Дискриминантный анализ

И ПРИМЕР

Задачи дискриминантного анализа

Задачи **первого типа** часто встречаются в торговле.

Допустим, что мы располагаем информацией о некотором числе покупателей и совершенных покупках. На основе этой информации нужно найти функцию, позволяющую поставить в соответствие новым покупателям характерные для их соц.-дем. статуса покупки.

Построение такой функции и составляет задачу дискриминации.

Задачи дискриминантного анализа

Второй тип задач относится к ситуации, когда признаки принадлежности покупателя к той или иной группе потеряны, и их нужно восстановить.

Примером может служить определение возрастной группы покупателей по истории совершенных покупок.

Задачи дискриминантного анализа

Задачи третьего типа связаны с предсказанием будущих событий на основании имеющихся данных.

Такие задачи возникают при прогнозе покупок, трафика. Например, прогноз вероятности посещения ТРЦ определенной группой покупателей

Задачи дискриминантного анализа

Целью задачи дискриминации является изучение основных процедур дискриминантного анализа:

- дискриминации и классификации
- построение и определение количества дискриминантных функций и их разделительной способности
- нахождение классифицирующих функций

Результаты ДА

- Основным результатом проведения дискриминантного анализа являются рассчитанные вероятности попадания каждого респондента в ту или иную группу, а также переменная, кодирующая принадлежность их к данным группам.
- Наряду с этой информацией по результатам дискриминантного анализа можно составить уравнение дискриминантной функции.

Дискриминантный анализ

Дискриминантный анализ			
Зависимые переменные		**Независимые переменные**	
Количество	**Тип**	**Количество**	**Тип**
Одна	Номинальная Порядковая	Любое	Любой

Зависимая переменная

- При выборе зависимой переменной (Y) для дискриминантного анализа следует помнить, что увеличение числа категорий в ней влечет уменьшение точности и надежности модели.
- Поэтому рекомендуется использовать в качестве зависимых переменные с малым количеством категорий (2-3), например:
  - Посетит – не посетит
  - Купит – не купит
  - Пол
  - Вид занятости

ПРИМЕР

Проводится маркетинговое исследование потенциального спроса на услуги нового ТРЦ.

Респонденты в ходе опроса отвечают на вопрос Будете ли Вы посещать новый комплекс? с вариантами ответа Да и Нет. В качестве независимых переменных, характеризующих респондентов, выделены:

- возраст ;
- занятость;
- среднемесячный доход на члена семьи;
- количество членов семьи;
- среднемесячные расходы на досуг;
- пол.

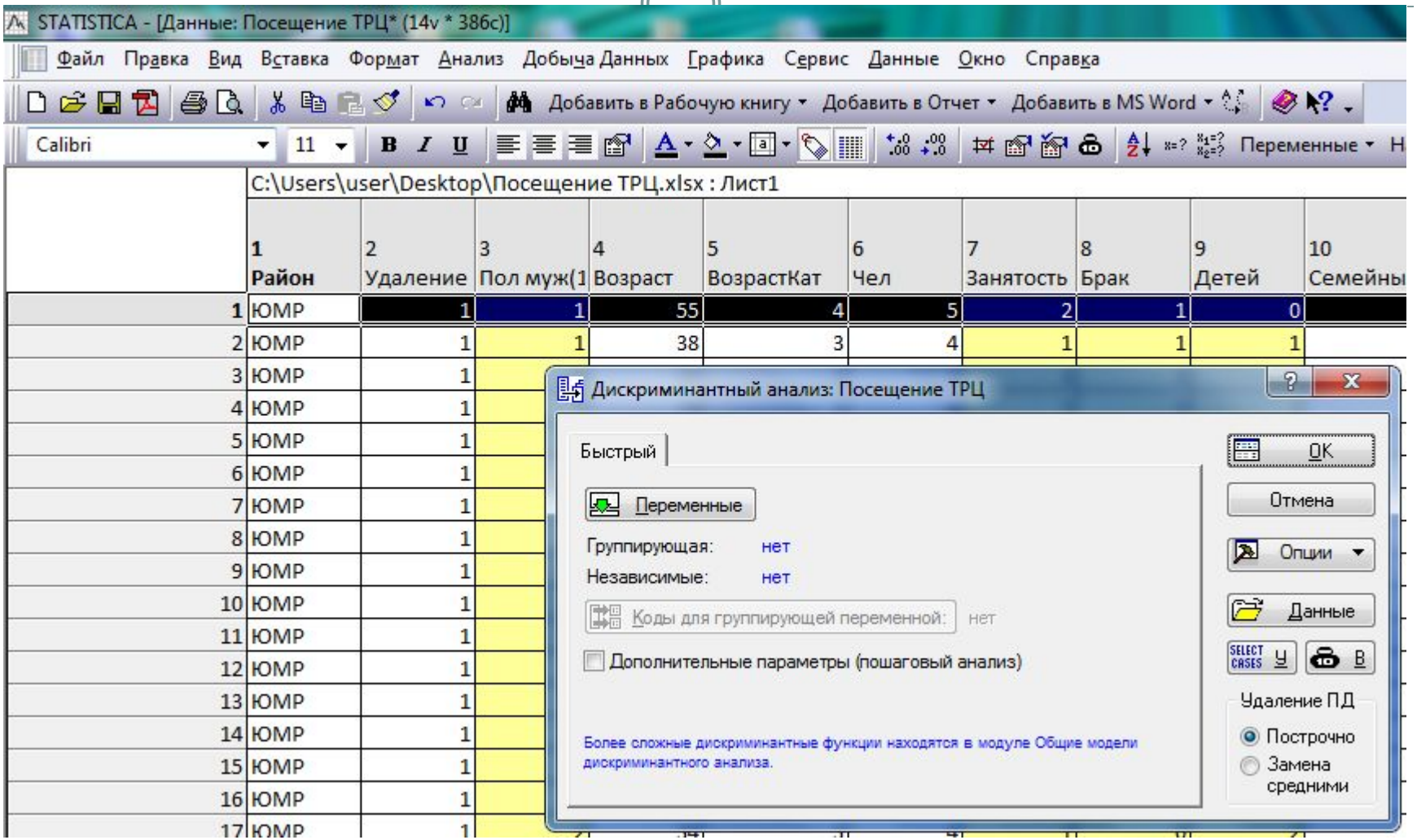
Зависимая переменная

В результате дискриминантного анализа разделим респондентов:

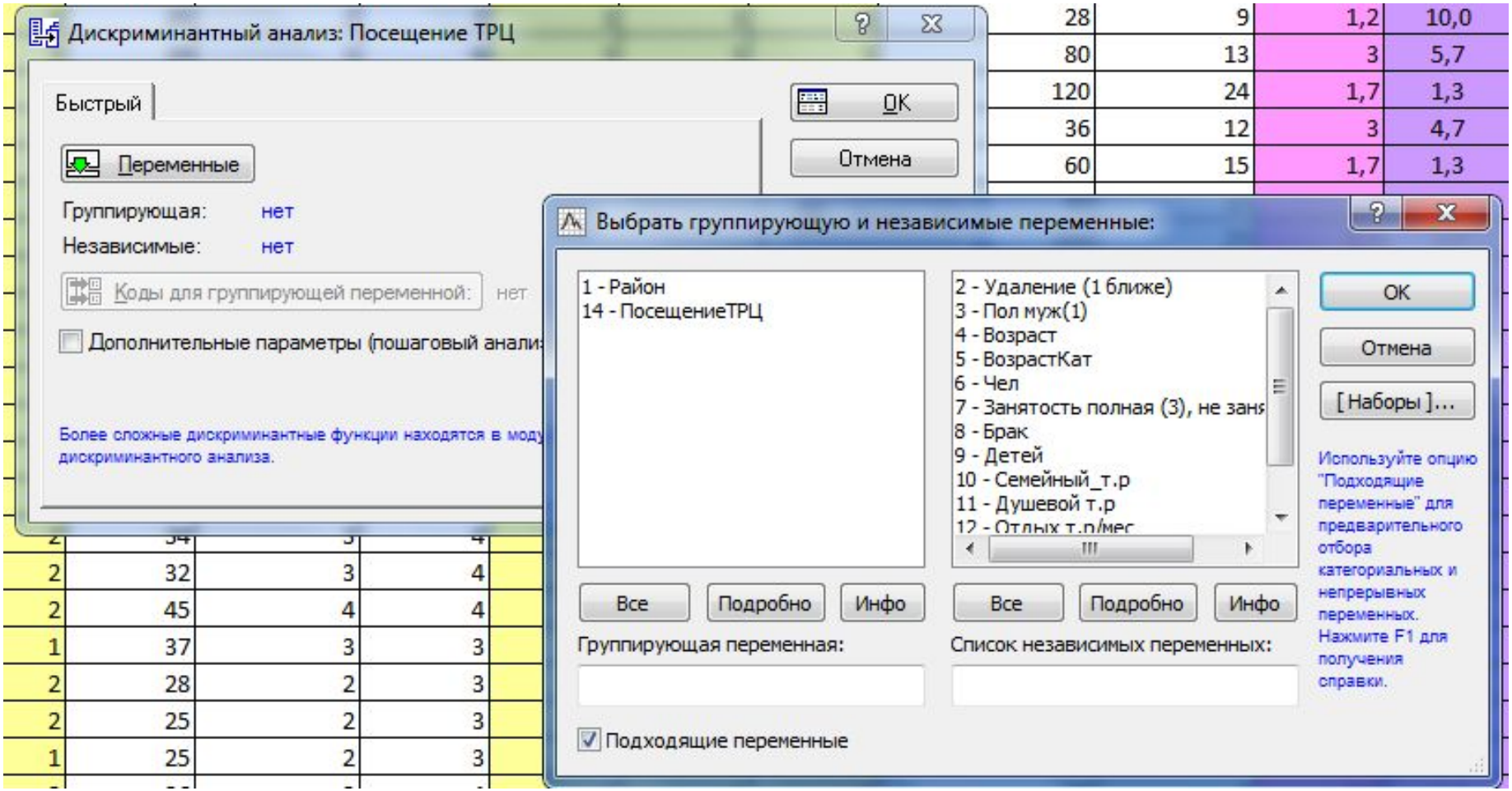
- на посетителей
- не посетителей нового центра

