Министерство образования и науки Украины
Севастопольский национальный технический университет

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ И ИНДИВИДУАЛЬНЫЕ ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ СТУДЕНТОВ ПО ИНТЕЛЛЕКТУАЛЬНОМУ АНАЛИЗУ ДАННЫХ

ДИСПЕРСИОННЫЙ АНАЛИЗ

Севастополь
2014

СОДЕРЖАНИЕ

1. ЦЕЛЬ И ЗАДАЧИ МЕТОДИЧЕСКИХ РЕКОМЕНДАЦИЙ

Закрепить теоретические знания и приобрести практические навыки в проведении дисперсионного анализа по экспериментальным данным с использованием программ STATISTICA и MS EXCEL.

По результатам наблюдений за функционированием объектов получены экспериментальные данные. Требуется провести дисперсионный анализ этих данных.

2. КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

2.1. Задача дисперсионного анализа

В любом эксперименте среднее значение наблюдаемых величин меняются с изменением основных факторов (качественных и количественных), определяющих условия опыта, а также и случайных факторов. Исследование влияние тех или иных факторов на изменчивость средних является задачей дисперсионного анализа.

Дисперсионный анализ использует свойство аддитивности дисперсии изучаемой случайной величины, обусловленной действием независимых факторов. Р.А. Фишер в 1938 г. впервые определил дисперсионный анализ как «отделение дисперсии, приписываемой одной группе причин от дисперсии, приписываемой другими группами». В зависимости от числа источников дисперсии различают однофакторный и многофакторный дисперсионный анализ.

Дисперсионный анализ особенно эффективен при изучении нескольких факторов. При классическом методе исследования варьируют только один фактор, а остальные оставляют постоянными. при этом для каждого фактора проводится серия наблюдений, не используемая при изучении других факторов. Кроме того, при таком методе исследования не удается определить взаимодействие факторов при одновременном их изменении. При дисперсионном анализе каждое наблюдение служит для одновременной оценки всех факторов и их взаимодействий.

Дисперсионный анализ состоит в выделении и оценке отдельных факторов, вызывающих изменчивость изучаемой случайной величины. Для этого производится разложение суммарной выборочной дисперсии на составляющие, обусловленные независимыми факторами. Каждая из этих составляющих представляет собой оценку дисперсии генеральной совокупности. Чтобы решить, значимо ли влияние данного фактора, необходимо оценить значимость соответствующей выборочной дисперсии в сравнении с дисперсией воспроизводимости, обусловленной случайными факторами. Проверка значимости оценок дисперсий проводится по критерию Фишера. Если рассчитанное значение критерия Фишера окажется меньше табличного, то влияние рассматриваемого фактора нет оснований считать значимым. Если же рассчитанное значение критерия Фишера окажется больше табличного, то рассматриваемый фактор влияет на изменчивость средних. В дальнейшем будем полагать, что выполняются следующие допущения:

- случайные ошибки наблюдений имеют нормальное распределение;
- факторы влияют только на изменение средних значений, а дисперсия наблюдений остается постоянной;
- эксперименты равноточны.

Требование нормального распределения определяет выбор основных факторов при исследовании процесса методом дисперсионного анализа. Если нужно получить нормальное распределение выходной величины, к случайным желательно относится только те факторы, влияние которых на выходную величину очень мало. Исключение можно делать лишь для тех факторов, которые сами по себе (из каких-либо других соображений) дают нормальное распределение результатов.

Факторы рассматриваемые в дисперсионном анализе, бывают двух родов:

- со случайными уровнями;
- с фиксированными.

В первом случае предполагается, что выбор уровней производится из бесконечной совокупности возможных уровней и сопровождаются рандомизацией. При этом результаты эксперимента имеют большее значение, поскольку выводы по эксперименту можно распространить на всю генеральную совокупность. Если все уровни выбираются случайным образом, математическая модель эксперимента называется моделью со случайными уровнями факторов (случайная модель). Когда все уровни фиксированы, модель называется моделью с фиксированными уровнями. Когда часть факторов рассматривается на

фиксированных уровнях, а уровни остальных выбираются случайным образом, модель называется моделью смешанного типа. Иногда отсутствие различие в критериях, применяемых для разных моделей, и единственное различие состоит в общности выводов, в других случаях существует различие в критериях.

Дисперсионный анализ может применяться в различных формах в зависимости от структуры исследуемого процесса; выбор соответствующей формы является обычно одной из главных трудностей в практическом применении анализа.

