

Министерство образования Республики Беларусь
Учреждение образования «Витебский государственный
университет имени П.М. Машерова»
Кафедра экологии и охраны природы

Г.Г. Сушко, И.А. Литвенкова

БИОМЕТРИЯ

*Методические указания
для проведения лабораторных работ*

В 2 частях

ЧАСТЬ 2

*Витебск
ВГУ имени П.М. Машерова
2019*

УДК 57(076.5)
ББК 28в631я73
С91

Печатается по решению научно-методического совета учреждения образования «Витебский государственный университет имени П.М. Машерова». Протокол № 6 от 26.06.2019.

Авторы: заведующий кафедрой экологии и охраны природы ВГУ имени П.М. Машерова, кандидат биологических наук, доцент **Г.Г. Сушко**; доцент кафедры экологии и охраны природы ВГУ имени П.М. Машерова, кандидат биологических наук **И.А. Литвенкова**

Рецензент:
заведующий кафедрой химии УО «ВГАВМ»,
кандидат биологических наук, доцент *В.П. Баран*

Сушко, Г.Г.
С91 Биометрия : методические указания для проведения лабораторных работ : в 2 ч. / Г.Г. Сушко, И.А. Литвенкова. – Витебск : ВГУ имени П.М. Машерова, 2019. – Ч. 2. – 47 с.

Методические указания подготовлены в соответствии с типовой и учебной программами по курсу «Биометрия». Приводятся методики регрессионного, корреляционного, дисперсионного и многофакторного анализа биологических данных.

Издание предназначено студентам, обучающимся по биологическим специальностям, а также лицам, занимающимся биологическими и экологическими исследованиями.

УДК 57(076.5)
ББК 28в631я73

© Сушко Г.Г., Литвенкова И.А., 2019
© ВГУ имени П.М. Машерова, 2019

СОДЕРЖАНИЕ

Введение	4
Модуль 2. Анализ данных	5
Лабораторная работа № 7. Основы дисперсионного анализа	5
Лабораторная работа № 8. Основы корреляционного анализа	21
Лабораторная работа № 9. Основы регрессионного анализа	32
Лабораторная работа № 10. Элементы многомерной статистики (много- факторный анализ)	39
Итоговое занятие по модулю 2	44
Рекомендуемая литература	46

ВВЕДЕНИЕ

Предлагаемое учебное издание включает практические рекомендации и задания по модулю 2 «Анализ данных» курса «Биометрия». Модуль состоит из методических рекомендаций по проведению лабораторных работ, которые содержат тему и цель занятия, перечень необходимого оборудования и программного обеспечения, список терминов и понятий, знание которых обязательно для выполнения лабораторного занятия, задания для выполнения лабораторных работ и контрольные вопросы. Студентам предложен пошаговый алгоритм с иллюстрациями действий в MS Excel, PAST, Statistica.

Для осуществления текущего контроля в конце каждого лабораторного задания приводится список контрольных вопросов. В конце модуля имеются вопросы и задания для итогового контроля по модулю.

Содержание данного издания предполагает освоение навыков и умений студентами в области основ регрессионного, корреляционного, дисперсионного и многофакторного анализа биологических данных.

При подготовке методических указаний применены собственный опыт авторов по статистической обработке данных, научная и методическая литература.

МОДУЛЬ 2. АНАЛИЗ ДАННЫХ

ЛАБОРАТОРНАЯ РАБОТА № 7

Основы дисперсионного анализа

Цель: получить практические навыки и закрепить на конкретных примерах знания о методиках проведения дисперсионного анализа.

Программное обеспечение: базы данных MS Excel, пакеты анализа Statistica, PAST.

Основные термины и понятия: зависимая и независимая переменные; однофакторный дисперсионный анализ (One-way ANOVA), двухфакторный дисперсионный анализ (Two-way ANOVA), понятие о многофакторном дисперсионном анализе; нулевая гипотеза при дисперсионном анализе; F-критерий Фишера, степени свободы (df); разведочный анализ: проверка на нормальность распределения (визуальный анализ гистограммы распределений и тесты Колмогорова-Смирнова, Шапиро-Уилка), проверка равенства групповых дисперсий (тесты Левене, Баррета, Кохрана); апостериорный (post-hoc) анализ: тесты Тьюки, Шеффе, Даннета; непараметрический дисперсионный анализ: тесты Крускала-Уолиса и Фридмана; поправка Бонферрони.

Задание 1. Выполнить параметрический однофакторный дисперсионный анализ в пакете Statistica. Выяснить наличие зависимости массы плодов при использовании трех различных удобрений.

- 1) Загрузить таблицу с данными из MS Excel (Рисунок 7.1).

No	удобрение	масса плодов, г
1	удобрение 1	50
2	удобрение 1	49
3	удобрение 1	51
4	удобрение 1	55
5	удобрение 1	56
6	удобрение 1	56
7	удобрение 1	57
8	удобрение 1	59
9	удобрение 1	60
10	удобрение 1	50
11	удобрение 2	51
12	удобрение 2	62
13	удобрение 2	62
14	удобрение 2	63
15	удобрение 2	59
16	удобрение 2	60
17	удобрение 2	64
18	удобрение 2	66
19	удобрение 2	71
20	удобрение 2	71
21	удобрение 3	64
22	удобрение 3	62
23	удобрение 3	62
24	удобрение 3	63
25	удобрение 3	68
26	удобрение 3	67
27	удобрение 3	69
28	удобрение 3	68
29	удобрение 3	73
30	удобрение 3	74

Рисунок 7.1 – Загрузка базы данных из MS Excel в Statistica

2) Запустить модуль One-way ANOVA: закладка Анализ/Дисперсионный анализ, выбрать однофакторный ДА (Рисунок 7.2)

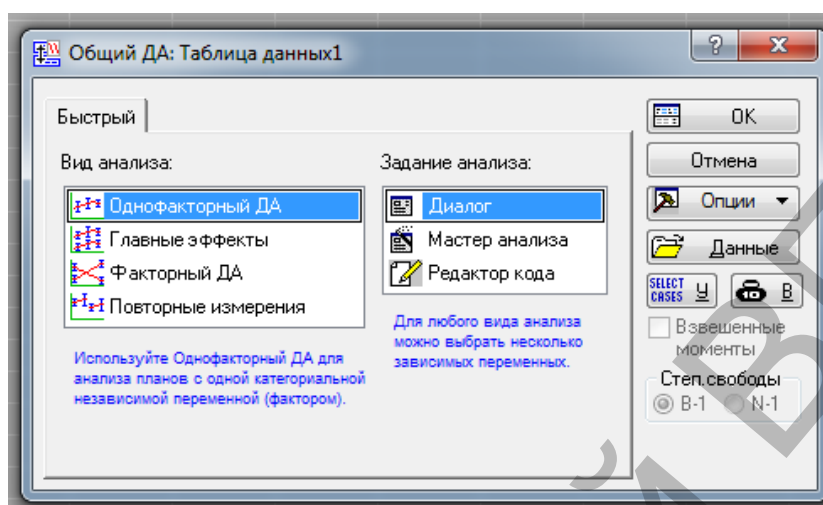


Рисунок 7.2 – Запуск модуля One-way ANOVA

3) Выбрать зависимую (масса плодов) и независимую (удобрение) переменные (Рисунок 7.3).

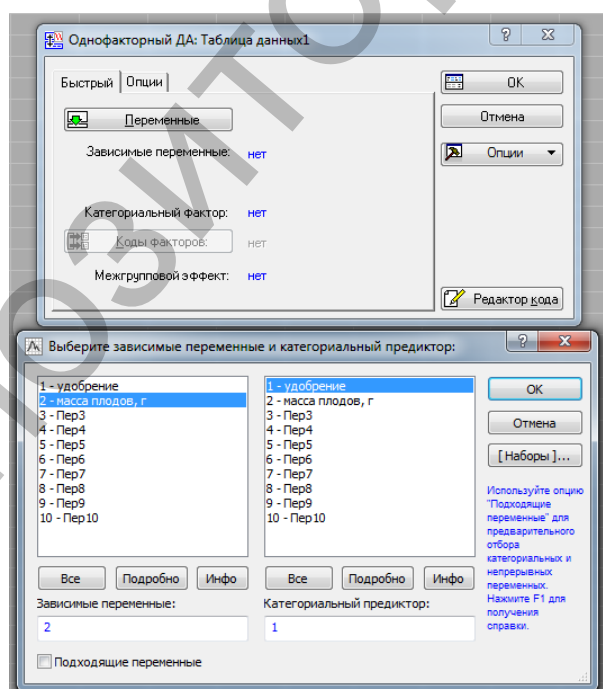


Рисунок 7.3 – Выбор зависимой и независимой переменных

4) провести разведочный анализ данных, указывающий на то, что наблюдаются следующие обязательные условия, позволяющие применить

параметрический дисперсионный анализ, такие как нормальность распределения признаков, однородность групповых дисперсий (т.е. между ними нет статистически значимой разницы). Следует учесть, что все сравниваемые выборки должны быть независимы. Если данные требования не соблюдаются, применяют непараметрический аналог One-way ANOVA, например тест Крускала-Уолиса.

Проверка нормальности распределения.

Откройте модуль описательные статистики, выберите закладку нормальность, отметьте необходимый критерий, например Шапиро-Уилка, и нажмите на Гистограммы (Рисунок 7.4).

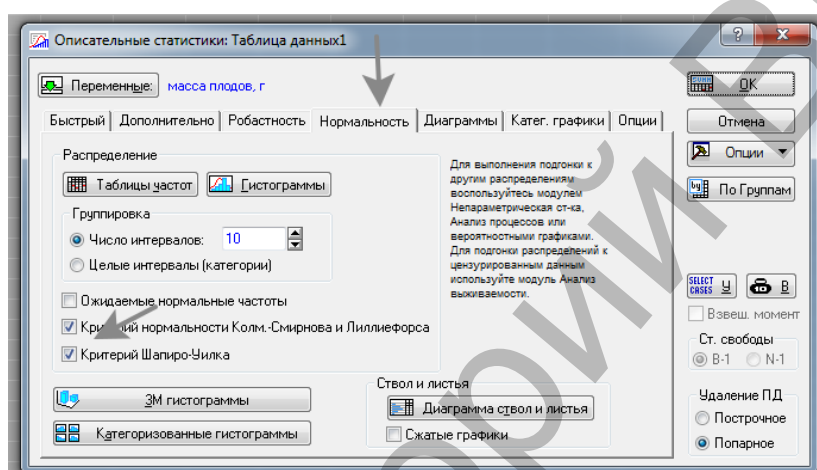


Рисунок 7.4 – Выбор критерия Шапиро-Уилка

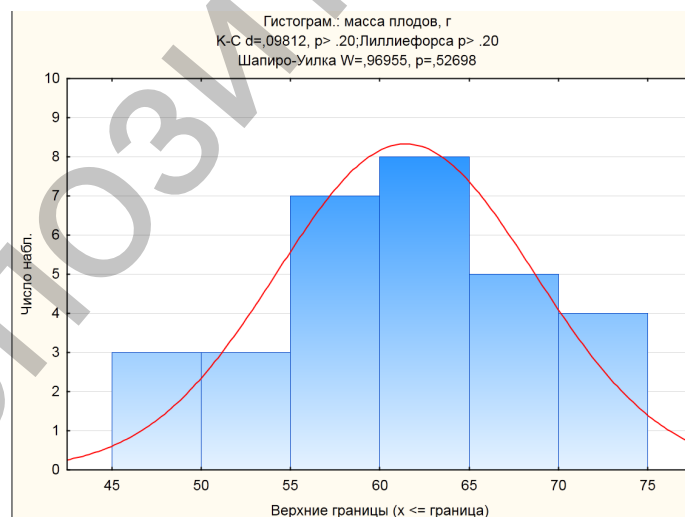


Рисунок 7.5 – Распределение данных показателя массы плодов

Как видно на графике (Рисунок 7.5) и по показателю критерия Шапиро-Уилка ($p > 0,05$), первое условие соблюдается.

Проверка однородности групповых дисперсий.

Возвращаемся в модуль Дисперсионный анализ, выбираем закладку Больше, а затем один из критериев, например Левене (Рисунок 7.6).

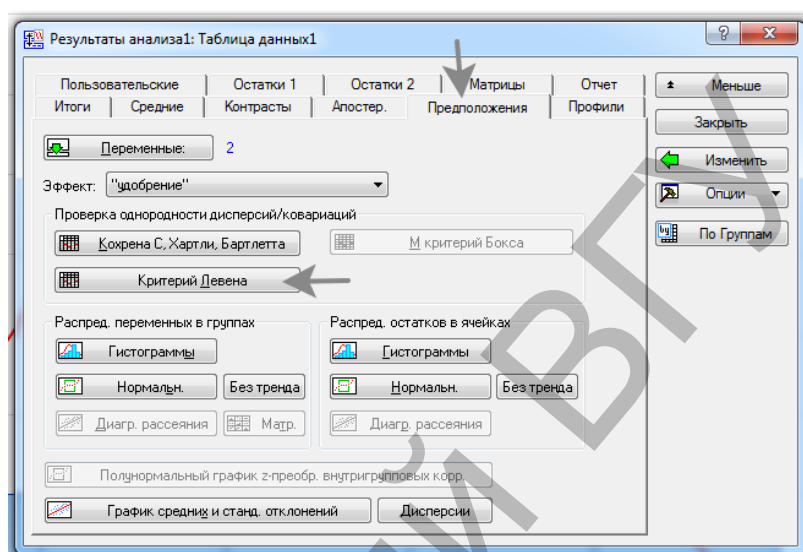


Рисунок 7.6 – Выбор критерия Левене

Итоги проверки на гомогенность дисперсии появятся в виде таблицы (Рисунок 7.7):

Критерий Левене однородности дисперсий (Таблица данных1)					
Эффект: "удобрение"					
Степени свободы для всех F: 2, 27					
	MS	MS	F	p	
Эффект	Ошибка				
масса плодов, г	2,052000	8,180148	0,250851	0,779932	

Рисунок 7.7 – Вывод данных итоговой проверки на гомогенность дисперсии по критерию Левене

Так как итоги проверки с помощью теста Левене статистически не значимы ($p > 0,05$), второе условие также соблюдается.

Таким образом, анализируемые данные удовлетворяют условиям, необходимым для параметрического дисперсионного анализа.

5) Выполнить анализ, выбрав Итоги/размеры эффектов (Рисунок 7.8).

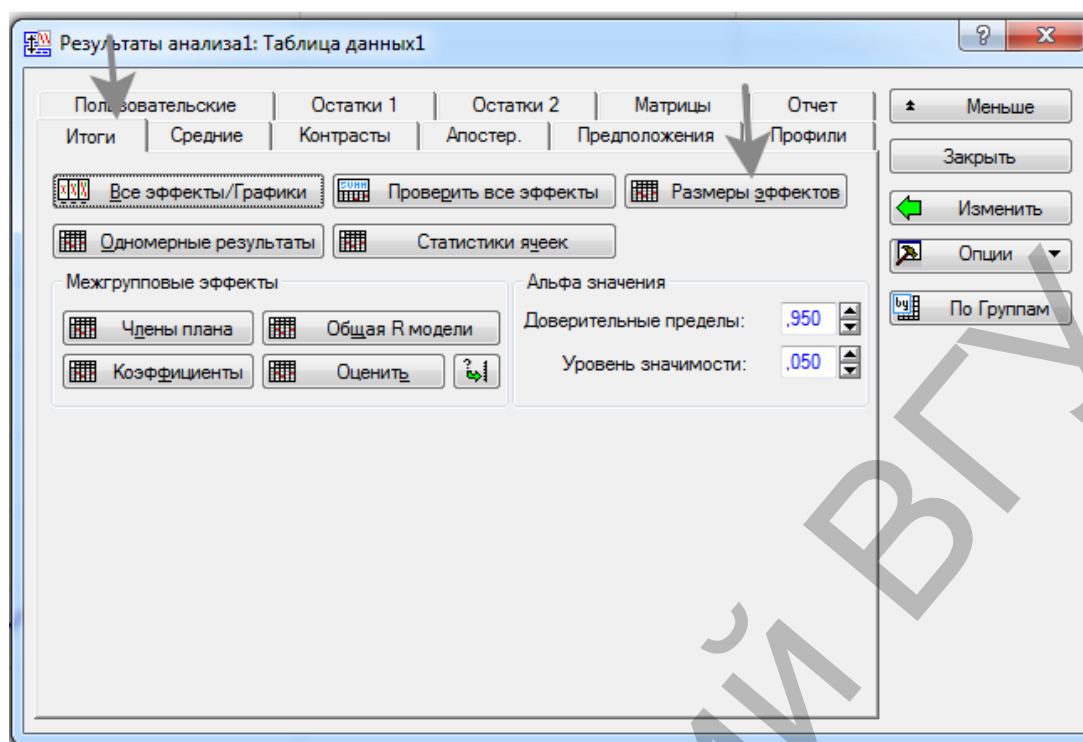


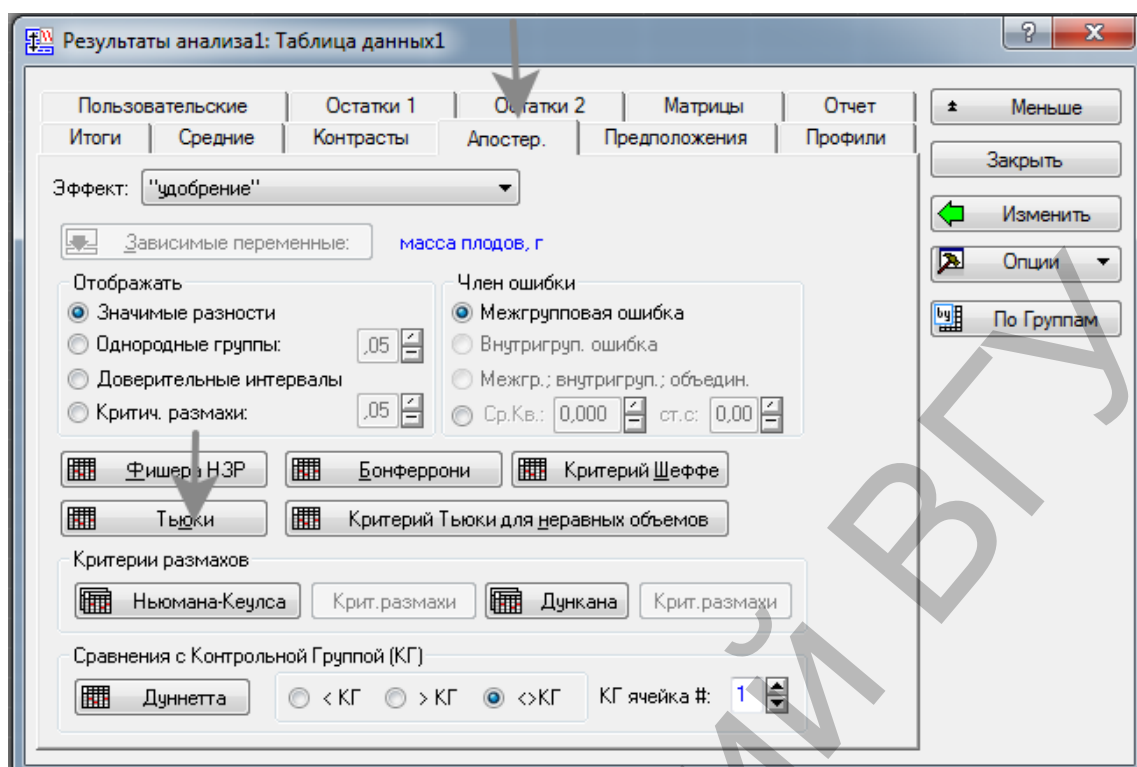
Рисунок 7.8 – Выбор опции Размеры эффектов