на основании выделенных социально-демографических характеристик опрошенных.

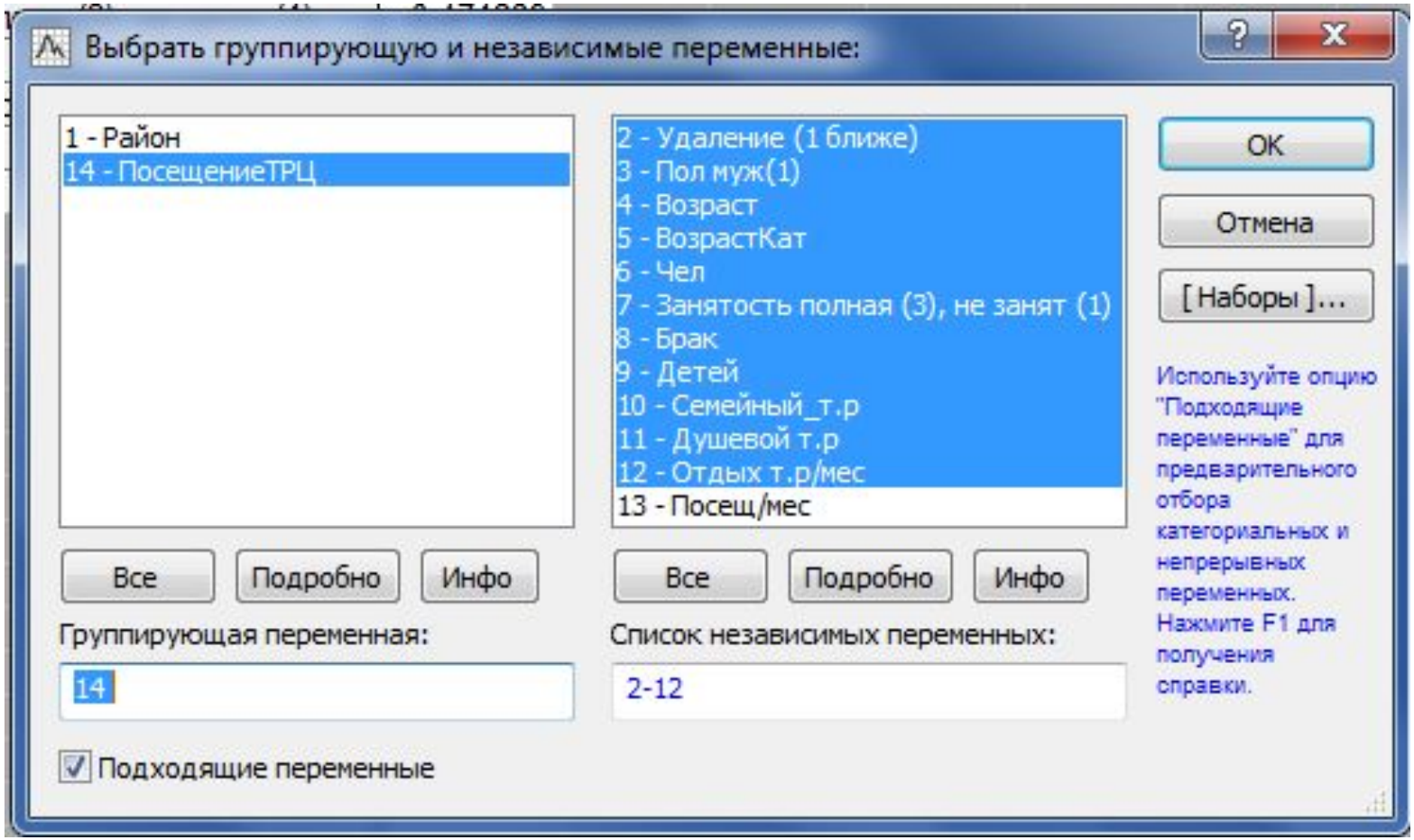
Скрин. Дискриминантный анализ



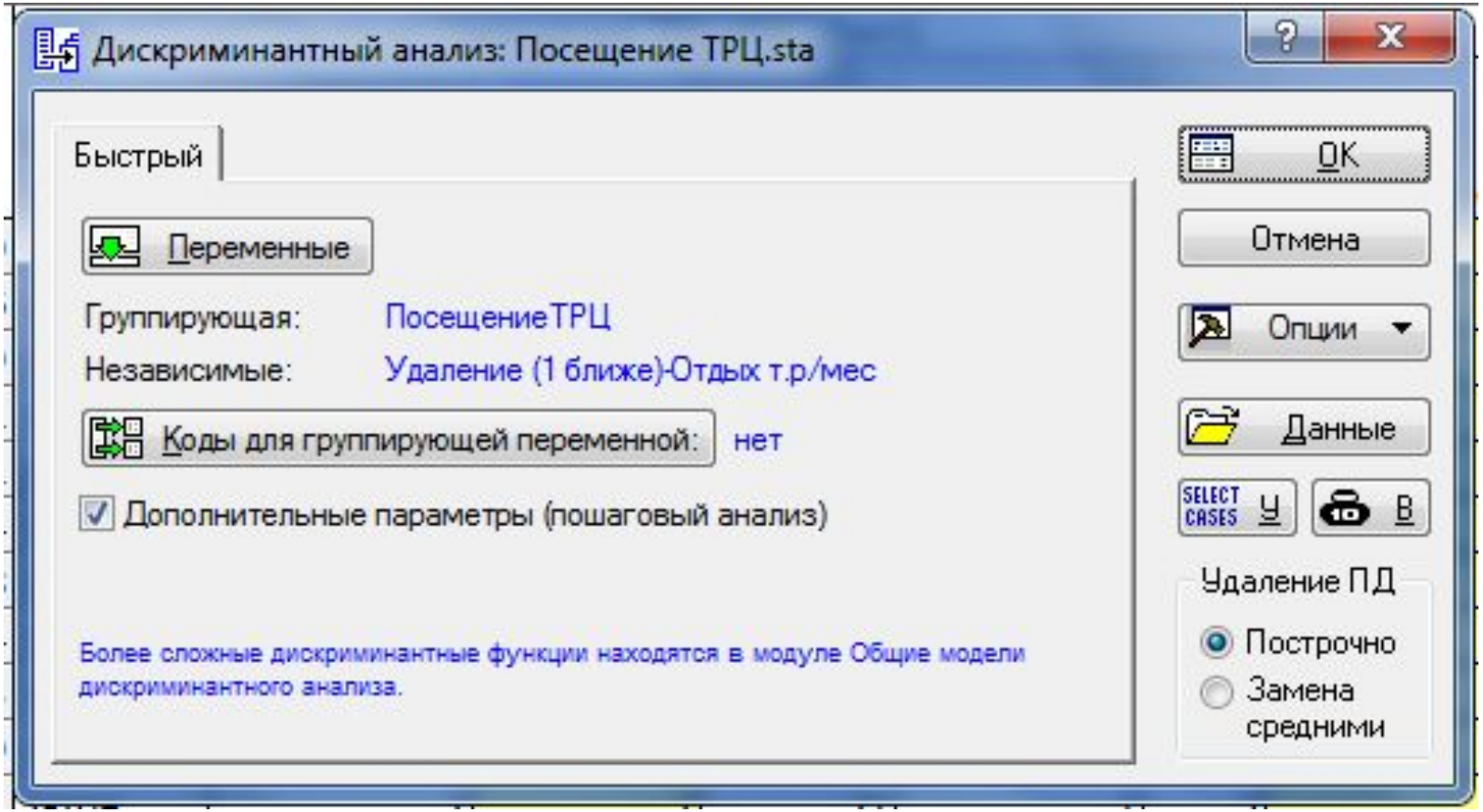