2.2. Однофакторный дисперсионный анализ

Задачей дисперсионного анализа является изучение влияния одного или нескольких факторов на рассматриваемый признак.

Однофакторный дисперсионный анализ используется в тех случаях, когда есть в распоряжении три или более независимые выборки, полученные из одной генеральной совокупности путем изменения какого-либо независимого фактора, для которого по каким-либо причинам нет количественных измерений.

Для этих выборок предполагают, что они имеют разные выборочные средние и одинаковые выборочные дисперсии. Поэтому необходимо ответить на вопрос, оказал ли этот фактор существенное влияние на разброс выборочных средних или разброс является следствием случайностей, вызванных небольшими объемами выборок. Другими словами если выборки принадлежат одной и той же генеральной совокупности, то разброс данных между выборками (между группами) должен быть не больше, чем разброс данных внутри этих выборок (внутри групп).

Пусть x_{ik} – i – элемент ($i = \overline{1, n_k}$) k -выборки ($k = \overline{1, m}$), где m – число выборок, nk – число данных в k -выборке. Тогда $\overline{x_{ik}}$ – выборочное среднее k -выборки определяется по формуле

$$\overline{x_k} = \frac{1}{n_k} \sum_{i=1}^{nk} x_{ik}$$

.

Общее среднее вычисляется по формуле

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{m} \sum_{i=1}^{nk} x_{ik}$$

,

где

$$n = \sum_{k=1}^{m} n_k \, .$$

Основное тождество дисперсионного анализа имеет следующий вид:

$$Q = Q_1 + Q_2 \, ,$$

где Q1 – сумма квадратов отклонений выборочных средних $\overline{x_{ik}}$ от общего среднего $\overline{x}$ (сумма квадратов отклонений между группами); Q2 – сумма квадратов отклонений наблюдаемых значений x_{ik} от выборочной средней $\overline{x_k}$ (сумма квадратов отклонений внутри групп); Q – общая сумма квадратов отклонений наблюдаемых значений x_{ik} от общего среднего $\overline{x}$.

Расчет этих сумм квадратов отклонений осуществляется по следующим формулам:

$$Q = \sum_{k=1}^{m}\sum_{i=1}^{nk}(x_{ik} - \bar{x})^2 = \sum_{k=1}^{m}\sum_{i=1}^{nk} x^2{}_{ik} - n\bar{x}^2,$$

$$Q_1 = \sum_{k=1}^{m} n_k(\bar{x}_k - \bar{x})^2 = \sum_{k=1}^{m} n_k \bar{x}_k^2 - n\bar{x}^2,$$

$$Q_2 = \sum_{k=1}^{m}\sum_{i=1}^{nk}(x_{ik} - \bar{x}_k)^2 = \sum_{k=1}^{m}\sum_{i=1}^{nk} x^2{}_{ik} - \sum_{k=1}^{m} n_k \bar{x}_k^2.$$

В качестве критерия необходимо воспользоваться критерием Фишера:

$$F = \frac{Q_1/(m-1)}{Q_2/(n-m)}.$$

Если расчетное значение критерия Фишера будет меньше, чем табличное значение $F_{\lambda; m-1; n-m}$ – нет оснований считать, что независимый фактор оказывает влияние на разброс средних значений, в противном случае, независимый фактор оказывает существенное влияние на разброс средних значений (λ– уровень значимости, уровень риска, обычно для экономических задач λ=0,05).

Недостаток однофакторного анализа: невозможно выделить те выборки, которые отличаются от других. Для этой цели необходимо использовать метод Шеффе или проводить парные сравнения выборок.

2.3. Дисперсионный анализ в MS EXCEL

Создать файл с исходными данными.

Запустить “Пакет анализа”.

В системе электронных таблиц Microsoft Excel имеется набор инструментов для анализа данных, называемый пакет анализа, который может быть использован для решения сложных статистических задач. Для использования одного из этих инструментов указать входные данные и выбрать параметры; анализ будет проведен с помощью подходящей статистической макрофункции, и результаты будут представлены в выходном диапазоне.

В меню Сервис выберите команду Анализ данных. Если такая команда отсутствует в меню Сервис, то необходимо установить в Microsoft Excel пакет анализа данных.

Установка производится следующим образом. В меню Сервис выберите команду Надстройки. Если в списке надстроек нет пакета анализа данных, то нажмите кнопку “Обзор” и задайте диск, каталог и имя файла для надстройки “Пакет анализа”, или запустите программу установки Microsoft Excel. Установите флажок “Пакет анализа” (надстройки, установленные в Microsoft Excel, остаются доступными, пока не будут удалены).

