ЛАБОРАТОРНАЯ РАБОТА № 12
ДИСКРИМИНАНТНЫЙ АНАЛИЗ

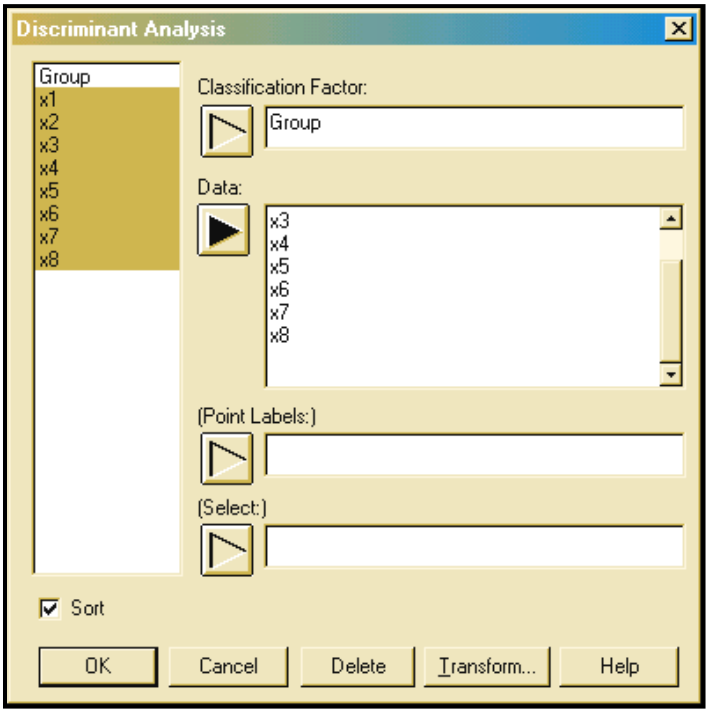
Предположим, что мы имеем совокупность объектов, разбитую на несколько групп (т. е. для каждого объекта мы можем сказать, к какой группе он относится). Пусть для каждого объекта имеются измерения нескольких количественных характеристик. Мы хотим найти способ, как на основании этих характеристик можно узнать группу, к которой принадлежит объект. Это позволит для новых объектов из той же совокупности предсказывать группы, к которой они относятся. Например, исследуемыми объектами могут быть пациенты — здоровые или больные той или иной болезнью, а характеристиками — результаты медицинских анализов. Если научиться по этим характеристикам узнавать, здоров ли пациент либо болен, это позволит значительно повысить эффективность медицинских обследований. Для решения такой задачи применяются методы дискриминантного анализа, они позволяют строить функции характеристик, значения которых и объясняют разбиение объектов на группы.

