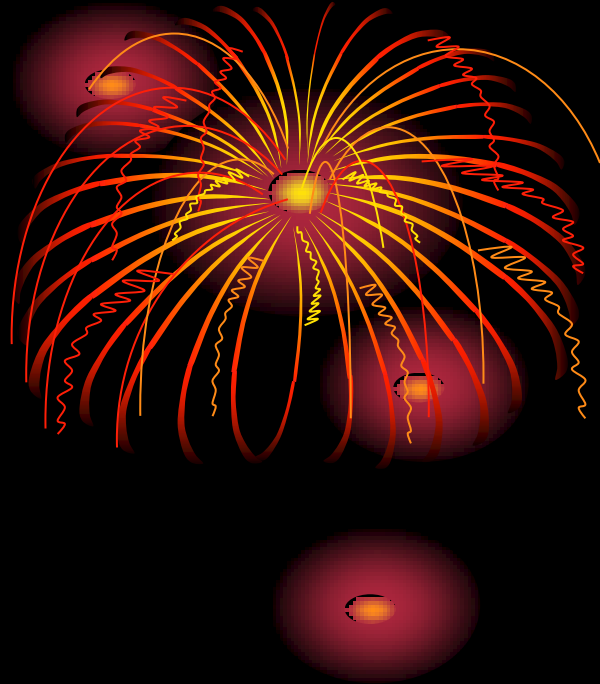




Лекция 7



Дисперсионный анализ.

Сравнение средних является одним из способов выявления зависимостей между переменными. Так, например, если при разбиении объектов исследования на подгруппы при помощи категориальной независимой переменной (предиктора) верна гипотеза о неравенстве средних некоторой зависимой переменной в подгруппах, то это означает, что существует стохастическая взаимосвязь между этой зависимой переменной и категориальным предиктором. Наиболее общим методом сравнения средних является дисперсионный анализ – **ANOVA (Analysis of Variance)**. В терминологии дисперсионного анализа категориальный предиктор называется **фактором**.

Таким образом, в дисперсионном анализе можно исследовать зависимость количественного признака (зависимой переменной) от одного или нескольких качественных признаков (факторов), например зависимость успеваемости студентов от пола, курса, факультета.

Рассмотрим сначала основные идеи однофакторного дисперсионного анализа. Представим файл исходных данных в виде таблицы, строки и столбцы которой отображают различные уровни фактора X_i , например, курс 1, ..., курс 4, в ячейках таблицы расположены значения анализируемого признака (зависимой переменной) для n объектов. В терминологии планирования экспериментов такая таблица называется планом эксперимента.

Фактор	Значения переменной			
Группа 1	X_1	x_{11}	x_{12}	$\dots x_{1n}$
Группа 2	X_2	x_{21}	x_{22}	$\dots x_{2n}$
.....				
Группа m	X_m	x_{m1}	x_{m2}	$\dots x_{mn}$

Однофакторная, дисперсионная модель имеет следующий вид:

$$x_{ij} = \mu + F_i + \varepsilon_{ij} ,$$

где x_{ij} — значение исследуемой переменной, соответствующей i -й группе (i -у уровню фактора X_i) с j -м порядковым номером объекта ($i = 1, \dots, m; j = 1, \dots, n$), μ — общая средняя, F_i — эффект, обусловленный влиянием i -го уровня фактора, ε_{ij} — случайная компонента, или возмущение, вызванное влиянием неконтролируемых факторов, т.е. вариацией переменных внутри отдельного уровня факторов.

Предположим, что элементы строк таблицы — реализации случайных величин X_1, X_2, \dots, X_m , имеющих **нормальный закон распределения** с математическими ожиданиями a_1, a_2, \dots, a_m и одинаковыми дисперсиями σ^2 для генеральной совокупности. Тогда задача сравнения средних в группах сведется к проверке нулевой гипотезы —

$$H_0: a_1 = a_2 = \dots = a_m .$$

Обозначим выборочные средние в группах $\bar{x}_{i\bullet}$, а общую выборочную среднюю — \bar{x} . Тогда

$$\bar{x}_{i\bullet} = \sum_{j=1}^n x_{ij}/n, \quad \bar{x} = \sum_{i=1}^m \sum_{j=1}^n x_{ij}/mn = \sum_{i=1}^m \bar{x}_{i\bullet}/m.$$

Можно показать, что сумму квадратов отклонений Θ наблюдений x_{ij} от общей средней \bar{x} можно представить следующим образом:

$$\begin{aligned} \Theta &= \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x})^2 = \sum_{i=1}^m \sum_{j=1}^n (\bar{x}_{i\bullet} - \bar{x})^2 + \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{i\bullet})^2 = \\ &= n \sum_{i=1}^m (\bar{x}_{i\bullet} - \bar{x})^2 + \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{i\bullet})^2. \end{aligned}$$

Обозначим слагаемые в правой части равенства, соответственно Θ_1 и Θ_2 . Получим соотношение $\Theta = \Theta_1 + \Theta_2$. Здесь Θ – общая, или полная, сумма квадратов отклонений, Θ_1 – межгрупповая (факторная) сумма квадратов отклонений, Θ_2 – внутригрупповая (остаточная) сумма квадратов отклонений. Полученное равенство показывает, что общая изменчивость признака, измеренная величиной Θ , состоит из двух компонент, одна из которых характеризует изменчивость признака Θ_1 между группами, вторая – изменчивость внутри групп Θ_2 . В дисперсионном анализе используются не сами суммы квадратов отклонений Θ_1, Θ_2 , а усредненные квадраты отклонений S_1, S_2 , получающиеся делением последних на число степеней свободы.

Число степеней свободы определяется как общее число наблюдений минус число связывающих их уравнений. Для Θ_1 число степеней свободы равно $l_1 = m - 1$, для $\Theta_2 - l_2 = mn - m$. Таким образом, $S_1 = \Theta_1 / m - 1$, $S_2 = \Theta_2 / mn - m$.

В терминах модуля **ANOVA**, Θ_1 называют эффектом, а Θ_2 называют ошибкой. S_1 , S_2 называют, соответственно, MS эффекта и MS ошибки. Можно показать, что проверка нулевой гипотезы сводится к проверке существенности различия MS эффекта и MS ошибки, которые являются оценками дисперсии σ^2 . MS эффекта и MS ошибки можно сравнить с помощью F -критерия. Гипотеза H_0 отвергается, если $F = S_1 / S_2$ больше табличного F_{α, l_1, l_2} , или уровень значимости p критерия меньше 0,05.

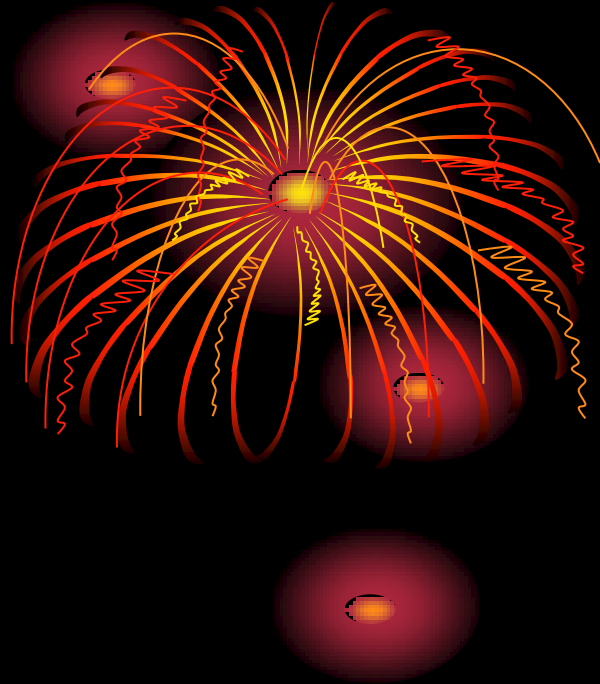
Если сравниваются средние в двух выборках, дисперсионный анализ даст тот же результат, что и обычный t -критерий для независимых выборок (если сравниваются две независимые группы наблюдений) или t -критерий для зависимых выборок (если сравниваются две переменные на одном и том же множестве наблюдений).

Основная причина, по которой использование дисперсионного анализа предпочтительнее повторного сравнения двух выборок при разных уровнях факторов с помощью серий t -критерия, заключается в том, что дисперсионный анализ существенно более эффективен и для малых выборок более информативен. Еще одно преимущество дисперсионного анализа по сравнению с *t*-критерием заключается в том, что он позволяет обнаруживать взаимодействия между факторами и, следовательно, изучать более сложные модели.

Идея однофакторного дисперсионного анализа перенесена в многофакторный анализ. Более сложными становятся факторный план эксперимента и процедуры вычисления MS эффекта и MS ошибки. Так, например, для двухфакторного дисперсионного анализа факторный план можно представить в виде табл.1.

Таблица 1			
Факторы	Группа 1*	Группа 2*	Группа k^*
Группа 1	x_{111}, \dots, x_{11j}	x_{121}, \dots, x_{12j}	$\dots x_{1k1}, \dots, x_{1kj}$
Группа 2	x_{211}, \dots, x_{21j}	x_{221}, \dots, x_{22j}	$\dots x_{2k1}, \dots, x_{2kj}$
.....	\dots	\dots	\dots
Группа m	x_{m11}, \dots, x_{m1j}	x_{m21}, \dots, x_{m2j}	x_{mk1}, \dots, x_{mkj}

Применение дисперсионного анализа целесообразно, если анализируемые признаки измерены минимум в интервальной шкале и имеют **нормальное распределение** внутри сравниваемых групп, дисперсии в группах однородны. Но следует заметить, что *F-критерий*, применяемый в дисперсионном анализе, **устойчив к незначительному отклонению от нормальности и однородности дисперсий**. Если условия применимости дисперсионного анализа не выполнены, **можно** воспользоваться непараметрическими критериями сравнения средних.



