МЕТОДИЧЕСКОЕ РУКОВОДСТВО ДЛЯ ПОЛЬЗОВАТЕЛЯ ПАКЕТА САИСИ

А.-М. Парринг, Е.-М. Тийт

1986

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Кафедра математической статистики

МЕТОДИЧЕСКОЕ РУКОВОДСТВО ДЛЯ ПОЛЬЗОВАТЕЛЯ ПАКЕТА САИСИ

А.-М. Парринг, Е.-М. Тийт

ТАРТУ 1986

Утверждено на заседании совета математического факультета ТГУ 20 июня 1986 года

МЕТОДИЧЕСКОЕ РУКОВОДСТВО ДЛЯ ПОЛЬЗОВАТЕЛЯ ПАКЕТОМ САИСИ.
Составители А.-И. П а р р и н г, Е.-М. Т и й т.
На русском языке.
Тартуский государственный университет.
ЭССР, 202400, г.Тарту, ул.Юликооли, 18.
Ответственный редактор Т. Мольс.
Подписано к печати 2.07.1986.
Формат 60x84/16.
Бумага ротаторная.
Машинопись. Ротапринт.
Условно-печатных листов 16,28.
Учетно-издательских листов 12,21. Печатных листов 17,5.
Тираж 400.
Заказ № 643.
Цена 40 коп.
Типография ТГУ, ЭССР, 202400, г.Тарту, ул.Тийги, 78.

ВВЕДЕНИЕ

В настоящем методическом руководстве даются указания к применению программ "СИСТЕМА АНАЛИЗА И ИНТЕРПРЕТАЦИИ СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ" (САИСИ) при решении практических задач анализа данных.

В первой части руководства даются общая характеристика пакета (глава I) и методические указания к созданию плана обработки (глава II). Следует описание управляющего языка пакета САИСИ (глава III и IУ), практическое указание к представлению заказа на языке САИСИ (глава У) и руководство к пользованию необходимыми управляющими картами операционной системы (глава УI).

Во второй части руководства описываются статистические процедуры, реализованные в пакете САИСИ. В главах УII и УIII исследуются процедуры первичной обработки, в главе IX - анализ зависимостей между признаками. Главы XI и XII посвящены методам вторичного анализа и построению статистических моделей, причём в главе XI рассматриваются описательные методы (факторный анализ, методы группирования), а в главе XII - прогностические методы (дискриминантный и регрессионный анализы).

В главе XIII излагается практический пример применения пакета.

Авторы выражают благодарность коллегам Калеву Пярна и Владимиру Фляйшеру за прочитание рукописи и за полезные замечитания, а также Айно Куренийт и Лийвии Йыесаар за оформление рукописи.

I. ОБЩАЯ ХАРАКТЕРИСТИКА ПАКЕТА САИСИ

Пакет прикладных программ САИСИ – "Система анализа и интерпретации статистической информации" является по своему составу программ и выбору возможностей пакетом среднего уровня, весьма подходящим для пользователей-практиков разных специальностей, таких как психологи, социологи, экономисты, а также медики, биологи и др.

Для характеристики пакета САИСИ можно отметить следующее.

1) Выбор статистических процедур пакета САИСИ покрывает практически все традиционные разделы как первичного статистического анализа, так и классического многомерного анализа (имеются программы для вычисления одно- и многомерных распределений, корреляционного, дисперсионного, факторного, канонического и дискриминантного анализа). Кроме того, существуют программы для некоторых методов таксономии, шкалирования по Гуттману и для анализа временных рядов.

2) Среди средств пакета САИСИ имеется много программ для оперирования с массивами данных и для преобразования признаков, в результате чего пакет отличается гибкостью и большим количеством возможностей для комбинирования статистических процедур.

Сюда относятся: вычисление новых признаков, перекодирование признаков, устранение, прибавление, перестановка признаков; выбор объектов, их взвешивание, сортирование, устранение; образование подфайлов, их комбинирование и устранение, обработка данных, содержащих неправильные или отсутствующие измерения

3) Пакет САИСИ имеет простой язык управления, базирующийся на обычном английском языке. Очень целесообразно выбран режим умолчания, с помощью которого можно пользоваться пакетом.при минимальном знании языка.

Пользование пакетом САИСИ при помощи его языка не требует от пользователя умения программировать. Методы математической статистики надо знать только в той степени, в какой ими пользуются.

4) Распечатки пакета САИСИ снабжены необходимыми комментариями и объяснениями (на английском языке), в общем хорошо органозованными, содержащими достаточно информации для интерпретации результатов, причем содержание выходной информации регулируется языком (карты "OPTIONS" и "STATISTICS"). Имеется возможность прибавлять комментарии со стороны пользователя.

Описание данных, названия задач и прогонов и комментарий, заданные пользователем, могут быть сделаны и на русском языке.

5) Все программы пакета САИСИ работают корректно. Их точность достаточно велика (в стандартных статистических процедурах до 5-6 знаков после запятой, а иногда даже больше).

II. ПОДГОТОВКА К АНАЛИЗУ ДАННЫХ

2.I. Подготовка данных

Начнем рассмотрение подготовки материала с того момента, когда исследователь имеет в своих руках протоколи наблюдений - т.е. тогда, когда эмпирический материал уже собран, и его невозможно больше исправить (но имеется достаточно возможностей испортить).

Чтобы этого не случилось, рекомендуется иметь в виду следующее:

I) Максимально сохранить имеющуюся информацию - обычно не надо вручную округлять данные, их классифицировать или перевычислить значения некоторых признаков.

2) Найденные ошибки измерений нельзя заменить "наверно правильными" результатами (или какими-нибудь средними). Их надо устранить, пробелы не мешают дальнейшей работе.

3) Нечисловые данные надо кодировать (лучше всего целыми числами, что дает впоследствии экономию в работе), сохраняя их упорядоченность. Все дихотомические (бинарные) признаки также целесообразно кодировать в числах, это расширяет возможности их обработки.

4) При кодировании пропусков для каждого признака надо точно выяснить, когда символ "0" обозначает соответствующее значение признака, а когда - недостаток информации.

5) Разумно сохранить каждый измеренный объект (даже в случае, когда по плану группы должны были равняться друг другу, но случайно одна группа оказалась больше, не стоит исключать объекты). Даже объекты, у которых много неизмеренных признаков, как правило, не вредят результатам обра-

ботки.

6) На этапе подготовки материала еще раз надо проверять, фиксированы ли все возможные факторы, влияющие на результаты исследования (например, время исследования, номер интервьюера, язык заполнения анкетов и т.д.).

7) Часто массив данных состоит из разных, более-менее самостоятельных частей. Тем не менее целесообразно их оформить как подмассивы одного большого массива с тем, чтобы иметь возможность их сравнивать при помощи статистических процедур.

2.2. Создание плана обработки данных

При создании плана обработки данных следует иметь в виду, что причиной неудачных результатов может быть как недостаточно обдуманный, так и чрезмерно детальный план обработки.

Хотя хороший исследователь всегда знает свои окончательные цели исследования, он должен быть готов корректировать свой план обработки по уже полученным результатам на любом этапе работы.

Следует сказать, что создание хорошего плана обработки по плечу только такому исследователю, который хорошо знает свой материал. Поэтому целесообразно, чтобы исследователь сам (а не его помощники) хотя бы часть материала готовил к машинной обработке (кодировал, создавал описание данных и т.д.).

В плане обработки, как правило, предусматриваются следующие этапы:

I) Работа всегда начинается с проверки материала. Для этого для каждого признака найдутся или описательные стати-

стики (в том числе max и min) или распределения (для качественных признаков). Изучая тщательно полученные результаты, можно обнаружить грубые ошибки (слишком большие, слишком маленькие или логически недопустимые значения). Если такие существуют, то придется их устранить (считать неправильные значения отсутствующими) и при необходимости повторить процедуру, пока исследователь не убедится, что грубых ошибок больше нет.

Кроме того, на этапе проверки материала необходимо устранить постоянные признаки, выяснить почти постоянные и другие малоинформативные признаки.

2) На предварительном этапе надо образовать новые признаки, правила которых известны заранее, исправить возможные ошибки кодирования, образовать из нескольких малоинформативных признаков новые, более информативные и т.д.

3) На этапе первичного анализа надо прежде всего проверять влияние мешающих и группирующих признаков, пользуясь для этого разными процедурами сравнения (t-тест, дисперсионный анализ, множественные сравнения). Мешающие и группирующие признаки, не имеющие существенного влияния, можно в дальнейшем устранить из списка обрабатываемых признаков. Здесь целесообразно определить и подмассивы - это группы объектов, которые похожи по группирующим признакам, но отличаются друг от друга по некоторым другим существенным признакам.

4) Следующим шагом является описание всех подмассивов по всем важным исследуемым признакам и (если это представляет интерес) проверка их различия.

На этом этапе используются описательные статистики, вы-

числение дву- и многомерных распределений, гистограмм и т.д.

Часто на этом этапе возможно пользоваться результатами предыдущих этапов. Имеются такие исследования, которые на этом этапе заканчиваются.

5) На дальнейшем этапе, т.н. вторичного анализа, надо прежде всего выяснить тип задачи.

а) Если основной задачей является описание исследуемого явления, выяснение структуры зависимостей между признаками, то мы имеем дело с описательной задачей.

б) Если основной целью исследования есть прогнозирование основных характеристик исследуемого явления, то наша задача прогностического типа.

Конечно возможны и задачи, в которых оба типа комбинированы.

6) Для задач описательного типа обычно характерно то, что измерено очень большое количество более-менее однотипных признаков, и проблемой является - как они связаны друг с другом, какие из этих признаков сильно влияют на другие, существуют ли какие-нибудь (неизмеренные) признаки, являющиеся общими причинами для измеренных признаков и т.д. Такие проблемы характерны для социологических исследований, и для решения этих задач целесообразно сделать корреляционный анализ (на основании линейных или некоторых непараметрических корреляций, пользуясь разными корреляционными графами), вычислить частные корреляции и сделать факторный анализ.

Именно факторный анализ, основывающийся на гипотезе существования небольшого количества неизмеренных причин-признаков - т.н. факторов - для большого количества измеренных признаков и есть ведущий метод решения описательных задач, со-

кращения числа и структурирования большого числа признаков.

7) При решении описательных задач иногда желательно структурировать и множество объектов, разбить их на классы, которые не известны заранее. Такие задачи решаются с помощью разных методов таксономии.

8) Для прогностических задач важной характерной чертой является существование т.н. функциональных (критериальных) признаков - эти признаки самые важные с точки зрения данного исследования, и требуют прогнозирования. Функциональными признаками являются в медицине - диагноз, в педагогике - успеваемость, в спортметодике - спортивный результат и т.д.

Кроме функциональных признаков в прогностических задачах всегда существует большое или малое количество аргументов. Это признаки, о которых предполагают, что они влияют на функциональный признак. Целью прогностических задач является выяснение зависимости функциональных признаков от аргументов и выражение этой зависимости в виде статистической модели.

9) Если как аргументы, так и функциональные признаки являются количественными, то методикой решения поставленной задачи является регрессионный анализ. В случае многих равноценных функциональных признаков можно применить и канонический анализ. Если число аргументов очень большое, то до начала регрессионного анализа целесообразно исследовать зависимости между функциональными и аргумент-признаками с помощью корреляций (четырехугольная матрица) или корреляционным отношений (для проверки линейности).

I0) Если аргументы не числовые, а функциональный признак числовой, то можно применить дисперсионный анализ. Дис-

персионный анализ применяется в том случае, если надо подробно (количественно) исследовать влияние отдельных аргументов и их всевозможных комбинаций. Для дисперсионного анализа существенным ограничением является то, что в этом случае аргументом может быть только один признак, в то время как для регрессионного анализа число аргументов может составлять несколько десятков и даже больше (но это, как правило, нецелесообразно).

II) Если аргументы числовые, а функциональный признак качественный (группирующий), то мы имеем дело с дискриминантным анализом. Эта методика типична для решения задач медицинской и технической диагностики, когда целью исследования является нахождение правил, распределяющих объекты в некоторые (известные заранее) группы.

I2) План обработки целесообразно сделать так, чтобы предусматривалось довольно тесное общение с ЭВМ, причем чтобы перед каждым новом заказом анализировались результаты предыдущего этапа и, при необходимости, корригировался заказ.

I3) Если задача решена, то исследователю полезно сохранить как исходные данные, так и результаты обработки с тем, чтобы было возможно в любое время продолжать обработку - как с целью сравнения с новыми исследованиями, или для решения некоторой новой задачи на основании этих-же исходных данных.

2.3. Представление данных в пакете САИСИ

Информация, представляемая в пакет обработки данных, состоит, как правило, из двух частей - из начальных данных (результаты измерения) и из описательной информации, харак-

теризующей эти начальные данные. Описательную информацию используют для управления обработкой данных, для их проверки, при оформлении распечаток и т.д.

