

ЛАБОРАТОРНАЯ РАБОТА № 6 ОДНОФАКТОРНЫЙ АНАЛИЗ

При исследовании зависимостей одной из наиболее простых является ситуация, когда можно указать **только один фактор**, влияющий на конечный результат, и этот фактор может принимать лишь конечное число значений (уровней). Такие задачи (называемые задачами однофакторного анализа) весьма часто встречаются на практике. Типичный пример — сравнение по достигаемым результатам нескольких различных способов действия, направленных на достижение одной цели, скажем, различных лекарств.

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Однофакторный дисперсионный анализ применяется для обнаружения влияния выделенного (контролируемого) набора факторов на результативный признак в случае, если величины имеют нормальное распределение с общей для всех дисперсией, которая нам неизвестна.

Задача 1. Для проверки того, влияет ли день недели на объем торгового оборота в универсаме, было проведено обследование оборота в выбранные наугад дни. Полученные данные приведены в следующей таблице (цифры условные).

Понедельник	Вторник	Среда	Четверг	Пятница	Суббота
1,2	1,1	1,5	1,6	1,2	1,5
1,0	1,4	1,0	1,4	1,2	2,0
0,8	1,1	1,1	1,5	1,8	1,9
1,1	1,2	1,3	1,3	1,3	1,7
1,3	1,5	1,2	1,4	1,4	1,6
1,0	1,8	1,3	1,5	1,7	1,8
0,9	1,3	1,4	1,6	1,5	2,1
1,1	1,2	1,2	1,7	1,6	1,6
1,2	1,1	1,1	1,4	1,3	1,8
1,0	1,2	1,2	1,5	1,4	1,9

0,9	1,3	1,3	1,3	1,3	1,8
-----	-----	-----	-----	-----	-----

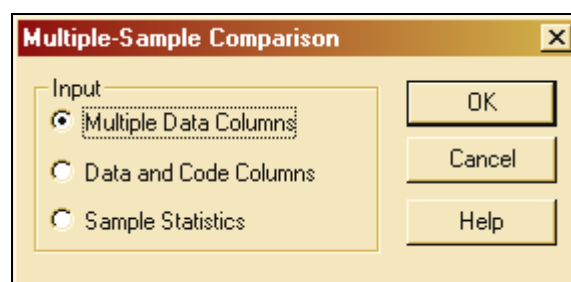
Предполагается, что выручка в каждый день недели имеет нормальное распределение (почему такое предположение естественно для нашей задачи?).

Используя однофакторный дисперсионный анализ, ответить на поставленный вопрос. Проверьте законность применения однофакторного дисперсионного анализа с помощью критерия Кочрена. Уровень значимости принять равным 0,05 (5 %).

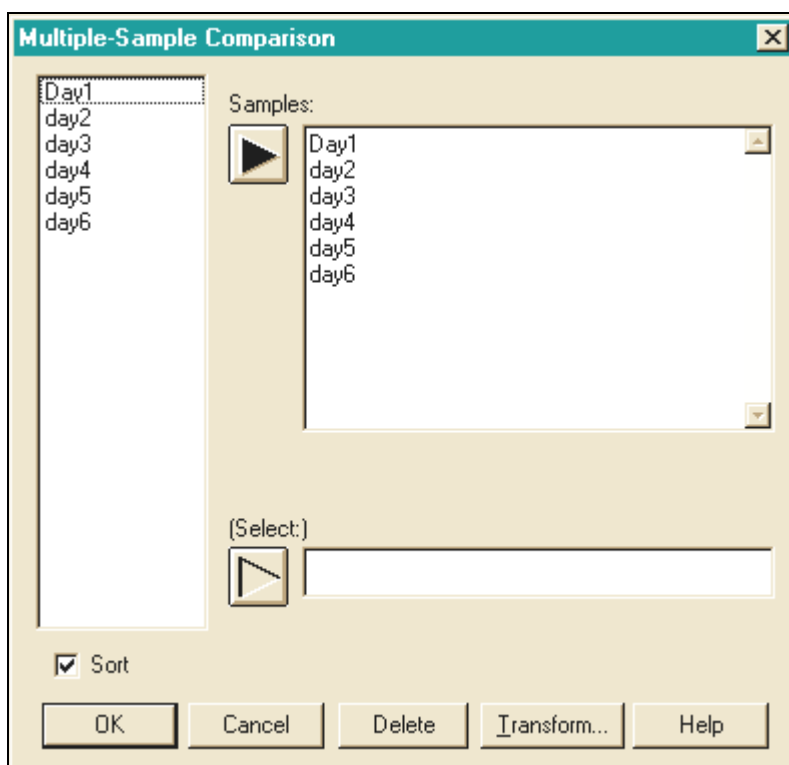
Введите данные, приведенные в таблице, следующим образом:

	Day1	day2	day3	day4	day5	day6
1	1,2	1,1	1,5	1,6	1,2	1,5
2	1,0	1,4	1,0	1,4	1,2	2,0
3	0,8	1,1	1,1	1,5	1,8	1,9

В строке меню выберите *Compare*, в раскрывшемся меню выберите *Multiple Samples, Multiple-Sample Comparison*,



в раскрывшемся диалоговом окне выберите все шесть столбцов



Нажмите кнопку OK.

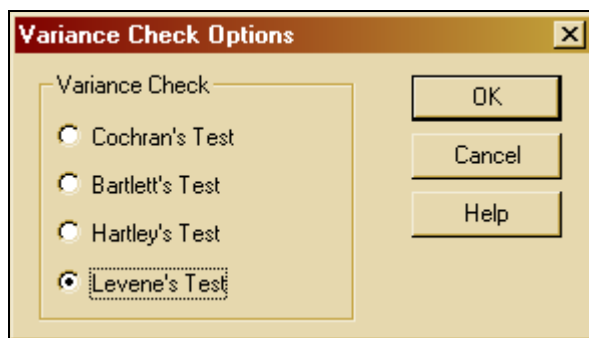
Из теории известно, что для того, чтобы применение однофакторного дисперсионного анализа было законным, необходимо проверить равенство дисперсий. Для проверки этой гипотезы щелкните по кнопке **Tables**, в раскрывшемся окне выберите **Variance Check**. Раскроется новое окно, которое после увеличения будет иметь следующий вид:

Variance Check

	Test	P-Value
Levene's	0,775316	0,571346

The StatAdvisor
 The statistic displayed in this table tests the null hypothesis that the standard deviations within each of the 6 columns are the same. Of particular interest is the P-value. Since the the P-value is greater than or equal to 0,05, there is not a statistically significant difference amongst the standard deviations at the 95,0% confidence level.

Щелкните по нему правой кнопкой, выберите **Pane Options**. Раскроется окно:

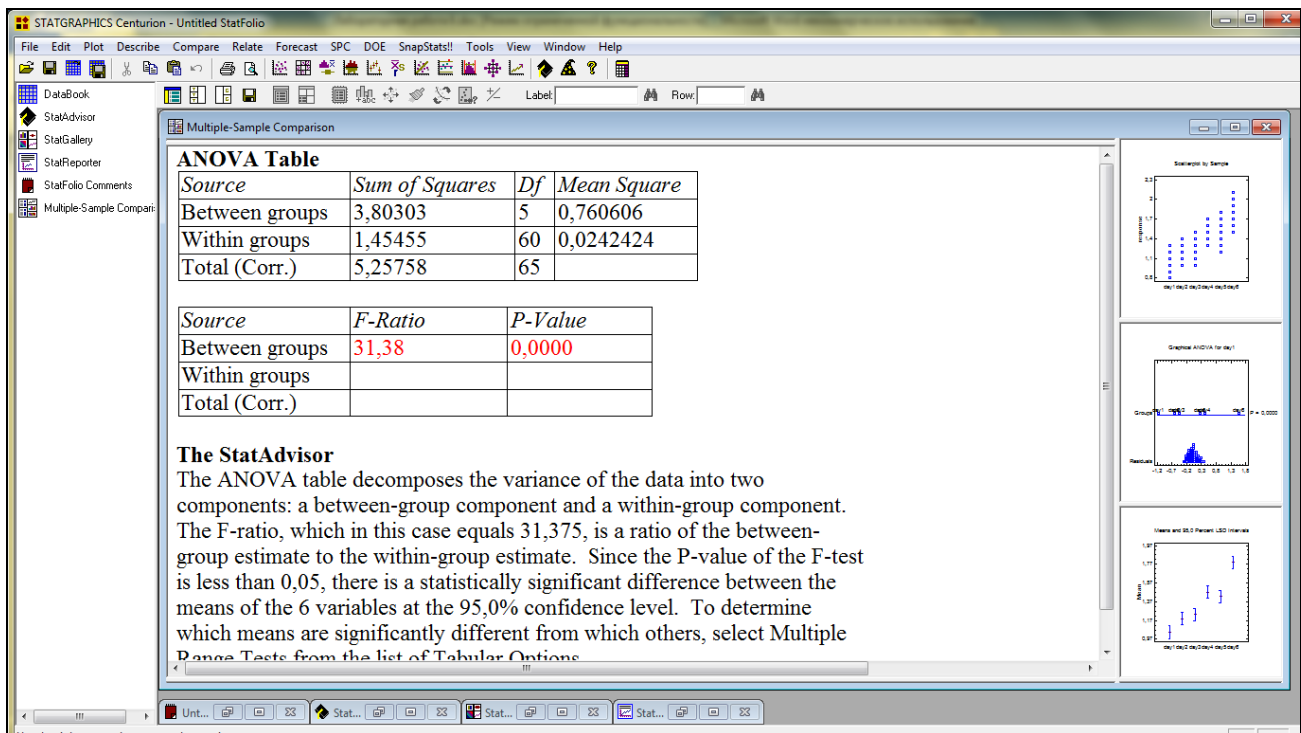


В этом окне можно видеть четыре варианта тестирования. Поскольку у нас объемы выборки равны, воспользуемся критерием Кочрена. **Если гипотезу о равенстве дисперсий следует отклонить, применение однофакторного дисперсионного анализа нельзя считать законным.**

Multiple-Sample Comparison		
Variance Check		
	<i>Test</i>	<i>P-Value</i>
Cochran's C	0,27625	0,395821

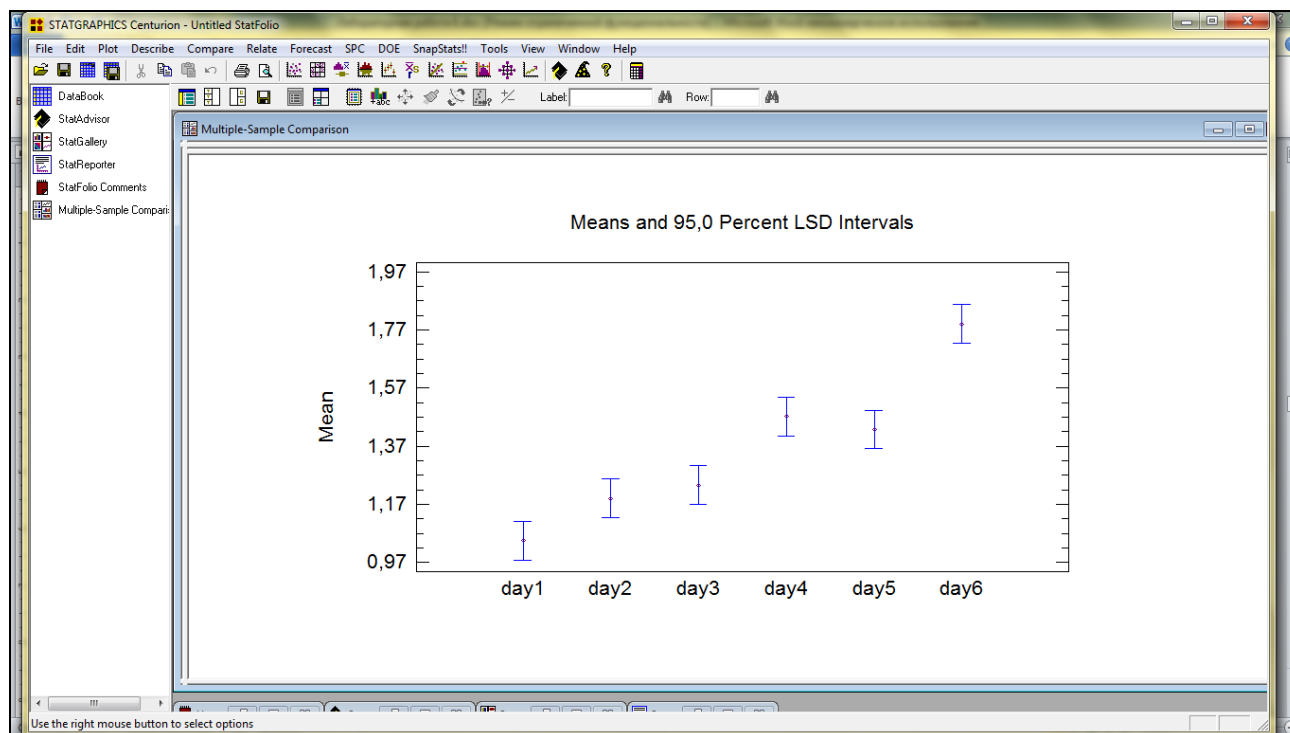
В нашем случае $P\text{-value} = 0,395821$ говорит о том, что нет основания отвергнуть нулевую гипотезу о равенстве дисперсий (альтернативная – двусторонняя). Следовательно, мы можем воспользоваться критерием *ANOVA*.

