Федеральное агентство по образованию РФ
ГОУ ВПО «АМУРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Факультет математики и информатики

Н.Н. Двоерядкина, А.Н. Киселева, Т.А. Юрьева

МАТЕМАТИЧЕСКИЕ МЕТОДЫ В ПСИХОЛОГИИ:

лабораторный практикум

Благовещенск, 2007

Печатается по решению
редакционно-издательского совета
факультета математики и информатики
Амурского государственного
Университета

Двоерядкина Н.Н., Киселева А.Н., ЮрьеваТ.А.

Математические методы в психологии: Лабораторный практикум. Благовещенск: Амурский гос. ун-т, 2007.

В практикуме рассматриваются основные методы обработки данных в программе Statistica. Приведены краткие теоретические сведения и типовые задачи, наиболее часто встречающиеся в экспериментальных психолого-педагогических следованиях. Содержатся практические задания для выполнения лабораторных работ.

Практикум предназначен для студентов очного отделения специальности «030301».

Введение

Современная социально-политическая жизнь России стимулирует появление большого количества специалистов, занимающихся изучением общественного мнения, консультированием политиков, проведением тренингов, имидж-мейкерством и другими многочисленными технологиями.

В информационном обществе анализ полученных данных крайне трудоемок без использования компьютерных программ. Именно поэтому представляется актуальным введение в содержание дисциплины «Математические методы в психологии» изучение различных специализированных компьютерных пакетов на базе задач экспериментальной и прикладной психологии, включающей психологическую диагностику, психологическое консультирование, психологическое воздействие.

Обработка экспериментальных данных предполагает полный и качественный статистический анализ информации. Статистические расчеты без помощи ЭВМ являются сложными и требуют применения многочисленных таблиц, функций и квантилей стандартных распределений. Поэтому они не дают возможности почувствовать элемент новизны в изучаемом материале, изменять произвольно условия задач и т. д. Наиболее мощными возможностями статистической обработки обладают специализированный пакет STATISTICA, который обладает широкими возможностями графического анализа, что, позволяет представить результаты исследования в наглядной форме. Эта англоязычная система нашла достаточно широкое распространение среди зарубежных специалистов. Она не так доступна, как Excel, но зато предоставляет ряд дополнительных возможностей, которые могут быть весьма полезны психологу в серьезной исследовательской работе.

Настоящий лабораторный практикум ориентирован на студентов специальности «Психология», но может быть полезен преподавателям, аспирантам и исследователям психолого-педагогических явлений, требующих статистической обработки.

Лабораторная работа № 1

Основы программы Statistica .

Программа *Statistica* состоит из отдельных модулей, каждый из которых располагается в отдельном окне. Для открытия нужного модуля выбирается команда Statistics в горизонтальном меню (рис.1). Основные модули:

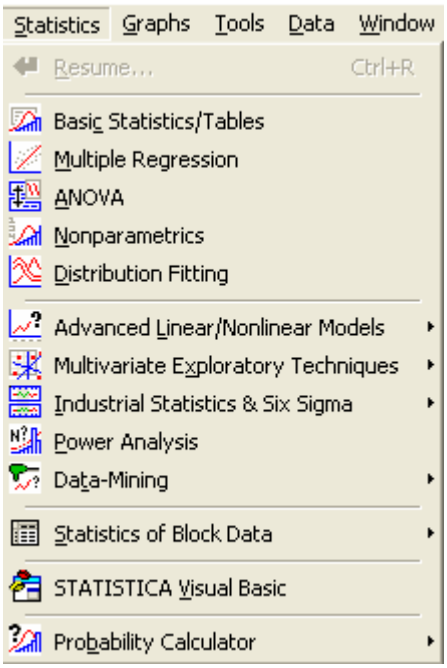

Рис.1. Окно основных команд

Basic Statistcs/Tables – Основные статистики и таблицы;

Multiple Regression – Множественная регрессия;

ANOVA – Однофакторный дисперсионный анализ;

Nonparametrics – Непараметрические критерии;

Distribution Fitting – Распределения

Probability Calculator - Вероятностный калькулятор

В каждом окне можно открыть или создать таблицу с данными (*Data)* (рис.2).

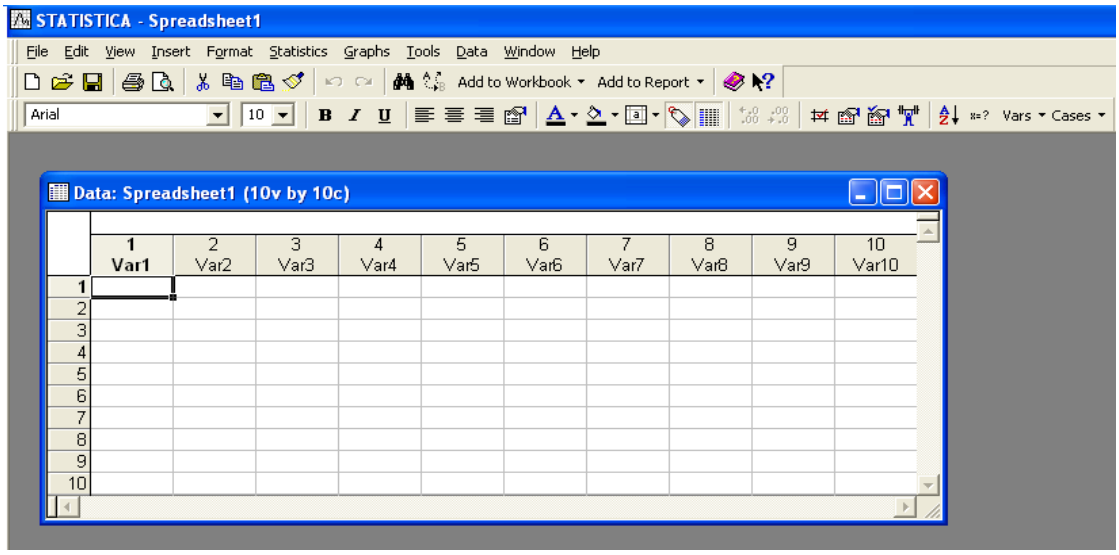

Рис.2. Рабочее окно программы *Statistica.*

Таблица данных состоит из строк и столбцов. Столбцы используют для задания имен переменных (*Variables*), строки - для заполнения наблюдений (*Cases*). Строки и столбцы можно редактировать, выполнив двойной щелчок мыши на названии строки, столбца.

Строки, столбцы в таблице можно добавлять (*Add*), удалять (*Delete*), перемещать (*Move*) и др. Данные действия выполняются при помощи команд *Edit/ Variables* (работа со столбцами), *Edit/ Cases* (работа со строками).

Диалоговое окно задания переменной позволяет:

а) *name* - задать имя переменной,

б) *column width* - ширину столбца в символах,

в) *decimals* - количество знаков после запятой,

г) *category* - тип данных (например, *number* - числовой).

д) *long name* - с помощью кнопки *Functions* автоматически заполнить один из столбцов как значение функции, аргументами которой являются переменные других столбцов.

Задания:

1. Выберите пункты меню Пуск/Программы/ Statistica / Basic Statistics. Откройте новый файл, выбрав пункт File/New Data. Сохраните файл с англоязычным именем.

2. Закройте окно пунктов меню, откройте в горизонтальном меню команду Statistics. Убедитесь, что пункты меню переместились в подменю Statistics.

3. В вашей таблице 10 строк. Добавьте две строки, чтобы у Вас получилось 12 наблюдений. Оставьте в таблице три столбца, остальные удалите. В итоге у Вас получится таблица размером 3x12.

4. Переименуйте переменные:

x1: ширина столбца - 6, количество знаков после запятой - 0;

x2: ширина столбца - 5. количество знаков после запятой - 1;

x3: ширина столбца - 4, количество знаков после запятой - 3.

5. В первый столбец введите значения: 20, 50, 10, 8, 30, 50, 140, 70, 20, 25, 20, 30.

6. Значения второй переменной рассчитайте по формуле: x2=log10(x1).

Лабораторная работа № 2

Процедура Descriptive statistics (Описательные статистики)

Расчет описательных статистик производится при помощи модуля *Basic Statistics/Tables*. В этом модуле объединены наиболее часто использующиеся на начальном этапе обработки данных процедуры. В стартовой панели модуля приводится перечень статистических процедур этого модуля (рис.3):

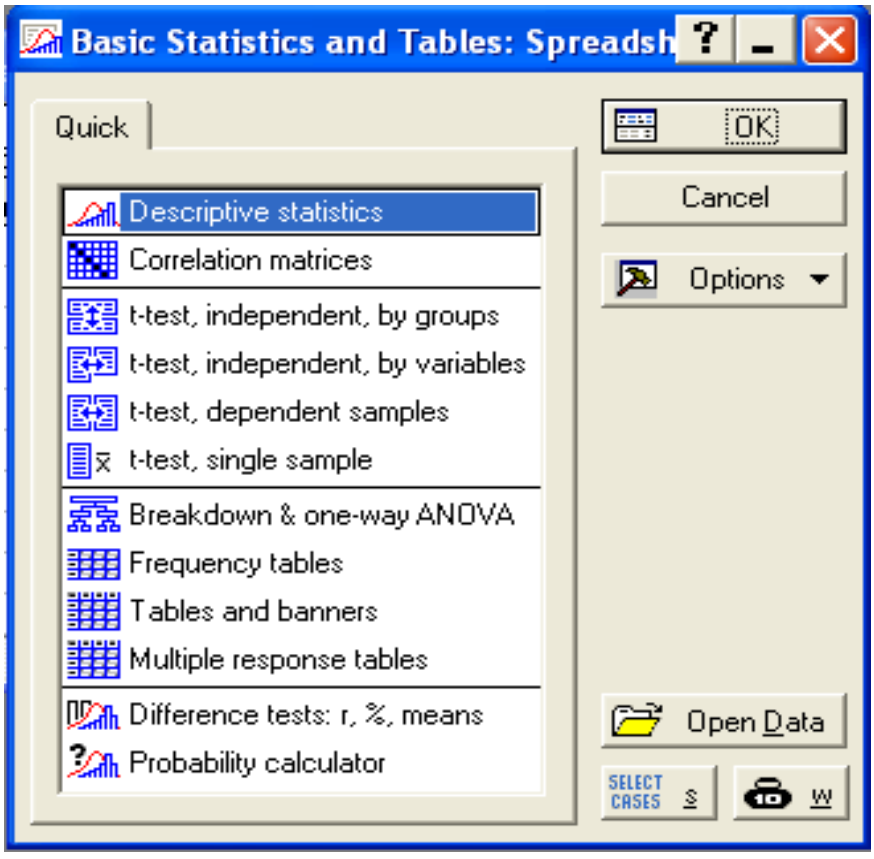

Рис.3. Стартовое окно модуля с перечнем статистических процедур

Descriptive statistics – Описательные статистики;

Correlation matrices- Корреляционные матрицы;

t-test for independent samples - t-тест для независимых выборок;

t-test for dependent samples - t-тест для зависимых выборок;

Breakdown & one-way ANOVA - Классификация и однофакторный дисперсионный анализ и др.

После выбора процедуры *Descriptive statistics* на экране появится одноименное диалоговое окно (рис. 4).

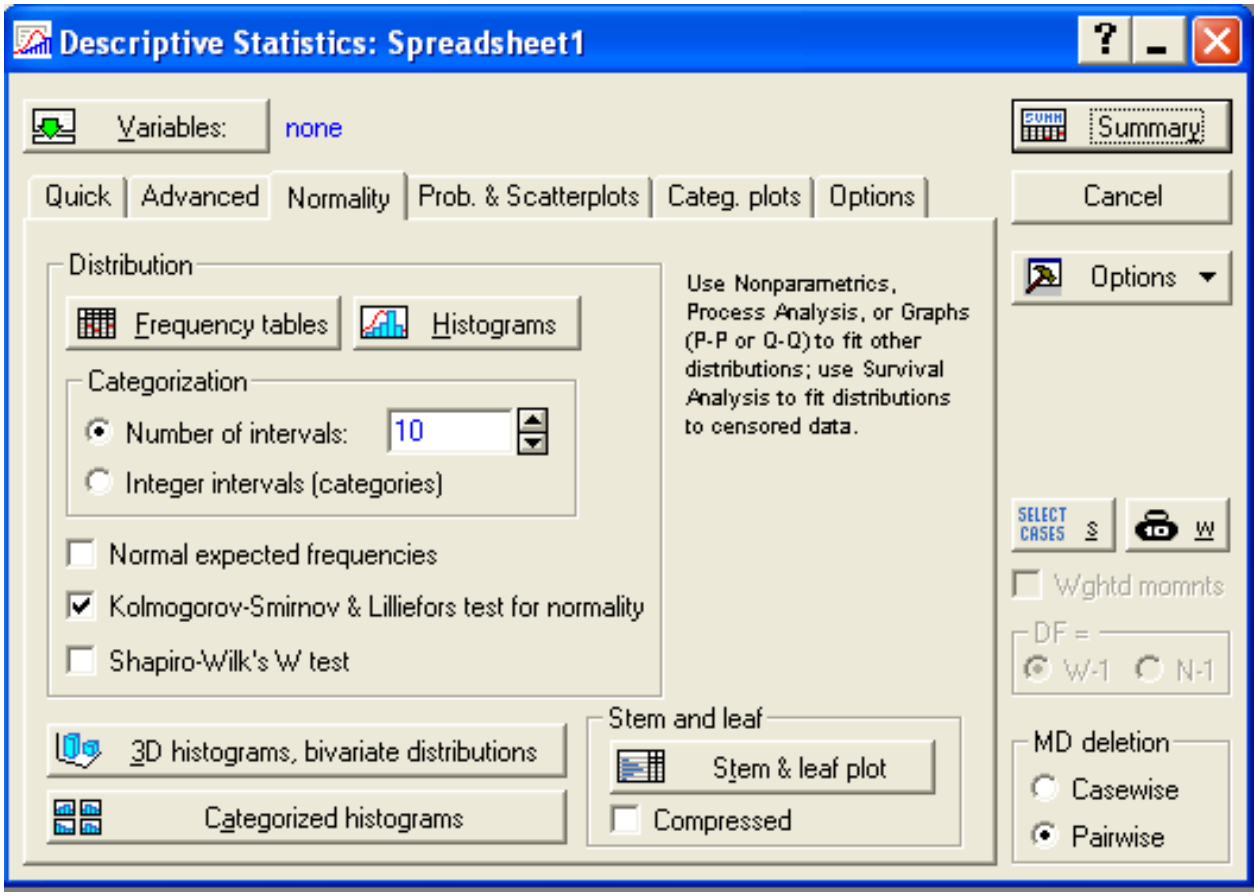

Рис. 4. Диалоговое окно «*Descriptive statistics*»

В этом окне при помощи кнопки *Variables* следует выбрать переменные для анализа.

На первом этапе обработки данных часто возникает необходимость в их группировке. Группировка позволяет представить первичные данные в компактном виде, выявить закономерности варьирования изучаемого признака. Количество классов можно приблизительно наметить, придерживаясь следующих правил: при количестве наблюдений 25-40 – 5-6 классов, при количестве наблюдений 40-60 – 6-8 классов, при количестве наблюдений 60-100 – 7-10 классов, при количестве наблюдений 100-200 – 8-12 классов, более 200 наблюдений – 10-15 классов.