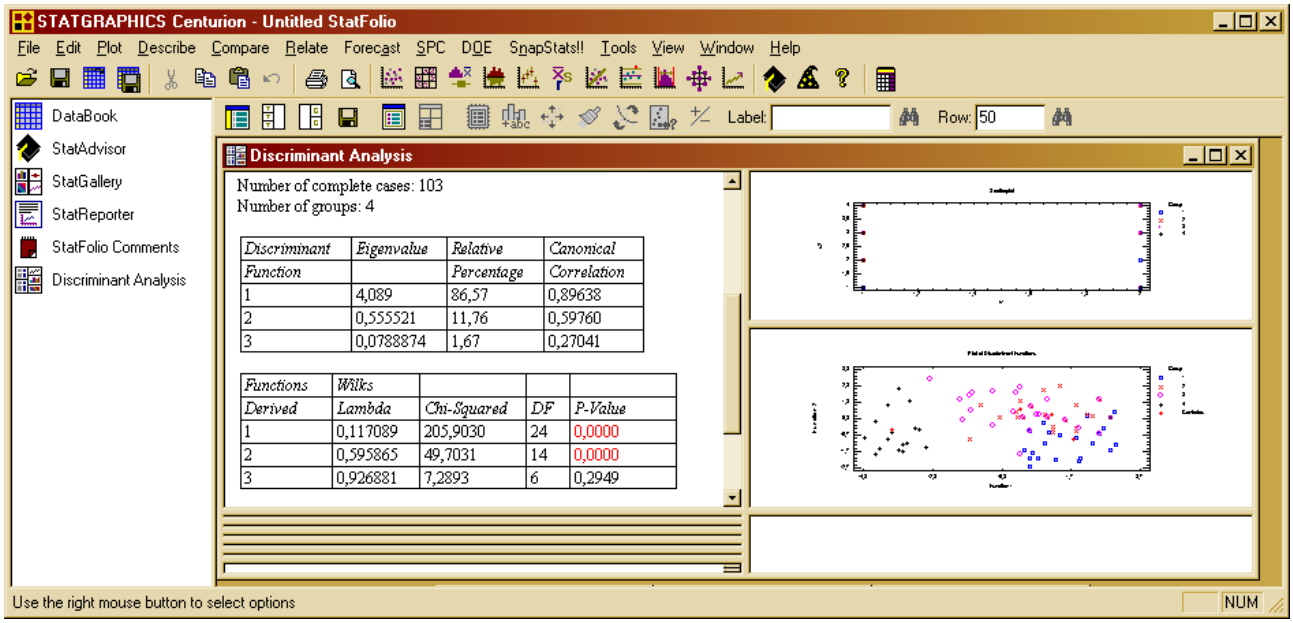
Пусть у нас есть файл, в котором содержатся данные — результаты обследования 103 человек с установленным диагнозом: группа 1 — гангренозный аппендицит (28 наблюдений); группа 2 — флегмонозный аппендицит 25 наблюдений); группа 3 — катаральный аппендицит (26 наблюдений) и группа 4 — неподтвержденный диагноз (24 наблюдения). Исходными данными служили 8 симптомов, заданных в таблице:

№ пп	Симптомы острого аппендицита	Выраженность	Код
*X*1	Боли в правой подвздошной области	незначительные	1
		выраженные	2
*X*2	Продолжительность болей	свыше 2-х суток	1
		25–48 часов	2
		13–24 часа	3
		до 12 часов	4
*X*3	Частота пульса	до 80 уд/мин	1
		81–100 уд/мин	2

		свыше 100 уд/мин	3
*X*4	Лейкоциты крови	до 8 тыс	1
		8-14 тыс	2
		свыше 14 тыс	3
*X*5	Изменения языка	не обложен	0
		обложен	1
*X*6	Симптом Щеткина-Блюмберга	отсутствует	0
		выражен	2
*X*7	Симптом Ровзинга	отсутствует	0
		выражен	2
*X*8	Защитное мышечное напряжение	отсутствует	0
		выражен	2

Для того чтобы сэкономить ваше время, эти данные набраны и находятся в файле *appendix.sf*, откройте этот файл. В строке меню выберите ***Relate***, в раскрывшемся меню выберите ***Classification Methods***, затем ***Discriminant Analysis***. Раскроется диалоговое окно, в этом диалоговом окне в поле ***Classification Factor*** введите переменную *Group*, в поле *Data* введите переменные *X*11, *X*2, *X*3, *X*4, *X*5, *X*6, *X*7, *X*8. Нажмите кнопку ОК. Раскроется сводка дискриминантного анализа:





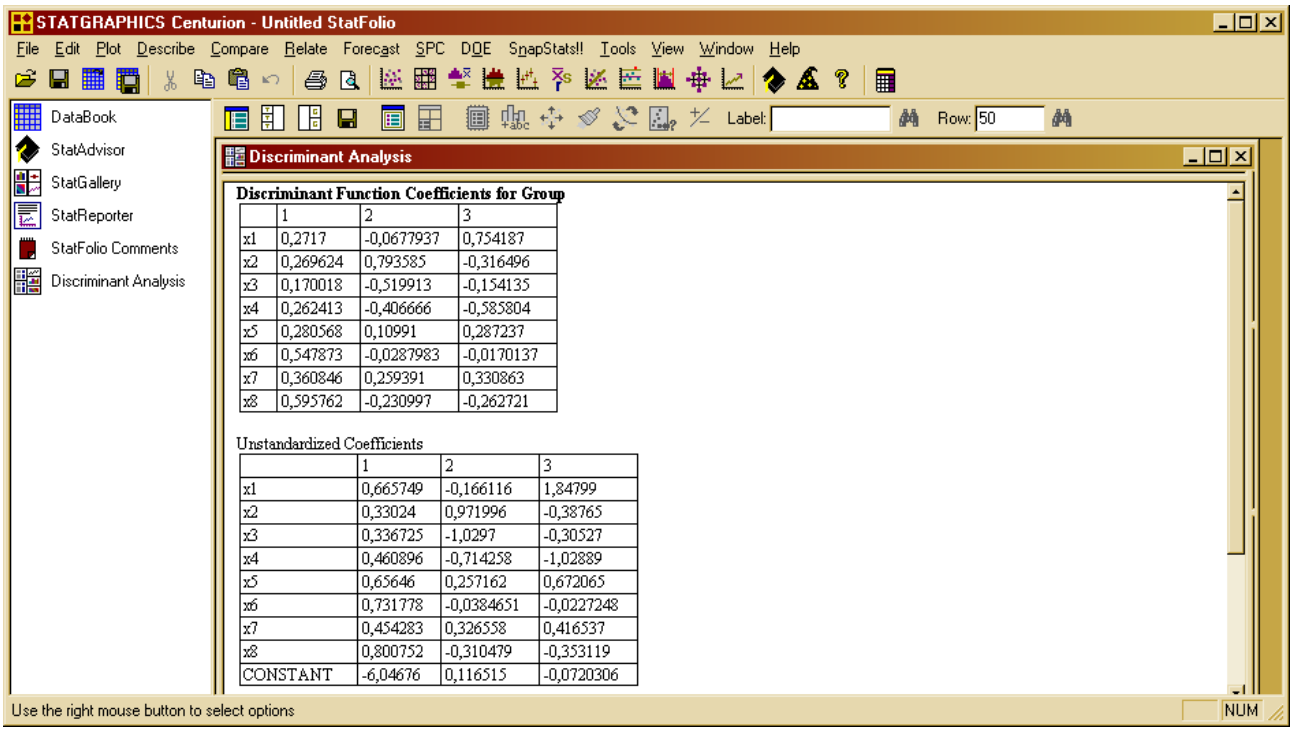
Discriminant Function	*Eigenvalue*	*Relative Percentage*	*Canonical Correlation*
1	4,089	86,57	0,89638
2	0,555521	11,76	0,59760
3	0,0788874	1,67	0,27041

Functions Derived	*Wilks Lambda*	*Chi-Squared*	*DF*	*P-Value*
1	0,117089	205,9030	24	0,0000
2	0,595865	49,7031	14	0,0000
3	0,926881	7,2893	6	0,2949

Эта таблица содержит характеристики трех выделенных дискриминантных функций (*Discriminant Functions*): собственные значения (*Eigenvalue*), вклад каждой функции в объяснение дисперсии симптомов (*Relative Percentage*) в %, канонические корреляции с классифицирующим фактором (*Canonical Correlation*) и оценки уровня значимости дискриминантных функций по критериям Лямбда и Хи-квадрат.

На первую дискриминантную функцию *F*1 приходится 86,57 % дисперсии симптомов, на вторую *F*2 — еще 11,76 %. На третью остается всего 1,67 % дисперсии, поэтому достаточно применять первые две — *F*1 и *F*2, на которые в сумме приходится 98,33 % дисперсии симптомов. В *StatAdvisor* написано, что эти две дискриминантные функции с *p-value*, меньшим 0,05, являются статистически значимыми с 95 %.