При совершенствовании пакетов статистической обработки значимость описательной информации растет. Относительно примитивные пакеты требуют мало описательной информации или вообще работают без нее.

Стандартной формой для представления начальных данных в пакете САИСИ является объект - признак-матрица, т.е. таблица, где результаты наблюдения размещены следующим образом:

$$\begin{pmatrix} x_{11} & x_{12} \ldots & x_{1m} \\ x_{21} & x_{22} \ldots & x_{2m} \\ & \ldots\ldots & \\ x_{n1} & x_{n2} \ldots & x_{nm} \end{pmatrix}$$

В каждой строке этой таблицы находятся результаты измерения одного объекта. Так как значения признаков представлены всегда в одном и том же порядке, каждый столбец таблицы содержит все значения одного наблюдаемого признака. Значения признака могут быть как численными, так и буквенными. Объект-признак-матрица вводится по строкам.

Максимальное число используемых признаков в пакете САИСИ 500, но в специальном режиме можно ввести объект-признак-матрицу с 1000 столбцами. Число объектов, в принципе, не ограничено.

Как отмечалось при описании плана обработки, в большинстве случаев обработка проводится в несколько этапов. Анализируя результаты одного этапа, выдвигаются новые гипотезы, уточняется дальнейший план обработки. Для облегчения

такого подхода в пакете САИСИ разрешено использовать в качестве начальных данных и результаты предыдущего этапа - корреляционные матрицы, средние, стандартные отклонения.

Из начальных данных и описательной информации в пакете САИСИ образуют системный файл. Для образования системного файла необходима следующая информация:

I) имя файла,

2) список имен признаков в файле,

3) число объектов в файле,

4) носитель начальных данных (перфокарты, магнитная лента и т.д.),

5) исходный формат, который уточняет размещение начальных данных на носителе.

При вводе начальных данных можно фиксировать специальную структуру объект-признак-матрицы, соединяя последовательные объекты в группы, которые называют подфайлами. В дальнейшем подфайлов можно использовать по-разному - начиная с параллельной обработки всех подфайлов и кончая игнорированием структуры подфайлов.

При наличии подфайлов в ходе описания данных указываются не число объектов в файле, а имена подфайлов и числа объектов в подфайлах.

При описании данных можно еще добавить следующую (необязательную) информацию:

I) перечень обозначений пропусков (для каждого признака можно определить до трех различных обозначений для пропусков);

2) метки признаков (т.е. длинные названия признаков);

3) метки значений признаков (т.е. длинные названия зна-

чений);

4) форматы печати признаков.

Если системный файл нужно использовать на нескольких этапах обработки, его следует записать. На каждом этапе обработки можно изменить содержание системного файла - добавить новые признаки или объекты, а также исключить некоторые использованные признаки или объекты. Можно изменить и описательную информацию.

Перед статистической обработкой признаками можно сделать разные вычисления, создать таким образом новые признаки, также можно перекодировать признаки или выбрать для обработки только часть объектов. Измененный системный файл в конце обработки можно снова записать.

III. УПРАВЛЯЮЩИЙ ЯЗЫК ПАКЕТА САИСИ

3.I. Общая характеристика управляющего языка

Как для статистических пакетов вообще, работой пакета САИСИ управляют при помощи управляющих карт. Для выполнения некоторой последовательности процедур нужно приготовить соответствующую последовательность управляющих карт. Управляющая программа вводит карту, дешифрирует ее и организует выполнение процедуры, указанной на карте.

Для того, чтобы управляющие карты можно было узнавать и дешифрировать, они должны быть оформлены по специальным правилам. Так возникает специальный язык - язык управления пакетом.

Для изучения этого языка нужно познакомиться с общими правилами оформления управляющих карт, с описаниями конкретных управляющих карт, допустимых в этом языке, и с правилами упорядочения управляющих карт.

3.2. Общие правила оформления управляющих карт

Каждая управляющая карта пакета САИСИ умышленно разделена на две части - управляющее поле и поле описаний.

Управляющее поле занимает колонки от I до I5. Оно содержит управляющее слово (или слова), которое идентифицирует нужную процедуру. Система использует только первые 8 символов управляющего поля.

Поле описании занимает колонки от I6 до 80 и при необходимости может быть продолжено в тех же колонках следующих карт. Оно содержит инструкции, нужные для выполнения процедуры, названной в управляющем поле.

Приведем некоторые примеры управляющих карт:

```
COMPUTE         ОДОХОД = ДОХОД/ЧИССЕМ
RECODE          ВОЗРАСТ (LOWEST THRU 16=1),
                (17 THRU 2Ø=2), (2Ø THRU HIGHEST = 3)
CODEBOOK        ALL
```

Управляющее поле Поле описаний

В поле описаний можно записать семь различных элементов. В следующих пунктах приведено описание всех элементов.

3.2.I. Имя

Каждый признак, каждый подфайл должен иметь имя, можно присваивать имя и системному файлу. Имя - это последовательность букв и цифр, которая начинается буквой. Длина имени ограничена - оно не должно содержать более 8 символов. Например:

ПОЛ, МЕСТЖИТ, ИНДЕКС25, X1, T1Y2

Для того, чтобы облегчить анализ распечаток, желательно выбрать такое имя, чтобы оно напоминало и содержательное значение признака.

3.2.2. Значение

При вычислении или перекодировании признаков очень часто нужно использовать некоторые значения. В пакете САИСИ значения могут быть либо численными, либо нечисленными.

Численными значениями являются целые и десятичные числа. Например:

2, Ø.2, .Ø5, -I4

Целая часть отделяется от дробной части точкой. Чтобы не перепутать нуль с буквой "O", его обозначают через "Ø".

Нечисленным значением может быть любая последовательность вводимых символов. Записанное на управляющую карту не-

численное значение должно всегда находиться между апострофами. Например:

'СЛУЖ', 'М', '+2'

3.2.3. Ключевое слово

Ключевое слово - это некоторое английское слово, или его сокращение. Например:

THRU, WITH, TO, BY

С каждой конкретной управляющей картой связаны конкретные ключевые слова. Поэтому ключевые слова описываются вместе с конкретными управляющими картами, на которых они встречаются.

3.2.4. Метка

Для более наглядного оформления распечаток признакам, значениям признака или системным файлам можно присваивать метку. Метка - это просто "длинное имя". Формально метка - произвольная последовательность вводимых символов. Например:

ОКОНЧЕННОЕ УЧЕБНОЕ ЗАВЕДЕНИЕ, СРЕДНЕЕ ОБРАЗОВАНИЕ, ОЦЕНКИ I ЧЕТВЕРТИ 6 и 7 КЛАССОВ.

Длина метки для разных управляющих карт разная; она определяется при описании конкретных управляющих карт.

3.2.5. Разделитель

В поле описаний имена, ключевые слова и метки должны быть отделены от других элементов разделителями, которыми являются пробел и запятая. Эти символы равноценны и взаимно заменяемы. Чтобы исправить читаемость управляющей карты, можно записать несколько таких разделителей подряд.

На конкретных управляющих картах в качестве разделителей надо использовать круглые скобки (), наклонную черту /

и знак равенства = . Но использовать эти разделители можно только в конкретных конструкциях, которые будут описаны вместе с соответствующими управляющими картами.

3.2.6. Арифметическое выражение

Арифметическое выражение образуется из соединенных знаками арифметических операций имен признаков, численных значений и стандартных функций. Допустимые операции:

1) возведение в степень (* *);

2) умножение (*), деление (/);

3) сложение (+), вычитание (-).

Порядок операций соответствует порядку в приведенном списке. Для изменения порядка операций используют круглые скобки. Ни одного знака арифметической операции нельзя пропускать.

Перечень возможных стандартных функций следующий:

Обозначение	Значение	Пример
SQRT	квадратный корень	SQRT(ПРИЗН)
LN	натуральный логарифм	LN(СКОР)
LG10	десятичный логарифм	LG10(B)
EXP	экспонента	EXP(A+B)
SIN	синус	SIN(Z)
COS	косинус	COS(T1)
ATAN	арктангенс	ATAN(F13)
RND	округление	RND(X1)
ABS	абсолютное значение	ABS(T1-T2)
TRUNC	отделение целой части	TRUNC(K/2)
MOD10	частное при делении на 10	MOD10(IND)

При вычислении арифметических выражений значения стандартных функций вычисляются первыми.

Примеры арифметических выражений:

```
X1
25
X ** 3 / BEETA
ALFA   +  EXP (SQRT (ABS (X1 + X2)))
0.5 * (ПР1/ПР2)
```

3.2.7. Логическое выражение

При обработке данных способ использования конкретного объекта может зависеть от некоторого логического условия. Такие условия представляются припомощи логического выражения.

Самым простым логическим выражением является сравнение. Сравнение - это пара арифметических выражений, соединенных знаком сравнения или пара из имени признака и нечисленного значения, соединенных знаком равенства. Возможные знаки сравнения приведены в следующей таблице:

Обозначение в САИСИ	Обычное обозначение
GT	$>$
GE	$\geqslant$
LT	$<$
LE	$\leq$
EQ	$=$
NE	$\neq$

Из сравнений при помощи логических операций AND и OR

можно образовать более сложные логические выражения.

Приведем несколько примеров логических выражений:

(ВОЗРАСТ LT 18)

((МЕ LT 5) AND (ВОЗРАСТ GT 50))

(SUBFIL EQ 'KL6')

Логическое выражение заключается всегда в скобки.

3.3. Классификация управляющих карт

По действиям, вызываемым управляющими картами, их можно разделить на следующие группы:

1) карты описания данных,

2) карты преобразования данных,

3) карты описания этапа обработки и задачи,

4) карты редактирования данных,

5) карты, организующие контрольную печать.

3.4. Обозначения, используемые для управляющих карт

Введем некоторые специальные обозначения для короткого и четкого описания управляющих карт.

3.4.1. Необязательный элемент

Некоторые элементы в определенных конструкциях являются необязательными. Такие элементы в дальнейших описаниях будут находиться в квадратных скобках. Например описание

FILE NAME имя [,метка]

разрешает использовать эту управляющую карту в виде, где в поле описаний написано только имя:

FILE NAME ОЦЕНКИ1

а также в виде, где в поле описаний имени следует метка

FILE NAME ОЦЕНКИ1, ОЦЕНКИ 1-ОЙ ЧЕТВЕРТИ

3.4.2. Выбираемый элемент

Иногда для определенных конструкций в некоторых местах нужно выбрать один из нескольких возможных элементов. В последующем такие выбираемые элементы будут находиться в фигурных скобках и разделены вертикальной чертой. Например, описание

INPUT MEDIUM {CARD | TAPE | DISK | OTHER}

определяет, что в поле описаний этой карты нужно написать одно из перечисленных слов, т.е.

INPUT MEDIUM CARD

или

INPUT MEDIUM DISK

и т.д.

3.4.3. Список признаков

В последующих описаниях список признаков коротко обозначим через **varlist**. Список признаков может состоять:

1) из одного имени признака;

2) из ТО-соглашения;

3) из разделенных пробелами или запятыми ТО-соглашений и имен признаков.

Конструкция ТО-соглашение дает удобную возможность перечислять признаки, которые в объект-признак-матрице находятся один за другим. Например, если признаки ВОЗРАСТ, ПОЛ, ДОХОД, ДЕЯТ, ОБРАЗ находятся в матрице начальных данных в последовательных столбцах, то можно записать список признаков следующим образом:

ВОЗРАСТ ТО ОБРАЗ

Для пакета этот список равносилен списку

ВОЗРАСТ, ПОЛ, ДОХОД, ДЕЯТ, ОБРАЗ.

Ключевое слово ТО должно обязательно быть отделено разделителями.

IУ. ОПИСАНИЯ УПРАВЛЯЮЩИХ КАРТ

4.I. Карты описания данных

Карты описания данных передают системе информацию, характеризующую начальные данные.

4.I.I. Карта FILE NAME

Пакет САИСИ может обрабатывать начальные данные только тогда, когда из данных образован системный файл. Поэтому сразу при вводе данных организуется такой системный файл.

Имя системного файла определяется картой, общий вид которой следующий:

FILE NAME имя, метка

Например:

FILE NAME ТЕСТ, ОЦЕНКИ ТЕСТОВ

FILE NAME СТУД

Длина метки не должна быть больше 64 символов.