Выберите необходимую строку в списке “Инструменты анализа”.

Введите входной и выходной диапазоны, затем выберите необходимые параметры. Для использования инструментов анализа исследуемые данные следует представить в виде строк или столбцов на листе. Совокупность ячеек, содержащих анализируемые данные, называется входным диапазоном.

Провести однофакторный дисперсионный анализ.

В меню Сервис выбираем команду Анализ данных.

В списке инструментов статистического анализа выбираем Однофакторный дисперсионный анализ (Рисунок 1).

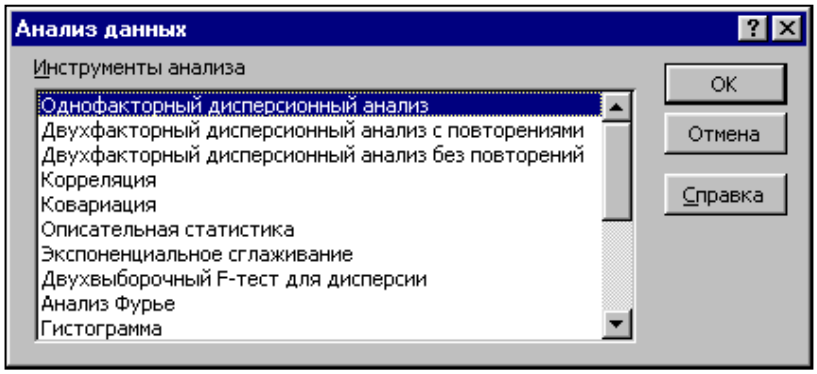

Рисунок 1 – Выбор инструмента анализа

В диалоговом окне режима (Рисунок 2) указываем входной интервал, способ группирования, выходной интервал, метки в первой строке/ Метки в первом столбце, альфа (уровень значимости).

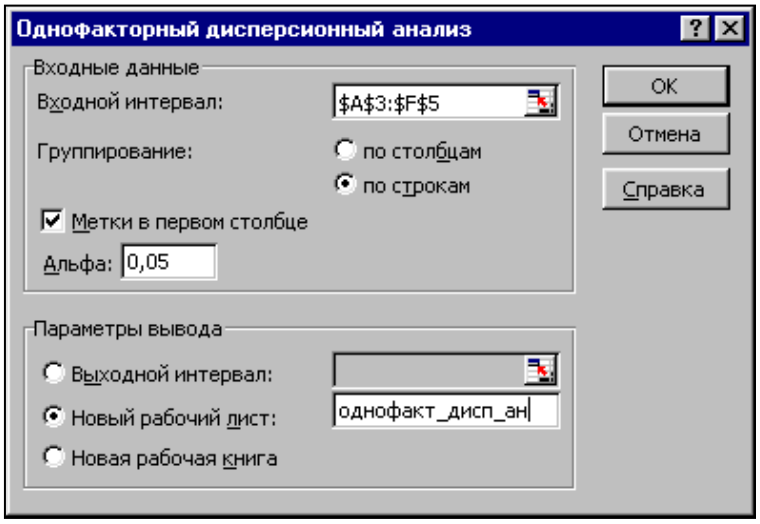

Рисунок 2 – Диалоговое окно однофакторного дисперсионного анализа

Входной диапазон – это ссылка на ячейки, содержащие анализируемые данные. Ссылка должна состоять как минимум из двух смежных диапазонов данных, организованных в виде столбцов или строк. Входной интервал можно задать при помощи мыши, или набрать на клавиатуре.

Группирование. Установите переключатель в положение “по столбцам” или “по строкам” в зависимости от расположения данных во входном диапазоне.

Метки в первой строке/ Метки в первом столбце. Установите переключатель в положение “Метки в первой строке”, если первая строка во входном диапазоне содержит названия столбцов. Установите переключатель в положение “Метки в первом столбце”, если названия строк находятся в первом столбце входного диапазона. Если входной диапазон не содержит меток, то необходимые заголовки в выходном диапазоне будут созданы автоматически.

Выходной диапазон. Введите ссылку на ячейку, расположенную в левом верхнем углу выходного диапазона. Размеры выходной области будут рассчитаны автоматически, и соответствующее сообщение появится на экране в том случае, если выходной диапазон занимает место существующих данных или его размеры превышают размеры листа.

Новый лист. Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенном напротив соответствующего положения переключателя.

Новая книга. Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку А1 на первом листе в этой книге.