Исходя из таблицы результатов (Рисунок 7.9) анализа видно, что $p < 0,05$. Следовательно, средняя масса плодов статистически значимо отличается в зависимости от использования удобрения.

Одномерные критерии значимости, размеров эффекта и мощности для масса плодов, г (Таблица дан Сигма-ограниченная параметризация Декомпозиция гипотезы								
Эффект	SS	Степени свободы	MS	F	p	Частичная эта-квадрат.	Нецентрированная	Наблюдаемая мощность. (альфа=0,05)
Св. член	113221,6	1	113221,6	4945,776	0,000000	0,994570	4945,776	1,000000
удобрение	877,3	2	438,6	19,160	0,000007	0,586657	38,321	0,999783
Ошибка	618,1	27	22,9					

Рисунок 7.9 – Вывод таблицы данных результатов анализа размеры эффектов

6) Провести апостериорные сравнения. Поскольку дисперсионный анализ показывает наличие или отсутствие различий между сравниваемыми переменными, с его помощью нельзя узнать какие именно группы признаков различаются. Для этого предусмотрены множественные попарные (апостериорные сравнения – post-hoc tests) сравнения средних величин. Для их выполнения откройте вкладку Апостер и выберите критерий Тьюки (Рисунок 7.10).



Крит. Тьюки ДЗР; перем. масса плодов, г (Таблица данных1)					
Приближенные вероятности для апостер. критериев					
Ошибка: Межгр. MS = 22,893, сс = 27,000					
Н ячейки	удобрение	{1}	{2}	{3}	
1	удобрение 1	54,200	62,900	67,200	
2	удобрение 2	0,001156		0,129280	
3	удобрение 3	0,000130	0,129280		

Рисунок 7.10 – Апостериорное сравнение
(парное сравнение средних величин)

Из таблицы результатов апостериорных сравнений (Рисунок 7.10) видна статистически значимая разница между парами сравниваемых признаков.

7) Сравнить средние значения графически. Для того чтобы увидеть различия средних значений по каждой группе сравниваемых признаков, можно воспользоваться вкладкой Средние (Рисунок 7.11). Полученный график демонстрирует отличия средних показателей рассматриваемых нами групп признаков.

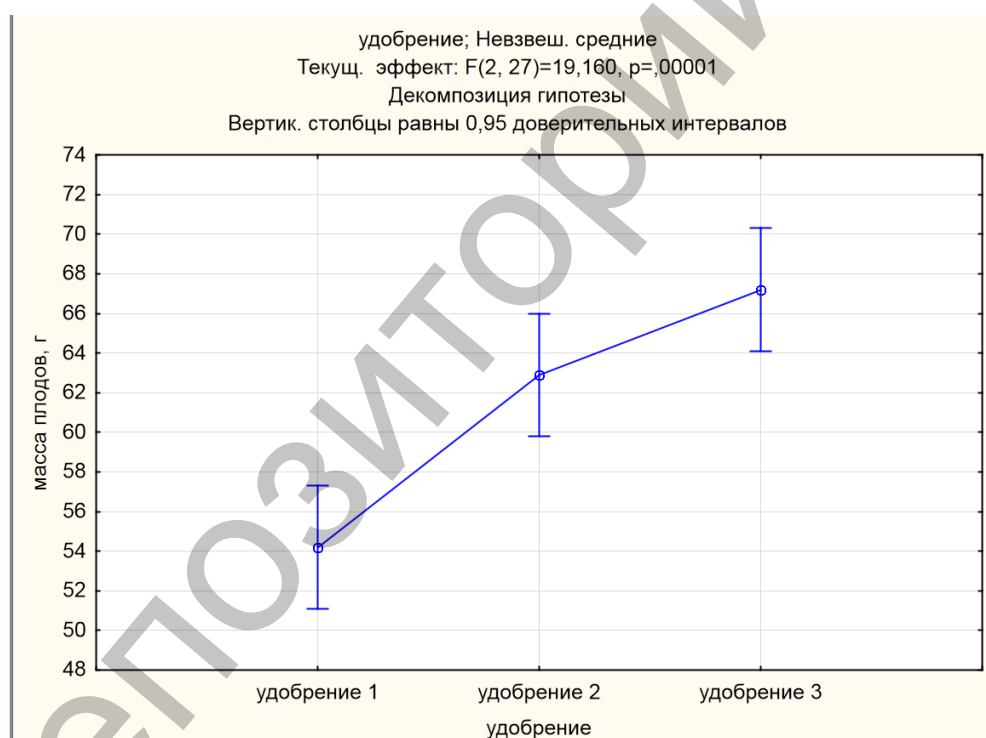
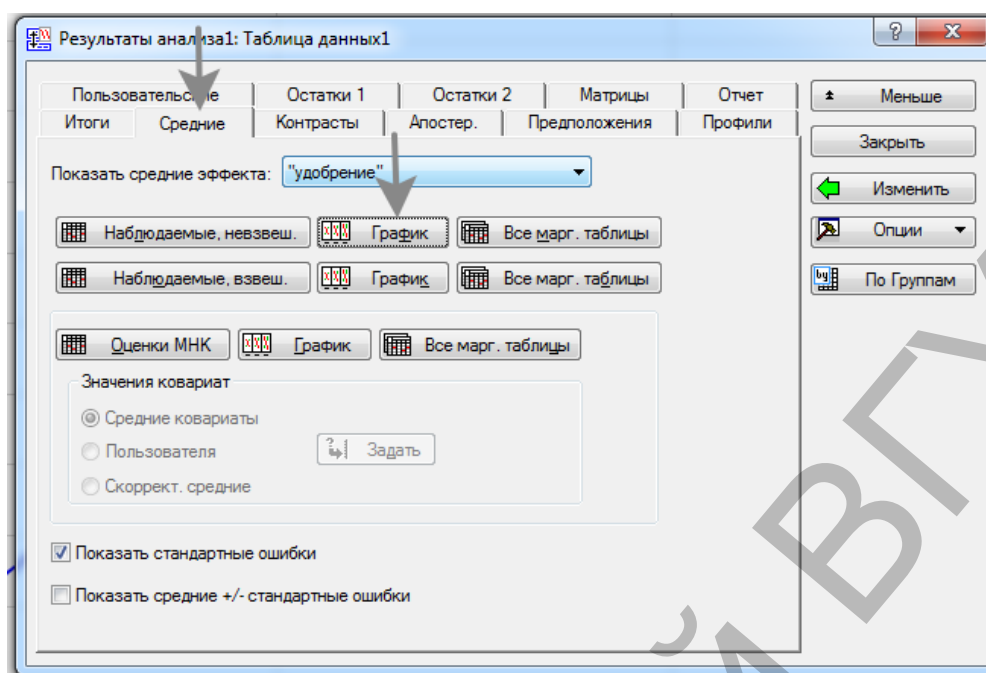


Рисунок 7.11 – Сравнение средних значений графически

Задание 2. Выполнить непараметрический однофакторный дисперсионный анализ в пакете Statistica. Выявить наличие зависимости числа отловленных экземпляров имаго стрекоз от типа биотопа.

Загрузить таблицу с данными из MS Excel (Рисунок 7.12).

№	биотоп	число экземпляров
1	1 сосновый лес	211
2	2 сосновый лес	161
3	3 сосновый лес	184
4	4 сосновый лес	184
5	5 сосновый лес	184
6	6 сосновый лес	184
7	7 сосновый лес	184
8	8 сосновый лес	184
9	9 сосновый лес	184
10	10 берег озера	954
11	11 берег озера	743
12	12 берег озера	791
13	13 берег озера	1098
14	14 берег озера	796
15	15 берег озера	1083
16	16 берег озера	629
17	17 берег озера	588
18	18 берег озера	928
19	19 луг	167
20	20 луг	172
21	21 луг	165
22	22 луг	212
23	23 луг	118
24	24 луг	171
25	25 луг	174
26	26 луг	92
27	27 луг	168

Рисунок 7.12 – Загрузка данных из MS Excel в пакет анализа Statistica

1) Запустить модуль One-way ANOVA: закладка Анализ/Дисперсионный анализ, выбрать однофакторный ДА (Рисунок 7.13).

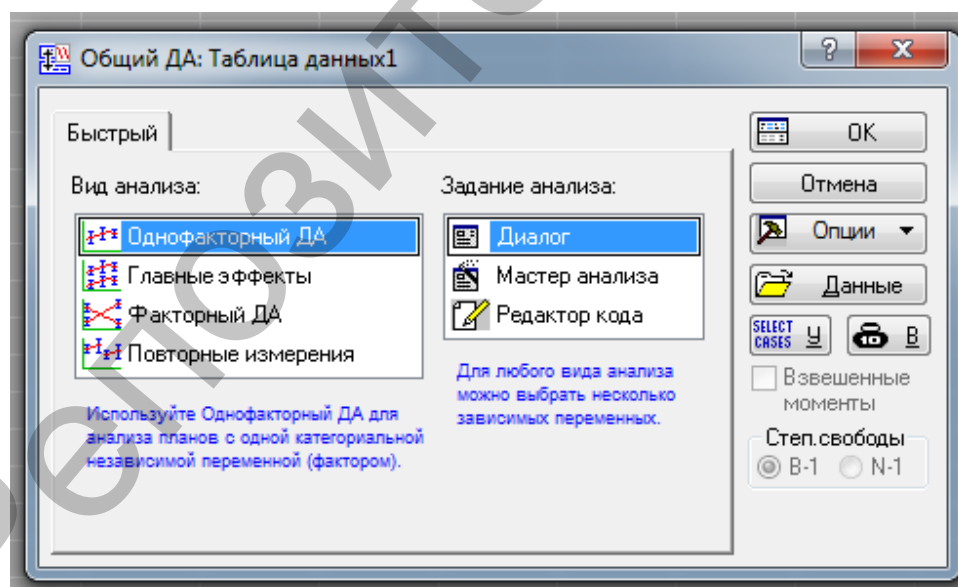


Рисунок 7.13 – Запуск модуля One-way ANOVA

2) Выбрать зависимую (число экземпляров) и независимую (биотоп) переменные (Рисунок 7.14).

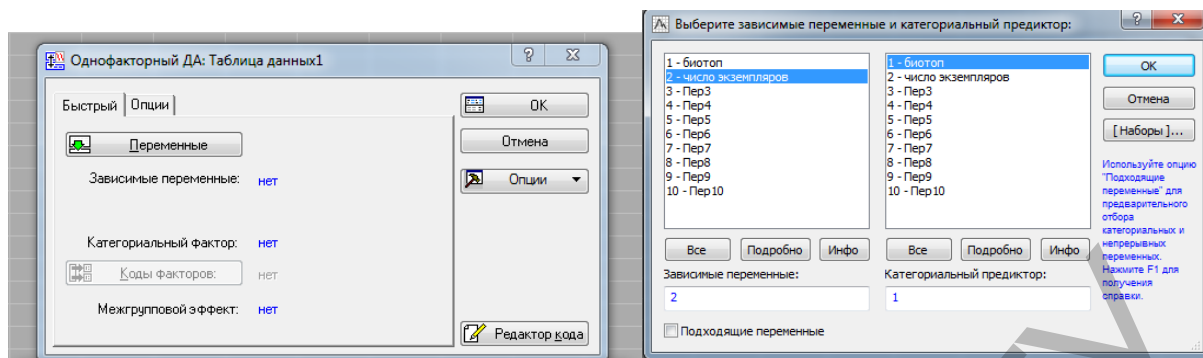


Рисунок 7.14 – Выбор зависимой и независимой переменных

3) провести разведочный анализ данных, указывающий на то, что соблюдаются следующие обязательные условия, позволяющие применить параметрический дисперсионный анализ, такие как нормальность распределения признаков, однородность групповых дисперсий (т.е. между ними нет статистически значимой разницы). Следует учесть, что все сравниваемые выборки должны быть независимы. Если данные требования не соблюдаются, применяют непараметрический аналог One-way ANOVA, например тест Крускала-Уолиса.

Проверка нормальности распределения.

Откройте модуль описательные статистики, выберите закладку нормальность, отметьте необходимый критерий, например Шапиро-Уилка, и нажмите на Гистограммы (Рисунок 7.15).

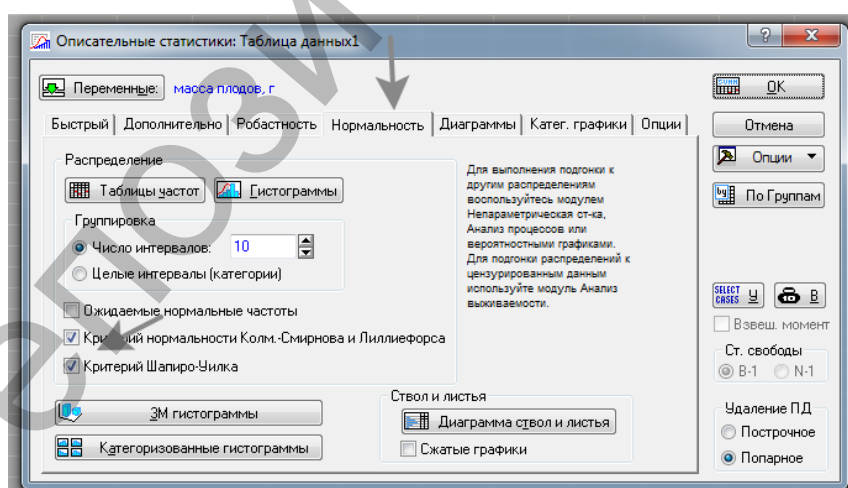


Рисунок 7.15 – Выбор критерия Шапиро-Уилка

Как видно на графике (Рисунок 7.16) и по показателю критерия Шапиро-Уилка ($p < 0,05$), первое условие не соблюдается, так как данные не подчиняются закону нормального распределения.