Скрин. Дискриминантный анализ



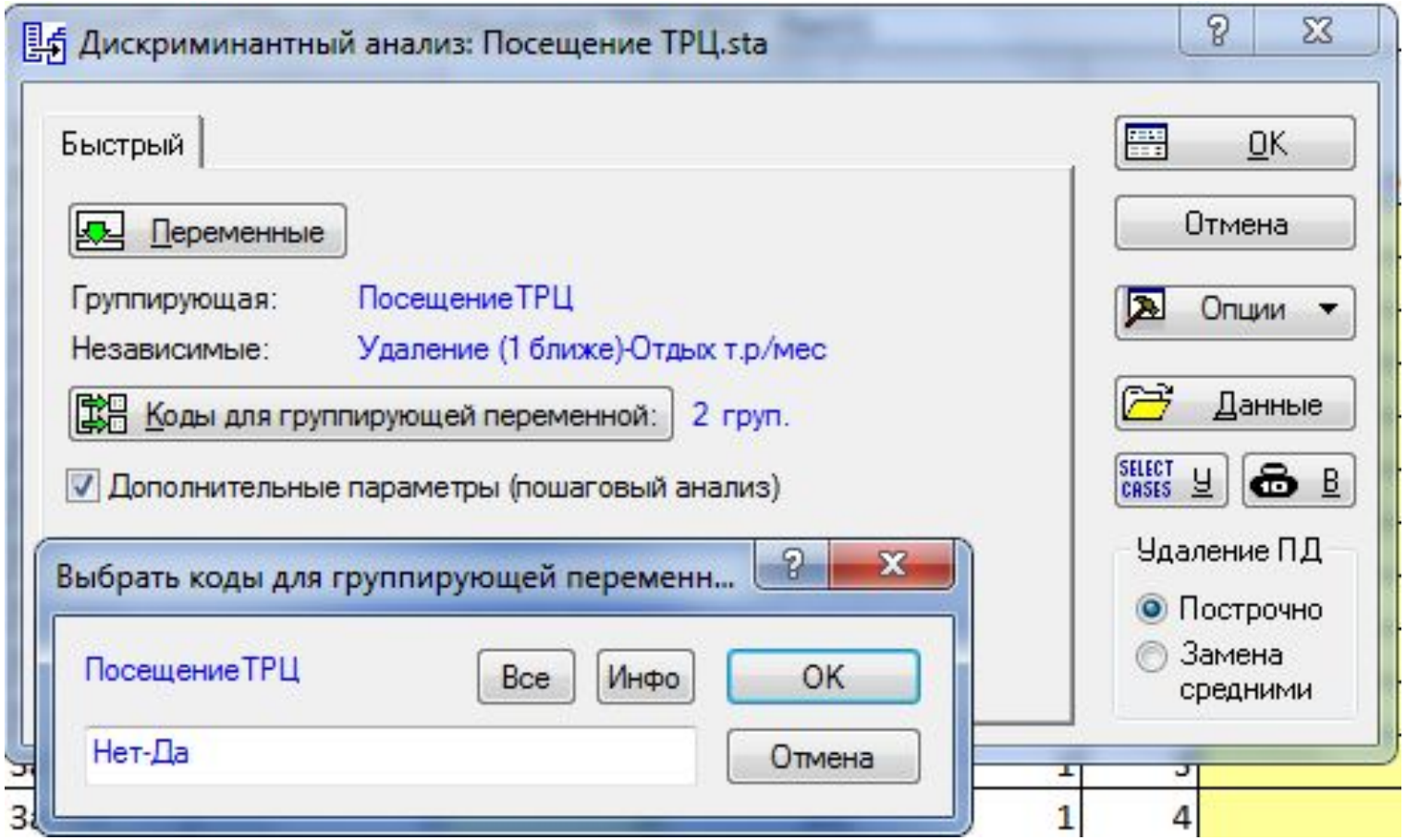
Скрин. Дискриминантный анализ



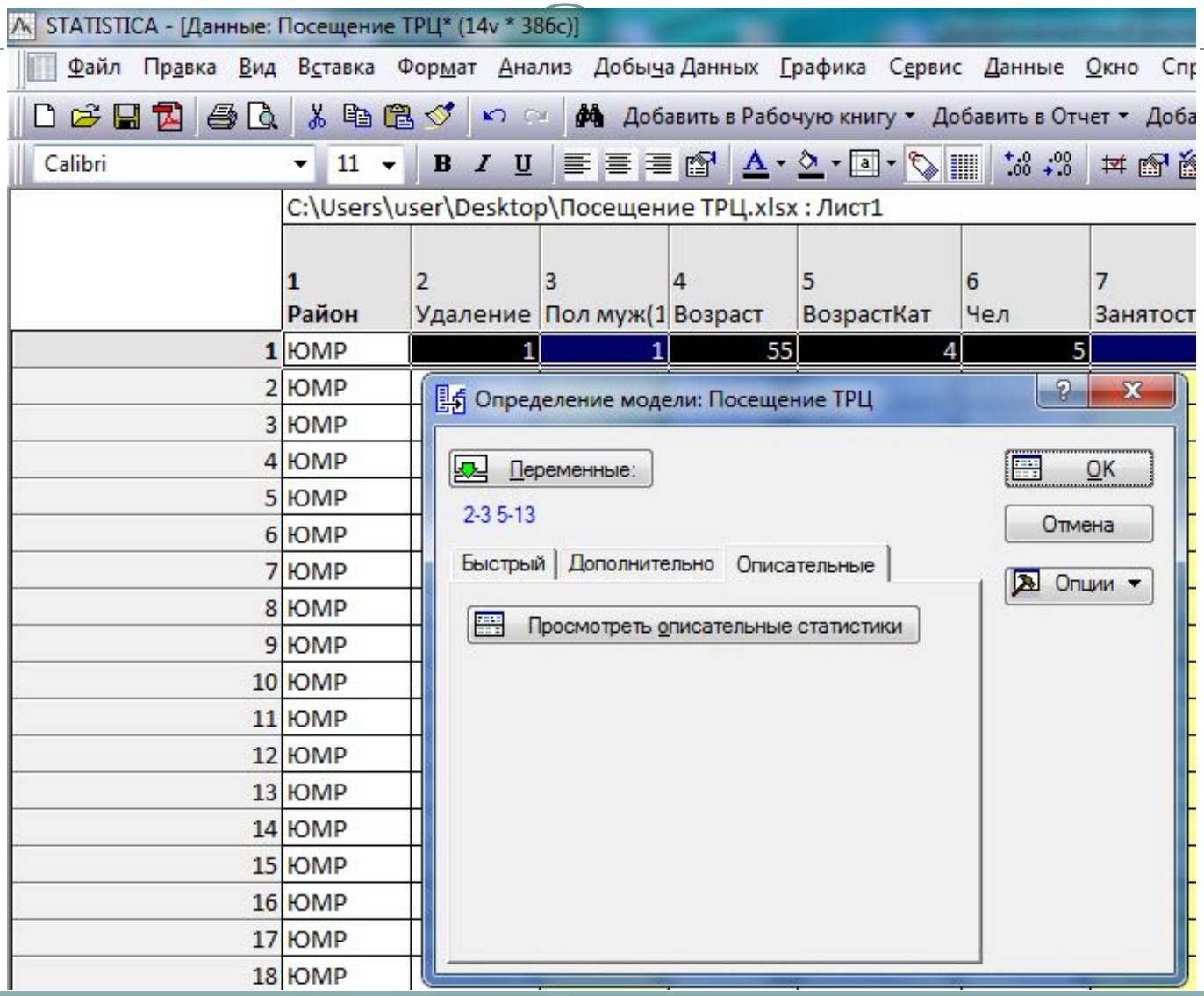
Скрин. Дискриминантный анализ



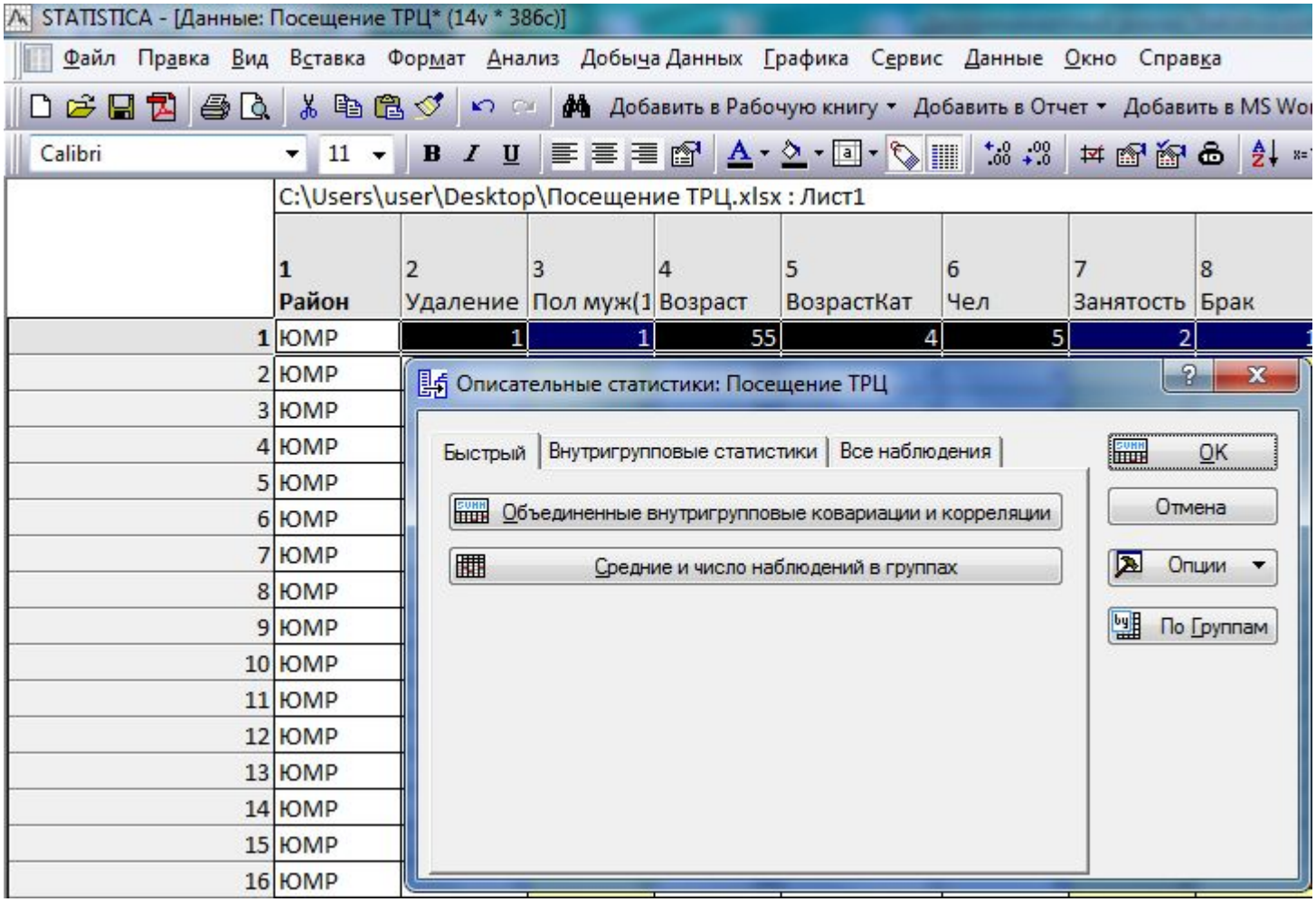