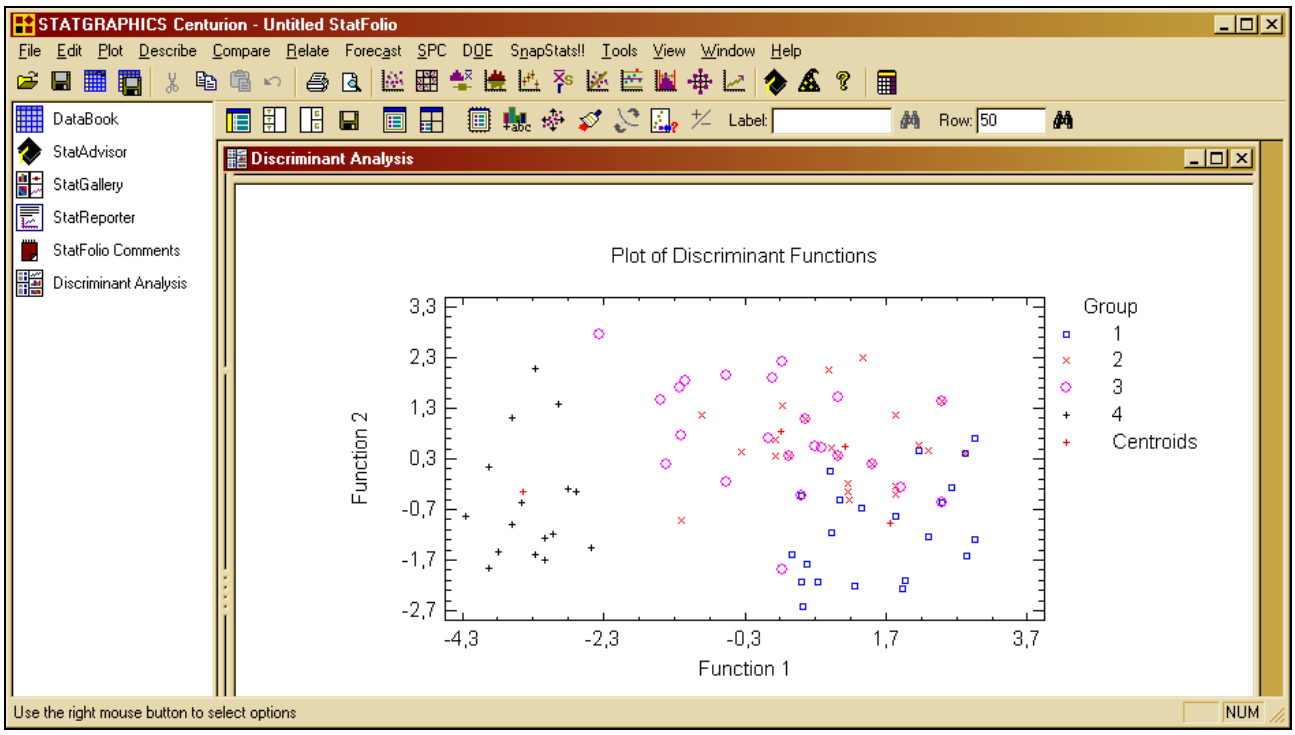
Нажмите кнопку ***Tables***, в раскрывшемся окне выберите ***Discriminant Functions*** (дискриминантные функции). Нажмите ОК. Раскроется окно с таблицами:



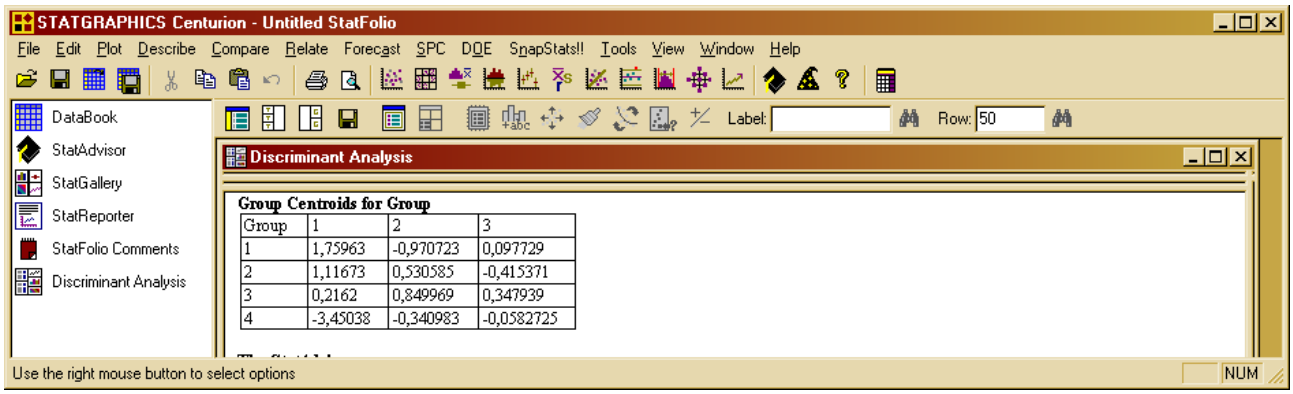
Первая таблица содержит коэффициенты трех дискриминантных функций в стандартизированном виде. В *StatAdvisor* вы можете увидеть эти функции, посмотрите. Для расчета по этим функциям в них следует подставлять стандартизированные значения исходных признаков. Вторая таблица включает константы и коэффициенты дискриминантных функций, зная которые можно выписать в явном виде дискриминантные функции, например, первая из них имеет вид