4.I.2. Карта VARIABLE LIST

Каждому используемому признаку нужно присвоить имя.Имена признаков определяются картой, общий вид которой следующий:

VARIABLE LIST имя, имя, ..., имя

Порядок имен признаков должен соответствовать порядку признаков в матрице начальных данных. Например:

VARIABLE LIST ВОЗРАСТ, ПОЛ, ДОХОД, ДЕЯТ, ОБРАЗ

Если пользователь считает возможным различать некоторые признаки только по порядковому номеру (например, X1, X2, X3, X4, X5), то на карте VARIABLE LIST можно написать TO-соглашение (X1 TO X5).

4.I.3. Карта SUBFILE LIST

Для определения структуры подфайлов нужно использовать пару управляющих карт, общий вид которых следующий:

SUBFILE LIST имя, имя, ..., имя

OF CASES n1, n2. ..., nk

Между символом # и ключевым словом OF находится пробел. Первая из этих карт определяет имена подфайлов, вторая - числа объектов, принадлежащих подфайлам. Так, например, карты:

SUBFILE LIST КЛ9, КЛ1Ø, КЛ11

OF CASES 2Ø, 23, 29

определяют 3 подфайла. Первому подфайлу, который носит имя КЛ9, принадлежит 20 первых объектов, второму с именем КЛ1Ø - 23 следующих объекта, а третьему с именем КЛ11 - 29 последних объекта.

Для определения подфайлов автоматически образуется новый признак SUBFIL, значениями котрого являются имена подфайлов, точнее - первые четыре символа от имени подфайла . В предыдущей обработке можно использовать этот признак.

4.I.4. Карта INPUT MEDIUM

При введении начальных данных нужно сообщить пакету, на каком носителе находятся начальные данные. Это можно сделать картой, общий вид которой следующий:

INPUT MEDIUM {CARD | TAPE | DISK | OTHER}

Ключевое слово в поле описаний нужно выбрать по следующим правилам:

I) CARD - если начальные данные находятся на перфокартах;

2) TAPE - если начальные данные находятся на магнитной ленте;

3) DISK - если начальные данные находятся на магнитном диске;

4) OTHER - если тип носителя начальных данных не описан выше.

4.I.5. Карта # OF CASES

Если подфайлы не используются, нужно все-таки сообщить пакету число объектов. Это делается картой, общий вид которой следующий (между # и OF находится пробел):

OF CASES целое число

Например:

OF CASES 150

4.I.6. Карта INPUT FORMAT

При введении начальных данных нужно сообщить пакету их точное размещение на носителе (например, размещение по колонкам перфокарты). Это можно сделать при помощи управляющей карты, общий вид которой следующий:

INPUT FORMAT FIXED(список форматов)

В скобках приводятся форматы для всех признаков. Форматом указывается тип признака, а также количество позиций на носителе данных, предназначенное для данного признака.

Тип признака определяется буквой, которой начинается формат. Численному признаку соответствует буква F, нечисленному признаку - буква A.

За буквой A следует целое число, указывающее, на скольких позициях (например, в колонках перфокарты) расположены значения этого признака. Например, если значения нечисленного признака занимают 4 колонки на перфокарте, то ему соответствует формат A4.

За буквой F следуют два целых числа, разделенные точ-

кой. Первое из них указывает, на скольких позициях расположены значения этого признака, второе - сколько из этих позиций предназначены для десятичных чисел. Например, если значения признака занимают 3 колонки на перфокарте и при этом последняя колонка предназначена для десятичных чисел, то формат этого признака F3.1.

Для того, чтобы начальные данные было легче читать, между значениями признаков можно ставить пробелы. Пробелы нужно также определить в формате - формат пробела начинается целым числом, за которым следует буква X. Например, 3X определяет 3 пробела.

Вместо выписывания нескольких одинаковых форматов подряд можно перед буквой, начинающей формат, написать число повторений. Например, вместо A2, A2, A2 можно написать 3A2.

Примеры:

```
INPUT FORMAT  FIXED (2X, 2F4.2, X, A2)
INPUT FORMAT  FIXED (F3.Ø, 2F2.Ø, 3A4)
```

4.I.7. Карта PRINT FORMATS

Для сообщения пакету, как печатать значения признаков, используют карту, общий вид которой следующий:

```
PRINT FORMATS        varlist ({A | целое число } )/[.../
                     varlist ({A | целое число } )/]
```

В скобках, после списка признаков, следует буква A, если признаки нечисленные, или, в случае численных признаков - целое число, которое указывает число выпечатываемых мест после запятой.

Например:

```
PRINT FORMATS  ВОЗРАСТ, ДОХОД (4)/ ДЕЯТ (A)/
```

Определение формата печати обязательно при нечисленных при-

знаках. Формат печати численных признаков можно не определять - в этом случае он определяется автоматически.

4.I.8. Карта VALUE LABELS

При помощи этой карты значениям признаков можно определить метки. Общий вид карты следующий:

```
VALUE LABELS   varlist   (значение) метка ,...,
               (значение) метка/[... /
               varlist   (значение) метка ,...,
               (значение) метка/ ]
```

В одном списке перечисляют признаки, которые имеют одинаковые значения. Длина метки не должна быть более 20 символов.

Примеры:

```
VALUE LABELS   ПОЛ (1) МУЖСКОЙ, (2) ЖЕНСКИЙ/
               ДЕЯТ ('С') СЛУЖАЩИЙ, ('Р') РАБОЧИЙ,
               ('К') РАБОТНИК СЕЛЬСКОГО ХОЗЯЙСТВА /
```

4.I.9. Карта VAR LABELS

При определении меток признаков используют карту, общий вид которой следующий:

```
VAR LABELS     имя, метка /[../ имя, метка/ ]
```

Метка признака не должна содержать более 40 символов.

Примеры:

```
VAR LABELS     ДЕЯТ, ВИД ПРОФЕССИОНАЛЬНОЙ ДЕЯТЕЛЬНОСТИ ОТВЕ-
               ЧАЮЩЕГО/ ОБРАЗ, ОБРАЗОВАНИЕ ГЛАВЫ СЕМЬИ/
```

4.I.IO. Карта MISSING VALUES

При помощи этой карты пакету сообщают обозначения пропусков (т.е. обозначения пропущенных значений). Пропускам, возникшим по разным причинам, можно придавать разные обозначения. Но число таких обозначений не должно быть больше трех.

Общий вид карты следующий:

MISSING VALUES varlist (обозначения пропусков) /[../
varlist (обозначения пропусков)/]

Например:

MISSING VALUES ВОЗРАСТ, ПОЛ (Ø)/ ДОХОД (Ø,1)/

4.2. Карты преобразования данных

В ходе обработки может оказаться необходимым преобразовать каким-то образом имеющиеся признаки. Также может оказаться желательным изучать отдельно некоторые объекты или для уточнения плана обработки образовать из существующей большой выборки маленькую подвыборку. Иногда возникает потребность присвоить объектам некоторые веса. Такие заказы представляются системе при помощи карт преобразования данных.

4.2.1. Карта RECODE

Для перекодирования признаков предназначена управляющая карта, общий вид которой следующий:

RECODE varlist (список значений = значение), ...,
(список значений = значение) / [.../
varlist (список значений = значение) ,...,
(список значений = значение)/]

Например:

RECODE ДЕЯТ ('С' = 1), ('Р' = 2), ('К' = 3)/
ОБРАЗ (2, 3 = 1),(4 THRU 8 = 2),
(ELSE = 3)/

При выполнении такой карты для всех признаков, перечисленных в списке, значения, приведенные слева от знака равенства заменяются значениями, приведенными справа от знака равенства.

Для оформления списка значений, которые находятся слева от знака равенства, есть три возможностей:

I) список значений состоит из одного значения или из нескольких значений, отделеных запятой. Например (Ø = 3) или ('A', 'B' = 1);

2) вместо перечисления значении, находящиеся в некотором интервале, указывают крайние точки этого интервала.

Например:

(1 THRU 1Ø = Ø) или (32.5 THRU 38.4 = 4)

Если интервал должен содержать все значения, меньше некоторой константы, то можно неизвестное минимальное значение заменить ключевым словом LOWEST. Например:

(LOWEST THRU 1Ø = Ø)

Если интервал должен содержать все значения, большие некоторой константы, можно неизвестное максимальное значение заменить ключевым словом HIGHEST. Например:

(4Ø.5 THRU HIGHEST = 5)

Напомним, что ключевое слово THRU должно быть отделено разделителями (пробелами или запятыми);

3) если после перекодирования некоторых значений всем остальным определяют одинаковые коды, то можно вместо их перечисления использовать ключевое слово ELSE. Например:

(LOWEST THRU 18 = 1), (18 THRU HIGHEST = 2), (ELSE = 3)

Для перекодирования пробела слева от знака равенства пишут не пробел, а ключевое слово BLANK. Например:

(BLANK = 9)

4.2.2. Карта COMPUTE

Эта управляющая карта дает возможность вычислять новые признаки, используя уже имеющиеся признаки. Общий вид карты следующий:

COMPUTE имя = арифметическое выражение

Например:

COMPUTE МЕСТ = СТАЖ/ ВОЗРАСТ

При выполнении этой карты для каждого объекта имена признаков в арифметическом выражении заменяются их численными значениями, затем вычисляется соответствующее арифметическое выражение и оно присваивается данному объекту в качестве нового признака.

4.2.3. Карта IF

Предписание вычисления нового признака может зависеть от некоторого логического условия. Такое предписание представляется управляющей картой, общий вид которой следующий:

IF (логическое выражение) имя =
арифметическое выражение

При выполнении этой карты вычисляется значение арифметического выражения только для тех объектов, при которых логическое выражение является истинным. Например:

IF (ВОЗРАСТ LT 18) ИНД = ВОЗРАСТ - 10

Здесь признак ИНД вычисляется только для тех объектов, у которых признак ВОЗРАСТ меньше, чем 18. У остальных объектов значения признака остаются прежними, если признак перевычисляется, или считается нулем, если вычисляется новый признак.

4.2.4. Карта COUNT

Эту карту используют для генерирования индекса перечисления. Предположим, что при выполнении некоторой анкеты был представлен некоторый список газет, например, "Правда", "Известия", "Спорт", "Литературная газета", "Неделя". Для каждой газеты нужно было ответить, читается ли она всегда (1), иногда (2) или не читается (3). (В скобках приведены коды соответствующих вариантов ответа). Предположим, что в ходе

обработки оказалось нужным определить для каждого отвечающего общее число читаемых им газет. Для определения этого числа нужно перечислить число единиц в ответах, приведенных для списка газет. Признак, полученный путем такого перечисления и называется индексом перечисления.

Вычисление индекса перечисления можно заказать картой, общий вид которой следующий:

```
COUNT     имя = varlist (список значений),...,
          varlist (список значений) /[.../
          имя = varlist (список значений),...,
          varlist (список значений)/]
```

Слева от знака равенства находится имя вычисляемого признака, справа от знака равенства находится список признаков, значения которых перечисляются. В скобках указываются перечисляемые значения. В приведенном примере с газетами соответствующая управляющая карта имеет вид:

```
COUNT     ИНД = ПРАВДА, ИЗВЕСТИЯ, СПОРТ, ЛИТГАЗ, НЕДЕЛЯ
          (1, 2)/
```

Перечисляемые значения у разных признаков могут быть разные. Например:

```
COUNT     ИНД = ПРАВДА, ИЗВЕСТИЯ (1),
          СПОРТ, ЛИТГАЗ, НЕДЕЛЯ (1, 2)/
```

4.2.5. Карта ASSIGN MISSING

Если при вычислении новых признаков у объекта отсутствует некоторое использованное в арифметическом или логическом выражении значение, то все нужные операции проделают с обозначениями пропусков. Специальное обозначение для отсутствующего значения нового признака можно определить картой, общий вид которой следующий:

ASSIGN MISSING varlist (обозначение пропуска)/[.../
varlist (обозначение пропуска)/]

Например:

ASSIGN MISSING ИНД, ИНДЕКС (99)/

Если отсутствует хоть одно значение, использованное в арифметическом или логическом выражении, то значением нового признака является теперь обозначение пропуска.

4.2.6. Карты DO и DOEND

Встречаются задачи, где приходится вычислять много новых признаков по схожим предписаниям. Например, предположим, что в данных медицинского характера зарегистрированы 20 физиологических показаний в начале лечения (признаки HA1 по HA20) и в конце лечения (признаки KO1 по KO20). Требуется вычислить для каждого признака разность между начальным и конечным показателями. Для заказа таких вычислений можно написать 20 одинаковых управляющих карт, но более рациональный пользователь ограничится тремя картами:

```
DO          Y = HA1 TO HA2Ø / Z = KO1 TO KO2Ø/
            W = P1 TO P2Ø/
COMPUTE     W = Y - Z
DOEND
```

Как видно, карта, требующая многократных повторений, размещается между картами DO и DOEND, а изменяющиеся имена признаков заменяются новыми фиктивными именами (в примере именами W, Y, Z). В поле описаний карты DO перечисляют для каждого фиктивного имени после знака равенства ее истинные значения, которые нужно использовать при повторении. Общий вид карты DO следующий:

DO имя = varlist/[.../ имя = varlist/]

Между картами DO и DOEND могут находиться кроме карт 4.2.I - 4.2.5 еще карты MISSING VALUES и SELECT IF. Различных повторяемых карт может быть больше, чем одна.