Скрин. Дискриминантный анализ



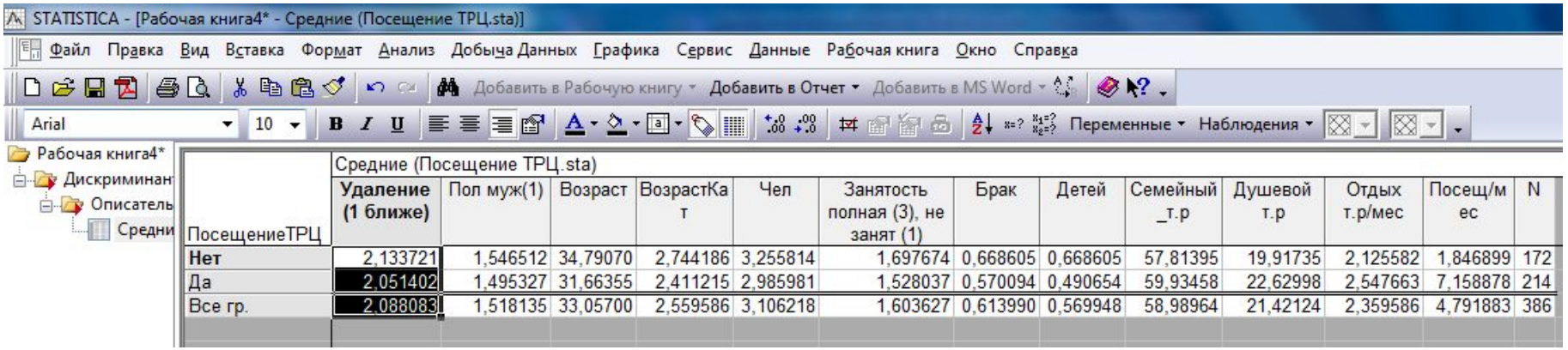
Скрин. Статистики



Скрин. Статистики



Скрин. Статистики



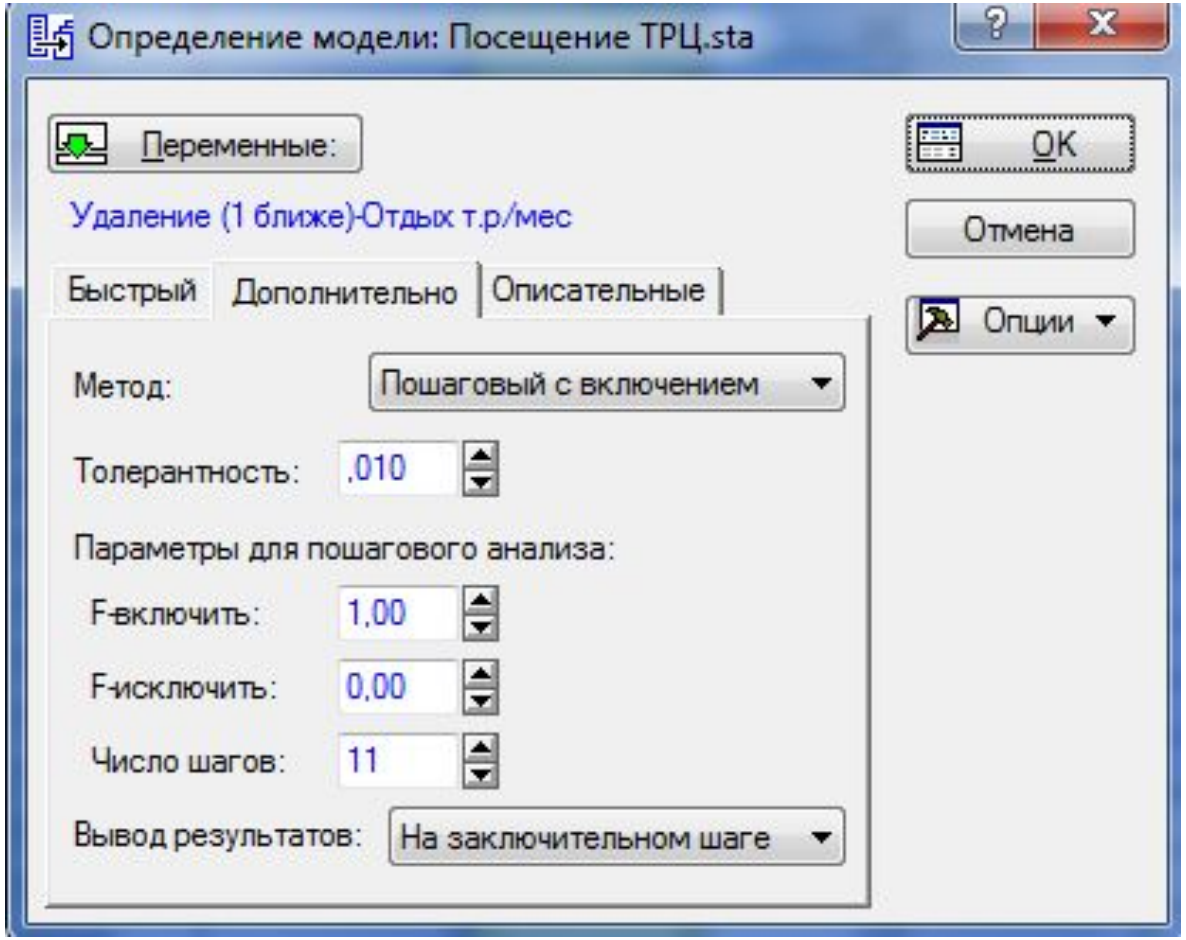
Средние (Посещение ТРЦ.sta)