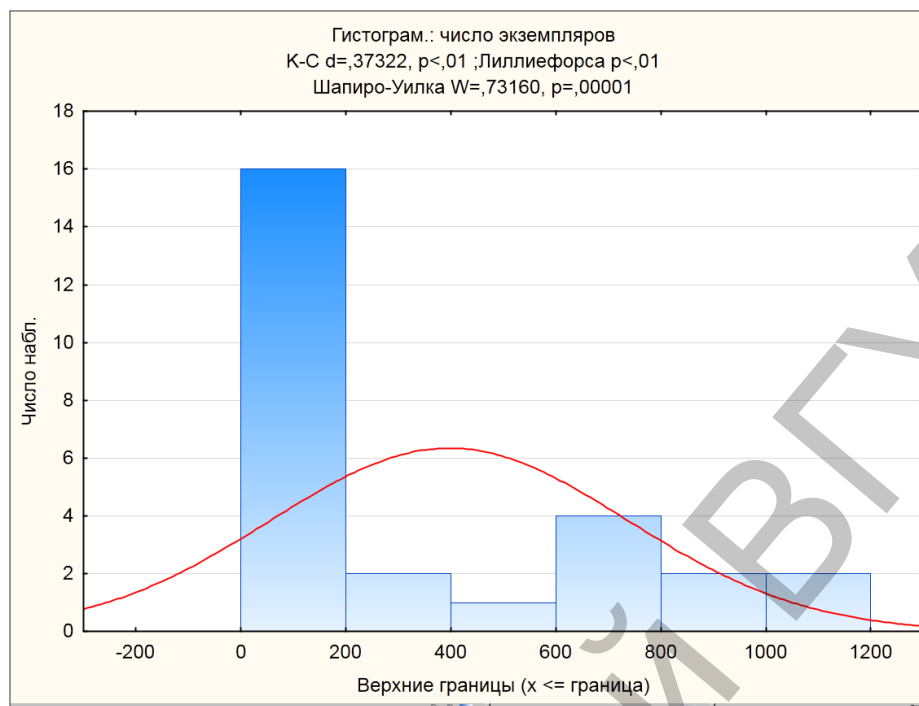


Рисунок 7.16 – Гистограмма распределения исследуемого показателя

Проверка однородности групповых дисперсий.

Возвращаемся в модуль Дисперсионный анализ, выбираем закладку Больше, а затем один из критериев для определения гомогенности дисперсии, например Левене (Рисунок 7.17).

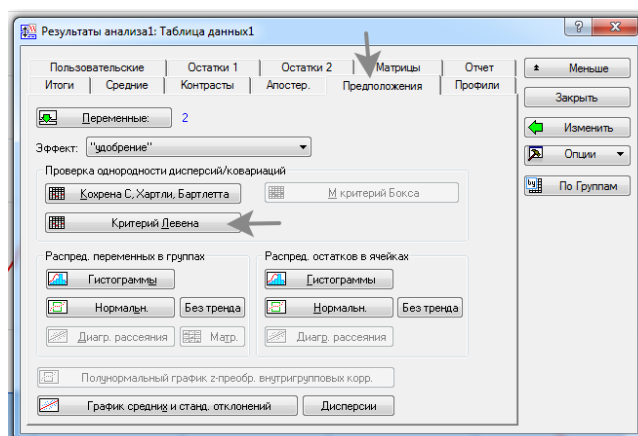


Рисунок 7.17 – Выбор критерия Левене для проверки однородности групповых дисперсий

Итоги проверки на гомогенность дисперсии появятся в виде таблицы (Рисунок 7.18).

Критерий Левена однородности дисперс				
Эффект: биотоп				
Степени свободы для всех F: 2, 24				
	MS	MS	F	p
	Эффект	Ошибка		
число экземпляров	56368,42	2797,597	20,14887	0,000007

Рисунок 7.18 – Таблица итогов проверки на гомогенность

Так как итоги проверки с помощью теста Левена статистически значимы ($p < 0,05$) и групповые дисперсии не являются однородными (гомогенными), второе условие также не соблюдается.

Таким образом, анализируемые данные удовлетворяют условиям, необходимым для непараметрического дисперсионного анализа.

4) выполнить анализ, в модуле Непараметрическая статистика / сравнение нескольких независимых групп, при этом снова выбрав зависимую и независимую переменные и вкладку ДА Крускала-Уолиса и медианный тест или нажав на ОК (Рисунок 7.19).

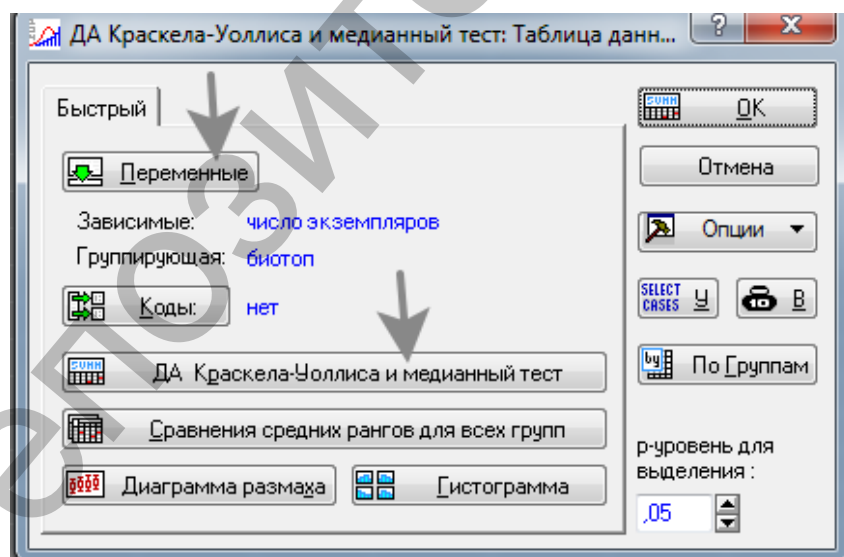


Рисунок 7.19 - Анализ в модуле Непараметрическая статистика

В первом случае вместе с результатом непараметрического дисперсионного анализа Крускала-Уолиса программа выдает и результаты медианного теста, который проверяет ту же нулевую гипотезу, однако является менее мощным (Рисунок 7.20).

		Медианный тест, общ. медиана = 184,000; число экземпляров (Таблица данных1)			
		Груп. (независ.) переменная: биотоп			
		Хи-квадрат = 19,63636 сс = 2 p = ,0001			
Зависимые:	число экземпляров	сосновый лес	берег озера	луг	Всего
<= Медианы:	наблюд.	8,00000	0,00000	8,00000	16,00000
	ожидаемые	5,33333	5,33333	5,33333	
	набл.-ожд.	2,66667	-5,33333	2,66667	
> Медианы:	наблюд.	1,00000	9,00000	1,00000	11,00000
	ожидаемые	3,66667	3,66667	3,66667	
	набл.-ожд.	-2,66667	5,33333	-2,66667	
Сумма: наблюд.		9,00000	9,00000	9,00000	27,00000

Рисунок 7.20 – Таблица с результатом медианного теста

Во втором случае мы видим только результаты Н-теста Крускала-Уолиса (Рисунок 7.21).

		р знач. (2-сторонние) для множест. сравнений; число экземпляров (Таблица данных1)		
		Груп. (независ.) переменная: биотоп		
		Кр.Краскала-Уоллиса: Н (2, N= 27) =19,99255 p =,0000		
Зависим.:	число экземпляров	сосновый лес R:12,333	берег озера R:23,000	луг R:6,6667
сосновый лес			0,013083	0,389712
берег озера		0,013083		0,000038
луг		0,389712	0,000038	

Рисунок 7.21 – Таблица с результатом Н-теста Крускала-Уолиса

Как видно из таблиц, $p < 0,05$. Следовательно, средние показатели числа экземпляров исследуемых организмов статистически значительно отличаются в зависимости от местообитаний. Из таблицы результатов также видна статистически значимая разница между парами сравниваемых признаков (выделены красным).

Различия между анализируемыми переменными можно увидеть и построив диаграмму размаха (Рисунок 7.22). Для этого нужно выбрать вкладку Диаграмма размаха в нижнем правом углу.

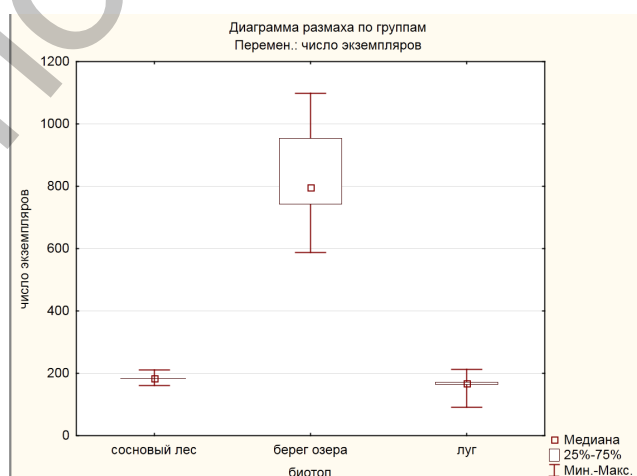
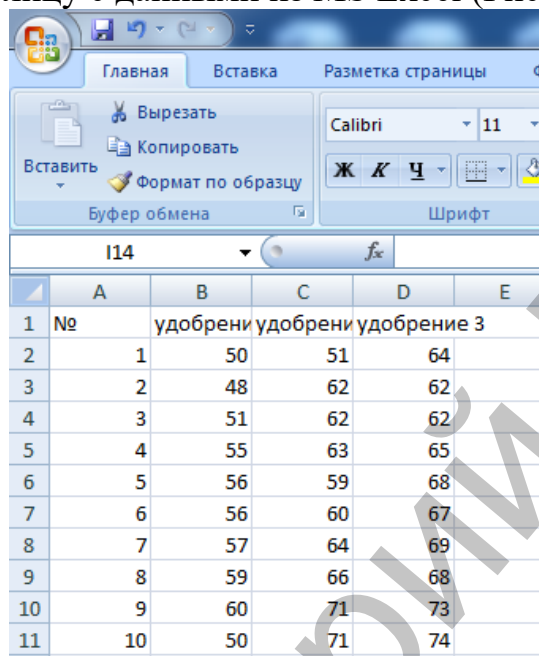


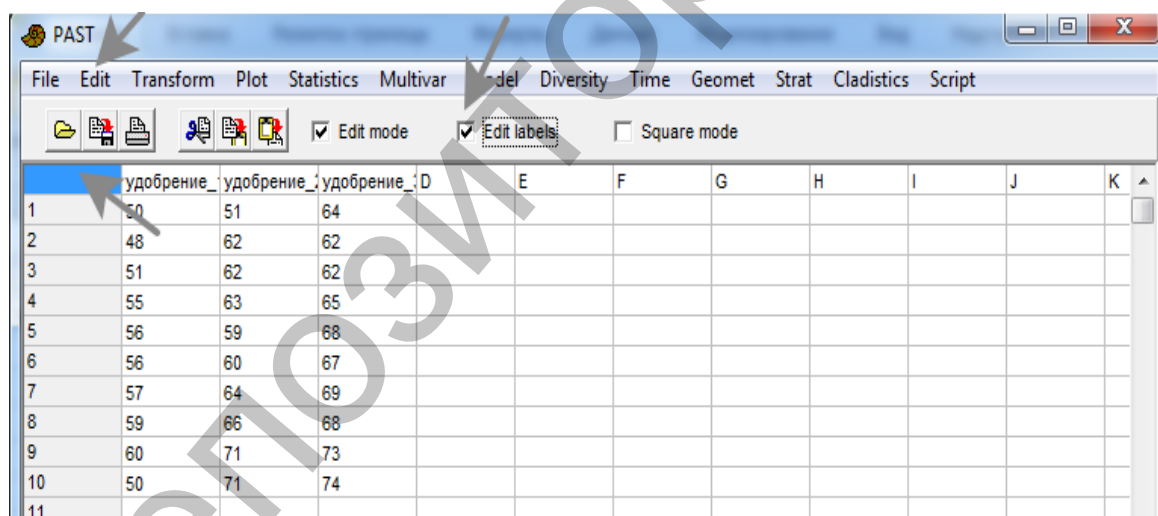
Рисунок 7.22 – Построение диаграммы размаха анализируемых признаков

Задание 3. Выполнить параметрический однофакторный дисперсионный анализ в пакете PAST. Выяснить наличие зависимости массы плодов при использовании трех различных удобрений.

1) Загрузить таблицу с данными из MS Excel (Рисунок 7.23).



	A	B	C	D	E
1	№	удобрение 1	удобрение 2	удобрение 3	
2	1	50	51	64	
3	2	48	62	62	
4	3	51	62	62	
5	4	55	63	65	
6	5	56	59	68	
7	6	56	60	67	
8	7	57	64	69	
9	8	59	66	68	
10	9	60	71	73	
11	10	50	71	74	

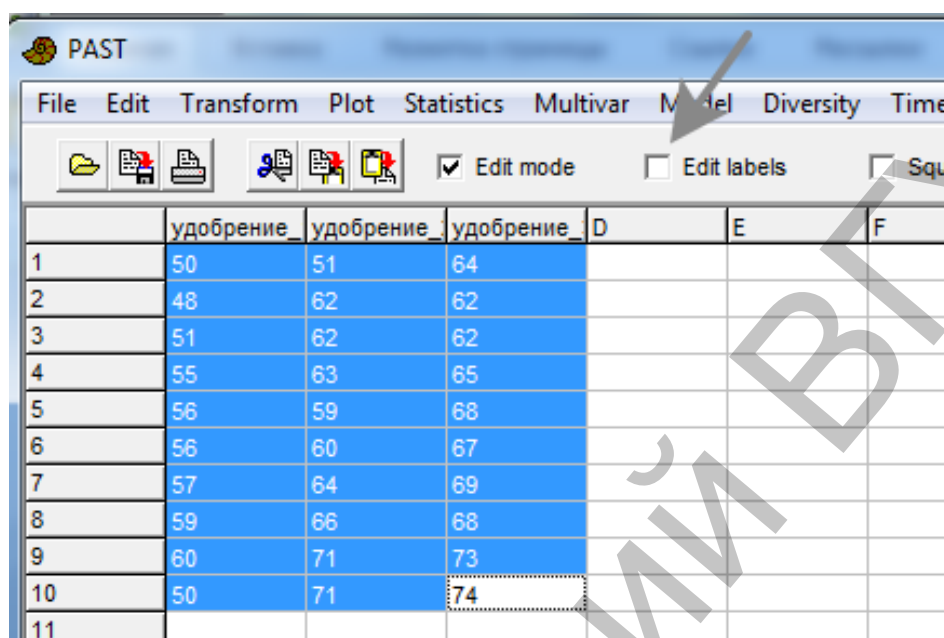


	удобрение_1	удобрение_2	удобрение_3	D	E	F	G	H	I	J	K
1	50	51	64								
2	48	62	62								
3	51	62	62								
4	55	63	65								
5	56	59	68								
6	56	60	67								
7	57	64	69								
8	59	66	68								
9	60	71	73								
10	50	71	74								
11											

Рисунок 7.23 – Загрузка данных из MS Excel в PAST

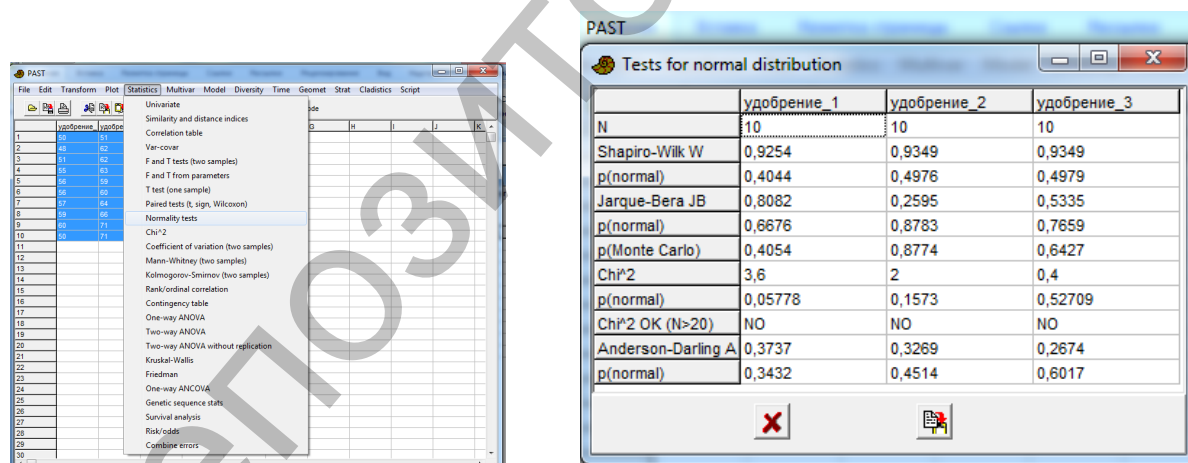
При этом обратите внимание, что в отличие от программы Statistica, данные в PAST заносятся в таблицу по столбцам, соответственно каждой анализируемой переменной. Для вставки нужно выбрать вкладку Edit/Paste. Предварительно необходимо поставить метку Edit labels и установить курсор в верхнюю левую ячейку, помеченную синим цветом.

2) Проверить данные на соответствие закону нормального распределения. Предварительно необходимо снять метку Edit labels и выделить данные (Рисунок 7.24).



	удобрение_1	удобрение_2	удобрение_3	D	E	F
1	50	51	64			
2	48	62	62			
3	51	62	62			
4	55	63	65			
5	56	59	68			
6	56	60	67			
7	57	64	69			
8	59	66	68			
9	60	71	73			
10	50	71	74			
11						

Рисунок 7.24 – Снятие метки Edit labels и выделение данных. Выберите закладку Statistics, а затем Normality test (Рисунок 7.25).



	удобрение_1	удобрение_2	удобрение_3
N	10	10	10
Shapiro-Wilk W	0,9254	0,9349	0,9349
p(normal)	0,4044	0,4976	0,4979
Jarque-Bera JB	0,8082	0,2595	0,5335
p(normal)	0,6676	0,8783	0,7659
p(Monte Carlo)	0,4054	0,8774	0,6427
Chi²	3,6	2	0,4
p(normal)	0,05778	0,1573	0,52709
Chi² OK (N>20)	NO	NO	NO
Anderson-Darling A	0,3737	0,3269	0,2674
p(normal)	0,3432	0,4514	0,6017

Рисунок 7.25 – Выбор модуля Normality test в закладке Statistics

Как видно из таблицы, по показателю критерия Шапиро-Уилка ($p > 0,05$), данные подчиняются закону нормального распределения.

