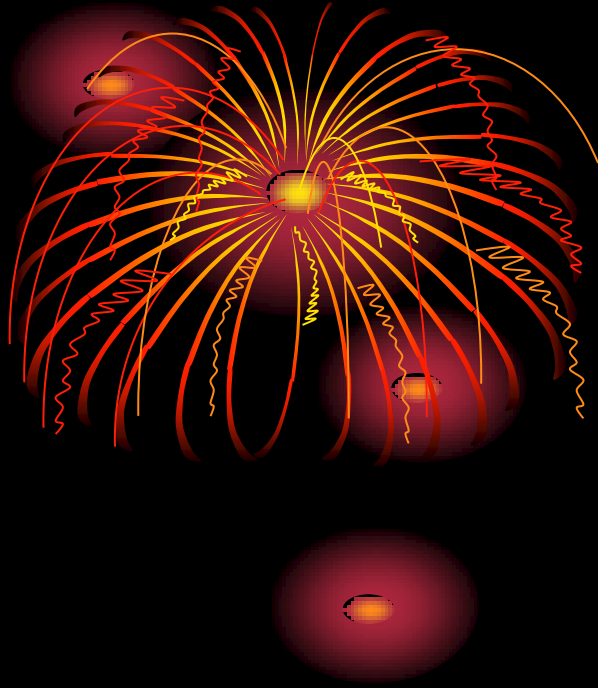
Лекция 6



Группировка и однофакторная ANOVA.

Как было сказано в предыдущей лекции для сравнения средних в двух группах применяют t-критерии Стьюдента. Это процедуры:

– **t-test, independent, by variables** (*t-критерий* для независимых выборок) применяется, если надо сравнить средние случайных величин, полученных по двум разным (независимым) выборкам;

– **t-test, independent, by groups** (*t-критерий* для независимых выборок с группирующей переменной) используется, если надо сравнить средние случайных величин двух независимых групп, полученных из одной выборки при помощи группирующей переменной;

– **t-test, dependent samples** (*t-критерий* для зависимых выборок) применяется, если надо сравнить средние случайных величин двух зависимых групп;

– t-test, single samples (простые выборки).

Напомним что в перечисленных процедурах в качестве нулевой гипотезы предполагается, что средние в группах равны.

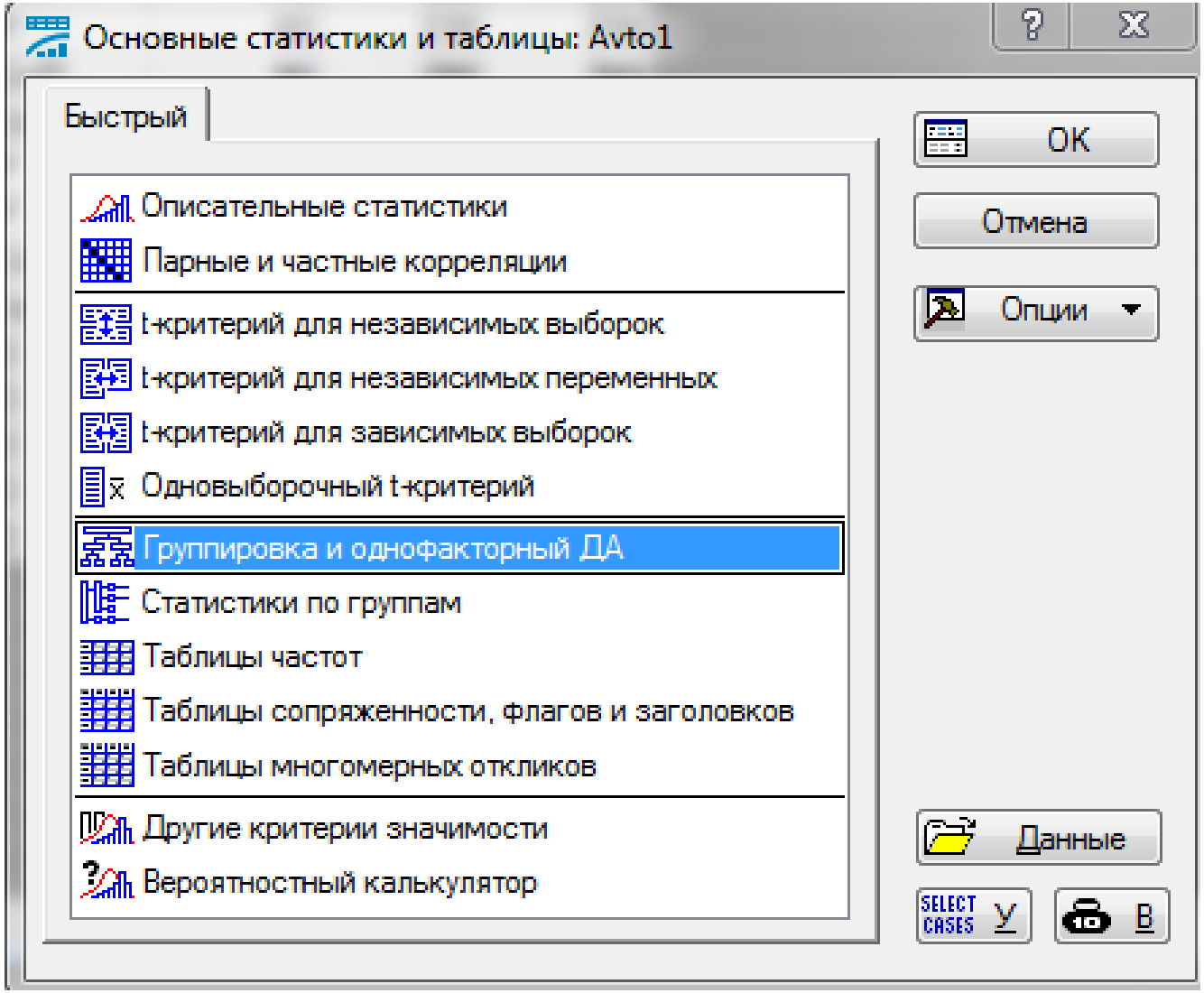
Модуль **Breakdown & one-way ANOVA** (группировка и однофакторный дисперсионный анализ *ANOVA*) определяет внутригрупповые описательные статистики и корреляции для зависимых переменных в каждой из нескольких групп, определенных группирующей переменной. Сравнивает средние и определяет, в каких именно группах отличие средних статистически значимо отличаются между собой. В качестве нулевой гипотезы предполагается, что средние в генеральной совокупности равны. В терминологии пакета группирующую переменную называют фактором, зависимыми называют количественные переменные, чьи средние следует сравнить.

Так как лекции составлялись по англоязычной версии пакета, а работать приходится в компьютерных классах на русскоязычной, то наряду со скринами на английском языке присутствуют скрины и на русском.

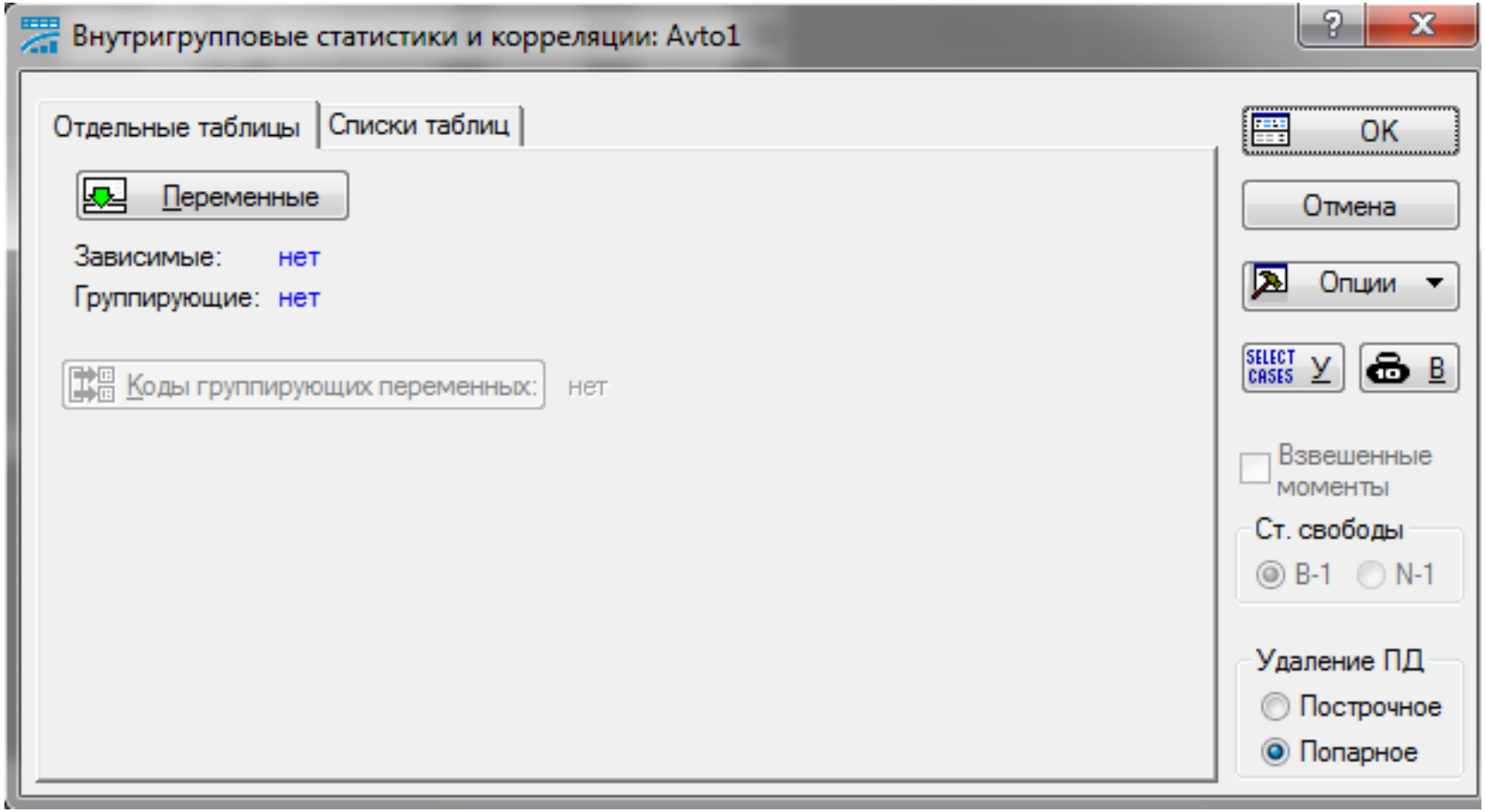
Работу данного модуля проследим на примере уже знакомых нам данных из файла **Auto1,** так как есть группирующая переменная Тип топлива, которая делит выборку более чем на 2 группы – 3 группы

