	Минимум	66,90

Рисунок 1 – Пример расчета статистических характеристик с помощью Пакета анализа Microsoft Excel

После расчета статистических характеристик заполняется итоговая выходная таблица, в которой указываются основные характеристики, необходимые для отчета. К ним чаще всего относятся средняя величина, ошибка средней, коэффициент вариации.

3.2 Приемы описательной статистики в пакете прикладных программ STATISTICA 6

3.2.1 Техника Box&Whisker Plot (коробочка с усиками) для предварительного (пилотного) анализа данных

Для визуализации описательных статистик можно построить статистические графики типа «коробок» (или «ящиков с усами»). Ящичковые диаграммы дают исследователю общее представление о распределении переменной: высота ящичка – разброс значений. Значения, не попавшие внутрь, изображаются отдельно вне ящика (рисунок 2).

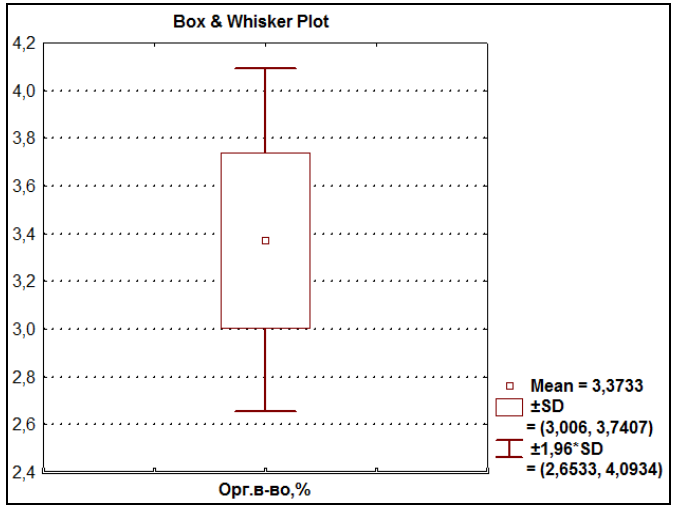


Рисунок 2 − Пример построения графика в технике Box&Whisker Plot

В пункте Discriptive Statistics (Описательная статистика) выберите Box & Whisker Plot и постройте графики для данных по органическому веществу: 1) медиана – квартили – лимиты (Median – Quart – Range); 2) среднее односигмовый предел – доверительные интервалы (Mean / SE / 1,96*SE). Перенесите полученный график в Microsoft Excel. Для этого щелкните по кнопке Файл в строке меню, выберите Copy, затем активизируйте Microsoft Excel и скопируйте туда график.

3.2.2 Построение гистограмм

Представим распределение переменных на гистограммах. Для этого предназаначена кнопка Histograms окна Descriptive statistics.

На гистограмму при необходимости можно наложить плотность нормального распределения, проверить близость распределения к нормальному виду при помощи критериев Колмогорова-Смирнова, Лилиефорса; вычислить статистику Шапиро-Уилкса. Для этого в группе опций Distribution необходимо установить флажок напротив соответствующих статистик. Значения статистик показываются прямо на гистограммах.

В пункте Statistics (Описательная статистика) проанализируйте распреление, например, органического вещества, (%). Вы получите график гистограммы и значения разных критериев с соответствующими им уровнями значимости (в данном случае на рисунке выведены значения критериев Колмогорова-Смирнова и Лиллифорса) (рисунок 3). Перенесите полученный график в Microsoft Excel. Для этого щелкните по кнопке Файл в строке меню, выберите Copy, затем активизируйте Microsoft Excel и скопируйте туда график.

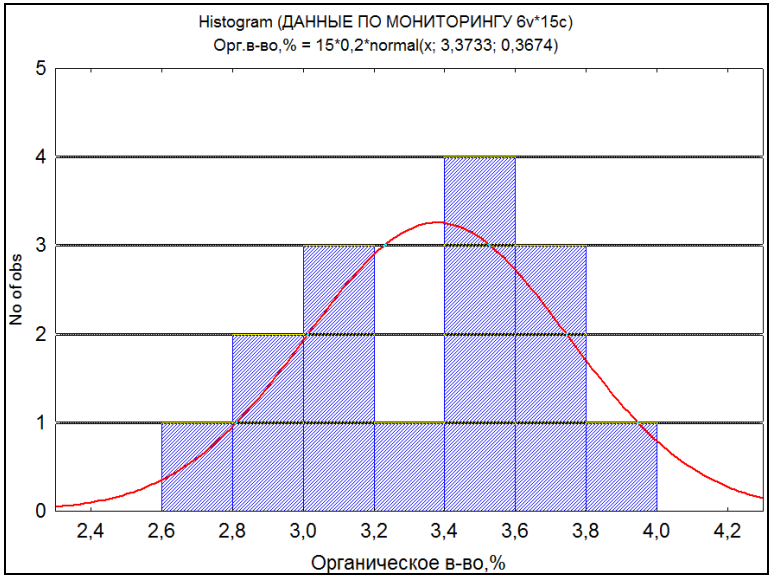
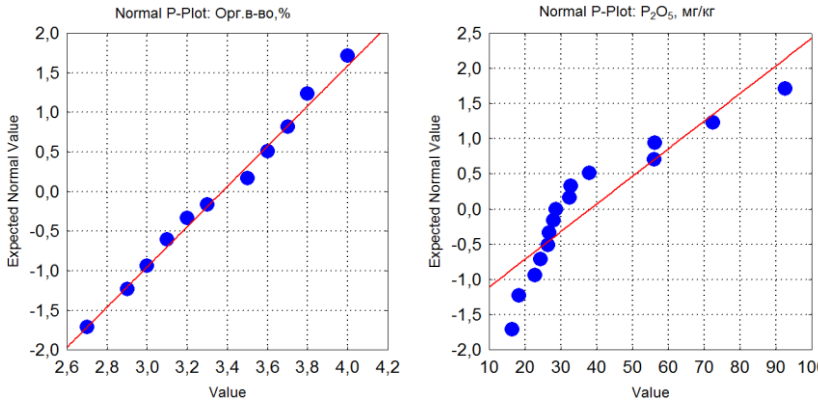


Рисунок 3 – Гистограмма распределения органического вещества (%) в почвах опытных участков

Критерий Колмогорова-Смирнова уместно применять в тех случаях, когда необходимо проверить, подчиняется ли наблюдаемая случайная величина некоторому закону распределения, известному с точностью до параметров. В случае неизвестных параметров гипотетического нормального распределения лучше пользоваться модификацией критерия Колмогорова − Смирнова, предложенной Стефенсом (Лиллифорсом). Отклонение от нормального распределения считается существенным при значении $p < 0,05$; в этом случае для соответствующих переменных следует применять непараметрические тесты. В рассматриваемом примере (значение $p > 0,2$), то есть вероятность ошибки является не значимой, поэтому значения переменной достаточно хорошо подчиняются нормальному распределению.

3.2.3 Техника Normal probability plot (NPP)

Для проверки гипотезы нормальности с помощью пакета прикладных программ Statistika 6.0 используется визуальный тест «график нормальных вероятностей» [Normal probability plot (NPP)]. Согласно применяемому тесту, идеально нормальным будет распределение, для которого точки, соответствующие наблюдаемым значениям, лежат точно на линии теоретической зависимости (как в случае с органическим веществом) (рисунок 4).



а б

Рисунок 4 – Тест на нормальный закон распределения данных по органическому веществу (а) и фосфатам (б)

Предположение о нормальном законе распределения используется во многих статистических методах обработки информации: в регрессионном анализе при установлении зависимостей между случайными величинами, в дисперсионном анализе при проверке статистических гипотез, поэтому при обработке данных рекомендуется обязательно его проверять.

Задания для самостоятельной работы

Задание 3. 1. На основании данных комплексного обследования полей рассчитать основные статистические характеристики, построить графические изображения и проверить распределение данных на нормальный закон.

Таблица 2 – Содержание цинка (Zn) в почвах сельхозугодий в зависмости от сезона года

№ п/п	Период проведения исследований		
	Весна, 2014	Лето, 2014	Осень, 2014
1	7,23	5,03	4,48
2	8,36	5,86	4,83
3	3,72	9,96	5,97
4	3,98	5,29	3,47
5	5,77	5,22	4,22
6	3,68	6,59	4,04
7	3,51	4,72	4,26
8	3,51	5,33	3,24
9	3,91	5,02	4,11
10	3,66	4,87	4,35
11	3,1	5,34	7,03
12	5,68	5,32	3,54
13	2,79	4,57	5,36
14	3,92	5,26	4,14
15	3,14	5,02	4,91
16	3,24	4,17	5,48
17	3,39	7,14	4,54
18	3,52	12,3	4,74
19	5,48	4,74	5,1
20	7,99	4,9	4,93
21	3,82	8,01	4,53
22	5,34	4,96	4,72
23	5,5	5,14	4,32
24	7,22	11,2	4,72

Тема 4. ПРОВЕРКА ГИПОТЕЗ О РАВЕНСТВЕ СРЕДНИХ

Одной из наиболее часто встречающихся задач при обработке данных является оценка достоверности отличий между двумя и более рядами значений. В математической статистике существует ряд способов для этого. Для использования большинства мощных критериев требуются дополнительные вычисления, обычно весьма развернутые.

Компьютерный вариант обработки данных стал в настоящее время наиболее распространенным. Во многих прикладных статистических программах есть процедуры оценки различий между параметрами одной выборки и разных выборок. Но ЭВМ дает исследователю принтерные распечатки, содержащие подсчитанные первичные статистики, результаты корреляционного анализа, иногда и факторного (компонентного). Основной анализ осуществляется позже, не в диалоге с ЭВМ. Перед экологом часто встает задача оценки достоверности различий, используя ранее вычисленные статистики. При сравнении средних значений признака говорят о достоверности (недостоверности) отличий средних арифметических, а при сравнении изменчивости показателей − о достоверности (недостоверности) отклонений сигм (дисперсий) и коэффициентов вариации.

Чтобы установить совпадение или различие характеристик экспериментальной и контрольной группы, формулируются статистические гипотезы:

– гипотеза об отсутствии различий (так называемая нулевая гипотеза);

– гипотеза о значимости различий (так называемая альтернативная гипотеза).

Для принятия решений о том, какую из гипотез (нулевую или альтернативную) следует принять, используют решающие правила – статистические критерии. То есть, на основании информации о результатах наблюдений (характеристиках чле-

нов экспериментальной и контрольной группы) вычисляется число, называемое эмпирическим значением критерия. Это число сравнивается с известным (например, заданным табличноо) эталонным числом, называемым критическим значением критерия.

Критические значения приводятся, как правило, для нескольких уровней значимости. Уровнем значимости называется вероятность ошибки, заключающейся в отклонении (не принятии) нулевой гипотезы, то есть вероятность того, что различия сочтены существенными, а они на самом деле случайны.

Обычно используют уровни значимости (α), равные 0,05, 0,01 и 0,001. В биологических и экологических исследованиях обычно ограничиваются значением 0,05, то есть, грубо говоря, допускается возможность ошибки не более чем на 5 %.

Если полученное исследователем эмпирическое значение критерия оказывается меньше или равно критическому, то принимается нулевая гипотеза – считается, что на заданном уровне значимости (то есть при том значении α, для которого рассчитано критическое значение критерия) характеристики экспериментальной и контрольной групп совпадают. В противном случае, если эмпирическое значение критерия оказывается строго больше критического, то нулевая гипотеза отвергается и принимается альтернативная гипотеза – характеристики экспериментальной и контрольной группы считаются различными с достоверностью различий 1 – α. Например, если α = 0,05 и принята альтернативная гипотеза, то достоверность различий равна 0,95 или 95 %.

Другими словами, чем меньше эмпирическое значение критерия (чем левее оно находится от критического значения), тем больше степень совпадения характеристик сравниваемых объектов. И наоборот, чем больше эмпирическое значение критерия (чем правее оно находится от критического значения), тем сильнее различаются характеристики сравниваемых объектов.

В дальнейшем мы ограничимся уровнем значимости $\alpha = 0{,}05$, поэтому, если эмпирическое значение критерия оказывается меньше или равно критическому, то можно сделать вывод, что «характеристики экспериментальной и контрольной групп совпадают с уровнем значимости 0,05».

Если эмпирическое значение критерия оказывается строго больше критического, то можно сделать вывод, что «достоверность различий характеристик экспериментальной и контрольной групп равна 95 %».

4.1 Критерий Стьюдента (t-тест)

Это параметрический метод, используемый для проверки гипотез о достоверности разницы средних при анализе количественных данных о совокупностях с нормальным распределением и с одинаковой дисперсией. К сожалению, метод Стьюдента слишком часто используют для малых выборок, не убедившись предварительно в том, что данные в соответствующих совокупностях подчиняются закону нормального распределения.

4.1.1 Метод Стьюдента для независимых выборок

Метод Стьюдента различен для независимых и зависимых выборок. Независимые выборки получаются при исследовании двух различных групп (например, данные с контрольного и фонового участка). С помощью критерия Стьюдента для независимых выборок можно было бы, например, проверить, существует ли достоверная разница между фоновыми уровнями содержания определенного вещества в почве и уровнями его содержания на экспериментальном участке.

В случае независимых выборок для выявления различий средних величин в больших выборках ($n > 30$) применяют формулу:

$$t_{набл} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{S^2_{\bar{x}_1} + S^2_{\bar{x}_2}}}. \qquad (3)$$

При сравнении двух групп с малыми выборками (n <30) величину критерия Стьюдента находят по формуле:

$$t_{набл} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{(n_1 - 1)S^2_{x_1} + (n_2 - 1)S^2_{x_2}}} \cdot \sqrt{\frac{n_1 \cdot n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}, \qquad (4)$$

где $\bar{x}_1$ и $\bar{x}_2$ – средние величины выборок; n_1 и n_2 – их объемы; $S^2_{x_1}$ и $S^2_{x_2}$ – дисперсии; $S_{\bar{x}_1}$; $S_{\bar{x}_2}$ – ошибки средних величин соответствующих выборок.

По специальной таблице (см. Г. Ф. Лакин «Биометрия», Приложение, табл. V стр. 270) или с помощью функции СТЬЮДРАСПРОБР мастера функций Microsoft Excel по принятому уровню значимости и числу степеней свободы $f=n_1+n_2-2$ находят *t* критическое и сравнивают эту величину с результатом расчета по формуле.