4.2.7. Карта SELECT IF

При помощи этой карты можно из имеющейся объект-признак-матрицы образовать новую, которой принадлежат только те объекты, для которых выполнено некоторое логическое условие. Общий вид карты следующий:

SELECT IF (логическое выражение)

Например:

SELECT IF (ВОЗРАСТ LT 18)

4.2.8. Карта SAMPLE

При помощи этой карты можно из данной объект-признак-матрицы образовать случайную выборку. Общий вид карты следующий:

SAMPLE число α

Например:

SAMPLE Ø.Ø1

Число α, написанное в поле описаний этой карты, должно находиться между нулем и единицей ($0 < \alpha < 1$). Это число определяет вероятность, с которой объект попадает в образуемую выборку. Число объектов в выборке невозможно заранее точно определить. Это число приблизительно будет равно αn, где n - число объектов в первоначальной матрице.

4.2.9. Карта WEIGHT

При помощи этой карты каждому объекту можно присвоить вес и далее во всех последующих вычислениях используются взвешенные значения признаков. Общий вид карты следующий:

WEIGHT имя

Например:

WEIGHT ИНДЕКС

В поле описаний карты находится имя признака, значениями которого являются веса объектов. Веса должны быть заранее определены - их можно ввести вместе с остальными начальными данными или вычислить при преобразовании данных.

4.3. Карты описания задачи и прогона

В соответствии с планом обработки использование пакета статистических программ происходит поэтапно. Каждый этап обработки будем назвать прогоном. На каждом прогоне можно обращаться к нескольким статистическим процедурам. Такое обращение в дальнейшем будем называть задачей. Таким образом, в одном прогоне может быть несколько задач.

Предусмотрен целый ряд управляющих карт, которые регулируют работу пакета в течение всего прогона. Также имеются некоторые общие правила при обращении к статистическим процедурам и карты, которые уточняют, как использовать начальные данные. Специальные карты предусмотрены для введения и записи начальных данных.

4.3.1. Карта RUN NAME

При помощи этой управляющей карты можно определить для каждого прогона "длинное имя". Это "длинное имя" будет указано на всех страницах распечатки, выдываемых на данном прогоне. Общий вид карты следующий:

RUN NAME метка

Например:

RUN NAME ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ

Максимальная допустимая длина метки - 64 символов.

4.3.2. Карта TASK NAME

При помощи этой карты можно для каждой задачи определить "длинное имя", которое вместе с "длинным именем" прогона будет указано на каждой странице распечатки. Общий вид карты следующий:

TASK NAME метка

Например:

TASK NAME ОДНОМЕРНЫЕ ТАБЛИЦЫ ЧАСТОТ

Максимальная допустимая длина метки - 64 символов

4.3.3. Карта COMMENT

Эта карта дает возможность на произвольном месте в заказе представить объясняющие комментарии. Общий вид карты следующий:

COMMENT метка

Например:

COMMENT В СЛЕДУЮЩЕЙ ОБРАБОТКЕ ИСПОЛЬЗУЮТСЯ ДАННЫЕ ГОРОЖАН

Длина использованной метки не ограничена.

4.3.4. Карта FINISH

Эта карта сообщает системе, что данный прогон окончен и является последней картой в колоде управляющих карт. Общий вид карты следующий:

FINISH

4.3.5. Карта NUMBERED

При помощи этой карты можно резервировать колонки от 73 по 80 управляющих карт для их порядковых номеров. Общий вид карты следующий:

NUMBERED YES

Нумерация управляющих карт будет полезной при их большом количестве.

4.3.6. Карты обращения к статистическим процедурам

При обращении к статистической процедуре можно использовать до трех управляющих карт.

Первая из этих карт определяет имя требуемой статистической процедуры, имена используемых признаков и способ их использования. В пакете имеется 21 различных статистических процедур, которые будут подробно описаны во втором разделе данного методического руководства.

Вторая из этих карт идентифицируется управляющим словом OPTIONS - варианты. В описательном поле карты находятся целые числа, которые вместе определяют режим работы используемой статистической процедуры (возможные варианты для различных процедур будут разными и поэтому будут описаны вместе с статистическими процедурами).

Карта OPTIONS не обязательна. У каждой статистической процедуры имеется "в среднем" оптимальный режим умолчания, который будет использован при отсутствии этой карты.

Третья из этих карт идентифицируется с управляющим словом STATISTICS. Она дает возможность регулировать объем выдаваемой процедурой информации. В поле описаний карты находится ключевое слово ALL, если требуется максимальное количество выдаваемой информации, или целые числа, которые определяют выдаваемые статистики или графики. Возможная выдаваемая информация будет описана вместе со статистическими процедурами.

4.3.7. Карта PROCESS SBFILES

При помощи этой карты можно определить, какие подфайлы или группы подфайлов в конкретной задаче нужно использовать. Общий вид карты следующий:

PROCESS SHFILES {EACH | ALL | список групп подфайлов}

При выписывании поля описаний нужно учитывать следующее:

1) если дальнейшую обработку нужно провести отдельно для каждого подфайла, в поле описаний записывают ключевое слово EACH;

2) если игнорируется структура подфайлов, в поле описаний карты записывают ключевое слово ALL;

3) если в обработке нужно использовать только некоторые подфайлы или группы подфайлов, соответствующие имена или списки имен заключаются в круглые скобки и из таких групп образуют список параллельно обрабатываемых групп.

Предположим, например, что объект-признак-матрица разделена на пять подфайлов с именами A1, A2, A3, A4 и A5. Пусть нужно провести одинаковую обработку параллельно для подфайла A3 и объединенных подфайлов A2 и A5. Такой заказ представляет карта:

PROCESS SHFILES (A3), (A2, A5)

Использование этой карты не является обязательным. Отсутствие карты вызывает игнорирование структуры подфайлов.

4.3.8. Временные версии карт преобразования данных

Могут встретиться ситуации, где какое-то преобразование начальных данных нужно использовать только в одной задаче. В этом случае перед соответствующей управляющей картой нужно записать символ "звездочка" * . Данные, преобразованные такой картой, используются только в задаче, непосредственно следующей за этой картой. Сразу после выполнения задачи восстанавливается первоначальный вид начальных данных.

Временные версии в пакете САИСИ допустимы при управля-

ющих картах:

* RECODE = SELECT IF
* COMPUTE = SAMPLE
* COUNT = WEIGHT
* IF

4.3.9. Карта READ INPUT DATA

Эта управляющая карта дает пакету распоряжение ввести начальные данные. Общий вид карты следующий:

READ INPUT DATA

Карта должна находиться непосредственно после первого обращения к статистической процедуре. Если начальные данные вводятся с перфокарт, они должны следовать сразу за картой READ INPUT DATA

4.3.10. Карты SAVE FILE и GET FILE

Для того, чтобы использовать уже образованный системный файл в следующих прогонах, его нужно записать на магнитный диск (или ленту). Для записи файла используют управляющую карту, общий вид которой следующий:

SAVE FILE

Эта карта должна непосредственно предшествовать карте FINISH.

Для вызова записанного системного файла, используют карту, общий вид которой следующий:

GET FILE имя

В поле описаний карты находится имя системного файла, которое было определено картой FILE NAME. Например:

GET FILE ДАННЫЕ

4.4. Карты редактирования данных

Начальные данные, записанные в конце одного прогона, можно дополнять и изменять в ходе следующих прогонов. Воз-

можно сделать следующие изменения:

I) выбросить признаки из системного файла;

2) добавить новые признаки;

3) переупорядочить признаки;

4) создать новую структуру подфайлов;

5) выбросить подфайлы;

6) добавить новые подфайлы;

7) переупорядочить объекты.

4.4.I. Карта DELETE VARS

Признаки, которые с некоторого прогона станут ненужными, можно выбросить из системного файла, а полученный файл снова записать. Общий вид карты выбрасывания признаков следующий:

DELETE VARS varlist

Например:

DELETE VARS ИНД1, ИНД5 ТО СТАТС

При выполнении этой карты из системного файла выбрасываются признаки, перечисленные в поле описаний карты. Карте DELETE VARS должна непосредственно следовать карта SAVE FILE или некоторая допустимая карта редактирования данных, в противном случае признаки не выбрасываются.

4.4.2. Карта KEEP VARS

При помощи этой карты, аналогично предыдущей, можно выбросить ненужные признаки из системного файла. Общий вид карты следующий:

KEEP VARS varlist

Например:

KEEP VARS ВОЗРАСТ, ОБРАЗ, ПОЛ

При выполнении карты из системного файла выбрасываются

признаки, которые не перечислены в поле описаний.

Карте KEEP VARS должна непосредственно следовать карта SAVE FILE или некоторая допустимая карта редактирования, иначе признаки не выбрасываются.

4.4.3. Карта ADD VARIABLES

На некотором этапе обработки может оказаться нужным добавить к существующей объект-признак-матрице новые признаки, измеренные у тех же самых объектов. Это можно сделать картой, общий вид которой следующий:

ADD VARIABLES имя, имя, ..., имя

Например:

ADD VARIABLES ВРЕМЯ,X1 TO X6

Значения новых признаков вводятся по объектам. Порядок значений признаков должен соответствовать порядку имен в управляющем поле карты.

4.4.4. Карта REORDER VARS

Признаки в системном файле можно переупорядочить. Это можно сделать картой, общий вид которой следующий:

REORDER VARS varlist

Например:

REORDER VARS Z, ALFA, X1, X7, X2 TO X6.

Признаки будут переупорядочены соответственно списку, приведенному в поле описаний карты.

За картой REORDER VARS должна непосредственно следовать карта SAVE FILE или некоторая допустимая карта редактирования, иначе переупорядочение не производится.

4.4.5. Карта NEW SUBFILE

При необходимости для данного системного файла можно определить новую структуру подфайлов. Это делается управляю-

щей картой, общий вид которой следующий:

NEW SUBFILE имя (n1)[, имя (n2), ..., имя (nK)]

В скобках после имени указывается число объектов в новом подфайле. Например:

NEW SUBFILE ПФ1 (30), ПФ2 (53)

Нужно помнить, что в один подфайл можно соединить только последовательные объекты. Приведенная в примере карта делит системный файл на два подфайла. В первый подфайл с именем ПФ1 принадлежит 30 первых объектов, во второй, с именем ПФ2, принадлежит 53 следующих объектов.

Новое разбиение должно обязательно хватывать все объекты системного файла.

4.4.6. Карта DELETE SUBFILES

Подфайлы, которые с некоторого этапа обработки окажутся ненужными, можно выбросить из системного файла. Общий вид соответствующей управляющей карты следующий:

DELETE SUBFILES имя[,имя, ..., имя]

Например:

DELETE SUBFILES КЛ7

При выполнении карты из системного файла выбрасывают подфайлы, имена которых указаны в поле описаний карты.

4.4.7. Карта ADD SUBFILES

На некотором этапе обработки может оказываться нужным добавить к существующей объект-признак-матрице новые объекты. Это можно сделать при помощи двух карт, общий вид которых следующий:

ADD SUBFILES имя [,имя, ..., имя]

OF CASES 20, 49

Первая из этих карт определяет имена введенных подфайлов, а вторая – числа объектов в каждом подфайле. Порядок признаков в новых подфайлах должен соответствовать порядку признаков в системном файле. Например:

```
ADD SUBFILES    АБИТУР, СТУД
# OF CASES      2Ø, 49
```

4.4.8. Карта SORT CASES

При необходимости объекты системного файла можно переупорядочить по значениям некоторого признака (или некоторых признаков). Общий вид соответствующей карты следующий:

SORT CASES varlist[({A|D})][..., varlist ({A|D})]

Например:

```
SORT CASES      ИНД1 (A), ИНД2 (D)
```

В поле описаний карты нельзя указать более шести признаков. За списком признаков следует буква A, если порядок их значений должен быть возрастающим, или буква D, если порядок их значений должен быть убывающим. При отсутствии буквы образуется возрастающий порядок значений.

4.5. Карты организования контрольной печати

Для проверки описательной информации или самих начальных данных в системе САИСИ имеется две карты.

4.5.I. Карта DUMP

При помощи этой карты можно заказать печать описательной информации системного файла. Общий вид карты следующий:

DUMP список ключевых слов

Ключевые слова, которые можно использовать в поле описаний карты вместе с описанием соответствующего действия, описаны в следующей таблице.