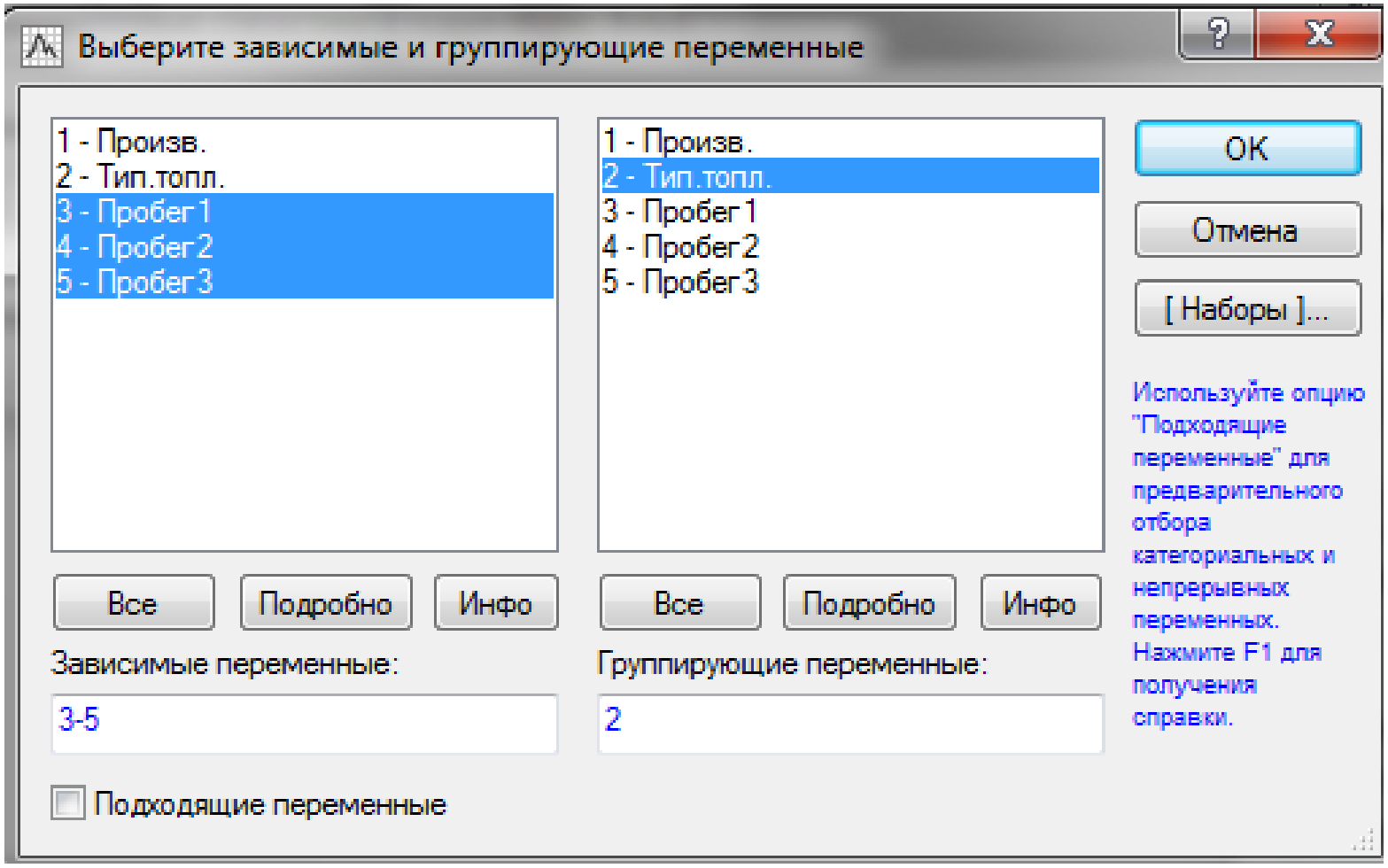
	Пробег до кап.рем., после кап. рем.				
	1 Произв.	2 Гип.топл	3 Пробег1	4 Пробег2	5 Пробег3
Opel Astra	Evropa	P	65	240	230
Skoda Fabia 1,2	Evropa	P	70	250	220
Mitsubishi Pinin	Japan	G+P	110	300	280
Skoda Ambiente 1,6	Evropa	P	60	230	230
Nissan Almera 1,5	Japan	G+P	90	280	260
Nissan Maxima 2,0 QX	Japan	G+P	100	300	280
Audi A4 2.0 MultiTronic	Evropa	P	80	250	230
Nissan Maxima 3,0 SE	Japan	P	110	310	310
Mitsubishi Pajero III	Japan	G+P	95	320	280
ToyotaCorolla	Japan	G+P	100	300	300
Toyota Carina	Japan	D	110	310	300
VW Passat1,8 T	Evropa	D	70	275	250
VW Bora 1,6	Evropa	D	80	260	230
Subaru Legacy	Japan	D	105	315	350
VW Golf 1,6	Evropa	D	75	250	240

Модуль **Breakdown & one-way ANOVA** (Группировка и однофакторный ДА) расположен в разделе (Basic statistics) Основные статистики, там же где и критерии Стьюдента.