Если полученный результат больше, чем значение для уровня достоверности 0,05 (вероятность 5 %), найденное в таблице, то можно отбросить нулевую гипотезу (H_0) и принять альтернативную гипотезу (H_1), т.е. считать разницу средних достоверной.

Если же, напротив, полученный при вычислении результат меньше, чем табличный, то нулевую гипотезу нельзя отбросить и, следовательно, разница средних считается недостоверной.

Рассмотрим решение задачи сравнения средних величин по t-критерию в пакете Statistica.

Пример. Пусть, например, имеются результаты определения водопроницаемости почвы на 4 площадках с различным характером напочвенного покрова (таблица 3).

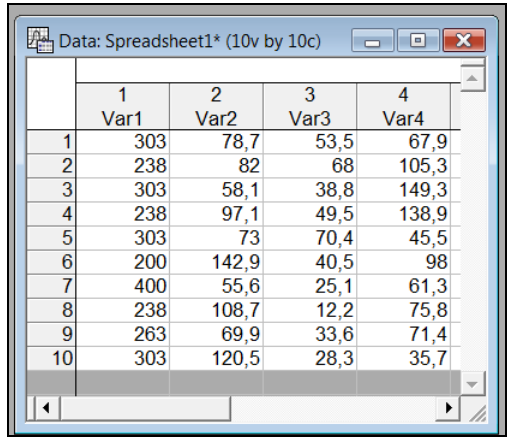
Таблица 3 – Данные водопроницаемости почвы с различным характером напочвенного покрова

Переменная			
VAR 1	VAR 2	VAR 3	VAR 4
303	78,7	53,5	67,9
238	82	68	105,3
303	58,1	38,8	149,3
238	97,1	49,5	138,9
303	73	70,4	45,5
200	142,9	40,5	98
400	55,6	25,1	61,3
238	108,7	12,2	75,8
263	69,9	33,6	71,4
303	120,5	28,3	35,7

Примечание: VAR 1 – водопроницаемость на площадке 1 (мертвый покров, лесная подстилка 2,5 см); VAR 2 – водопроницаемость на площадке 2 (травяной покров, проективное покрытие 40–50 %, задернение 10 %); VAR 3 – водопроницаемость на площадке 3 (травяной покров, проективное покрытие 100 %, задернение 70 %); VAR 4 – водопроницаемость на площадке 4 (травяной покров, проективное покрытие 30−40 %, задернения нет).

Используя этот критерий, можно определить, например, влияет ли характер напочвенного покрова на водопроницаемость почвы с ее поверхности.

Запустив программу Statistica, создадим файл с имеющимися данными. Для этого нажмите < Файл >, затем < Новый > и в открывшемся окне укажите число переменных – 4 (количество вариантов), число регистров – 10 (количество измерений на каждом из участков). Окно с файлом данных этого примера приводится на рисунке 4.



	1 Var1	2 Var2	3 Var3	4 Var4
1	303	78,7	53,5	67,9
2	238	82	68	105,3
3	303	58,1	38,8	149,3
4	238	97,1	49,5	138,9
5	303	73	70,4	45,5
6	200	142,9	40,5	98
7	400	55,6	25,1	61,3
8	238	108,7	12,2	75,8
9	263	69,9	33,6	71,4
10	303	120,5	28,3	35,7

Рисунок 4 − Окно с файлом данных по водопроницаемости почвы

Воспользуемся процедурой t-test dependent samples (меню второго уровня в закладке < Статистика > основного меню программы Statistica) для расчета средних величин водопроницаемости по вариантам опыта и одновременно проверим достоверность различий между средними значениями. Для этого необходимо при помощи кнопки Variables выбирать переменные для попарного сравнения. При этом должны быть выбраны переменные в обоих списках. Окно с файлом данных этого примера приводится на рисунке 5.

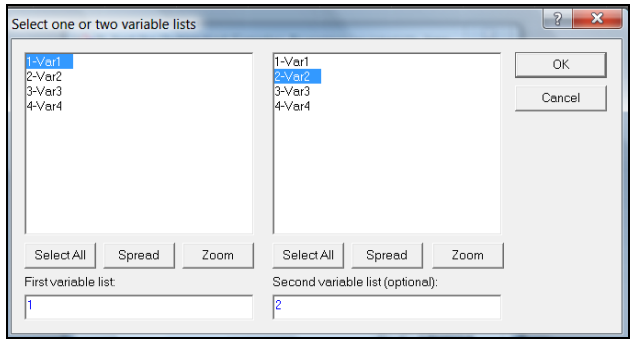


Рисунок 5 − Выбор переменных для попарного сравнения

После нажатия на кнопку Summary на экране появляется таблица с результатами сравнения по t-критерию. Если вероятность нулевой гипотезы (p) (различий в средней величине нет) меньше 5 % ($p < 0,05$), то с вероятность 0,95 нулевую гипотезу можно отбросить и считать, что различия достоверны на принятом уровне значимости.

Фрагменты окон с результатами проведения процедуры сравнения между 1-й и 2-й вариантами, а также между 2-й и 4-й приводятся на рисунках 6 и 7.

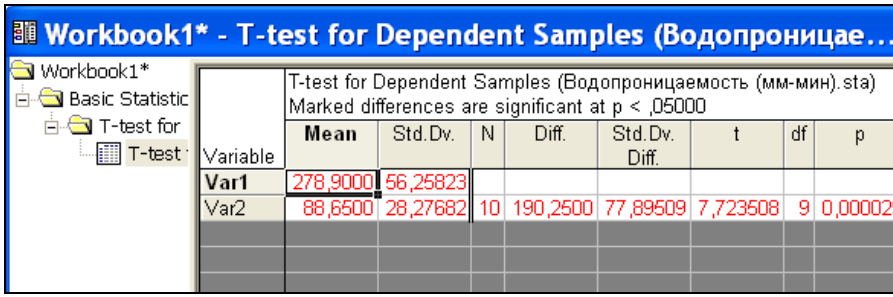


Рисунок 6 – Сравнение средних величин водопроницаемости почвы на 1-м и 2-м участках

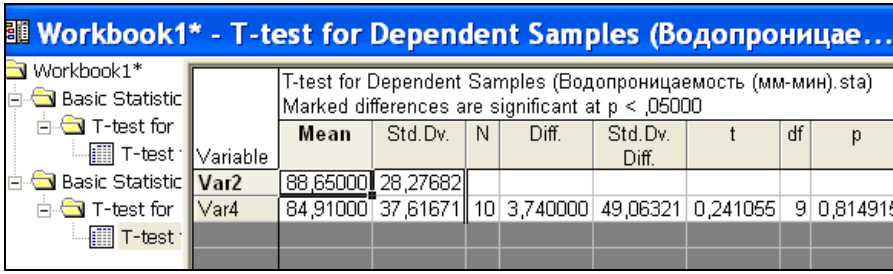


Рисунок 7 – Сравнение средних величин водопроницаемости почвы на 2-м и 4-м участках

Обычно достоверные различия средних величин в программе Statistica выделяются красным шрифтом. Попарное сравнение средних величин водопроницаемости показало достоверное различие между всеми вариантами опыта, кроме ва-

риантов 2 и 4. Нулевую гипотезу в последнем случае отбросить нельзя, так как ее вероятность достаточно высока ($p \approx 0{,}8149$).

Таким образом, окончательный вывод решения задачи о влиянии характера напочвенного покрова на водопроницаемость почвы с ее поверхности можно формулировать следующим образом: Водопроницаемость на площадке 2 (травяной покров, проективное покрытие 40−50 %, задернение 10 %) и 4-ой (травяной покров, проективное покрытие 30−40 %, задернения нет) статистически не отличается, а во всех остальных случаях различия средней водрпроницаемости на площадках статистически достоверны.

4.1.2 Метод Стьюдента для зависимых выборок

Наиболее полезным t-тест оказывается при проверке гипотезы о достоверности разницы средней между результатами опытной и контрольной групп после воздействия, т. е. для зависимых выборок.

К зависимым выборкам относятся, например, результаты одной и той же группы испытуемых до и после воздействия независимой переменной. В нашем случае с помощью статистических методов для зависимых выборок можно проверить гипотезу о достоверности разницы между фоновым уровнем и уровнем после воздействия отдельно для опытной и для контрольной группы.

Метод позволяет проверить гипотезу о том, что средние значения двух генеральных совокупностей, из которых извлечены сравниваемые зависимые выборки, отличаются друг от друга. Допущение зависимости чаще всего значит, что признак измерен на одной и той же выборке дважды, например, до воздействия и после него. В общем же случае каждому представителю одной выборки поставлен в соответствие представитель из другой выборки (они попарно объединены) так, что два ряда данных положительно коррелируют друг с другом.

Для определения достоверности разницы средних, в случае зависимых выборок (следовательно, равных по объему) применяется следующая формула:

$$S_d = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n \cdot (n-1)}}, \quad (5)$$

где d_i − разность между результатами в каждой паре; $\sum d_i$ − сумма этих частных разностей; $\sum d_i^2$ − сумма квадратов частных разностей.

Полученные результаты сверяют с таблицей распределения Стьюдента, отыскивая в ней значения, соответствующие n – 1 степени свободы; n − это в данном случае число пар данных наблюдений.

Пример. Пусть, например, требуется узнать, изменяется ли содержание P_2O_5 в почве на 12-ти опытных площадках спустя полгода после внесения фосфогипса (таблица 4).

Таблица 4 – Данные по содержанию P_2O_5 (мг/кг) в почве до и после внесения фосфогипса

Номер площадки	Фон	После воздействия
1	16,3	16,5
2	28,6	28,7
3	24,4	24,4
4	18,3	18,4
5	32,7	31.9
6	26,8	27,1
7	92,6	93,1
8	27,9	28,1
9	26,4	26,3
10	56,1	57,1
11	37,9	40,2
12	72,5	72,7

Важно отметить, что при использовании данного критерия распределение признака и в одной, и в другой выборке должно существенно не отличаться от нормального.

Результаты t-теста представлены в таблице 5.

Таблица 5 – Парный t-тест для зависимых выборок

Парный двухвыборочный t-тест для средних		
Статистический показатель	Переменная 1	Переменная 2
Среднее	38,38	38,71
Дисперсия	547,25	555,68
Наблюдения	12,00	12,00
Корреляция Пирсона	1,00	1,00
Гипотетическая разность средних	0,00	0,00
df	11,00	11,00
t-статистика	–1,55	–1,55
P (T <= t) одностороннее	0,07	0,07
t критическое одностороннее	1,80	1,80
P (T <=t) двухстороннее	0,15	0,15
t критическое двухстороннее	2,20	2,20

Величина t по модулю равная 1,55, ниже той, которая необходима для уровня значимости 0,05 при 12 степенях свободы (2,2). Иными словами, порог вероятности для такого t выше 0,05 − в нашем случае p двухстороннее приблизительно равно 0,15.

Таким образом, нулевая гипотеза не может быть отвергнута, и разница между выборками недостоверна. В сокращенном виде это записывается следующим образом: t = 1,55; df = 11; p > 0,05; недостоверно. Следовательно, можно считать, что в нашем случае разница в содержании P_2O_5 в почве на 12 опытных площадках до и после внесения фосфогипса не достоверна.

Задния для самостоятельной работы

Задание 1. Проверить гипотезу о равенстве среднего содержания свинца в листьях одуванчика (мг/кг) на придорожной территории до и после введения в эксплуатацию новой дорожной полосы (таблица 6).

Таблица 6 – Содержание Pb (мг/кг) в листьях одуванчика придорожной территории до и после введение в эксплуатацию новой дорожной полосы

Номер участка	До строительства дорожной полосы	После строительства дорожной полосы
1	9,47	9,49
2	2,62	4,8
3	3,49	6,12
4	8,09	9,82
5	9,38	9,8
6	2,54	4,18
7	7,77	17,05
8	2,55	8,12
9	5,24	13,02
10	11,4	12,1
11	2,44	3,76
12	2,39	5,41
13	13,01	18,1
14	3,12	16,2
15	5,91	12,94
16	9,63	13,59
17	4,97	19,9
18	11,49	12,17
19	3,98	6,4
20	1,39	9,16
21	4,14	5,96
22	0,92	3,72
23	3,56	3,74
24	1,27	3,51
25	1,84	4,61

Задание 2. Проверить гипотезу о равенстве средних длин стебля овса (мм) для различных вариантов опыта: вариант 1 – внесение в почву куриного помета (2т/га); вариант 2 – внесение в почву куриного помета (2т/га) совместно с фосфогипсом (5 т/га) (таблица 7).

Таблица 7 – Данные по длине стебля овса в различных вариантах опыта

Номер растения	Вариант 1	Вариант 2	Номер растения	Вариант 1	Вариант 2
1	29,5	28,5	16	30,5	28
2	30,5	29,5	17	30	28,5
3	29	28,5	18	26	27
4	27	26	19	26	29,5
5	28,5	32	201	26	29,5
6	26	32	21	28	31
7	32	29,5	22	27	27,5
8	30,5	30	23	29,5	28,5
9	29	31	24	30	26
10	29	29,5	25	33,5	32
11	29	29	26	27	28,5
12	29,5	30	27	24,5	29
13	30,5	26,5	28	29	30,5
14	28	24	29	30,5	28,5
15	32,5	33	30	32	31,5

Отчет по заданию 2 должен содержать распечатку фрагмета окна электронных таблиц Microsoft Excel с выполненными расчетами, а также интерпритацию результатов.

Тема 5. ДИСПЕРСИОННЫЙ АНАЛИЗ

Важнейшее свойство живой системы − изменчивость ее различных признаков, которая обусловлена как природой самой живой системы, так и влиянием факторов окружаюшей среды. Количественно изменчивость характеризуется такими статистическими показателями, как дисперсия, среднеквадратичное отклонение вариант от среднего значения признака, а также коэффициентом вариации признака. При этом изменчивость определенного признака биологического объекта можно рассматривать как результат воздействия на обьект всей совокупности биотических и абиотнческих факторов. В биологических экспериментах одни факторы, например, свет, температура, влажность почвы, концентрация химических агентов и т. д. строго учитываются количественно, т. е. строго контролируются, регулируются, тогда как другая *группа* факторов не учитывается, строго не контролируется.