3) Выполнить однофакторный дисперсионный анализ. Запустить модуль One-way ANOVA: закладка Statistics/ One-way ANOVA (Рисунок 7.26).

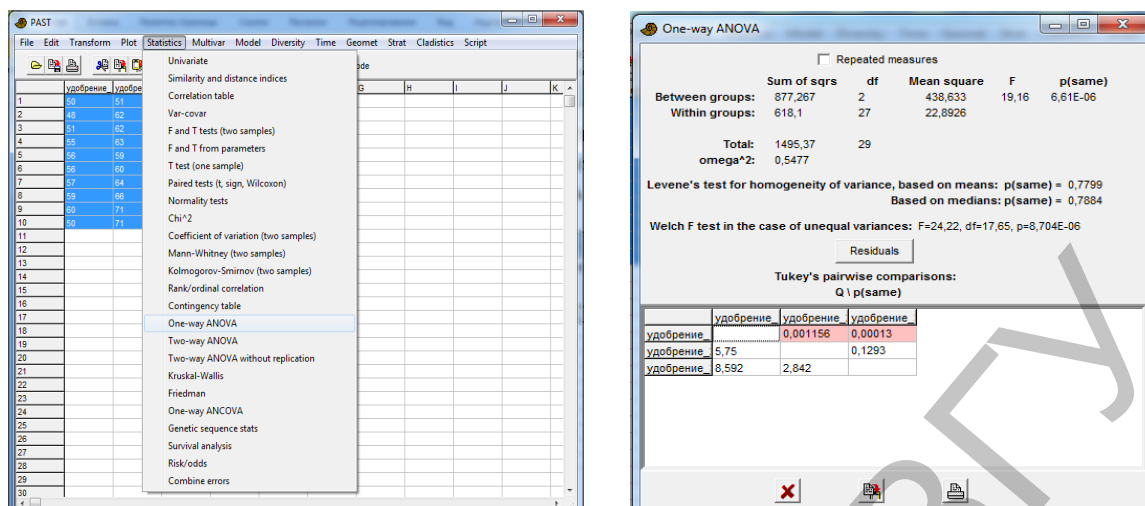


Рисунок 7.26 – Запуск модуля One-way ANOVA для выполнения однофакторного дисперсионного анализа

Перед анализом полученных результатов нужно убедиться в однородности групповых дисперсий. На это указывает тест Левена, значение показателя которого статистически не значимо ($p > 0,05$). Следовательно, требование однородности групповых дисперсий выполняется.

Результаты дисперсионного анализа показаны сверху. Поскольку $p < 0,05$, между анализируемыми данными есть статистически значимые различия. Ниже в таблице приведены результаты попарных апостериорных сравнений (тест Тьюки). Статистически значимые различия выделены красным.

Задание 4. Выполнить непараметрический однофакторный дисперсионный анализ в пакете PAST. Выявить наличие зависимости числа отловленных экземпляров имаго стрекоз от типа биотопа.

1) Загрузить таблицу с данными из MS Excel (Рисунок 7.27).

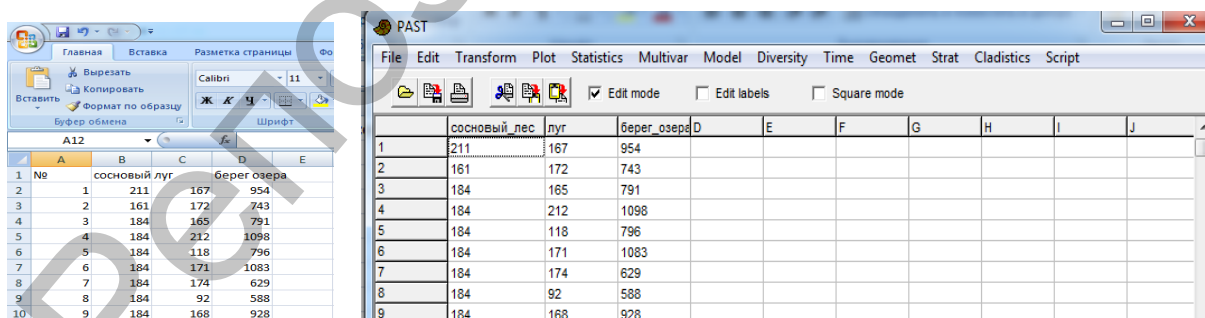


Рисунок 7.27 – Загрузка данных из MS Excel в PAST

Проверить данные на соответствие закону нормального распределения (Рисунок 7.28).

Tests for normal distribution			
	сосновый_лес	луг	берег_озера
N	9	9	9
Shapiro-Wilk W	0,6896	0,8464	0,9419
p(normal)	0,001073	0,06802	0,6023
Jarque-Bera JB	1,144	0,8403	0,5678
p(normal)	0,5644	0,657	0,7528
p(Monte Carlo)	0,1684	0,3489	0,5981
Chi²	13,667	9,2222	0,33333
p(normal)	0,0002183	0,002391	0,5637
Chi² OK (N>20)	NO	NO	NO
Anderson-Darling A	1,641	0,832	0,2348
p(normal)	0.0001172	0.01891	0.7084

Рисунок 7.28 – Проверка данных на нормальность распределения

2) Проверить данные на гомогенность дисперсии по критерию Левена. Запустить модуль One-way ANOVA: закладка Statistics/ One-way ANOVA (Рисунок 7.29).

One-way ANOVA					
<input type="checkbox"/> Repeated measures					
Between groups:	Sum of sqrs	df	Mean square	F	p(same)
	2,72343E06	2	1,36171E06	117,2	4,112E-13
Within groups:	278793	24	11616,4		
Total:	3,00222E06	26			
omega²:	0,8959				
Levene's test for homogeneity of variance, based on means: p(same) = 7,315E-06					
Based on medians: p(same) = 0,0002375					
Welch F test in the case of unequal variances: F=57,75, df=11,92, p=7,387E-07					
Residuals					
Tukey's pairwise comparisons:					
Q \ p(same)					
	сосновый_лес	луг	берег_озера		
сосновый_лес		0,8799	0,0001291		
луг	0,6835		0,0001291		
берег_озера	18,4	19,09			

Рисунок 7.29 – Проверка данных на гомогенность дисперсии по критерию Левена

Результаты тестов показали, что ряд переменных не подчиняются закону нормального распределения, как по критерию Шапиро-Уилка, так по результатам других тестов ($p < 0,05$), а критерий Левена выявил гетерогенность дисперсии ($p < 0,05$), т.е. групповые дисперсии не однородны. Следовательно, параметрический дисперсионный анализ проводить не корректно.

Выполнить непараметрический однофакторный дисперсионный анализ Крускала-Уолиса (Рисунок 7.30).

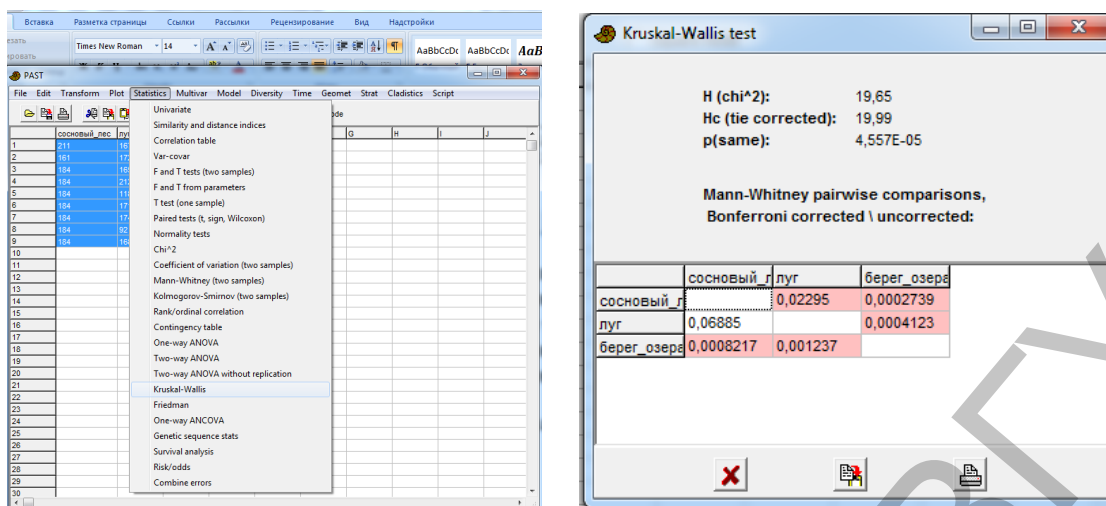


Рисунок 7.30 – Непараметрический однофакторный дисперсионный анализ Крускала-Уолиса

Результаты анализа демонстрируют достоверные различия ($p < 0,05$) между переменными. Ниже в таблице приведены результаты по парным апостериорных сравнений. Статистически значимые различия выделены красным.

Задание 5. Выполнить самостоятельно дисперсионный анализ, используя приведенные выше схемы, в пакетах Statistica и PAST для данных, предложенных преподавателем.

Контрольные вопросы:

1. Понятие зависимой и независимой переменной.
2. Понятие дисперсионного анализа.
3. Условия применения дисперсионного анализа. Разведочный анализ данных.
4. Параметрический и непараметрический дисперсионный анализ.
5. Апостериорные тесты и их значение.
6. Объяснить результаты дисперсионного анализа на предложенном примере.

ЛАБОРАТОРНАЯ РАБОТА № 8

Основы корреляционного анализа

Цель: получить практические навыки и закрепить на конкретных примерах знания о методиках проведения корреляционного анализа.

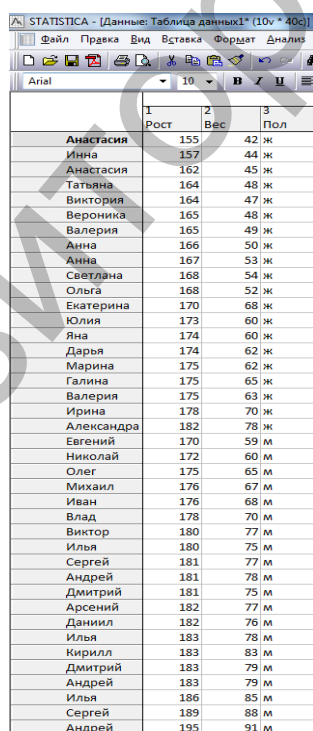
Программное обеспечение: базы данных MS Excel, пакеты анализа Statistica, PAST.

Основные термины и понятия: понятие о функциональной и корреляционной зависимости, степень и направление корреляционной зависимости, положительная и отрицательная корреляция, линейная и нелинейная корреляция, коэффициент корреляции Пирсона, коэффициент ранговой корреляции Спирмена, оценка статистической значимости коэффициентов корреляции.

Задание 1. Выполнить корреляционный анализ в пакете Statistica, используя коэффициент корреляции Пирсона. Провести анализ связи между ростом и весом студентов биологического факультета. Данные получены случайно, в результате опроса 20 девушек и 20 юношей, студентов 3–4 курса.

Нужно помнить, что для использования коэффициента корреляции Пирсона, отражающего степень связи между двумя переменными, необходимо выполнение обязательных условий, таких как нормальность распределения и наличие линейной связи между признаками. В противном случае используются коэффициенты ранговой корреляции, например коэффициент Спирмена.

- 1) Загрузить таблицу с данными из MS Excel (Рисунок 8.1).



	1 Рост	2 Вес	3 Пол
Анастасия	155	42	ж
Инна	157	44	ж
Анастасия	162	45	ж
Татьяна	164	48	ж
Виктория	164	47	ж
Вероника	165	48	ж
Валерия	165	49	ж
Анна	166	50	ж
Анна	167	53	ж
Светлана	168	54	ж
Ольга	168	52	ж
Екатерина	170	68	ж
Юлия	173	60	ж
Яна	174	60	ж
Дарья	174	62	ж
Марина	175	62	ж
Галина	175	65	ж
Валерия	175	63	ж
Ирина	178	70	ж
Александра	182	78	ж
Евгений	170	59	м
Николай	172	60	м
Олег	175	65	м
Михаил	176	67	м
Иван	176	68	м
Влад	178	70	м
Виктор	180	77	м
Илья	180	75	м
Сергей	181	77	м
Андрей	181	78	м
Дмитрий	181	75	м
Арсений	182	77	м
Даниил	182	76	м
Илья	183	78	м
Кирилл	183	83	м
Дмитрий	183	79	м
Андрей	183	79	м
Илья	186	85	м
Сергей	189	88	м
Андрей	195	91	м

Рисунок 8.1 – Загрузка данных из MS Excel в Statistica

- 2) Проверьте данные на соответствие закону нормального распределения сначала для переменной рост, затем для переменной вес, используя модуль Анализ/основные статистики и таблицы/нормальность (Рисунок 8.2).

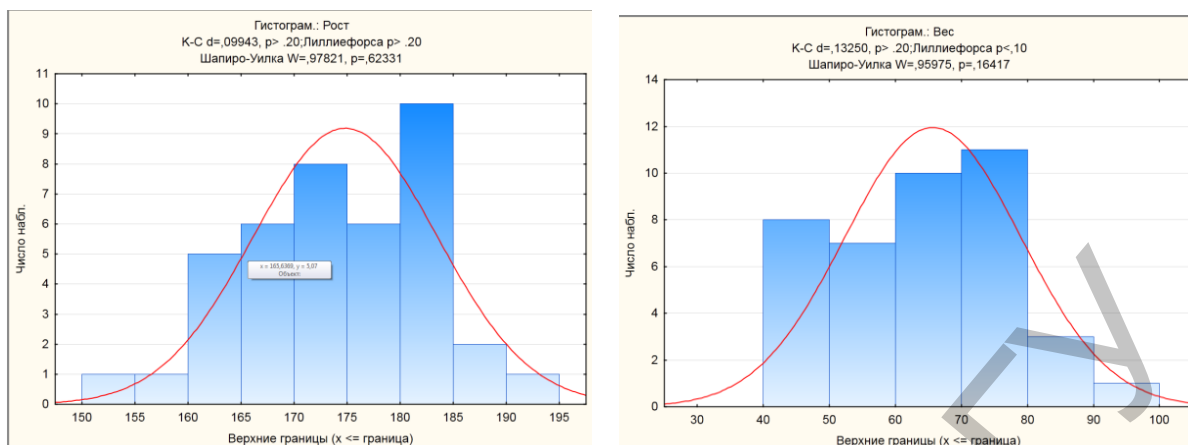


Рисунок 8.2 – Гистограммы распределения переменных роста и веса

Как видно из рисунков, графический метод не дает достаточно точную оценку, однако показатель критерия Шапиро-Уилка ($p > 0,05$) указывает на распределении близкое к нормальному. При необходимости можно применить подгонку данных, используя преобразования Бокса-Кокса или прологарифмировав данные.

3) Проверить наличие линейной связи между признаками. Для этого запустите модуль Анализ/основные статистики и таблицы/парные и частные корреляции. Выберите прямоугольные матрицы и соответствующие переменные (Рисунок 8.3).

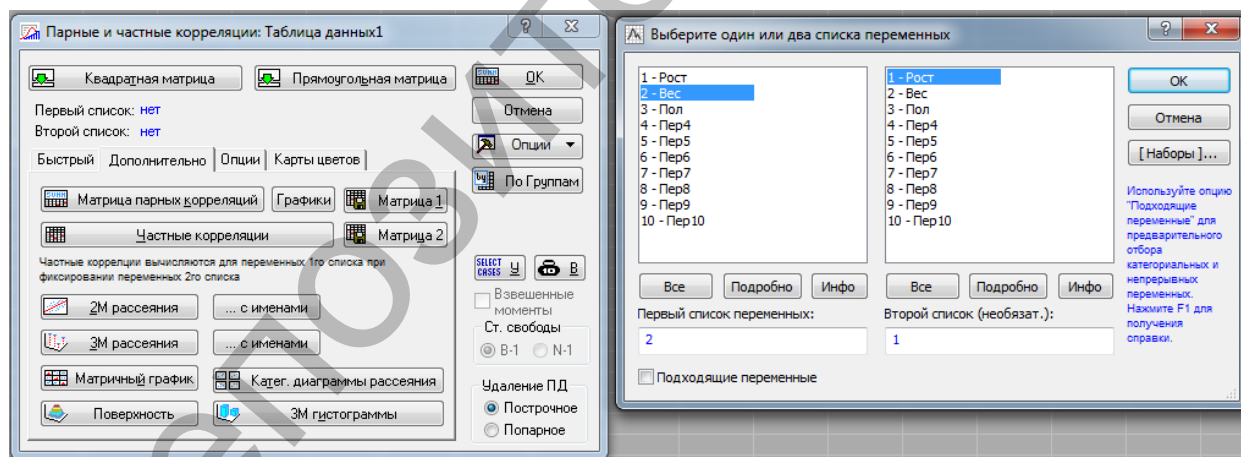


Рисунок 8.3 – Запуск модуля Анализ и выбор соответствующих переменных

Затем выберите вкладку дополнительно/графики и нажимайте на ОК (Рисунок 8.4).