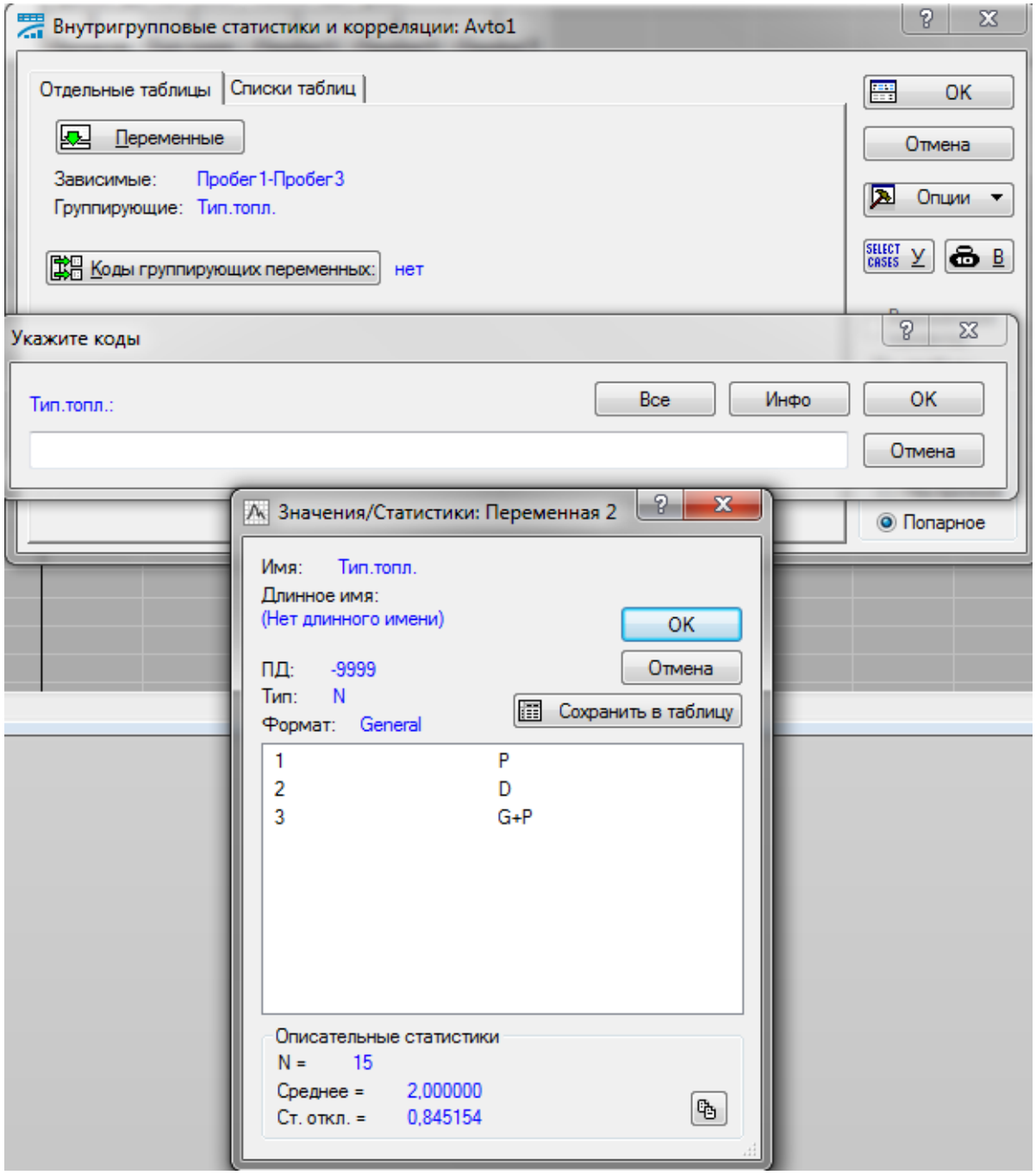


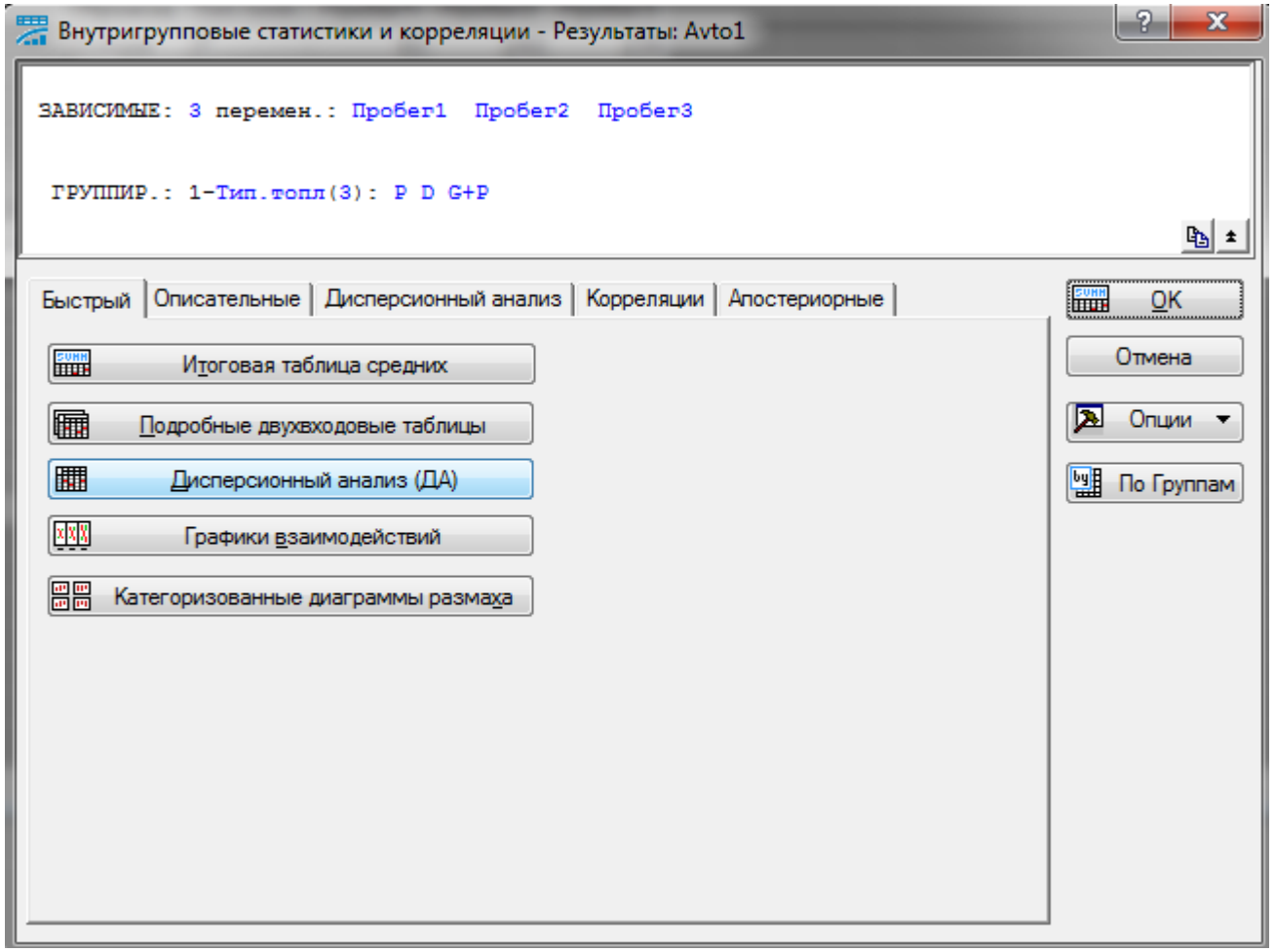
Выделим в модуле **Basic Statistic Tables** процедуру **Breakdown & one-way ANOVA,** откроется интерфейсное окно команды. Нажмите кнопку **Variables** и выберите **Grouping variables** (группирующие переменные) *Тип. топл.* и **Dependent variables** (зависимые переменные) *Пробег1 – Пробег3.*




Выберите зависимые и группирующие переменные
1 - Произв.
2 - Тип.топл.
3 - Пробег1
4 - Пробег2
5 - Пробег3
1 - Произв.
2 - Тип.топл.
3 - Пробег1
4 - Пробег2
5 - Пробег3
OK
Отмена
[Наборы]...
Используйте опцию "Подходящие переменные" для предварительного отбора категориальных и непрерывных переменных. Нажмите F1 для получения справки.
Все
Подробно
Инфо
Все
Подробно
Инфо
Зависимые переменные:
3-5
Группирующие переменные:
2
Подходящие переменные

Если мы проигнорируем кнопку **Коды группирующих переменных,** то программа автоматически в анализе использует все группы, в нашем случае их 3 по типу топлива. Если нас какая либо из групп не интересует, то следует нажать на кнопку, в появившемся окне воспользоваться кнопкой **Инфо** и выделить интересующие нас значения категориальной переменной. Если выбрать любые 2 группы, то результаты анализа совпадут с критерием Стьюдента **t-test, independent, by groups.** Если в правом верхнем углу диалога щелкнуть по кнопке ОК, то программа перейдет в окно результатов



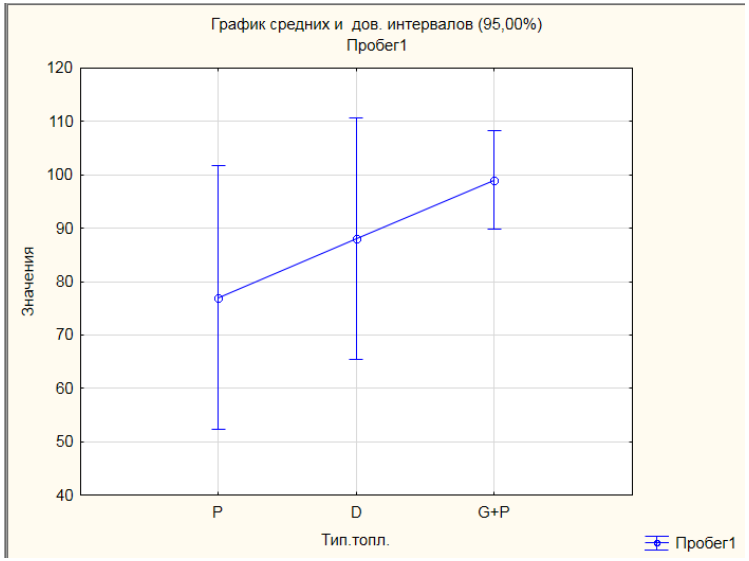


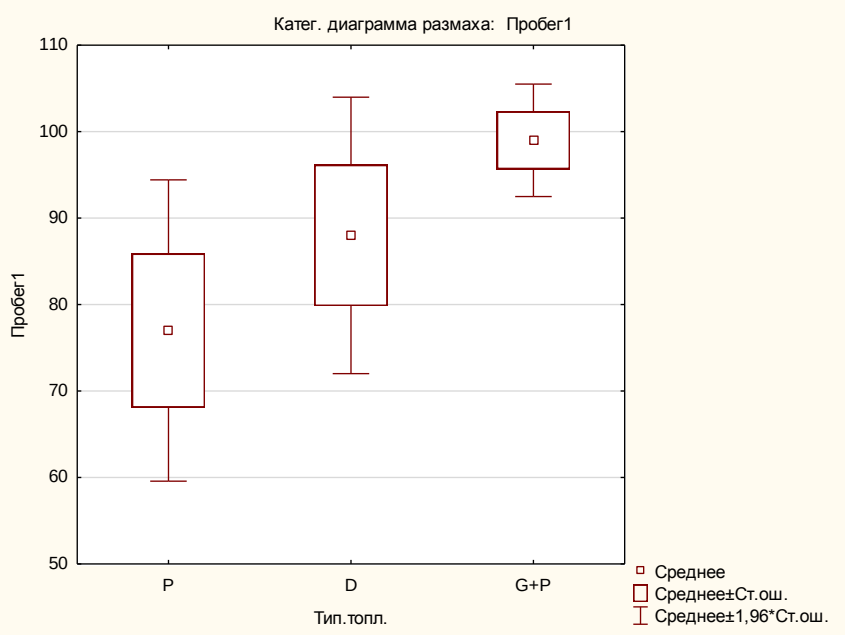
Щелкнем последовательно по кнопам **Итоговая таблица средних**, **График взаимодействий** и **Категорированная диаграмма размаха**. В данном случае кнопка **Подробные двухвходовые таблицы** строит таблицу идентичную таблице, построенной кнопкой **Итоговая таблица средних**

Итоговая таблица средних (Avto1)
N=15 (Нет пропусков в завис. перем.)