Поэтому совокупность факторов, воздействующих на живой объект, можно делить на две *группы:* учитываемые (или регулируемые) и неучитываемые (или не регулируемые). На основе этого можно прийти к заключению, что обшая дисперсия в принципе должна состоять из двух слагаемых: дисперсия, обусловленная воздействием регулируемых, учитываемых в опыте факторов, т. е. так называемая факториальная дисперсия и дисперсия, обусловленная влиянием случайных, не учитываемых и не регулируемых в опыте факторов, что называется остаточной дисперсией.

Сущность этого метода заключается в том, чтобы определить, является ли разброс средних для различных выборок, относительно общей средней для всей совокупности данных, достоверно отличным от разброса данных относительно средней в пределах каждой выборки. Если все выборки принадлежат одной и той же совокупности, то разброс между ними должен быть не больше, чем разброс данных внутри их самих.

В методе Снедекора в качестве показателя разброса используют вариансу (дисперсию).

Критерием опенки влияния регулируемых в эксперименте факторов на результативный признак служит отношение факториальной (межгрупповой) дисперсии к остаточной (внутригрупповой) дисперсии, что называют критерием Фишера:

$$F_{набл} = \frac{S^2_{1\ межгрупповая}}{S^2_{2\ внутригрупповая}}, \qquad (6)$$

где S_1^2 – дисперсия между группами; S_2^2 – дисперсия внутри групп.

F-критическое определяют по таблице распределения Фишера для принятого уровня значимости α и чисел степеней свободы k_1 (для большей дисперсии) и k_2 (для меньшей дисперсии):

$$F_{крит.}(\alpha = 0{,}05; k_1 = m-1; k_2 = mn-m). \qquad (7)$$

Если $F_{набл} > F_{крит}$ нулевая гипотеза отклоняется, и следует считать, что среди средних значений имеются хотя бы два не равных друг другу.

Дисперсионный анализ завершается оценкой силы влияния факторов.

Дисперсионный анализ содержит такие понятия, как дисперсионный (статистический) комплекс, результативный признак, факторы регулируемые и не регулируемые, градации фактора, общая дисперсия, факторнальная дисперсия, остаточная дисперсия, а также сила влияния того или иного фактора.

По числу факторов, влияющих на результативный признак, дисперсионный анализ бывает однофакторным, двухфакторным, трехфакторным и многофакторным.

5.1 Реализация процедуры дисперсионного анализа в Microsoft Excel

Пример 5.1. Проанализируем результаты полевого опыта по влиянию органоминерального удобрения, приготовленного из отходов сельскохозяйственного и химического производства, на показатели роста побегов овса (см) (таблица 7). Опыт проводился с различными дозами внесения куриного помета в почву (доза фосфогипса при этом оставалась без изменения, т.е. регулируемым был один фактор – куриный помет).

Как видно из таблицы 8, внесение удобрений в почву в двух различных дозах куриного помета сказалось на длине ростка овса. Чтобы убедиться в этом утверждении, необходимо подвергнуть исходные данные дисперсионному анализу.

Таблица 8 – Исходная таблица для дисперсионного анализа по изучению влияния органоминерального удобрения на показатели роста овса

Варианты опыта	1	2	3	Среднее
Контроль (почва)	2,3	2,5	2,1	2,3
2т/га кур. пом.+5 т/га ФГ	3,8	4,1	3,9	3,9
4 т/га кур. пом. +5 т/га ФГ	3,7	4,3	4,2	4,1

В Microsoft Excel для проведения однофакторного дисперсионного анализа используется процедура **Однофакторный дисперсионный анализ.**

Для ее реализации необходимо:

1) ввести данные в таблицу, так чтобы в каждой строке (или столбце) оказались данные, соответствующие одному значению исследуемого фактора, а строки (столбцы) располагались в порядке возрастания (убывания) величины исследуемого фактора;

2) выполнить команду **Сервис** – **Анализ данных**;

3) в появившемся диалоговом окне **Анализ данных** в списке **Инструменты анализа** выбрать процедуру **Однофакторный дисперсионный анализ**, затем нажать кнопку **ОК**;

4) в появившемся диалоговом окне задать **Входной интервал**, то есть ввести ссылку на диапазон анализируемых данных, содержащий все столбцы данных;

5) в разделе **Группировка** переключатель установить в положение *по строкам (по столбцам)*;

6) указать **выходной интервал**, то есть ввести ссылку на ячейку, в которой будут показаны результаты анализа. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные. Нажать кнопку **ОК**.

Влияние исследуемого фактора определяется по величине значимости критерия Фишера, которая находится в таблице **Дисперсионный анализ** на пересечении строки **Между группами** и столбца **Р–значение**. В случаях, когда Р–значение < 0,05, критерий Фишера значим, и влияние исследуемого фактора можно считать доказанным.

Используя технологию решения задачи с помощью программы Microsoft Excel, получим следующие результаты (рисунок 8).

6	Однофакторный дисперсионный анализ						
7							
8	ИТОГИ						
9	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
10	Строка 1	3,000	6,900	2,300	0,040		
11	Строка 2	3,000	11,800	3,933	0,023		
12	Строка 3	3,000	12,200	4,067	0,103		
13							
14							
15	Дисперсионный анализ						
16	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
17	Между группами	5,807	2,000	2,903	52,260	0,000	5,143
18	Внутри групп	0,333	6,000	0,056			
19							
20	Итого	6,140	8,000				

Рисунок 8 − Результаты дисперсионного анализа

Выходной диапазон будет включать в себя результаты дисперсионного анализа: средние, дисперсии, критерий Фишера и другие показатели.

Фактическое значение F-критерия Фишера (52,26) заметно превосходит его стандартное (критическое) значение (5,143) на принятым уровне значимости (0,05), поэтому влияние регулируемлого фактора (доза куриного помета в органоминеральном удобрении) на результативный признак (длину ростка) считается статистически достоверным. Другими словами, изменчивость результативного признака, обнаруженная в опытах, есть действие влияния регулируемого фактора.

Пример 5.2. Рассмотрим данные пятикратного (n = 5) измерения валового содержания свинца в почвах на трех (m = 3) участках, удаленных на разном расстоянии от источника загрязнения (таблица 9).

Таблица 9 – Валовое **с**одержание свинца в почвах на разном расстоянии от источника загрязнения (мг/кг)

Номер варианта	**Номер эксперимента**				
	1	2	3	4	5
i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
1	12,5	15,4	17,2	13,1	16,9
2	20,1	17,5	16,3	25,3	14,2
3	10,3	12,3	11,2	13,5	8,4

Требуется проверить, влияет ли удаленность участка от источника загрязнения на содержание свинца в почве. Решим этот пример на компьютере, используя технологию, изложенную выше (рисунок 9).

7	Однофакторный дисперсионный анализ						
8	ИТОГИ						
9	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
10	Строка 1	5	75,1	15,02	4,617		
11	Строка 2	5	93,4	18,68	18,242		
12	Строка 3	5	55,7	11,14	3,783		
13							
14							
15	Дисперсионный анализ						
16	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
17	Между группами	142,169	2	71,085	8,004429	0,0061846	3,885293835
18	Внутри групп	106,568	12	8,8807			
19							
20	Итого	248,737	14				

Рисунок 9 − Результаты дисперсионного анализа

Произведем проверку нулевой гипотезы с помощью F-критерия:

$$F_{набл} = \frac{S_1^2}{S_2^2} = \frac{71{,}08}{8{,}88} = 8\,.$$

При двух степенях свободы большей дисперсии ($k_1 = 2$) и двенадцати степенях свободы меньшей дисперсии ($k_2 = 12$) по таблице критерия Фишера находим критические границы для F, равные 3,89 при 5 % уровне значимости и 6,93 при 1 % уровне значимости.

Полученное нами из наблюдений $F_{набл}$ превышает указанные границы, и поэтому нулевая гипотеза должна быть отвергнута, т. е. содержание валового свинца на рассматриваемых категориях почвы не одинаково, а это значит, удаленность от источника загрязнения влияют на его содержание.

Контрольные вопросы

1. Основные числовые характеристики случайной величины: математическое ожидание, дисперсия, среднее квадратическое отклонение, коэффициент корреляции, линейная регрессия. Мода и медиана случайной величины.
2. Понятие интервального оценивания. Доверительная вероятность и предельная ошибка выборки.
3. Доверительный интервал. Схема построения доверительного интервала
4. Анализ первичных статистик. Оценка достоверности отличий.
5. Статистическая гипотеза и общая схема ее проверки
6. Представления о диаграммах рассеивания, ковариации, корреляции.
7. Основные статистические распределения и их оценка
8. Основы теории общей линейной модели, однофакторный и двухфакторный дисперсионный анализ

Задания для самостоятельной работы

В таблицах 10 и 11 представлены данные по содержанию подвижных форм марганца, меди и цинка в пахотном слое почвы при различных вариантах агротехнологии. Методом однофакторного дисперсионного анализа выявить, влияет ли способ обработки на содержание тяжелых металлов в пахотном слое.

Таблица 10 − Данные по содержанию подвижных форм тяжелых металлов (Mn, Cu, Zn) в пахотном слое почвы при различных вариантах агротехнологии

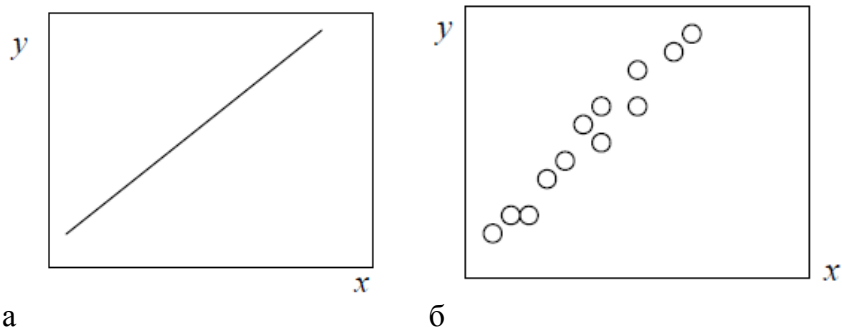
Варианты	Способы обработки								
	Марганец (Mn)			Медь (Cu)			Цинк (Zn)		
	$Д_1$	$Д_2$	$Д_3$	$Д_1$	$Д_2$	$Д_з$	$Д_1$	$Д_2$	$Д_з$
000	13,5	17,5	15.4	0,9S	0,71	0,69	0,99	1,06	1,21
111	28,0	15,8	15,1	1,42	0,79	1.45	1,63	1,05	1,06
222	18,6	18,6	12,5	1,34	0,88	1,32	1,24	0,76	0,76
333	26,1	25,8	21,6	2,0	1,18	1,4	1,33	1,57	1,3
Среднее	21,6	19,4	16,2	1,44	0,89	1,22	1,30	1,2	1,08

Таблица 11 − Данные по содержанию подвижных форм тяжелых металлов (Pb, Co, Cd) в пахотном слое почвы при различных вариантах агротехнологии

Варианты	Способы обработки								
	Свинец (Pb)			Кобальт (Co)			Кадмий (Cd)		
	$Д_1$	$Д_2$	$Д_3$	$Д_1$	$Д_2$	$Д_з$	$Д_1$	$Д_2$	$Д_з$
000	1,035	0,91	0,94	0,17	0,15	0,26	0,089	0,071	0,049
111	0,65	1,03	0,53	0,17	0,071	0,13	1,078	0,072	0,082
222	0,76	1,31	0,74	0,08434	0,041	0,03	1,065	0,054	0,051
333	0,82	1,66	1,02	0,101	1,07	0,08	1,06	0,076	0,062
Среднее	0,82	1,23	0,82	0,13	0,08	1,12	0,068	0,076	0,061

Тема 6. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Между величинами может существовать более точная, или функциональная связь, когда одному значению аргумента *x* соответствует одно определенное значение функции *y* (рисунок 10 а), и менее точная − корреляционная связь, когда одному конкретному значению аргумента соответствует приближенное значение или некоторое множество значений функции, в той или иной степени близких друг к другу (рисунок 10 б):



а б

Рисунок 10 − Функциональная (а) и корреляционная связь (б) переменных *x* и *y*

Близость в этом множестве значений функции друг к другу соответствует понятию силы, или тесноты, корреляционной связи. Чем больше сила корреляционной связи, тем ближе эта связь к функциональной.

Объекты экологических исследований всегда в той или иной степени неоднородны, имеют некоторые индивидуальные особенности. Кроме того, в любом, даже тщательно поставленном эксперименте его объекты испытывают не учитываемые воздействия многих факторов внешней среды. Поэтому между признаками объектов экологических исследований бывают исключительно корреляционные связи.

Сложность изучения корреляционных связей заключается также и в том, что все признаки (например, у растения) в различной степени взаимосвязаны и при исследовании методом парной корреляции можно считать доказанной лишь ту связь между двумя признаками, механизм которой понятен. В противном случае можно обнаружить корреляцию там, где ее нет, ибо изменения двух признаков могут более сильно зависеть от изменения какого-то иного, третьего признака или от совокупности некоторых других признаков. Например, количество атмосферных осадков определяет и количество луж на асфальте, и количество пешеходов с зонтиками. Последние два показателя будут тесно коррелировать между собою, однако при этом не будут находиться в причинно-следственной связи: они оба определяются первым показателем.

Задача корреляционного анализа состоит в том, чтобы определить характер связи между сопряженными признаками, убедиться в статистической достоверности найденного количественного значения связи, выяснить корреляционное отношение между признаками с тем, чтобы в дальнейшем подвергать эти данные регрессионному анализу.

Графическое представление зависимости между переменными можно получить с помощью диаграммы рассеяния.

6.1 Коэффициенты корреляции

Мерой корреляционной зависимости служит коэффициент корреляции (r). Он может принимать значения от – 1 до +1. Знак «–» означает, что связь обратная, «+» – прямая. Чем ближе коэффициент к 1, тем теснее линейная связь. При величине коэффициента корреляции менее 0,3 связь оценивается как слабая, от 0,31 до 0,5 – умеренная, от 0,51 до 0,7 – значительная, от 0,71 до 0,9 – тесная, 0,91 и выше – очень тесная.