Ключевое слово	Выдаваемая информация
VARLIST	Печатаются имена признаков подфайла в порядке их размещения
SORT VARS	Печатаются имена признаков подфайла в алфавитном порядке
VARINFO	Представляются формат представления и формат печати каждого признака
LABELS	Печатаются метки всех признаков и всех значений
SUBDIRECTORY	Печатается список подфайлов вместе с числом объектов в подфайле
COMPLETE	Для каждого признака из системного файла печатается вся описательная информация
TRANSFORM	Печатаются все предписания вычислений
RECODES	Печатаются все предписания перекодирования, а также список перекодированных признаков

В поле описаний допустима произвольная комбинация перечисленных ключевых слов.

Например:

```
DUMP          LABELS, SUBDIRECTORY
```

4.5.2. Карта LIST CASE

При помощи этой карты можно печатать значения признаков системного файла. Общий вид карты следующий:

LIST CASE CASES= n/[VARIABLES = {varlist | ALL}]

За ключевым словом CASES, после знака равенства записывается число печатаемых объектов. Ключевому слову VARIABLES, после знака равенства, следует список имен печатаемых признаков. Если желательно печатать значения всех признаков, в поле описаний карты записывается ключевое слово ALL.

Пусть, например, требуется для 10 первых объектов печатать значения признаков ВОЗРАСТ, ИНД1 , ИНД2, ..., ИНД10 Такой заказ представляется картой:

```
LIST CASE          CASES= 10/VARIABLES=
                   ВОЗРАСТ, ИНД1 ТО ИНД10 /
```

Нужно помнить, что эта карта работает только вместе с обращением к какой-либо статистической процедуре, она должна быть расположена непосредственно перед картой обращения к процедуре.

V. ПРЕДСТАВЛЕНИЕ ЗАКАЗА И ПОРЯДОК УПРАВЛЯЮЩИХ КАРТ

В последующем изложении будем называть заказом последовательность (колоду) управляющих карт, которая определяет действия пакета САИСИ в течение одного прогона. В ходе прогона можно обращаться к нескольким статистическим процедурам, каждое такое обращение называем задачей.

5.I. Прогон, в ходе которого образуется системный файл

5.I.I. Начало прогона

Имеются две карты, начинающие прогон. Если необходимо нумеровать управляющие карты, то прогон должен начинаться картой NUMBERED. Затем определяется имя заказа картой RUN NAME. Использование этих карт не является обязательным.

5.I.2. Описание данных

Создание системного файла начинается с определения имени файла картой FILE NAME. Затем нужно представить список имен используемых признаков (карта VARIABLE LIST).

Самый простой, но не самый разумный, способ присваивать признакам имена - это различать признаки только по порядковым номерам X1, X2, ... Но такой способ делает трудным интерпретацию распечаток и часто является источником ошибок. Поэтому лучше избрать такое имя признака, чтобы оно напоминало содержательное значение признака. Например, ПРОФССИЯ для признака, значением которого является профессия отвечающего, НАЦИОН - для национальности, ШКОЛА и т.д. Хотя имя признака может включать до 8 символов, оно может быть и короче.

Носитель начальной информации (перфокарты, магнитный

диск и т.д.) определяется картой INPUT MEDIUM. Если объект-признак-матрица разделена на подфайлы, то нужно представить список подфайлов (карты SUBFILE LIST и # OF CASES). При отсутствии подфайлов системе сообщают число вводимых объектов картой # OF CASES.

Затем описывают вводный формат (карта INPUT FORMAT). Вводный формат определяет тип признака и точное расположение значений признаков на носителе исходной информации.

Вводный формат численного признака начинается буквой F. Так, например, если значение численного признака измерено с точностью до двух знаков после запятой и при этом целая часть меньше тысячи, то выбираем формат F5.2. (Первое из этих чисел определяет общее число мест, нужных для представления значения признака, второе - число десятичных знаков).

Если признак является целочисленным, то после точки в описании формата находится нуль. Такие признаки часто полезны, так как в пакете имеются некоторые процедуры, которые с целочисленными признаками работают особенно быстро.

Вводный формат нечисленного признака начинается буквой A. Следует число мест, предусмотренных для значения данного признака. Например, нужно записать A1, если значения признака представлены при помощи одного символа. (Этим символом может быть и цифра, но никакие арифметические операции с этой цифрой тогда не возможны.) При нечисленных значениях, образованных из нескольких символов, используются форматы A2, A3 и т.д.

Порядок форматов должен соответствовать порядку признаков на носителе исходной информации.

Дальше может следовать описательная информация, которая является необязательной. Заказчик по своим конкретным данным всегда может решить, добавить эту информацию, или нет.

Может случиться, что не удалось определить всех значений всех признаков - наши данные содержат пробелы. Так как пробелы могут возникнуть по разным причинам, может понадобиться их обозначить по-разному. Обозначения пропусков определяют картой MISSING VALUES, для одного признака можно определить до трех разных обозначений. Обозначением пропуска может быть определенное число или нечисленное значение, которое не совпадает ни с одним возможным значением признака, но соответствует типу и формату признака. Очень часто при признаках, значением которых не может быть нуль, в качестве обозначения пропуска выбирается нуль. (Этот символ при перфорации можно заменить пробелом). Если такой выбор невозможен (нуль может быть значением признака) часто обозначением пропуска выбирается максимальное число, которое разместится на месте, предусмотренном форматом.

При желании, можно еще ввести метки признаков; ведь имя признака может не полностью определять содержание признака. Тогда для облегчения интерпретации распечаток для признака можно определить "длинное имя" - метку. Эту метку можно использовать во всех задачах при оформлении распечаток. Метка определяется картой VAR LABELS. Длина "длинного имени" признака не может быть больше 40 символов, включая пробелы между словами. Примеры "длинных имен": МЕСТЖИТ - МЕСТО ЖИТЕЛЬСТВА ОТВЕЧАЮЩЕГО, СПОРТИВНЫЙ РАЗРЯД и т.д.

Как часть описательной информации в системный файл

можно включить и предписания кодирования признаков. Метки для значений признаков определяются картой VALUE LABELS, они и объяснят значения кодов. Максимальная длина такой метки - 20 символов. Эти метки можно использовать при оформлении распечаток. Например, значения 1 и 2 признака ПОЛ можно снабжать метками - (1) ЖЕНСКИЙ, (2) МУЖСКОЙ и т.д.

Картой PRINT FORMATS можно определить печатный формат признака, т.е. определить число печатаемых мест после запятой. Для численных признаков использование этой карты не обязательно. Печатный формат нечисленного признака нужно обязательно указать.

5.I.3. Обращение к статистической процедуре

Каждое обращение к статистической процедуре мы называем задачей. Задачу желательно начать с определения его имени (карта TASK NAME). Этой карте следует, при необходимости, руководство использования подфайлов (карта PROCESS SBFILES). Эта карта необязательна - при ее отсутствии структура подфайлов игнорируется.

К статистической процедуре обращаются процедурной картой, на которой указываются имя статистической процедуры, список используемых признаков и способ их использования. Дополнительная информация зависит от используемого статистического метода. Процедурные карты будут подробно описаны в следующих разделах методического руководства.

За процедурной картой могут следовать карты OPTIONS и STATISTICS.

5.I.4. Карта OPTIONS

Каждая статистическая процедура имеет несколько вариантов использования, так называемых опций, которые различа-

ются способом использования начальных данных и объемом выдаваемой информации. Разные опции часто являются невзаимноисключаемыми - несколько разных опций можно использовать одновременно. Номера желаемых опций перечисляются в поле описаний карты OPTIONS, вместе они определяют режим использования статистической процедуры.

Так как разные статистические процедуры имеют разные опции, номера опций и их подробное описание будут представлены вместе с описанием конкретных статистических процедур. Здесь мы рассматриваем только разные возможности использования объектов с пропусками. В большинстве процедур этим возможностям соответствуют разные опции.

Рассмотрим в качестве примера следующие данные:

ИМЯ	ВОЗРАСТ	ВЕС	ПОЛ
АНТС	25	70	М
ЭНН	99	65	М
МАЙ	44	0	Ж
ЛИЗ	37	62	Н
Х	56	80	Ж

Как видно, данные включают пробелы. Для признака ИМЯ пробел обозначен символом Х, для признака ВОЗРАСТ - числом 99, для признака ВЕС - цифрой 0, для признака ПОЛ - буквой Н.

Опция А. Обозначение пропуска игнорируется, т.н. всеобщая обработка.

При такой обработке включаются пять объектов, при этом возрастом отвечающего ЭНН считается 99 лет, весом отвечающего МАЙ - 0 килограммов и т.д. Понятно, что такая всеобщая обработка в большинстве случаев не имеет смысла. Исключением является ситуация, где систематически изучаются объекты с

пропусками. Например, при анкетировании отсутствие ответа может характеризовать отношение отвечающего к данному вопросу.

При большинстве статистических процедур опция А носит номер I.

Опция Б. (Опция, которая применяется при обработке, где признаки будут просмотрены парами, например, при вычислении коэффициентов корреляции).

При использовании этой опции для каждой пары признаков образуется подвыборка, которая не содержит пробелов. Например, для пары ВОЗРАСТ и ВЕС используют объекты с именами АНТС, ЛИЗ и Х, для пары ПОЛ и ВЕС используют объекты с именами АНТС, ЭНН и Х, для пары ВОЗРАСТ и ПОЛ - объекты с именами АНТС, МАЙ и Х. Так для каждой пары получаем 3-объектную выборку, но эти выборки являются разными.

Таким способом максимально используют существующую информацию. Но если вычисляется, например, корреляционная матрица, эта матрица может оказаться "плохой" - она может не быть неотрицательно определенной, частные корреляции могут быть больше единицы и т.д.

В большинстве статистических процедур опция Б носит номер 2.

Опция В. При использовании этой опции для данного списка признаков образуют выборку, которая не содержит пропуски. Для данных, приведенных для иллюстрации, такая подвыборка без пропусков при признаке ВОЗРАСТ состоит из объектов с именами АНТС, МАЙ, ЛИЗ и Х, а при одновременной обработке признаков ВОЗРАСТ, ВЕС и ПОЛ - из объектов с именами АНТС и Х.

В большинстве статистических процедур опция В носит номер 3.

Использование карты OPTIONS при обращении к статистической процедуре не обязательно. Ее отсутствию наиболее часто соответствует опция В - работа с подвыборкой без пропусков.

Кроме перечисленных, некоторые статистические процедуры имеют комбинированные опции, где пропуски аргументов-признаков используют иначе, чем пропуски функций-признаков.

5.I.5. Карта STATISTICS

Объем информации, выдаваемой при выполнении статистической процедуры, регулируется в системе САИСИ картой STATISTICS, в поле описаний которой перечисляются номера желаемых статистик. Так как при разных статистических процедурах номера возможных статистик разные, статистики описываются вместе со статистическими процедурами.

Если желательно напечатать все результаты, получаемые при выполнении статистических процедур, вместо списка номеров в поле описаний карты STATISTICS можно написать ключевое слово ALL. Но использования этой возможности просто "на всякий случай" следовало бы избежать. Неопытному заказчику намного труднее искать нужную ему информацию из большого числа результатов, чем читать лаконичную и компактную распечатку.

5.I.6. Введение начальных данных

Начальные данные, нужные для образования системного файла, вводятся после первого обращения к статистической процедуре. Начальным данным должна предшествовать карта READ INPUT DATA. Если начальные данные находятся на перфо-

картах, то они непосредственно следуют за этой картой. Если за первой задачей следуют еще другие, то соответствующие карты находятся после начальных данных.

При образовании системного файла, как правило, вдобавок введенным признакам, автоматически генерируются три новых. Имена этих признаков SUBFIL, SEQNUM и CASWGT. Первый из этих - SUBFIL - определяет структуру подфайлов и является нечисленным. Его значением для конкретного объекта будет имя подфайла, к которому этот объект принадлежит. Так как форматом этого признака является 4 символа, подфайлы различаются только по первым четырем символам.

Значением признака SEQNUM является порядковой номер объекта в подфайле. Этот признак численный.

Значением признака CASWGT является вес данного объекта. Если в заказе не используют карту WEIGHT, все веса считаются равными единице.

Три перечисленных признака можно использовать в ходе обработки в арифметических и логических выражениях. Но перекодирование или перевычисление этих признаков запрещено.

5.I.7. Завершение прогона

В конце прогона желательно сохранить образованный системный файл картой SAVE FILE. За этой картой должна непосредственно следовать карта конца - карта FINISH.