Найдите у себя в окне *ANOVA*, если его нет в окне -- щелкните по кнопке *Tables*, в раскрывшемся меню выберите *ANOVA Table*, раскроется окно, в котором нужно дважды щелкнуть для увеличения его размера.

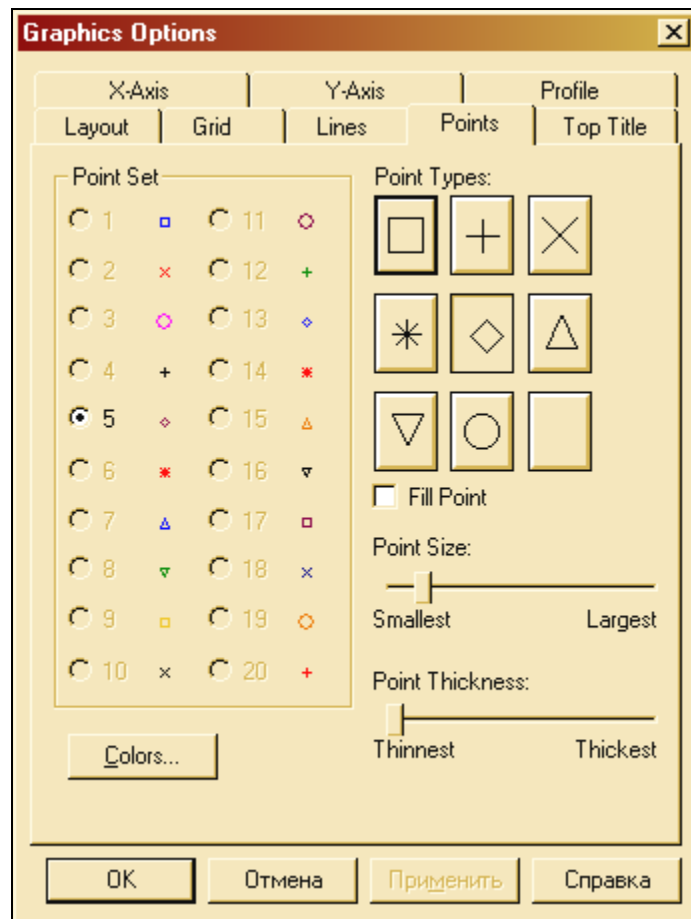


Обратите внимание, в **StatAdvisor** можно прочитать о методе и посмотреть результаты. Посмотрим внимательнее на результаты. В столбце *Sum of Squares* $S_{\text{факт}}=1,185$; $S_{\text{ост}}=0,68$; $S_{\text{общ}}=1,865$. *Df* – число степеней свободы. *F-Ratio* — $F_{\text{набл}}$. В данном случае $F\text{-Ratio}=4,18$. Можете самостоятельно сравнить эту величину с критическим значением $F_{0,05,5,12}=3,106$ (находится по таблице для распределения Фишера). Если таблиц нет, то мы сравниваем *p-value* с α и отвергаем нулевую гипотезу, если $p\text{-value} < 0,05$. В нашем случае как раз $p\text{-value} = 0,0000 < 0,05$, отсюда следует, что гипотезу о равенстве математических ожиданий следует отвергнуть. Следовательно, день недели влияет на выручку.

Посмотрим на графическую иллюстрацию решения этой задачи. Щелкните по кнопке **Graphs**, раскроется окно, в котором вы должны выбрать **Means Plot** (график средних). Перед вами раскроется окно, щелкните дважды по графику, для того, чтобы он развернулся во весь экран.