Значение коэффициента корреляции может быть высоким, но не достоверным, случайным. Чтобы проверить статистическую значимость коэффициента корреляции, необходимо

расчитать эмпирическое (наблюдаемое) значение t-критерия. Для малых выборок n < 100 оно рассчитывается по формуле:

$$t_{набл} = \frac{|r| \cdot \sqrt{n-2}}{\sqrt{1-r^2}} . \quad (8)$$

При n > 100 формула для расчета t-наблюдаемого следующая:

$$t_{набл} = \frac{|r|\sqrt{n}}{1-r^2} . \quad (9)$$

Критическое значение t-критерия Стьюдента рассчитывается по таблице или с помощью встроенной функции СТЬЮДРАСПОБР в соответствии с принятым уровнем значимости α и числом степеней свободы, расчитываюшиеся по формуле f = n – 2 : $t_{крит}(\alpha, f = n-2)$.

При $t_{набл} > t_{крит}$ нулевая гипотеза о равенств нулю коэффициента корреляции между изучаемыми признаками в генеральной совокупности отвергается, и r считается статистически значим на принятом уровне значимости.

Для использования коэффициента корреляции Пирсона необходимо, чтобы все переменные были непрерывными и данные являлись бы случайной выборкой из генеральной совокупности с нормальным распределением. Корректное применение коэффициента корреляции для оценки зависимости какой-либо переменной Y от какой-либо переменной X возможно только в том случае, если эта зависимость близка к линейной.

Если какое-либо из этих условий не выполняется, применяются так называемые непараметрические критерии и, в частности, коэффициент ранговой корреляции Спирмена. Его значение также заключено между –1 и +1, интерпретация такая же, как и интерпретация значений коэффициента Пирсона.

6.2 Множественная корреляция

При большом числе наблюдений, когда коэффициенты корреляции необходимо последовательно вычислять для нескольких выборок, для удобства получаемые коэффициенты сводят в таблицы, называемые **корреляционными матрицами**.

Корреляционная матрица − это таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими параметрами.

6.2.1 Выполнение процедуры Корреляция в MS EXCEL и STATISTICA 6

Используя инструмент анализа «Корреляция» **пакета анализа *MS EXCEL***, построим корреляционную матрицу по данным, полученным при анализе экологического состояния почв на территории 1-го отделения учебного хозяйства «Кубань» Кубанского ГАУ. Перед тем, как вызвать эту процедуру, на свободном листе электронных таблиц необходимо ввести матрицу исходных данных (таблица 12).

В MS EXCEL для вычисления корреляционной матрицы используется процедура **Корреляция** из пакета **Анализ данных.**

Для реализации процедуры необходимо:

1) выполнить команду **Сервис** – **Анализ данных**;

2) в появившемся списке **Инструменты анализа** выбрать строку **Корреляция** и нажать кнопку **ОК** (если пункт ***«Анализ данных»*** в меню ***«Сервис»*** отсутствует, то следует обратиться к пункту ***«Надстройки»*** того же меню и установить флажок ***«Пакет анализа»***);

3) в появившемся окне указать **Входной интервал**, то есть ввести ссылку на ячейки, содержащие анализируемые данные. Входной интервал должен содержать не менее двух столбцов.

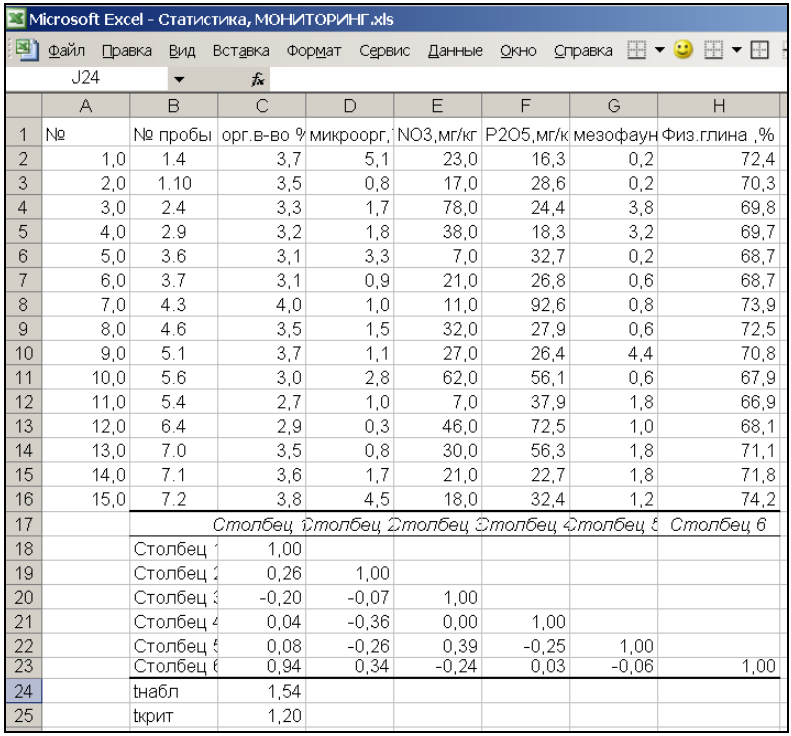
4) в разделе **Группировка** переключатель установить в соответствии с введенными данными (по столбцам или по строкам);

5) указать **выходной интервал**, то есть ввести ссылку на ячейку, с которой будут показаны результаты анализа. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные. Нажать кнопку **ОК**.

Таблица 12 – Данные экологического мониторинга почв на территории 1-го отделения учхоза «Кубань» Кубанского ГАУ

№ п/п	№ пробы	Орг. вещ-во, %	Микроорг., $*10^9$ экз./г	NO_3, мг/кг	P_2O_5, мг/кг	Мезо-фауна, экз./10 кг	Физ. глина, %
1	1.4	3,7	5,1	23,0	16,3	0,2	72,4
2	1.10	3,5	0,8	17,0	28,6	0,2	70,3
3	2.4	3,3	1,7	78,0	24,4	3,8	69,8
4	2.9	3,2	1,8	38,0	18,3	3,2	69,7
5	3.6	3,1	3,3	7,0	32,7	0,2	68,7
6	3.7	3,1	0,9	21,0	26,8	0,6	68,7
7	4.3	4,0	1,0	11,0	92,6	0,8	73,9
8	4.6	3,5	1,5	32,0	27,9	0,6	72,5
9	5.1	3,7	1,1	27,0	26,4	4,4	70,8
10	5.6	3,0	2,8	62,0	56,1	0,6	67,9
11	5.4	2,7	1,0	7,0	37,9	1,8	66,9
12	6.4	2,9	0,3	46,0	72,5	1,0	68,1
13	7.0	3,5	0,8	30,0	56,3	1,8	71,1
14	7.1	3,6	1,7	21,0	22,7	1,8	71,8
15	7.2	3,8	4,5	18,0	32,4	1,2	74,2

В выходной диапазон будет выведена корреляционная матрица, в которой на пересечении каждых строки и столбца находится коэффициент корреляции между соответствующими параметрами (рисунок 11).



Microsoft Excel - Статистика, МОНИТОРИНГ.xls

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

J24

	A	B	C	D	E	F	G	H
1	№	№ пробы	орг.в-во %	микроорг.	NO3,мг/кг	P2O5,мг/к	мезофаун	Физ.глина ,%
2	1,0	1.4	3,7	5,1	23,0	16,3	0,2	72,4
3	2,0	1.10	3,5	0,8	17,0	28,6	0,2	70,3
4	3,0	2.4	3,3	1,7	78,0	24,4	3,8	69,8
5	4,0	2.9	3,2	1,8	38,0	18,3	3,2	69,7
6	5,0	3.6	3,1	3,3	7,0	32,7	0,2	68,7
7	6,0	3.7	3,1	0,9	21,0	26,8	0,6	68,7
8	7,0	4.3	4,0	1,0	11,0	92,6	0,8	73,9
9	8,0	4.6	3,5	1,5	32,0	27,9	0,6	72,5
10	9,0	5.1	3,7	1,1	27,0	26,4	4,4	70,8
11	10,0	5.6	3,0	2,8	62,0	56,1	0,6	67,9
12	11,0	5.4	2,7	1,0	7,0	37,9	1,8	66,9
13	12,0	6.4	2,9	0,3	46,0	72,5	1,0	68,1
14	13,0	7.0	3,5	0,8	30,0	56,3	1,8	71,1
15	14,0	7.1	3,6	1,7	21,0	22,7	1,8	71,8
16	15,0	7.2	3,8	4,5	18,0	32,4	1,2	74,2
17			*Столбец 1*	*Столбец 2*	*Столбец 3*	*Столбец 4*	*Столбец 5*	*Столбец 6*
18		Столбец 1	1,00					
19		Столбец 2	0,26	1,00				
20		Столбец 3	-0,20	-0,07	1,00			
21		Столбец 4	0,04	-0,36	0,00	1,00		
22		Столбец 5	0,08	-0,26	0,39	-0,25	1,00	
23		Столбец 6	0,94	0,34	-0,24	0,03	-0,06	1,00
24		tнабл	1,54					
25		tкрит	1,20					

Рисунок 11 − Пример построения корреляционной матрицы по данным мониторинга

Заметим, что ячейки выходного диапазона, имеющие совпадающие координаты строк и столбцов, содержат значение 1, так как каждый столбец во входном диапазоне полностью коррелирует сам с собой.

В данном случае наиболее тесная корреляционная взаимосвязь ($r \approx 0{,}96$) проявляется между органическим веществом (%) и физической глиной (%), о чем наглядно демонстрирует и диаграмма рассеяния (рисунок 12), выполненная в пакете Statistica.

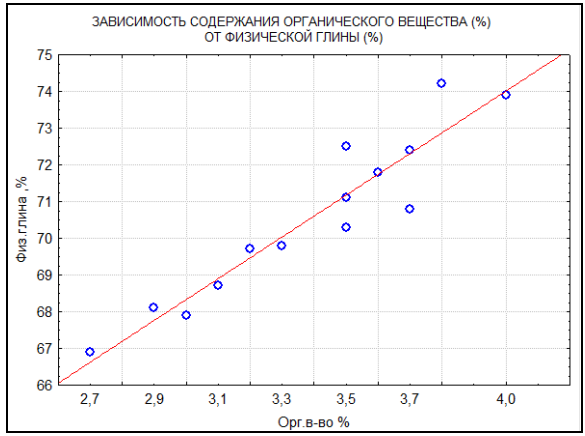


Рисунок 12 − Зависимость содержания органического вещества (%) от физической глины (%)

Диаграмма позволяет на глаз оценить зависимость двух переменных. Для построения диаграммы рассеяния используются пункты меню: Графики – Графики рассеяния – Переменные – Выбор переменных. Поверх уже созданной диаграммы в окне вывода можно наложить линию наименьших квадратов. Чтобы его вызвать, необходимо в окне Редактора графиков поставить птичку напртив Linear fit.

Для вычисления коэффициента корреляции Пирсона используются пункты меню: Статистика – Основная статистика/таблицы – Correlation matrices – выбор переменных – Correlation Coefficients – Pearson. Для каждой выбранной пары переменных принимается нулевая гипотеза о том, что линейная зависимость между ними отсутствует.

Рассматривается отдельно каждый коэффициент корреляции между соответствующими параметрами. Отметим, что хотя в результате будет получена треугольная матрица, корреляционная матрица симметрична. Подразумевается, что в пустых клетках в правой верхней половине таблицы находятся те же коэффициенты корреляции, что и в нижней левой (симметрично расположенные относительно диагонали).

Окно выходной таблицы представлено на рисунке 13. Здесь красным шрифтом обозначены статистически достоверные значения коэффициента корреляции.

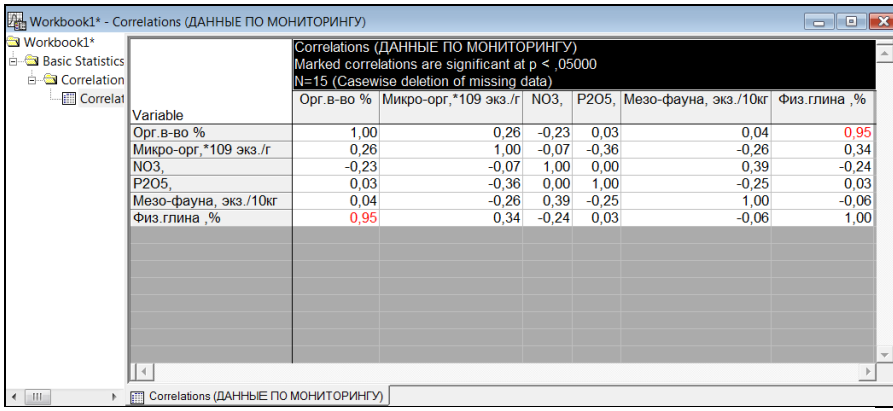


Рисунок 13 − Корреляционная матрица зависимости почвенных характеристик

Прокомментируем выходную информацию. Pearson Correlation – коэффициент корреляции; significant at – уровень значимости коэффициента; N − количество записей в файле данных, по которым делался расчет. Особое внимание следует обратить на уровень значимости – любая значимость выше 0,05 (5 %) подтверждает нулевую гипотезу (о том, что в генеральной совокупности значение коэффициента корреляции равно нулю).

По результатам анализа данных таблицы 9 статистически достоверной на 5 % уровне значимости получилась только взаимосвязь органического вещества и физической глины.

6.3 Построение множественной линейной регрессионной модели с помощью MS EXCEL

Регрессионный анализ позволяет получить предсказание значений зависимой переменной на основе значений независимых переменных. Процедура построения регрессионной модели, а особенно оценка ее адекватности является достаточно

сложной статистической процедурой, поэтому здесь ограничимся рассмотрением случая линейной регрессии, уравнение которой в общем виде имеет вид: $y = m_1x_1 + m_2x_2 + ... + m_kx_k + m_0$.