5.I.8. Пример

В качестве примера рассмотрим колоду управляющих карт для образования системного файла из данных, приведенных в пункте 5.I.4. Так как в каждом прогоне должно быть хоть одно обращение к статистической процедуре, то для всех признаков заказывают одномерные таблицы частот. Это также дает

возможность проверить введенные данные.

```
1                 16                                                  73
NUMBERED          YES                                                  1
RUN NAME          НАЧАЛЬНАЯ ОБРАБОТКА                                  2
FILE NAME         ПРИМЕР                                               3
VARIABLE LIST     ИМЯ, ВОЗРАСТ, ВЕС, ПОЛ                              4
INPUT MEDIUM      CARD                                                 5
# OF CASES        5                                                    6
INPUT FORMAT      FIXED (T3, A4, X, 2F3.Ø, X, A1)                      7
MISSING VALUES    ИМЯ (*)/ ВОЗРАСТ (99)/                               8
                  ВЕС (Ø)/ ПОЛ ('Н')/                                  9
VAR LABELS        ИМЯ, ИМЯ ОТВЕЧАЮЩЕГО                                 1Ø
VALUE LABELS      ПОЛ ('М') МУЖСКОЙ,                                   11
                  ('Ж') ЖЕНСКИЙ/                                       12
PRINT FORMATS     ИМЯ, ПОЛ (A)/                                        13
                  ВОЗРАСТ, ВЕС (2)/                                    14
TASK NAME         ОДНОМЕРНЫЕ ТАБЛИЦЫ ЧАСТОТ                            15
CODEBOOK          ИМЯ ТО ПОЛ                                           16
OPTIONS           4                                                    17
STATISTICS        1, 2                                                 18
READ INPUT DATA                                                        19
     АНТС    25    7Ø    М                                            2Ø
     ЭНН     99    65    М                                            21
     МАЙ     44    Ø     Ж                                            22
     ЛИЗ     37    62    Н                                            23
     *       56    8Ø    Ж                                            24
SAVE FILE                                                              25
FINISH                                                                 26
```

Карта с номером 1 сообщает системе, что управляющие карты занумерированы. В данном случае это нужно только для того, чтобы колоду было удобнее описать.

Карты 2 и 3 сообщают системе имена прогона и создаваемого системного файла. Карты от 4 до 7 передают обязательную информацию для создания системного файла, карты от 8 по I4 - дополнительную информацию.

Карты от I5 по I8 представляют задачу, карты от I9 по 24 - начальные данные.

Карта 25 организует запись системного файла. Карта 26 кончает прогон, сообщая, что колода управляющих карт исчерпана.

5.I.9. Порядок управляющих карт

Для уточнения правил упорядочения управляющих карт картам, встречающимся в описанном прогоне, присваиваются номера предпочтения. Колоду управляющих карт нужно упорядочить так, чтобы номера предпочтения находились в возрастающем порядке. Порядок карт с одинаковыми номерами предпочтения не имеет значения.

Номера предпочтения приведены в следующей таблице. Соблюдение номеров предпочтения, находящихся в таблице в скобках, является обязательным только в рамках одной задачи. За картами начальных данных может следовать описание новой задачи, которое снова начинается картой TASK NAME. Внутри этой задачи номера предпочтения опять являются обязательными. Но дальше может опять следовать карта TASK NAME и описание новой задачи. В конце описания последней задачи находится карта SAVE FILE, а за ней карта FINISH.

Если в прогоне используют карту COMMENT, то она может находиться в любом месте колоды управляющих карт.

Управляющая карта	Номер предпочтении
NUMBERED	1
RUN NAME	2
FILE NAME	3
VARIABLE LIST	4
INPUT MEDIUM	5
SUBFILE LIST	6
# OF CASES	6
INPUT FORMAT	7
MISSING VALUES	7
VAR LABELS	7
VALUE LABELS	7
PRINT FORMATS	7
TASK NAME	(8)
PROCESS SRFILES	(9)
Карта стат. процедуры	(10)
OPTIONS	(11)
STATISTICS	(11)
READ INPUT DATA	(12)
Карты нач. данных	(13)
SAVE FILE	14
FINISH	15

5.2. Прогон с преобразованием признаков

5.2.1. Преобразование признаков

В ходе обработки часто возникает надобность перекодировать или заново вычислить первоначальные признаки.

Арифметические перевычисления можно заказать картой COMPUTE. Как было сказано выше, кроме обычных арифметических операций можно использовать и целый ряд функций преобразования. Но можно и просто дублировать признаки - это полезно для того, чтобы в ходе преобразований сохранить и оригинальный вид признака.

Типичным перевычислением является, например, нахождение разностей признаков для характеризации динамику наблюдаемого явления в течение периода наблюдений. С той же целью вычисляется иногда отношение признаков. Нахождение логарифма, возведение в степень часто применимы при регрессионном анализе для линеаризации связи. В анкетном опросе часто могут формулировать один и тот же вопрос в различных формах, чтобы увеличить достоверность ответа. В таком случае разумно использовать в обработке сумму или какую-то другую комбинацию ответов.

В некоторых случаях может понадобиться перевычислить значение признака для разных объектов по разному предписанию, при этом предписание может зависеть от логического условия, представимого при помощи других признаков. Такое условное перевычисление реализуется при помощи карты IF. Условие представляется в виде логического выражения. Допустим, например, что значение 1 индекса телостроения обозначает стройных людей, т.е. людей, у которых разность между ростом

и весом больше, чем 105. Вычисление этого значения индекса можно тогда представить картой:

```
IF              (РОСТ - ВЕС GT 105)    ИТС = 1
```

Метки для значений нового признака можно потом определить картой VALUE LABELS.

Для перекодирования признаков предназначена управляющая карта RECODE. Для такого перекодирования могут быть разные причины. Это нужно сделать тогда, когда первоначальное перекодирование было в некотором смысле неудачным (например, не сохранилось содержательное упорядочение значений), или в случае, если обнаружилась ошибка в введенных данных, которую нужно исправить или заменить пропуском, или при дискретизации непрерывного признака. При дискретизации область возможных значений непрерывного признака делится на отрезки и каждый такой отрезок кодируется отдельным числом.

Если при перекодировании окажется желательным сохранить и исходные значения признака, его нужно дублировать, т.е. образовать новый признак с теми же значениями, но с новым именем. (Такой признак можно заказать картой COMPUTE). При этом новый признак перекодируется, а прежний сохраняется без изменений.

При переработке материала социологических опросов, а иногда и при других проблемах анализа данных, полезно использовать т.н. индекс перечисления. Такой индекс перечисления можно образовать управляющей картой COUNT. Пусть, например, в анкетном опросе для некоторых приборов домашнего быта опрошено их наличие или перспективы покупки со шкалой: 1 - есть, 2 - покупаем в ближайшее время, 3 - желаем,

но не хватает денег, 4 - не интересует. Подсчитывая ответы "1" для всех перечисленных приборов домашнего быта, найдем их число для каждого отвечающего. Подсчитывая ответы "1" и "2", найдем их число для каждого отвечающего в ближайшем будущем.

Если возникает потребность в преобразовании целого ряда признаков с одинаковым предписанием, то можно уменьшить число требуемых управляющих карт при помощи карт DO и DOEND.

Независимо от того, с каким предписанием преобразуются признаки, нужно быть внимательным к пробелам. Как правило, все преобразования проделываются и с обозначениями пропусков. Так пропуски новых признаков могут приобрести много разных значений. Специальное обозначение для пропуска нового признака можно определить картой ASSIGN MISSING.

При очень объемистой выборке для проверки некоторых первоначальных гипотез может оказаться полезным образовать новую случайную выборку из первоначальных данных. Для образования случайной выборки предназначена управляющая карта SAMPLE.

Из начальных данных можно выбрать только те объекты, для которых выполнено некоторое логическое условие. Для такого условного выбора предназначена управляющая карта SELECT IF.

Для присвоения объектам разных весов предназначена управляющая карта WEIGHT.

5.2.2. Размещение карт преобразования данных

Карты преобразования данных нужно размещать за картами описания данных, перед обращением к первой статистической процедуре. Если обрабатывается системный файл, то карты пре-

образования данных следуют за картой GET FILE. Карты временного преобразования данных размещаются непосредственно перед той статистической процедурой, к которой они принадлежат.

Преобразования начальных данных носят кумулятивный характер - признак, заказанный некоторой картой можно использовать во всех следующих картах. Преобразования выполняются в порядке размещения управляющих карт.

Карты, описывающие новые, преобразованные признаки (ASSIGN MISSING, VAR LABELS, VALUE LABELS, PRINT FORMATS), должны находиться за картами преобразования признаков, перед обращением к статистической процедуре.

При надобности системный файл с новыми, преобразованными признаками можно записать картой SAVE FILE.

5.2.3. Пример

Рассмотрим в качестве примера колоду управляющих карт, представляющую заказ составления регрессионного уравнения. Вид искомого уравнения следующий:

$$У = \alpha \exp\left(\frac{A}{2x_1 + 0{,}5x_2}\right).$$

Уравнение линеаризуется (преобразуется в линейное по отношению к неизвестным коэффициентам α и A) при помощи логарифмирования:

$$\ln Y = C + AW,$$

где $C = \ln\alpha$, а $W = \frac{1}{2x_1 + 0{,}5x_2}$.

Пусть требуется образовать уравнение отдельно для двух групп объектов. При выборе объектов используют значения признака РЕЖИМ, который перед этим дискретизируется. В ходе дискретизации первоначальный пробел заменяется значением 999.

```
1                 16                                          73
NUMBERED          YES                                          1
RUN NAME          РЕГРЕССИОННЫЙ АНАЛИЗ                         2
GET FILE          АБ1                                          3
COMPUTE           ДРЕЖИМ=РЕЖИМ                                 4
RECODE            ДРЕЖИМ (BLANK = 999),                        5
                  ( LOWEST THRU 100 = 1),                      6
                  ( 101 THRU HIGHEST = 2)                      7
COMPUTE           LY = LN (Y)                                  8
COMPUTE           W = 1/ (2* X1 + 0.5 * X2)                    9
ASSIGN MISSING    LY, W (999)                                  10
VAR LABELS        LY, НАТУРАЛЬНЫЙ ЛОГАРИФМ                     11
                  ДОБЫЧИ/W  НОВЫЙ АРГУМЕНТ/                    12
VALUE LABELS      ДРЕЖИМ (1) НИЗКИЙ, (2) ВЫСОКИЙ/              13
TASK NAME         УРАВНЕНИЕ РЕГРЕССИИ                          14
                  НА НИЗКОМ РЕЖИМЕ                             15
*SELECT IF        (ДРЕЖИМ EQ 1)                                16
REGRESSION        VARIABLES = LY, W/                           17
                  REGRESSION = LY WITH W (2)/                  18
TASK NAME         УРАВНЕНИЕ РЕГРЕССИЙ                          19
                  НА ВЫСОКОМ РЕЖИМЕ                            20
*SELECT IF        (ДРЕЖИМ EQ 2)                                21
REGRESSION        VARIABLES = LY, W/                           22
                  REGRESSION = LY WITH W (2)/                  23
SAVE FILE                                                      24
FINISH                                                         25
```

Карты 1 и 2 подготавливают прогон, карта 3 вызывает системный файл, записанный в предыдущем прогоне. Карты от 4 по 9 заказывают новые признаки, а карты от 10 до 13 описы-

вают эти новые признаки. Карта I6 требует образования новой матрицы начальных данных, содержащей только те объекты, у которых значением признака ДРЕЖИМ является 1.

Карты I7 и I8 представляют обращение к процедуре регрессионного анализа. Карты от I9 до 23 представляют новую задачу. Картой 24 записывается системный файл, содержащий, кроме первоначальных признаков, признаки ДРЕЖИМ,LY, W, созданные в ходе прогона и описывающую их информацию. Карта 25 сообщает о конце прогона.

5.2.4. Порядок управляющих карт

Взаимный порядок между картами преобразования данных определяют следующие номера предпочтения:

Управляющая карта	Номер предпочтения
SAMPLE	1
SELECT IF	2
RECODE	2
COMPUTE	2
IF	2
COUNT	2
WEIGHT	3
*SAMPLE	4
*SELECT IF	5
*RECODE	5
*COMPUTE	5
*IF	5
*COUNT	5
*WEIGHT	6

5.3. Прогон редактирования

5.3.I. Возможности для редактирования

В ходе обработки может возникнуть потребность изменить объект-признак-матрицу начальных данных, например, добавить к уже существующим новые результаты измерения. В пакете САИСИ можно добавлять новые признаки (карта ADD VARIABLES) и новые объекты (карта ADD SUBFILES).

Добываемые признаки должны быть измерены у тех же объектов, как признаки, принадлежащие в системный файл. Имена новых признаков должны быть отличными от уже существующих.

Добываемые объекты образуют всегда новый подфайл, который должен содержать в точности те же признаки, какие уже существуют. Но значения признаков измерены у новых объектов.