Хорошо видно, что оборот товара растет к концу недели. На графике отображены также доверительные интервалы. Кстати, обратите внимание на цвет фона, цвет подписей графика. Попробуйте самостоятельно изменить у себя на экране цвет фона, цвет шрифта, а также тип маркера. Для этого щелкните правой кнопкой по графику, выберите **Graphs**. Перед вами окно, в котором изображены различные типы маркеров, которые вы можете использовать.

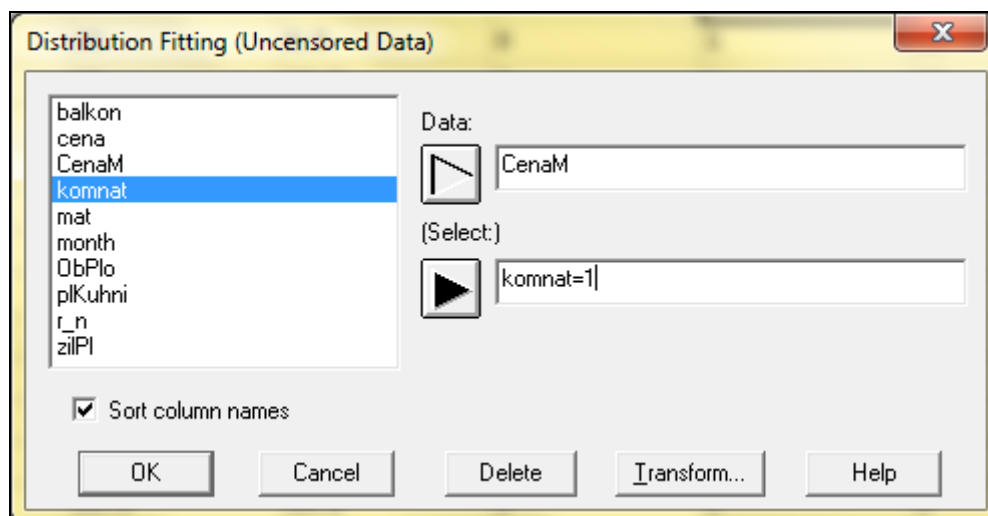


Задача 2. Откройте файл Питер.sf. В нем набраны данные о рынке строящегося жилья в Санкт-Петербурге (декабрь 2000 г.). Здесь *Cena* — цена квартиры (тыс. дол.), *Komnat* — количество комнат, *r_n* — район города (1 – Приморский, Шувалово–Озерки, 2 – Гражданка, 3 – Юго–Запад, 4 – Красносельский); *ObPlo* — общая площадь квартиры (кв. м.); *ZilPlo* — жилая площадь квартиры (кв. м.); *PlKuchni* — площадь кухни (кв. м.); *Mat* — материал (1 – кирпичный дом, 2 – другой); *Balkon* — наличие балкона (1 – есть балкон, 0 – нет); *Month* — число месяцев до окончания строительства. Проанализируем, зависит ли цена одного метра общей площади от количества комнат в квартире.

В этой задаче данные введены не в виде таблицы, а в виде двух столбцов, в одном из которых элементы выборки, а во втором — значения фактора. *StatGraphics* позволяет обрабатывать и таким образом введенные данные.