ПосещениеТРЦ	**Удаление (1 ближе)**	Пол муж(1)	Возраст	ВозрастКат	Чел	Занятость полная (3), не занят (1)	Брак	Детей	Семейный_т.р	Душевой т.р	Отдых т.р/мес	Посещ/мес	N
Нет	2,133721	1,546512	34,79070	2,744186	3,255814	1,697674	0,668605	0,668605	57,81395	19,91735	2,125582	1,846899	172
Да	2,051402	1,495327	31,66355	2,411215	2,985981	1,528037	0,570094	0,490654	59,93458	22,62998	2,547663	7,158878	214
Все гр.	2,088083	1,518135	33,05700	2,559586	3,106218	1,603627	0,613990	0,569948	58,98964	21,42124	2,359586	4,791883	386

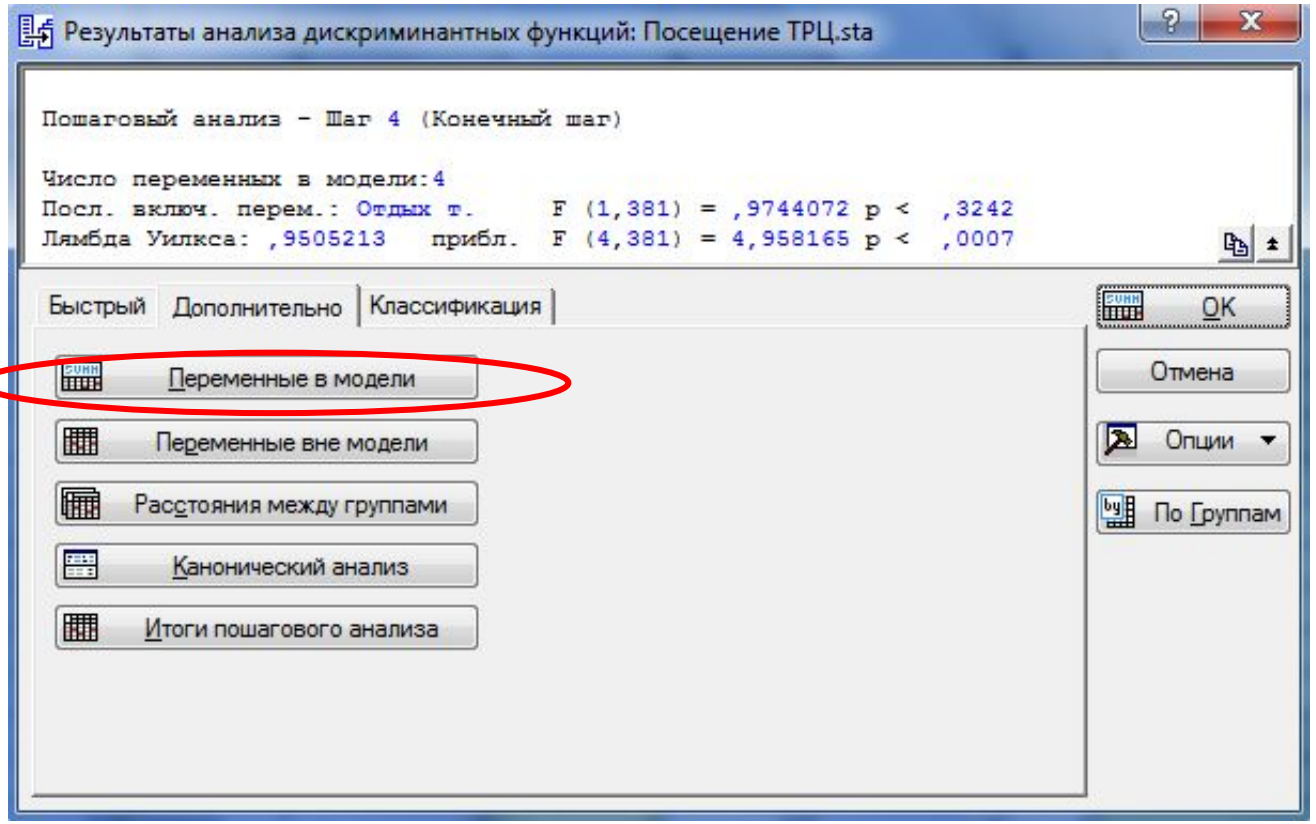
Выбор анализа дискриминантной функции

- Теперь вернемся к первичной цели анализа. Нажмем на кнопку Отмена в диалоговом окне Описательные статистики для того, чтобы вернуться к диалоговому окну Определение Модели.
- Для того чтобы увидеть, что происходит на каждом шаге дискриминантного анализа, необходимо выполнить пошаговый анализ.
- Во вкладке Дополнительно, в списке Метод установите значение Пошаговый с включением. При такой установке программа будет вводить переменные одну за другой, каждый раз выбирая переменную, вносящую наибольший вклад в дискриминацию.

Анализ


Определение модели: Посещение ТРЦ.sta
Переменные:
Удаление (1 ближе)-Отдых т.р/мес
OK
Отмена
Опции
Быстрый
Дополнительно
Описательные
Метод:
Пошаговый с включением
Толерантность:
,010
Параметры для пошагового анализа:
F-включить:
1,00
F-исключить:
0,00
Число шагов:
11
Вывод результатов:
На заключительном шаге

Анализ



Теперь нажмем на кнопку Переменные в модели
для обзора независимых вкладов каждой переменной
в общую дискриминацию

Анализ

Итоги анализа дискриминантн. функций (Посещение ТРЦ.sta)
Шаг 4, Переменных в модели: 4; Группир.: ПосещениеТРЦ (2 гр.)
Лямбда Уилкса: ,95052 прибл. F (4,381)=4,9582 p< ,0007

N=386	**Уилкса Лямбда**	Частная Лямбда	F-исключ (1,381)	p-уров.	Толер.	1-толер. (R-кв.)
ВозрастКат	0,972580	0,977320	8,841620	0,003132	0,944316	0,055684
Занятость полная (3), не занят (1)	0,960199	0,989921	3,879169	0,049613	0,917113	0,082887
Чел	0,957166	0,993058	2,663491	0,103501	0,956353	0,043647
Отдых т.р/мес	0,954816	0,995502	1,721506	0,190289	0,974487	0,025513