В результате обработки данных получили следующее:

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	I группа (контр.)	5	1673	334,6	56,8		
6	II группа	5	1812	362,4	220,8		
7	III группа	5	1885	377	276,5		
8							
9	ANOVA						
10	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
11	Между группами	4640	2	2319,8	12,55983	0,0011415	3,885290312
12	Внутри групп	2216	12	184,7			
13							
14	Итого	6856	14				
15							
16							
17							

однофакт_дисп_ан | Лист1 | Лист2 | Лист3 | Лист4 | Лист5

Рисунок 3 – Результаты однофакторного дисперсионного анализа

Таблица ИТОГИ:

“Счет” – число повторностей. “Сумма” – сумма значений показателя по строкам. “Дисперсия” – частная дисперсия показателя.

Таблица ANOVA представляет результаты дисперсионного анализа однофакторного комплекса, в котором первая колонка “Источник вариации” содержит наименование дисперсий. Графа “SS” - это сумма квадратов отклонений, “df” - степень свободы, графа “MS” - средний квадрат, “F” - критерий фактического F – распределения. “P - значение” - вероятность того, что дисперсия, воспроизводимая уравнением, равна дисперсии остатков. Определяет вероятность того, что полученная количественная определенность взаимосвязи между факторами и результатом может считаться случайной. “F - критическое” - это значение F – теоретического, которое впоследствии сравнивается с F – фактическим.

Рассчитать эмпирическое корреляционное отношение и коэффициент детерминации. Сформулировать выводы.

Учитывая специфику исходных данных, провести двухфакторный дисперсионный анализ с повторениями или без повторений в той же последовательности.

2.4. Дисперсионный анализ в пакете STATISTICA

Пример. Три группы продавцов продавали штучный товар, расфасованный в различные упаковки. После окончания срока распродажи был произведен тестовый контроль над случайно отобранными продавцами из каждой группы. Были получены следующие результаты (табл.1).

Т а б л и ц а 1

Номер группы	Число продаж, которые сделали	Общее количество продаж	Количество продавцов, n_k

	продавцы, x_{ik}		
1	1 3 2 1 0 2 1	10	7
2	2 3 2 1 4 - -	12	5
3	4 5 3 - - - -	12	3

Рассмотрим процедуру решения рассмотренной задачи методом дисперсионного анализа в системе **STATISTICA**.

- Запустите пакет **STATISTICA**.
- Появится диалоговое окно **Statistica Module Switcher** (рис.4).

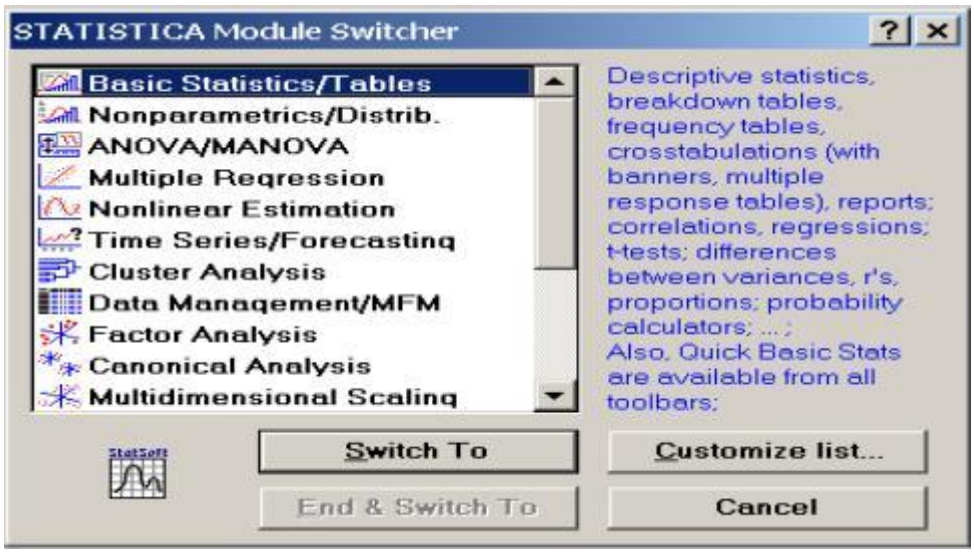

Рис. 4

- Выделите модуль **ANOVA/MANOVA** и нажать кнопку **Switch To**.
- Откроется окно **GENERAL ANOVA/MANOVA**. Если кто-то работал до Вас с этим пакетом, появятся исходные данные предыдущей работы. В любом случае закройте все окна и начните работу сначала. Дайте команду *File/New Data*. Появится электронная таблица Data: new.sta для ввода исходных данных и их преобразования, размерностью 10 столбцов (Vars – переменных) и 10 строк (Cases – случаи).
- Введите исходные данные для переменных в столбцы VAR1 и VAR2 в следующем виде (придется добавить 5 Cases) (рис. 5) или воспользоваться файлом DATA3.

