

Лабораторная работа № 4

«ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ В PAST»

Дисперсионный анализ (лат. «*dispersio*» – рассеивание; в английском варианте – *analysis of variance* (ANOVA) – анализ переменных) применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну зависимую количественную переменную.

В основе дисперсионного анализа лежит предположение о том, что одни переменные могут рассматриваться как причины (факторы, независимые переменные), а другие – как следствия (зависимые переменные).

Цель дисперсионного анализа – исследование значимости различия между средними с помощью сравнения (анализа) дисперсий. Необходимо иметь в виду, что в случае простого сравнения средних двух выборок, дисперсионный анализ даст тот же результат, что и обычный t-тест.

По количеству выявляемых регулируемых факторов дисперсионный анализ может быть:

- *однофакторным* – изучается влияние одного фактора на результаты эксперимента;
- *двухфакторным* – изучается влияние двух факторов;
- *многофакторным* – позволяет оценить не только влияние каждого из факторов в отдельности, но и их взаимодействие.

Дисперсионный анализ относится к группе параметрических методов и поэтому его следует применять только тогда, когда доказано, что распределение подчиняется закону нормального распределения.

Рассмотрим проведение однофакторного дисперсионного анализа на следующем примере.

Задание 1. Имеются данные по численности беспозвоночных в почвенных ловушках в окрестностях нефтяных скважин и на контрольном участке (таблица 1).

Таблица 1 – Численность беспозвоночных в почвенных ловушках

скважина 127	38	13	97	68	16	67	39	75	38	14
скважина 174	240	205	139	116	27	39	79	109	60	150
контроль	114	35	98	55	200	101	140	58	39	56
скважина 127	19	36	50	135	44	48	31	52	86	16
скважина 174	431	328	115	88	66	27	81	103	59	24
контроль	37	24	23	13	1	15	142	121	82	191

Решение:

Создадим электронную таблицу с данными (рисунок 1)

уч_пример.dat

File Edit Transform Plot Univariate Multivariate

Show

☐ Row attributes

☐ Column attributes

Click mode

☒ Select

☐ Drag rows/columns

	скважина 12	скважина 17	контроль
1	38	240	114
2	13	205	35
3	97	139	98
4	68	116	55
5	16	27	200
6	67	39	101
7	39	79	140
8	75	109	58
9	38	60	39
10	14	150	56
11	19	431	37
12	36	328	24
13	50	115	23
14	135	88	13
15	44	66	1
16	48	27	15
17	31	81	142
18	52	103	121
19	86	59	82
20	16	24	191

Рисунок 1 – Электронная таблица с данными

Далее, предварительно выделив диапазон данных необходимо выбрать в меню однофакторный дисперсионный анализ (ANOVA). Для этого необходимо перейти в меню по следующему пути: **Univariate** (Одномерные данные) → **ANOVA etc. (several samples)** (Дисперсионный анализ и т.д. (несколько переменных)) → **Several-sample tests (ANOVA, Kruskal-Wallis)** (Тест для нескольких переменных (параметрический дисперсионный анализ, Краскел-Уоллис)) – рисунок 2.