Для построения множественной линейной регрессионной модели необходимо:

1) подготовить список из n строк и m столбцов, содержащий экспериментальные данные (столбец, содержащий выходную величину y должен быть либо первым, либо последним в списке);

2) обратиться к меню ***Сервис/Анализ данных/Регрессия***

3) в диалоговом окне «Регрессия» (рисунок 14) задать:

– входной интервал Y;

– входной интервал X;

– выходной интервал − верхняя левая ячейка интервала, в который будут помещаться результаты вычислений (рекомендуется разместить на новом рабочем листе);

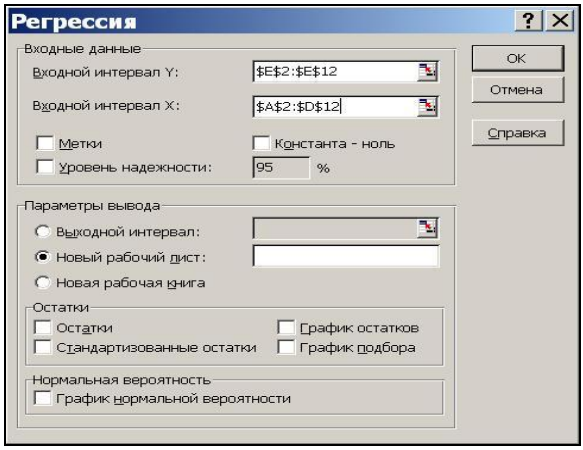


Рисунок 14 − Окно «Регрессия»

4) нажать «ОК» и проанализировать результаты.

Пример. По данным о зависимости урожайности У (ц/га) от величины гумусового слоя X_1 (м) и количества минеральных удобрений X_2 (т/га) построить множественную регрессионную модель.

Решение. Внесем данные как показано в верхней части листа электронной таблицы на рисунке 15. Выполняя процедуру регрессионного анализа, как указано выше, получим ВЫВОД ИТОГОВ, который содержит три таблицы.

1	N п/п	X_1	X_2	У						
2	1	1	0,2	18,2						
3	2	1	0,2	18,6						
4	3	1	0,2	18,7						
5	4	2	0,4	21,6						
6	5	2	0,4	23,4						
7	6	2	0,4	23,7						
8	7	2,5	0,3	22						
9	8	2,5	0,3	23						
10	9	2,5	0,3	22,5						
11	ВЫВОД ИТОГОВ									
12										
13	*Регрессионная статистика*									
14	Множественный	0,9575394								
15	R-квадрат	0,9168818								
16	Нормированный	0,8891757								
17	Стандартная ош	0,7325754								
18	Наблюдения	9								
19										
20	Дисперсионный анализ									
21		*df*	*SS*	*MS*	*F*	*Значимость F*				
22	Регрессия	2	35,52	17,76	33,093168	0,000574234				
23	Остаток	6	3,22	0,536666667						
24	Итого	8	38,74							
25										
26		*Коэффициенты*	*дартная ош*	*t-статистика*	*P-Значение*	*Нижние 95%*	*рхние 95%*	*ижние 95,0%*	*рхние 95,0%*	
27	Y-пересечение	14,1	0,9457507	14,90879102	5,732E-06	11,78583133	16,4142	11,7858313	16,4142	
28	Переменная X 1	1,8	0,518009	3,474843041	0,0132254	0,532477619	3,06752	0,53247762	3,06752	
29	Переменная X 2	13	3,9563592	3,285849309	0,0167001	3,31913791	22,6809	3,31913791	22,6809	

Рисунок 15− Результаты множественного регрессионного анализа

Информация о вкладе каждой независимой переменной показана в нижней таблице, а информация о модели в целом показана над этой таблицей.

6.3.1 Интерпретация результатов регрессионного анализа

В начале просмотрите информацию о модели в целом.

R − коэффициент множественной корреляции. Он характеризует тесноту линейной связи между зависимой и всеми независимыми переменными и может принимать значения от 0 до 1.

Задачей построения регрессионной зависимости является нахождение вектора коэффициентов модели, при котором коэффициент R принимает максимальное значение.

Коэффициент множественной корреляции R. В рассмотренном примере оценка множественного коэффициента корреляции между случайной величиной У и двумя остальными составила $R \approx 0{,}96$, что указывает на тесную зависимость урожайности (у) от величины гумусового слоя (x_1) и количества минеральных удобрений (x_2).

Коэффициент детерминации R^2. Он численно выражает долю вариации зависимой переменной, объясняемую с помощью регрессионного уравнения. Чем больше R^2, тем большую долю вариации объясняют переменные, включенные в модель. В рассматриваемом примере значение R^2 (0,9168) указывает на то, что около 92% дисперсии функции отклика (урожайности) объясняется вариацией линейной комбинации факторов x_1 и x_2.

Нормированный R − скорректированный коэффициент множественной корреляции. Этот коэффициент лишен недостатков коэффициента множественной корреляции. Включение новой переменной в регрессионное уравнение увеличивает его не всегда, а только в том случае, когда частный F-критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение нормированного коэффициента.

Значимость F-критерия (вторая таблица) точность аппроксимации исследуемой зависимости линейной зависимостью. В рассматриваемом примере $p \approx 0{,}00057$, что меньше 0,05, по-

этому точность аппроксимации зависимости урожайности от содержания гумуса и количества внесенных удобрений линейной функцией является достоверной.

Информация о каждой независимой переменной содержится в третьей нижней таблице, заголовки строк в которой являются названиями независимых переменных и свободного члена.

Направление связи между переменными определяется на основании знаков (отрицательный или положительный) коэффициентов регрессии. Если знак при коэффициенте регрессии положительный, связь зависимой переменной с независимой будет положительной. Если знак при коэффициенте регресси отрицательный, связь зависимой переменной с независимой является отрицательной (обратной).

m_0 – среднее значение Y, если каждая независимая переменная равна 0.

m_i – среднее изменение Y на единицу измерения X_i, когда воздействия остальных переменных постоянны.

В рассматриваемой задаче $m_1 = 1{,}8$; $m_2 = 13$; $m_0 = 14{,}1$.

Таким образом, получено следующее регрессионное уравнение: $y = 14{,}1 + 1{,}8m_1 + 13m_2$.

6.3.2 Оценка влияния отдельной независимой переменной (НП) на колебания зависимой переменной (ЗП)

К сожалению, определение относительного влияния разных независимых переменных не тождественно простому сравнению их коэффициентов регрессии. В тех случаях, когда независимая переменная измеряются в разных единицах, коэффициенты регрессии не отражают относительного воздействия их на зависимую переменую. Одним из возможных путей обойти это – стандартизировать переменные так, чтобы они были измерены в одних и тех же единицах, и снова произвести подсчеты коэффициента регрессии.

Когда числовые ряда заменены в уравнении регрессии на стандартизованные ряды, уравнение приходит к общей формуле:

$$Y' = \beta_1 X_1^* + \beta_2 X_2^* + \beta_3 X_3^* + \ldots + \beta_n X_n^* + e,$$

где β представляет частный коэффициент стандартизованной регрессии и называется **бета-вес**, или **бета-коэффициент**.

Он корректирует частный нестандартизованный коэффициент регрессии путем деления стандартного отклонения независимой переменной на стандартное отклонение зависимой переменной и может быть посчитан по формуле:

$$\beta_i = b_i \frac{S_x}{S_y}. \qquad (10)$$

Выводы о вкладе независимых переменных в дисперсию зависимой переменной можно делать по нормализованным (Beta) и ненормализованным (B) угловым коэффициентам и по уровню значимости вклада каждой независимой переменной (p-level). Чем более статистически значимой является независимая переменная, тем больший вклад она вносит в дисперсию зависимой переменной. Оценить долю вклада каждой независимой переменной Xi в суммарное влияние всех факторов позволяет дельта-коэффициент, который рассчитывается по формуле:

$$\Delta i = ri\ (\beta i / R^2), \qquad (11)$$

где βi – бета-коэффициент i-го фактора Xi; ri – коэффициент парной корреляции i-го фактора Xi и переменной Y; R^2 – коэффициент множественной детерминации.

В случае необходимости оценки влияния независимых переменных на зависимую требуется в пакете MS EXCELпо преведенным формулам найти Δ-коэффициенты.

6.4 Порядок выполнения корреляционно-регрессионного анализа в системе STATISTICA

6.4.1 Оценка парных коэффициентов регрессии

Существенно больше возможностей в проведении корреляционно-регрессионного анализа дает пакет STATISTICA.

Для обработки данных необходимо должным образом перенести их в пакет STATISTICA.

Импортируем данные из программы Excel, предварительно выделив их и скопировав в буфер. Кнопкой Paste вставим данные из буфера в окно пакета STATISTICA.

Стандартный вид исходной таблицы с данными пакета STATISTICA содержит 10 строк (cases) и 10 столбцов (variables). Если исходная информация представлена матрицей большей размерности, тогда открыв меню **Cases**, выбираем позицию **Add** (добавить), переходим к следующему окну, где указываем количество добавляемых строк и столбцов (рисунок 16).

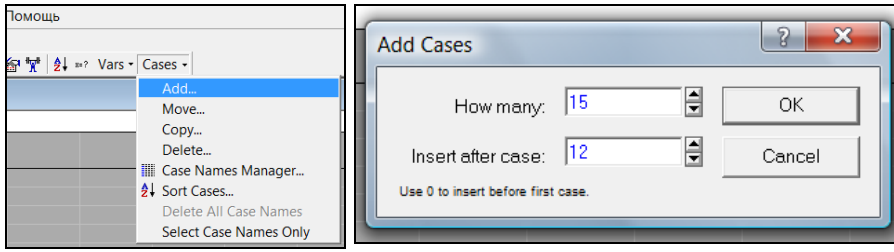
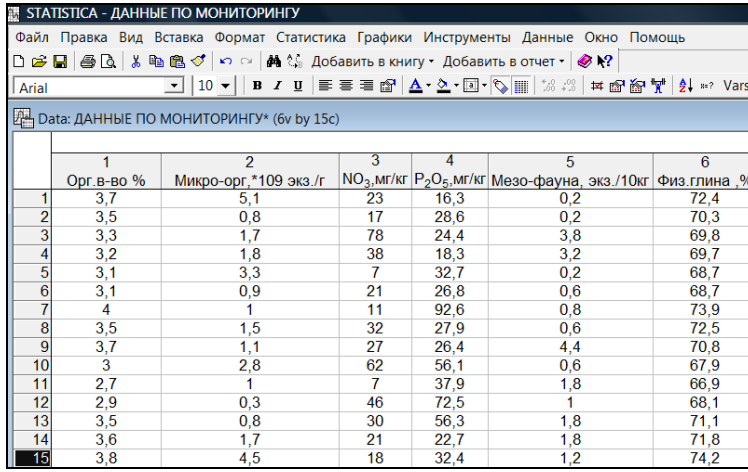


Рисунок 16 − Окно корректировки таблицы с входными данными

Рассмотрим пример решения задачи корреляционно-регрессионного анализа для данных экологического мониторинга почв на территории 1-го отделения учебного хозяйства «Кубань» Кубанского ГАУ.

Окно с исходными данными представлено на рисунке 17.



	1 Орг.в-во %	2 Микро-орг,*109 экз./г	3 NO_3,мг/кг	4 P_2O_5,мг/кг	5 Мезо-фауна, экз./10кг	6 Физ.глина ,%
1	3,7	5,1	23	16,3	0,2	72,4
2	3,5	0,8	17	28,6	0,2	70,3
3	3,3	1,7	78	24,4	3,8	69,8
4	3,2	1,8	38	18,3	3,2	69,7
5	3,1	3,3	7	32,7	0,2	68,7
6	3,1	0,9	21	26,8	0,6	68,7
7	4	1	11	92,6	0,8	73,9
8	3,5	1,5	32	27,9	0,6	72,5
9	3,7	1,1	27	26,4	4,4	70,8
10	3	2,8	62	56,1	0,6	67,9
11	2,7	1	7	37,9	1,8	66,9
12	2,9	0,3	46	72,5	1	68,1
13	3,5	0,8	30	56,3	1,8	71,1
14	3,6	1,7	21	22,7	1,8	71,8
15	3,8	4,5	18	32,4	1,2	74,2

Рисунок 17 − Окно исходной таблицы для корреляционно-регрессионного анализа

Оценки парных коэффициентов корреляции производится в блоке **Основная статистика/Таблицы**. При запуске этого блока на экране появляется меню (рисунок 18) :

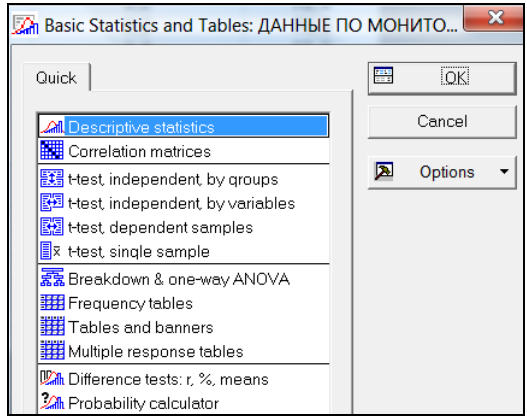


Рисунок 18 − Меню блока Основной Статистики

Выбираем позицию **Correlation matrices**. После вызова данного пункта появляется окно корреляционного анализа (рисунок 19).

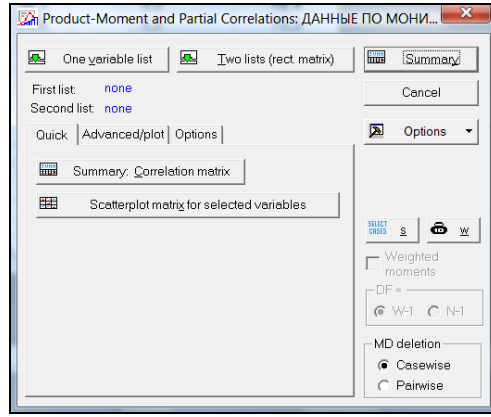


Рисунок 19 − Окно Корреляционного анализа

Корреляционный анализ начинается с выбора переменных, между которыми будут оцениваться парные коэффициенты корреляции. Вызов режима выбора осуществим с помощью кнопки **One variable list.** Далее появляется следующий экран (рисунок 20), в котором производится выбор необходимых переменных.

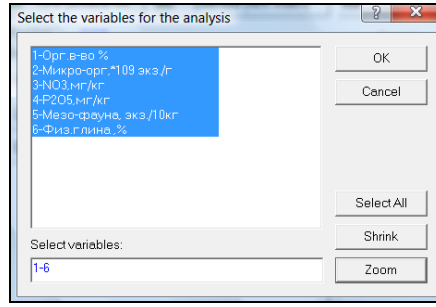


Рисунок 20 − Выбор переменных для корреляционного анализа

Если при анализе используются все переменные, то проще нажать кнопку **Select All.** Если надо выбрать только часть переменных, то можно выделить с помощью мыши. Если необходимо выделить отдельные переменные, то используется комбинация клавиш **[Ctrl]** и левая клавиша мыши.