Для построения гистограмм и таблиц часто используется группа кнопок *Distribution* окна *Descriptive statistics*. Число классов (интервалов) группировки данных устанавливается при помощи счетчика переключателя *Number of inter-*

vals в закладке *Normality* окна *Descriptive statistics*. Ниже кнопок *Distribution* находятся две опции *Categorization* (Группировка), позволяющие задать число интервалов группировки или установить величину интервала равную целому числу. Если заактивировать переключатель *Integer intervals (categories)*, то классы (интервалы) группировки будут представлять собой целые числа. Результаты группировки представляются в виде таблицы 1:

Таблица 1

Интервал	Count (количество)	Cumul. Count (кол-во с накоплением)	Percent of Valid (%)	Cumul. % of Valid (% с накоплением)	% of all Cases (% от общего кол-ва)

Для распределения переменных на гистограммах предназначена кнопка *Histograms* окна *Descriptive statistics*. На гистограмму при необходимости можно наложить плотность нормального распределения, проверить близость распределения к нормальному виду при помощи критериев Колмогорова-Смирнова, Лилиефорса; Вычислить статистику Шапиро – Уилкса. Для этого в группе опций *Distribution* необходимо установить флажок напротив соответствующих статистик. Значения статистик показываются прямо на гистограммах.

Чтобы выбрать статистики, подлежащие вычислению, удобнее всего пользоваться закладкой *Advanced* окна *Descriptive statistics* (рис. 5).

1) Valid N - объем выборки;

2) Mean - средняя арифметическая;

Среднее значение случайной величины представляет собой наиболее типичное, наиболее вероятное ее значение, своеобразный центр, вокруг которого разбросаны все значения признака.

3) Sum **-** сумма**;**

4) Median - медиана**;**

Медианой является такое значение случайной величины, которое разделяет все случаи выборки на две равные по численности части.

5) Standard Deviation - стандартное отклонение;

Стандартное отклонение (или среднее квадратическое отклонение) является мерой изменчивости (вариации) признака. Оно показывает, на какую величину в среднем отклоняются случаи от среднего значения признака.

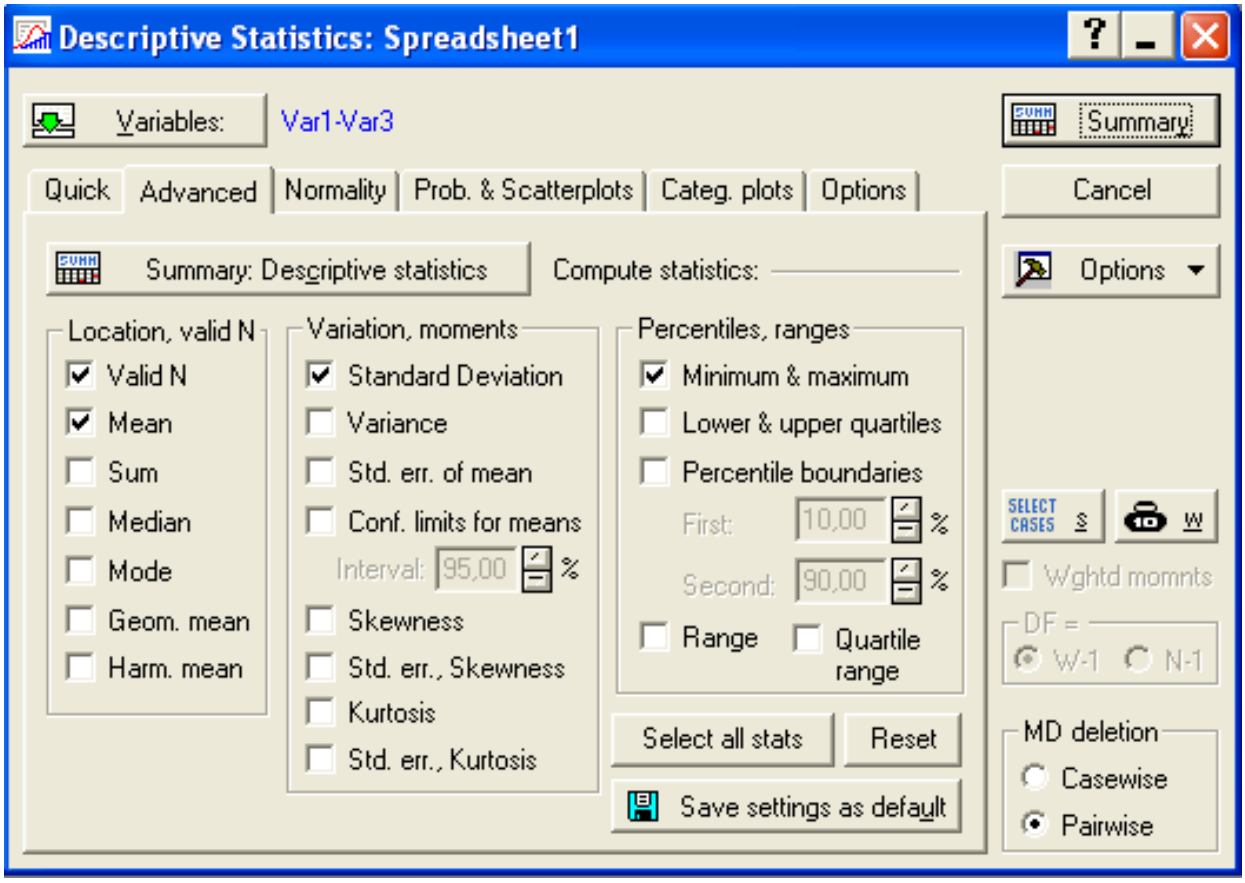

Рис. 5. Окно выбора статистик.

Особенно большое значение среднее квадратическое отклонение имеет при исследовании нормальных распределений. В нормальном распределении 68% всех случаев лежит в интервале ± одного отклонения от среднего, 95% - ± двух стандартных отклонений от среднего и 99,7% всех случаев - в интервале ± трех стандартных отклонений от среднего.

6) Variance - дисперсия;

Дисперсия является мерой изменчивости, вариации признака и представляет собой средний квадрат отклонений случаев от среднего значения признака. В отличие от других показателей вариации дисперсия может быть разложена на составные части, что позволяет тем самым оценить влияние различных факторов на вариацию признака. Дисперсия - один из существеннейших показателей,

характеризующих явление или процесс, один из основных критериев возможности создания достаточно точных моделей.

7) Standard error of mean - стандартная ошибка среднего;

Стандартная ошибка среднего это величина, на которую отличается среднее значение выборки от среднего значения генеральной совокупности при условии, что распределение близко к нормальному. С вероятностью 0,68 можно утверждать, что среднее значение генеральной совокупности лежит в интервале ± одной стандартной ошибки от среднего, с вероятностью 0,95 – в интервале ± двух стандартных ошибок от среднего и с вероятностью 0,99 – среднее значение генеральной совокупности лежит в интервале ± трех стандартных ошибок от среднего.

8) 95% *confidence limits of mean* - 95%-ый доверительный интервал для среднего;

Интервал, в который с вероятностью 0,95 попадает среднее значение признака генеральной совокупности.

9) Minimum, maximum - минимальное и максимальное значения;

10) Lower, upper quartiles - нижний и верхний квартили;

Нижний квартиль это такое значение случайной величины, больше которого по величине 25% случаев выборки. Верхний квартиль это такое значение случайной величины, меньше которого по величине 25% случаев выборки.

11) Range - размах;

Расстояние между наибольшим (maximum) и наименьшим (minimum) значениями признака.

12) Quartile range - интерквартильная широта;

Расстояние между нижним и верхним квартилями.

13) Skewness -асимметрия;

Асимметрия характеризует степень смещения вариационного ряда относительно среднего значения по величине и направлению. В симметричной кривой коэффициент асимметрии равен нулю. Если правая ветвь кривой, начиная от вершины) больше левой (правосторонняя асимметрия), то коэффициент

асимметрии больше нуля. Если левая ветвь кривой больше правой (левосторонняя асимметрия), то коэффициент асимметрии меньше нуля. Асимметрия менее 0,5 считается малой.

14) Standard error of Skewness -стандартная ошибка асимметрии;

15) Kurtosis - эксцесс**;**

Эксцесс характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутости кривой. В кривой нормального распределения эксцесс равен нулю. Если эксцесс больше нуля, то кривая распределения характеризуется островершинностью, т.е. является более крутой по сравнению с нормальной, а случаи более густо группируются вокруг среднего. При отрицательном эксцессе кривая является более плосковершинной, т.е. более пологой по сравнению с нормальным распределением. Отрицательным пределом величины эксцесса является число -2, положительного предела нет.

16) Standard error of Kurtosis - стандартная ошибка эксцесса.

Напротив статистик, подлежащих вычислению (рис. 5) следует поставить флажок.

После нажатия на кнопку ОК окна *Descriptive statistics* на экране появится таблица с результатами расчетов описательных статистик.

К сожалению, пакет *Statistica* не рассчитывает такие часто применяемые статистики, как коэффициент вариации и относительная ошибка среднего значения (точность опыта). Но их определение не представляет большого труда. Коэффициент вариации (%) есть отношение стандартного отклонения к среднему значению, умноженное на 100%:

Коэффициент вариации = (*Standard Deviation* / *Mean*)• 100%. Коэффициент вариации, как дисперсия и стандартное отклонение, является показателем изменчивости признака. Коэффициент вариации не зависит от единиц измерения, поэтому удобен для сравнительной оценки различных статистических совокупностей. При величине коэффициента вариации до 10% изменчивость оценивается как слабая, 11-25% - средняя, более 25% - сильная.

Относительная ошибка среднего значения (%) - отношение стандартной ошибки среднего к среднему значению, умноженное на 100% (для вероятности 0,68).

Относительная ошибка среднего значения = (*Standard error of mean/ Mean)·* 100%. Это процент расхождения между генеральной и выборочной средней, показывает, на сколько процентов можно ошибиться, если утверждать, что генеральная средняя равна выборочной средней. Если относительная ошибка не превышает 5%, то точность исследований (точность опыта) оценивается как хорошая, до 10% -удовлетворительная. Точность 3-5% при вероятности 0,95, а в некоторых случаях и при вероятности 0,68, является вполне достаточной для большинства задач.

Задание:

В плане комплексного исследования личности у студентов психологического факультета были получены оценки социометрического статуса в своей учебной группе (x_1), нейротизма по Айзенку (x_2) и эмоциональной экспансивности 42 студентов (x_3), представленные в таблице 2.

1. *Создать файл данных.*
2. *Сгруппировать данные.*
3. *Сравнить распределение социометрического статуса с нормальным с помощью гистограммы.*
4. *Вычислить числовые характеристики распределения показателей и сделать выводы по каждой.*

Таблица 2

№	X_1	X_2	X_3	№	X_1	X_2	X_3
1	1	18	3	22	8	7	-8
2	-9	20	-15	23	8	5	18
3	22	18	15	24	-2	17	1
4	-11	22	-1	25	-28	12	3
5	-6	9	-26	26	2	9	-1
6	-0,5	12	11	27	15	12	34
7	5	13	2	28	-10	8	13
8	9	16	10	29	4	9	9

Продолжение таблицы 2

9	0	14	-4	30	19	3	12
10	2	16	13	31	2	15	22
11	11	16	-17	32	6	14	29
12	15	17	9	33	17	9	22
13	-8	21	-27	34	7	14	7
14	0	22	25	35	8	7	40
15	9	7	61	36	4	13	40
16	8	23	33	37	7	6	20
17	15	11	-2	38	10	16	-30
18	4	12	34	29	4	12	19
19	-8	15	-4	40	0,5	20	10
20	-0,7	12	-6	41	8	12	0
21	-15	13	-4	42	9	8	24

Лабораторная работа № 3

Корреляционный анализ.

Задача корреляционного анализа сводится к установлению направления и формы между варьирующими признаками, измерению тесноты, и, наконец, к проверке значимости коэффициентов корреляции. Варьирующие признаки могут быть измерены в разных шкалах, что определяет выбор соответствующего коэффициента корреляции.

В программе *STATISTICA* предусмотрены процедуры позволяющие оценивать коэффициенты корреляции Пирсона, Кендала, Спирмена и др.

Процедура *Correlation matrices*, которая вызывается при помощи пункта модуля Basic Statistic/Tables, предназначена для нахождения линейного коэффициента корреляции Пирсона и установления тесноты линейной связи между переменными, измеренными в интервальной шкале.

В стартовом окне этой процедуры (рис. 6) для расчета квадратной матрицы используется кнопка *One variable list*. В списке переменных выбирают переменные, между которыми будут рассчитаны парные коэффициенты корреляции Пирсона. После нажатия на кнопку ОК и *Summary: Correlation matrix* на экране появится корреляционная матрица.

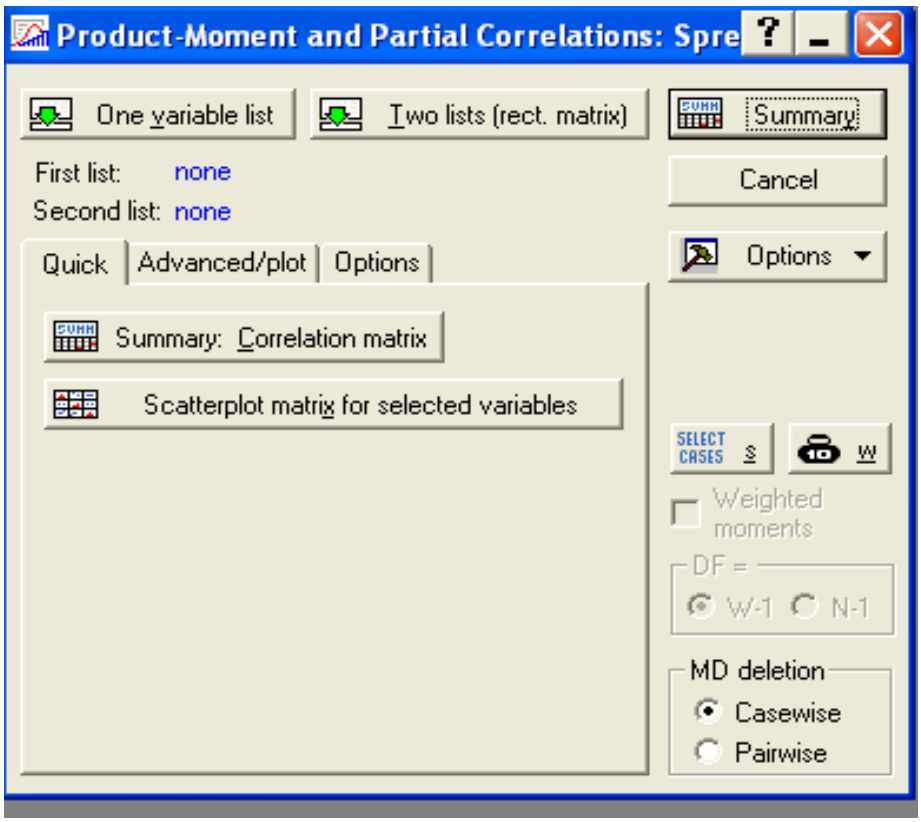

Рис. 6. Диалоговое окно корреляционного анализа.