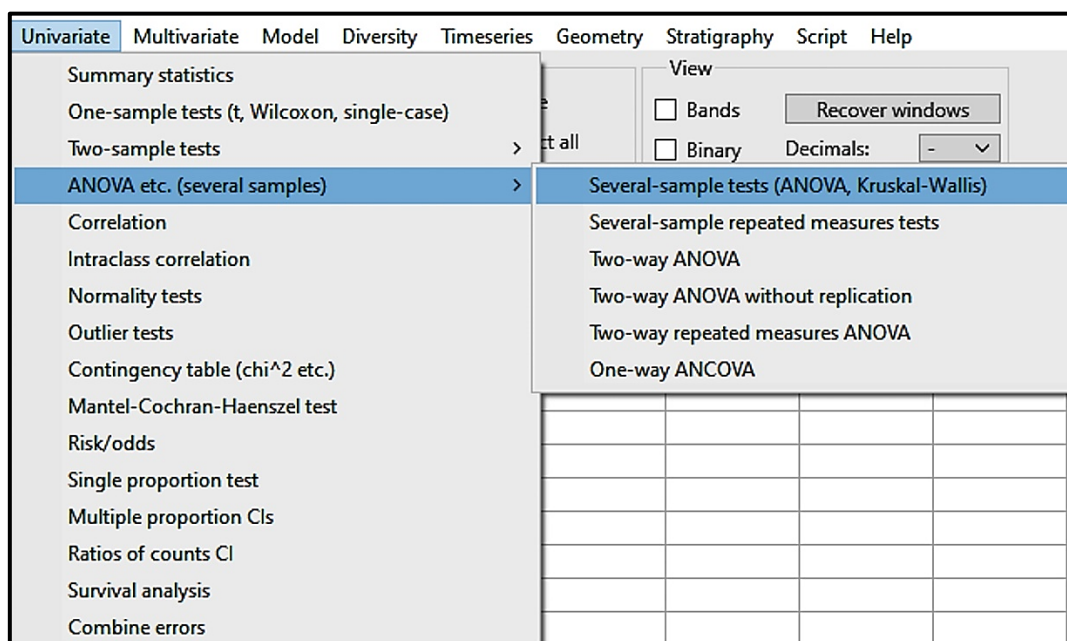


Рисунок 2 – Выбор в меню пункта **Several-sample tests (ANOVA, Kruskal-Wallis)**

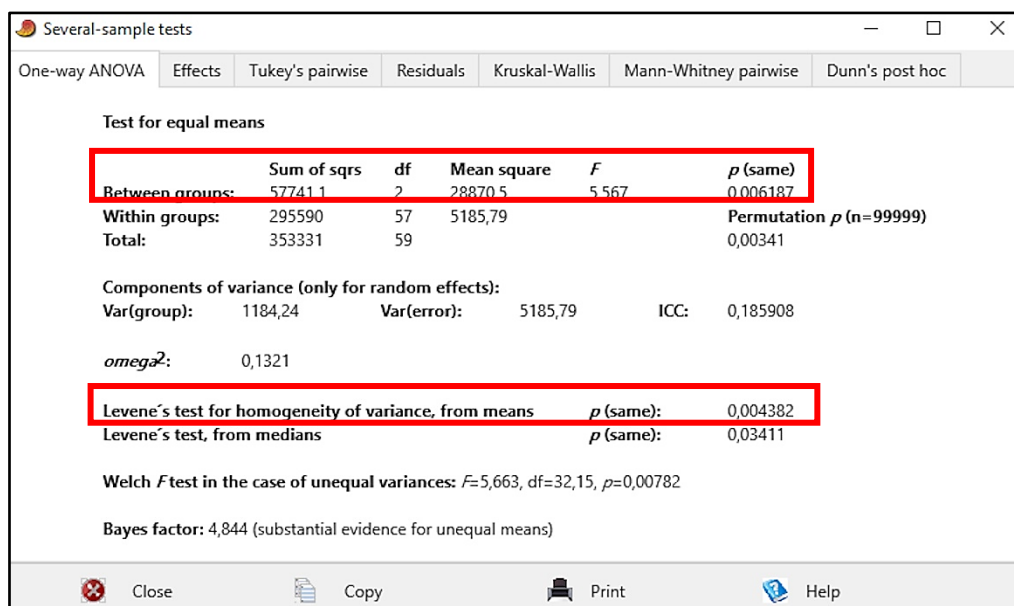


Рисунок 3 – Окно результатов однофакторного дисперсионного анализа

В появившемся окне результатов (рисунок 3) имеется несколько закладок:

- **One-way ANOVA** (Однофакторный дисперсионный анализ)
- **Effects** (Эффекты)
- **Tukey's pairwise** (Попарный тест Тьюки)
- **Residuals** (Остатки)
- **Kruskal-Wallis** (Непараметрический анализ Краскела-Уоллиса)
- **Mann-Whitney pairwise** (Непараметрический попарный тест Манна-Уитни)
- **Dunn's post hoc** (Пост хок анализ Данна).

Рассмотрим каждую из них.

Закладка **One-way ANOVA** (Однофакторный дисперсионный анализ)

В данной закладке (рисунок 3) собственно отражены результаты анализа, где нас интересуют несколько параметров. Это результаты сравнения между группами (**Between groups**), т.е. оценка факториального влияния и тест Левена на гомогенность выборок (**Leven's test for homogeneity of variance, from means**). Первый из рассмотренных параметров отражает рассчитанный критерий Фишера (**F**) для наших переменных и его уровень значимости (**p(same)**). Рассчитанный экспериментальный критерий Фишера оказался выше табличного, о чем свидетельствует высокий уровень *p* (значительно меньше, чем минимальный 0,05 – 0,006187). Это свидетельствует о достоверном различии между оцениваемыми выборками и, следовательно, о влиянии фактора на исследуемый нами признак – численность беспозвоночных.

Другой важный показатель – тест Левена на гомогенность выборок. Если результат этого теста указывает на отсутствие различий между дисперсиями ($p > 0,05$), то применение параметрического варианта дисперсионного анализа является обоснованным. В нашем случае это не так ($p=0,004382$), следовательно, необходимо воспользоваться для дисперсионного непараметрическим тестом Краскела-Уоллиса.