После выбора переменных для анализа можно уточнить, в какой форме пользователь желает получить информацию. Для этого необходимо нажать кнопку Opcions и поставить птичку в соответствие с вариантом вывода информации:

Display simple matrix (highlight p's) – наиболее краткий вид, показываются только парные коэффициенты корреляции, красным выделяются те, гипотеза о незначимости которых отвергается.

Display r, p-levels and N's – аналогична предыдущей, но кроме значения коэффициента показывается вероятность принятия гипотезы о незначимости коэффициента, в таблице данная вероятность обозначается символом **p**.

Display detailed table of results – наиболее подробная таблица. Переменные сгруппированы в пары, для каждой переменной выводится ее среднее, дисперсия, а также парный коэффициент корреляции, как и в предыдущем случае: вероятность принятия гипотезы о незначимости коэффициента (**p**), а также коэффициенты для линейного уравнения регрессии. Для проведения корреляционного анализа достаточно **Corr.Matrix (display p&N).** Результаты расчетов приведены на рисунке 21.

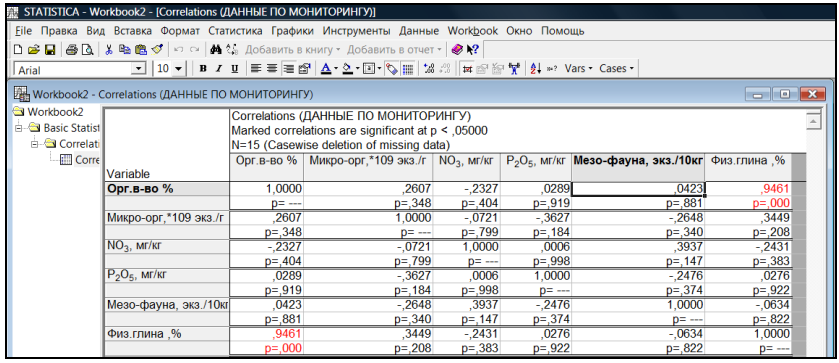


Рисунок 21 − Окно результатов корреляционного анализа

В данном примере красным подсвечены оценки парного коэффициента корреляции r_{16} = 0,9461, для которого гипотеза

о незначимости отвергается. Все остальные коэффициенты корреляции оказались статистически незначимыми.

6.4.2 Построение уравнения множественной регрессии

Найдем параметры регрессионного уравнения линейной связи содержания органического вещества (У) с остальными почвенными характеритсиками: численностью микроорнизмов (X_1), содержанием P_2O_5 (X_2), плотностью мезофауны (X_3), pH почвы (X_4) и содержанием физической глины (X_5) по данным таблицы 9.

Для построения уравнения множественной регрессии вызываем модуль **Статистика Множественная регрессия.** Прежде всего, необходимо выбрать переменные для анализа. Делается это после нажатия кнопки Variables. В качестве зависимой переменной выбираем органическое вещество, а в качестве независимых переменных выбираем все остальные (как на рисунке 22).

После выбора переменных Для вывода общей характеристики регрессионной модели ставим птичку в пункт Extended precision computations, как показано на рисунке 23.

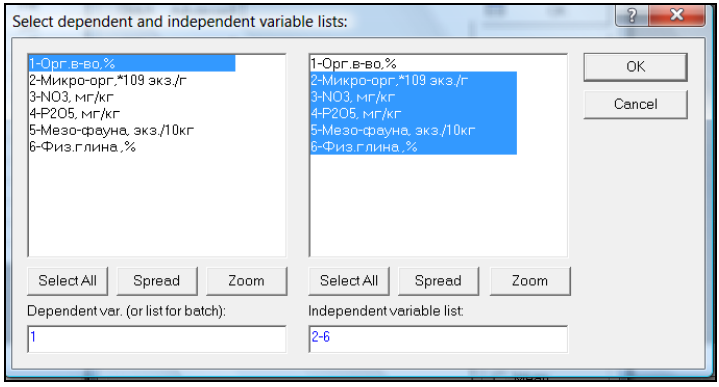


Рисунок 22 − Выбор переменных

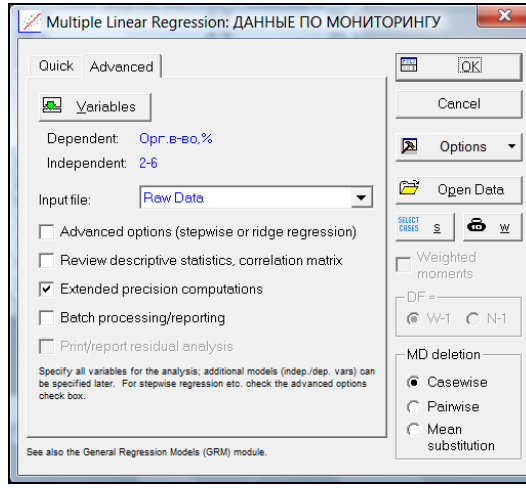


Рисунок 23 – Установка рехима вывода общих характеристик регрессионного анализа

После подтверждения выбора данного режима, на экран выводятся результаты анализа (рисунок 24).

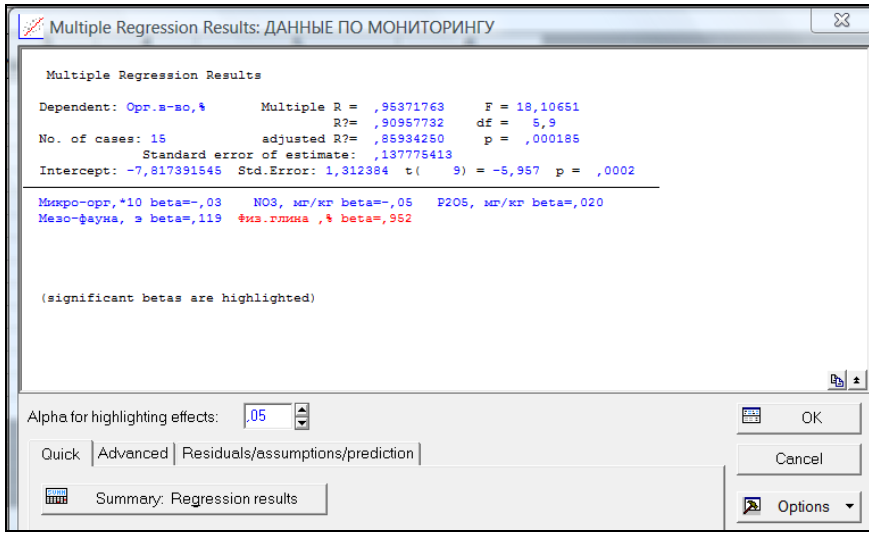


Рисунок 24 − Общая регрессионная статистика регрессионной модели

Рассмотрим информационную часть.

− Dependent – имя зависимой переменной;

− No of cases – объем выборки, n = 15;

− Multiple R – оценка коэффициента множественной корреляции;

− R^2– оценка коэффициента детерминации;

− adjusted R^2 – скорректированный коэффициент детерминации;

− F – значение F – критерия расчетный;

− df – число степеней свободы F – критерия;

− p – вероятность принятия гипотезы H_0;

− Partial correlations – оценки частных коэффициентов корреляции.

− Std. Error of estimate – стандартная ошибка оценки, оценивает меру рассеяния наблюдаемых значений относительно регрессионной прямой;

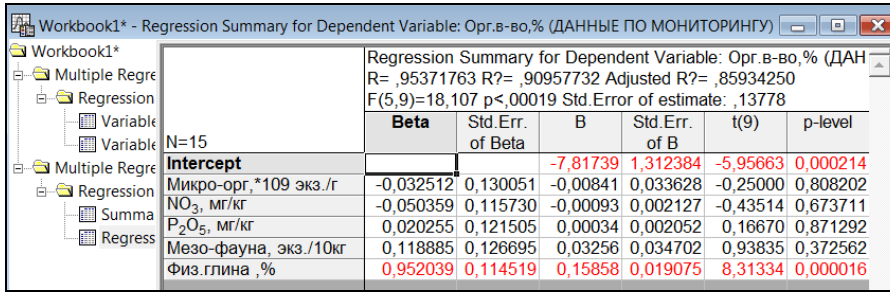
− Intercept – оценка свободного члена, значение коэффициента b_0 в уравнении регрессии;

− Std.Error – стандартная ошибка коэффициента b_0 в уравнении регрессии;

− t(df) and p-value – значение t-критерия и уровня p.

В рассмотренном примере оценка множественного коэффициента корреляции между случайной величиной У и всеми остальными составила 0,9537. Вероятность принятия гипотезы H_0 : $\rho_{1/23456} = 0$ о незначимости множественного коэффициента составила p = 0,000195, следовательно, гипотеза Но отвергается и множественный коэффициент корреляции значимо отличен от нуля. Коэффициент детерминации составил примерно 0,9096.

Оценки коэффициентов уравнения регрессии (рисунок 25) могут быть получены нажатием на кнопку **Summary Regression result**.



Regression Summary for Dependent Variable: Орг.в-во,% (ДАН
R= ,95371763 R?= ,90957732 Adjusted R?= ,85934250
F(5,9)=18,107 p<,00019 Std.Error of estimate: ,13778

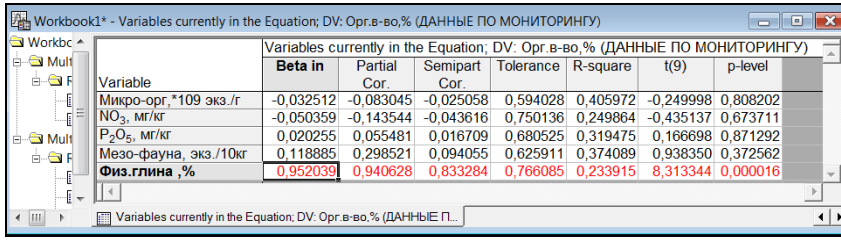
N=15	Beta	Std.Err. of Beta	B	Std.Err. of B	t(9)	p-level
Intercept			-7,81739	1,312384	-5,95663	0,000214
Микро-орг,*109 экз./г	-0,032512	0,130051	-0,00841	0,033628	-0,25000	0,808202
NO_3, мг/кг	-0,050359	0,115730	-0,00093	0,002127	-0,43514	0,673711
P_2O_5, мг/кг	0,020255	0,121505	0,00034	0,002052	0,16670	0,871292
Мезо-фауна, экз./10кг	0,118885	0,126695	0,03256	0,034702	0,93835	0,372562
Физ.глина ,%	0,952039	0,114519	0,15858	0,019075	8,31334	0,000016

Рисунок 25 − Оценки коэффициентов уравнения регрессии

В данном случае статистически значимыми получились только свободный член и коэффициент при переменной X_5 (физическая глина). Остальные переменные в регрессионную модель включать не следует, так они мало влияют на изменчивость зависимой переменной. Таким образом, адекватная регрессионная модель будет иметь вид: $У = -7{,}82 + 0{,}952\ X_5$.

6.4.3 Оценка частных коэффициентов корреляции

Для получения коэффициентов частной корреляции надо нажать в окне регрессионной статистики (рисунок 25) сначала Advanced, а затем на кнопку **Partial correlations**. Появится окно с результатами (рисунок 26).



Variable	Beta in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(9)	p-level
Микро-орг,*109 экз./г	-0,032512	-0,083045	-0,025058	0,594028	0,405972	-0,249998	0,808202
NO_3, мг/кг	-0,050359	-0,143544	-0,043616	0,750136	0,249864	-0,435137	0,673711
P_2O_5, мг/кг	0,020255	0,055481	0,016709	0,680525	0,319475	0,166698	0,871292
Мезо-фауна, экз./10кг	0,118885	0,298521	0,094055	0,625911	0,374089	0,938350	0,372562
Физ.глина ,%	0,952039	0,940628	0,833284	0,766085	0,233915	8,313344	0,000016

Рисунок 26 – Вывод частных коэффициентов корреляции

Частные коэффициенты корреляции между двумя случайными величинами при фиксированных остальных характеризуют тесноту связи между этими двумя величинами, лишенными влияния остальных величин. Поэтому, если парный

коэффициент корреляции между теми же двумя случайными величинами оказался больше соответствующего частного коэффициента, то делается вывод о том, что остальные фиксированные величины усиливают взаимосвязь между изучаемыми величинами, т. е. более высокое значение парного коэффициента обусловлено присутствием остальных величин. Более низкое значение парного коэффициента корреляции в сравнении с соответствующими частными свидетельствует об ослаблении связи между изучаемыми величинами вследствие действия фиксируемых величин.

Так, в наших расчетах оценки частных коэффициентов корреляции (Semipart cor.) больше соответствующих парных (Partial cor.), следовательно делаем вывод об усилении корреляционных связей между соответствующими парами при фиксированных остальных. Значимость частного коэффициента определяется пользователем в зависимости от выбранного уровня значимости. Если указанное значение p в окне частных корреляций меньше выбранного уровня значимости, то H_0 : $\rho = 0$ отвергается. В нашем случае статистически значимым является только частный коэффициент корреляции между органическим веществом и физической глиной $r_{16\backslash 2345} = 0{,}9250$, т. к. p = 0,000016, а остальные частные коэффициенты корреляции являются статистически незначимыми.

Задания для самостоятельной работы

Задание 6.1. Известны удельные показатели выбросов вредных веществ в атмосферу от объектов энергетики в городах Краснодарского края, кг/т (таблица 13).

Найти парные коэффициенты корреляции между z и x; z и y, x и y. Сделать выводы. Построить регрессионную модель зависимости z от x и y.