Коэффициент корреляции - это показатель, оценивающий тесноту линейной связи между признаками. Он может принимать значения от -1 до +1. Знак «–» означает, что связь обратная, «+» – прямая. Чем ближе модуль коэффициента к 1 тем теснее линейная связь. При величине модуля коэффициента корреляции (по Дворецкому) менее 0,3 связь оценивается как слабая, от 0,31 до 0,5 - умеренная, от 0,51 до 0,7 - значительная, от 0,71 до 0,9 - тесная, 0,91 и выше - очень тесная. Для практических целей Дворецкий рекомендует использовать значительные, тесные и очень тесные связи. Процедура *Correlation matrices* сразу же дает возможность проверить достоверность рассчитанных коэффициентов корреляции. Значение коэффициента корреляции может быть высоким, но не достоверным, случайным, чтобы увидеть вероятность нулевой гипотезы (p), гласящей о том, что коэффициент корреляции равен 0, нужно в *Options* окна установить переключатель на вторую строку *Display r, p-ievels, and N's*. Но даже если этого не делать и оставить переключатель в первом положении статистически значимые на 5-% уровне коэффициенты корреляции будут выделены в корреляционной матрице на экране монитора цветом, а при распечатке по-

мечены звездочкой. Третье положение переключателя опции позволяет просмотреть результаты корреляционного анализа в деталях.

Процедура *Correlations (Spearman, Kendal tau, gamma)*, которая вызывается с помощью модуля *Nonparametrics* (Непараметрические критерии), находящегося в пункте меню *Statistica* (рис.7) позволяет оценивать связь между переменными, измеренными в разных шкалах.

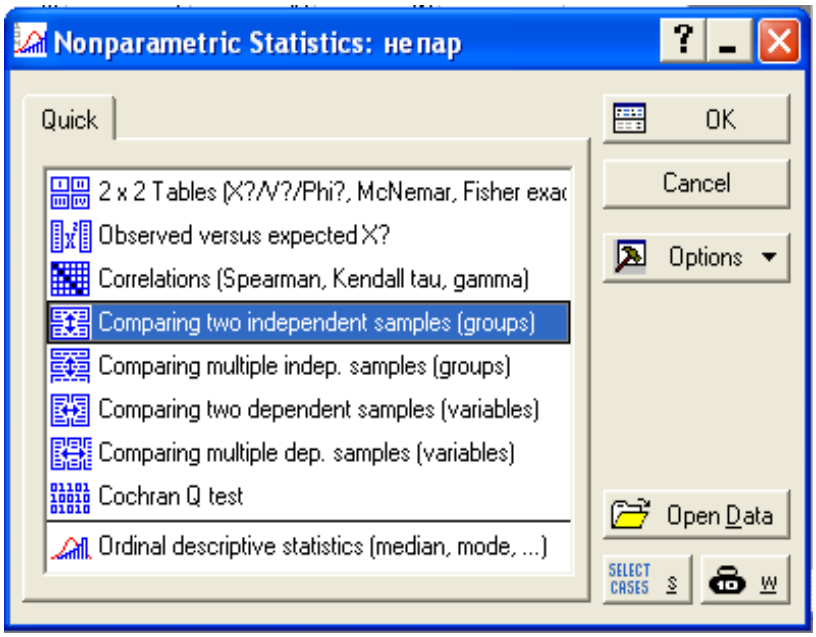

Рис.7. Диалоговое окно непараметрических критериев

Другой вариант рассмотрения взаимосвязи между переменными выделяет одну из величин как независимую *x,* а другую как зависимую *y.* И изучает влияние этих переменных друг на друга.

Зависимость *y=f(x)* называется функцией регрессии *y* на *x.*

Если рассматривается зависимость двух величин, то регрессия называется парной.

Для определения вида парной регрессии в декартовой системе координат строят точки наблюдений и соединяют их отрезками. Полученную линию называют эмпирической линией регрессии.

По внешнему виду эмпирической линии регрессии определяют плавную кривую, около которой группируются все точки наблюдений. Эту кривую называют теоретической линией регрессии или регрессией.

Самой простой парной регрессией является линейная регрессия с уравнением: $y = a + bx + e$, где a и b - коэффициенты регрессии, которые находятся с помощью метода наименьших квадратов, e - случайный член.

b - показывает на сколько изменяется y при изменении переменной x на 1 единицу;

a - это первоначальное значение y при $x=0$.

Используя кнопку *2D scatter plot* вкладки *Advanced/plot* можно получить график теоретической линии регрессии и уравнение регрессии.

Задание:

Согласно условию задачи, сформулированной в лабораторной работе №2 выполнить:

1. *Вычислите коэффициенты корреляции между всеми переменными и проверить их значимость. Сделайте выводы.*
2. *Переименуйте строки, присвоив им значения: «м» (мужчина) первым 21 строкам и «ж» (женщина) последним 21 строкам. Определите существует ли связь между полом и оценками социометрического статуса.*
3. *Постройте уравнение регрессии x_1 на x_2 и x_1 на x_3. Сделайте выводы.*
4. *Постройте теоретическую линию регрессии. Измените название графика и вид линии. Измените фон рисунка*.

Лабораторная работа №4

Модель множественной регрессии

1. Линейная модель

На любой статистический показатель действует чаще всего не один, а несколько факторов. Все их следует включить в модель, т.е. построить уравнение множественной регрессии. Рассмотрим самую употребляемую и наиболее простую из моделей множественной регрессии - модель множественной линейной регрессии вида:

$y = \alpha + \beta x_1 + \beta x_2 + \ldots + \beta x_n + \varepsilon,$

где y - зависимая переменная,

$x_1, x_2, \ldots, x_n$ - независимые переменные,

ε -случайная величина.

Модель множественной регрессии в *Statistica* строится при помощи модуля *Multiple Regression*.

В стартовом окне модуля (рис. 8) при помощи кнопки *Variabies* указывается зависимая (dependent) и независимые (independent) переменные.

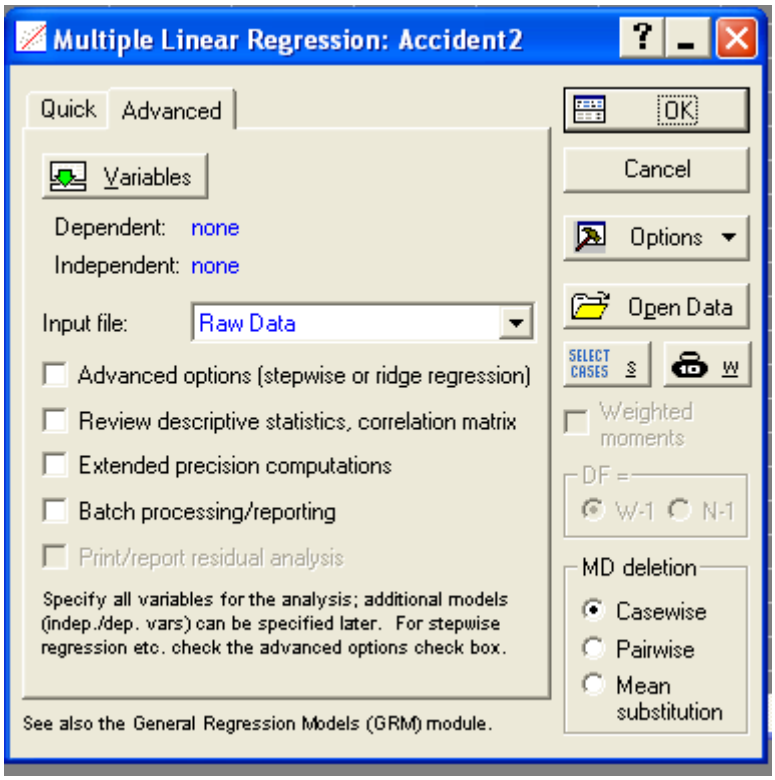

Рис.8. Диалоговое окно построения линейной регрессии.

В поле *Input file* указывается тип файла с данными:

Raw Date - данные в виде строчной таблицы;

Correlation matrix - данные в виде корреляционной матрицы.

В поле *MD deletion* указывается способ исключения из обработки недостающих данных:

Casewise – игнорируется вся строка, в которой есть хотя бы одно пропущенное значение;

Pairwise – попарное исключение данных с пропусками из тех переменных, корреляция которых вычисляется;

Mean Substitution – взамен пропущенных данных подставляется среднее значение переменных.

С помощью кнопки *Select cases* можно установить условие включения строк в статистическую обработку (рис.9).

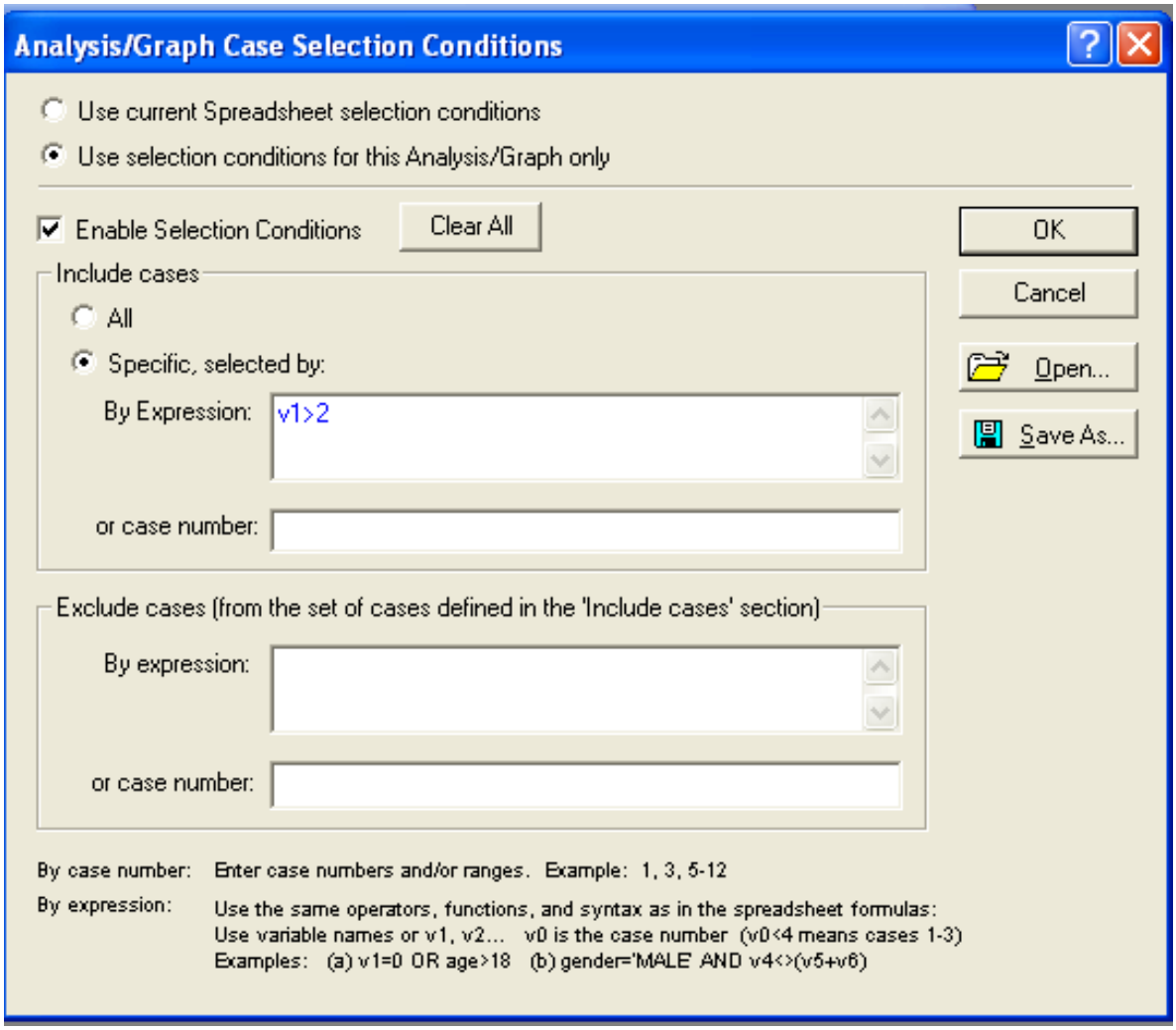

Рис. 9. Окно установки условий для переменных.

После того, как все опции стартового диалогового окна регрессионного анализа выставлены, нажатие на кнопку OK приведет к появлению окна *Multiple Regressions Results* (результаты регрессионного анализа), с помощью которого результаты анализа можно просмотреть в деталях (рис. 10).

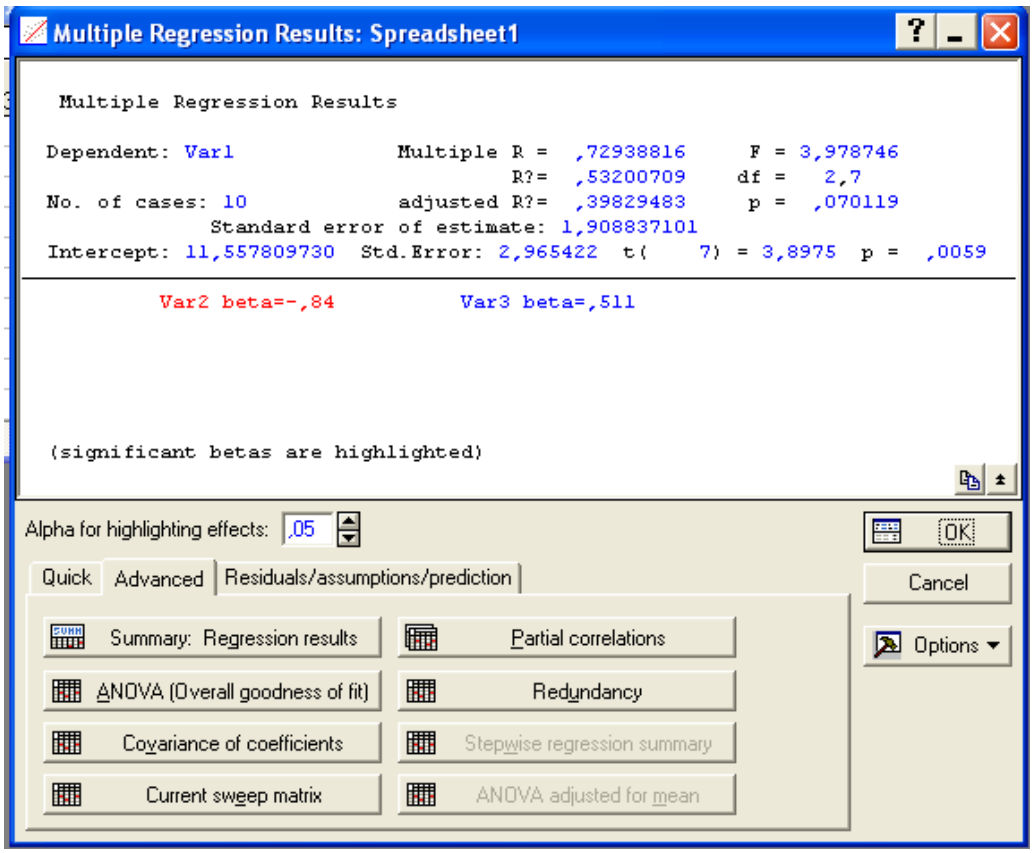

Рис. 10. Окно результатов регрессионного анализа.