Закладка **Effects** (Эффекты)

Данная закладка (рисунок 4) показывает 95% доверительный интервал значений для групповых различий (множественный t-критерий Стьюдента) в рядах и столбцах.

Several-sample tests				
One-way ANOVA	Effects	Tukey's pairwise	Residuals	Kruskal-Wa
95% confidence intervals for group differences (Multiple-t), row-column				
	скважина 127	скважина 174	контроль	
скважина 127		-131.4; -19.03	-84.32; 28.02	
скважина 174	19.03; 131.4		-9.122; 103.2	
контроль	-28.02; 84.32	-103.2; 9.122		

Рисунок 4 – Закладка **Effects** (Эффекты)

В нашем случае наиболее сильные различия отмечены между 174 и 127 скважинами.

Закладка **Tukey's pairwise** (Попарный тест Тьюки)

Результаты теста Тьюки отражены в виде таблицы (рисунок 5).

Several-sample tests				
One-way ANOVA	Effects	Tukey's pairwise	Residuals	Kruskal-W
Tukey's Q below the diagonal, p(same) above the diagonal. Significant comparisons are pink.				
	скважина 127	скважина 174	контроль	
скважина 127		0.004659	0.4371	
скважина 174	4.67		0.1061	
контроль	1.748	2.922		

Рисунок 5 – Закладка **Tukey's pairwise** (Попарный тест Тьюки)

Над диагональю указаны уровни значимости сравнения выборок, а под диагональю – значения Тьюки (Q). В случае достоверной разницы соответствующее значение p подсвечивается. В нашем случае это достоверная разница между численностью беспозвоночных около скважины 127 и 174 – $p=0,004659$.

Закладка **Residuals** (*Остатки*)

В данной закладке можно посмотреть насколько точки наблюдений укладываются вдоль теоретически ожидаемой прямой (рисунок 6).

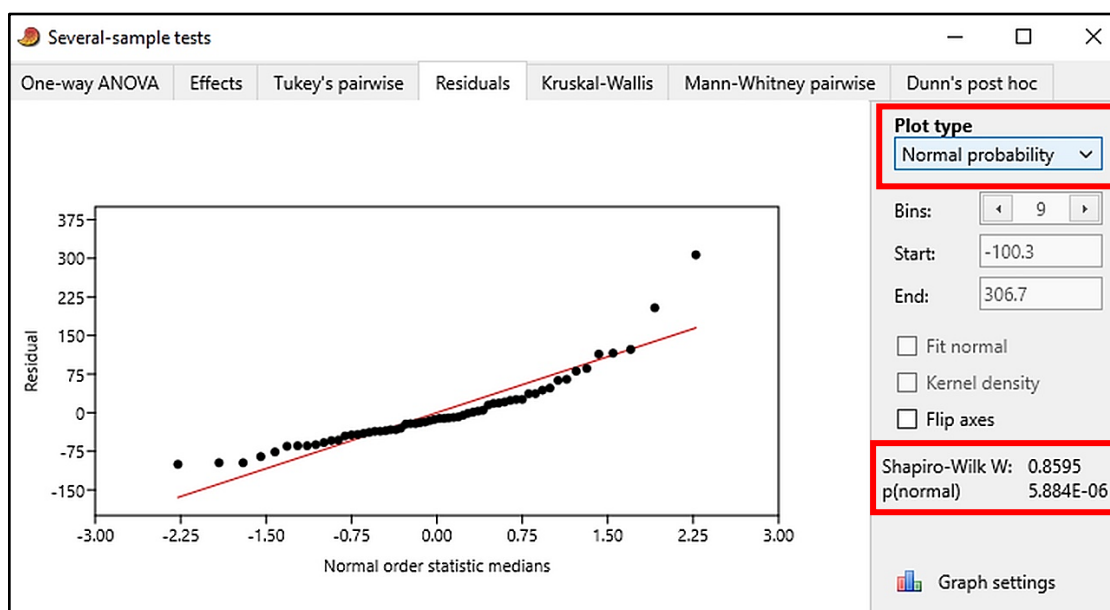


Рисунок 6 – Закладка **Residuals** (*Остатки*)

Для этого справа в разделе **Plot type** (*Тип графика*) в ниспадающем списке необходимо выбрать позицию **Normal probability** (*Нормальное распределение*). В том случае, если точки плотно лежат вдоль теоретически ожидаемой прямой можно говорить о нормальном распределении значений признака. Дополнительно внизу справа выводится результат анализа Шапиро-Уилка (**Shapiro-Wilk W** – само значение теста и его уровень значимости – **p(normal)**). В нашем случае о нормальном распределении говорить не приходится, поэтому использовать ANOVA в качестве дисперсионного анализа нельзя. Данный график используется если объем каждой выборки менее 30. Если он имеет больше 30 значений, то лучше использовать диаграмму в виде гистограммы распределений (опция **Histogram** в том же ниспадающем списке **Plot type**).