$$F1=-6.04676+0{,}665749*x1+0{,}33024*x2+0{,}336725*x3+0{,}460896*x4+$$
$$+0{,}65646*x5+0{,}731778*x6+\ 0{,}454283*x7+0{,}800752*x8.$$

Посмотрим теперь на графическое отображение результатов. Нажмите кнопку ***Graphs***, в окне выберите ***Discriminant Functions***, нажмите кнопку ОК. Перед вами откроется график:

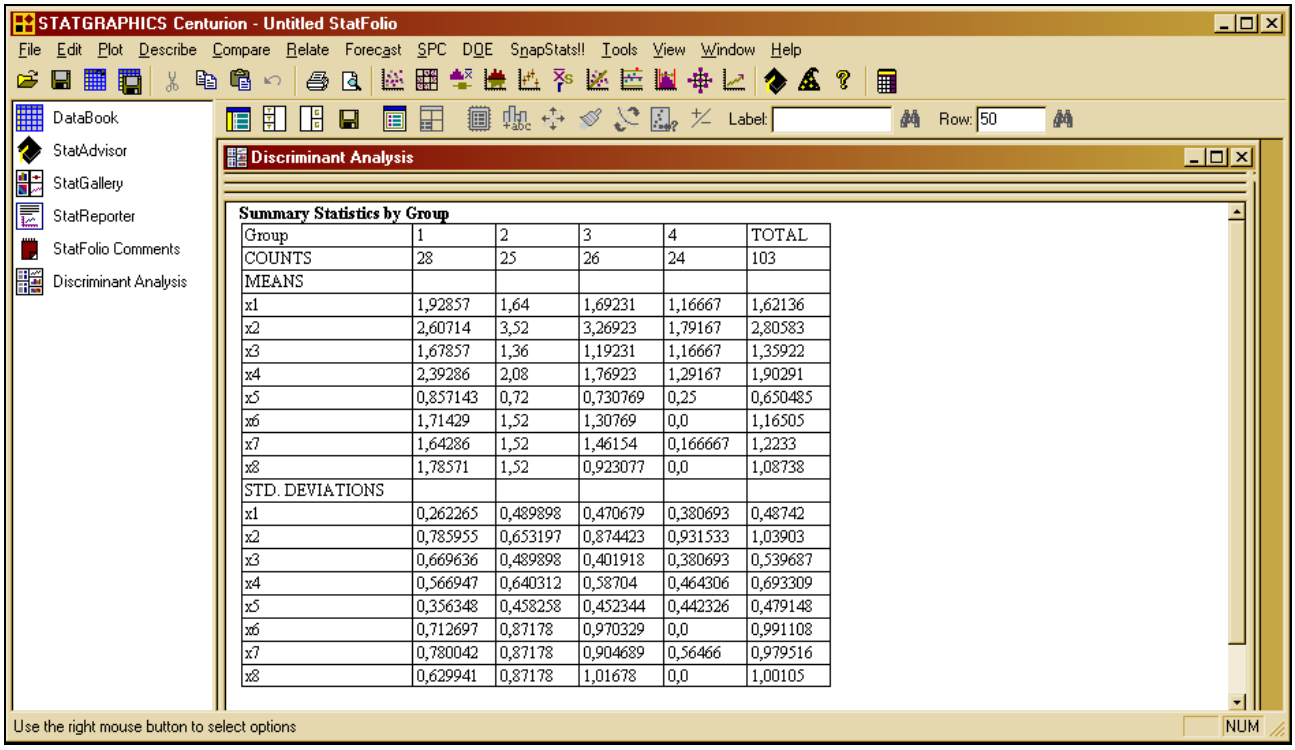


На диаграмме хорошо видно, что объекты 4-го класса (неподтвержденный диагноз обозначается плюсом) образуют самостоятельную, четко выраженную группировку, не пересекающуюся с другими классами. Остальные же классы имеют значительные пересечения в пространстве дискриминантных функций. Обратите внимание на центроиды групп (*Centroids*), на диаграмме они обозначены красными плюсами. Выведем на экран результаты расчета координат центроидов групп. Нажмите кнопку ***Tables***, в окне выберите ***Group Centroids***, нажмите кнопку ОК.



Диагностическое правило заключается в вычислении расстояния от диагностируемого объекта до центроидов классов в пространстве канонических дискриминантных функций и отнесению нового объекта к тому классу, для которого это расстояние минимально.