Таблица 13 − Данные по выбросам вредных веществ в атмосферу от объектов энергетики в г. Краснодаре

№ п|п	Пыль неорганическая (z)	SO_2 (x)	NO_3 (y)
1	20,4	16,8	4,7
2	13,9	9,4	8,0
3	29,3	10,2	*53*
4	26,3	4,4	*2,6*
5	5,0	5,7	1,6
6	109,2	24,4	4,1
7	23,8	13,1	1,6
8	18,2	11,4	4,3
9	12,2	4,6	0,8
10	8,2	12,9	3,9
11	21,5	12,5	4,6
12	25,2	4,3	2,5
13	41,2	11,1	1,7
14	9,5	14,3	1,7
15	17,9	5,6	*3,2*

Задание 6.2. Определите, имеется ли взаимосвязь между рождаемостью и смертностью (количество на 1000 человек) в г. Краснодаре (таблица 14).

Таблица 14 – Данные по смертности и рождаемости в г. Краснодаре (количество на 1000 человек)

Годы	Рождаемость	Смертность
2001	9,3	12,5
2002	7,4	13,5
2003	6,6	17,4
2004	7,1	17,2
2005	7,0	15,9
2006	6,6	14,2
2007	7,1	16
2008	8,2	13,4

Тема 7. КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ ДИНАМИКИ ЧИСЛЕННОСТИ ПОПУЛЯЦИЙ

7.1 Модели неограниченного роста

Чтобы реализовать модели популяционного роста на компьютере, их необходимо записать в виде рекуррентной формулы, связывающей численность популяции следующего момента времени (t + 1) с численностью текущего момента времени (t).

Модель Мальтуса можно записать следующим рекуррентным соотношением:

$$x(t+1) = x(t) + \Delta x = Rx(t), \qquad (13.1)$$

где $R = e^r$.

Построим компьютерную модель.

Заполним лист EXCEL так, как показано на рисунке 29.

	A	B	C	D	E	F
1		МОДЕЛЬ НЕОГРАНИЧЕННОГО РОСТА МАЛЬТУСА				
2						
3	Параметры модели			x(t+1)=R*x(t)		
4	R	0,25				
5	x0	100				
6	t(время)	численность				
7	0	100				
8	1	125				
9	2	156,25				
10	3	195,3125				
11	4	244,140625				
12	5	305,175781				
13	6	381,469727				
14	7	476,837158				
15	8	596,046448				
16	9	745,05806				
17	10	931,322575				
18	11	1164,15322				

Рисунок 29 – Компьютерная реализация модели Мальтуса

1. Дадим имена ячейкам B4 : B5. Для этого выделим блок A4 : B5, выберем в меню команду «Вставка / Имя / Создать», выберем вариант «В столбце слева», нажмем «ОК».

2. В столбце А будем формировать время, в столбце В – численность популяции. Для этого зададим начальное время и начальную численность: наберем в в ячейке A7 ноль, и поместив в ячейке B7 знак равенства, кликнем на ячейку B5, содержащую значение нулевой численности.

3. В ячейке A8 будем пересчитывать время, набрав формулу : A7 + 1.

4. В ячейке B8 наберем рекуррентную формулу, выражающую численность популяции через ее значение в предыдущий момент времени, кликая мышкой на значении коэффициента роста, который помещен в ячейке B4. Так, в ячейке B8 содержится формула: = B7 + R×B7.

5. Растянем маркером заполнения формулы в ячейках A8 и B8 до 20-строки.

6. Построим график по полученным значениям численности.

Ответим на вопрос: через какой промежуток времени численность популяции удвоится?

7.2 Модель ограниченного роста Ферхюльста

Наиболее известная формула дискретного представления уравнения Ферхюльста имеет вид:

$$x(t+1) = x(t) \cdot e^{R(1-\frac{x(t)}{K})}. \qquad (13.2)$$

Здесь R – константа собственной скорости, К-емкость среды.

Рассмотрим динамику численности популяции в зависимости от величины R.

Таблицу для эксперимента заполняем так, как показано на рисунке 30.

	A	B	C	D	E	F
1	ТИПЫ ДИНАМИКИ ЧИСЛЕННОСТИ ПОПУЛЯЦИИ					
2	параметры модели					
3	R	0,5		t(время)	x(численность)	
4	K	50000		0	100	
5	x_0	100		1	=(E4*EXP(r_*(1-E4/K)))	
6				2	271,1095852	
7	$x(t+1) = x(t) \cdot e^{R\left(1-\frac{x(t)}{K}\right)}$			3	445,7739642	
8				4	731,6880611	
9				5	1197,555167	
10				6	1950,930749	
11				7	3154,396703	
12				8	5039,229993	
13				9	7899,986032	

Рисунок 30 – Компьютерная реализация модели Ферхюльста

1. В ячейках D4 и E4 задаем начальные значения для времени и численности.

2. В ячейке D5 пересчитываем время, задавая формулу пересчета : = D4 + 1.

3. В ячейке E5 пересчитываем численность по рекуррентной формуле : E4·EXP (r·× 1 – E4 / K).

4. Растянем маркером заполнения сразу две формулы: для времени и для численности.

5. Построим график по полученным значениям численности популяции.

6. Изучим динамику численности популяции в зависимости от коэффициента прироста R, задавая ему значения 0,5; 1,5; 2,2; 3 и построив графики.

Графическое изображение модели для различных значений R представлено на рисунке 31.

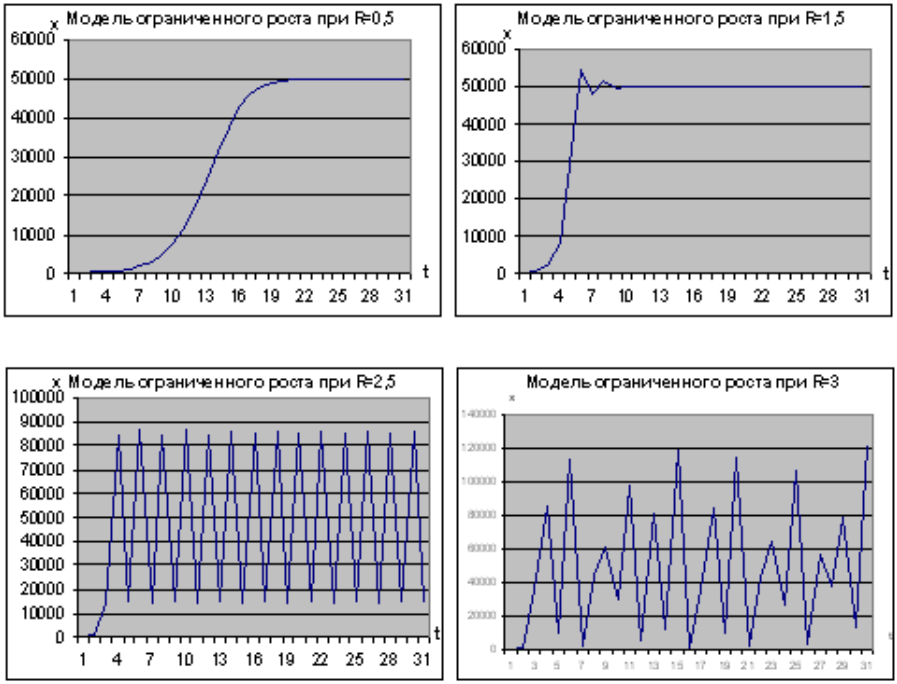


Рисунок 31 – Модель ограниченного роста для различных значений R

Контрольные вопросы

1. Как изменяется характер роста популяции при возрастании R?

2. Определите, сколько поколений проходит до «популяционного взрыва» при различных значениях R?

3. Исследуйте поведение модели при значениях R<1, например: 0,9; 0,8; 0,7; 0,3; 0,1 при достаточно большой исходной численности, например, 1000. Нажмите F4, чтобы видеть на экране предыдущую кривую роста.

4. При каждом значении R перейдите к полулогарифмическому масштабу и обратно. Что изменяется на графике?

Задания для самостоятельной работы

1. При помощи электронных таблиц построить график изменения заготовок шкурок зайца-беляка последовательно за 30 лет (объем заготовок в баллах). Баллы : 2, 1, 2, 3, 3, 4, 5, 15, 30, 80, 100, 60, 55, 0, 1, 1, 1, 2, 8, 90, 100, 100, 130, 10, 2, 1, 2, 2, 1, 2. Сколько лет длится один цикл в динамике численности зайца?

2. Постройте модель роста популяции с ежегодным приростом 6 %, а начальная численность 200 экз. Через сколько лет численность удвоится?

3. Постройте модель роста популяции, если R = 1,5; K = 10000; N_0 = 300. Через сколько лет численность стабилизируется?

4. Постройте модель роста популяции, если r = 0,25; K = 2000; N_0 = 100. Через сколько лет популяция увеличится в 5 раз?

5. Попробуйте, используя соотношение N (t + 1) / N (t), доказать, что наибольшая скорость роста достигается при численности приблизительно равной половине емкости среды.

6. Известно, что каждую минуту на земле рождается 240 человек, а умирает 120. В настоящее время население земного шара равно 6,5 млрд человек. Емкость среды нашей планеты по оценкам ряда ученых (при прогрессивном и грамотном ведении хозяйства) приблизительно равно 20 млрд человек. Используя модель Ферхюльста, попытайтесь спрогнозировать, через сколько лет должен прекратиться рост населения, и каким оно будет.

ТЕСТЫ

1. Применение в экологических исследованиях методов математической статистики определяется тем, что ...
а) экосистемы являются стохастически-детерминированными системами;
б) экосистемы являются динамическими системами;
в) экосистемы являются саморегулирующими системами.

2. В большинстве случаев компьютерную обработку данных целесообразно начинать с ...
а) классификации данных;
б) шкалирования данных;
в) составления сводных таблиц.

3. После создания входной электронной таблицы необходимо проверить ...
а) формат таблицы;
б) количество столбцов;
в) количество строк;
г) качество полученных данных.

4. Объектом статистического наблюдения является …
а) совокупность элементов, подлежащих обследованию;
б) первичный элемент, от которого получают информацию;
в) первичный элемент, признаки которого регистрируются;
г) общественное явление, подлежащие обследованию.

5. По полноте охвата единиц совокупности различают наблюдение ...
а) сплошное и несплошное;
б) периодическое;
в) единовременное;
г) текущее.

6. Для характеристики номинальных данных наиболее часто используются …
а) пропорция и процентное отношение;
б) абсолютные величины;
в) логарифмы чисел.

7. Порядковые шкалы соответствуют таким качественным переменным, для которых характерна …
а) упорядоченность;
б) непрерывность;
в) дискретность.

8. Ошибка при оценке плотности популяции должна составлять не более …
а) 5 %;
б) 10 %;
в) 20 %;
г) 30 %.

9. Правильные значения, оценивающие показатель обилия популяции следующие ...
а) %;
б) экз / час;
в) штук;
г) кг / $м^2$.

10. Гистограмма применяется для графического изображения ...
а) дискретных рядов распределения;
б) интервальных рядов распределения;
в) ряда накопленных частот;
г) прерывного ряда распределения.

11. Основанием группировки может быть признак …
а) результирующий;
б) количественный;
в) качественный;
г) как качественный, так и количественный.

12. Средняя величина признака равна 20, а коэффициент вариации – 25 %.Дисперсия признака равна ...
а) 20;
б) 25;
в) 125;
г) 45.

13. Медианой называется …

а) среднее значение признака в ряду распределения;

б) наиболее часто встречающееся значение признака в данном ряду;

в) значение признака, делящее совокупность на две равные части;

г) наиболее редко встречающееся значение в данном ряду.

14. Модой называется …

а) среднее значение признака в данном ряду распределения;

б) наиболее часто встречающееся значение признака в данном ряду;

в) значение признака, делящее данную совокупность на две равные части;

г) наиболее редко встречающееся значение в данном ряду.

15. Абсолютные показатели вариации:

а) размах вариации;

б) коэффициент корреляции;

в) коэффициент осцилляции;

г) коэффициент вариации.

16. К относительным показателям вариации относятся...

а) размах вариации;

б) дисперсия;

в) коэффициент вариации;

г) среднее линейное отклонение.

17. Для значений признака: 3, 5, 6, 9, 11, 12, 13. Мода …

а) отсутствует;

б) 3;

в) 13;

г) 9.

18. Для следующих значений признака: 3, 3, 3, 4, 4, 6, 7, 9, 9 мода …

а) отсутствует;

б) 3;

в) 13;

г) 9.

19. Средний квадрат отклонений вариантов от средней величины – это …
а) коэффициент вариации;
б) размах вариации;
в) дисперсия;
г) среднее квадратическое отклонение.

20. Величина ошибки измерения параметров популяции зависит от ...
а) изменчивости признака;
б) размеров выборки;
в) численности популяции;
г) величины генеральной совокупности.

21. Параметры моделей ранговых распределений оцениваются с помощью …
а) регрессионного анализа;
б) дисперсионного анализа;
в) дискриминантного анализа;
г) факторного анализа.

22. Если принят уровень значимости p = 0,05, а значение критерия t составило 1,5 то анализируемая варианта ...
а) принадлежит генеральной совокупности;
б) не принадлежит генеральной совокупности.

23. Если принят уровень значимости 0,05 и значение критерия Стьюдента больше 3, то ...
а) данное значение не относится к анализируемой совокупности (выборке), включающей 95 %;
б) данное значение относится к анализируемой совокупности (выборке), включающей 95 %.

24. Частное от деления стандартного отклонения на среднюю и умноженное на 100 % называется ...
а) коэффициент вариации;
б) дисперсия;
в) ошибка средней;
г) погрешность средней.

25. Чтобы сравнить по уровню изменчивости признаки любой размерности (выраженные в различных единицах измерения), применяют ...
а) коэффициент вариации;
б) стандартное отклонение;
в) дисперсию.

26. По дисперсии можно сравнивать изменчивость ...
а) одних и тех же показателей;
б) разных признаков по абсолютной величине.

27. Характер распределения при котором в интервале от M – 1,96 ; M + 1,96 лежат 95 % вариант близок к ...
а) нормальному;
б) биноминальному;
в) Пуассоновскому.

28. Показатель скошенности распределения в левую или правую сторону по оси абсцисс называется ...
а) коэффициент асимметрии;
б) эксцесс;
в) коэффициент вариации.

29. Показатель островершинности кривой распределения данных называется ...
а) эксцесс;
б) коэффициент асимметрии;
в) коэффициент вариации.

30. Эксцесс у признаков с нормальным распределением обычно принимает значение в диапазоне ...
а) 1–2;
б) 2–4;
в) 4–6.