Закладка **Kruskal-Wallis** (*Непараметрический анализ Краскела-Уоллиса*)

Если распределение в выборке не подчиняется закону нормального распределения, то оценить влияние фактора на признак можно при помощи теста Краскела-Уоллиса, результаты которого выводятся вместе с результатами ANOVA (рисунок 7)

Kruskal-Wallis test for equal medians	
$H(ch^2)$:	9.505
H_c (tie corrected):	9.508
p (same):	0.008619
There is a significant difference between sample medians	

Рисунок 7 – Результаты непараметрического анализа Краскела-Уолиса (**Kruskal-Wallis**) в соответствующей закладке

Результаты анализа говорят о достоверном влиянии фактора на признак, о чем свидетельствует как значение уровня значимости ($p=0,008619$), так и соответствующая надпись под результатами – «**There is a significant difference between sample medians**» (*Существует значительное различие между медианами выборок*).

Закладка **Mann-Whitney pairwise**

(Непараметрический попарный тест Манна-Уитни)

Данная закладка является своеобразным аналогом теста Тьюки, только используется для непараметрического анализа (рисунок 8).

Several-sample tests				
One-way ANOVA	Effects	Tukey's pairwise	Residuals	Kruskal-Wallis
Mann-Whitney pairwise				
Raw p values, uncorrected significance				
	скважина 127	скважина 174	контроль	
скважина 127		0.00204	0.1805	
скважина 174	0.00204		0.1017	
контроль	0.1805	0.1017		

Рисунок 8 – Результаты непараметрического парного теста Манна-Уитни (**Mann-Whitney pairwise**) в соответствующей закладке

Результаты, отраженные в этой закладке, также подтверждают наиболее существенные отличия в численности беспозвоночных между теми, которые обитают вблизи скважин 127 и 174.

Закладка **Dunn's post hoc** (*Пост хок анализ Данна*)

Эта закладка показывает результаты пост хок анализа, в котором можно подкорректировать уровень значимости сравнений, используя поправку Бонферрони для чего необходимо ее выбрать в ниспадающем списке вверху окна результатом, над таблицей (рисунок 9).

Several-sample tests

One-way ANOVA

Effects

Tukey's pairwise

Residuals

Kruskal-Wallis

Mann-Whitney pairwise

Dunn's post hoc

Bonferroni corrected p values

▼

	скважина 127	скважина 174	контроль
скважина 127		0.006239	0.4937
скважина 174	0.006239		0.2738
контроль	0.4937	0.2738	

Рисунок 9 – Результаты непараметрического парного теста Манна-Уитни (**Mann-Whitney pairwise**) в соответствующей закладке

Всякий раз, когда выполняется проверка гипотезы, всегда есть шанс совершить ошибку первого рода. Это когда отвергается нулевая гипотеза, когда она на самом деле верна. Иногда это называется «ложноположительным» – когда утверждается, что есть статистически значимый эффект, но на самом деле его нет.

Когда проводится проверка одной гипотезы, частота ошибок типа I равна уровню значимости (p), который обычно выбирается равным 0,01, 0,05 или 0,10. Однако, когда мы проводим сразу несколько проверок гипотез, вероятность получения ложного срабатывания возрастает.

Поправка Бонферрони относится к процессу корректировки уровня p для семейства статистических тестов таким образом, чтобы мы контролировали вероятность совершения ошибки I рода.

Формула поправки Бонферрони выглядит следующим образом:

$$p_{\text{новый}} = p_{\text{ориг}} / n,$$

где $p_{\text{ориг}}$ – исходный уровень p ; n – общее количество выполненных сравнений или тестов

Например, в нашем случае выполняется сразу три статистических теста и используется $p = 0,05$ для каждого теста, поправка Бонферрони говорит, что необходимо использовать $p_{\text{новый}} = 0,01667$:

$$p_{\text{новый}} = p_{\text{ориг}} / n = 0,05 / 3 = 0,01667$$

Таким образом, в нашем случае необходимо отклонять нулевую гипотезу каждого отдельного теста, только если p -значение теста меньше 0,01667, а так как значения p с учетом поправки Бонферрони составило 0,006239 (что значительно меньше, чем 0,01667) разница между численностью беспозвоночных около скважин 127 и 174 абсолютно достоверно.

В качестве заключения дисперсионного анализа используется диаграмма в виде боксплота («ящики и усы»).