Описание процедуры Factorial ANOVA .

Рассмотрим работу процедуры **ANOVA**, используя файл **Crabs** (крабы) из библиотеки **Examples** (рис. 1). В файле приведены данные по количеству спутников (*SATELLTS*) – особей мужского пола у особей женского пола в зависимости от их цвета (*COLOR*), состояния клешней (*SPINE*), размеров (*CATWIDTH*, *WIDTH* – ширина) и веса (*WEGHT*). Если число спутников больше 0, то переменная *Y* в первом столбце принимает значение 1, в противном случае – 0. Общее число наблюдений (крабов) равно 173.

	Number of crab satellites by female's color, spine condition, width, and weight						
	Y	COLOR	SPINE	WIDTH	SATELLTS	WEIGHT	CATWIDTH
1	1	medium	bothworn	28,3	8	3,05	28,75
2	0	darkmed	bothworn	22,5	0	1,55	22,75
3	1	lightmed	bothgood	26,0	9	2,30	25,75
4	0	darkmed	bothworn	24,8	0	2,10	24,75
5	1	darkmed	bothworn	26,0	4	2,60	25,75
6	0	medium	bothworn	23,8	0	2,10	23,75
7	0	lightmed	bothgood	26,5	0	2,35	26,75
8	0	darkmed	oneworn	24,7	0	1,90	24,75
9	0	medium	bothgood	23,7	0	1,95	23,75
10	0	darkmed	bothworn	25,6	0	2,15	25,75
11	0	darkmed	bothworn	24,3	0	2,15	24,75
12	0	medium	bothworn	25,8	0	2,65	25,75
13	1	medium	bothworn	28,2	11	3,05	27,75
14	0	dark	oneworn	21,0	0	1,85	22,75
15	1	medium	bothgood	26,0	14	2,30	25,75

Рис. 1

Для запуска программы в верхнем меню **Statistics** надо выбрать команду **ANOVA**, что переводится как анализ вариаций или дисперсионный анализ. Появится стартовая панель **General ANOVA/MANOVA** (рис. 2).

Данный диалог содержит два списка **Type of analysis** (вид анализа) и **Specification method** (задание метода). Список **Type of analysis** состоит из четырех элементов, представляющих собой различные модели дисперсионного анализа:

- **One-way ANOVA** (однофакторный дисперсионный анализ);
- **Main effects ANOVA** (дисперсионный анализ главных эффектов);
- **Factorial ANOVA** (многофакторный дисперсионный анализ);
- **Repeat measures ANOVA** (дисперсионный анализ повторных измерений).

Список **Specification method** позволяет задать три типа интерфейса дисперсионного анализа в STATISTICA:

- **Quick Specs Dialog** (диалог быстрых спецификаций);
- **Analysis Wizard** (мастер анализа);
- **Analysis syntax editor** (редактор кода).

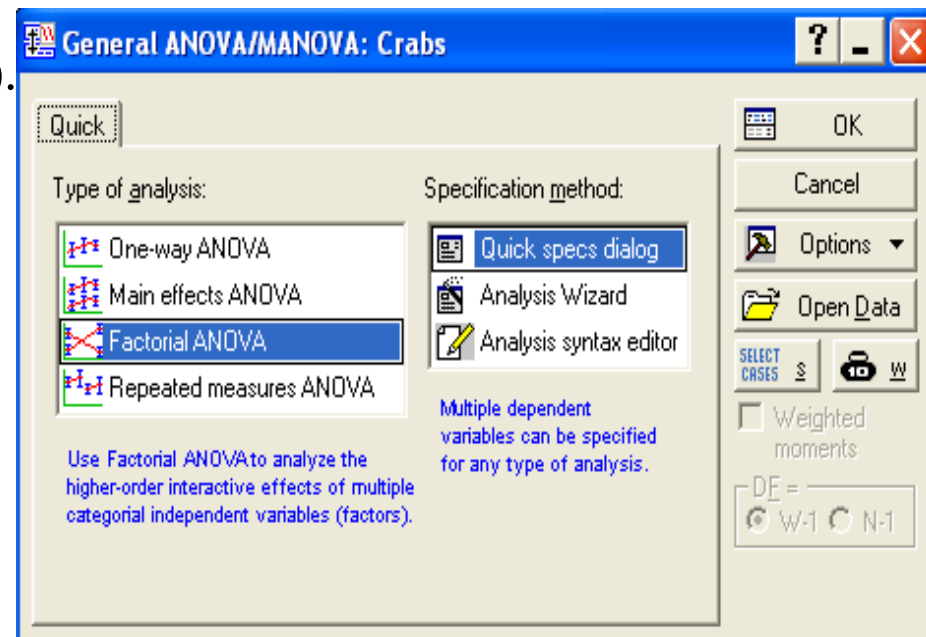


Рис. 2

В диалоге **Quick Specs Dialog** (диалог быстрых спецификаций) можно задать зависимые переменные и категориальные переменные (предикторы). Вариация числа и типа переменных зависит от выбранного вида анализа в списке **Type of analysis**.

Диалог **Analysis Wizard** (мастер анализа) предназначен для задания по шагам интересующего анализа в рамках выбранной модели. В конце анализа можно вычислить результаты или использовать **Analysis syntax editor** (редактор кода) для дальнейшей настройки при помощи встроенных команд, открыть существующий файл с командами или сохранить для дальнейшего использования. Диалог **Analysis syntax editor** позволяет полностью настроить параметры вычислительных процедур. В случае необходимости можно сохранить файл с готовым кодом анализа для дальнейшего использования или открыть уже существующий.

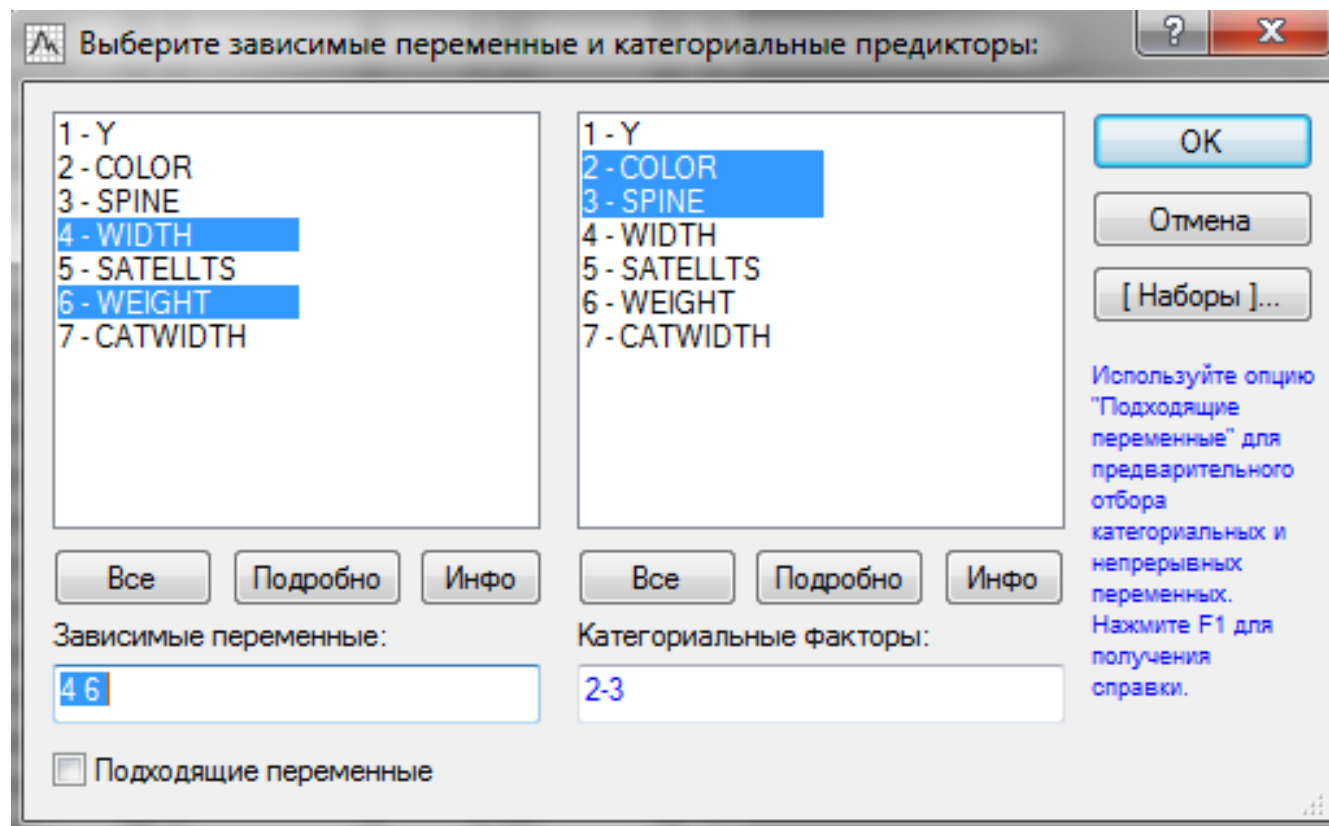
После выбора диалога **Specification method** можно задать **Type of analysis**.

One-way ANOVA позволяет оценить эффект одной группирующей переменной (одного межгруппового фактора) на одну или более зависимых переменных.