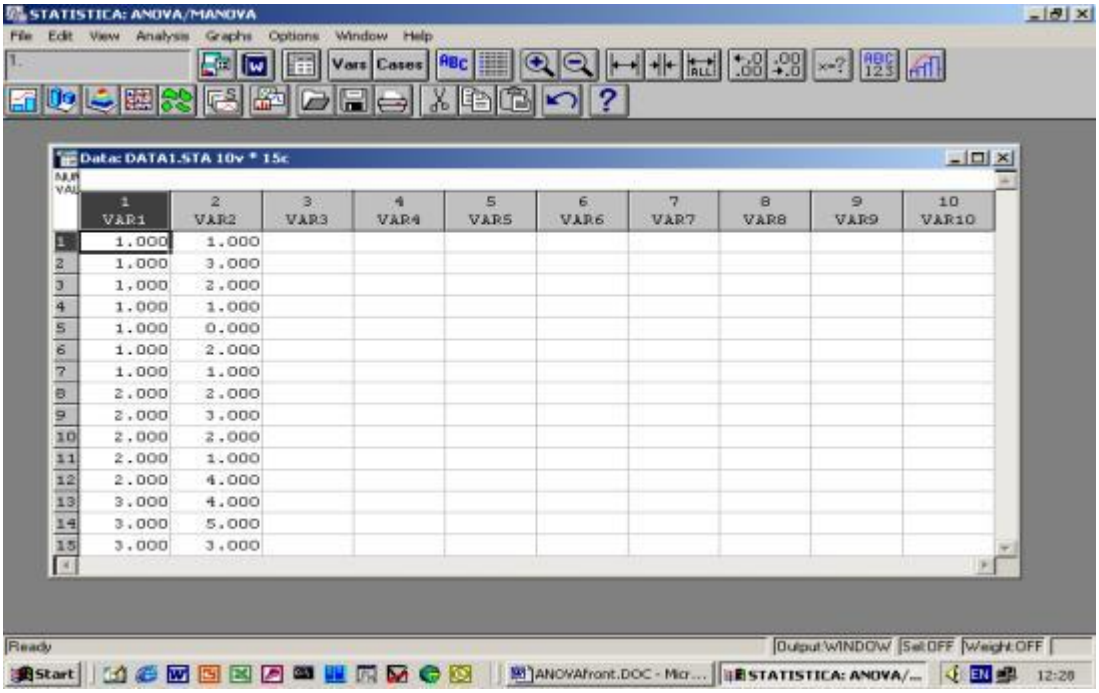

Рис. 5

- Нажимая кнопку **Vars/Cases** удалите лишние переменные *from* VAR3 *to* VAR10.
- Щелкая правой клавишей по столбцам VAR1 и VAR2, выберите контекстное меню, выделите пункт **Variable Specs…** и поменяйте имена переменных, если в этом есть необходимость.
- *File/Save As* – сохраните полученный файл в нужном директории с именем *data*1.*sta*.

- Выполните команду *Analysis/Resume Analysis.* Появится меню **General ANOVA/MANOVA** (рис. 6).