Для создания такого рода диаграммы необходимо в разделе меню **Plot** (*Диаграмма*) выбрать пункт **Barchart/Boxplot** (*Столбчатая диаграмма/Боксплот*) – как на рисунке 10.

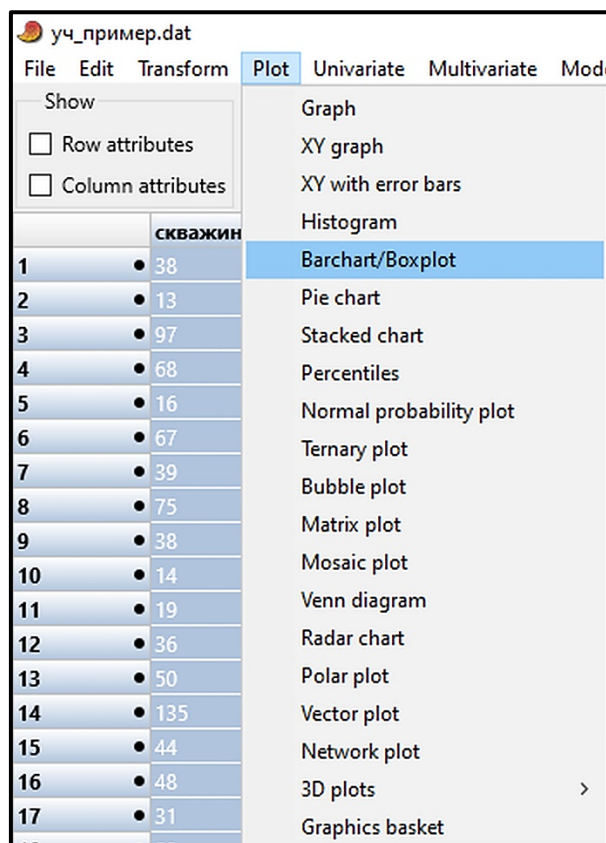


Рисунок 10 – Выбор пункта меню для построения боксплота

Затем в получившемся окне по умолчанию будет отображена столбчатая диаграмма, которую нужно изменить на боксплот в разделе **Plot type** (*Тип диаграммы*) – рисунок 11.

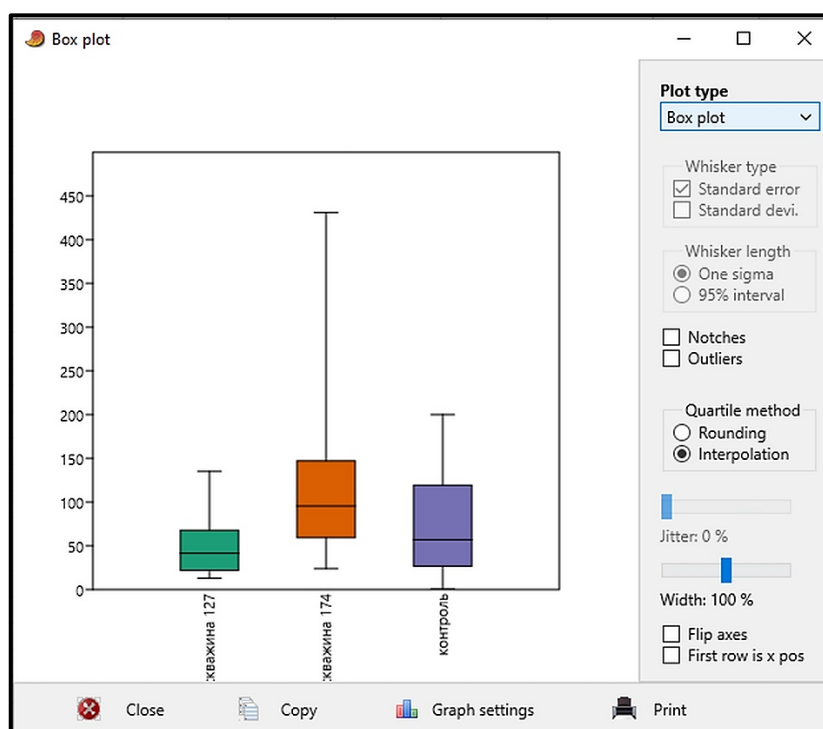


Рисунок 11 – Предварительная диаграмма – боксплот

Получившийся боксплот необходимо настроить, изменив толщину линий, шрифта, размеров самой диаграммы и подписать названия осей. Выполните настройки согласно рисунку 12, предварительно нажав кнопку **Graph settings** (*Настройки графика*).