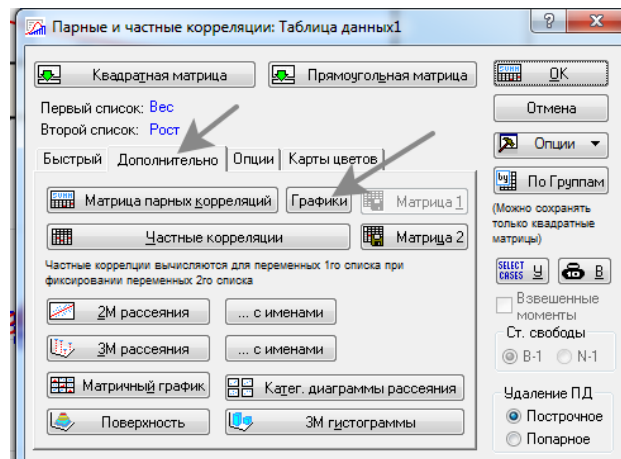


Рисунок 8.4 – Выбор опций для построения диаграммы рассеяния

Программа построит диаграмму рассеяния (Рисунок 8.5), на которой по оси x отложены значения переменной вес, а по оси y – значения переменной рост. Диагональная линия данного графика служит для оценки линейности связи. Как видно, точки соответствующих данных расположены вдоль этой линии на близком расстоянии. Поэтому можно утверждать о наличии линейной зависимости между переменными. Вместе с точечной диаграммой рассеяния программа строит гистограммы для анализируемых переменных, по которым можно проверить условие о нормальности распределения, которое, как видно из нашего примера, выполняется. Поскольку условия для применения коэффициента корреляции Пирсона выполняются.

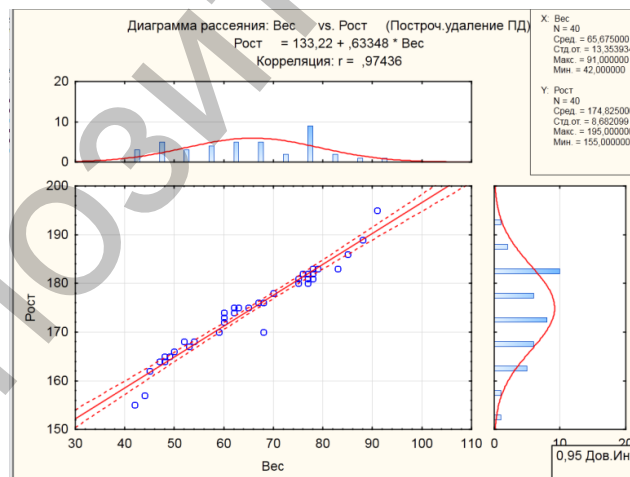


Рисунок 8.5 – Вывод гистограммы и диаграммы рассеяния

4) Рассчитать коэффициент корреляции Пирсона. Для этого продолжим анализ во вкладке Парные и частные корреляции, нажав на кнопку Матрица парных корреляций (Рисунок 8.6).

В результате появляется таблица результатов анализа из которой видно, что между переменными существует высокая ($r=0,971$) достоверная ($p<0,05$) связь.

Корреляции (Таблица данных1)	
Отмеченные корреляции значимы на уровне $p < ,05000$	
N=40 (Построчное удаление ПД)	
Переменная	Рост
Вес	0.974359

Рисунок – 8.6 Таблица с данными коэффициента корреляции Пирсона

Задание 2. Выполнить корреляционный анализ в пакете Statistica, используя коэффициент ранговой корреляции Спирмена. Проведем анализ связи между числом выявленных особей имаго стрекоз в различных биотопах и температурой воздуха во время полевых исследований.

Для использования коэффициента ранговой корреляции Спирмена, отражающего степень связи между двумя переменными, выполнение таких условий как нормальность распределения и наличие линейной связи между признаками не требуется. Расчет данного коэффициента предполагает распределение значений исследуемой переменной на отдельные группы – ранги. При этом анализируется наличие связи не между отдельными значениями, а между их рангами.

1) Загрузить таблицу с данными из MS Excel (Рисунок 8.7).

STATISTICA - [Данные: Таблица данных1* (10v * 27c)]			
Файл Правка Вид Вставка Формат Анализ Данные			
Arial 10 B I U			
	1	2	3
	биотоп	среднее числ	средняя тем
1	сосновый лес	97	17,5
2	сосновый лес	94	16
3	сосновый лес	192	18
4	сосновый лес	188	18
5	сосновый лес	190	18,5
6	сосновый лес	191	18
7	сосновый лес	82	17,3
8	сосновый лес	184	19
9	сосновый лес	190	20,7
10	берег озера	928	24
11	берег озера	629	22,3
12	берег озера	743	22,3
13	берег озера	1098	25
14	берег озера	796	24
15	берег озера	588	22
16	берег озера	954	24,6
17	берег озера	1083	24,3
18	берег озера	791	23,8
19	луг	192	19
20	луг	168	20
21	луг	171	19,5
22	луг	172	21,5
23	луг	219	21
24	луг	215	22
25	луг	292	21
26	луг	167	20,5
27	луг	321	23,6

Рисунок 8.7 – Загрузка данных из MS Excel в Statistica

2) Для того, чтобы убедиться, что использование корреляции Пирсона не корректно, проверьте данные на соответствие закону нормального распределения сначала для переменной рост, затем для переменной вес, используя модуль Анализ/основные статистики и таблицы/нормальность (Рисунок 8.8).

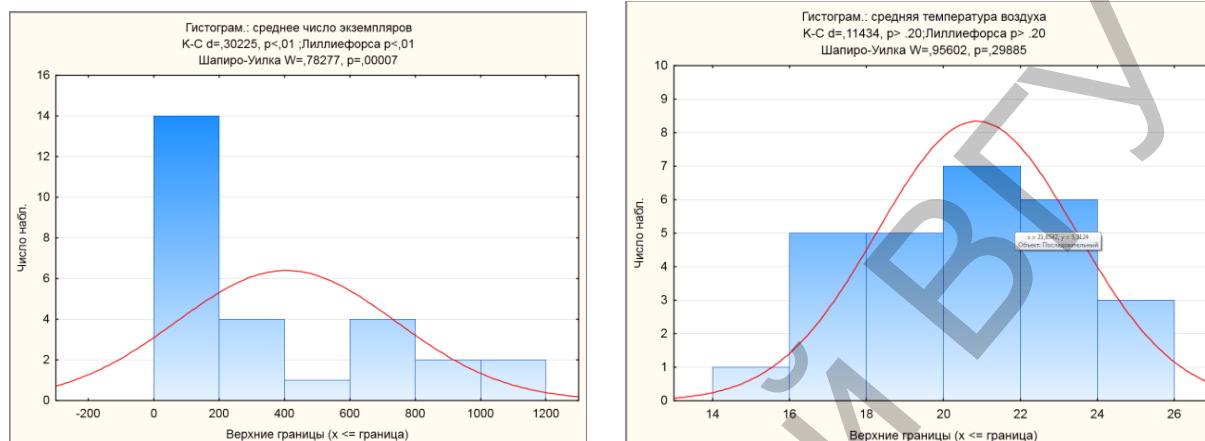


Рисунок 8.8 – Гистограммы распределения переменных роста и веса

3) Как видно из рисунков, для переменной среднее число экземпляров условие нормальности распределения не выполняется. Это подтверждает и показатель критерия Шапиро-Уилка ($p < 0,05$). Поэтому требуется применение коэффициента ранговой корреляции Спирмена.

4) Рассчитать коэффициент ранговой корреляции Спирмена. Выбираем модуль Анализ/непараметрическая статистика/Корреляция Спирмена (Рисунок 8.9).

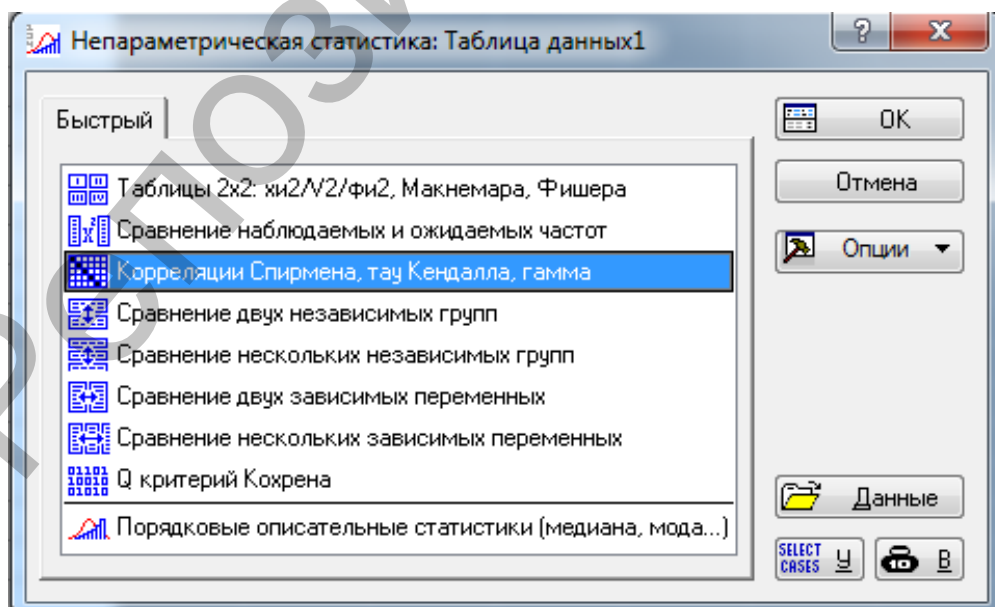


Рисунок 8.9 – Выбор коэффициента ранговой корреляции Спирмена

Далее выбираем матрицу двух списков и анализируемые переменные и нажимаем на кнопку Спирмена R (Рисунок 8.10).

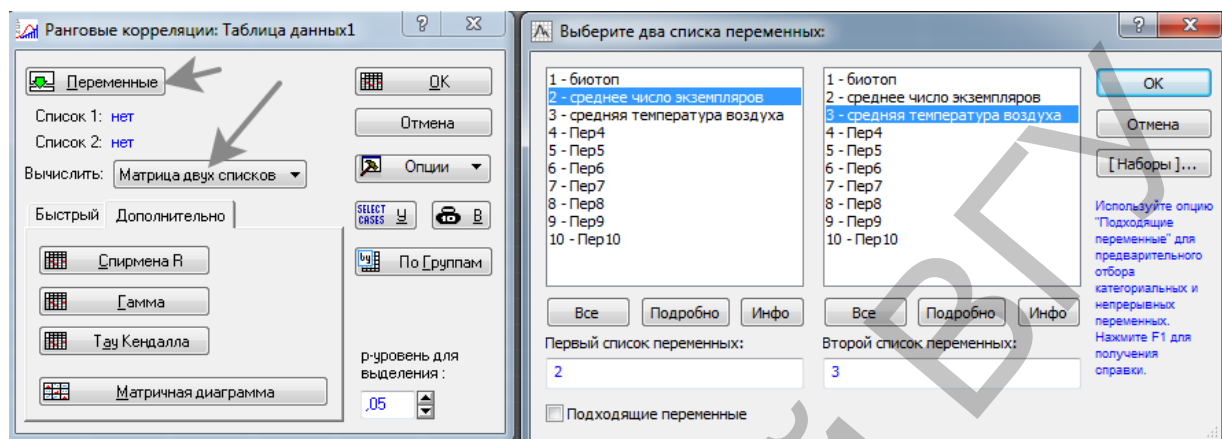


Рисунок 8.10 – Выбор матрицы двух списков и анализируемых переменных

В результате появляется таблица результатов анализа (Рисунок 8.11), из которой видно, что между переменными существует высокая ($rs=0,874$) достоверная ($p<0,05$) связь.

Ранговые корреляции Спирмена (Таблица данных1) ПД попарно удалены Отмеченные корреляции значимы на уровне $p < ,05000$	
Перем.	средняя температура воздуха
среднее число экземпляров	0,874332

Рисунок 8.11 – Таблица результата анализа ранговой корреляции Спирмена

Задание 3. Выполнить корреляционный анализ в пакете PAST, используя коэффициент корреляции Пирсона. Провести анализ связи между ростом и весом студентов биологического факультета.

- 1) Загрузить таблицу с данными из MS Excel (Рисунок 8.12).

	Рост	Вес	Пол	D	E	F	G
Анастасия	155	42	ж				
Ирина	157	44	ж				
Анастасия	162	45	ж				
Татьяна	164	46	ж				
Виктория	164	47	ж				
Вероника	165	48	ж				
Валерия	165	49	ж				
Анна	166	50	ж				
Анна	167	53	ж				
Светлана	168	54	ж				
Ольга	168	52	ж				
Екатерина	170	68	ж				
Юлия	173	60	ж				
Яна	174	60	ж				
Дарья	174	62	ж				
Марина	175	62	ж				
Галина	175	65	ж				
Валерия	175	63	ж				
Ирина	178	70	ж				
Александра	182	78	ж				
Евгений	170	59	м				
Николай	172	60	м				
Олег	175	65	м				
Михаил	176	67	м				
Иван	176	68	м				
Влад	178	70	м				
Виктор	180	77	м				
Илья	180	75	м				
Сергей	181	77	м				
Андрей	181	78	м				
Дмитрий	181	75	м				
Арсений	182	77	м				
Даниил	182	76	м				
Илья	183	78	м				
Кирилл	183	83	м				
Дмитрий	183	79	м				
Андрей	183	79	м				
Илья	186	85	м				
Сергей	189	88	м				
Андрей	195	91	м				

Рисунок 8.12 – Загрузка данных из MS Excel в PAST

2) Выполнить проверку на соответствие данных закону нормально-го распределения, выбрав вкладку Statistics/Normality tests (Рисунок 8.13).

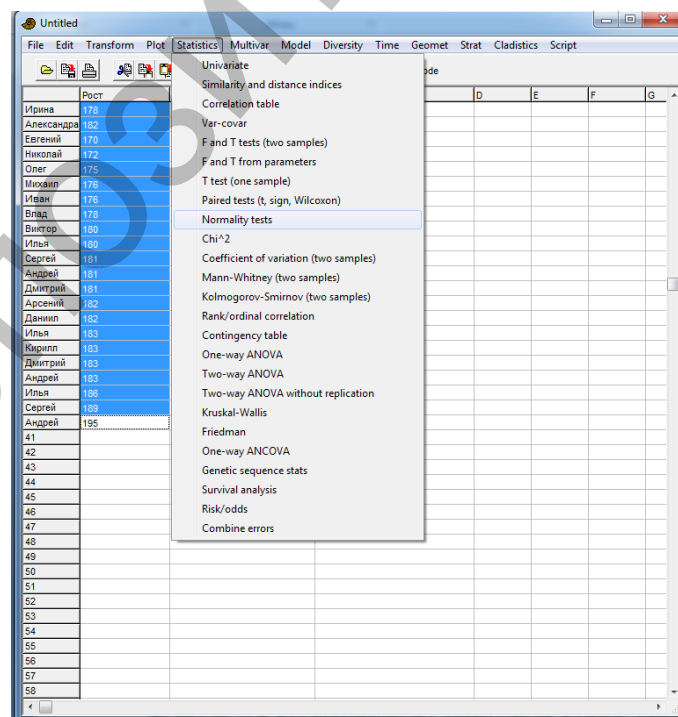


Рисунок 8.13 – Выбор модуля Normality tests во вкладке Statistics

Как видно из таблиц (Рисунок 8.14), для обоих переменных условие нормальности распределения выполняется, что подтверждает и показатель критерия Шапиро-Уилка ($p > 0,05$).

	Рост
N	40
Shapiro-Wilk W	0,9782
p(normal)	0,6233
Jarque-Bera JB	0,3821
p(normal)	0,8261
p(Monte Carlo)	0,805
Chi^2	2,6
p(normal)	0,10686
Chi^2 OK (N>20)	YES
Anderson-Darling A	0,4102
p(normal)	0,328

	Вес
N	40
Shapiro-Wilk W	0,9597
p(normal)	0,1642
Jarque-Bera JB	1,897
p(normal)	0,3872
p(Monte Carlo)	0,2248
Chi^2	5,2
p(normal)	0,022587
Chi^2 OK (N>20)	YES
Anderson-Darling A	0,5488
p(normal)	0,148

Рисунок 8.14 – Результаты анализа по критерию Шапиро-Уилка

3) Рассчитать коэффициент корреляции Пирсона, используя вкладку Statistics/Correlation table. При этом в нижнем левом углу нужно выставить метку Linear correlation (Рисунок 8.15).