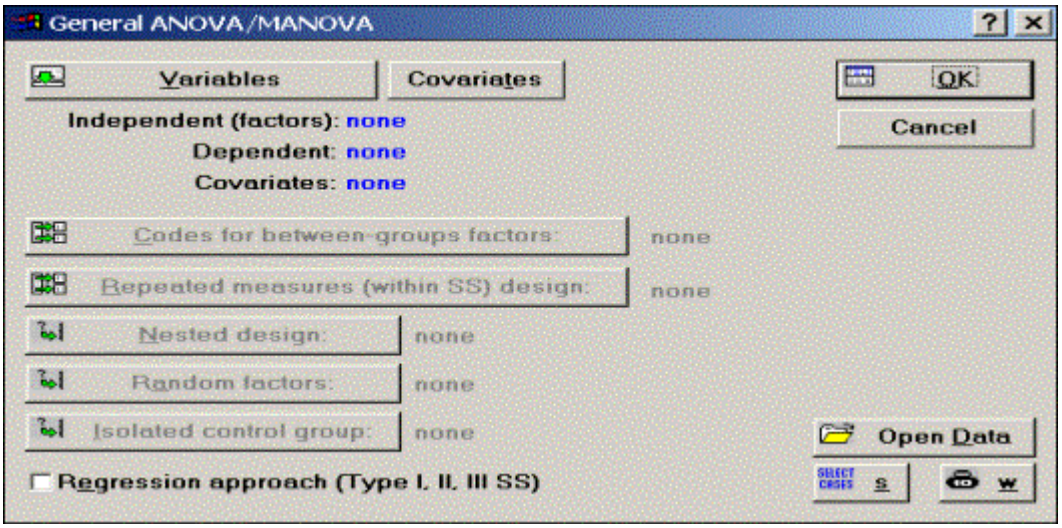

Рис. 6

- Нажмите кнопку **Variables** и определите независимую (VAR1) и зависимую (VAR2) переменные. После определения переменных вы вернетесь в меню **General ANOVA/MANOVA.** Нажмите**OK**. Появится панель **ANOVA Results** (рис. 7).

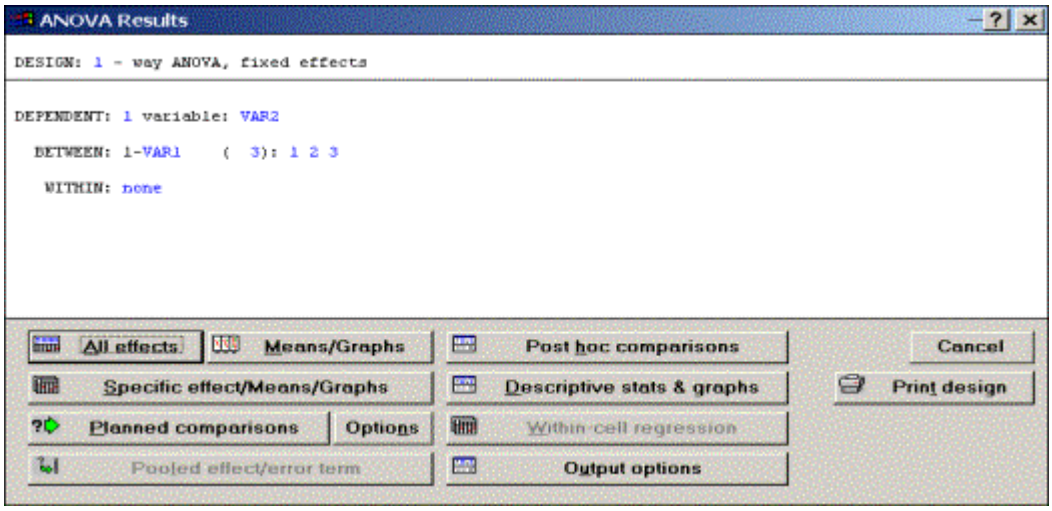

Рис. 7

Для решения данной задачи достаточно нажать кнопку **All effects** и на экране появятся результаты общего дисперсионного анализа (рис. 8). Если эти результаты выделены красным цветом – фактор оказывает существенное влияние, что мы и наблюдаем на экране Более точный вывод можно сделать, применив критерий Фишера.

Summary of all Effects; design: (data1.sta)

Continue.. 1-VAR1

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	2	7.009524	12	1.076190	6.513274	.012153

Рис. 8

3. КОНТРОЛЬНЫЕ ВОПРОСЫ

1.Что такое факторы в дисперсионном анализе?
2.С какой целью применяют дисперсионный анализ?
3.Что называется факторной дисперсией в дисперсионном анализе?
4.Что такое остаточная дисперсия в дисперсионном анализе?

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Айвазян С. А. Прикладная статистика. Основы эконометрики: Учебник для вузов/ С. А. Айвазян, В. С. Мхитарян. М. : ЮНИТИ-ДАНА, 2001. –656 с.

2. Боровиков В.П., Боровиков И.П. STATISTICA. Статистический анализ и обработка данных в среде Windows. М.: Филин, 1997.

3. Боровиков В.П. Популярное введение в программу STATISTICA. М., 1998

4. www.statsoft.ru (сайт компании StatSoft Russia – документация по STATISTICA).

5. www.exponenta.ru (примеры решения практических задач в ППП STATISTICA).

6. Шанченко, Н. И. Эконометрика: лабораторный практикум : учебное пособие / Н. И. Шанченко. – Ульяновск : УлГТУ, 2011. – 117 с.