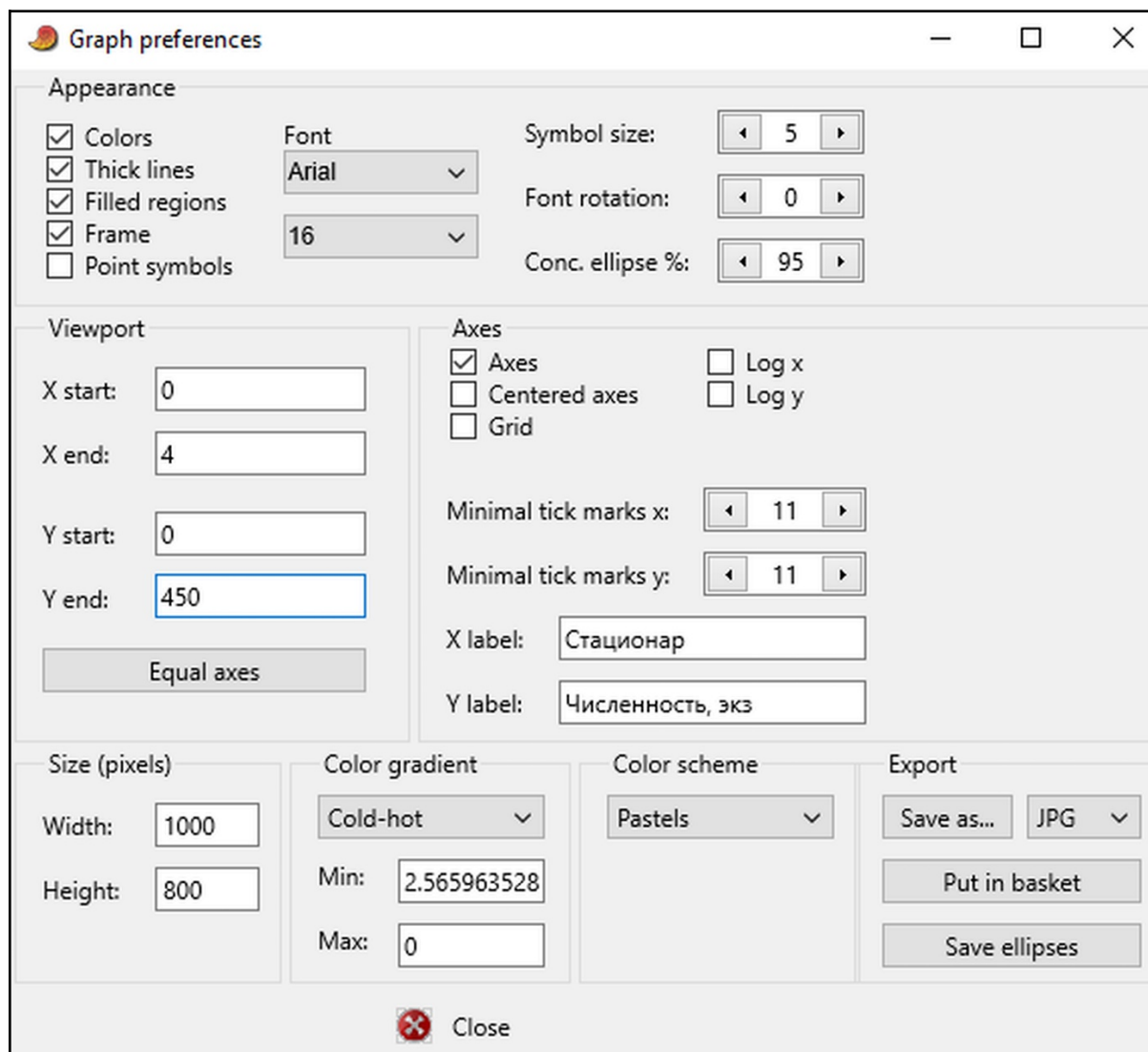


Рисунок 12 – Окно настроек графика

Получившуюся диаграмму сверьте с рисунком 13. Если в окне с получившимся рисунком поставить справа маркер в боксе **Notches** (*Вырезы*) программа сделает вырезы у боксов по линии медианы, а в случае маркировки бокса **Outliers** (*Выбросы*) можно увидеть значения признака, которые резко выделяются из остальной выборки и могут быть артефактами

Такую диаграмму можно использовать в качестве иллюстративного материала результатов исследований при оценке либо различий признаков, либо оценке влияния.