В верхней части окна приводятся наиболее важные параметры полученной регрессионной модели:

Multiple R – коэффициент множественной корреляции (характеризует тесноту линейной связи между зависимой и всеми независимыми переменными. Может принимать значения от 0 до 1);

R? или *R2* – коэффициент детерминации (численно выражает долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения. Чем больше *R2* , тем большую долю вариации объясняют переменные, включенные в модель);

adjusted R? или *adjusted R2* – скорректированный коэффициент детерминации;

F – F-критерий Фишера;

df – число степеней свободы для F-критерия;

p – вероятность нулевой гипотезы для F-критерия;

Standard error of estimate – стандартная ошибка оценки уравнения;

Intercept – свободный член уравнения;

Std. Error – стандартная ошибка свободного члена уравнения;

t - t – критерий Стьюдента для свободного члена уравнения;

p – вероятность нулевой гипотезы для свободного члена уравнения;

beta – β-коэффициенты уравнения. Это стандартизированные регрессионные коэффициенты, рассчитанные по стандартизированным значениям переменных. По их величине можно сравнить и оценить значимость зависимых переменных, так как β-коэффициент показывает на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной при условии постоянства остальных независимых переменных. Свободный член в таком уравнении равен 0.

При помощи кнопок диалогового окна *Multiple Regressions Results* результаты регрессионного анализа можно просмотреть более детально.

Кнопка *Summary Regressions Results* позволяет просмотреть результаты регрессионного анализа (рис. 11):

B – оценки коэффициентов регрессии;

St. Err. of B – стандартные ошибки коэффициентов;

t(4) – *t*-критерии для коэффициентов уравнения регрессии;

p-level – вероятность нулевой гипотезы для коэффициентов уравнения регрессии.

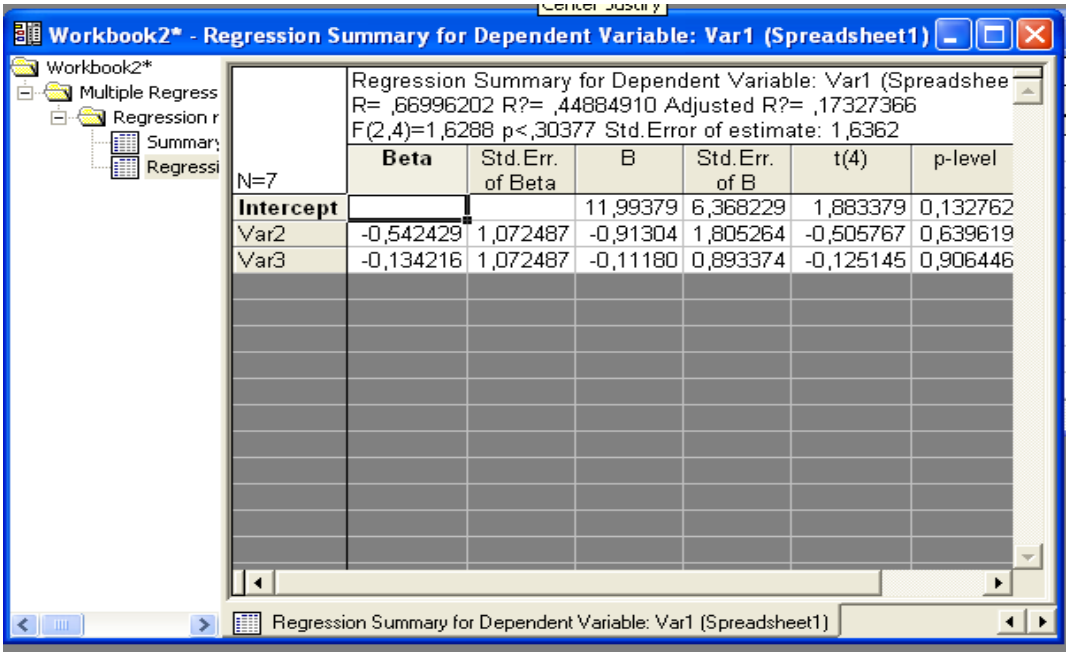

Рис. 11. Окно анализа коэффициентов регрессии.

Согласно данным, представленным на рисунке 11, уравнение регрессии выглядит следующим образом: v1=11,99379-0,91304v2 - 0,11180v3. Все коэффициенты уравнения незначимы на 5% уровне значимости. Это уравнение объясняет всего 44,88 % вариации зависимой переменной (коэффициент детерминации равен 0,4488491).

Кнопка *ANOVA* позволяет ознакомиться с результатами дисперсионного анализа уравнения регрессии (рис.12):

Regress – дисперсия, обусловленная регрессией;

Residual – остаточная дисперсия;

Total – общая дисперсия;

F – F-критерий;

p-level – вероятность нулевой гипотезы для F-критерия.

F-критерий полученного уравнения на 5% уровне значимости: Вероятность нулевой гипотезы больше 0,05, что говорит о незначимости регрессии.

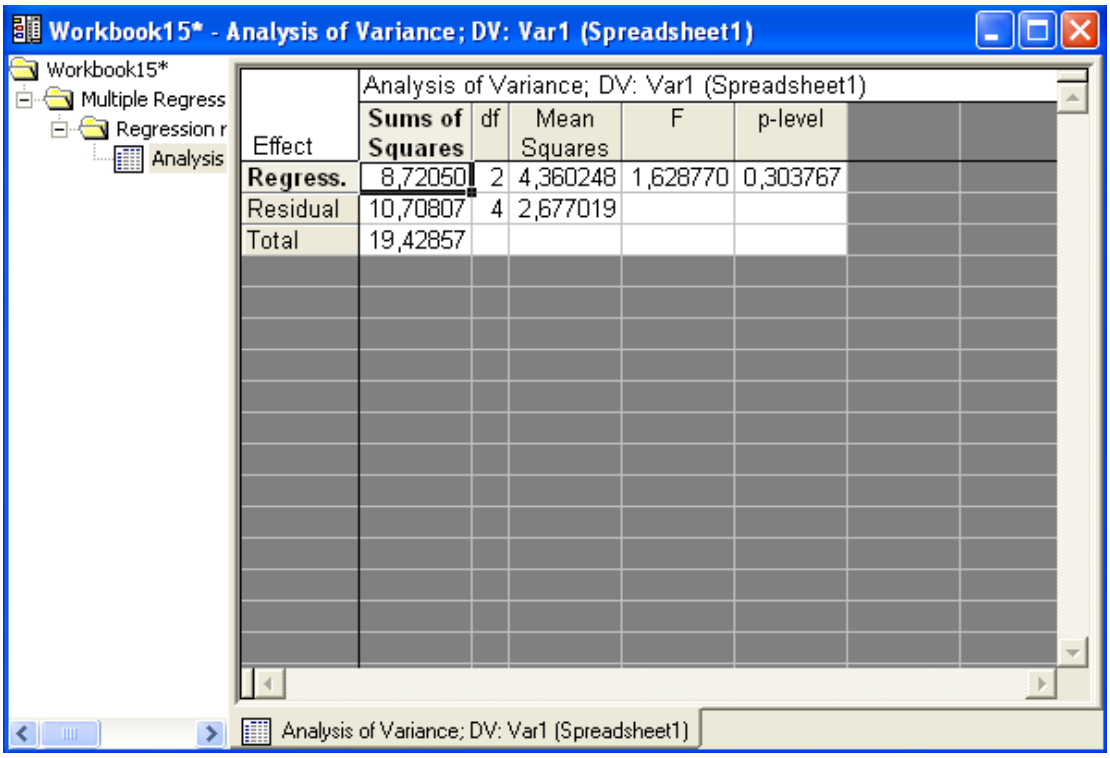

Рис. 12. Окно результатов дисперсионного анализа

Используя закладку *Predicting Values* окна *Multiple Regressions Results* можно прогнозировать значение зависимой переменной при заданных значениях независимых переменных.

Задание:

1. Откройте файл лабораторной работы № 2

2. Постройте модели зависимости переменной x_1 от переменных x_2 и x_3 отдельно для каждого пола и в целом для группы.

3. Запишите полученные модели. Объясните значение коэффициентов детерминации. Проверьте значимость коэффициентов уравнений регрессии используя t – критерий Стьюдента и значимость уравнений в целом используя F-критерий.. Запишите проверяемые гипотезы.

4. Спрогнозируйте значения зависимых переменных. Значения x_2=16, x_3=-8.

5.Сделайте выводы.

2. Нелинейная модель.

В силу многообразия и сложности реальных процессов многие зависимости не являются линейными и, следовательно, требуют моделирования нелинейными уравнениями.

Рассмотрим наиболее часто встречающиеся модели. Для простоты изложения ограничимся моделями парной нелинейной регрессии.

1. Логарифмические модели.

– $\ln Y = \beta_0 + \beta \cdot \ln X + \varepsilon$ - *двойная логарифмическая модель*

коэффициент β в данной модели определяет процентное изменение Y для данного процентного изменения X.

– $\ln y = a + b \cdot x + \varepsilon$ - *лог-линейная модель*

коэффициент b в данной модели характеризует отношение относительного изменения *y* к абсолютному изменению *x*.

– $y = a + b \cdot \ln x + \varepsilon$ - *линейно-логарифмическая модель*, используется, когда необходимо исследовать влияние процентного изменения независимой переменной на абсолютное изменение зависимой переменной.

2. *Гиперболическая модель:* $y = a + b \cdot \frac{1}{x} + \varepsilon$ - применяется в тех случаях, когда неограниченное увеличение значений объясняющей переменной x асимптотически приближает зависимую переменную y к некоторому пределу a.

3. *Полиномиальная модель:* $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$

4. *Степенная модель:* $y = ax^b$

5. *Показательная модель:* $y = a \cdot e^{b \cdot x}$

Каждая модель путем математических преобразований может быть приведена к линейному виду.

Для случая парной регрессии подбор модели обычно осуществляется по виду расположения наблюдаемых точек на корреляционном поле. В случае если зависимость может быть описана несколькими функциями, необходимо выбрать ту из них, которая обладает наилучшим качеством. Но следует помнить, что чем сложнее модель, тем менее интерпретируемы ее параметры.

Для построения точек в системе координат выберите в строке меню *Graphs* (Графики) пункт *Scatterplots*. В открывшемся окне, используя закладку *Advanced* можно указать тип графика и предполагаемый вид функциональной зависимости. В поле *Graph type* (Тип графика) выбираем *Regular*, в поле *Fit* (Подгонка) выбираем *Off* (неопределенный). Далее задаем переменные и нажимаем кнопку *OK* (рис.13).

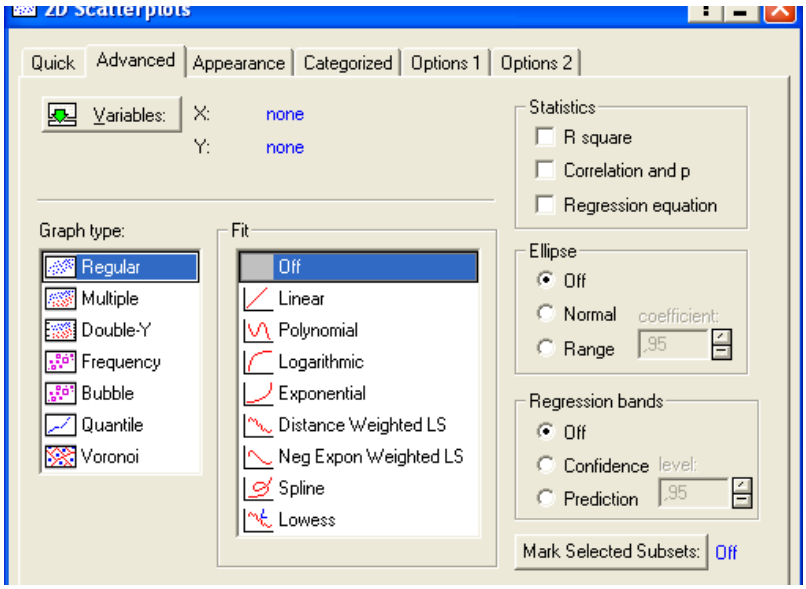

Рис.13. Построение корреляционного поля.

Построить модель нелинейной регрессии можно с помощью пункта *Fixed Nonlinear Regression* (фиксированная нелинейная модель, т.е. нелинейная модель, но она может быть приведена к линейному виду путем преобразования переменных) модуля *Advanced Linear/ Nonlinear Models*. После указания в этом окне переменных, активировав *Extended precision computation* (вычисление с расширенной точностью) и *OK*, откроем окно *Non-linear Components Regression*, в котором можно выбрать типы преобразования переменных в поле *Non-linear Transformation Functions* (рис.14).

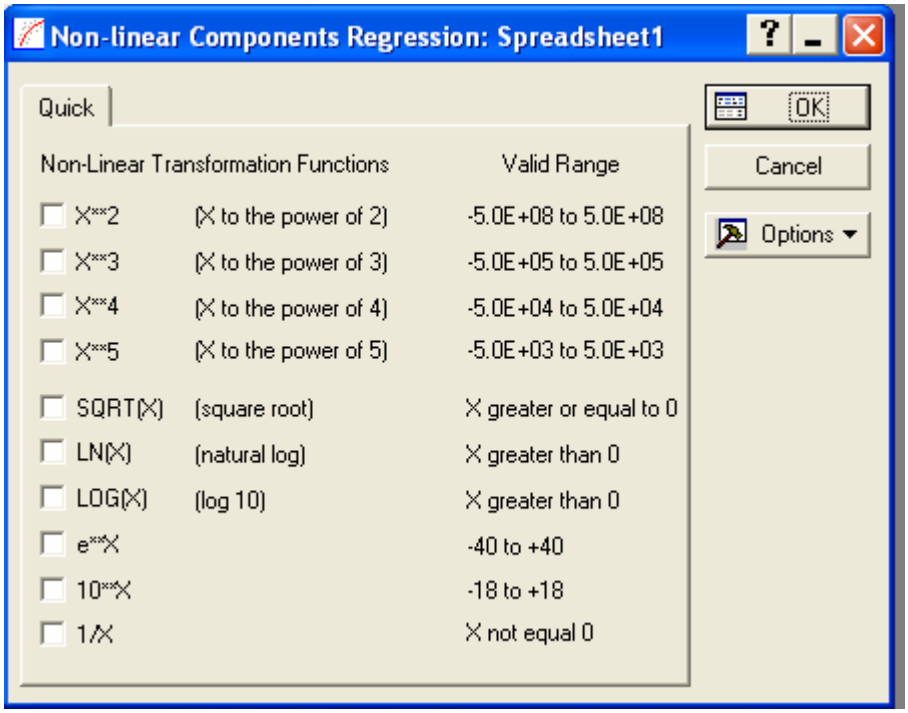

Рис. 14. Окно выбора преобразования переменных.

Для построения квадратичной модели выбирается преобразование X**2 .

После нажатия *OK* откроется окно *Model Definition*, в котором нужно указать переменные для построения модели.

Напомним вид квадратичной функции: $y=ax^2+bx+c$, т.е. ее можно рассматривать как модель множественной регрессии с независимыми переменными x и x^2.