4)

	Рост	Вес
Рост		3,1835E-26
Вес	0,97436	

Correlation statistic

- ☒ Linear correlation r
- ☐ Spearman's D
- ☐ Spearman's rs
- ☐ Kendall's tau
- ☐ Partial linear correlation

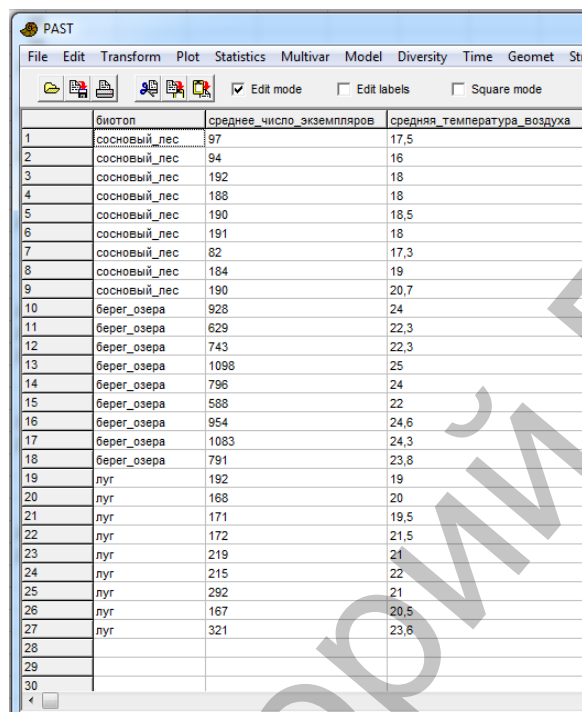
Рисунок 8.15 - Расчет коэффициента корреляции Пирсона

Из таблицы результатов анализа (Рисунок 8.15) видно, что между переменными существует высокая ($r=0,974$) достоверная ($p < 0,05$) связь.

Задание 4. Выполнить корреляционный анализ в пакете PAST, используя коэффициент ранговой корреляции Спирмена. Провести

анализ связи между числом выявленных особей имаго стрекоз в различных биотопах и температурой воздуха во время полевых исследований.

- 1) Загрузить таблицу с данными из MS Excel (Рисунок 8.16).



	биотоп	среднее число экземпляров	средняя температура воздуха
1	сосновый_лес	97	17,5
2	сосновый_лес	94	16
3	сосновый_лес	192	18
4	сосновый_лес	188	18
5	сосновый_лес	190	18,5
6	сосновый_лес	191	18
7	сосновый_лес	82	17,3
8	сосновый_лес	184	19
9	сосновый_лес	190	20,7
10	берег_озера	928	24
11	берег_озера	629	22,3
12	берег_озера	743	22,3
13	берег_озера	1098	25
14	берег_озера	796	24
15	берег_озера	588	22
16	берег_озера	954	24,6
17	берег_озера	1083	24,3
18	берег_озера	791	23,8
19	луг	192	19
20	луг	168	20
21	луг	171	19,5
22	луг	172	21,5
23	луг	219	21
24	луг	215	22
25	луг	292	21
26	луг	167	20,5
27	луг	321	23,6
28			
29			
30			

Рисунок 8.16 – Загрузка данных из MS Excel в PAST

- 2) Выполнить проверку на соответствие данных закону нормально-го распределения, выбрав вкладку Statistics/Normality tests (Рисунок 8.17).

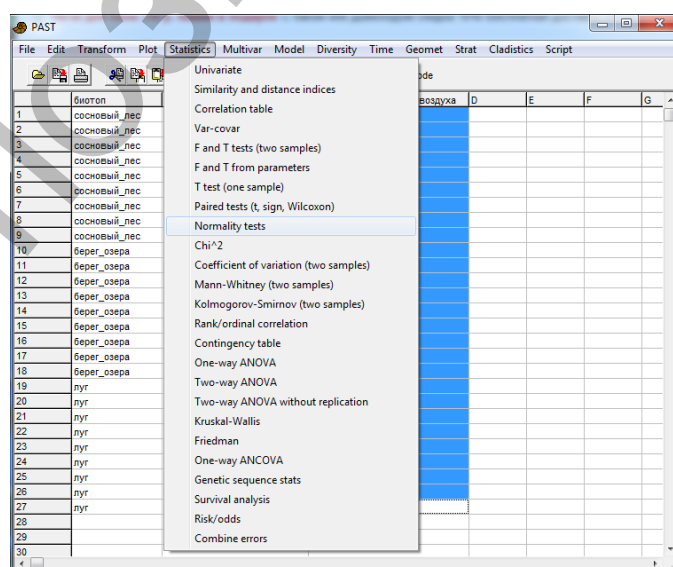


Рисунок 8.17 - Выбор модуля Normality tests во вкладке Statistics

Как видно из таблиц (Рисунок 8.18), для переменной «среднее число экземпляров» условие нормальности распределения не выполняется: показатель критерия Шапиро-Уилка ($p < 0,05$).

Test	Value
N	27
Shapiro-Wilk W	0,7828
p(normal)	6,984E-05
Jarque-Bera JB	4,412
p(normal)	0,1102
p(Monte Carlo)	0,0469
Chi²	6,037
p(normal)	0,014009
Chi² OK (N>20)	YES
Anderson-Darling A	2,642
p(normal)	7,415E-07

Test	Value
N	27
Shapiro-Wilk W	0,956
p(normal)	0,2989
Jarque-Bera JB	1,466
p(normal)	0,4806
p(Monte Carlo)	0,2716
Chi²	1,8889
p(normal)	0,16933
Chi² OK (N>20)	YES
Anderson-Darling A	0,3835
p(normal)	0,3717

Рисунок 8.18 – Результаты анализа нормальности распределения для переменной «среднее число экземпляров» по критерию Шапиро-Уилка

3) Выполнить корреляционный анализ, используя вкладку Rank/ordinal correlation (Рисунок 8.19).

Variable 1	Variable 2	rs (uncorr)	Permutation p:
среднее число экземпляров	средняя температура воздуха	0,8743	0,0001

Рисунок 8.19 – Результаты корреляционного анализа

Как видно из таблицы результатов анализа (Рисунок 8.19), между переменными существует высокая ($r_s = 0,8743$) достоверная ($p < 0,05$) связь.

Задание 5. Выполнить самостоятельно корреляционный анализ, используя приведенные выше схемы, в пакетах Statistica и PAST для данных, предложенных преподавателем.

Контрольные вопросы:

1. Понятие корреляционного анализа. Виды корреляций.
2. Отличия корреляционного и дисперсионного анализов.
3. Условия применения корреляционного анализа с использованием коэффициента Пирсона.
4. Непараметрический корреляционный анализ. Понятие ранговой корреляции.
5. Объяснить результаты корреляционного анализа на предложенном примере.

ЛАБОРАТОРНАЯ РАБОТА № 9

Основы регрессионного анализа

Цель: получить практические навыки и закрепить на конкретных примерах знания о методиках проведения регрессионного анализа.

Программное обеспечение: базы данных MS Excel, пакет анализа Statistica.

Основные термины и понятия: зависимая и независимая переменные (предиктор); уравнение регрессионного анализа; нулевая гипотеза при регрессионном анализе; коэффициент детерминации; понятие о нелинейной и множественной регрессионной зависимости; разведочный анализ: проверка на нормальность распределения (визуальный анализ гистограммы распределений и тесты Колмогорова-Смирнова, Шапиро-Уилка), выявление линейной или нелинейной зависимости, проверка нормальности распределения остатков, оценка величины остаточной дисперсии.

Задание 1. Выполнить регрессионный анализ в пакете Statistica. Проанализировать зависимость величины систолического артериального давления от возраста человека.

Прежде чем приступить к анализу, нужно помнить, что обе переменные должны подчиняться закону нормального распределения и зависимость между ними должна носить линейный характер.

- 1) Загрузить таблицу с данными из MS Excel (Рисунок 9.1).

STATISTICA - [Данные: Таблица данн			
Файл Правка Вид Вставка Ф			
Arial 10			
	1	2	
	возраст,	давлени	П
1	30	110	
2	30	106	
3	40	120	
4	40	118	
5	40	125	
6	50	135	
7	50	133	
8	50	134	
9	60	150	
10	60	148	
11	60	151	
12	60	152	
13	70	164	
14	70	160	
15	70	162	
16	70	161	

Рисунок 9.1 – Загрузка данных из MS Excel в Statistica

2) Выполнить проверку данных на соответствие закону нормального распределения с помощью модуля Описательные статистики.

Как видно по показателю критерия Шапиро-Уилка ($p > 0,05$), в обоих случаях данные распределены нормально (Рисунок 9.2).

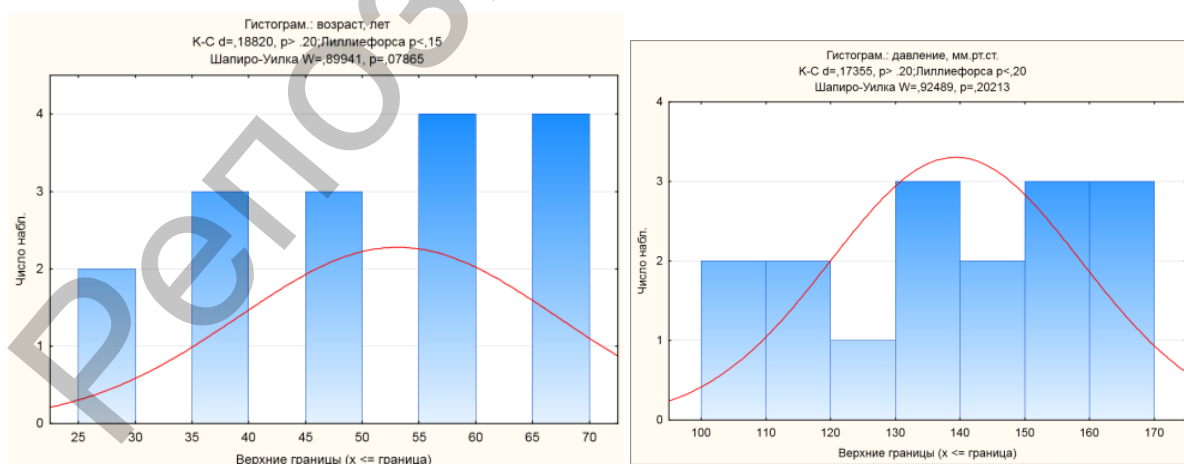


Рисунок 9.2 – Гистограммы распределения исследуемых показателей

3) Запустить модуль Анализ/Множественная регрессия и выбрать переменные (Рисунок 9.3).

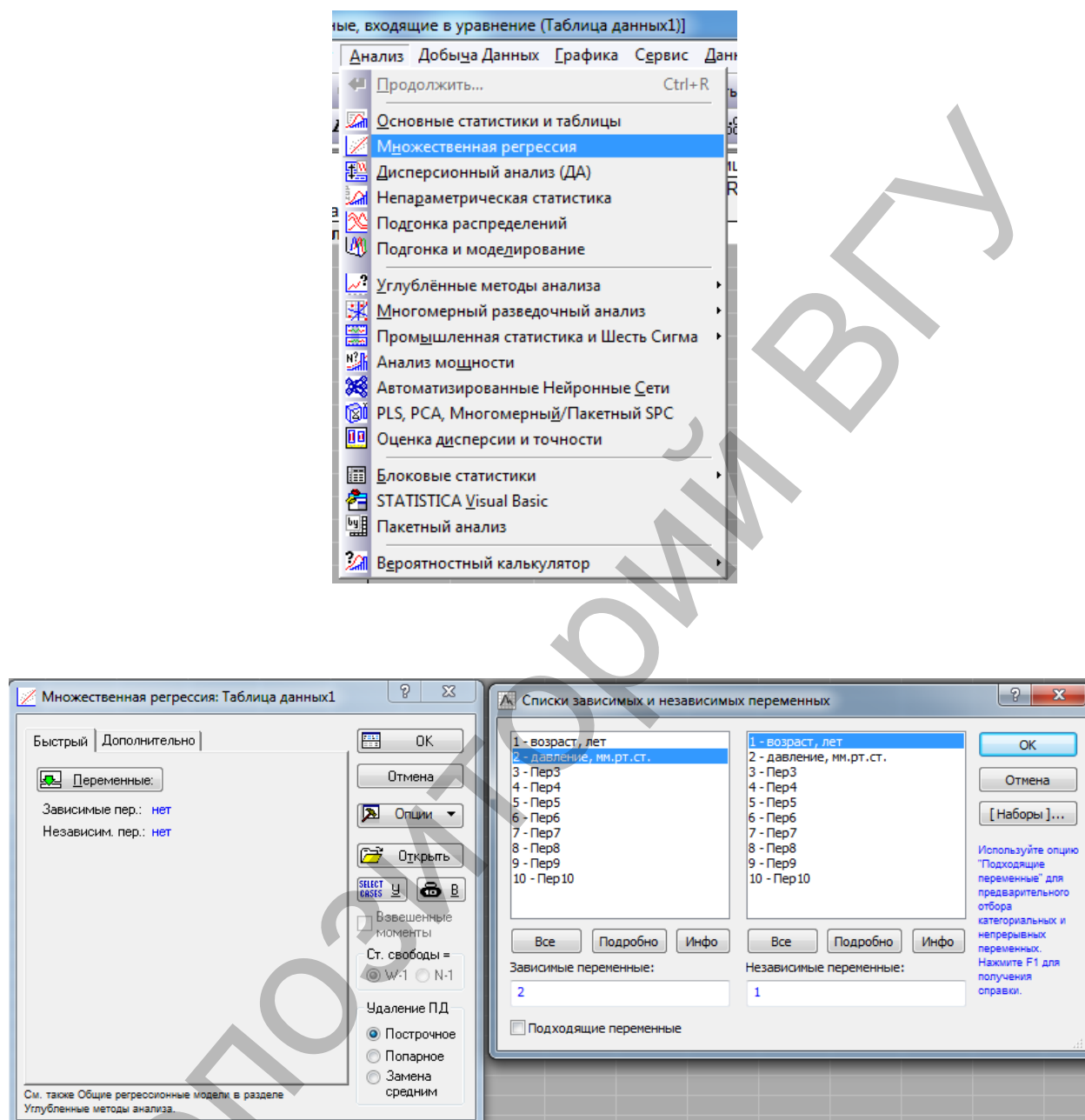


Рисунок 9.3 – Выбор переменных для проведения множественной регрессии

Нажать на ОК. Но прежде чем начать анализировать результаты регрессионного анализа, нужно убедиться, что зависимость между признаками носит линейный характер. Для этого выбираем вкладку Остатки/предсказанные /наблюдаемые значения и нажимаем на кнопку Описательные статистики (Рисунок 9.4).

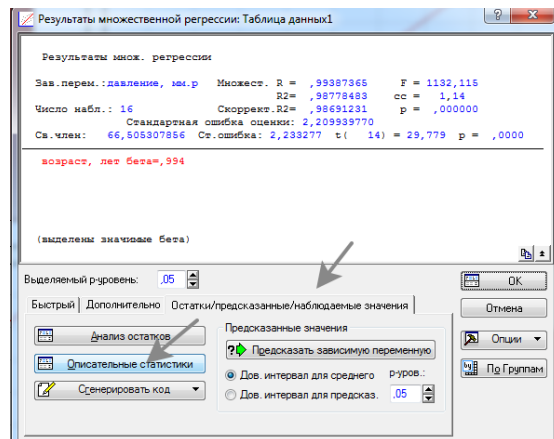


Рисунок 9.4 – Оценка характера зависимости между признаками

В появившемся окне выбираем Матричный график, после чего выбрав переменные, нажимаем на ОК (Рисунок 9.5).

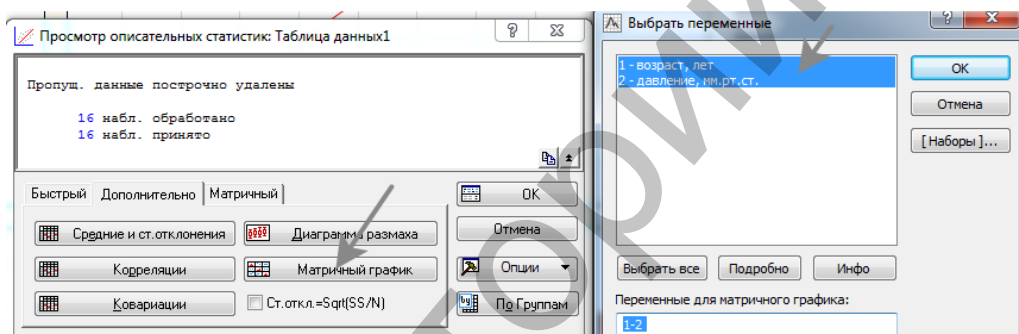


Рисунок 9.5 – Выбор опции «Матричный график» и анализируемых Переменных

Как видно из графиков (Рисунок 9.6), переменные имеют линейную зависимость, то есть данное условие выполняется.

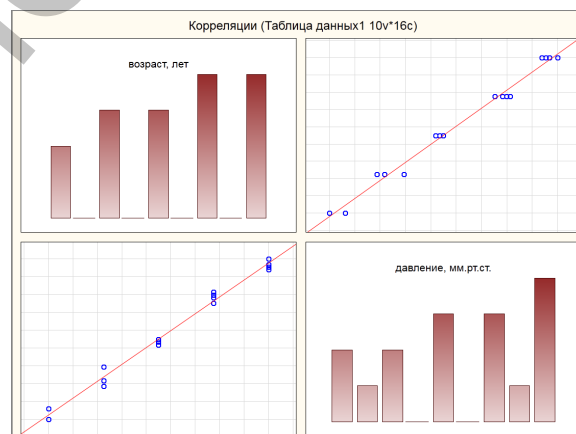


Рисунок 9.6 – Графики линейной зависимости исследуемых переменных

4) Выполнить регрессионный анализ, вернувшись в модуль Анализ/множественная регрессия/текущий анализ (Рисунок 9.7).