Снова нажмите кнопку ***Tables***, в окне выберите ***Group Statistics***.

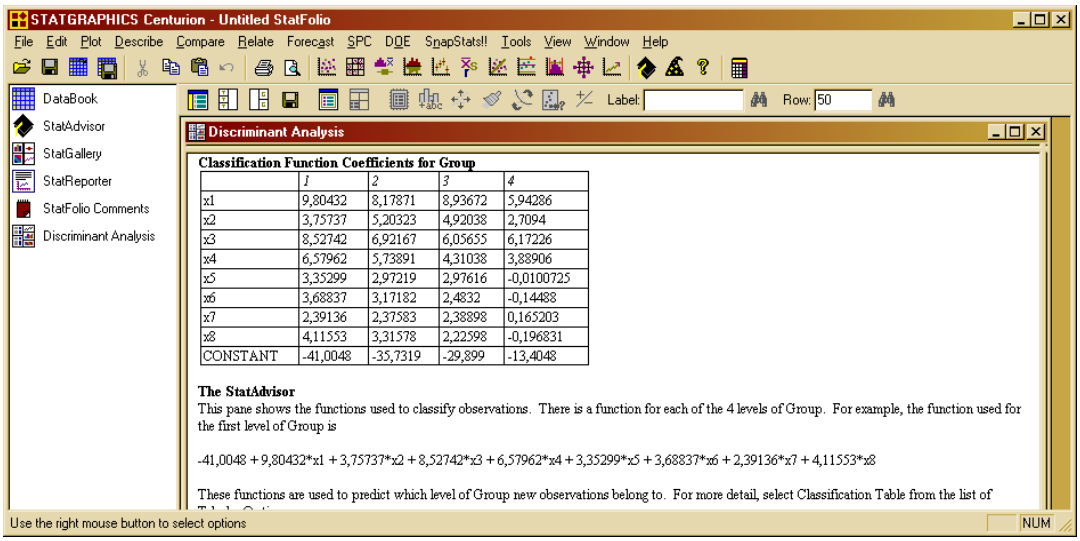


Summary Statistics by Group

Group	1	2	3	4	TOTAL
COUNTS	28	25	26	24	103
MEANS					
x1	1,92857	1,64	1,69231	1,16667	1,62136
x2	2,60714	3,52	3,26923	1,79167	2,80583
x3	1,67857	1,36	1,19231	1,16667	1,35922
x4	2,39286	2,08	1,76923	1,29167	1,90291
x5	0,857143	0,72	0,730769	0,25	0,650485
x6	1,71429	1,52	1,30769	0,0	1,16505
x7	1,64286	1,52	1,46154	0,166667	1,2233
x8	1,78571	1,52	0,923077	0,0	1,08738
STD. DEVIATIONS					
x1	0,262265	0,489898	0,470679	0,380693	0,48742
x2	0,785955	0,653197	0,874423	0,931533	1,03903
x3	0,669636	0,489898	0,401918	0,380693	0,539687
x4	0,566947	0,640312	0,58704	0,464306	0,693309
x5	0,356348	0,458258	0,452344	0,442326	0,479148
x6	0,712697	0,87178	0,970329	0,0	0,991108
x7	0,780042	0,87178	0,904689	0,56466	0,979516
x8	0,629941	0,87178	1,01678	0,0	1,00105

Здесь можно посмотреть средние значения симптомов в каждой группе больных и какова их вариация относительно средних. Видно, что по отдельно взятым разрозненным симптомам невозможно добиться постановки удовлетворительного диагноза.