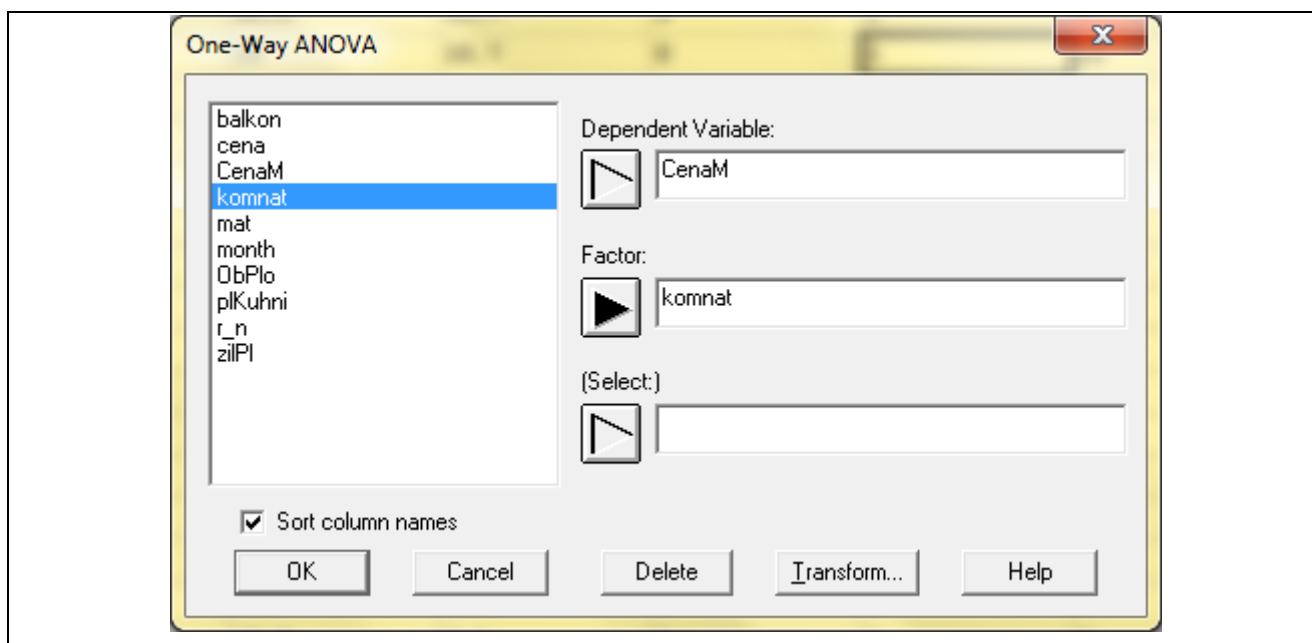
Прежде всего самостоятельно вычислите цену метра общей площади (вспомните, в пакете *StatGraphics* есть калькулятор, посмотрите работу № 2). Занесите цену 1 кв. метра в столбец *CenaM*. Проверьте нормальность этой выборки



(по количеству комнат). Например, для однокомнатных квартир окно выглядит так:

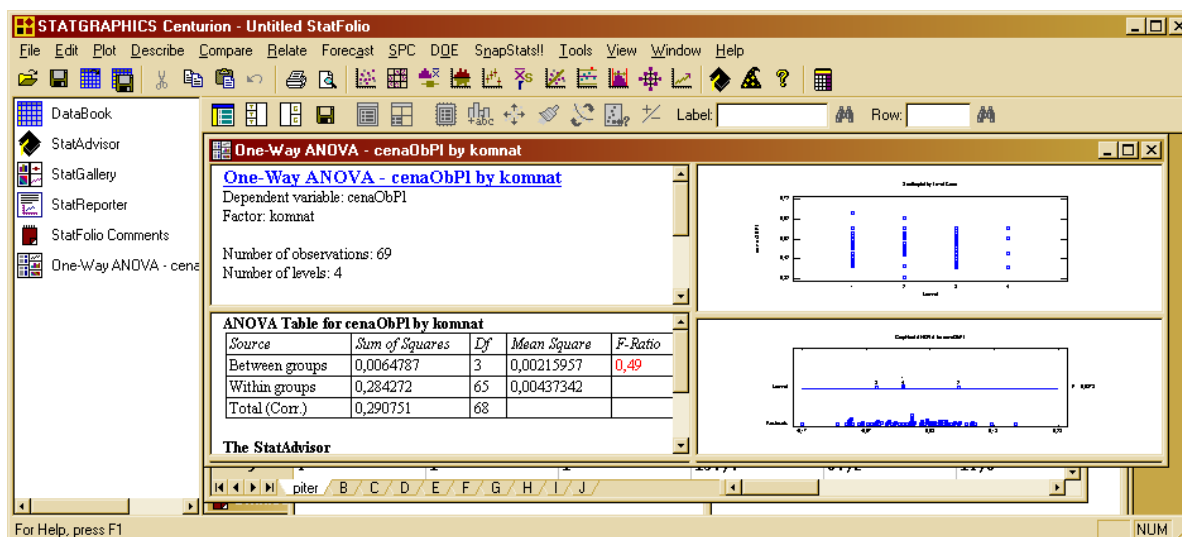


Результат покажите преподавателю.

В строке меню выберите **Compare**, в раскрывшемся меню выберите **Analyses of Variance**, затем **One-Way ANOVA**.