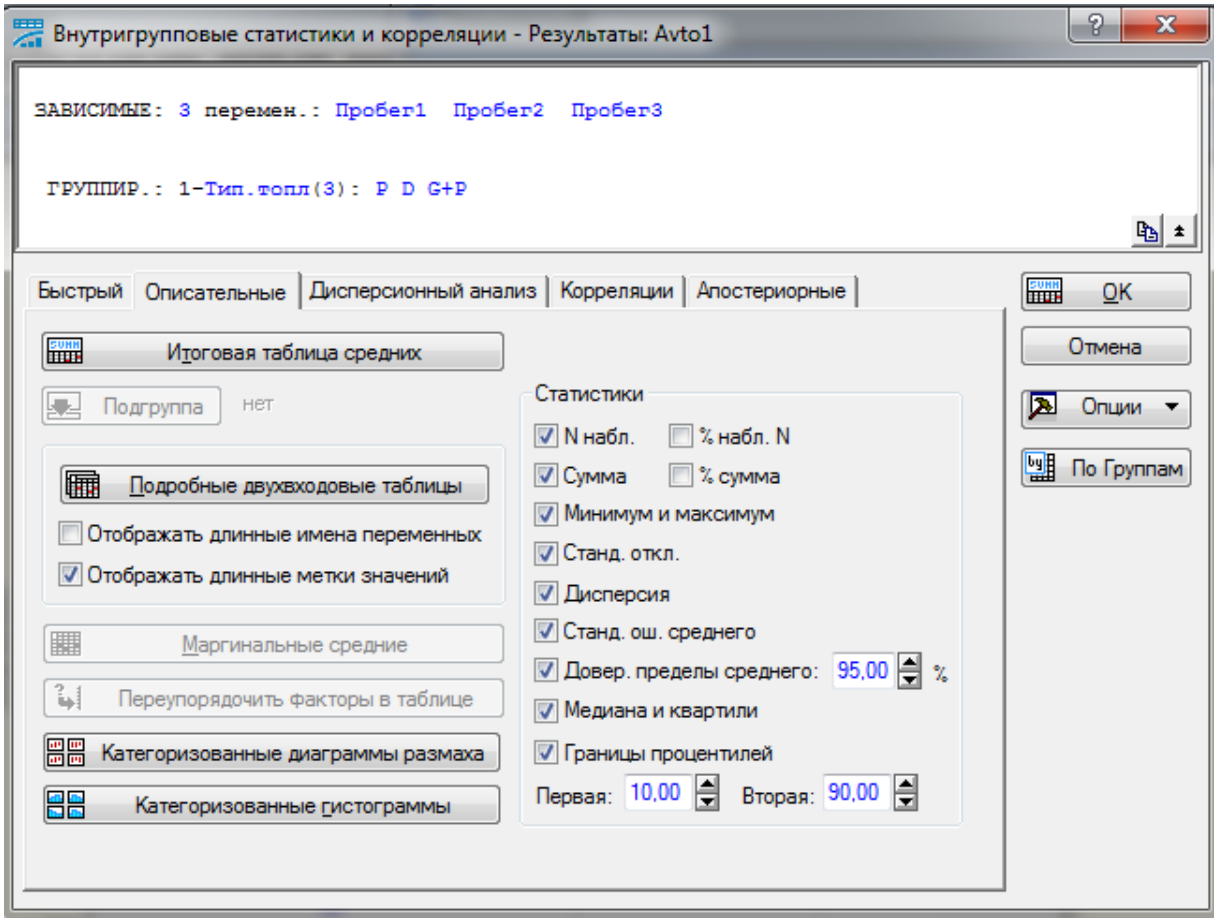
Тип.топл.	Пробег1 Среднее	Пробег1 N	Пробег1 Ст.откл.	Пробег2 Среднее	Пробег2 N	Пробег2 Ст.откл.	Пробег3 Среднее	Пробег3 N	Пробег3 Ст.откл.
P	77,00000	5	19,87461	256,0000	5	31,30495	244,0000	5	37,14835
D	88,00000	5	18,23458	282,0000	5	29,28310	274,0000	5	50,29911
G+P	99,00000	5	7,41620	300,0000	5	14,14214	280,0000	5	14,14214
Всего	88,00000	15	17,60682	279,3333	15	30,52322	266,0000	15	37,94733

В таблице отображены средние значения, объемы групп, стандартные отклонения для все выбранных зависимых переменных *Пробег1-Пробег2* в 3-х группах. На рисунках то же самое в графическом представлении



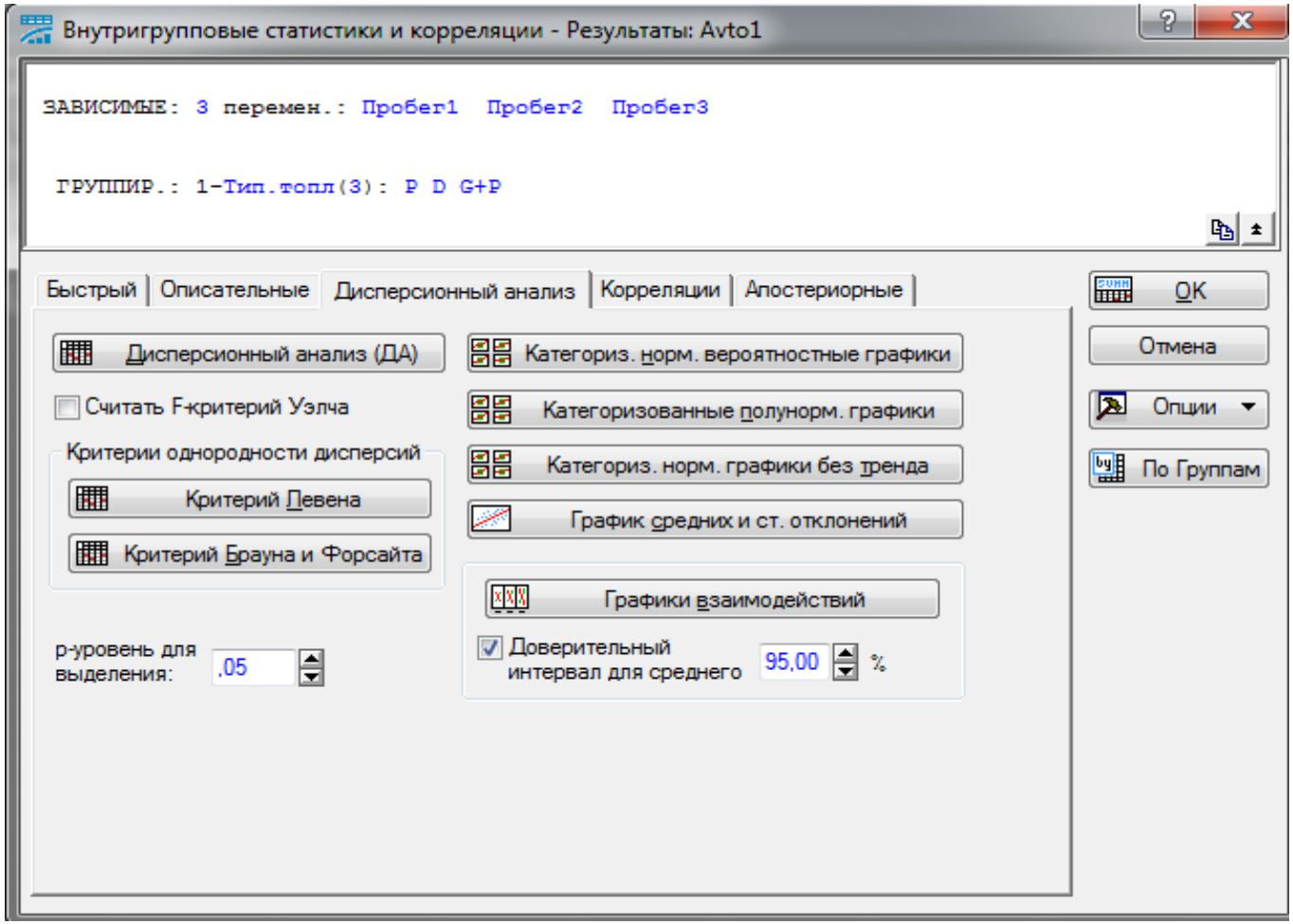


Можно на вкладке **Описательные** в правой части окна добавить дополнительные описательные статистики и щелкнуть по кнопке **Итоговая таблица средних.** Появится таблица с вычисленными описательными статистиками для всех переменных



Тип.топл.	Пробег1 Среднее	Доверит. -95,000%	Доверит. +95,000%	Пробег1 N	Пробег1 Сумма	Пробег1 Ст.откл.	Пробег1 Дисперсия	Пробег1 Стд.ош.	Пробег1 Минимум	Пробег1 Максим.	Пробег1 25%	Пробег1 Медиана
P	77,00000	52,32242	101,6776	5	385,000	19,87461	395,0000	8,888194	60,00000	110,0000	65,00000	70,0000
D	88,00000	65,35878	110,6412	5	440,000	18,23458	332,5000	8,154753	70,00000	110,0000	75,00000	80,0000
G+P	99,00000	89,79157	108,2084	5	495,000	7,41620	55,0000	3,316625	90,00000	110,0000	95,00000	100,0000
Всего	88,00000	78,24967	97,7503	15	1320,000	17,60682	310,0000	4,546061	60,00000	110,0000	70,00000	90,0000

На вкладке **Дисперсионный анализ** можно посмотреть результаты дисперсионного анализа, проверяется гипотеза о равенстве средних во всех (у нас 3) группах. Нулевая гипотеза: средние во всех группах равны. Если p критерия Фишера Дисперсионного анализа <0,05, то это означает, что могут быть группы объектов, для которых не верна гипотеза о равенстве средних! Критерии Левена и Брауна-Форсайта проверяют гипотезу о равенстве (однородности) дисперсий в 3 группах. Нулевая гипотеза: дисперсии в группах равны. Нажмем последовательно на 3 кнопки, соответствующие 3 критериям.