В ходе обработки может возникнуть потребность интенсивно заниматься с некоторым подмножеством признаков. Для ускорения работы в этом случае разумно образовать новый системный файл, который не содержит лишних признаков.

Для выбрасывания лишних признаков из системного файла имеются две управляющие карты - карты DELETE VARS и KEEP VARS. Первая дает возможность перечислить выбрасываемые признаки, вторая - перечислить сохраняемые признаки. Разумно использовать ту карту, на которой список признаков короче.

Для выбрасывания лишних подфайлов имеется карта DELETE SUBFILES. В ходе одного прогона лишние подфайлы можно выбросить, а вместо них добавить новые подфайлы.

Можно заменить существующую структуру подфайлов новой или создать структуру подфайлов в системном файле, где она до сих пор отсутствовала. Это делается картой NEW SUBFIL.

Заказать переупорядочение признаков можно картой REOR-

DER VARS. Такое переупорядочение может оказаться полезным, например, для использования ТО-соглашения, при представлении списка признаков. Заказать переупорядочение объектов можно картой SORT CASES. Переупорядочение объектов может оказаться необходимым, например, при создании новой структуры подфайлов, поскольку в один подфайл можно соединить только последовательные объекты.

Кроме того в пакете имеется статистическая процедура AGGREGATE, для использования которой нужно переупорядочивать объекты по значениям группирующих признаков. Этой процедурой можно вычислить описательные статистики для групп и добавить их в качестве новых признаков объект-признак-матрице.

5.3.2. Размещение карт редактирования

Местоположение карт редактирования определено двумя следующими правилами:

1) карты ADD VARIABLES, ADD SUBFILES, DELETE SUBFILES, NEW SUBFILES должны следовать сразу за картой GET FILE;

2) карты KEEP VARS, DELETE VARS, REORDER VARS, SORT CASES должны прямо предшествовать карте SAVE FILE.

При добавлении новых данных нужно ввести и описательную информацию - имена новых признаков, метки, вводный формат и т.д. Соответствующие карты следуют за картами редактирования в начале колоды.

Карты ADD VARIABLES и ADD SUBFILES должны непосредственно следовать за картой GET FILE. Поэтому на одном и том же прогоне нельзя добавлять признаки и подфайлы.

5.3.3. Пример

Рассмотрим ситуацию, где из-за ошибок, возникших по разным причинам, подфайл КЛ6 нужно исключить из системного

файла. Добавляется новый подфайл, содержащий теперь правильные данные. Для удобной проверки правильности новых данных образуются одномерные таблицы частот для всех признаков нового подфайла. В конце обработки признаки переупорядочиваются.

```
1                   16                                          73
NUMBERED            YES
RUN NAME            ПРИМЕР РЕДАКТИРОВАНИЯ                       2
GET FILE            ОЦЕНКИ                                      3
ADD SUBFILES        КЛ6К                                        4
DELETE SUBFILES     КЛ6                                         5
# OF CASES          6                                           6
INPUT MEDIUM        CARD                                        7
INPUT FORMAT        FIXED (8F3.Ø)                               8
TASK NAME           ПРОВЕРКА ПОДФАЙЛА                           9
PROCESS SBFIL       (КЛ6К)                                      IØ
FASTMARGINALS       T1 TO T8  (2,5)                             II
OPTIONS             3                                           I2
STATISTICS          1                                           I3
READ INPUT DATA                                                 I4
     3  4  4  5  3  3  3  4                                     I5
     5  5  5  5  5  5  5  5                                     I6
     4  4  5  5  5  5  5  5                                     I7
     4  4  3  3  4  3  4  4                                     I8
     2  3  3  3  3  4  3  3                                     I9
     5  5  5  5  3  3  5  5                                     2Ø
FILE NAME           ОЦЕНКИ2                                     2I
REORDER VARS        T1 TO T4, T7, T8, T5, T6                    22
SAVE FILE                                                       23
FINISH                                                          24
```

Карты 4 и 5 представляют заказ исключения подфайла КЛ6 и введения нового подфайла КЛ6К. Карты от 6 по 8 передают описательную информацию для введения подфайла. Карта I0 сообщает, что работать нужно только с новым подфайлом. Картами II по I3 обращаются к статистической процедуре для образования одномерных таблиц частот. Карты от I4 по 20 передают вводимые данные. Картой 2I исправленному системному файлу определяется новое имя. Такое переименование может оказаться полезным во избежание путаницы в дальнейшем. Карта 22 определяет новый порядок признаков. Карта 23 записывает преобразованный системный файл; карта 24 заканчивает заказ.

5.3.4. Порядок управляющих карт

Совместный порядок карт редактирования определяют следующие номера предпочтения:

Управляющая карта	Номер предпочтения	Комментарии
ADD VARIABLES	I	Должен непосредственно следовать за картой GET FILE
ADD SUBFILES	I	Должен непосредственно следовать за картой GET FILE
DELETE SUBFILES	2	
NEW SUBFILE	3	
KEEP VARS	4	
DELETE VARS	4	
REORDER VARS	5	
SORT CASES	6	Должен непосредственно предшествовать карте SAVE FILE

5.4. Прогон, включающий контрольную печать

5.4.I. Возможности контрольной печати

Как уже было сказано, системный файл включает информацию двух видов - информацию, описывающую данные и сам начальные данные.

Описывающую информацию можно вывести картой DUMP, объем вывода определяет список ключевых слов в поле описаний карты.

Карта DUMP может находиться в произвольном месте в колоде управляющих карт, но здесь надо учитывать два условия:

I) если в ходе прогона создается системный файл, карта DUMP должна находиться после карт описания данных;

2) если используют уже существующий системный файл, то карта DUMP должна следовать за картой GET FILE.

В ходе обработки может также возникнуть потребность проверить начальные данные или правильность преобразований, проделанных с ними. Такую контрольную печать можно заказать картой LIST CASES. Эта карта должна обязательно находиться перед обращением к некоторой статистической процедуре. Объекты печатаются так, как они будут использованы в процедуре.

5.4.2. Пример

Предположим, что требуется проверить, расположены ли признаки правильно во всех подфайлах. Для такой проверки достаточно печатать все значения признаков для первых двух объектов каждого подфайла. Кроме того, заказывают печать всей описательной информации для всех признаков. В ходе

прогона вычисляется корреляционная матрица для всех признаков.

1	16	73
NUMBERED	YES	I
RUN NAME	ПРИМЕР КОНТРОЛЬНОЙ ПЕЧАТИ	2
GET FILE	ЭКСПЕРИМЕНТ	3
DUMP	COMPLETE	4
TASK NAME	КОРРЕЛЯЦИОННАЯ МАТРИЦА	5
PROCESS SERIES	EACH	6
LIST CASES	CASES=2/VARIABLES=TEMP TO ДОБ	7
PEARSON CORR	TEMP TO ДОБ	8
OPTION	3	9
FINISH		IØ

Карта 4 представляет заказ для вывода описательной информации, карта 7 организует контрольную печать значений всех признаков из списка признаков на карте для первых двух объектов каждого подфайла. Карты 8 и 9 представляют обращении к статистической процедуре для вычисления корреляционной матрицы.

УІ. УПРАВЛЯЮЩИЕ КАРТЫ ОПЕРАЦИОННОЙ СИСТЕМЫ

Использование пакета САИСИ происходит под руководством операционной системы ОС. Поэтому при оформлении заказа управляющим картам пакета САИСИ должны предшествовать управляющие карты операционной системы. В данной главе мы дадим краткое описание этих карт, описывая только те параметры, которых пользователь пакета может изменить.

6.I. Представление заказа САИСИ

Для выполнения заказа САИСИ (как и для всех заказов, представленных операционной системе) нужно использовать операторы JOB, JOBLIB и EXEC. Названные три оператора можно использовать, например, в виде:

```
//SAISI JOB U-ØØ8,   Л.КАСК
//JOBLIB DD DSN=SAISI.LOAD,UNIT=SYSDA,
// VOL=SER=SYSSSS,DISP=SHR
//S1 EXEC PGM=SAISI,PARM=1ØØK
```

При помощи оператора JOB представляется шифр задачи (в данном примере U-ØØ8) и фамилия пользователя (в данном случае Л. КАСК). Конечно, каждый пользователь представляет этой картой свой шифр и свою фамилию.

Оператор с именем JOBLIB представляет имя библиотеки, в которой хранится программа SAISI. Этот оператор имеет всегда стандартный вид.

Оператор EXEC сообщает, что нужно решать программу SAISI. Параметром, который пользователь может изменить на этой карте, является параметр PARM, определяющий требование на дополнительный объем памяти. Для пакета САИСИ автоматически выделяют I40 килобайт оперативной памяти. При отсутст-

вии параметра PARM (режим умолчания) выделяются дополнительно 80 килобайт оперативной памяти. При присутствии параметра PARM объем дополнительной оперативной памяти определяет число, написанное после знака равенства. В приведенном примере требуется дополнительно 100 килобайтов оперативной памяти.

Эти три оператора должны присутствовать в приведенном в примере порядке во всех заказах.

6.2. Операторы определения места для вводной и выводной информации

Для ввода и вывода информации нужны некоторые DD операторы с разными именами.

6.2.1. Оператор DD с именем FT01F001

Этот оператор нужен для создания временного файла, в котором сохраняется информация об использованных метках. Использование оператора обязательно во всех заказах, где некоторым объектам присваиваются метки.

В создаваемом файле информация хранится в 800 байтовых блоках. Как правило, число блоков считается равным числу использованных признаков.

Описываемый оператор может, например, встретиться в виде:

```
//FT01F001 DD UNIT=SYSDA,SPACE=(800,(500,100))
```

Объем создаваемого файла определяется параметром SPACE. Первое число в скобках (в примере 800), определяет объем блока и является неизменным. Изменить можно только числа, которые находятся во внутренних скобках. Первое из них (в примере - 500), определяет количество использованных блоков, второе (в примере - 100) - количество дополнительных при

необходимости блоков. Указанное количество блоков можно дополнительно добавить, повторно, до 15 раз.

Если, например, в заказе используют только 10 признаков оператор FT01F001 имеет вид:

```
//FT01F001 DD UNIT=SYSDA,SPACE=(800,(10,1))
```

6.2.2. Оператор DD с именем FT02F001

Этот оператор нужен для создания временного файла, где будут размещены представленные в заказе начальные данные. Использование оператора обязательно во всех заказах, где вводятся начальные данные. Оператор FT02F001 может, например, встретиться в виде:

```
//FT02F001 DD UNIT=SYSDA,SPACE=(2012,(40,10))
```

Начальные данные будут размещены в блоки объемом 2012 байт. Как правило, количество использованных блоков определяется формулой:

(4 х число признаков х число объектов)/ 2012.

Например, при 40 признаках и 500 объектах потребуется 40 блоков, а при 10 признаках и 60 объектах только 2 блока.

Объем создаваемого файла определяется параметром SPACE, способом, описанным в предыдущем пункте.

6.2.3. Оператор DD с именем FT05F001

Этот оператор определяет входной файл для размещения управляющих карт пакета САИСИ. Как правило, этот оператор встречается в виде:

```
//FT05F001 DD *
```

Оператор сообщает, что непосредственно за ним следуют управляющие карты пакета САИСИ. В числе управляющих карт операционной системы он должен быть последним. Использование этой карты обязательно во всех заказах.

6.2.4. Оператор DD с именем FTØ6FØØ1

Этот оператор определяет выходной файл для SAISI в выходном потоке операционной системы. Как правило, оператор FTØ6FØØ1 имеет вид:

```
//FTØ6FØØ1 DD SYSOUT=A
```

Оператор обязателен во всех заказах, без него всякая печать окажется невозможной.

6.2.5. Оператор DD с именем FTØ9FØØ1

При помощи этого оператора описывается файл промежуточных результатов. Такой файл может потребоваться в некоторых опциях многих статистических процедур (Например, во всех процедурах, где можно вывести специальным файлом корреляционную матрицу). Если промежуточные результаты выдаются на печать, этот оператор имеет, как правило, вид:

```
//FTØ9FØØ1 DD SYSOUT=A
```

Если промежуточные результаты выдаются на магнитный диск, то нужно определить имя записываемого файла, а также имя использованного диска. В этом случае оператор может, например, иметь следующий вид:

```
//FTØ9FØØ1 DD DSN=VFAIL,VOL=SER=OSØØØ1,
  UNIT=SYSDA,DISP=(NEW,CATLG,DELETE)
```

Имя записываемого файла определяется параметром DSN (в примере именем выбрано VFAIL), имя использованного магнитного диска определяется параметрами VOL и SER (в примере именем диска является OSØØØ1).