Для анализа **Main effects ANOVA** в диалоге **Quick Specs Dialog** можно задать до четырех категориальных предикторов. Затем программа произведет оценку модели главных эффектов. Данный тип планов часто используется в анализе и планировании промышленных экспериментов для оценки большого набора факторов в сильно раздробленных планах. Также данный тип планов используется при анализе сбалансированных неполных планов.

В отличие от рассмотренных типов анализа, в **Factorial ANOVA** учитывается еще один возможный источник изменчивости — взаимодействие факторов.

Для того чтобы задать план факторного дисперсионного анализа, надо выбрать **Factorial ANOVA** в качестве вида анализа и **Quick Spec Dialog** в списке **Specification method** на вкладке **Quick** стартовой панели дисперсионного анализа. Откроется диалоговое окно **ANOVA/MANOVA Factorial ANOVA**. На вкладке **Quick** нажмем кнопку **Variables**. В появившемся окне выберем группирующие переменные *COLOR* и *SPINE*, зависимые *WIDTH*, *WEIGHT* и нажмем на ОК, появится диалог на рис.3



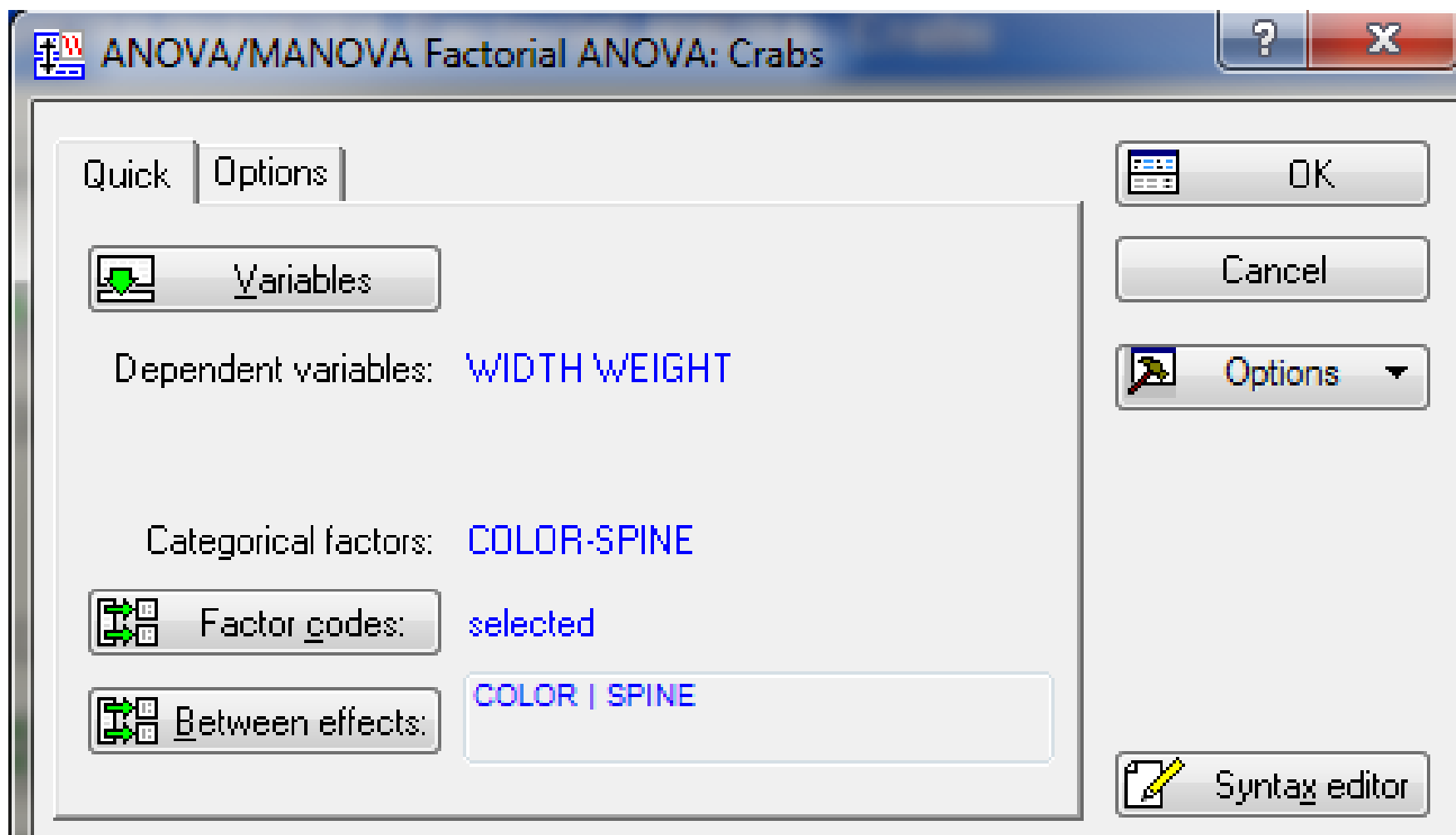
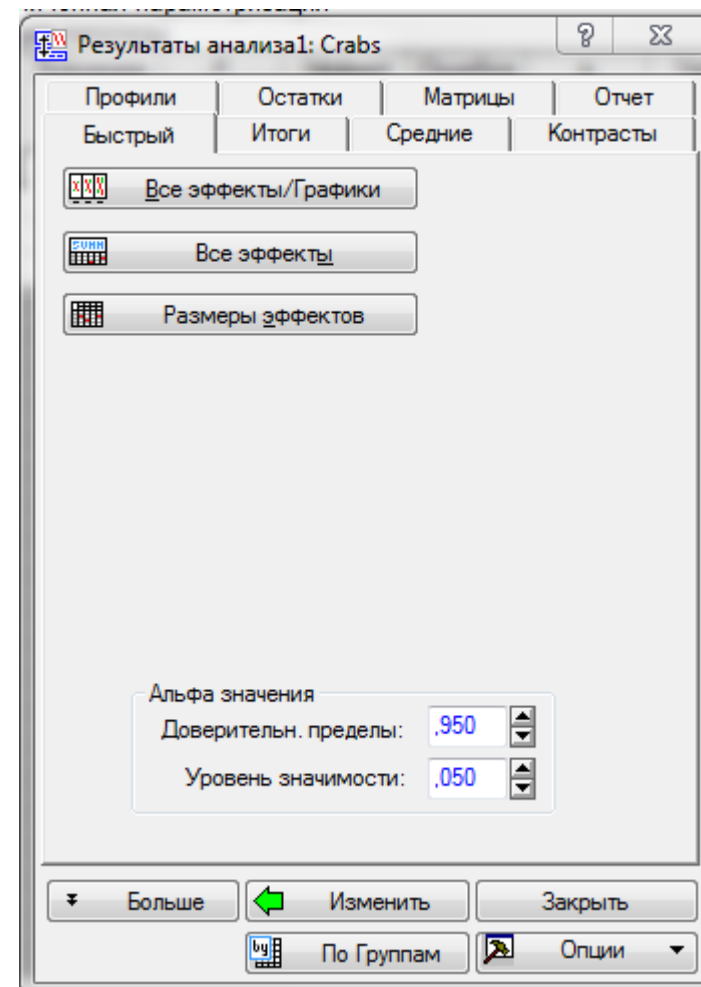
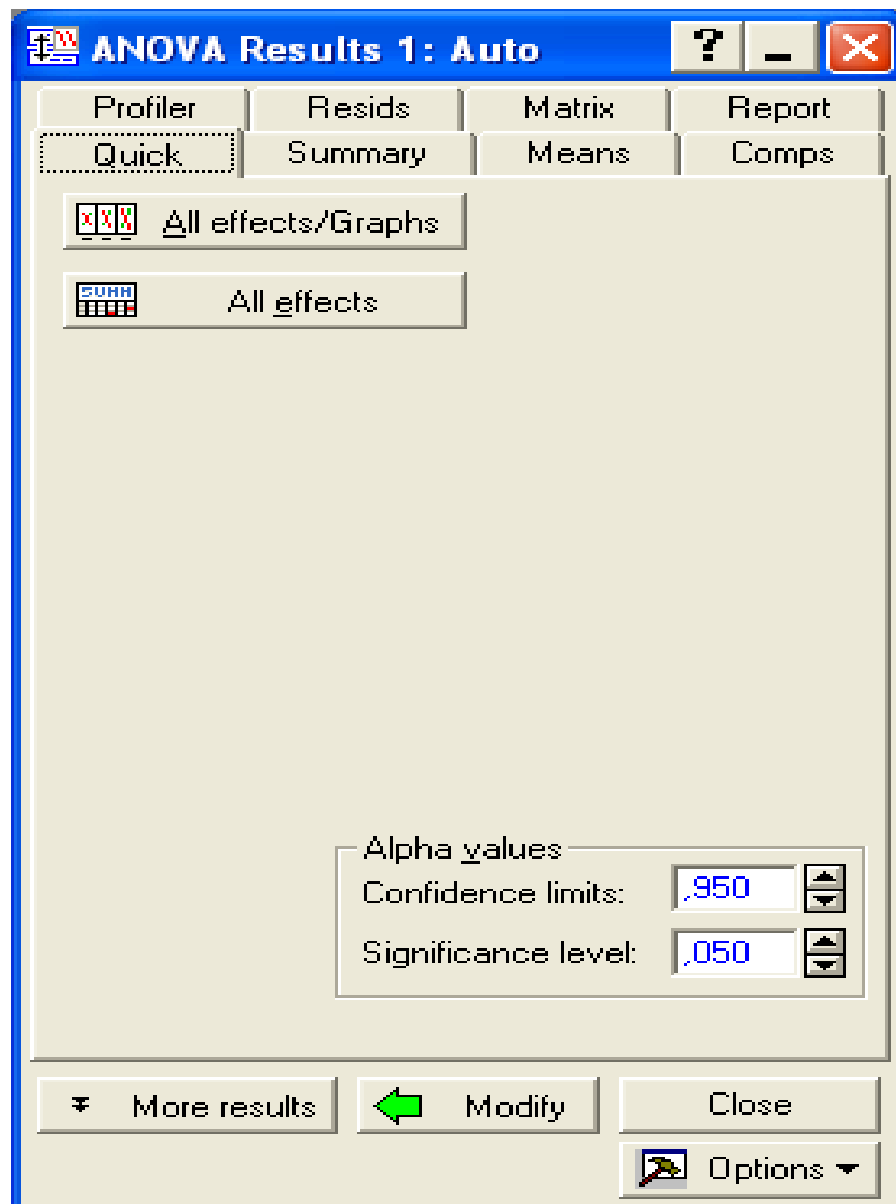