Дисперсионный анализ (Avto1)
Отмечены эффекты,значимые на уров. p < ,05000

Переменная	Сум.квад эффект	Ст.св. эффект	Ср.квад. эффект	Сум.квад ошибки	Ст.св. ошибки	Ср.квад. ошибки	F	p
Пробег1	1210,000	2	605,000	3130,00	12	260,833	2,319489	0,140711
Пробег2	4893,333	2	2446,667	8150,00	12	679,167	3,602454	0,059513
Пробег3	3720,000	2	1860,000	16440,00	12	1370,000	1,357664	0,294082

Критерий Левена однородности дисперсий (Avto1)
Отмечены эффекты,значимые на уров. p < ,05000

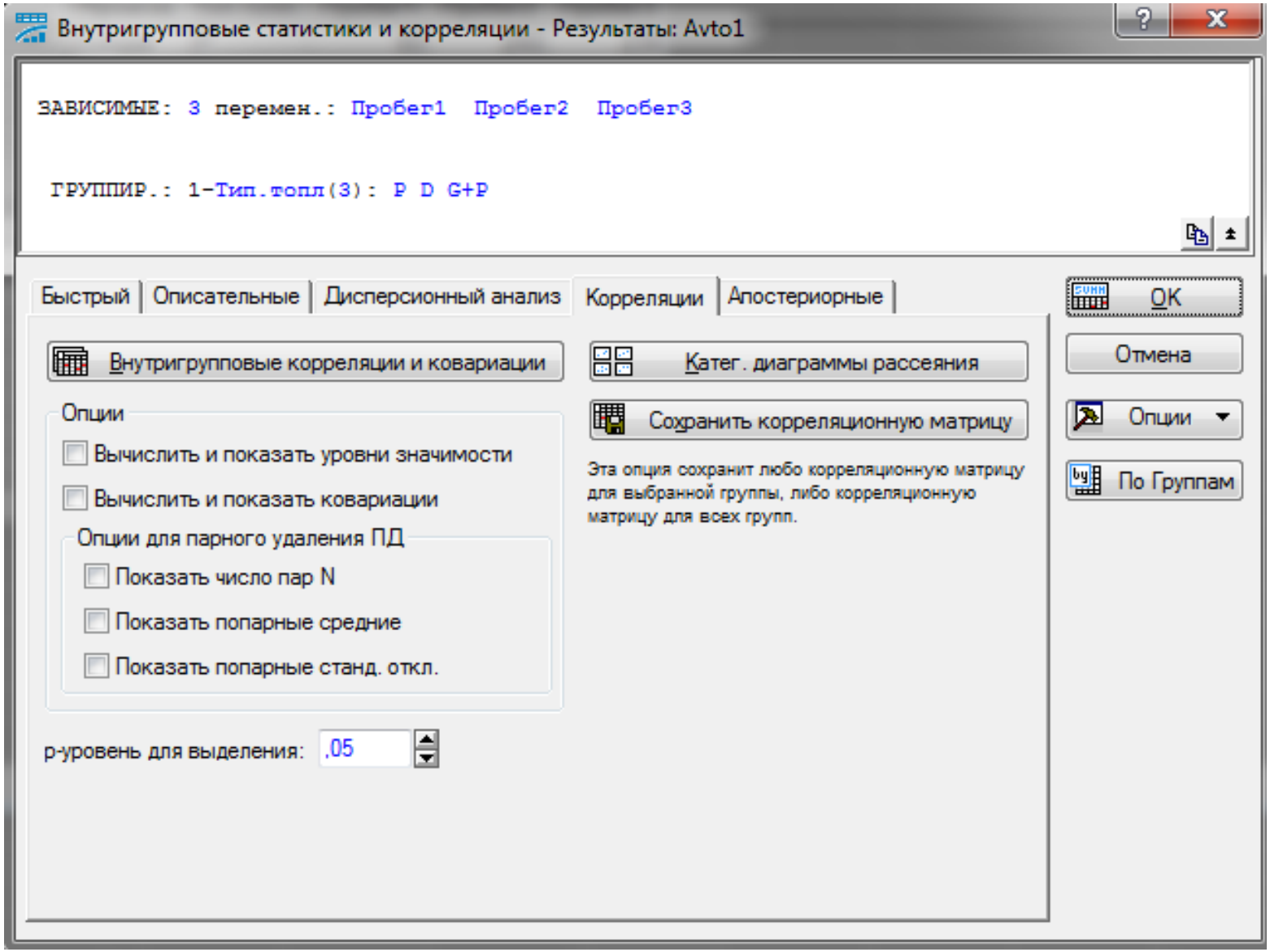
Переменная	Сум.квад эффект	Ст.св. эффект	Ср.квад. эффект	Сум.квад ошибки	Ст.св. ошибки	Ср.квад. ошибки	F	p
Пробег1	323,733	2	161,867	741,200	12	61,7667	2,620615	0,113678
Пробег2	769,600	2	384,800	2520,400	12	210,0333	1,832090	0,202135
Пробег3	2702,933	2	1351,467	4312,000	12	359,3333	3,761039	0,053943

Брауна-Форсайта критерий однород. дисперсий (Avto1)
Отмечены эффекты,значимые на уров. p < ,05000

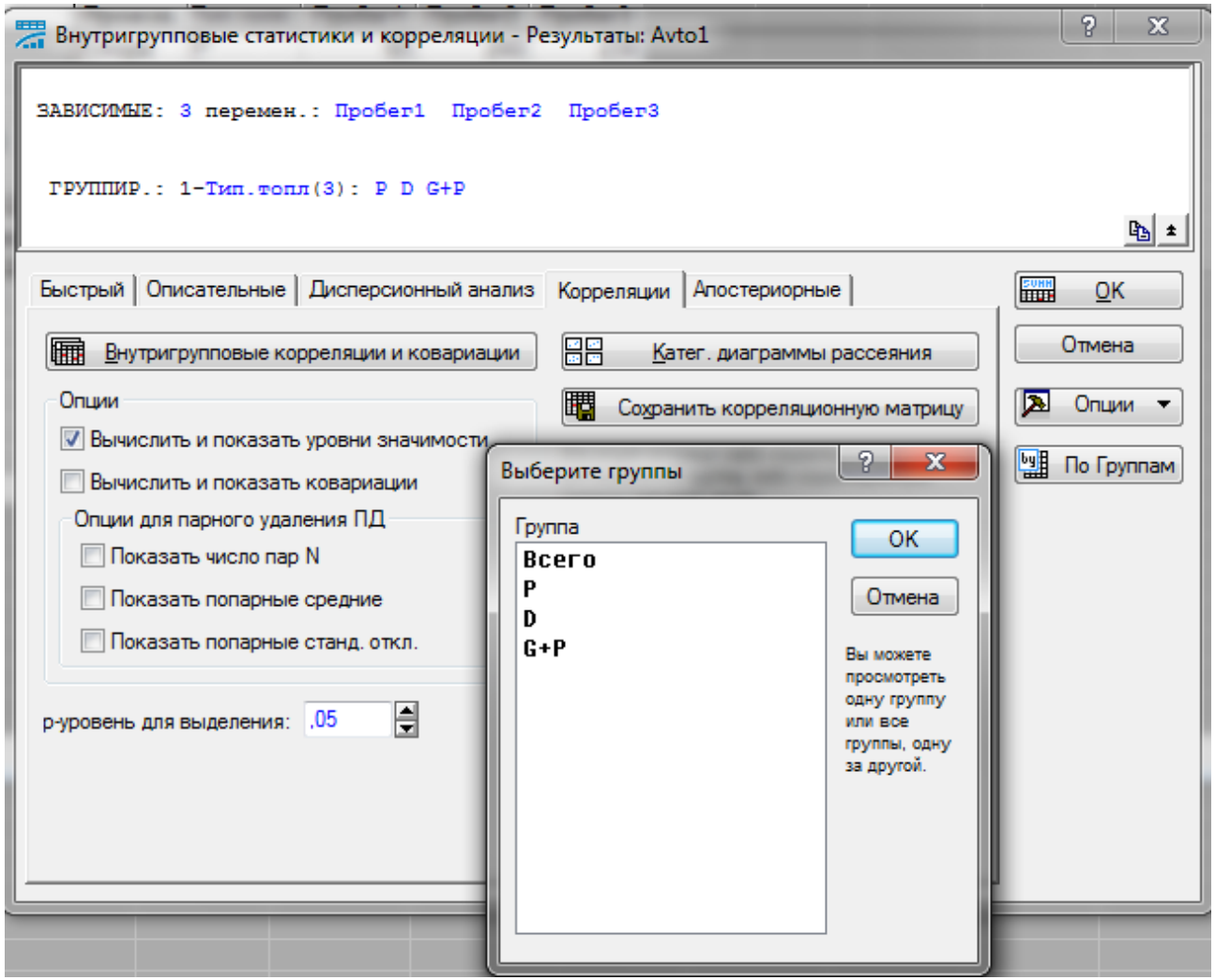
Переменная	Сум.квад эффект	Ст.св. эффект	Ср.квад. эффект	Сум.квад ошибки	Ст.св. ошибки	Ср.квад. ошибки	F	p
Пробег1	243,333	2	121,667	1750,00	12	145,8333	0,834286	0,457877
Пробег2	583,333	2	291,667	3990,00	12	332,5000	0,877193	0,441002
Пробег3	2013,333	2	1006,667	11880,00	12	990,0000	1,016835	0,390895

Из первой таблицы следует, что по критерию Фишера верна гипотеза о равенстве средних генеральных совокупностей во всех 3 группах, но для *Пробег 2* уровень p критерия Фишера принимает близкое к 0,05 значений, поэтому могут быть группы для которых гипотеза не верна. Из 2-й и 3-й таблиц следует, что верны гипотезы об однородности дисперсий.