Перед Вами раскрылось окно, в нем выберите **CenaM**, нажмите кнопку , слово перейдет в поле **Dependent Variable** (зависимая переменная), затем выберите **Komnat**, нажмите кнопку  в поле **Factor**, слово перейдет в соответствующее поле. Затем нажмите кнопку OK. Щелкните по кнопке **Tables**, в раскрывшемся меню выберите **Variance Check** и **ANOVA Table**, затем раскроется окно



Самостоятельно проанализируйте результаты. Покажите преподавателю. Самостоятельно проверьте, есть ли зависимость цены метра общей площади от района. Можно ли в этом случае пользоваться ANOVA? Покажите результаты преподавателю.

НЕПАРАМЕТРИЧЕСКИЕ КРИТЕРИИ ПРОВЕРКИ ОДНОРОДНОСТИ.

Критерий Краскела–Уоллиса

Если мы ничего не знаем о распределении наблюдений, можно воспользоваться непараметрическими критериями проверки однородности.

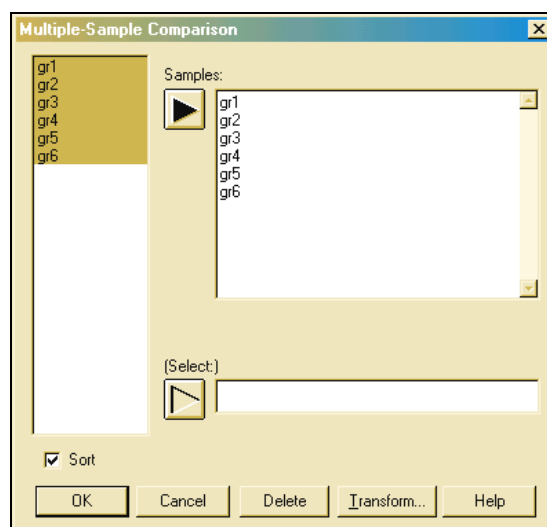
Задача 3. Для выяснения влияния денежного стимулирования на производительность труда шести однородным группам из пяти человек были предложены задачи одинаковой трудности. Задачи предлагались каждому испытуемому независимо от всех остальных. Группы отличались между собой величиной денежного вознаграждения за решаемую задачу. В таблице приведено число решенных задач членами каждой группы.

Группа 1	Группа 2	Группа 3	Группа 4	Группа 5	Группа 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Введите данные следующим образом:

	gr1	gr2	gr3	gr4	gr5	gr6
1	10	8	12	12	24	19
2	11	10	17	15	16	18
3	9	16	14	16	22	27
4	13	13	9	16	18	25
5	7	12	16	19	20	24

В строке меню выберите **Compare**, в раскрывшемся меню выберите **Multiple-Sample**, раскроется диалоговое окно, в котором в поле **Sample** введите данные, как показано на рисунке, затем нажмите кнопку OK.



Нажмите на кнопку **Tables**, в раскрывшемся диалоговом окне выберите **Kruskal-Wallis and Friedman Test** нажмите кнопку OK. Раскроется окно:

STATGRAPHICS Centurion - Untitled StatFolio

File Edit Plot Describe Compare Relate Forecast SPC DOE SnapStats!! Tools View Window Help

DataBook StatAdvisor StatGallery StatReporter StatFolio Comments Multiple-Sample Compari...

Multiple-Sample Comparison

Kruskal-Wallis Test

	Sample Size	Average Rank
gr1	5	5,7
gr2	5	9,0
gr3	5	12,5
gr4	5	16,1
gr5	5	23,4
gr6	5	26,3

Test statistic = 21,219 P-Value = 0,000736395

The StatAdvisor

The Kruskal-Wallis test tests the null hypothesis that the medians within each of the 6 columns is the same. The data from all the columns is first ranked and ranked from smallest to largest. The average rank is then computed for the data in each column. Since the P-value is less than 0.05, the null hypothesis is rejected.

Use the right mouse button to select options

NUM

Так как $p\text{-value}=0,0007363$ меньше чем 0,05, то существует статистически значимое различие между медианами с 95 % доверительным интервалом. Следовательно, можно уверенно отвергнуть гипотезу об однородности.

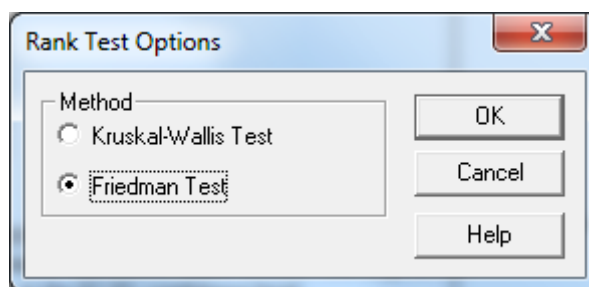
Критерий Фридмана

Задача 4. Деканат заметил, что занятия по одним предметам студенты прогуливают чаще, а по другим – реже. В конце семестра была составлена таблица с пропусками студентов одной и той же группы. Здесь А, В, С – коды предметов. Можно ли на основании этих данных сказать, что студенты прогуливают разные предметы выборочно, в зависимости от предмета?