Задание: Психолог у восьми подростков сравнивает баллы по третьему, математическому, субтесту теста Векслера и оценки по алгебре. Данные представлены в таблице 3:

Таблица 3

№ испытуемого	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Баллы	*8*	*18*	*18*	*10*	*16*	*10*	*8*	*14*
Оценки	*2*	*3*	*4*	*5*	*4*	*4*	*3*	*5*

1. *Оцените линейную связь между переменными. Сделайте вывод.*
2. *Постройте точечный график данных (по оси абсцисс – баллы по тесту Векслера, по оси ординат – оценки по алгебре).*
3. *По построенному точечному графику сделайте предположение о виде зависимости (квадратичная, логарифмическая и т.д.).*
4. *Оцените полученную модель. Сделайте выводы.*

Лабораторная работа № 5

Непараметрические критерии

Непараметрические критерии не включают в формулу расчета параметров распределения, основаны на оперировании частотами или рангами, присваемыми разным показателям группой экспертов.

Рассмотрим некоторые непараметрические критерии.

1. Критерии различий

U- критерий Манна-Уитни - предназначен для оценки различий между двумя выборками по уровню какого-либо признака, количественно измеренного.

Эмпирическое значение критерия рассчитывается по формуле:

$$U_{эмп} = (n_1 \cdot n_2) + \frac{n_x \cdot (n_x + 1)}{2} - T_x,$$

где n_1, n_2 - количество испытуемых в выборках 1 и 2 соответственно,

T_x – большая из ранговых сумм,

n_x – количество испытуемых в группе с большей суммой рангов.

Критическое значение критерия определяется для данных n_1 и n_2 и выбранного уровня значимости. Если $U_{эмп}$ больше $U_{кр}$, то принимается H_o.

H - критерий Крускала – Уоллиса - предназначен для оценки различий между тремя и более выборками по уровню какого-либо признака. Он позволя-

ет установить, что уровень признака изменяется при переходе от группы к группе, но не указывает направление этих изменений.

Эмпирическое значение критерия H подсчитывается по формуле:

$$H_{эмп} = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{c} \frac{T_i^2}{n_i} - 3 \cdot (N+1),$$

где N – общее количество испытуемых в объединенной выборке,

n_i – количество испытуемых в каждой выборке,

T_i^2 – квадраты сумм рангов по каждой i–ой выборке.

Если эмпирическое значение критерия меньше критического значения, то H_o принимается.

2. Критерии изменений

Парный критерий T – Вилкоксона - применяется для сопоставления показателей, измеренных в двух разных условиях на одной и той же выборке испытуемых. Он позволяет установить не только направленность изменений, но и их выраженность. С его помощью мы определяем, является ли сдвиг показателей в каком-то одном направлении более интенсивным, чем в другом.

Эмпирическое значение критерия подсчитывают по формуле:

$$T_{эмп} = \sum R_r ,$$

где R_r – ранговые значения сдвигов с более редким знаком.

Критическое значение критерия $T_{кр}$ определяется для данного объема выборки и выбранного уровня значимости. Если $T_{эмп}$ не превосходит $T_{кр}$, то гипотеза H_0 отвергается.

Критерий χ_r^2 Фридмана - применяется для сопоставления показателей, измеренных в трех или более условиях на одной и той же выборке испытуемых. Он позволяет установить, что величины показателей от условия к условию изменяются, но при этом не указывает на направление изменений.

Эмпирическое значение критерия вычисляется по формуле:

$$\chi^2_{r\,эмп} = \left[\frac{12}{n \cdot c \cdot (c+1)} \cdot \sum_{i=1}^{c} T_i^2 \right] - 3 \cdot n \cdot (c+1) ,$$

где c – количество условий,

n – количество испытуемых,

T_i – суммы рангов по каждому из условий.

Критическое значение критерия $\chi_{r\,кр}^2$ определяем при выбранном уровне значимости и данном объеме выборки. Если $\chi_{r\,эмп}^2$ не меньше $\chi_{r\,кр}^2$ то гипотеза H_0 отклоняется.

Для проверки гипотез в программе STATISTICA с помощью непараметрических критериев используется модуль *Nonparametrics* (Непараметрические критерии), находящийся в пункте меню *Statistica* (рис.15).

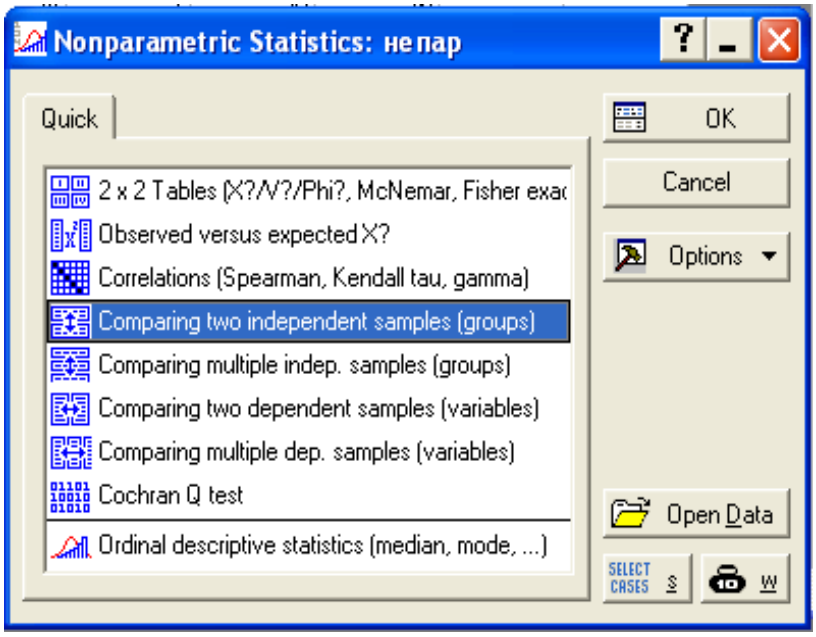

Рис.15. Диалоговое окно непараметрических критериев.

Пункт *Comparing two independent samples* (Сравнение двух независимых выборок), позволяет реализовать критерий Манна-Уитни *(Mann-Whitney U test)*.

Пункт *Comparing multiple independent samples* (Сравнение множества независимых выборок) – критерий Крускала-Уоллиса (*Kruskal-Wallis)*

Пункт *Comparing two dependent samples* (Сравнение двух зависимых выборок) - критерий Вилкоксона (*T Wilcoxon)*.

Пункт *Comparing multiple dependent samples* (Сравнение множества зависимых выборок) – критерий Фридмана *(Friedman)*.

Появившееся в процессе анализа диалоговое окно позволяет определить эмпирическое значение соответствующего критерия и вероятность нулевой гипотезы *p-level* (при $p \le 0{,}05$ нулевая гипотеза отвергается)

Задание: Решите с помощью программы STATISTICA следующие задачи.

Задача 1. Используя критерий Манна-Уитни определить, можно ли утверждать, что студенты-психологи превосходят студентов-физиков по уровню невербального интеллекта. Данные, полученные с помощью методики Д. Векслера, приведены в таблице 4.

Таблица 4

Студенты-физики		Студенты - психологи	
Код имени испытуемого	Показатель невербального интеллекта	Код имени испытуемого	Показатель невербального интеллекта
1. И.А.	111	1. Н.Т.	113
2. К.А.	104	2. ОБ.	107
3. К.Е.	107	3. Е.В.	123
4. ПА.	90	4. Ф.О.	122
5. С.А.	115	5. И.Н.	117
6. Ст.А.	107	6. И.Ч.	112
7. Т.А.	106	7. И.В.	105
8. ФА.	107	8. К.О.	108
9. Ч.И.	95	9. Р.Р.	111
10. Ц.А.	116	10. Р.И.	114
11. См.А.	127	11. О.К.	102
12. К.Ан.	115	12. Н.К.	104
13. Б.Л.	102		
14. Ф.В.	99		

Замечание: Данные внести в один столбец. Показатели физиков обозначить во втором столбце цифрой 1, психологов – 2.

Задача 2. Четыре группы испытуемых выполняли тест Бурдона в разных экспериментальных условиях. Используя критерий Крускала – Уоллиса Н установить – зависит ли эффективность выполнения теста от условий или существуют ли статистически достоверные различия в успешности выполнения теста между группами. В каждую группу входило четыре испытуемых. Число ошибок показателя переключаемости внимания в процентах дано в таблице 5.

Таблица 5

№ испытуемых	1 группа	2 группа	3 группа	4 группа
1	23	45	34	21
2	20	12	24	22
3	34	34	25	26
4	35	11	40	27
Суммы	112	102	123	96

Данные ввести аналогично предыдущей задаче, кодируя номер выборки. Сделать вывод.

Задача 3. Психолог проводит с младшими школьниками коррекционную работу по формированию навыков внимания, используя для оценки результатов корректурную пробу. Используя критерий Т–Вилкоксона определить, будет ли уменьшаться количество ошибок внимания у младших школьников после специальных коррекционных упражнений? Для решения этой задачи психолог у 19 детей определяет количество ошибок при выполнении корректурной пробы до и после коррекционных упражнений. В таблице 6 приведены соответствующие экспериментальные данные.

Таблица 6

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
До	24	12	42	30	40	55	50	52	50	22	33	78	79	25	28	16	17	12	25
После	22	12	41	31	32	44	50	32	21	34	56	78	23	22	12	16	17	18	25

Замечание: указать переменные: в первом столбце «до», во втором – «после». Далее использовать кнопку *Wilcoxon*.

Задача 4**.** Шести школьникам предъявляют тест Равенна. Фиксируется время решения каждого задания. Выясняется вопрос – будут ли найдены статистически значимые различия между временем решения первых трех заданий теста? Психолог измерил время решения первых трех заданий теста у шести испытуемых. Результаты этих измерений приведены в таблице 7:

Таблица 7

№ испытуемых	Время решения первого задания теста в сек.	Время решения второго задания теста в сек.	Время решения третьего задания теста в сек.
1	8	3	5
2	4	1	12
3	6	23	15
4	3	6	6
5	7	12	3
6	15	24	12

Сформулировать гипотезы. Выполнить их проверку с помощью критерия Фридмана.

Лабораторная работа №6

Параметрические критерии

1. t-критерий Стьюдента используется

a) для сравнения выборочной средней $\bar{x}$ с некоторым известным числовым значением a_0.

H_0: $\bar{x} = a_0$ выборочная средняя генеральной совокупности равна заданному числу a_0.

H_1: $\bar{x} \neq a_0$ ($\bar{x} < a_0$, $\bar{x} > a_0$)выборочная средняя генеральной совокупности не равна (меньше, больше) заданному числу a_0.

b) для обнаружения различия между средними значениями $\bar{x}$, $\bar{y}$ двух выборок.

H_0: $\bar{x} = \bar{y}$ средние значения двух выборок равны,

H_1: $\bar{x} \neq \bar{y}$ средние значения двух выборок не равны.

2. F — критерий Фишера-Снедекора используют

а) для сравнения разброса значений двух выборок, т.е. для проверки гипотезы о равенстве дисперсий.

Возможны гипотезы:

H_0 : $S_x^2 = S_y^2$ - разброс значений признака относительно среднего одинаковый в обеих выборках.

H_1: $S_x^2 \neq S_y^2$ - разброс значений признака не совпадает.

b) для выявлений тенденций изменения признака в трех и более выборках при переходе от условия к условию (однофакторный дисперсионный анализ).

для независимых выборок: (влияние разных условий на разных испытуемых)

H_0 : разные условия не влияют на изменение значений признака;

H_1: условия влияют на изменение значений признака.

для зависимых выборок (одни и те же испытуемые, но в разных условиях) возможно два варианта гипотез:

а) H_0: условия не влияют на изменение признака;

H_1: условия влияют на изменение значений признака.

б) H_0: индивидуальные различия испытуемых не влияют на изменение значений признака;

H_1: индивидуальные различия между испытуемыми влияют на изменение значений признака.

В программе *STATISTICA* можно выполнять при помощи модуля *Basic statistic/tables* следующие процедуры:

t-test for independent samples - t-тест для независимых выборок;

t-test for dependent samples - t-тест для зависимых выборок;

Breakdown & one-way ANOVA - классификация и однофакторный дисперсионный анализ.

t-test for independent samples и *t-test for dependent samples* используются для установления достоверной статистической разницы между средними значениями выборок на основе t-критерия Стьюдента.

Для определения критических значений критериев Стьюдента и Фишера используется пункт меню *Statistics/ Probability Calculator/Distributions.* В диалоговом окне данного пункта необходимо задать вероятность $p=1-\alpha$ и соответствующие степени свободы *df* при нажатии на кнопку *Compute* в окошке для критерия выведется его критическое значение (рис.16).

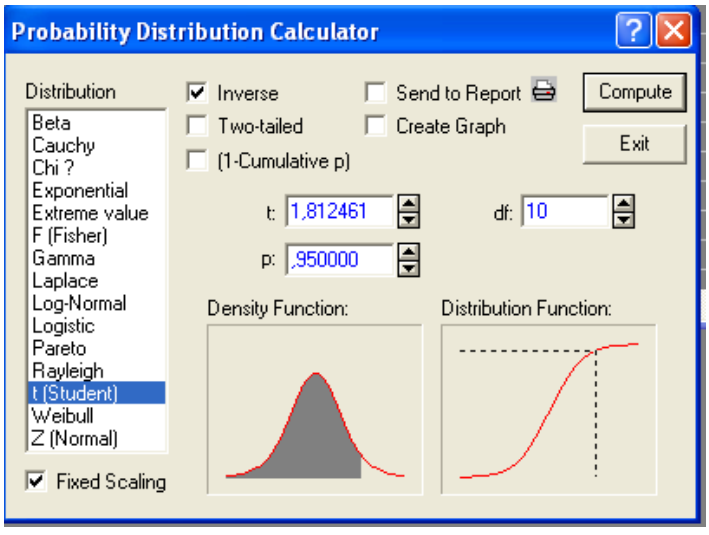

Рис. 16. Окно для определения критических значений критериев.

Задание: Решите с помощью программы STATISTICA следующие задачи.

Задача 1. Два университета (А и В) готовят специалистов аналогичных специальностей. Министерство образования решило проверить качество подготовки в обоих университетах, организовав для этого объемный тестовый экза-

мен для студентов пятого курса. Отобранные случайным образом студенты показали следующие результаты:

A: 41, 50, 35, 45, 53, 30, 57, 20, 50, 44, 36, 48, 55, 28, 40, 50;

B: 40, 57, 52, 38, 20, 25, 47, 52, 48, 55, 48, 53, 39, 49, 46, 45, 55, 43, 51, 55, 40.

Можно ли утверждать при уровне значимости $\alpha = 0{,}05$, что один из университетов обеспечивает лучшую подготовку.

Задача 2. В условиях предыдущей задачи определите, есть ли основания считать, что разброс оценок у студентов одного университета больше чем у другого.

Задача 3**.** Три различные группы из шести человек получили списки из десяти слов. Первой группе слова предъявлялись с низкой скоростью (1 слово в 5 секунд), второй группе – со средней скоростью (1 слово в 2 секунды), третьей – с высокой (1 слово в секунду). Количество воспроизведенных слов представлено в таблице 8. Определить влияет ли скорость предъявления слов на их воспроизведение.