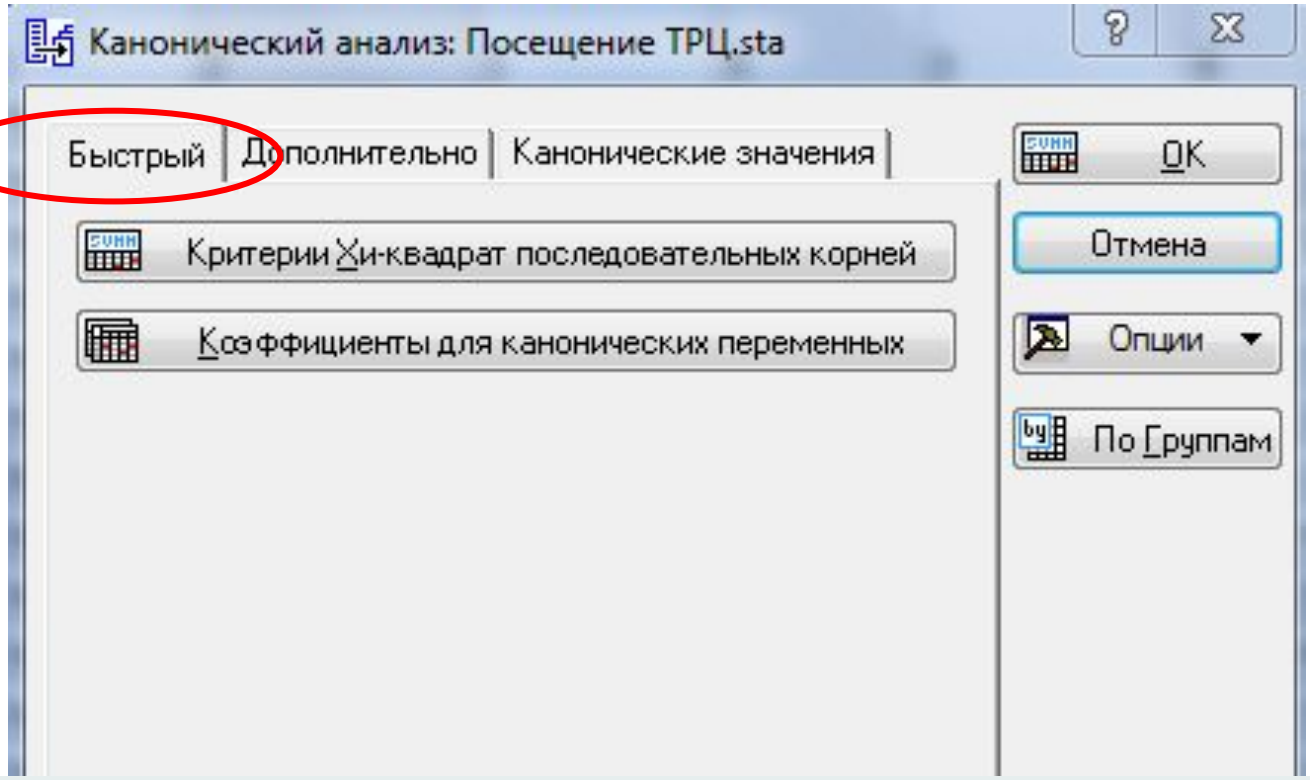
Наибольшее влияние на готовность посетить ТРЦ оказывают:
ВозрастКат, Занятость, Человек в семье и Затраты на отдых

Чем меньше частная статистика Уилкса лямбда,
тем больше вклад в общую дискриминацию

Анализ

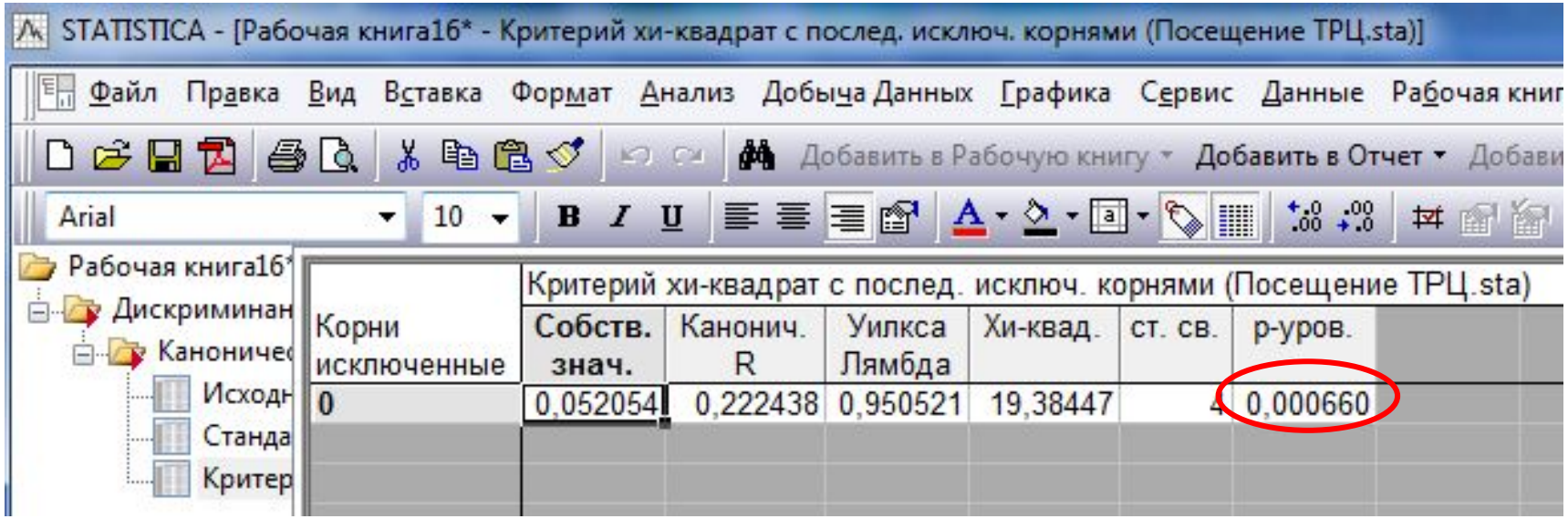
- Если толерантность имеет значение, меньшее, чем значение по умолчанию 0.01 (или установленное специально пользователем), то эта переменная признается неинформативной и не включается в модель, поскольку не несет дополнительной информации по сравнению с остальными переменными.

Проверка значимости



Значимость корней. Сначала определим, является ли дискриминантная функция статистически значимой. Нажмите на кнопку Критерий Хи-квадрат последовательных корней и увидите следующую таблицу:

Анализ дискриминирующих переменных

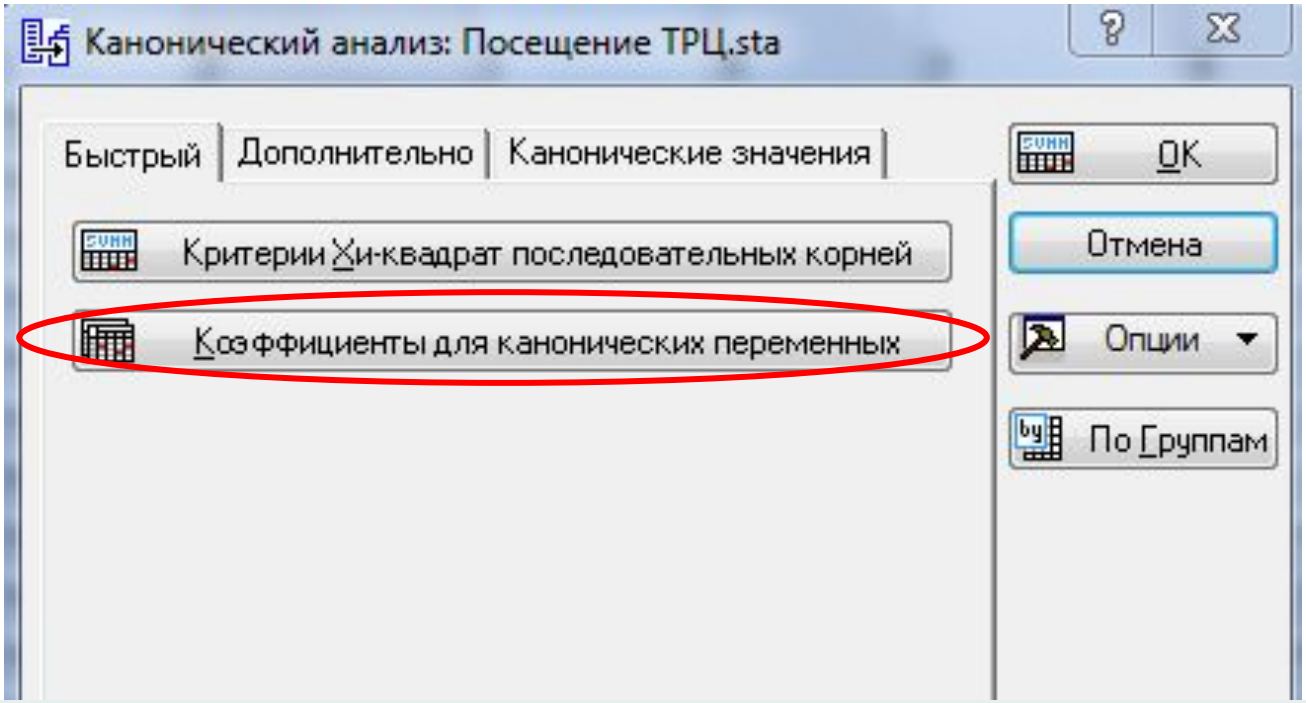


Критерий хи-квадрат с послед. исключ. корнями (Посещение ТРЦ.sta)

Корни исключенные	Собств. знач.	Канонич. R	Уилкса Лямбда	Хи-квад.	ст. св.	р-уров.
0	0,052054	0,222438	0,950521	19,38447	4	0,000660

Дискриминантная функция значима $p << 0{,}01$.