На вкладке Корреляции можно посмотреть корреляции между переменными отдельно в группах объектов



Если нажать на кнопку Внутригрупповые корреляции, то откроется окно с предложением выбора группы. В рамке опции можно указать дополнительную информацию к коэффициентам корреляции. Выберем, например последовательно P, D, G+P



В первой таблице отображены коэффициенты корреляции Пирсона, во второй их уровни значимости

Переменные	Внутригрупповые корреляции (Avto1) Группа:Тип.топл.:P Отмеченные корреляции значимы на уровне p <,05000		
	Пробег1	Пробег2	Пробег3
Пробег1	1,000000	0,980434	0,917638
Пробег2	0,980434	1,000000	0,941589
Пробег3	0,917638	0,941589	1,000000

Переменные	p-уровни для внутригрупповых корр. (Avto1) Группа:Тип.топл.:P Отмеченные корреляции значимы на уровне p <,05000		
	Пробег1	Пробег2	Пробег3
Пробег1		0,003276	0,028021
Пробег2	0,003276		0,016797
Пробег3	0,028021	0,016797	

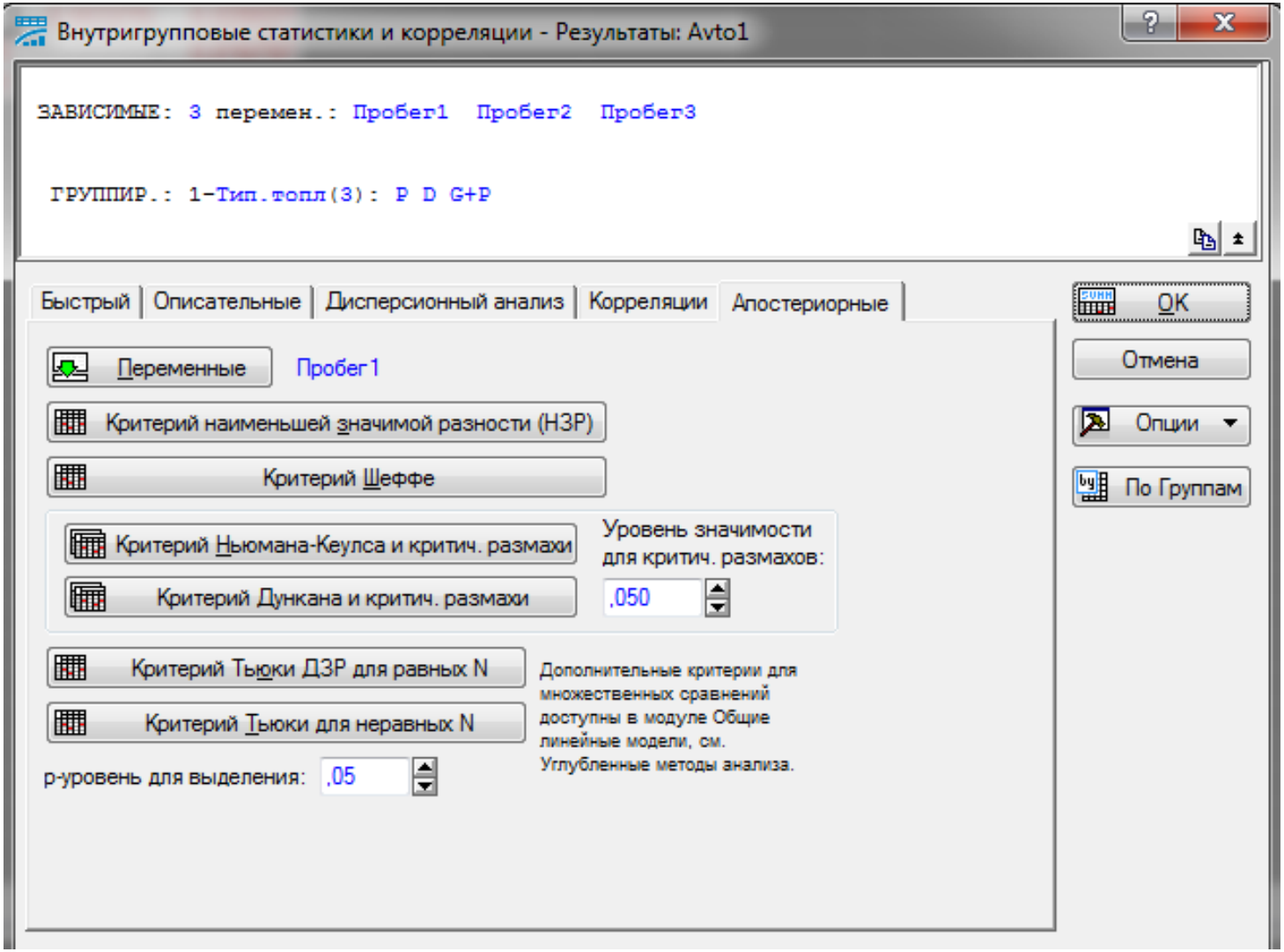
Переменные	Внутригрупповые корреляции (Avto1) Группа:Тип.топл.:D Отмеченные корреляции значимы на уровне p <,05000		
	Пробег1	Пробег2	Пробег3
Пробег1	1,000000	0,887230	0,842253
Пробег2	0,887230	1,000000	0,926734
Пробег3	0,842253	0,926734	1,000000

Переменные	p-уровни для внутригрупповых корр. (Avto1) Группа:Тип.топл.:D Отмеченные корреляции значимы на уровне p <,05000		
	Пробег1	Пробег2	Пробег3
Пробег1		0,044682	0,073405
Пробег2	0,044682		0,023542
Пробег3	0,073405	0,023542	

Переменные	Внутригрупповые корреляции (Avto1) Группа:Тип.топл.:G+P Отмеченные корреляции значимы на уровне p <,05000		
	Пробег1	Пробег2	Пробег3
Пробег1	1,000000	0,238366	0,476731
Пробег2	0,238366	1,000000	0,500000
Пробег3	0,476731	0,500000	1,000000

Переменные	p-уровни для внутригрупповых корр. (Avto1) Группа:Тип.топл.:G+P Отмеченные корреляции значимы на уровне p <,05000		
	Пробег1	Пробег2	Пробег3
Пробег1		0,699402	0,416855
Пробег2	0,699402		0,391002
Пробег3	0,416855	0,391002	

Для решения ключевой задачи дисперсионного анализа, следует проверить статистическую значимость отличия средних значений переменных в группах при помощи критериев на вкладке **Апостериорные.** Проверим для *Пробег 1*



Обратите внимание, что на слайде 13 показано, что по критерию Фишера для переменной *Пробег1* верна гипотеза о равенстве средних в генеральных совокупностях. Все 3 критерия в приведенных ниже таблицах показывают статистическую незначимость отличия средних в группах, что подтверждает справедливость вывода по дисперсионному анализу!

Крит. НЗР; перем.: Пробег1 (Avto1)
Отмечены разности, значимые на уров

Тип.топл.	{1} M=77,000	{2} M=88,000	{3} M=99,000
P {1}		0,302693	0,052285
D {2}	0,302693		0,302693
G+P {3}	0,052285	0,302693	

Крит.Шеффе; Переменная: Пробег1 (Av
Отмечены разности, значимые на уров

Тип.топл.	{1} M=77,000	{2} M=88,000	{3} M=99,000
P {1}		0,574914	0,140711
D {2}	0,574914		0,574914
G+P {3}	0,140711	0,574914	

Крит.Тьюки ДЗР; Перемен.: Пробег1 (A
Отмечены разности, значимые на уров

Тип.топл.	{1} M=77,000	{2} M=88,000	{3} M=99,000
P {1}		0,545556	0,120525
D {2}	0,545556		0,545556
G+P {3}	0,120525	0,545556	

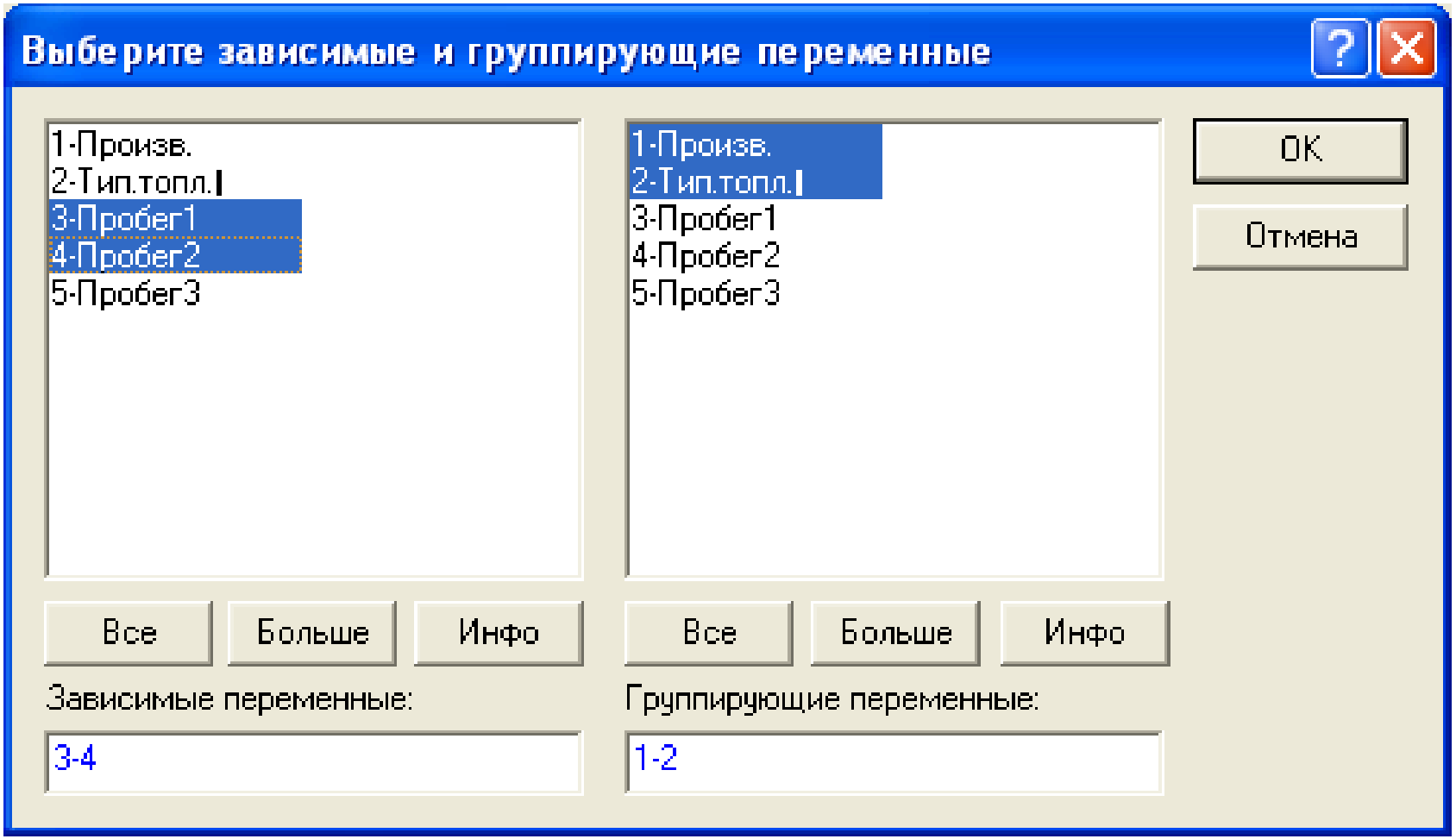
Несколько иная картина для *Переменной 2*, есть подгруппы в которых отличие средних статистически значимо сразу по 2 критериям! При этом на слайде 13 в результатах дисперсионного анализа уровень значимости критерия Фишера чуть больше 0,05!