Таблица 8

№	1 группа	2 группа	3 группа
1	8	7	4
2	7	8	5
3	9	5	3
4	5	4	6
5	6	6	2
6	8	7	4

Лабораторная работа № 7

Кластерный анализ.

Назначение кластерного анализа состоит в объединении объектов в достаточно большие кластеры, используя некоторую меру сходства между объектами.

В качестве меры сходства в кластерном анализе применяют расстояние между объектами. Наиболее часто употребляются следующие функции расстояний:

$\rho = \sqrt{\sum (x - y)^2}$ – евклидово расстояние, наиболее общий тип расстояния. Оно является геометрическим расстоянием в многомерном пространстве;

$\rho = \sum (x - y)^2$ – квадрат евклидова расстояния используется для того, чтобы придать большие веса более отдаленным друг от друга объектам;

$\rho = \sum |x - y|$ – расстояние городских кварталов (манхэттенское расстояние) для этой меры влияние отдельных больших разностей (выбросов) уменьшается;

$\rho = \max(x - y)$ – расстояние Чебышева полезно, когда желают определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением);

$\rho = \left(\sum |x - y|^p\right)^{\frac{1}{k}}$ степенное расстояние используют, когда хотят увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Параметры r и p определяются пользователем Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами;

$\rho = \frac{количество\ x_i \neq y_i}{i}$ – процент несогласия используется в тех случаях, когда данные являются категориальными;

Методы связи кластеров.

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако когда связываются вместе несколько объектов необходимо правило объединения или связи для двух кластеров. Существует множество методов объединения кластеров, перечислим наиболее распространенные:

Одиночная связь (метод ближайшего соседа) – расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами.

Полная связь (метод наиболее удаленных соседей) – расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах.

Невзвешенное попарное среднее – расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них.

Взвешенное попарное среднее – идентично методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента.

Невзвешенный центроидный метод – расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

Взвешенный центроидный метод (медиана) – идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них)

Метод Варда. – отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов (SS) для любых двух кластеров, которые могут быть сформированы на каждом шаге. В целом метод представляется очень эффективным, однако он стремится создавать кластеры малого размера.

Типичным результатом кластеризации является иерархическое дерево – дендрограмма. Она начинается с каждого объекта в классе. Постепенно ослабляется критерий о том, какие объекты являются уникальными, а какие нет, т.е. понижается порог, относящийся к решению об объединении двух или более объектов в один кластер. В результате связывается вместе всё большее и большее число объектов и объединяется все больше и больше кластеров, состоящих из всё сильнее различающихся элементов. Окончательно, на последнем шаге все объекты объединяются вместе (рис.17). На этих диаграммах горизонтальные оси представляют расстояние объединения – для каждого узла в

графе можно увидеть величину расстояния, для которого соответствующие элементы связываются в новый единственный кластер.

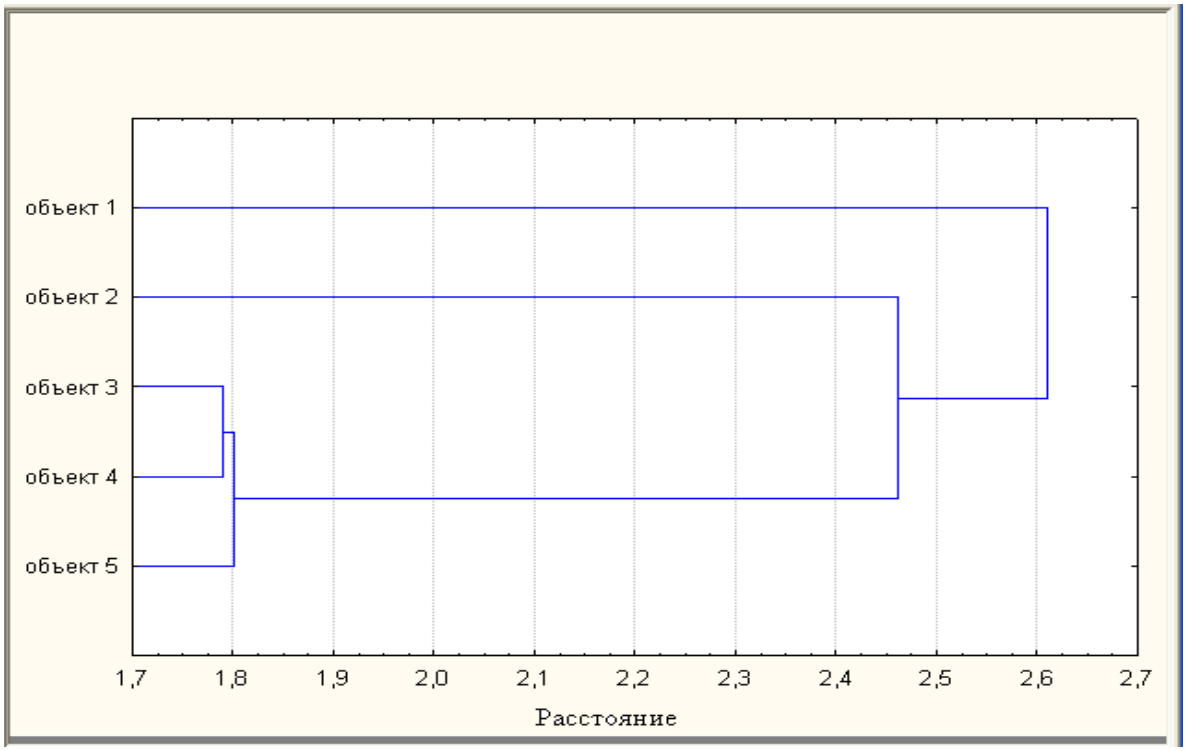

Рис.17. Горизонтальная дендрограмма.

Рассмотрим реализацию методов кластерного анализа с помощью компьютерного пакета *STATISTICA*. Данная программа позволяет осуществлять иерархический агломеративный метод *(joining (tree clustering)),* результатом которого является дендрограмма, и последовательный итерационный метод *(k-means clustering)* с заранее заданным количеством кластеров.

Перед вызовом процедуры кластеризации необходимо стандартизировать данные, для того чтобы привести значения всех переменных к единому диапазону значений. Стандартизация осуществляется путем выбора процедуры *Standardize* (стандартизация) в пункте меню *Data* (переменные). Открывшееся диалоговое окно (рис.18) позволяет выбрать переменные (*variables)* для стандартизации и при необходимости придать вес (*weight*) некоторым переменным.

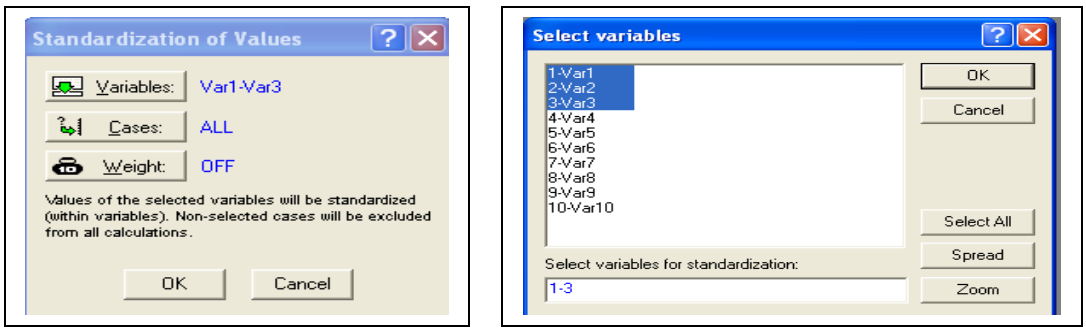

Рис.18. – Диалоговое окно для стандартизации выбранных переменных.

Вызов процедуры кластерного анализа осуществляется путем выбора пунктов меню *Statistics / Multivariate Exploratory Techniques / Cluster Analysis*, в результате чего появляется диалоговое окно, позволяющее выбрать метод кластеризации (рис.19).

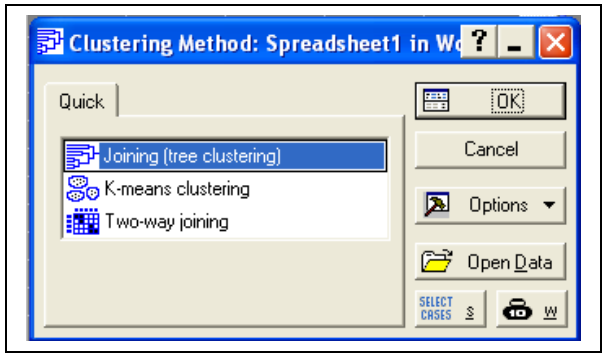

Рис.19. Выбор методов кластеризации.

При реализации каждого из этих методов необходимо выбрать переменные, требующие группировки, а также указать тип кластеризации, т.к. в программе *STATISTICA* предусмотрена кластеризация как наблюдений (*cases*) так и переменных (*variables*).

В процедуре *joining (tree clustering)*(иерархический метод) каждое наблюдение образовывает сначала свой отдельный кластер. На первом шаге два соседних кластера объединяются в один; этот процесс может продолжаться до тех пор, пока не останутся только два кластера.

В диалоговом окне Joining(tree clustering) необходимо выбрать переменные для анализа, указать по каким данным проводить кластеризацию (по строкам или столбцам), выбрать тип расстояния (Distance measure) и меру связи (Amalgamation (linkage) rule) (рис 20).

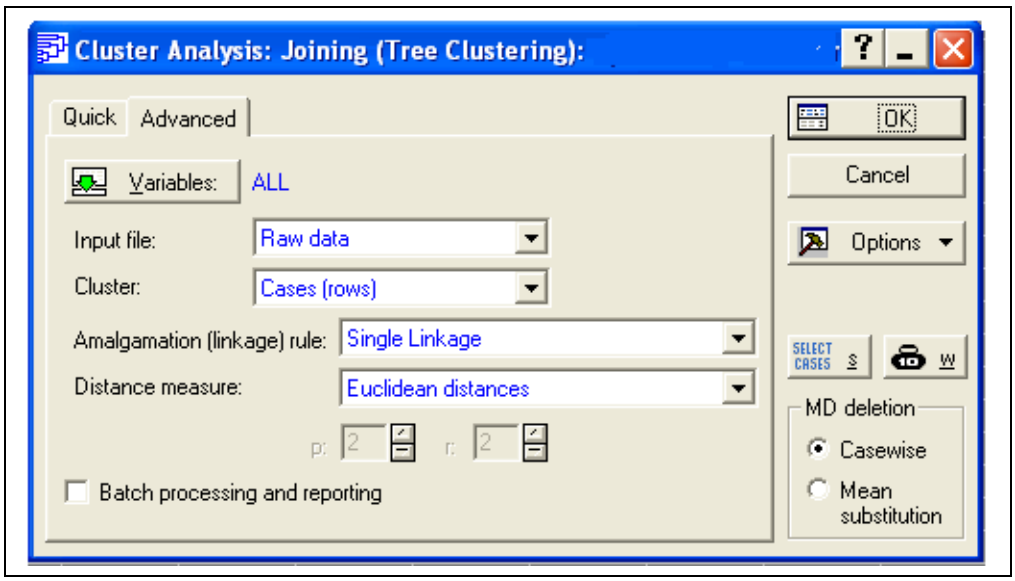

Рис.20. – Диалоговое окно иерархического кластерного анализа.

Расстояния:

- Euclidean distances – евклидово расстояние;
- Squared euclidean distances – квадрат евклидова расстояния;
- City-block (Manhattan) distances – расстояние городских кварталов (манхэттенское расстояние);
- Chebychev distance metric – расстояние Чебышева;
- Power distance – степенное расстояние;
- Percent disagreement – процент несогласия;
- 1-Pearson r – (1- коэффициент корреляции Пирсона).

Методы связи кластеров:

- Single Linkage – одиночная связь (метод ближайшего соседа);
- Complete Linkage – полная связь (метод наиболее удаленных соседей);
- Unweighted pair-group average – невзвешенное попарное среднее**;**
- Weighted pair-group average – взвешенное попарное среднее;
- Unweighted pair-group centroid – невзвешенный центроидный метод;
- Weighted pair-group centroid (median) – взвешенный центроидный метод (медиана);
- Ward's method – метод Варда.

После нажатия на кнопке *Ok* появится окно результатов кластерного анализа, которое позволяет просмотреть основные результаты иерархического кла-

стерного анализа представленые в форме матрицы расстояний и горизонтального или вертикального дерева (дендрограммы), которые вызываются путем нажатия соответствующих кнопок на диалоговом окне результатов кластерного анализа (рис.21).

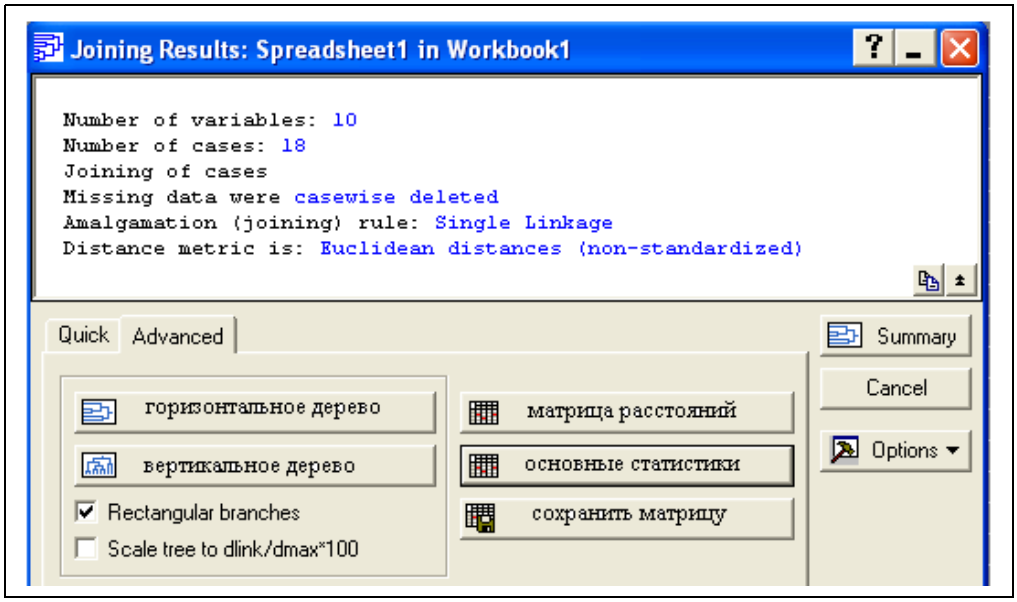

Рис. 21. – Диалоговое окно результатов иерархического кластерного анализа

Разобраться в значении кластеров помогают кластерные профили; они представляют собой средние значения переменных, которые включены в анализ, распределённые по кластерной принадлежности (основные статистики).

Недостатком иерархического кластерного анализа является сложность, а зачастую невозможность интерпретации результатов полученной дендрограммы. Поэтому иерархический анализ удобно использовать лишь при небольшом количестве наблюдений (не более 10). При большем количестве наблюдений иерархический анализ является разведочным методом для последовательного итерационного анализа. Он позволяет определить примерное количество кластеров разбиения.