Студент	А	В	С
1	3	5	7
2	5	2	3
3	2	6	4
4	6	7	5
5	7	1	3
6	5	0	2
7	0	4	3
8	4	5	6
9	1	2	3
10	5	7	6
11	2	4	1
12	2	0	3
13	5	3	0
14	0	3	4
15	3	7	5
16	0	5	4
17	3	4	6
18	1	6	4
19	3	5	2
20	5	1	2

Решим задачу с помощью критерия Фридмана. Введите данные, в строке меню выберите **Compare**, в раскрывшемся меню выберите **Multiple-Sample**, раскроется диалоговое окно, в котором в поле **Sample** введите данные, затем нажмите кнопку ОК. Нажмите на кнопку **Tables**, в раскрывшемся диалоговом окне выберите **Kruskal-Wallis and Friedman Test** нажмите кнопку ОК. В раскрывшемся

окне щелкните правой кнопкой, выберите *Pane Options*, выберите *Friedman Test*, нажмите OK



раскроется окно

Friedman Test

	Sample Size	Average Rank
A	20	1,7
B	20	2,2
C	20	2,1

Test statistic = 2,8 P-Value = 0,246597

Т.К. $P\text{-Value} = 0,246597 < 0,05$ нет оснований отвергнуть нулевую гипотезу (которая состояла в том, что медианы равны). Следовательно, по данным этой выборки нельзя сказать, что какие-то предметы студенты прогуливают чаще.

Можно ли эту задачу решать с помощью критерия ANOVA? С помощью критерия Краскела-Уоллеса? Обсудите результат с преподавателем.

ЗАДАНИЕ 1. Бригада рабочих производит однотипную продукцию. Количество изготовленных за смену деталей в зависимости от стажа приведено в таблице

Стаж работы		
до 10 лет	10-15 лет	15-25 лет
135	176	155
156	196	160
165	204	149
	180	171
		140

Зависит ли производительность труда от стажа? Какой критерий вы будете использовать и почему?

ЗАДАНИЕ 2. За десять лет пребывания А.М.Чернецкого в должности мэра Екатеринбурга в городе было построено определенное количество жилья, школьных мест и т.д. В таблице для сравнения приведены данные по сопоставимым по населению городам за эти же годы.

	Екатеринбург	Пермь	Челябинск	Новосибирск	Омск	Н.Новгород	Самара
Жилье(1000 кв.м.)	2821	2037	2106	2193	1783	2077	1810
Школы (мест)	10413	2832	8970	6628	8396	10397	5020
Больницы (кол-во коек)	982	1124	653	525	200	860	100
Водопровод(км)	80,3	64,8	2,6	33,8	8,8	3,7	14,2
Тепловые сети(км)	42	3,2	1,7	26	4,7	0,2	12,3
Трам/Трол пути (км)	29,7	16,3	13,8	15,7	4	9,2	5,8

1. Есть ли существенные различия в показателях развития?

2. Есть ли различие между показателями Екатеринбурга и Новосибирска?

Какие тесты будете использовать и почему?

ВОПРОСЫ К ЛАБОРАТОРНОЙ РАБОТЕ

1. Какому условию должны удовлетворять выборки, чтобы можно было воспользоваться однофакторным дисперсионным анализом?
2. Что дают критерии Барлетта и Кочрена для однофакторного анализа?
3. Как проверить нормальность для выборки малого объема?
4. Каким критерием следует воспользоваться, если при однофакторном анализе вы обнаружили, что нет нормальности?
5. Что вы предпримите, если нельзя проинтерпретировать числа, записанные в столбце, как значения одной случайной величины?
6. В чем состоят основная и альтернативная гипотезы в однофакторном дисперсионном анализе?
7. Таблицы какого распределения используются для принятия решения в одно-(много) факторном дисперсионном анализе?
8. Должны ли совпадать объемы выборок по каждой случайной величине в однофакторном дисперсионном анализе?