Тип топлива		Крит. НЗР; перем.: Пробег2 (Avto1) Отмечены разности, значимые на уровне p < ,05000		
		1 M=256,00	2 M=282,00	3 M=300,00
P	1		0,140676	0,020429
D	2	0,140676		0,296239
G+P	3	0,020429	0,296239	

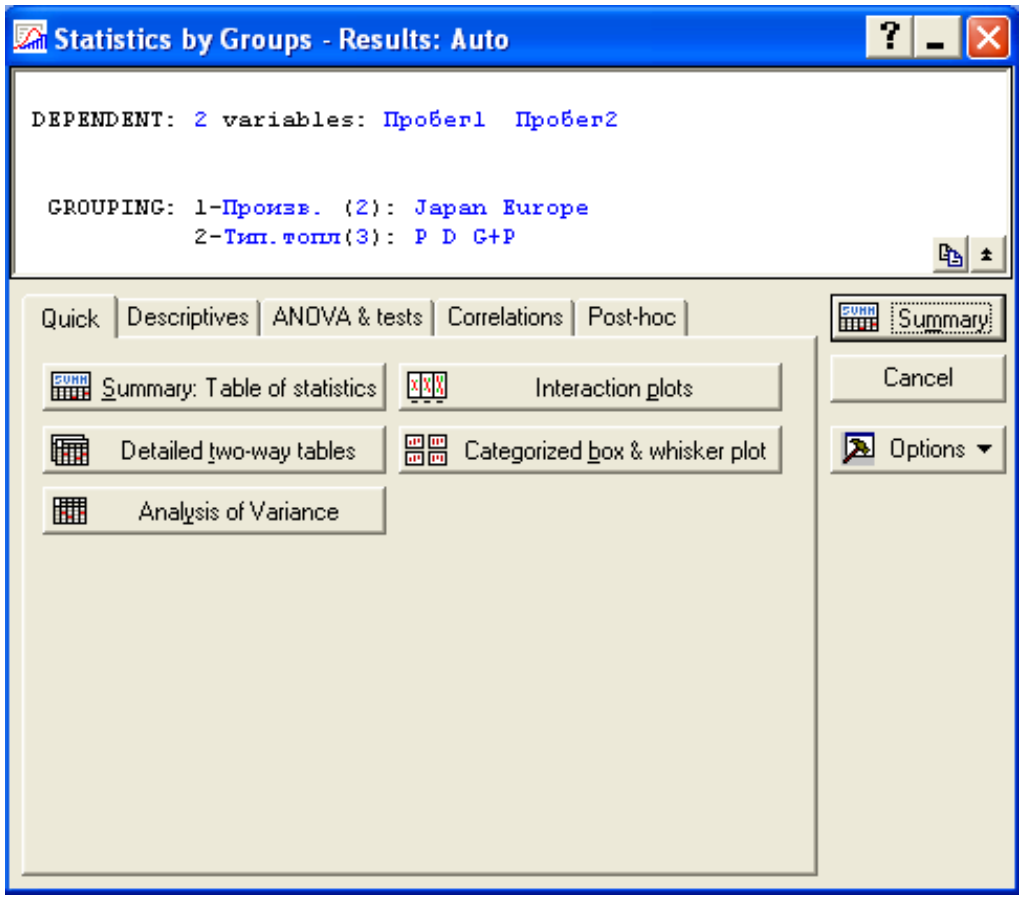
Тип топлива		Крит.Шеффе; Переменная: Пробег2 Отмечены разности, значимые на уровне p < ,05000		
		1 M=256,00	2 M=282,00	3 M=300,00
P	1		0,322831	0,060995
D	2	0,322831		0,566367
G+P	3	0,060995	0,566367	

Тип топлива		Крит.Тьюки ДЗР; Перемен.: Пробег2 Отмечены разности, значимые на уровне p < ,05000		
		1 M=256,00	2 M=282,00	3 M=300,00
P	1		0,292363	0,049971
D	2	0,292363		0,536745
G+P	3	0,049971	0,536745	

Но однофакторный дисперсионный анализ может быть использован когда объекты разбиваем на группы по нескольким группирующим переменным. Выберем в поле **Группирующие переменные** *Производитель* и *Тип топлива*



Если нажать на **OK** в диалоговом окне **Statistics by Groups (Breakdown)**, то откроется диалоговое окно **Statistics by Groups-Results** (внутригрупповые описательные статистики – результаты), которое предоставляет различные процедуры и настройки для анализа данных внутри групп. Цель такого анализа – лучшее понимание различий между группами. Информационная часть окна сообщает, что зависимых – две переменные: *Пробег1*, *Пробег2*; группирующих – две переменные: *Произв.* с двумя кодами (*Europe*, *Japan*) и *Тип. топл.* с тремя кодами (*P*, *G + P*, *D*). На рисунке активизирована вкладка **Quick.**



Если нажать на **Summary: Table of statistics**, появится таблица результатов. В приведенной таблице имеются описательные статистики для выбранных переменных, разбитых на 6 групп. Так в столбцах 1 и 2 отображены подгруппы, в столбцах 3 и 5 показаны средние (*means*) переменных *Пробег1*, *Пробег2*, в столбцах 4 и 7 – количество автомобилей, в столбцах 5, 8 – среднеквадратические отклонения *(Std.Dev).*

Итоговая таблица средних (Avto1)
N=15 (Нет пропусков в завис. перем.)

Тип.топл.	Произв.	Пробег1 Среднее	Пробег1 N	Пробег1 Ст.откл.	Пробег2 Среднее	Пробег2 N	Пробег2 Ст.откл.
P	Japan	110,0000	1		310,0000	1	
P	Evropa	68,7500	4	8,53913	242,5000	4	9,57427
D	Japan	107,5000	2	3,53553	312,5000	2	3,53553
D	Evropa	75,0000	3	5,00000	261,6667	3	12,58306
G+P	Japan	99,0000	5	7,41620	300,0000	5	14,14214
G+P	Evropa		0			0	
Все груп.		88,0000	15	17,60682	279,3333	15	30,52322

Если нажать на кнопку Подробные двухвходовые таблицы, то появится другая форма таблицы:

Тип.топл. Произв.	Подробные двухвходовые таблицы (Avto1) N=15 (Нет пропусков в завис. перем.) Пробег1 Среднее	Пробег1 N	Пробег1 Ст.откл.	Пробег2 Среднее	Пробег2 N	Пробег2 Ст.откл.
P	77,0000	5	19,87461	256,0000	5	31,30495
Japan	110,0000	1	0,00000	310,0000	1	0,00000
Evropa	68,7500	4	8,53913	242,5000	4	9,57427
D	88,0000	5	18,23458	282,0000	5	29,28310
Japan	107,5000	2	3,53553	312,5000	2	3,53553
Evropa	75,0000	3	5,00000	261,6667	3	12,58306
G+P	99,0000	5	7,41620	300,0000	5	14,14214
Japan	99,0000	5	7,41620	300,0000	5	14,14214
Evropa		0			0	
Все груп.	88,0000	15	17,60682	279,3333	15	30,52322