Кластерный анализ при большом количестве наблюдений
(Кластерный анализ методом к-средних)

Иерархические методы объединения, хотя и точны, но трудоёмки: на каждом шаге необходимо выстраивать дистанционную матрицу для всех текущих кластеров. Расчётное время растёт пропорционально третьей степени количе-

ства наблюдений, что при наличии нескольких тысяч наблюдений может утомить и серьёзные вычислительные машины.

Поэтому при наличии большого количества наблюдений применяют другие методы. Недостаток этих методов заключается в том, что здесь необходимо заранее задавать количество кластеров, а не так как в иерархическом анализе, получить это в качестве результата. Эту проблему можно преодолеть проведением иерархического анализа со случайно отобранной выборкой наблюдений и, таким образом, определить оптимальное количество кластеров. Если количество кластеров указать предварительно, то появляется следующая проблема: определение начальных значений центров кластеров. Их также можно взять из предварительно проведённого иерархического анализа, в котором для каждого наблюдения рассчитывают средние значения переменных, использовавшихся при анализе, а потом в определённой форме сохраняют их в некотором файле. Этот файл может быть, затем прочитан методом, который применяется для обработки больших количеств наблюдений. Если нет желания проходить весь этот длинный путь, то можно воспользоваться методом, предлагаемым для данного наблюдения программой *STATISTICA*.

Если количество кластеров k, которое необходимо получить в результате объединения, задано заранее, то первые k наблюдений, содержащихся в файле, используются как первые кластеры. На последующих шагах кластерный центр заменяется наблюдением, если наименьшее расстояние от него до кластерного центра больше расстояния между двумя ближайшими кластерами. По этому правилу заменяется тот кластерный центр, который находится ближе всего к данному наблюдению. Таким образом, получается новый набор исходных кластерных центров. Для завершения шага процедуры рассчитывается новое положение центров кластеров, а наблюдения перераспределяются между кластерами с изменёнными центрами. Этот итерационный процесс продолжается до тех пор, пока кластерные центры не перестанут изменять свое положение или пока не будет достигнуто максимальное число итераций.

Выберите в меню *Statistica* (Статистика) *Multivariate Exploratory Techniques* (Многомерные методы исследования) *Cluster Analysis* (Кластерный анализ) *k-means clustering* (Кластерный анализ методом к - средних). Откроется диалоговое окно *K-Means Cluster Analysis* (Кластерный анализ методом к - средних) (рис.22).

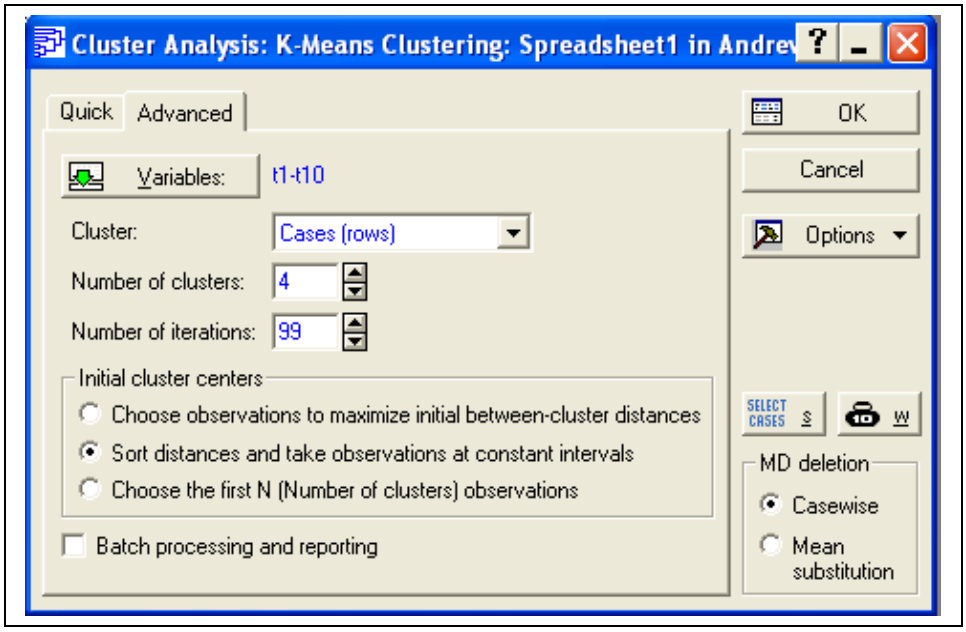

Рис.22.– Окно Кластерного анализа методом к-средних.

Выберите переменные. В поле Cluster укажите, что кластеризация проводится по наблюдениям (cases). Теперь необходимо указать количество кластеров (Number of cluster) и количество итераций (Number of iteration). Укажите число итераций равное 99; установленное по умолчанию количество итераций равное 10, иногда оказывается недостаточным.

Для определения количества кластеров *k* можно провести несколько опытных, пробных расчётов с различным количеством кластеров и после этого определиться с подходящим вариантом решения. Но наиболее подходящим вариантом определения k является предварительное проведение иерархического кластерного анализа для произвольно выбранных наблюдений и получившееся количество кластеров принять за оптимальное. Щёлкните на *OK*, чтобы начать расчёт (рис 23).

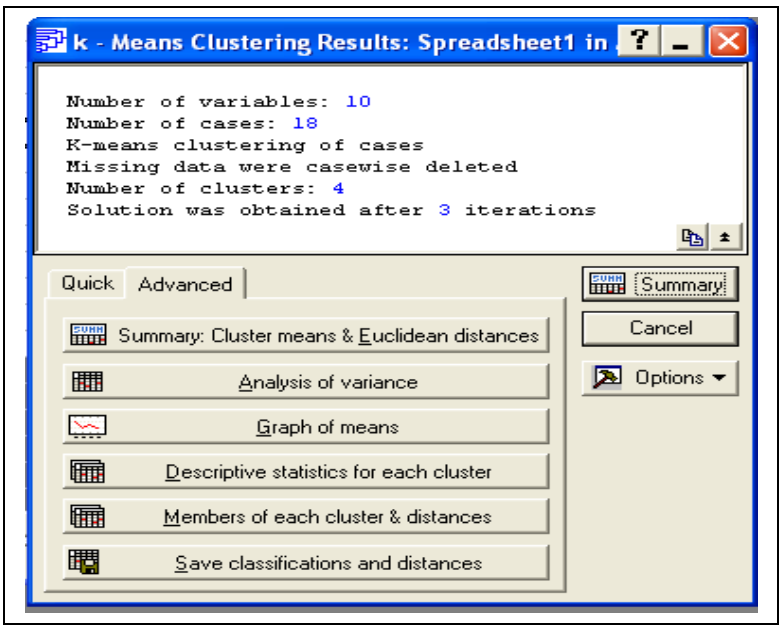

Рис.23.– Окно просмотра результатов кластерного анализа методом *к*-средних.

Сначала приводятся первичные кластерные центры (*cluster means*) и евклидовы расстояния между кластерами (*Euclidean distances)*.

Далее выводятся показатели, позволяющие анализировать каждый кластер: анализ разброса значений (*Analysis of variance*), графическое представление кластерных центров (*Graph of mean*), основные статистики (среднее, стандартное отклонение и дисперсия) для каждого кластера (*Descriptive statistics for each cluster*).

В заключение можно просмотреть и сохранить наблюдения, относящиеся к каждому из кластеров и расстояния между наблюдениями. Полученная классификация и расстояния сохраняются в отдельном файле, в который для удобства пользователя и возможно для дальнейшей работы можно добавить необходимые переменные из исходного рабочего документа, путем выделения соответствующих строк в открывающемся диалоговом окне.

Кластерный анализ методом *к*-средних дополняет и уточняет картину, полученную с помощью иерархического кластерного анализа. Однако конфигурация кластеров не поддается представлению в графическом виде.

Задание: 18 претендентов прошли 10 различных тестов в кадровом отделе предприятия.

Таблица9

Номер теста	*обозначение*	*Предмет теста*
1	*t1*	*Память на числа*
2	*t2*	*Математические задачи*
3	*t3*	*Находчивость при прямом диалоге*
4	*t4*	*Тест на составление алгоритмов*
5	*t5*	*Уверенность во время выступления*
6	*t6*	*Командный дух*
7	*t7*	*Находчивость*
8	*t8*	*Сотрудничество*
9	*t9*	*Признание в коллективе*
10	*t10*	*Сила убеждения*

Максимальная оценка, которую можно было получить на каждом из тестов, составляет 10 балов. Результаты теста для 18 претендентов находятся в таблице 10 в переменных t1-t10. Каждое наблюдение является характеристикой тестируемых кандидатов.

Таблица 10

инициалы	*t1*	*t2*	*t3*	*t4*	*t5*	*t6*	*t7*	*t8*	*t9*	*t10*
А.П.	*10*	*9*	*3*	*10*	*4*	*5*	*1*	*7*	*6*	*5*
М.К.	*6*	*5*	*8*	*6*	*9*	*6*	*2*	*6*	*5*	*9*
С.Т.	*7*	*4*	*5*	*6*	*6*	*7*	*5*	*5*	*4*	*6*
Т.Т.	*8*	*8*	*5*	*6*	*6*	*4*	*6*	*7*	*7*	*5*
А.А.	*5*	*4*	*8*	*3*	*9*	*2*	*9*	*9*	*8*	*9*
М.Р.	*6*	*4*	*4*	*5*	*6*	*8*	*7*	*4*	*2*	*6*
А.С.	*9*	*10*	*7*	*8*	*6*	*9*	*3*	*6*	*5*	*7*
Д.Н.	*4*	*3*	*7*	*3*	*6*	*5*	*6*	*10*	*9*	*6*
Ю.Т.	*8*	*5*	*4*	*9*	*5*	*8*	*4*	*6*	*7*	*7*
К.А.	*8*	*6*	*4*	*6*	*5*	*7*	*8*	*6*	*8*	*5*
Ч.Н.	*7*	*7*	*6*	*7*	*7*	*9*	*9*	*3*	*4*	*7*
М.Т.	*5*	*5*	*9*	*5*	*10*	*5*	*10*	*6*	*5*	*9*
Л.Г.	*4*	*5*	*10*	*4*	*10*	*4*	*5*	*8*	*9*	*10*
С.П.	*10*	*10*	*6*	*9*	*5*	*8*	*4*	*8*	*10*	*5*
Р.Д.	*6*	*5*	*4*	*6*	*4*	*4*	*6*	*5*	*6*	*4*
Ф.А.	*6*	*7*	*3*	*7*	*2*	*3*	*8*	*4*	*3*	*3*
К.С.	*7*	*8*	*5*	*7*	*3*	*6*	*7*	*4*	*4*	*4*
Г.Т.	*8*	*9*	*6*	*8*	*5*	*10*	*7*	*10*	*8*	*5*

С использованием результатов теста соответствия провести кластерный анализ и обнаружить группы кандидатов, близких по своим качествам.

Лабораторная работа № 8

Факторный анализ

Факторный анализ – статистический метод, который используется при обработке больших массивов экспериментальных данных. Задачами факторного анализа являются: сокращение числа переменных (редукция данных) и определение структуры взаимосвязей между переменными, т.е. классификация переменных, поэтому факторный анализ используется как метод сокращения данных или как метод структурной классификации.

Материалом для факторного анализа служат корреляционные связи, а точнее – коэффициенты корреляции Пирсона, которые вычисляются между переменными, включенными в обследование. Иными словами, факторному анализу подвергают корреляционные матрицы или матрицы интеркорреляций.

Главное понятие факторного анализа – фактор. Это искусственный статистический показатель, возникающий в результате специальных преобразований таблицы коэффициентов корреляции между изучаемыми психологическими признаками, или матрицы интеркорреляций.

Основные результаты факторного анализа выражаются в наборах факторных нагрузок и факторных весов.

Факторные нагрузки – это значения коэффициентов корреляции каждого из исходных признаков с каждым из выявленных факторов. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значение факторной нагрузки. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак – на обратную) связь данного признака с фактором. Таблица факторных нагрузок содержит m строк (по числу признаков) и k столбцов (по числу факторов).

Для построения матрицы факторных нагрузок необходимо найти собственные числа и собственные векторы корреляционной матрицы. Нормированные координаты собственных векторов являются элементами матрицы факторных нагрузок.

Факторными весами называют количественные значения выделенных факторов для каждого из *n* имеющихся объектов. Объекту с большим значением факторного веса присуща большая степень проявления свойств, определяемых данным фактором.

Поэтому положительные факторные веса соответствуют тем объектам, которые обладают степенью проявления свойств больше средней, а отрицательные факторные веса соответствуют тем объектам, для которых степень проявления свойств меньше средней. Таблица факторных весов содержит *n* строк (по числу объектов) и *k* столбцов (по числу факторов).

Таким образом, данные о факторных нагрузках позволяют сформулировать выводы о наборе исходных признаков, отражающих тот или иной фактор, и об относительном весе отдельного признака в структуре каждого фактора. В свою очередь, данные о факторных весах определяют ранжировку объектов по каждому фактору.

Набор методов факторного анализа в настоящее время достаточно велик, насчитывает десятки различных подходов и приемов обработки данных. В основе каждого метода факторного анализа лежит математическая модель, описывающая соотношения между исходными признаками и обобщенными факторами. Рассмотрим наиболее распространенные методы.

Центроидный метод. Этот метод основан на предположении о том, что каждый из исходных признаков X_i может быть представлен как функция небольшого числа общих факторов $F_1,F_2,\ldots,F_k$ и характерного фактора U_i.:

$$X_i = \sum a_{ik} F_k + U_i$$

Факторы F построены так, чтобы наилучшим способом (с минимальной погрешностью) представить X. В этой модели «скрытые» переменные F_k называются общими факторами, а переменные U_i специфическими («характерными», «уникальными») факторами.

При этом считается, что каждый общий фактор имеет существенное значение для анализа всех исходных признаков, т.е. фактор F_i -общий для всех X_i. В то же время изменения в специфическом факторе U_i воздействуют на значе-

ния только соответствующего признака X_i. Таким образом, специфический фактор U_i отражает ту специфику признака X_i, которая не может быть выражена через общие факторы.

Метод главных компонент. В основе модели для выражения исходных признаков через факторы здесь лежит предположение о том, что число общих факторов равно числу исходных признаков ($k=m$), а специфические факторы вообще отсутствуют: $X_i = \sum a_{ik} F_k$

Уравнения определяют систему преобразования одних параметров в другие. Поскольку число факторов равно числу исходных параметров, задача искомого преобразования решается однозначно, т.е. факторные нагрузки определяются в этом методе однозначно.

Каждая из переменных F_j называется здесь *i-й* главной компонентой. Метод главных компонент состоит в построении факторов – главных компонент, каждый из которых представляет линейную комбинацию исходных признаков.

Первая главная компонента F_1 определяет такое направление в пространстве исходных признаков, по которому совокупность объектов (точек) имеет наибольший разброс (дисперсию). Вторая главная компонента F_2 строится с таким расчетом, чтобы ее направление было ортогонально направлению F_1 и она объясняла как можно большую часть остаточной дисперсии, и т.д. вплоть до *m*-й главной компоненты F_m. Так как выделение главных компонент происходит в убывающем порядке с точки зрения доли объясняемой ими дисперсии, то признаки, входящие в первую главную компоненту оказывают максимальное влияние на дифференциацию изучаемых объектов.

Для определения количества факторов имеется несколько критериев:

1. Критерий Кайзера предлагает отобрать только факторы, с собственными значениями, большими *1*.