Рис. 3

Различные цвета и состояния клешней крабов являются межгрупповыми факторами. Чтобы вручную задавать коды для межгрупповых факторов, надо нажать кнопку **Factor Codes** (коды факторов). Необязательно коды задавать вручную, так как программа задаст по умолчанию все коды выбранных переменных. Кодами предиктора *COLOR* являются цвета крабов: *medium* (серый); *lightmed* (светло-серый); *dark* (темный); *darkmed* (темно-серый). Кодами предиктора *SPINE* являются состояния клешней крабов: *bothgood* (обе клешни целые); *oneworr* (одна клешня повреждена); *bothworr* (обе клешни повреждены).

Кнопка **Options** используется для задания параметров вычисления, кнопка **Syntax editor** (редактор кода) позволяет произвести дальнейшие настройки модели при помощи синтаксиса анализа.

Щелчком по кнопке **OK**, откроется диалоговое окно (рис. 4) **ANOVA Results 1** (результаты анализа) с набором вкладок, которые позволяют всесторонне отобразить результаты анализа в виде таблиц и графиков.



В более современных версиях пакета добавлена кнопка *Размеры эффекта*

Рис. 4

Если нажать на кнопку All affects, то появится таблица в строках которой будут указаны названия эффектов, значения критерия Уилкса, критерия Фишера дисперсионного анализа и в последнем столбце отображены уровни значимости критерия Фишера. Если $p < 0,05$, то это означает, что при разбиении на группы по данному эффекту наверняка будут группы со статистически значимыми отличиями средних

Эффект	Многомерные критерии значимости (Crabs) Сигма-ограниченная параметризация Декомпозиция гипотезы					
	Крит.	Знач.	F	Эффект ст.св.	Ошибка ст.св.	p
Св. член	Уилкса	0,012423	6359,569	2	160	0,000000
COLOR	Уилкса	0,925037	2,119	6	320	0,050860
SPINE	Уилкса	0,929532	2,977	4	320	0,019492
COLOR*SPINE	Уилкса	0,879619	1,766	12	320	0,052765

На вкладке **Quick** нажмем кнопку **All effects/Graphs** (все эффекты/графики). Данный диалог **Table of All Effects** (таблицы всех эффектов) (рис. 5) содержит результаты и используется для просмотра выбранных из данной таблицы эффектов в виде графиков средних или таблиц.

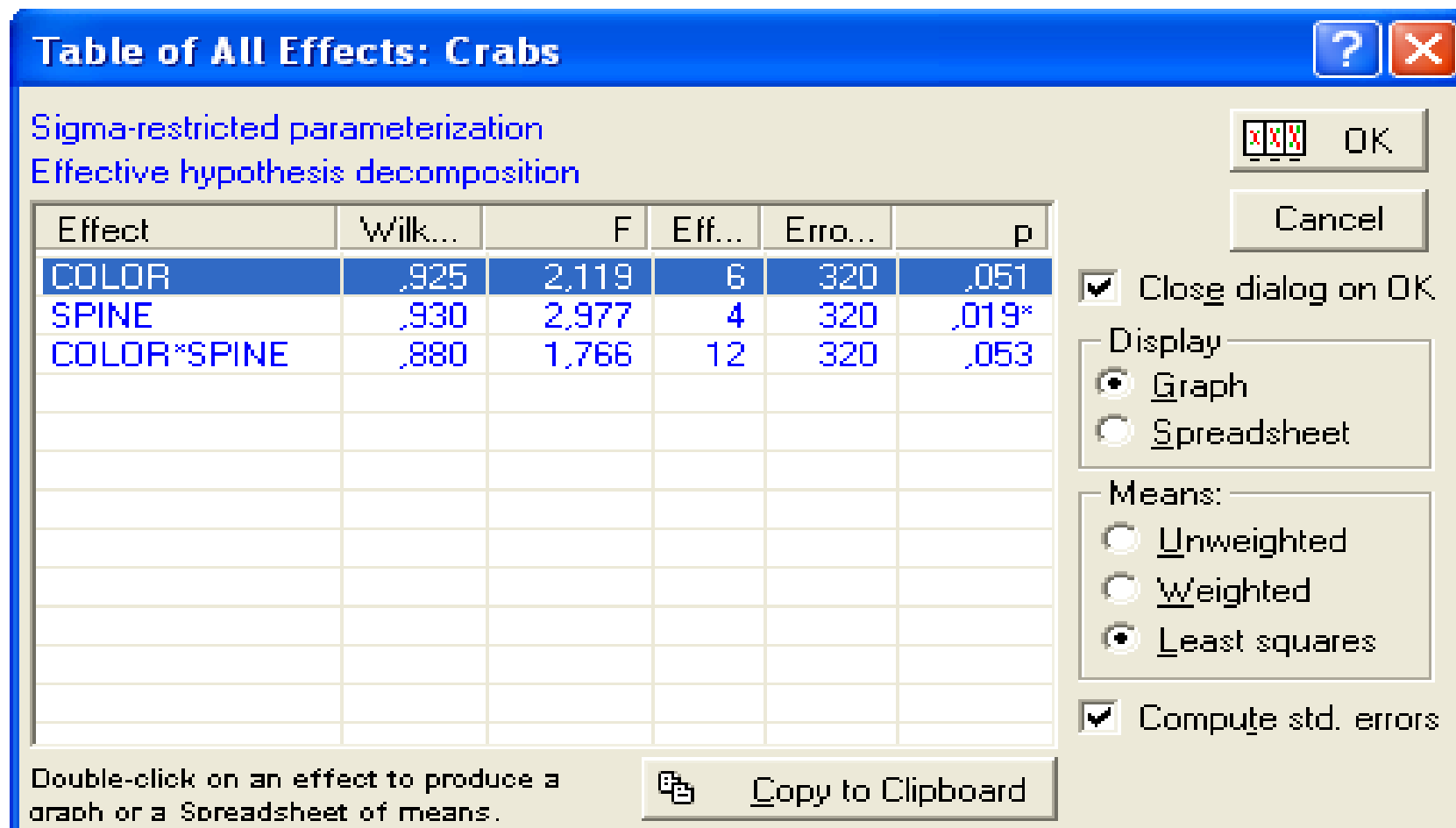


Рис. 5

Значимые эффекты ($p < 0,05$) в таблице *Table of All Effects* помечены *. Видно, что гипотеза о неравенстве средних верна только для предиктора *SPINE*. Для предиктора *COLOR* и взаимодействия предикторов *COLOR*SPINE* уровень значимости незначительно превосходит 0,05.

Можно изменить значимость критерия, введя необходимое значение параметра *Alpha* (альфа) в поле **Significance level** (уровень значимости) вкладки **Quick** окна **ANOVA Results 1**.

Выделим, например, опцию *Spreadsheet* (таблица) в рамке **Display** (отображать) и два раза щелкнем на эффекте *SPINE* или, выделив эффект *SPINE*, нажмем **ОК**. Появится таблица (рис. 6) со значениями средних всех зависимых переменных и другими статистиками в группах, соответствующих трем уровням категориального предиктора *SPINE* – *bothgood*; *oneworr*; *bothworr*.

SPINE; МНК средние (Crabs) Лямбда Уилкса=,92953, F(4, 320)=2,9770, p=,01949 Декомпозиция гипотезы										
N ячеек	SPINE	WIDTH Средне	WIDTH Стд. ош.	WIDTH -95,00%	WIDTH +95,00%	WEIGHT Средне	WEIGHT Стд. ош.	WEIGHT -95,00%	WEIGHT +95,00%	N
1	bothgood	26,80451	0,598481	25,62263	27,98640	2,721788	0,168747	2,388546	3,055030	37
2	oneworr	24,10625	0,672241	22,77870	25,43380	2,120313	0,189544	1,745999	2,494626	15
3	bothworr	25,93783	0,513233	24,92429	26,95137	2,391723	0,144710	2,105948	2,677498	121

Рис. 6

Вернемся в окно **Table of All Effects** и выделим опцию *Graph* (график) в рамке **Display**, нажмите **OK**. В появившемся окне выберем, например, зависимую переменную *WIDTH*. Программа построит график средних переменной *WIDTH* (рис. 7).