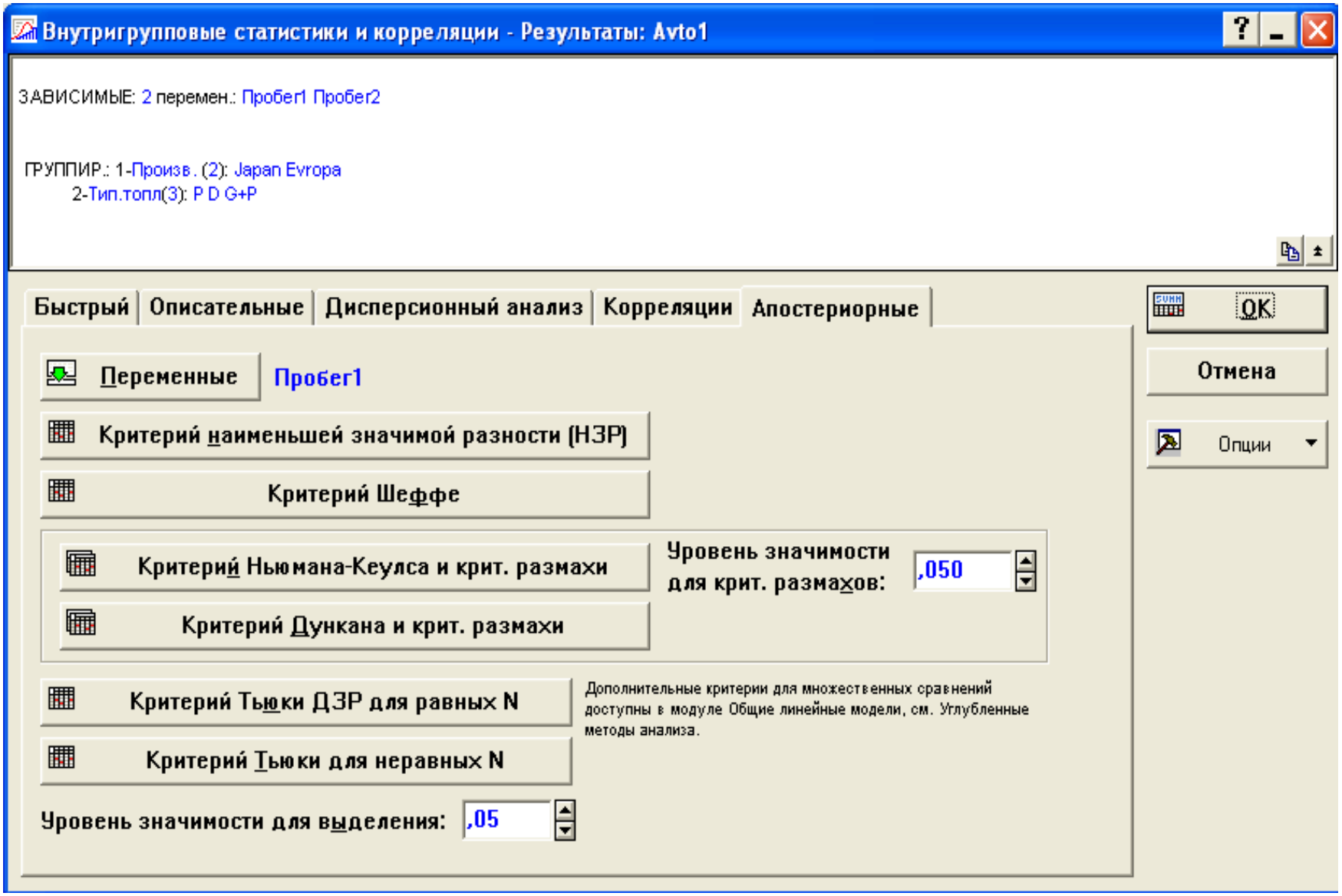
Для проверки гипотезы о равенстве средних в 6 группах генеральной совокупности надо использовать процедуру **Analysis of Variance** (анализ дисперсий). Щелкнем кнопкой **Analysis of Variance** на вкладке **ANOVA & tests**. Откроется таблица результатов **Analysis of Variance** (рис. 5). Из таблицы видно, что можно отвергнуть гипотезу о равенстве средних переменных *Пробег1*, *Пробег2* в группах. Так как число групп более двух, то из таблицы не видно, какие группы вызвали статистически значимое отличие средних. Процедура **Post-hoc** (апостериорные сравнения средних) позволяет устранить этот недостаток.

Analysis of Variance (Auto)
Marked effects are significant at p < ,05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error	MS Error	F	p
Пробег1	3838,7	4	959,68	501,25	10	50,125	19,1458	0,00011
Пробег2	11639,1	4	2909,79	1404,16	10	140,416	20,7225	0,00007

На вкладке реализовано 6 критериев:

– **LSD test or planned comparison** (критерий наименьшей значимости (НЗР));
– **Scheffe test** (критерий Шеффе);
– **Newman – Keuls test & critical ranges** (критерий Ньюмана – Кеулса и критические размахи);
– **Duncan's multiple range test & critical ranges** (критерий Дункана и критические размахи);
– **Tukey honest significant difference (HSD)** (критерий Тьюки ДЗР);
– **Tukey HSD for unequal N (Spjotvoll/Stoline)** (критерий Тьюки ДЗР для неравных *N*).



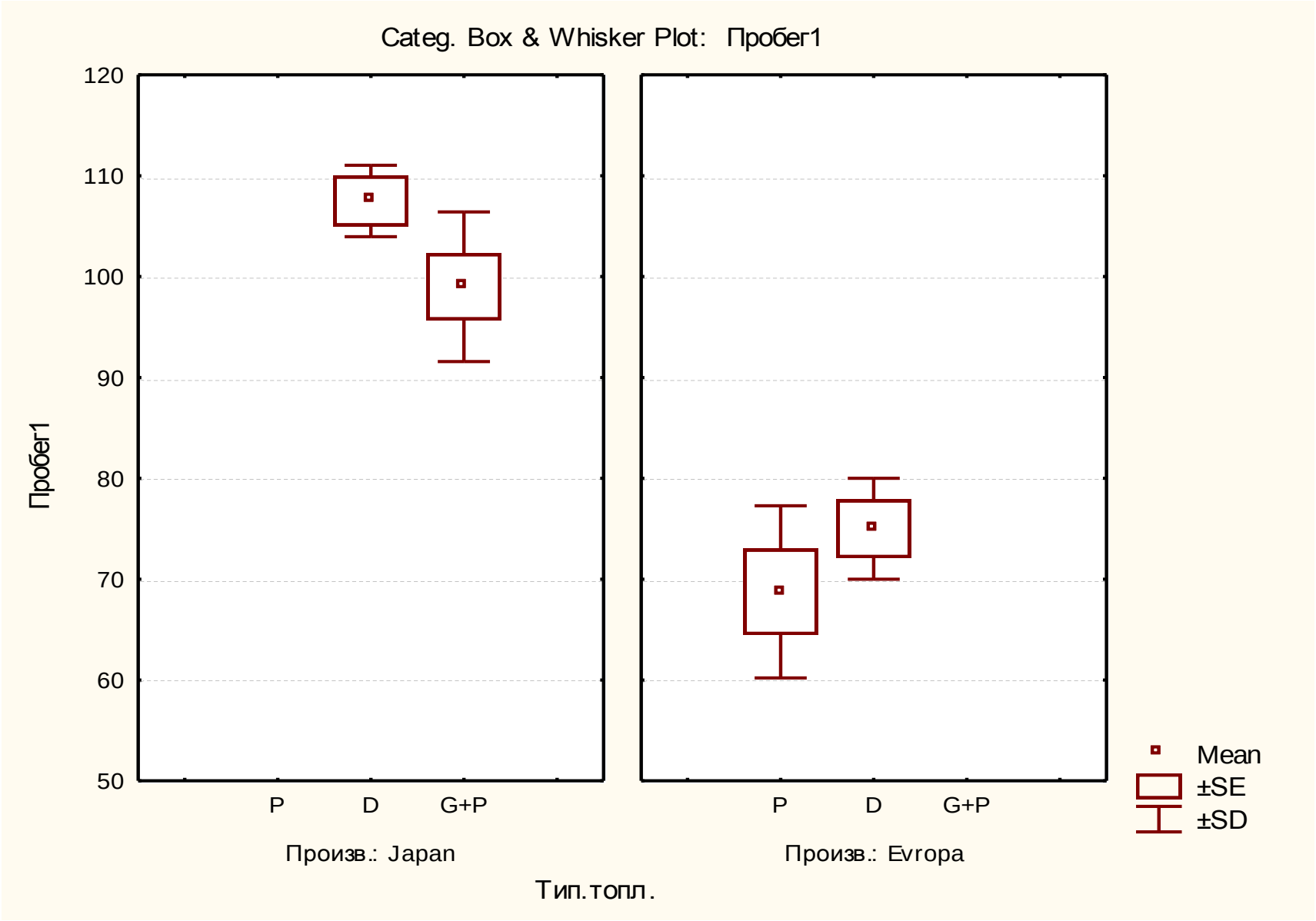
В нижней части окна пользователь может назначить **p-level for highlighting** (*p-уровень* значимости для выделения).

Если воспользоваться кнопкой **Критерий наименьшей значимой разности p**, то программа построит таблицу с уровнями значимости критерия, из которой видно между какими подгруппами выполняется условие статистической значимости отличия средних ($p<0,05$) и статистической не значимости ($p\geq0,05$), то есть верности нулевой гипотезы

	LSD Test; Variable: Пробег1 (Auto) Marked differences are significant at p < ,05000					
Произв. Тип.топл	{1} M=110,0(	{2} M=107,5(	{3} M=99,00(	{4} M=68,75(	{5} M=75,00(	{6} M=0,0
JapanP {1}		0,77899	0,18650	0,00039	0,00160	
JapanD {2}	0,77899		0,18182	0,00008	0,00051	
JapanG+P {3}	0,18650	0,18182		0,00008	0,00091	
EuropeP {4}	0,00039	0,00008	0,00008		0,27461	
EuropeD {5}	0,00160	0,00051	0,00091	0,27461		
EuropeG+P {6}						

Из таблицы видно, что верна гипотеза о равенстве средних в группах: {1, 2}; {1, 3}; {2, 3}; {4, 5}. Не верна гипотеза о равенстве средних в группах: {4, 1};{4, 2};{4, 3}; {5, 1};{5, 2};{5, 3}.

Различия средних можно увидеть на графиках, доступных в диалоговом окне **Statistics by Groups-Results**. Например, щелкните по кнопке **Categorized box & whisker plot**, которая находится на вкладках **Descriptives или Quick**. Откроется диалоговое окно **Box-Whisker Type.** В этом окне выделите одну из опций, например *Mean/SE/SD*. Программа построит диаграммы размаха (рис. 8), визуализирующие степень сходства и различия средних в анализируемых группах. Из приведенных результатов можно сделать вывод применительно к генеральной совокупнсти, что средний пробег японских автомобилей до обращения на СТО примерно одинаков для различных типов топлива. Аналогичный вывод справедлив для автомобилей европейского производства. Но средний пробег японских автомобилей в среднем больше пробега европейских автомобилей для любых типов топлива. Другими словами, пробег автомобилей до обращения на СТО не зависит от типа топлива, но зависит от страны производителя.


Categ. Box & Whisker Plot: Пробег1
120
110
100
90
80
70
60
50
Пробег1
P
D
G+P
P
D
G+P
Произв.: Japan
Произв.: Evropa
Тип.топл.
Mean
±SE
±SD