2. Критерий каменистой осыпи является графическим методом, впервые предложенным Кэттелем (Cattell). Необходимо изобразить собственные значения, расположенные в убывающем порядке в виде графика (по оси абсцисс порядковый номер числа, по оси ординат его значение) (рис. 24).

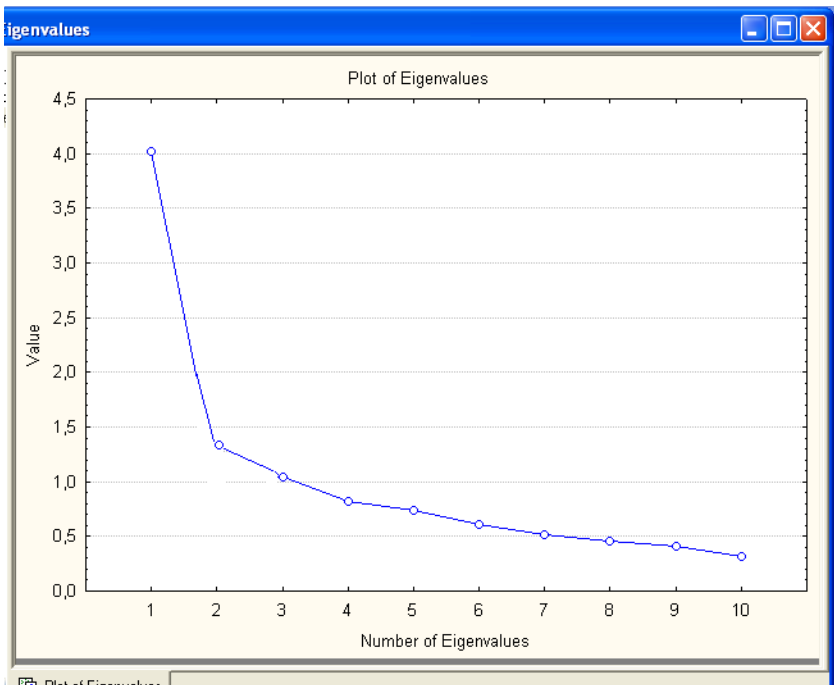

Рис.24. Критерий каменистой осыпи.

Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь» – «осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. В соответствии с этим критерием можно оставить в этом примере 2 или 3 фактора.

Первый критерий (*критерий Кайзера*) иногда сохраняет слишком много факторов, в то время как второй критерий (*критерий каменистой осыпи*) иногда сохраняет слишком мало факторов; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. Обычно исследуется несколько решений с большим или меньшим числом факторов, и затем выбирается одно наиболее интерпретируемое.

В программе *Statistica* факторный анализ проводится с помощью модуля *Multivariate Exploratory Techniques*. Переход к процедуре факторного анализа осуществляется посредством пункта *Factor Analysis*. В открывшемся окне необходимо указать тип формата данных (*Raw Data* или *Correlation Matrix*), переменные для анализа и нажать *OK*. Далее выбирают конкретный метод фактори-

зации корреляционной матрицы – *Extraction method* (например, *Principal components* – метод главных компонент), максимальное число факторов (*maximum of factors*) обычно на первом этапе равное числу переменных, минимальное собственное число (*minimum eigenvalue)* равное 0 и *OK* (рис.25).

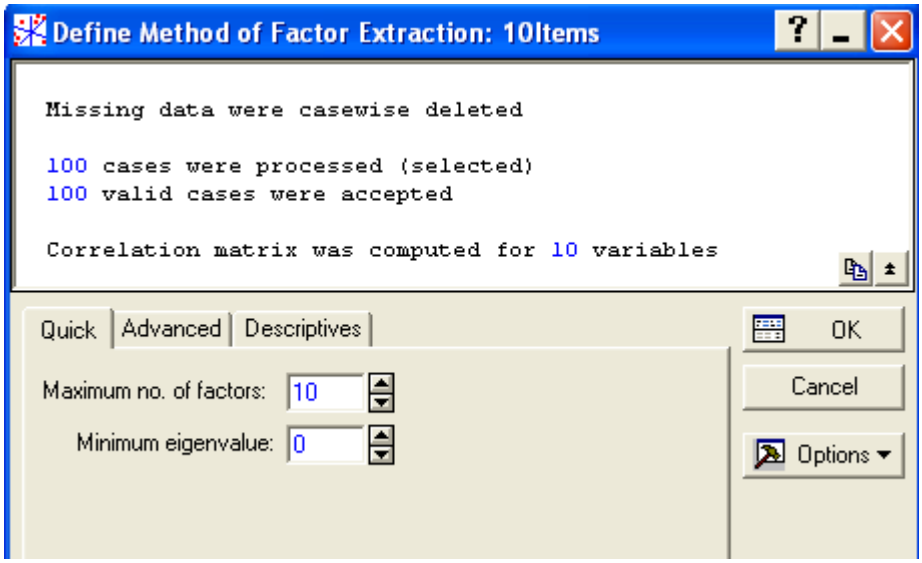

Рис.25. Диалоговое окно факторного анализа.

После этого появится окно результатов факторного анализа, которое позволяет просмотреть собственные значения корреляционной матрицы *(Eigenvalues),* графическое изображение собственных чисел *(Scree plot)* матрицу факторных нагрузок (*Summary: Factor loading)* и факторных весов (*Factor scores).*

Задание: Используя данные, полученные у случайной выборки учащихся первого класса (x1 – вес тела, x2 – рост, x3 – количество слов, читаемых в минуту, x4 – оценка по чтению за год, x5 – длина руки, x6 – количество книг прочитанных за год, x7 – количество выученных стихотворений) провести факторный анализ данных. Сделать выводы.

Таблица 11

	X1	X2	X3	X4	X5	X6	X7
1	20	120,00	22	3	10	12	12
2	31	122,00	26	3	13	13	13
3	22	123,00	45	5	12	23	19
4	37	126,00	38	5	14	19	19
5	42	127,00	44	5	14	24	19
6	45	123,00	32	4	15	17	17
7	38	123,00	31	4	13	16	16
8	44	124,00	33	4	13	18	18
9	50	126,00	39	5	14	19	19

Продолжение таблицы 11

10	20	127,00	35	4	10	18	18
11	26	125,00	27	3	11	11	11
12	30	129,00	28	3	12	13	13
13	50	124,00	26	3	13	13	13
14	26	125,00	27	3	12	12	12
15	32	124,00	36	4	13	18	18
16	28	125,00	42	5	11	19	19
17	20	122,00	35	4	10	18	18
18	25	127,00	34	4	12	16	16
19	20	122,00	26	3	11	13	13
20	30	123,00	15	6	15	15	15

Лабораторная работа № 9

Многомерное шкалирование

Главная задача многомерного шкалирования – найти минимальное число субъективных признаков, определяющих различение стимулов человеком, и вычислить значение признаков, которыми характеризуются данные стимулы. Решение задачи многомерного шкалирования основано на использовании понятия психологического пространства, точки которого представляют исходные стимулы. Аналогично геометрическим представлениям вводится система координат, число которых определяется числом простых субъективных признаков. Это число задает размерность психологического пространства. Оси координат представляют собой шкалы соответствующих субъективных признаков, и положение точек-стимулов в пространстве задано шкальными значениями признаков. Число субъективных шкал и шкальные значения стимулов характеризуют пространственную модель многомерного шкалирования. Модель строится, когда определяется субъективное расстояние на основе сходства и различия между стимулами в психологическом пространстве.

Формально общая задача многомерного шкалирования выражается следующим образом. По заданной симметричной матрице различий между стимулами $D = \begin{pmatrix} D_{11} & \dots & D_{1n} \\ \dots & \dots & \dots \\ D_{n1} & \dots & D_{nn} \end{pmatrix}$ нужно построить метрическую и пространственную мо-

дели стимулов, т.е. определить размерность пространства и координаты точек-стимулов в этом пространстве $X = \begin{pmatrix} X_{11} & \dots & X_{1n} \\ \dots & \dots & \dots \\ X_{n1} & \dots & X_{nn} \end{pmatrix}$ таким образом, чтобы матрица расстояний, вычисленных между точками на основании метрической модели расстояния $d = \begin{pmatrix} d_{11} & \dots & d_{1n} \\ \dots & \dots & \dots \\ d_{n1} & \dots & d_{nn} \end{pmatrix}$ была бы в смысле некоторого критерия возможно более близка к исходной матрице различий.

Реализация методов многомерного шкалирования в программе *Statistica* проводится с помощью модуля *Multivariate Exploratory Techniques*. Переход к процедуре многомерного шкалирования осуществляется посредством пункта *Multidimensional Scaling*. Исходные данные должны представлять собой матрицу расстояний. В открывшемся окне необходимо указать переменные для анализа, количество осей (Number of dimensions) – чаще всего выбирают 2 оси и нажать *OK* (рис.26).

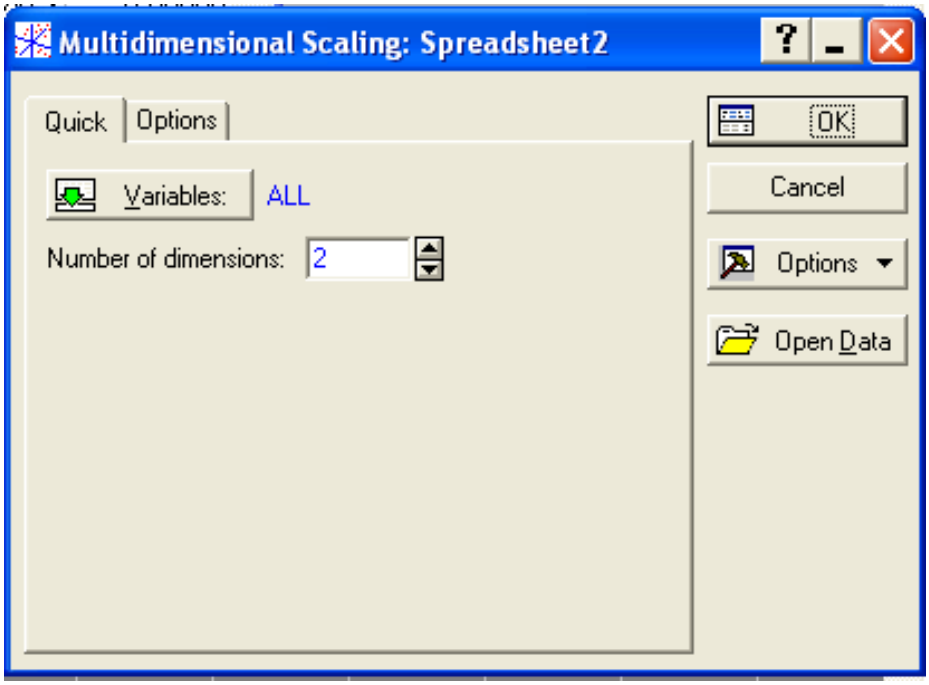

Рис.26. Диалоговое окно многомерного шкалирования.

Результат многомерного анализа представлен в виде расположения точек наблюдения в системе координат, которое вызывается кнопкой *Graph final configuration 2D* (рис.27).

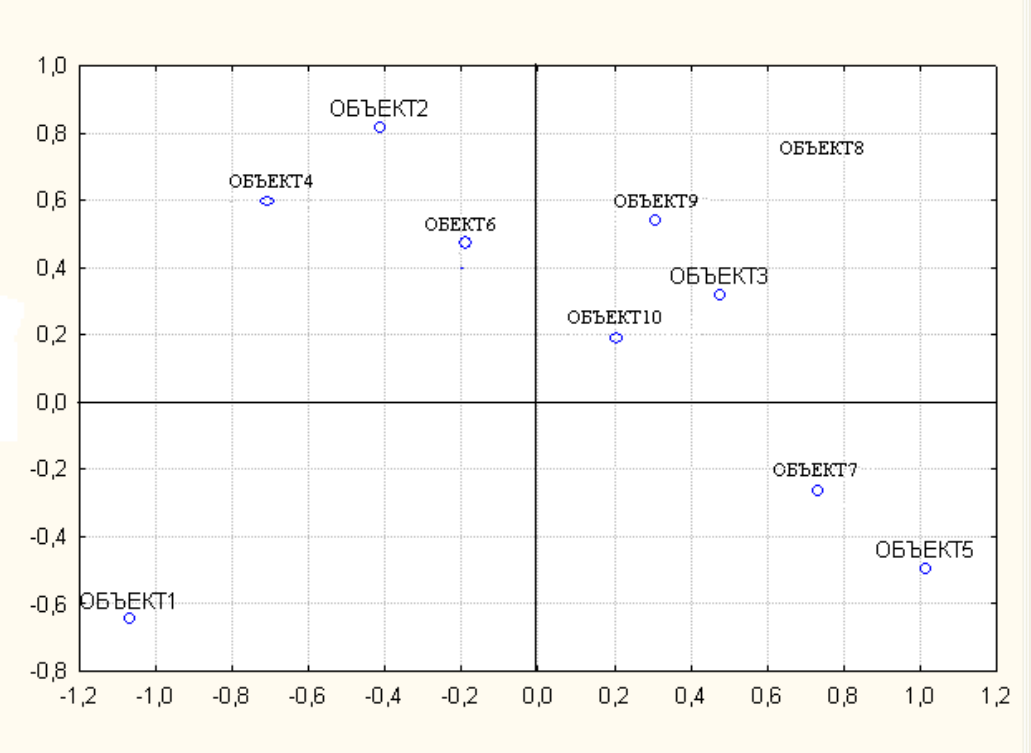

Рис.27. Графическая интерпретация результатов многомерного шкалирования.

Модели различий в многомерном шкалировании достаточно широко применяются в психологии, главным образом для изучения индивидуальной специфики оценок сложных стимулов различными людьми.

Задание: Создайте файл, в котором введите бальную оценку 15-20 респондентов нескольким произвольным телепередачам (порядка 10).

Таблица 12

Новости	*Аншлаг*	*…*	*Время*

Проведите для данных кластерный анализ и сохраните матрицу расстояний при помощи кнопки Matrix. А теперь для ***МАТРИЦЫ РАССТОЯНИЙ*** *осуществите многомерное шкалирование. И распределите передачи на несколько групп. По возможности интерпретируйте полученные оси.*

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ермолаев О. Ю. Математическая статистика для психологов. – М.: Московский психолого-социальный институт: Флинта, 2002.
2. Суходольский Г. В. Основы математической статистики для психологов. – Л.: ЛГУ, 1972.
3. Двоерядкина Н.Н., Макарчук Т.А., Киселева А.Н. Эконометрика. Лабораторный практикум. – Благовещенск: АмГУ, 2004.
4. Гусев А.Н., Измайлов Ч.А., Михалевская М.Б. Измерение в психологии: общий психологический практикум. – М.: Смысл, 1997.

СОДЕРЖАНИЕ

Наталья Николаевна Двоерядкина,

кандидат педагогических наук,

доцент кафедры общей математики и информатики АмГУ;

Алена Николаевна Киселева,

ст. преподаватель кафедры общей математики и информатики АмГУ;

Татьяна Александровна Юрьева,

ст. преподаватель кафедры общей математики и информатики АмГУ.

Математические методы в психологии: Лабораторный практикум.