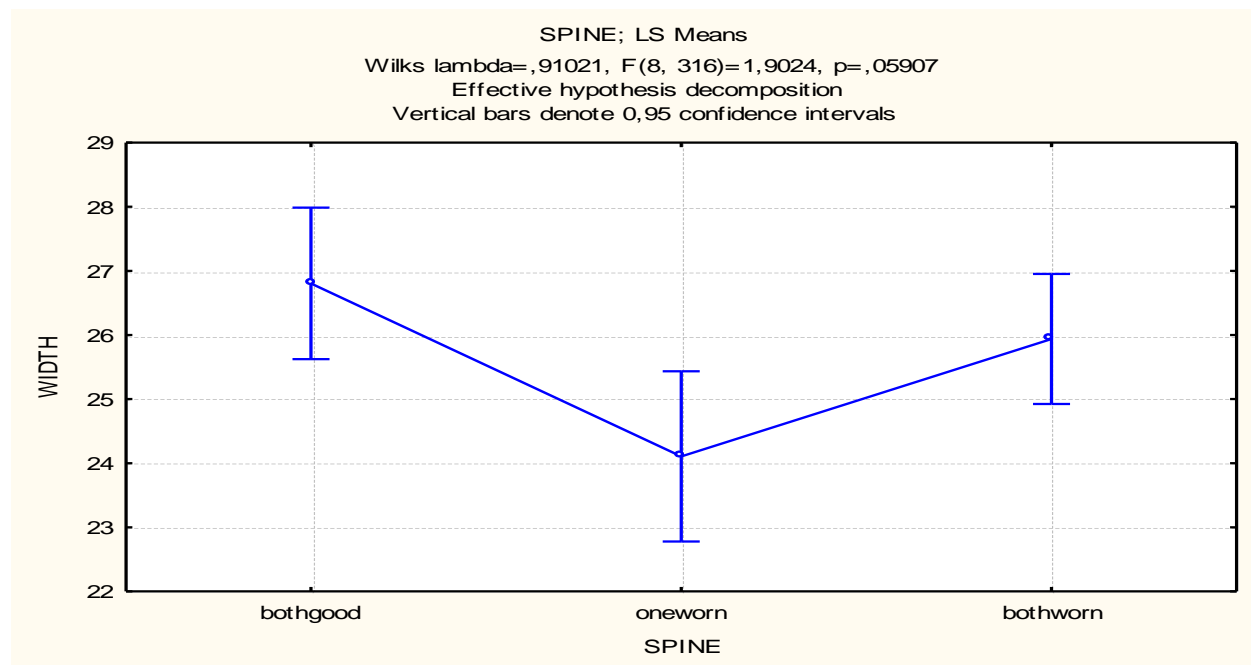
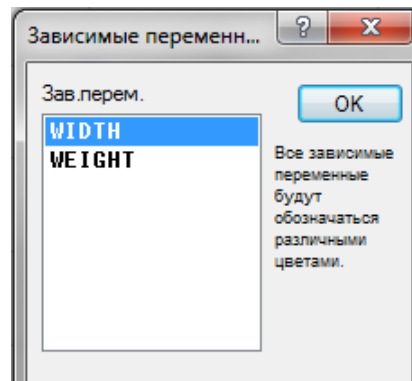
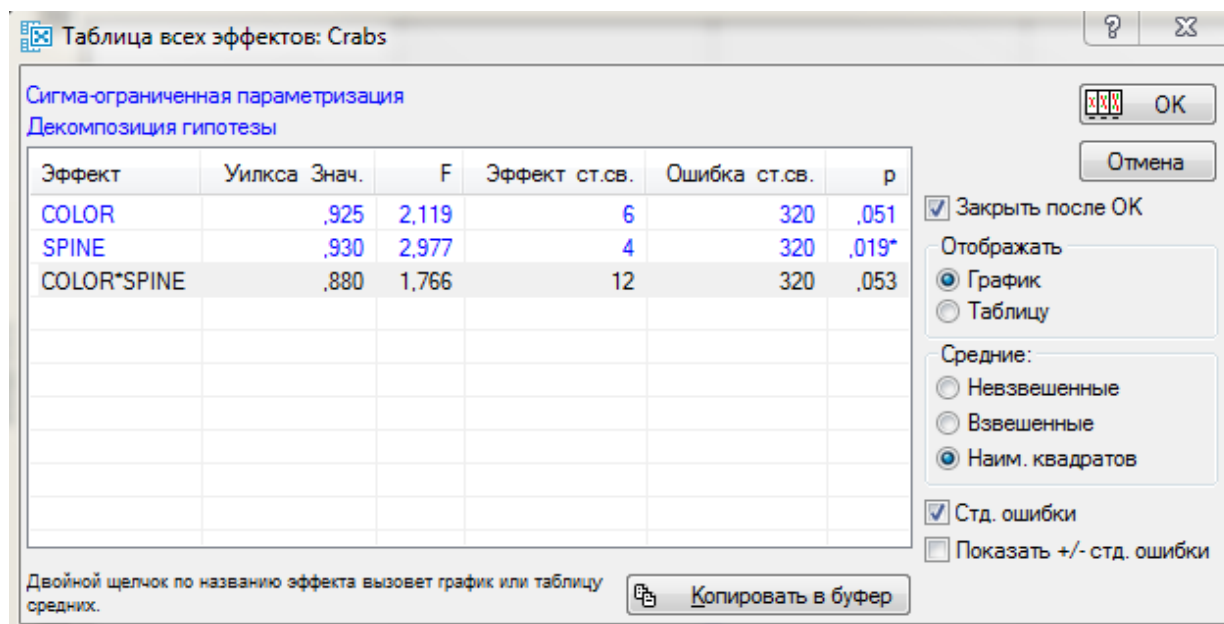


Рис. 7

Из графика и таблицы видно, что средняя ширина крабов с двумя целыми клешнями превосходит ширину крабов с двумя поврежденными клешнями и значительно превосходит ширину крабов с одной поврежденной клешней. По-видимому, более широкие крабы обладают большей силой, и это позволяет им сохранить клешни в целости.

Выделим опцию *Graph* в рамке **Display**, эффект **color*spine** и нажмем **OK**/



В открывшемся окне (рис. 8) **Dependent vars for the...** укажем имя переменной *WIDTH*. Щелкнем **OK**, появится окно (рис. 9) **Arrangement of Factors** (расположение факторов), в котором можно указать порядок выбора взаимодействующих факторов. Выберем *COLOR* под ось *X*, верх и *SPINE* под шаблоном линии. Нажмем кнопку **OK**, появятся графики средних (рис. 10).