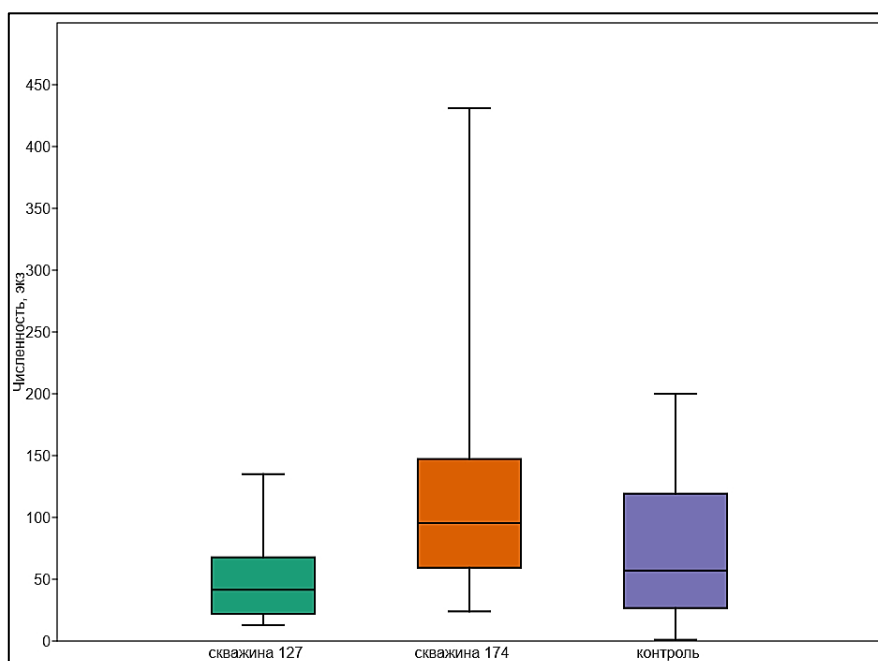


Рисунок 13 – Отредактированный боксплот

Задания для самоконтроля

1) Были получены данные по внесению смеси удобрений (*a*) и их влиянию на прирост ели в питомнике, в см. В качестве контроля использовались саженцы ели участка, где удобрения не вносили (*b*).

<i>a</i>	12,5	12,1	12,1	15,4	2,0	9,5	12,1	11,0	5,8	16,0	12,1
<i>b</i>	21,0	5,0	24,1	14,8	29,1	20,1	11,2	19,5	20,5	22,9	17,7
<i>a</i>	27,8	19,8	18,1	28,2	43,7	19,1	28,0	20,1	15,0	20,7	12,6
<i>b</i>	1,6	0,9	1,1	0,7	1,6	0,7	1,0	1,1	1,1	0,8	0,9
<i>a</i>	16,6	15,5	11,5	9,3	14,4	12,2	13,2	6,2	8,1	31,3	10,8
<i>b</i>	13,8	9,3	20,4	16,7	18,3	29,1	13,1	19,5	20,0	1,1	0,7

Выясните параметрическим, так и непараметрическим способом проведения дисперсионного анализа, влияет ли фактор внесения данной смеси удобрений на прирост ели. Постройте график «ящики и усы» для наглядной демонстрации модели.

2) Были получены данные по внесению смеси экспериментальных кормов (*a*) и их влиянию на привес телят, в кг. В качестве контроля выступали телята, которых кормили обычным кормом (*b*).

<i>a</i>	24,4	15,8	19,5	16,5	16,5	15,0	21,4	17,5	17,2	11,5	11,5	12,0	15,3	30,2	33,3	15,4
<i>b</i>	21,5	23,2	16,8	20,5	8,2	19,4	28,6	21,3	19,8	12,7	23,5	20,2	19,1	29,8	25,9	22,5
<i>a</i>	17,3	22,8	11,0	17,8	23,5	11,9	15,0	24,6	19,9	20,8	22,8	29,2	25,4	19,8	25,6	10,8
<i>b</i>	33,0	14,2	18,3	27,2	5,6	22,6	31,4	26,4	33,9	26,6	14,1	20,7	25,8	21,7	29,4	27,0

Выясните как параметрическим, так и непараметрическим способом проведения дисперсионного анализа, влияет ли фактор внесения данных кормов на прирост телят. Постройте график «ящики и усы» для наглядной демонстрации модели.

3) Было проведено внесение удобрений разных смесей (b , c , d) под посевы ячменя, на одно поле удобрения не вносились (a) для увеличения урожайности (в ц/га):

a	9,4	8,4	4,9	13,4	11,1	3,6	15,1	13,2	12,0	11,6	7,3
b	21,1	8,2	16,6	10,5	18,4	11,3	18,3	11,2	19,4	10,0	18,2
c	12,5	12,1	12,1	15,4	2,0	9,5	12,1	11,0	5,8	16,0	8,1
d	21,0	5,0	24,1	14,8	29,1	20,1	11,2	19,5	20,5	22,9	20,0

Определите, влияет ли внесение данных марок удобрений на урожайность ячменя. Постройте графики «ящики и усы». Дайте обоснованный ответ.