Более точные результаты диагностики дает применение линейных **дискриминантных функций Фишера**. В *Statgraphics* они называются ***Classification Functions*** (классифицирующие функции). Нажмите на кнопку ***Tables***, в окне выберите ***Classification Functions,*** нажмите кнопку ОК. Раскроется следующее окно:



Classification Function Coefficients for Group

	1	2	3	4
x1	9,80432	8,17871	8,93672	5,94286
x2	3,75737	5,20323	4,92038	2,7094
x3	8,52742	6,92167	6,05655	6,17226
x4	6,57962	5,73891	4,31038	3,88906
x5	3,35299	2,97219	2,97616	-0,0100725
x6	3,68837	3,17182	2,4832	-0,14488
x7	2,39136	2,37583	2,38898	0,165203
x8	4,11553	3,31578	2,22598	-0,196831
CONSTANT	-41,0048	-35,7319	-29,899	-13,4048

The StatAdvisor

This pane shows the functions used to classify observations. There is a function for each of the 4 levels of Group. For example, the function used for the first level of Group is

-41,0048 + 9,80432*x1 + 3,75737*x2 + 8,52742*x3 + 6,57962*x4 + 3,35299*x5 + 3,68837*x6 + 2,39136*x7 + 4,11553*x8

These functions are used to predict which level of Group new observations belong to. For more detail, select Classification Table from the list of

В *StatAdvisor* можно прочитать, что в этой таблице представлены коэффициенты классифицирующих функций. Например, первая функция имеет вид

$$-41{,}0048+9{,}80432*x1+3{,}75737*x2+8{,}52742*x3+6{,}57962*x4+$$
$$3{,}35299*x5+3{,}68837*x6+ 2{,}39136*x7+4{,}11553*x8.$$

Для более детального анализа нажмите на кнопку ***Tables***, в окне выберите ***Classification Table***, нажмите кнопку ОК. Появится окно, в котором в верхней части находится таблица.

Classification Table

Actual Group	*Group Size*	*Predicted Group 1*	*2*	*3*	*4*
1	28	22	5	1	0
		(78,57%)	(17,86%)	(3,57%)	(0,00%)
2	25	1	16	7	1
		(4,00%)	(64,00%)	(28,00%)	(4,00%)
3	26	3	6	17	0
		(11,54%)	(23,08%)	(65,38%)	(0,00%)
4	24	0	0	0	24
		(0,00%)	(0,00%)	(0,00%)	(100,00%)

Percent of cases correctly classified: 76,70%

Из этой таблицы можно получить сведения об итоговых результатах диагностики. Первая группа имеет размер 28, точность диагностики составляет 78,574 % (22 человека из 28). Вторая группа — 25 человек, точность диагностики 64,004 % (16 человек из 25) – берем данные с главной диагонали матрицы. Третья группа — точность диагностики 65,384 %, 17 человек из 26 и четвертая — 100 %, 24 человека из 24. Таким образом, результаты не слишком точные, но в какой-то мере они могут содействовать окончательному заключению специалиста. Вместе с тем констатация отсутствия острого аппендицита (4-я группа, неподтвержденный диагноз) осуществляется с надежностью 100 %.

Рассмотрим вторую часть таблицы, там дается подробный разбор результатов диагностики посредством полученных классифицирующих функций.