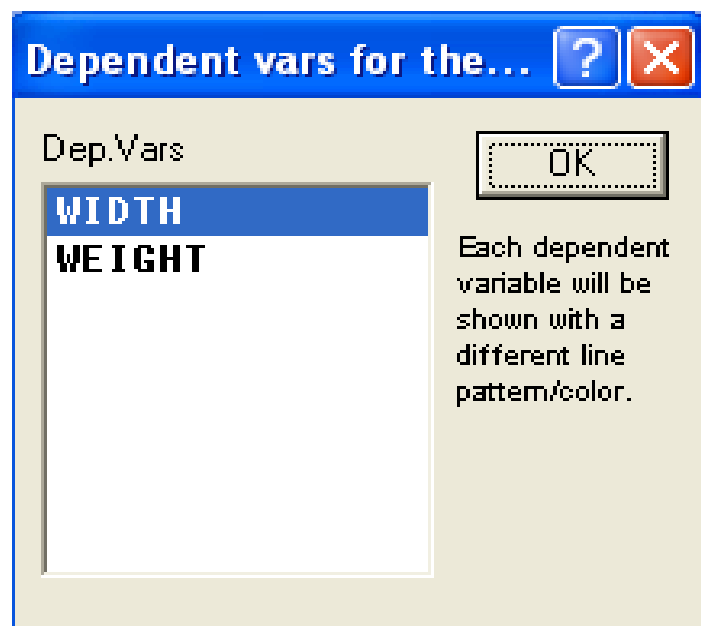


Рис. 8

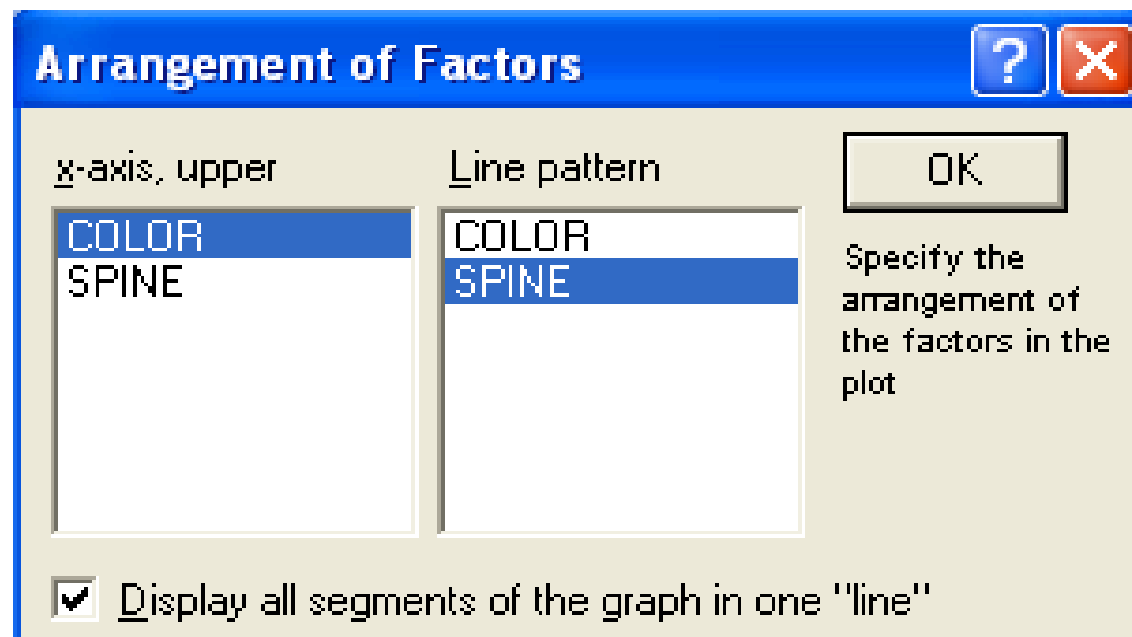


Рис. 9

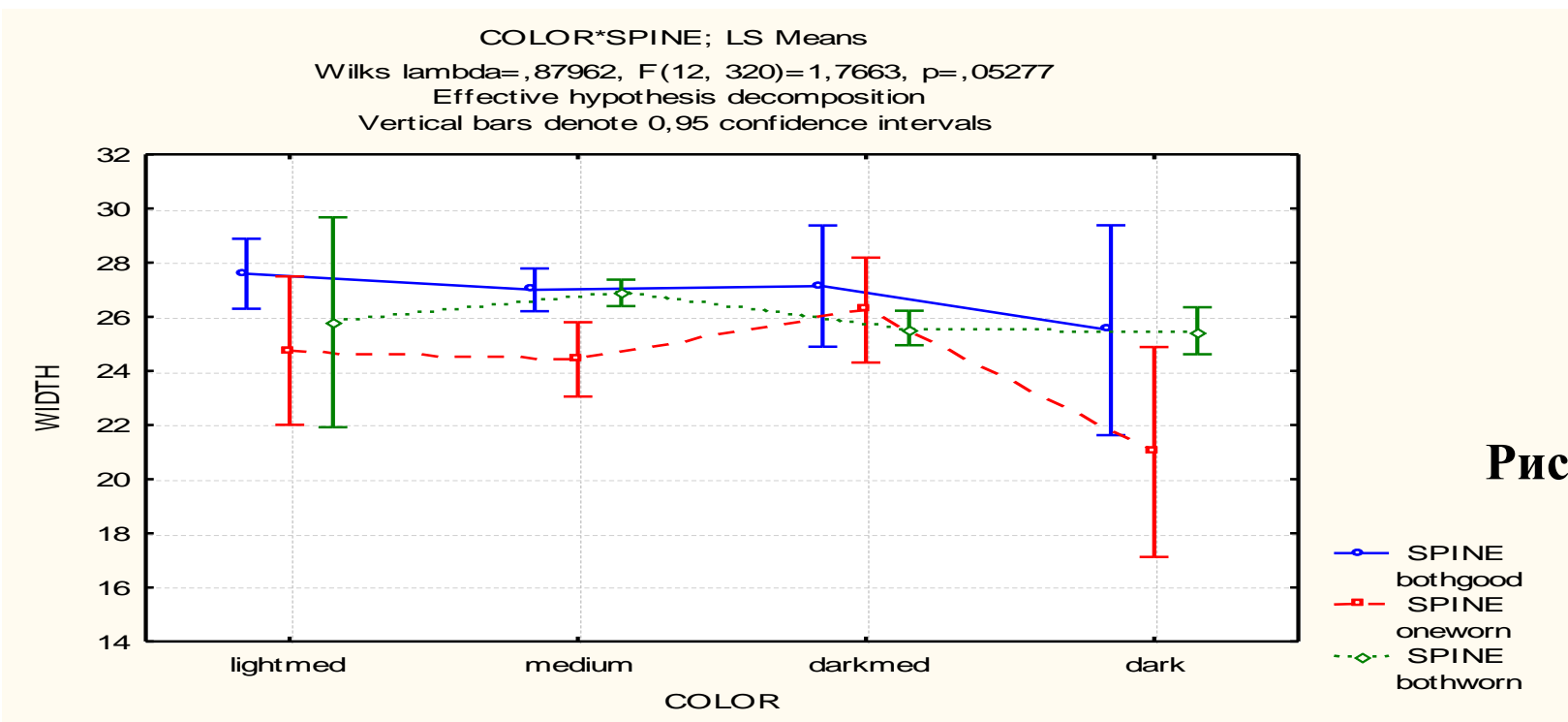
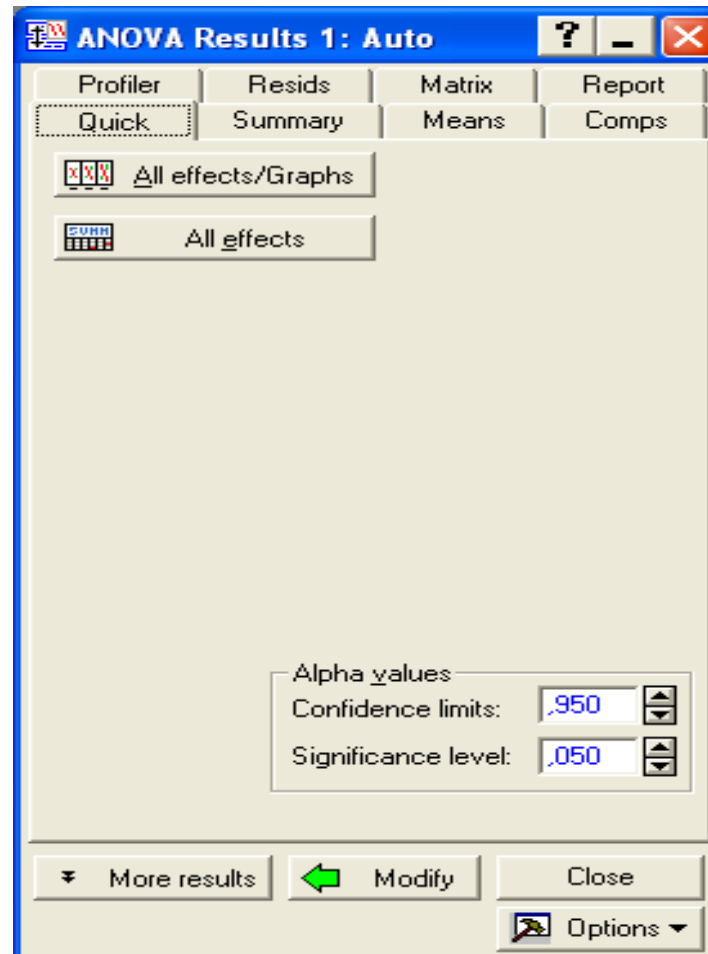


Рис. 10

Видно, что у крабов светло-серого цвета с двумя целыми клешнями и у крабов темного цвета с одной поврежденной клешней средняя ширина соответственно больше и меньше, чем во всех остальных группах. Независимо от цвета средняя ширина крабов с одной поврежденной клешней меньше, чем средняя ширина крабов с двумя целыми клешнями. Приведенные результаты показывают, что существуют различия между средними в группах, соответствующих различным межгрупповым факторам. Но значимы ли эти различия? Для ответа на этот вопрос нужно использовать апостериорные сравнения для проверки разности средних.

В диалоге **ANOVA Results 1** нажмем кнопку **More results**



в открывшемся окне выберем вкладку **Post-hoc**, на которой представлены различные апостериорные критерии (рис. 11). Все эти критерии позволяют сравнивать средние при отсутствии априорной гипотезы относительно этих средних. Большое количество критериев минимизирует вероятность случайных результатов.