4) Изучали продолжительность развития эмбрионов (в днях) кроликов различных пород:

Породы	Продолжительность развития отдельных крольчат									
Альбиносы	30	36	31	30	34	32	34	32	33	32
	35	32	31	33	33	35	31	33	32	33
Шиншилла	31	32	30	34	32	31	30	31	30	31
	30	32	31	32	30	31	33	32	32	33
Голландские	30	29	30	31	30	30	30	31	31	31
	30	31	29	32	31	31	30	31	31	31
Польские	30	31	29	30	29	30	29	31	29	30
	30	30	31	30	30	30	31	30	31	30

Выясните, влияет ли породность развития эмбрионов на кроликов. Постройте графики «ящики и усы». Дайте обоснованный ответ.

5) У кубышки (*Nymphaea*) 4 раза в сутки проверяли концентрацию каротиноидов:

Часы суток	Определения									
	1	2	3	4	5	6	7	8	9	10
18	1,42	1,30	1,68	1,59	1,49	1,62	1,36	1,26	1,58	1,66
24	1,45	1,38	1,49	1,71	1,54	1,57	1,34	1,32	1,66	1,39
06	1,48	1,42	1,58	1,67	1,50	1,80	1,35	1,36	1,67	1,49
12	1,43	1,38	1,47	1,33	1,22	1,35	1,10	1,08	1,34	1,11

Определите, влияет ли время суток на содержание каротиноидов в листьях кубышки. Постройте графики «ящики и усы». Дайте обоснованный ответ.

6) Получены следующие данные о содержании хлорофилла а (в мг/дм³) в листьях канатника (*Abutilon*) в разное время суток:

Часы суток	Определения			
	1	2	3	4
15	3,06	2,88	2,83	2,41
18	3,20	2,97	2,50	3,03
21	1,82	1,73	1,33	2,25
24	1,67	1,26	1,52	1,36
6	2,76	1,26	1,46	1,32
9	2,78	2,70	2,49	1,66
12	2,41	3,22	1,90	2,00

Определите, влияет ли время суток на содержание хлорофилла в листьях канатника. Постройте графики «ящики и усы». Дайте обоснованный ответ.

7) Изучали живой вес ягнят-одиночек при рождении (в кг), ношенных разное число дней:

Длительность беременности	Живой вес ягнят									
145	3,8	2,9	3,3	3,6	3,8	3,7	4,8	5,1	3,4	3,3
146	3,7	2,9	3,3	3,6	3,9	3,7	4,7	5,0	3,4	3,2
147	3,9	4,1	4,4	5,0	3,0	2,9	4,0	3,2	4,2	4,3
148	4,0	5,2	4,3	2,9	4,1	3,9	3,2	3,9	4,2	4,0
149	4,0	5,3	4,2	3,0	4,0	3,9	4,2	3,3	4,0	4,1
150	4,1	4,3	5,4	3,1	4,0	4,0	4,3	3,9	4,0	4,1
151	4,3	4,2	5,5	4,2	4,1	4,1	4,4	3,5	4,1	3,6
152	4,3	3,6	4,4	5,5	4,0	4,1	4,5	4,1	4,2	4,3
153	4,4	4,7	3,9	4,6	5,7	4,3	4,8	4,9	4,7	4,7

Примените метод дисперсионного анализа для выяснения влияния длительности плодоношения на живой вес ягнят. Постройте графики «ящики и усы». Дайте обоснованный ответ.

8) Получены следующие данные о плодовитости мышей при облучении рентгеновыми лучами:

Группы	Число мышат от отдельных самок			
Контроль	10	12	11	10
Доза 100 Р	8	10	7	9
Доза 200 Р	7	9	6	4

Влияет ли облучение на плодовитость мышей? Постройте графики «ящички и усы». Дайте обоснованный ответ.

9) Были получены следующие данные о содержании хлорофилла *b* в листьях томата (в услов. ед.) в различные часы суток:

Часы суток	Определения									
	1	2	3	4	5	6	7	8	9	10
18	0,087	0,106	0,091	0,097	0,077	0,103	0,118	0,104	0,099	0,108
24	0,086	0,092	0,095	0,100	0,086	0,096	0,101	0,113	0,113	0,092
6	0,093	–	0,091	0,089	0,081	–	0,118	0,106	0,088	0,096
12	0,088	0,091	0,104	0,113	0,079	0,097	0,114	0,107	0,103	0,093

Определите методами дисперсионного анализа, оказывают ли влияние различные часы суток на содержание хлорофилла *b* в листьях томата. Постройте графики «ящички и усы». Дайте обоснованный ответ.