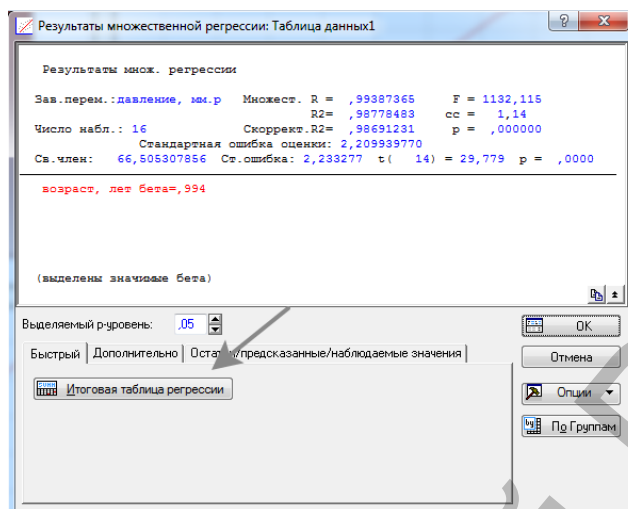


Рисунок 9.7 – Выбор опций для выполнения регрессионного анализа

Исходя из результатов регрессионного анализа (Рисунок 9.8), можно утверждать, что между анализируемыми переменными существует достоверная связь ($p < 0,05$). Значение коэффициента детерминации указывает ($R^2 = 0,986$) свидетельствует высокой точности данной регрессионной модели и она хорошо описывает связь между двумя переменными.

Итоги регрессии для зависимой переменной: давление, мм.рт.ст. (Таблица данных1)						
R= .99387365 R²= .98778483 Скоррект. R²= .98691231						
F(1,14)=1132,1 p<.00000 Станд. ошибка оценки: 2,2099						
N=16	БЕТА	Ст.Ош. БЕТА	B	Ст.Ош. B	t(14)	p-знач.
Св.член			66,50531	2,233277	29,77925	0,000000
возраст, лет	0,993874	0,029538	1,37049	0,040731	33,64692	0,000000

Рисунок 9.8 – Результаты регрессионного анализа

5) Проанализировать регрессионное уравнение.

Уравнение имеет следующий вид: $y = a + bx$, где y – зависимая переменная, x – независимая переменная, a – свободный член (Intercept), b – коэффициент регрессии.

В нашем случае уравнение будет выглядеть так: $АД = 66,50531 + 1,37049 * В$

где, АД (зависимая переменная) – систолическое артериальное давление, В – возраст (независимая переменная), 66,50531 – свободный член, 1,37049 – коэффициент регрессии.

Подставив в уравнение определенный возраст, можно предсказать показатели артериального давления.

6) Проверить остатки на нормальность распределения. Важным компонентом регрессионного анализа в нашем случае является нормальность распределения остатков. Остатки представляют собой разность между наблюдаемыми значениями зависимой переменной и другими, предсказанными значениями. Чем лучше регрессионная модель согласуется с анализируемыми переменными, тем меньше величина остатков и тем меньше влияние других, неучтенных переменных. Выполнить проверку можно, используя: Анализ остатков/ Нормальный график остатков (Рисунок 9.9).

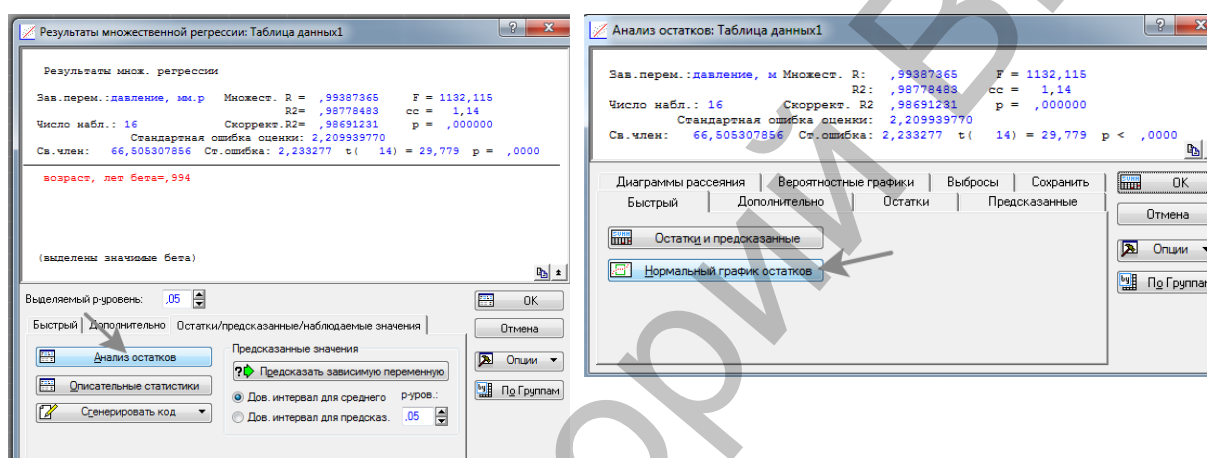


Рисунок 9.9 – Проверка остатков на нормальность распределения

Как видно из рисунка (Рисунок 9.10), точки достаточно тесно выстраиваются вдоль воображаемой прямой, что позволяет предположить о нормальности распределения.

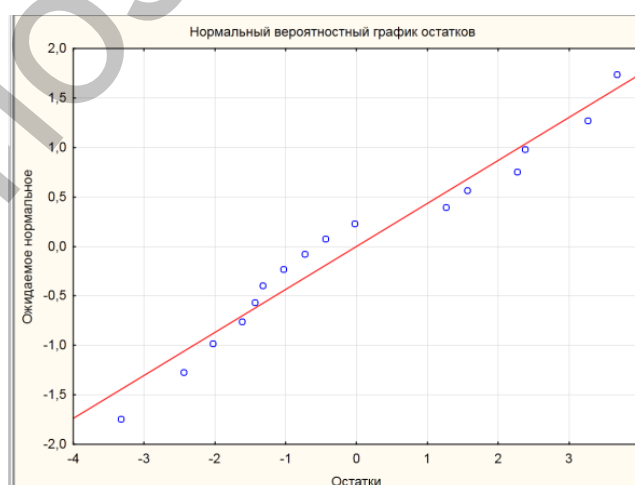


Рисунок 9.10 – Графический анализ нормальности распределения

Проверить гомогенность дисперсии остатков. Гомогенность (неизменность) остатков еще одно необходимое условие регрессионного анализа. Выполнить проверку можно с помощью диаграммы рассеяния (Рисунок 9.11).

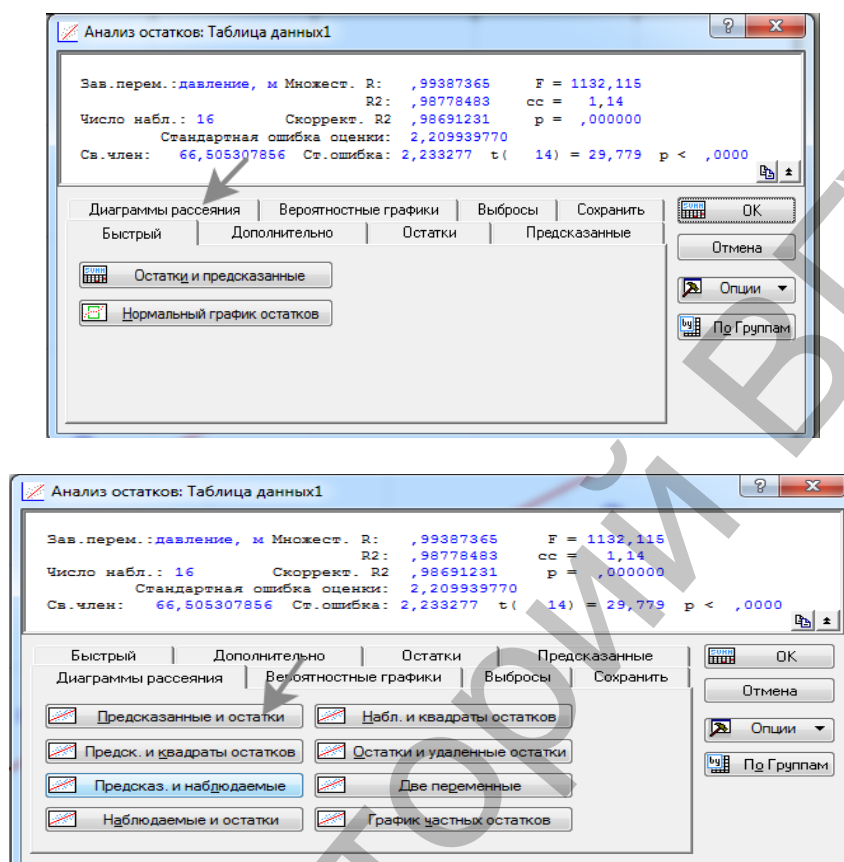


Рисунок 9.11 – Выбор опций для построения диаграммы рассеяния

Точки на данном графике должны быть разбросаны хаотично, то есть без всякой закономерности, что и наблюдается в нашем случае.

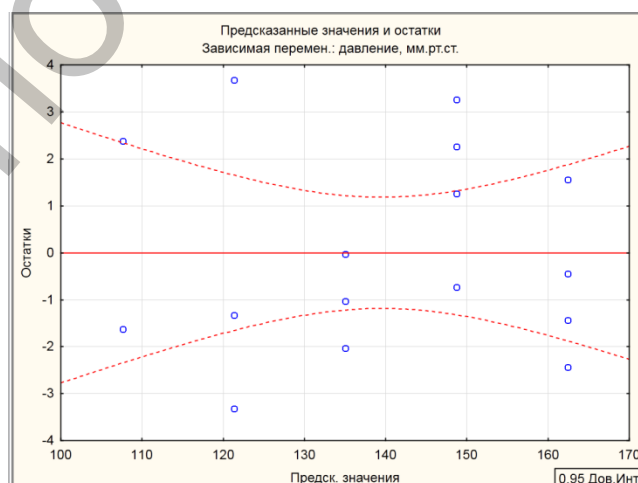


Рисунок 9.11 Диаграмма рассеяния для проверки гомогенности дисперсии остатков

Контрольные вопросы:

1. Понятие регрессионного анализа.
2. Условия применения регрессионного анализа. Разведочный анализ данных.
3. Объяснить результаты регрессионного анализа на предложенном примере.
4. Охарактеризовать уравнение регрессии.
1. Понятие остатков и их оценка.
2. Какие типы дисперсионного анализа применяются при отсутствии нормального распределения и (или) линейной зависимости между переменными.

ЛАБОРАТОРНАЯ РАБОТА № 10

Элементы многомерной статистики (многофакторный анализ)

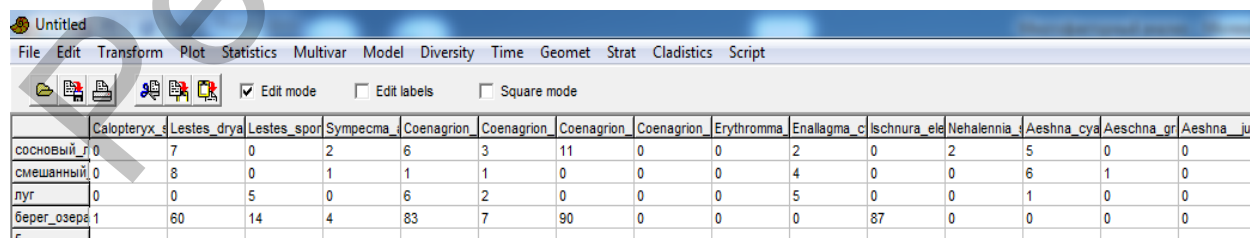
Цель: получить практические навыки и закрепить на конкретных примерах знания о методиках проведения многомерного анализа данных.

Программное обеспечение: базы данных MS Excel, пакет анализа Past.

Основные термины и понятия: многомерная совокупность и многомерное пространство; ординация и ординационные диаграммы; кластерный анализ, правила объединения объектов в кластеры и меры расстояния, графическое изображение результатов кластерного анализа; понятие дискриминантного анализа; прямой и не прямой градиентный анализ, анализ главных компонент, кумулятивная объясненная дисперсия; многомерное шкалирование.

Задание 1. Выполнить кластерный анализ в пакете Past. Исследовать сходство комплексов стрекоз различных биотопов по количественным данным.

- 1) Загрузить таблицу с данными из MS Excel (Рисунок 10.1).



	Calopteryx_s	Lestes_drya	Lestes_spor	Sympetma_	Coenagrion	Coenagrion	Coenagrion	Coenagrion	Erythromma	Enallagma_c	Ischnura_ele	Nehalennia	Aeshna_cya	Aeshna_gr	Aeshna_jur
сосновый_г	0	7	0	2	6	3	11	0	0	2	0	2	5	0	0
смешанный	0	8	0	1	1	1	0	0	0	4	0	0	6	1	0
луг	0	0	5	0	6	2	0	0	0	5	0	0	1	0	0
берег_озера	1	60	14	4	83	7	90	0	0	0	87	0	0	0	0

Рисунок 10.1 – Загрузка данных из MS Excel в Past

2) Запустить модуль Multivar/Cluster analysis (Рисунок 10.2).

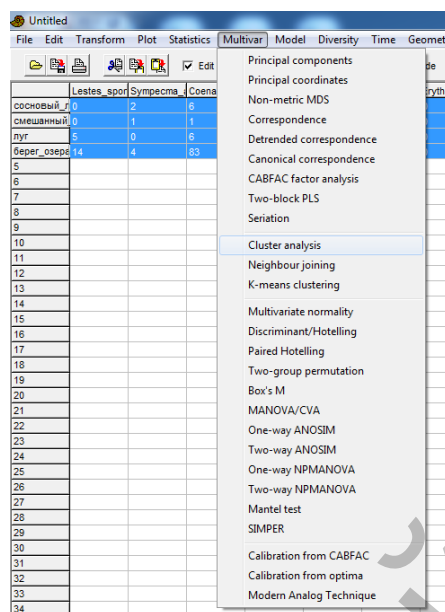


Рисунок 10.2 – Запуск модуля Multivar для кластерного анализа

Нажать на ОК. В появившемся окне будет видна кластерная дендрограмма (Рисунок 10.3), построенная по параметрам автоматически заданным в программе. Пользователь может сам выбирать параметры, такие как алгоритм (Paired group – по парного сравнения, Single linkage – одиночной связи, Wards method – метод минимизации внутригрупповой дисперсии Уорда) а также меру сходства (Similarity measure). Выберите алгоритм Single linkage и меру сходства Брэя-Кертиса.

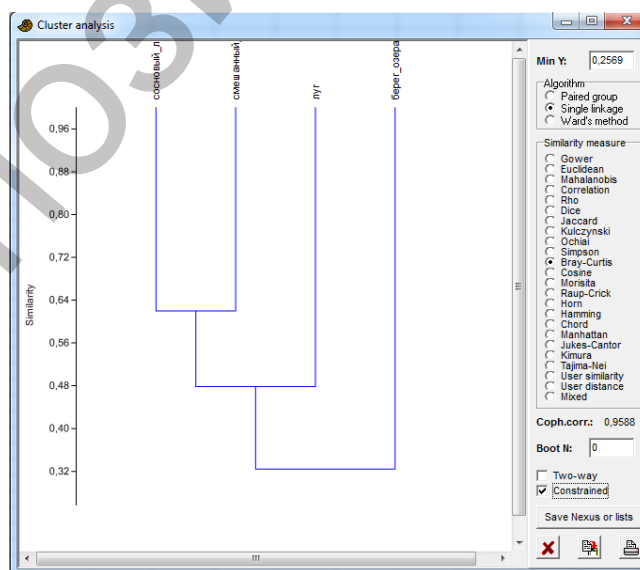
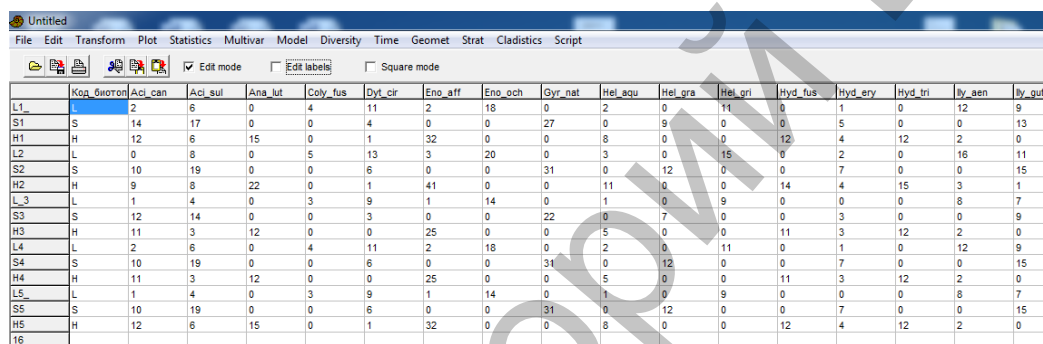


Рисунок 10.3 – Полученная кластерная дендрограмма

Как видно из дендрограммы, наибольшим сходством обладают комплексы стрекоз соснового и смешанного лесов, а наиболее отличаются от них группировки стрекоз, зарегистрированные по берегу озера.