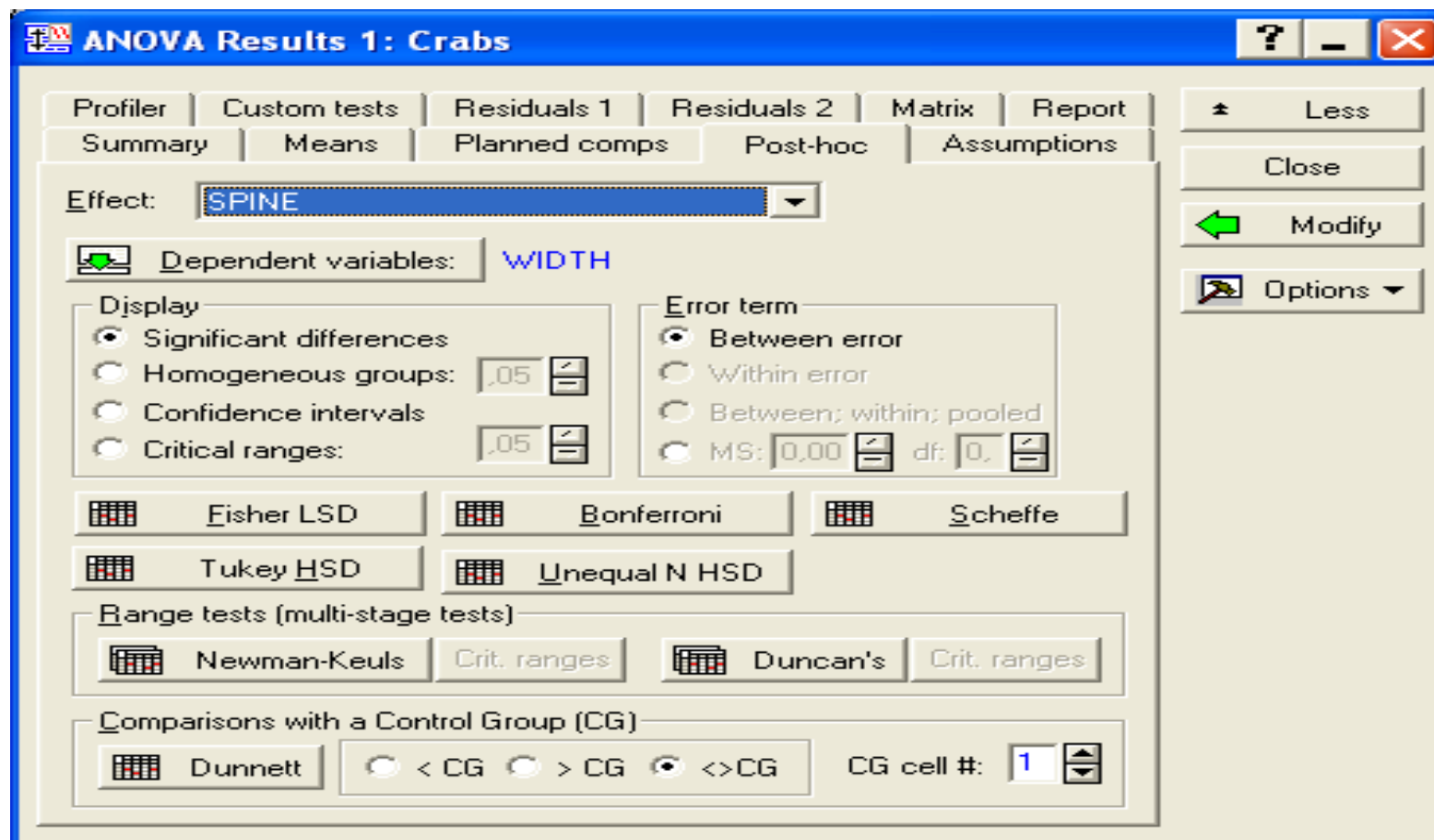


Рис. 11

Выберем зависимую переменную *WIDTH*, эффект *SPINE* и нажмем кнопку **Fisher LSD**. В открывшейся таблице (рис. 12) в первой строке приведены значения средних, в столбце 1 – названия групп, в остальных ячейках – уровни значимости. Из таблицы следует, что неверна гипотеза о равенстве средних, т.е. средняя ширина крабов статистически значимо отличается во всех группах, соответствующих различным уровням предиктора *SPINE*.

		LSD test; variable WIDTH (Crabs) Probabilities for Post Hoc Tests Error: Between MS = 3,8563, df = 161,		
Cell No.	SPINE	{1} 27,111	{2} 24,727	{3} 26,245
1	bothgood		0,00011	0,02020
2	oneworried	0,00011		0,00532
3	bothworried	0,02020	0,00532	

Рис. 12

Более интересный результат получится, если в диалоге **ANOVA Results 1** для исследования взаимодействия предикторов выбрать эффект *COLOR*SPINE*. Так, из таблицы, изображенной на рис. 13, следует, что средняя ширина крабов светло-серого цвета с обеими целыми клешнями (27,58) значимо больше, чем средняя ширина крабов серого цвета с одной поврежденной клешней (24,42). Средняя ширина крабов темного цвета с одной поврежденной клешней (27,58) значимо больше, чем средняя ширина крабов умеренного цвета с обеими целыми клешнями (26,99).

LSD test; variable WIDTH (Crabs) Probabilities for Post Hoc Tests Error: Between MS = 3,8563, df = 161,00														
Cell	COLOF	SPINE	{1} 27,58	{2} 24,75	{3} 25,80	{4} 26,99	{5} 24,42	{6} 26,88	{7} 27,13	{8} 26,25	{9} 25,58	{10} 25,50	{11} 21,00	{12} 25,485
1	lightmed	bothgood		0,07	0,39	0,44	0,00	0,31	0,73	0,26	0,01	0,31	0,00	0,01
2	lightmed	oneworn	0,07		0,66	0,12	0,83	0,13	0,19	0,38	0,56	0,76	0,12	0,61
3	lightmed	bothworn	0,39	0,66		0,55	0,51	0,59	0,56	0,84	0,91	0,91	0,09	0,88
4	medium	bothgood	0,44	0,12	0,55		0,00	0,81	0,91	0,48	0,01	0,46	0,00	0,01
5	medium	oneworn	0,00	0,83	0,51	0,00		0,00	0,04	0,13	0,13	0,61	0,10	0,20
6	medium	bothworn	0,31	0,13	0,59	0,81	0,00		0,83	0,53	0,00	0,49	0,00	0,01
7	darkmed	bothgood	0,73	0,19	0,56	0,91	0,04	0,83		0,56	0,19	0,47	0,01	0,18
8	darkmed	oneworn	0,26	0,38	0,84	0,48	0,13	0,53	0,56		0,52	0,73	0,02	0,48
9	darkmed	bothworn	0,01	0,56	0,91	0,01	0,13	0,00	0,19	0,52		0,97	0,02	0,86
10	dark	bothgood	0,31	0,76	0,91	0,46	0,61	0,49	0,47	0,73	0,97		0,11	0,99
11	dark	oneworn	0,00	0,12	0,09	0,00	0,10	0,00	0,01	0,02	0,02	0,11		0,03
12	dark	bothworn	0,01	0,61	0,88	0,01	0,20	0,01	0,18	0,48	0,86	0,99	0,03	

Рис. 13

Для проверки предположений, лежащих в основе метода дисперсионного анализа, необходимо воспользоваться вкладкой **Assumptions** (предположения) в окне **ANOVA Results 1** (рис. 14). На вкладке представлены различные критерии проверки гипотезы однородности дисперсий (критерий Кохрана, Хартли, Бартлетта, критерий Левена, М критерий Бокса), графические средства проверки соответствия закона распределения переменной нормальному закону (гистограммы, диаграммы рассеяния, нормальные вероятностные графики).

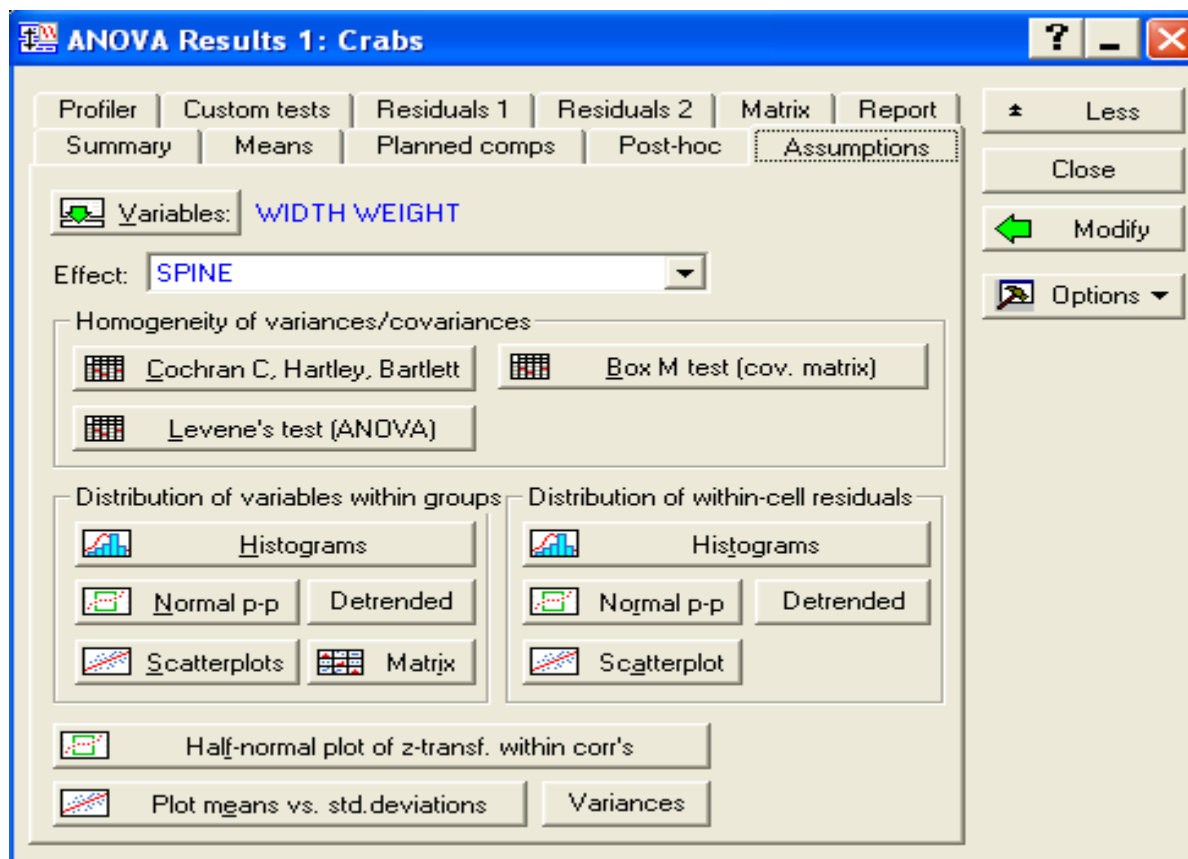
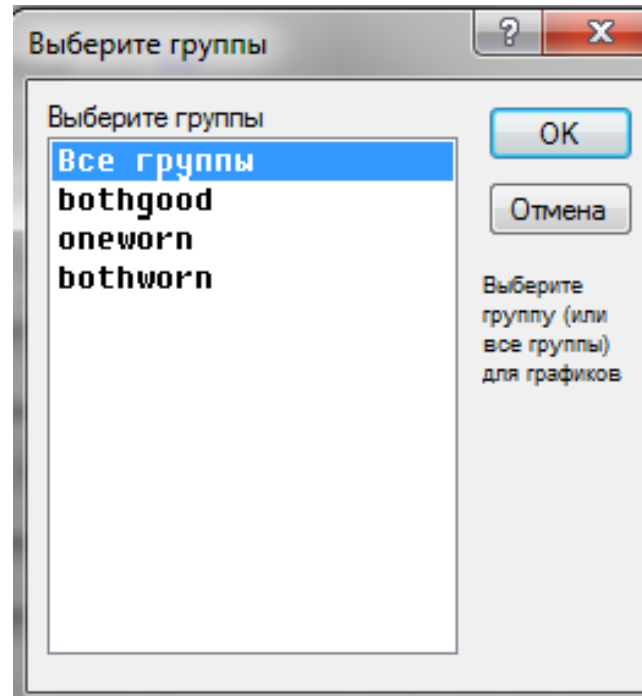


Рис. 14

Выберем эффект *SPINE* и нажмем кнопку **Histograms**. В появившемся окне выберем переменную *WIDTH* и укажем группу, если нужно проанализировать распределение внутри каждой группы.