	Actual	Highest	Highest	Squared		2nd Highest	2nd Highest	Squared	
Row	Group	Group	Value	Distance	Prob.	Group	Value	Distance	Prob.
1	1	*2	35,3337	2,70742	0,4088	1	35,3061	2,76268	0,3977
2	1	1	35,2934	1,75972	0,9407	2	32,3005	7,74559	0,0472
3	1	1	41,8857	1,32991	0,6694	2	41,0726	2,95607	0,2969
4	1	1	34,6472	4,62564	0,8469	3	32,5014	8,91736	0,0991
5	1	*3	36,19	1,93146	0,4709	2	35,6238	3,06377	0,2673
6	1	*2	31,7568	1,46753	0,5008	1	31,156	2,66918	0,2746
7	1	1	35,2934	1,75972	0,9407	2	32,3005	7,74559	0,0472
8	1	*2	46,2759	3,43537	0,6186	1	45,6431	4,70091	0,3285
9	1	1	36,8515	5,47235	0,9011	2	34,6124	9,95051	0,0960
10	1	1	43,8335	0,906759	0,8016	2	42,2554	4,06301	0,1654
11	1	1	26,1399	3,78058	0,7475	3	24,6031	6,8543	0,1608
12	1	1	43,8335	0,906759	0,8016	2	42,2554	4,06301	0,1654
13	1	1	29,9926	4,37252	0,8468	2	28,1456	8,0667	0,1335
14	1	1	36,4568	0,988913	0,4907	2	35,9118	2,07895	0,2845
15	1	1	47,5909	3,0215	0,4981	2	47,4586	3,28604	0,4364
16	1	1	38,1283	0,367121	0,8927	2	35,8694	4,885	0,0933
17	1	1	47,5782	2,16198	0,9558	2	44,4254	8,46765	0,0408
18	1	1	23,1615	5,91017	0,9096	2	20,6981	10,837	0,0774
19	1	1	50,4131	1,88351	0,9147	2	47,9943	6,72118	0,0814
20	1	1	26,9942	3,21866	0,8895	2	24,3226	8,56204	0,0615
21	1	1	52,3609	1,58457	0,9570	2	49,1771	7,95233	0,0396
22	1	1	42,1821	0,26668	0,6147	2	41,3627	1,90536	0,2709
23	1	1	26,766	0,831189	0,7130	2	25,3788	3,60555	0,1781

Для каждого объекта приведены значения двух наибольших дискриминантных функций и результат отнесения к тому или иному классу. Неправильно классифицированные объекты помечены звездочкой.

В задачах такого типа обычно возникает необходимость произвести дискриминацию какого-то набора данных, т. е. отнесение нового объекта к одному из существующих классов. В нашем случае после обследования больного и получения результатов осмотра определить, к какой группе он относится. Допустим, данные его осмотра имеют следующий вид:

$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$
1	1	2	1	1	0	2	0

К какой группе больных его следует отнести?

Для решения этой задачи введем данные в наш файл *appendix.sf*, оставив пустой клетку в поле ***Group***. Должна получиться следующая таблица

98	4	1	2	1	1	0	0	0
99	4	1	1	1	2	1	0	0
100	4	1	2	1	1	0	0	0
101	4	1	1	2	1	0	0	0
102	4	2	1	1	1	0	0	2
103	4	1	2	1	1	0	0	0
104		1	1	2	1	1	0	2

Запустите снова дискриминантный анализ. В последних строках ***Classification Table*** получим:

100	4	4	8,01816	0,540525	0,9998	3	-0,754562	18,086	0,0002
101	4	4	11,481	2,53537	1,0000	3	0,381614	24,7342	0,0000
102	4	4	11,582	8,39891	0,9716	3	8,03975	15,4835	0,0281
103	4	4	8,01816	0,540525	0,9998	3	-0,754562	18,086	0,0002
104		4	11,8014	3,28987	0,9735	3	8,13574	10,6211	0,0249

Из этой таблицы можно сделать вывод о том, что больной относится к четвертой группе, т. е. у него, по-видимому, нет аппендицита.

ЗАДАНИЯ

1. В файле SobakiVolki.sf указаны размеры челюстей и зубов 30 собак (№ 1–30) и 12 волков (№ 31–42) и ископаемого черепа неизвестного животного. Смысл переменных можно посмотреть в комментариях к ним в файле. Требуется решить, к какому классу (волков или собак) следует отнести неизвестное животное.
2. Рассмотрим задачу о рынке ценных бумаг из работы № 10. Обратите внимание на последний столбец в файле *Growth.sf*. В нем по первым шестнадцати фондам приводятся рекомендации экспертов (*Buy* – купить, *Sell* – продать, *Hold* – придержать). Используя методы дискриминантного анализа сформулируйте рекомендации по акциям последних четырех фондов.