31. Степень соответствия выборочных показателей генеральным параметрам называется ...
а) репрезентативностью;
б) адекватностью;
в) выборкой;
г) генеральной совокупностью.

32. Характеристика, показывающая, в каких пределах могут отклоняться от параметров генеральной совокупности частные определения, называется ...
а) ошибкой достоверности;
б) выборочной ошибкой;
в) ошибкой репрезентативности.

33. Достаточным уровнем достоверности в экологических исследованиях принято считать ...
а) $p < 0,05$;
б) $p < 0,1$;
в) $p < 0,001$;
г) $p < 0,3$.

34. Достоверность различий средних арифметических можно оценить по критерию ...
а) Стьюдента;
б) Фишера;
в) Колмогорова-Смирнова.

35. Достоверность различий дисперсий можно оценить по критерию ...
а) Стьюдента;
б) Фишера;
в) критерию хи-квадрат.

36. Решение о достоверности различий средних арифметических принимается в случае, если ...
а) величина $t_{набл}$ превышает табличное значение для данного числа степеней свободы;
б) величина $t_{набл}$ меньше табличного значения для данного числа степеней свободы.

37. Метод статистического вывода, который применяется в отношении параметров генеральной совокупности, называется ... методом
а) параметрическим;
б) непараметрическим;
в) искусственным.

38. Главным условием для параметрических методов является ...
а) нормальность распределения переменных;
б) небольшой объем выборки;
в) большой объем выборки.

39. Если непараметрические методы применяются для количественных данных, то необходимо применить следующие дополнительные вычисления ...
а) ранжирование переменных;
б) применение весовых сравнений;
в) сравнение частот;
г) проверка на биноминальный закон распределения
ограничение выборки.

40. К непараметрическим критериям сравнения выборок относятся следующие...
а) t-критерий Стьюдента;
б) критерий Вилкоксона;
в) критерий знаков;
г) критерий Фишера.

41. Репрезентативностью выборки – это ...
а) степень соответствия характеристик выборки истинным характеристикам биологического объекта;
б) степень несоответствия характеристик выборки истинным характеристикам биологического объекта;
в) точность исследования;
г) способ, когда изучается лишь небольшая группа объектов.

42. Из всех перечисленных статистических характеристик параметров популяции меры вариации следующие ...
а) средняя арифметическая;
б) дисперсия;
в) коэффициент вариации;
г) стандартное отклонение;
д) мода;
е) медиана.

43. Величина ошибки измерения параметров популяции зависит от ...
а) изменчивости признака;
б) размеров выборки;
в) численности популяции;
г) величины генеральной совокупности.

44. Равенство средних величин для нормально распределенных совокупностей в случае равенства их дисперсий можно определить с помощью …
а) F-критерия;
б) критерия хи-квадрат;
в) t-критерия Стьюдента.

45. Согласованность эмпирического распределения с теоретическим можно определить с помощью …
а) F-критерия;
в) критерия хи-квадрат;
г) t-критерия Стьюдента.

46. Вариационный ряд – это ряд распределения, построенный по ... признаку
а) количественному;
б) качественному;
в) непрерывному;
г) количественному и качественному.

47. Аналогом t-критерия сравнения двух независимых выборок в непараметрической статистике является ...
а) Критерий Манна-Уитни;
б) Тест Колмогорова-Смирнова;
в) Критерий Спирмена.

48. При реализации метода Манна-Уитни необходимым объемом выборки является ...
а) 12–40;
б) 2–10;
в) 50–100;
г) 70–100.

49. Условием применения теста Колмогорова-Смирнова является ...
а) количество категорий для тестируемых переменных ограничено;
б) количество категорий для тестируемых переменных неограничено;
в) широкая шкала изменчивости переменной.

50. Метод, применяющийся для исследования влияния одной или нескольких качественных переменных на одну зависимую количественную переменную называется ...
а) дисперсионный анализ;
б) метод зависимости;
в) метод независимости;
г) аналитическим методом.

51. Исходным материалом для дисперсионного анализа не могут быть ...
а) равные выборки;
б) неравные выборки;
в) связные выборки;
г) несвязные выборки;
д) выборки объемом <10.

52. В дисперсионном анализе независимая переменная должна иметь ...
а) одну градацию;
б) две градации;
в) три и более градации.

53. Ограничениями дисперсионного анализа являются следующие положения ...
а) дисперсии выборок должны быть однородны;
б) численность выборок не должна быть меньше двух объектов;
в) численность выборок не должна быть больше 20;
г) численность выборок не должна быть больше 100.

54. Вид дисперсионного анализа, в котором разным градациям фактора соответствует одна и та же выборка (зависимые выборки), называется ...
а) ANOVA с повторными измерениями;
б) однофакторным ANOVA;
в) двухфакторным ANOVA.

55. В дисперсионном анализе результативный признак называют ...
а) зависимым признаком;
б) независимым признаком.

56. В дисперсионном анализе влияющие факторы называют...
а) зависимым признаком;
б) независимыми признаками.

57. При наличии одного фактора, влияние которого исследуется, дисперсионный анализ называется ...
а) однофакторным;
б) двухфакторным.

58. Если исследуется одновременное воздействие двух или более факторов, дисперсионный анализ называется ...
а) многофакторным;
б) множественным;
в) однофакторным.

59. Обобщенно задача дисперсионного анализа состоит в том, чтобы из общей вариативности признака выделить следующие частные вариативности ...
а) вариативность, обусловленную действием каждой из исследуемых независимых переменных (факторов);
б) вариативность, обусловленную взаимодействием исследуемых независимых переменных;
в) вариативность случайную, обусловленную всеми неучтенными обстоятельствами;
г) вариативность, обусловленную внешними факторами.

60. Долю общей вариативности результативного признака, обусловленную действием регулируемых факторов можно оценить с помощью критерия ...
а) Фишера;
б) Стьюдента;
в) Колмогорова.

61. Фактическое значение критерия Фишера определяется как ...
а) отношение межфакторной дисперсии к остаточной;
б) отношение межфакторной дисперсии к общей.

62. Табличное значение критерия Фишера определяется исходя из ...
а) заданного уровня значимости;
б) числа степеней свободы для межгрупповой и остаточной дисперсии;
в) отношения межфакторной дисперсии к общей.

63. Степень влияния независимой переменной на зависимую переменную оценивается при помощи ... коэффициента детерминации
а) корреляционного отношения;
б) коэффициента корреляции;
в) критерия Фишера.

64. Корреляционное отношение по своему абсолютному значению колеблется в пределах ...
а) [0; 1];
б) [0; -1];
в) [-1; 1];
г) [-2; 2].

65. Корреляционное отношение расчитывается как ...
а) отношение межгрупповой (факторной) дисперсии к общей;
б) отношение межфакторной дисперсии к остаточной.

66. Чем ближе корреляционное отношение к 1, тем ... влияние оказывает факторный признак на результативный
а) больше;
б) меньше.

67. Вид взаимосвязи между переменными, когда одному конкретному значению аргумента соответствует приближенное значение или некоторое множество значений функции, в той или иной степени близких друг к другу, называется ...
а) корреляционная связь;
б) функциональная связь;
в) линейная связь;
г) нелинейная связь.

68. Сложность изучения корреляционных связей в экологических исследованиях объясняется тем, что ...
а) все признаки в различной степени взаимосвязаны;
б) все признаки не взаимосвязаны.

69. Корреляционный анализ используется для изучения...
а) взаимосвязи явлений;
б) развития явления во времени;
в) структуры явлений;
г) формы взаимосвязи явлений.

70. Парный коэффициент корреляции показывает тесноту …
а) линейной зависимости между двумя признаками на фоне действия остальных, входящих в модель;
б) линейной зависимости между двумя признаками при исключении влияния остальных, входящих в модель;
в) тесноту нелинейной зависимости между двумя признаками связи между результативным признаком и остальными, включенными в модель.

71. Коэффициент корреляции принимает значения в промежутке ...
а) [0; 1];
б) [0; -1];
в) [-1; 1];
г) [-2; 2].

72. Статистическую значимость коэффициента корреляции Пирсона оценивают с помощью критерия ...
а) Фишера;

б) Стьюдента;
в) Колмогорова.

73. Таблица, в которой на пересечении соответствующих строк и столбцов находится коэффициент корреляции между соответствующими параметрами называется ...
а) корреляционная матрица;
б) матрица регрессии.

74. В результате проведения регрессионного анализа получают функцию, описывающую ...
а) взаимосвязь показателей;
б) соотношение показателей;
в) структуру показателей;
г) темпы роста показателей.

75. Статистический метод исследования зависимости между зависимой переменной и одной или несколькими независимыми переменными называется ...
а) регрессионным анализом;
б) дисперсионным анализом;
в) факторным анализом.

76. Существуют следующие виды регрессионного анализа ...
а) одномерный;
б) многомерный;
в) линейный;
г) нелинейный;
д) факторный.

77. Для оценки параметров регрессий, линейных по параметрам, используют ...
а) метод наименьших квадратов;
б) метод временных рядов.

78. Задача регрессионного анализа понимается как задача выявления такой функциональной зависимости, которая ...
а) наилучшим образом описывает имеющиеся экспериментальные данные;
б) худшим образом описывает экспериментальные данные.

79. Основной функцией методов многомерного анализа является ...
а) выявление скрытой структуры экологического явления;
б) определение элементов экологической системы.

80. По исходным представлениям о числе общих факторов выделяют анализ ...
а) двухфакторный;
б) многофакторный;
в) однофакторный;
г) многокритериальный.

81. Позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие анализ ...
а) двухфакторный дисперсионный;
б) множественный регрессионный.

82. Показателями множественного регрессионного анализа являются ...
а) коэффициенты регрессии;
б) стандартизированные коэффициенты регрессии;
в) коэффициент множественной корреляции;
г) коэффициент множественной детерминации;
д) критерий Фишера и его достоверность;
е) критерий Стьюдента.

83. Наиболее простой среди множества вариантов кластеризации, не требующий обязательного использования ЭВМ – это метод ...
а) «ближайшего соседа»;
б) «дальнего соседа»;
в) «сходства».

84. Нахождение «расстояния» (меры различия) между объектами по всей совокупности параметров и их изображение графически – суть анализа ...
а) кластерного;
б) факторного;
в) регрессионного.

85. Плотность кластеров – это …
а) относительное скопление точек по сравнению с другими;
б) неравномерное распределение точек;
в) стохастическое распределение точек.

86. При многомерном статистическом анализе используется параметрический критерий …
а) знаков;
б) Фишера;
в) Спирмена.

87. Первый этап кластерного анализа предусматривает ...
а) отбор выборки для кластеризации;
б) выбор критерия сравнения средних величин.

88. Последовательное объединение объектов в так называемые в группы, где сходство между объектами выше,чем с другими объектами – это метод ...
а) кластеризации;
б) регрессии;
в) дисперсионного анализа.

89. Раздел статистики, содержанием которого является разработка методов решения задач различения (дискриминации) объектов наблюдения по определенным признакам называется ...
а) дискриминантный анализ;
б) регресссионный анализ;
в) дисперсионный анализ.

90. Для обнаружения факторов, влияющих на измеряемые переменные, используются методы ...
а) факторного анализа;
б) регресссионного анализа;
в) дисперсионного анализа.

91. Задачей факторного анализа является ...
а) представление наблюдаемых параметров в виде линейных комбинаций факторов;
б) отбор выборки для кластеризации;
в) выбор критерия сравнения средних величин.

92. Метод нахождения канонической корреляции, основанный на построении таких линейных комбинаций признаков (в двух заданных группах признаков), что обычный коэффициент парной корреляции между этими комбинациями достигает наибольшего значения называется ...

а) канонический анализ;

б) регресссионный анализ;

в) дисперсионный анализ.

93. Метод исследования, при котором рассматривается более двух факторов одновременно называется ...

а) многофакторный анализ;

б) регресссионный анализ.

в) дисперсионный анализ.

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян А. М. Прикладная статистика и основы эконометрики : учебник / А. М. Айвазян, В. С. Мхитарян. – М. : ЮНИТИ, 1998. – 1022 с.

2. Баканов А. И. Количественная оценка доминирования в экологических сообществах / А. И. Баканов // ИБВВ АН СССР. – 1987. – 63 с.

3. Баканов А. И. О некоторых методологических вопросах применения системного подхода для изучения структур водных экосистем / А. И. Баканов // Биология внутренних вод. – 2000. – № 2. – С. 5–18.

4. Системная экология : практикум / И. С. Белюченко, Е. И. Муравьев, Е. В. Попок, Л. Б. Попок. – Краснодар : КГАУ, 2007. – 184 с.

5. Бигон М. Экология: особи, популяции, сообщества / М. Бигон, Дж. Харпер, К. Таунсенд. – М. : Мир, 1989. – Т. 1. – 667 с.

6. Боровиков В. П. Статистический анализ и обработка данных в среде Windows / В. П. Боровиков, И. П. Боровиков. – М.: Филин, 1998. – 608 с.

7. Василевич В. И. Статистические методы в геоботанике / В. И. Василевич. – М. : Наука, 1969. – 136 с.

8. Грейг-Смит Р. Количественная экология растений / Р. Грейт-Смит. – М. : Мир, 1967. – 254 с.

9. Гринин А. С. Математическое моделирование в экологии : учеб. пособие / А. С. Гринин, Н. А. Орехов, В. Н. Новиков. – М. : ЮНИТИ-ДАНА, 2003. – 269 с.

10. Джонгман Р. Г. Анализ данных в экологии сообществ и ландшафтов / Р. Г. Джонгман. – М. : РАСХН, 1999. – 306 с.

11. Лакин Г. Ф. Биометрия : учеб. пособие / Г. Ф. Лакин. – М. : Высшая школа, 1980. – 293 с.

СОДЕРЖАНИЕ

У ч е б н о е и з д а н и е

Никифоренко Юлия Юрьевна

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ЭКОЛОГИИ И ПРИРОДОПОЛЬЗОВАНИИ

Учебное пособие

В авторской редакции
Дизайн обложки – Н. П. Лиханская

Подписано в печать 27.12.2019. Формат 60 × 84 $^{1}/_{16}$.
Усл. печ. л. – 5,1. Уч.-изд. л. – 4,1.
Тираж 50 экз. Заказ № 4

Типография Кубанского государственного
аграрного университета.
350044, г. Краснодар, ул. Калинина, 13