Если выбрать *All* (все), то программа построит (рис. 15) гистограмму частот для всех групп. Видно, что общее распределение соответствует нормальному закону.

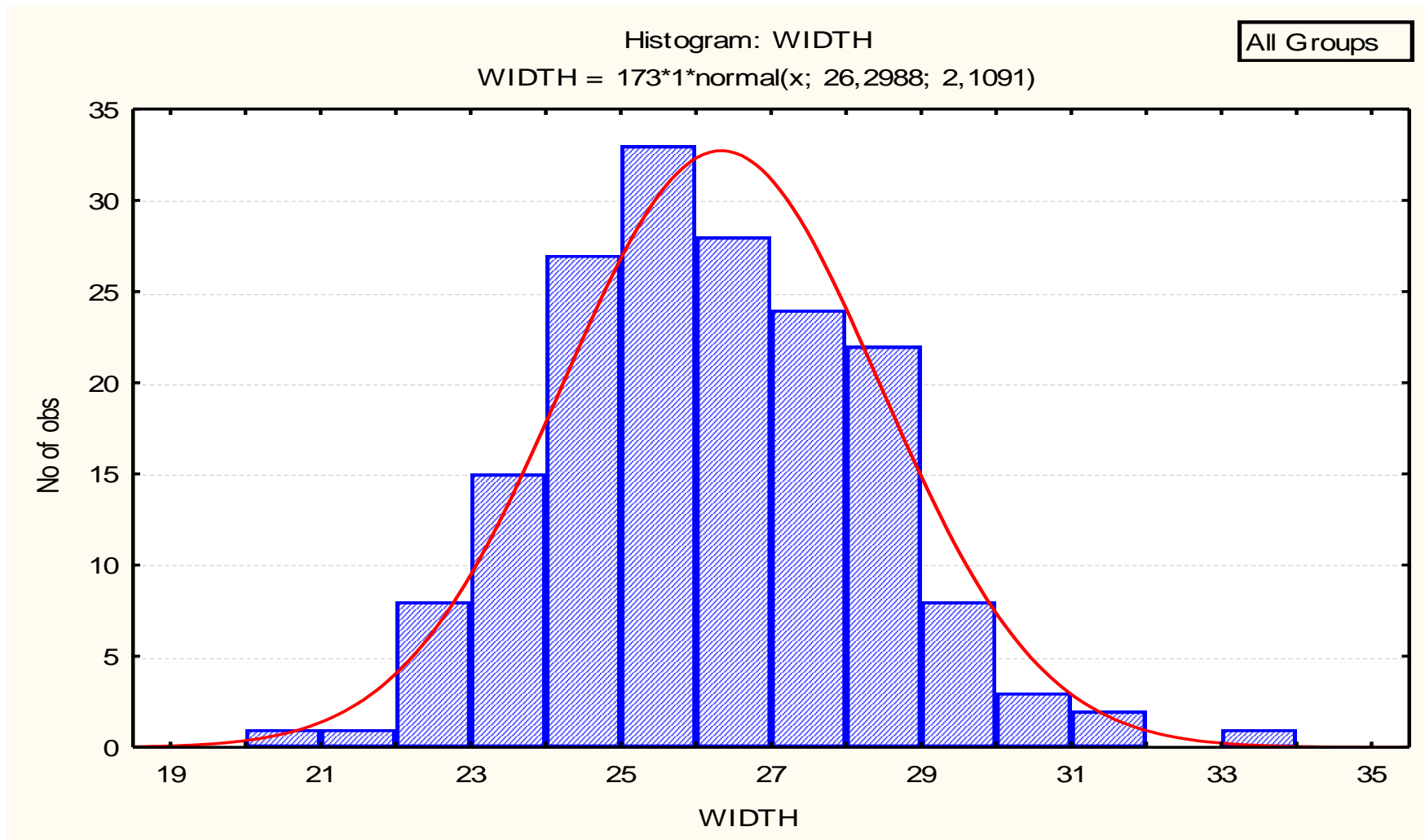


Рис. 15

Нажмем кнопку **Leven's test (ANOVA)**, появится таблица (рис. 16) с результатами проверки гипотезы об однородности дисперсий для зависимых переменных *WIDTH* и *WEGHT*. Из таблицы следует, что во всех группах, соответствующих уровням категориального предиктора *SPINE* дисперсии однородны, т.е. верна гипотеза о равенстве дисперсий.

Levene's Test for Homogeneity of Variance				
Effect: SPINE				
Degrees of freedom for all F's: 2, 170				
	MS Effect	MS Error	F	p
WIDTH	2,95	1,55	1,91	0,15
WEIGHT	0,30	0,11	2,65	0,07

Рис. 16

Еще одним дополнительным условием применимости дисперсионного анализа является отсутствие корреляции между средними и стандартными отклонениями. На вкладке **Assumptions** нажмем кнопку **Plot means vs.std.deviation**. Из диаграммы рассеяния, изображенной на рис. 17, видно, что средние и стандартные отклонения коррелируют незначительно.

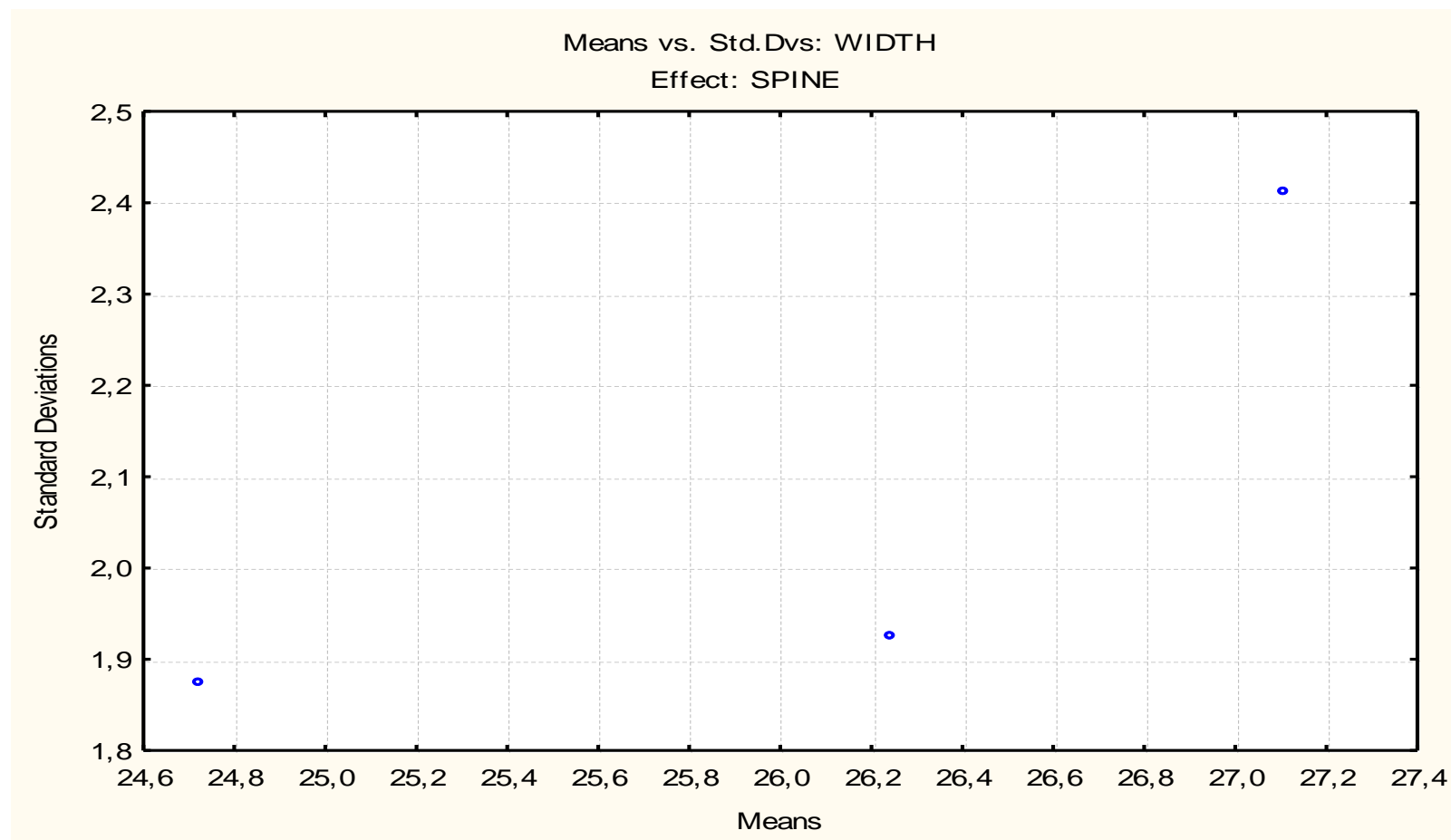


Рис. 17

Таким образом, основные условия применимости дисперсионного анализа выполнены, что подтверждает достоверность полученных результатов.