Анализ дискриминирующих переменных



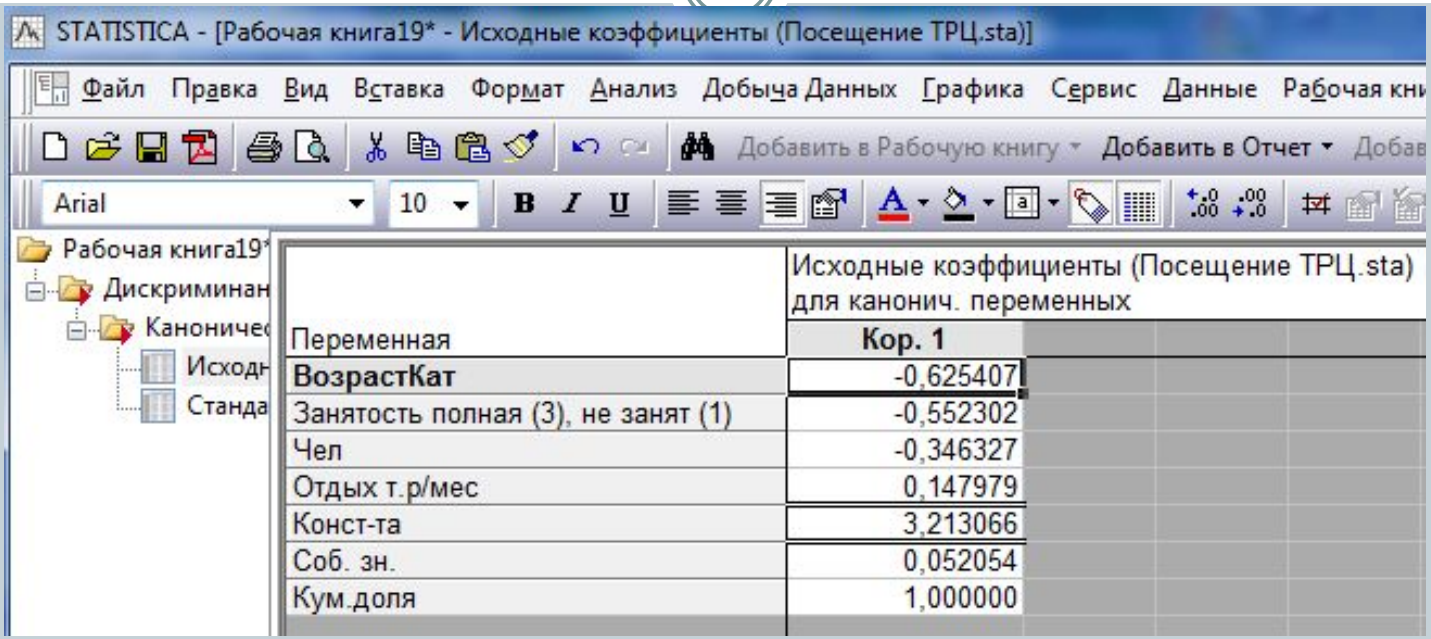
Канонический анализ. Чтобы увидеть, как четыре переменные разделяют посетителей, вычислим действительную дискриминантную функцию.

Нажмите на кнопку →Коэффициенты для канонических переменных.

Коэффициенты дискриминантной функции

- Нажмите на кнопку Коэффициенты для канонических переменных в диалоговом окне Канонический анализ.
- Будут получены две таблицы, одна для Исходных коэффициентов и другая для Стандартизованных коэффициентов.

Решение

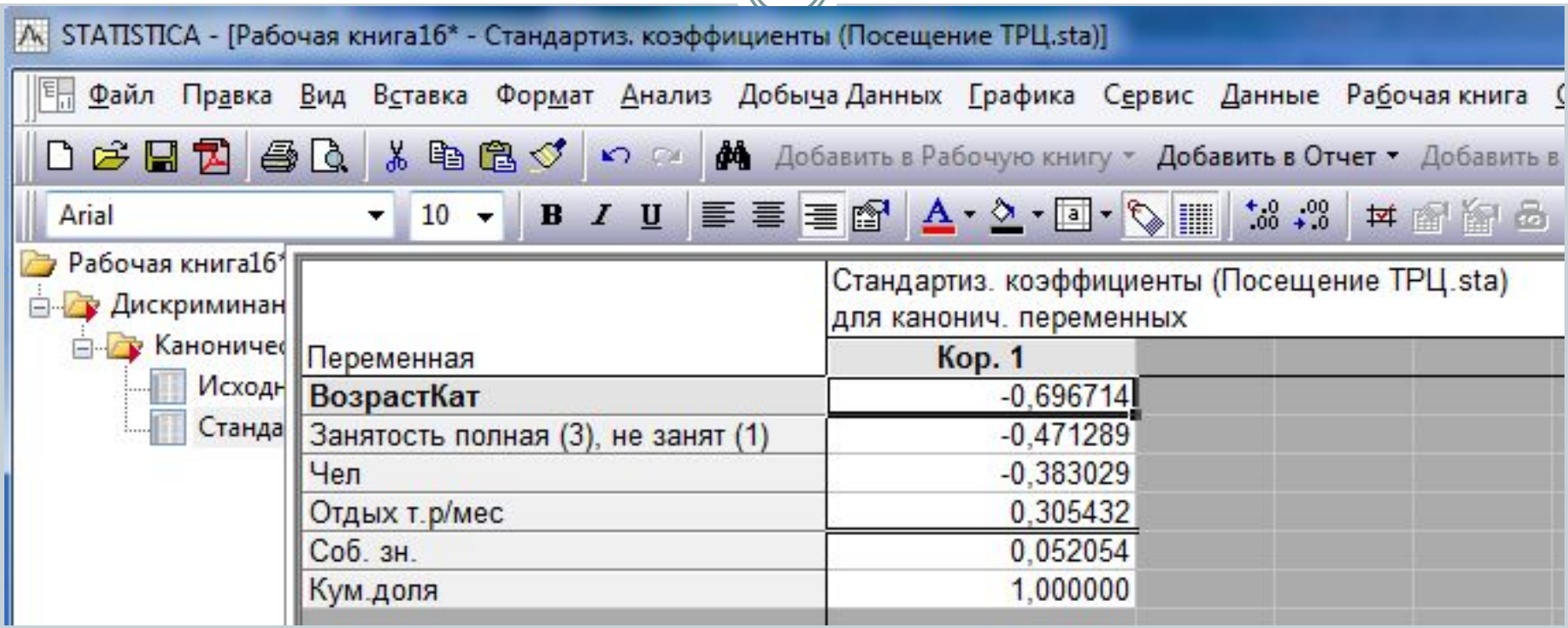


Переменная	Исходные коэффициенты (Посещение ТРЦ.sta) для канонич. переменных Кор. 1
ВозрастКат	-0,625407
Занятость полная (3), не занят (1)	-0,552302
Чел	-0,346327
Отдых т.р/мес	0,147979
Конст-та	3,213066
Соб. зн.	0,052054
Кум.доля	1,000000

- Исходные коэффициенты
- Уравнение дискриминантной функции имеет вид:

План=-0,625×ВозрастКат - 0,552×Занятость - 0,346 ×Чел + 0,148×Отдых +3,213

Решение



Переменная	Стандартиз. коэффициенты (Посещение ТРЦ.sta) для канонич. переменных Кор. 1
ВозрастКат	-0,696714
Занятость полная (3), не занят (1)	-0,471289
Чел	-0,383029
Отдых т.р/мес	0,305432
Соб. зн.	0,052054
Кум.доля	1,000000