Задание 2. Выполнить анализ главных компонент (PCA – Principal Component Analysis) в пакете Past.

1) Загрузить таблицу с данными из MS Excel (Рисунок 10.4). Предварительно, для лучшего графического представления результатов анализа, нужно ввести аббревиатуры названий анализируемых параметров в таблице данных. Например, латинские названия видов сокращают, как правило, до 6 букв (три первые буквы названия вида и три первые буквы названия рода). Также во многих случаях, с целью снижения так называемого «шума» удаляют некоторые второстепенные значения, например, виды, представленные в сборах менее, чем 5 экземплярами и т.д.



	Код биотоп	Асi_сan	Асi_суl	Ана_лut	Сол_fus	Дыт_сir	Ено_аff	Ено_сch	Огр_нат	Нел_аqu	Нел_гра	Нел_гри	Ныд_fus	Ныд_ery	Ныд_три	Ны_аen	Ны_gut
L1	L	2	6	0	4	11	2	18	0	2	0	11	0	1	0	12	9
S1	S	14	17	0	0	4	0	0	27	0	9	0	0	5	0	0	13
H1	H	12	6	15	0	1	32	0	0	8	0	0	12	4	12	2	0
L2	L	0	8	0	5	13	3	20	0	3	0	15	0	2	0	16	11
S2	S	10	19	0	0	6	0	0	31	0	12	0	0	7	0	0	15
H2	H	9	8	22	0	1	41	0	0	11	0	0	14	4	15	3	1
L3	L	1	4	0	3	9	1	14	0	1	0	9	0	0	0	8	7
S3	S	12	14	0	0	3	0	0	22	0	0	0	0	3	0	0	9
H3	H	11	3	12	0	0	25	0	0	5	0	0	11	3	12	2	0
L4	L	2	6	0	4	11	2	18	0	2	0	11	0	1	0	12	9
S4	S	10	19	0	0	6	0	0	31	0	12	0	0	7	0	0	15
H4	H	11	3	12	0	0	25	0	0	5	0	0	11	3	12	2	0
L5	L	1	4	0	3	9	1	14	0	1	0	9	0	0	0	8	7
S5	S	10	19	0	0	6	0	0	31	0	12	0	0	7	0	0	15
H5	H	12	6	15	0	1	32	0	0	8	0	0	12	4	12	2	0
L6																	

Рисунок 10.4 – Загрузка данных из MS Excel в Past

2) Выполнить анализ, используя модуль Multivar/Principal Components (Рисунок 10.5).

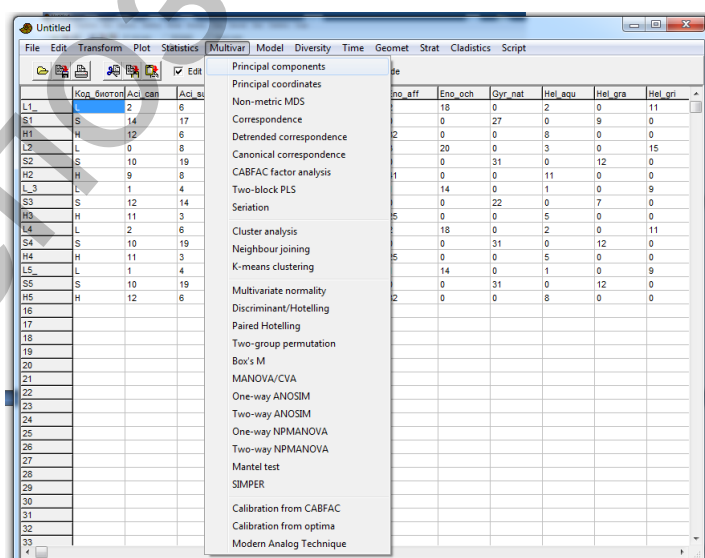


Рисунок 10.5 – Выбор вкладки Principal Components в модуле Multivar

В открывшемся окне будет таблица, которая показывает значения факторных нагрузок и процент их дисперсии на оси (главные компоненты – PC) (Рисунок 10.6). Наибольшие показатели дисперсии, в нашем случае, приходится на первую (61,542 %) и вторую компоненты (35,547 %).

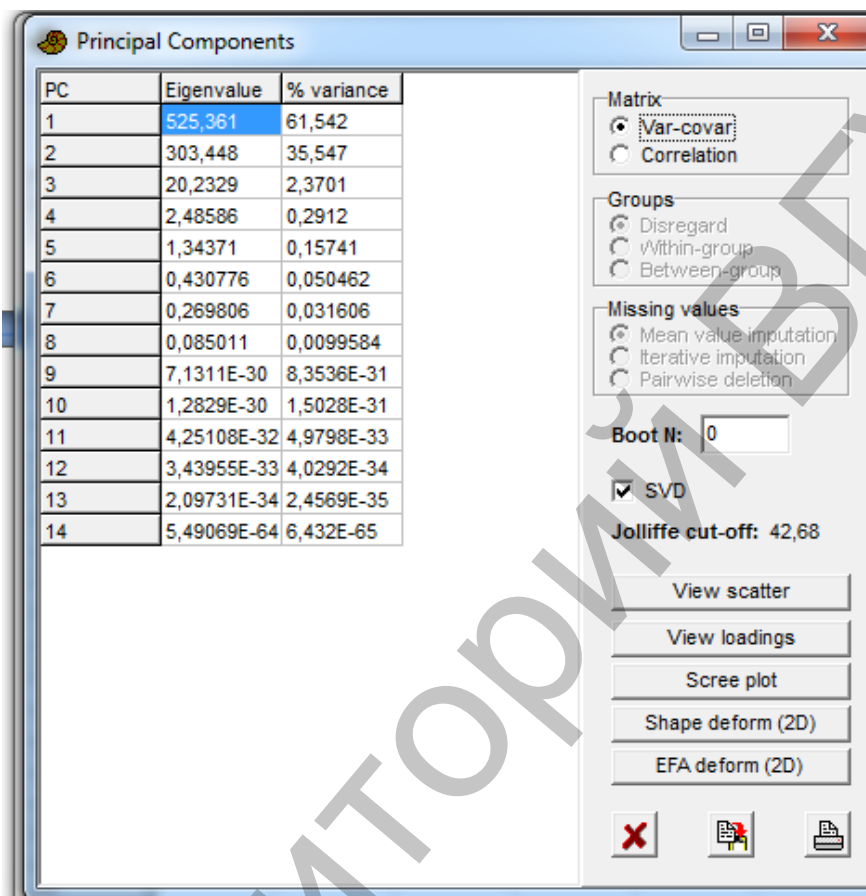


Рисунок 10.6 – Таблица значений факторных нагрузок и процент их дисперсии на оси

Поскольку анализ главных компонент относится к группе не прямых градиентных анализов, мы можем только предположить, что есть два основных фактора в наибольшей степени влияющих на биотопическое распределение анализируемых видов, а остальные факторы оказывают меньшее влияние. Их отражают остальные компоненты (с 3 по 14).

3) Построить ординационную диаграмму. Выберите в открытом окне анализа View scatter. Затем справа обозначьте тип диаграммы (Biplot) и подписи данных (Row labels) (Рисунок 10.7). Изменить параметры диаграммы можно с помощью контекстного меню, выбрав тип и размеры шрифта, введя подписи осей и др., а также формат для сохранения в виде графического файла (Рисунок 10.8).

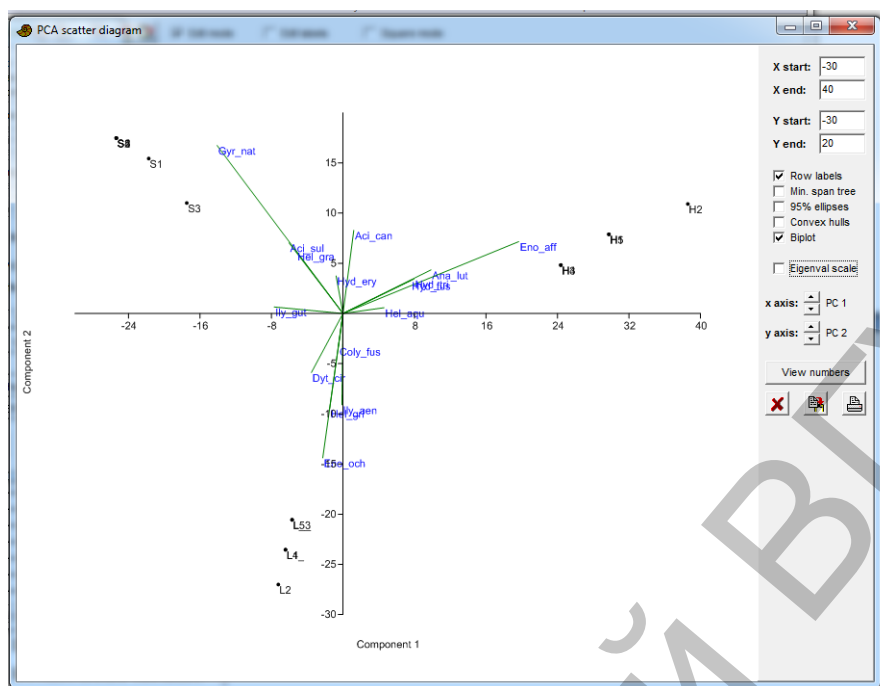


Рисунок 10.7 – Ординационная диаграмма

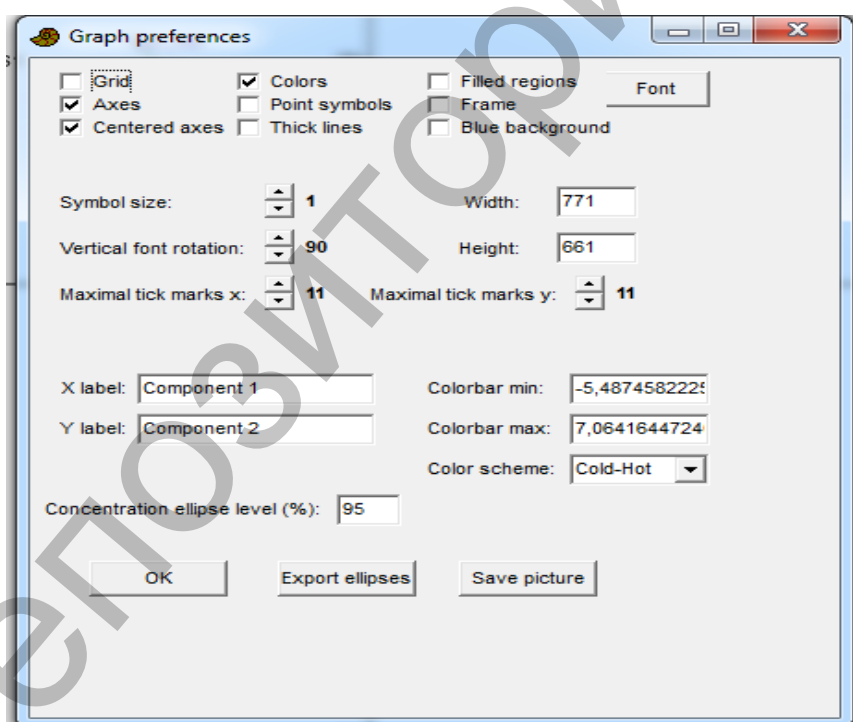


Рисунок 10.8 – Изменение параметров диаграммы

Анализ ординационной диаграммы второй важный этап, позволяющий предположить, какие факторы оказывают влияние на анализируемые переменные. В предложенном примере координаты одних видов в большей степени расположены вдоль оси X (первая главная компонента), а координаты других видов в большей степени расположены вдоль оси Y (вторая

главная компонента). Зная экологические преферендумы данных видов, исследователь предполагает возможные причины такой группировки. В тоже время это требует высокой профессиональной подготовки. Для исследовательской студенческой работы наибольшую ценность будет представлять ординация (упорядочивание) одних объектов по отношению к другим. В частности из построенной диаграммы можно увидеть, какие виды наиболее ассоциированы с определенными местообитаниями. Виды здесь представлены векторами, а биотопы точками. Чем ближе расположен и длиннее вектор, тем больше связь.

Какие конкретно факторы имеют ту или иную связь с анализируемыми переменными, можно выяснить, применив прямой градиентный анализ, в частности многомерное шкалирование или канонический анализ соответствий, введя дополнительно вторую таблицу данных с измерениями, например, абиотических факторов.

Задание 3. Выполнить самостоятельно кластерный анализ, используя приведенные выше схемы, в пакете Past для данных, предложенных преподавателем.

Задание 4. Выполнить самостоятельно анализ главных компонент, используя приведенные выше схемы, в пакете Past для данных, предложенных преподавателем.

Контрольные вопросы:

- 1. Понятие многомерного анализа и ординации.*
- 2. Прямой и не прямой градиентный анализы.*
- 3. Кластерный анализ и его визуализация.*
- 4. Анализ главных компонент и его визуализация.*
- 5. Преобразование данных при многомерном анализе.*

ИТОГОВОЕ ЗАНЯТИЕ ПО МОДУЛЮ 2

Отчет о выполнении заданий по модулю «Анализ данных».

Вариант №.....

Выполнил студент (Ф.И.О., № группы/подгруппы)

1. В соответствии с типом распределения анализируемых данных для сравнения двух независимых выборок выбран параметрический (непараметрический) тест.....

Таблица 1. Результаты проверки выборочных совокупностейс использованием теста.....

Рисунок 3. Диаграмма размаха средних величин

Вывод. Между выборками установлены статистически достоверные различия или не установлены и согласно чему (значение коэффициента (t , U),....., p =.....). Значениядостоверно выше.....

2. В соответствии с типом распределения анализируемых данных для сравнения двух независимых выборок использован параметрический (непараметрический) дисперсионный анализ.....

Таблица 1. Результаты проверки выборочных совокупностейс использованием дисперсионного анализа

Таблица 2 Результаты апостериорных сравнений

Рисунок 4. Диаграмма размаха средних величин

Вывод. Между выборками установлены статистически достоверные различия или не установлены и согласно чему (значение коэффициента, p =.....). Значениядостоверно выше для.....

3. В соответствии с типом распределения анализируемых данных для выявления зависимости между двумя переменными использован коэффициент корреляции.....

Таблица 3 Результаты корреляционного анализа

Вывод. Между и выявлена положительная (отрицательная) достоверная (p =) корреляционная зависимость. Связь между переменными сильная, средняя и т.д. (r (r_s)= ...). Или зависимость не установлена (p =).

4. Для выявления сходства между анализируемыми данными проведен кластерный анализ с использованием меры сходства.....

Рисунок 5. Дендрограмма сходства.....

Вывод. Наибольшим сходством обладают....., тогда как наименьшим.....

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Введение в статистическое обучение с примерами на языке R / Е. Джеймс, Д. Уитгон, Т. Хасты, Р.В. Тибширани; пер. с англ. С.Э. Мاستицкого. – 2-е изд., испр. – Москва: ДМК Пресс, 2017. – 456 с.
2. Дворецкий, М.Л. Пособие по вариационной статистике / М.Л. Дворецкий. – М.: «Лесная промышленность», 1971. – 134 с.
3. Ивантер, Э.В. Элементарная биометрия: учеб. пособие / Э.В. Ивантер, А.В. Коросов. – Петрозаводск: Изд-во ПетрГУ, 2010. – 104 с.
4. Лакин, Г.Ф. Биометрия / Г.Ф. Лакин. – Изд. четвертое, перераб. и доп. – Москва: «Высшая школа», 1990. – 350 с.
5. Орлов, А.И. Математика случая. Вероятность и статистика – основные факторы / А.И. Орлов. – М.: МЗ-Пресс, 2004. – 158 с.
6. Рокицкий, П.Ф. Биологическая статистика / П.Ф. Рокицкий. – Изд. 3-е, испр. – Минск: «Вышэйш. школа», 1973. – 320 с.
7. Халафян, А.А. Statistica 6. Статистический анализ данных / А.А. Халафян. – Москва: ООО Бином-Пресс, 2007. – 512 с.
8. Чайковская, Н.А. Биометрия: курс лекций: в 2 ч. / Н.А. Чайковская. – Гродно: ГрГУ, 2012. – 56 с.
9. Hammer, Ø., Harper, D.A.T., and P.D. Ryan, 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica 4(1): 9pp.

Учебное издание

СУШКО Геннадий Геннадьевич
ЛИТВЕНКОВА Инна Александровна

БИОМЕТРИЯ

Методические указания
для проведения лабораторных работ

В 2 частях

Часть 2

Технический редактор
Компьютерный дизайн

Г.В. Разбоева
Е.В. Крайло

Подписано в печать 03.10.2019. Формат 60x84 ¹/₁₆. Бумага офсетная.
Усл. печ. л. 2,73. Уч.-изд. л. 1,21. Тираж 60 экз. Заказ 106.

Издатель и полиграфическое исполнение – учреждение образования
«Витебский государственный университет имени П.М. Машерова».

Свидетельство о государственной регистрации в качестве издателя,
изготовителя, распространителя печатных изданий

№ 1/255 от 31.03.2014 г.

Отпечатано на ризографе учреждения образования
«Витебский государственный университет имени П.М. Машерова».
210038, г. Витебск, Московский проспект, 33.