- Стандартизованные коэффициенты - это те коэффициенты, которые обычно используются для интерпретации, так как они относятся к нормированным переменным и поэтому должны находиться в сравнимых масштабах

Применение

-0,697 ВозрастКат
-0,471 Занятость
- 0,1217 Чел
+ 0,305 Отдых

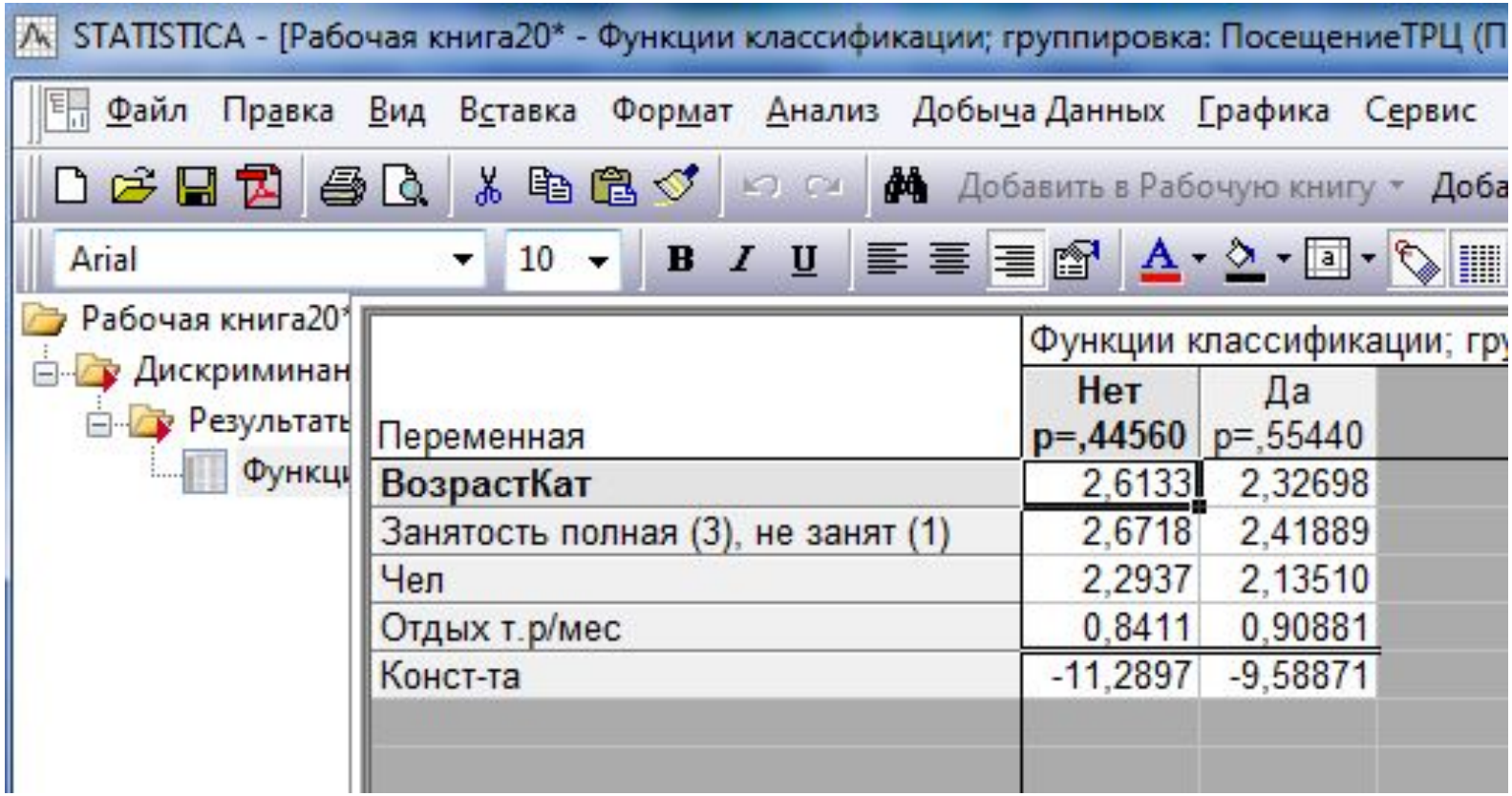
Интерпретация вычисленных по формуле значений:

- Можно сделать вывод о том, что наиболее вероятные посетители ТРЦ: **более молоды, не имеют полной занятости, из малой семьи и больше других тратят на отдых**

Вывод результатов

- Нажмите на кнопку Функции классификации во вкладке Классификация диалогового окна Результаты анализа дискриминантных функций для того, чтобы увидеть эти функции.

Вывод результатов

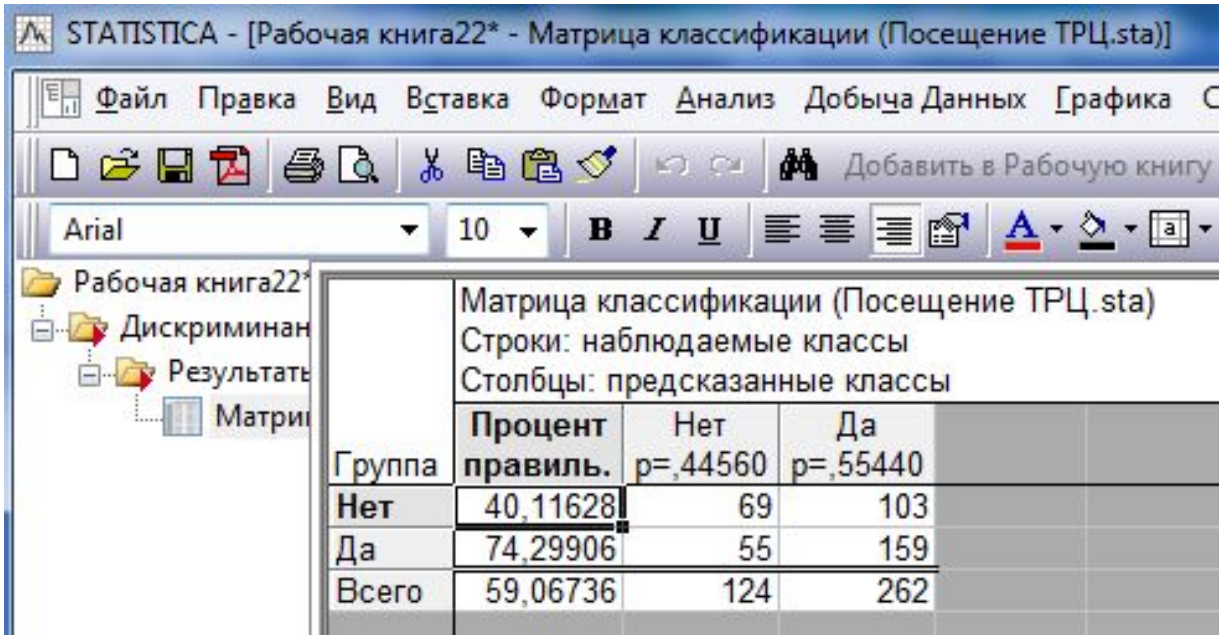


Переменная	Функции классификации; гру Нет p=,44560	Да p=,55440
ВозрастКат	2,6133	2,32698
Занятость полная (3), не занят (1)	2,6718	2,41889
Чел	2,2937	2,13510
Отдых т.р/мес	0,8411	0,90881
Конст-та	-11,2897	-9,58871

Вывод результатов

- Нажмите на кнопку Матрица классификации во вкладке Классификация диалогового окна Результаты анализа дискриминантных функций

Вывод результатов



Матрица классификации (Посещение ТРЦ.sta)
Строки: наблюдаемые классы
Столбцы: предсказанные классы

Группа	Процент правиль.	Нет p=,44560	Да p=,55440
Нет	40,11628	69	103
Да	74,29906	55	159
Всего	59,06736	124	262

В первом столбце таблицы вы видите процент наблюдений, которые были правильно классифицированы для каждой совокупности полученными функциями классификации.

- Потенциальные посетители нового ТРЦ (Да) определяются функцией с вероятностью 74% 159/(55+159)

Точность модели

- Результаты оценки корректности классификации варьируют в пределах от 50 % до 100 %.
- Получен результат для потенциальных посетителей — 74 %

Задание

- Какие зависимые переменные следует проверить на возможность применения дискриминантной модели?
- Проведите дискриминантный анализ Ваших переменных