6.2.6. Оператор DD с именем FTØ8FØØ1

Этот оператор используется при вводе ранее записанных вспомогательных результатов (корреляционные матрицы, другие промежуточные результаты, данные, сгенерированные вне сис-

темы). Если нужные данные записаны на магнитный диск, оператор FTØ8FØØ1 имеет, например, следующий вид:

```
//FTØ8FØØ1 DD DSN=VFAIL,UNIT=SYSDA,
// DISP=OLD
```

Имя нужного файла определяется параметром DSN. Если файл был образован некоторой процедурой САИСИ, то это имя определено параметром DSN оператора DD с именем FTØ9ØØ1.

6.2.7. Ввод системного файла

Вводимый системный файл определяется оператором DD именем FTØ3FØØ1. Использование этого оператора обязательно во всех заказах, где встречается карта GET FILE.

Обычно системный файл хранится на магнитном диске. В этом случае типичный вид этого оператора следующий:

```
//FTØ3FØØ1 DD DSN=FAIL1,UNIT=SYSDA,DISP=OLD
```

Имя использованного файла определяется параметром DSN. Это имя не должно совпадать с именем, указанным на карте GET FILE, а определяется при записи файла параметром DSN оператора DD с именем FTØ4FØØ1.

Если в заказе нужно использовать несколько различных системных файлов, то эти файлы определяются операторами FTØ3FØØ2, FTØ3FØØ3 и т.д.

6.2.8. Вывод системного файла

Если образованный системный файл нужно записать, то его необходимо описать оператором DD с именем FTØ4FØØ1. Использование этого оператора обязательно во всех заказах, где встречается карта SAVE FILE.

Если системный файл записывается на магнитный диск, то типичный вид этого оператора следующий:

```
//FT04F001 DD UNIT=SYSDA,VOL=SER=OS0001,
// SPACE=(800,(10,2)),DISP=(NEW,CATLG,DELETE),
// DSN=FAIL2
```

Параметры VOL и SER определяют имя использованного диска, параметр DSN- имя записываемого файла. Значение параметра SPACE описано в пункте 6.2.I.

6.3. Операторы сортирования

Как описано выше, перед записью в системный файл объекты можно переупорядочить. Соответствующий заказ представляют картой SORT CASES. Для выполнения такого заказа нужно добавить еще пять управляющих карт операционной системы.

6.3.I. Операторы DD с именами SORTLIB и SYSOUT

Оператор с именем SORTLIB определяет адрес программы переупорядочивания и используется, как правило, в виде:

```
//SORTLIB DD DSN=SYS1.SORTLIB,DISP=SHR
```

Так как программа переупорядочивания должна иметь свой выход, то нужно добавить еще следующий оператор DD:

```
//SYSOUT DD SYSOUT=A
```

6.3.2. Операторы DD с именами SORTWK01, SORTWK02, SORTWK03

Все эти операторы имеют одинаковый вид и одинаковое предназначение - описать временные файлы на магнитном диске. Типичный вид оператора является следующим:

```
//SORTWK01 DD UNIT=SYSDA,SPACE=(3600,,(250),CONTIG)
```

Из параметров оператора можно изменить только среднее число находящееся в скобках в определении параметра SPACE.

6.4. Примеры

6.4.I. Заказ, в ходе которого образуется системный файл

Приведем управляющие карты операционной системы, которыми нужно дополнить заказ, представленный в пункте 5.I.8:

```
//SAISI JOB K-4Ø5,N.TELLIJA
//JOBLIB DD DSN=SAISI.LOAD,UNIT=SYSDA,
// VOL=SER=SYSSSS,DISP=SHR
//TELL1 EXEC PGM=SAISI,PARM=1K
//FTØ1FØØ1 DD UNIT=SYSDA,SPACE=(8ØØ,(4,1))
//FTØ2FØØ1 DD UNIT=SYSDA,SPACE=(2Ø12,(1,Ø))
//FTØ4FØØ1 DD UNIT=SYSDA,VOL=SER=OSØØØ1,
// SPACE=(8ØØ,(1,Ø)),
// DISP=(NEW,CATLG,DELETE),
// DSN=FNIM
//FTØ6FØØ1 DD SYSOUT=A
//FTØ5ØØ1 DD *
```

Операторы JOB, JOBLIB и EXEC обязательны для каждого заказа. Операторы DD с именами FTØ1FØØ1 и FTØ2FØØ1 нужны для записи вводимой информации. Оператор DD с именем FTØ4FØØ1 определяет точный адрес нового файла, который был образован в ходе обработки. Параметром DSN определяется имя нового файла FNIM.

Операторы DD с именами FTØ6FØØ1 и FTØ5FØØ1 обязательны для каждого заказа для вывода результатов и ввода управляющих карт соответственно.

6.4.2. Заказ, который использует системный файл

Приведем управляющие карты операционной системы, которыми нужно дополнить заказ, представленный в пункте 5.2.3:

```
//SAISI JOB K-705, N.TELLIJA
//JOBLIB DD DSN=SAISI.LOAD,UNIT=SYSDA,
// VOL=SER=SYSSSS,DISP=SHR
//S1 EXEC PGM=SAISI
//FT01F001 DD UNIT=SYSDA,SPACE=(800,(4,2))
//FT02F001 DD UNIT=SYSDA,SPACE=(2012,(1,0))
//FT03F001 DD DSN=AB1,UNIT=SYSDA,DISP=OLD
//FT06F001 DD SYSOUT=A
//FT09F001 DD SYSOUT=A
//FT05F001 DD *
```

Оператор DD с именем FT03F001 определяет адрес вводимого системного файла, оператор DD с именем FT09F001 нужен для вывода промежуточных результатов на печатающем устройстве.

УII. ПЕРВИЧНЫЙ АНАЛИЗ ПРИЗНАКА

Первичным анализом признака считается описание признака при помощи описательных статистик и таблицы частот. Для вычисления описательных статистик в пакете САИСИ предназначена процедура CONDESCRIPTIVE. Составить таблицу частот и вычислить описательные статистики можно при помощи процедур CODEBOOK, MARGINALS и FASTMARG. Существование такого количества процедур оправдывается тем, что первичный анализ признаков требует относительно большую часть машинного времени и объема распечаток. Всякая экономия, связанная с учетом типа признаков (процедура FASTMARG обрабатывает только целочисленные признаки) или более компактной формой распечатки (процедура MARGINALS) может при задачах оказаться полезной. При маленьких задачах можно больше внимания уделить наглядности распечатки (процедура CODEBOOK).

7.I. Процедура CONDESCRIPTIVE

7.I.I. Цель и методические указания

При помощи данной процедуры найдутся описательные статистики для заданных численных признаков. Перечень возможных статистик задан в пункте 7.I.4.

Напомним, что коэффициент асимметрии и эксцесс характеризуют форму распределения, их значения равняются нулю в случае нормального распределения. Чем больше эти статистики отличаются от нуля, тем больше рассматриваемое распределение отличается от нормального.

Описательные статистики имеют смысл только в случае численных признаков, но формально их можно вычислять и для

всех таких признаков, значения которых кодированы числами. Этот факт позволяет пользоваться программами, предназначенными для вычисления описательных статистик, с целью проверки правильности данных. Например, в качестве первого заказа после ввода данных целесообразно заказать некоторые описательные статистики (например, минимум и максимум) всех признаков, значения которых числа или которые закодированы с помощью чисел. Такую же процедуру рекомендуется повторять каждый раз после образования новых признаков или исправления ошибок.

Преимуществом процедуры CONDESCRIPTIVE перед другими процедурами, вычисляющими описательные статистики, является большая скорость работы и компактность распечаток.

7.I.2. Процедурная карта

К процедуре обращаются при помощи процедурной карты, общий вид которой следующий:

CONDESCRIPTIVE {varlist | ALL}

В управляющем поле карты находится имя процедуры, в поле описаний - список тех признаков, для которых требуется вычислить описательные статистики. Если они вычисляются для всех признаков, можно этот список заменить ключевым словом ALL.

Учитывая, что управляющее слово идентифицируется по первым 8 символам, можно имя процедуры в управляющем поле записать в виде CONDESCR.

7.I.3. Опции

Основная опция процедуры (используемая по умолчанию, т.е. при отсутствии карты OPTIONS) не использует при вывычислении обозначений пропусков; для каждого признака,

вместе с заданными статистиками, указывается число присутствующих и отсутствующих значений, а также метка признака.

За картой CONDESCRIPTIVE может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций, отделенные запятой. Возможные опции (описаны лишь их отличия от основной опции) следующие:

I) обозначение пропуска используется в вычислениях как значение признака;

2) подавляется печать меток признаков;

3) из результатов образуется специальный файл, адрес которого определяется оператором DD с именем FTØ9FØØ1.

7.I.4. Статистики

За процедурной картой может следовать и карта STATISTICS, на которой указывается, какие статистики нужно вычислить для признаков, перечисленных на процедурной карте. Можно выбрать следующие статистики:

I) среднее значение (MEAN);

2) ошибка среднего (стандартное отклонение среднего) (STD ERROR);

3) отсутствует;

4) отсутствует;

5) стандартное отклонение (STD DEV);

6) дисперсия (VARIANCE);

7) эксцесс (KURTOSIS);

8) коэффициент асимметрии (SKEWNESS);

9) размах (RANGE);

IO) минимум (MIN);

II) максимум (MAX).

В поле описаний карты STATISTICS перечисляются номе-

ра желаемых статистик, разделяя их запятой. Если нужно вычислить все статистики одновременно, в поле описаний можно написать ключевое слово ALL.

При отсутствии карты STATISTICS печатаются только числа присутствующих и отсутствующих значений.

7.I.5. Пример

Предположим, что в заказе нужно вычислить для признаков ВЕС, РОСТ и ВОЗРАСТ средние значения, стандартные отклонения, максимальные и минимальные значения. Указать метки признаков не считается нужным. Заказ представляется при помощи карт:

```
CODESCR      ВЕС, РОСТ, ВОЗРАСТ
OPTIONS      2
STATISTICS   1,5,10,11
```

7.2. Процедура CODEBOOK

7.2.I. Цель и методические указания

При помощи этой процедуры найдутся таблица частот и эмпирическая функция распределения (т.н. кумулятивные или накопленные частоты) для признаков любого типа (численных или качественных). При желании заказчика вычисляются и описательные статистики (их список задан в пункте 7.2.4) и распечатывается гистограмма распределения.

Для того, чтобы получить достаточно компактную обозримую таблицу частот, целесообразно исследуемый непрерывный признак дискретизировать, т.е. присваивать каждому интервалу значений код. Чтобы сохранить полную исходную информацию, дублируется перед кодированием рассматриваемый признак.

Например, с тем чтобы найти таблицу частот для признака ВОЗРАСТ, мы прежде всего дублируем его с помощью карты

```
COMPUTE      ВОЗКЛАС=ВОЗРАСТ
```

а затем кодирует новый признак с помощью карты

```
RECODE      ВОЗКЛАС (LOWEST THRU 20=1),(21 THRU 30=2),
            (31 THRU 40=3),(ELSE=4).
```

Только после этих предварительных шагов обращаемся к процедуре CODEBOOK.

Следует запомнить, что после перекодирования для признака ВОЗКЛАС описательные статистики не нужные: при их вычислении используются коды классов значений. Правильные значения их можно вычислять с помощью процедуры CONDESCRIPTIVE на основании исходного признака ВОЗРАСТ.

Иногда целесообразно перекодировать и дискретные признаки перед образованием таблиц частот. Такой случай имеет место тогда, когда исходный признак имеет очень много значений, частоты которых чрезмерно малы. В качестве примера рассмотрим признак "Количество детей" КОЛДЕТЕЙ, при котором значения больше 5 встречаются чрезвычайно редко. Для получения компактной таблицы частот следует объединить все значения больше 5. Это реализуется с помощью следующих карт:

```
COMPUTE     ДКОЛДЕТ=КОЛДЕТЕЙ
RECODE      ДКОЛДЕТ ( 6 THRU HIGHEST = 5)
```

Разумеется, что описательные статистики, вычисленные для признака ДКОЛДЕТ, не пригодны для характеризации количества детей: их значения меньше истинных.

7.2.2. Процедурная карта

К процедуре обращаются при помощи процедурной картой, общий вид которой следующий:

```
CODEBOOK    {varlist | ALL}
```

В управляющем поле карты находится имя процедуры, в поле описаний - список тех признаков, для которых требуется сос-

тавить таблицы частот. Если они нужны для всех признаков, список можно заменить ключевым словом ALL.

7.2.3. Опции

Основная опция процедуры не использует при вычислениях обозначения пропусков. В виде таблицы печатаются частоты и относительные частоты значений признаков. Последние вычисляются двумя способами - учитывая обозначения пропусков и не учитывая их. Печатается также эмпирическая функция распределения. Указываются метки признаков и значений, сообщается число присутствующих и отсутствующих значений.

За картой CODEBOOK может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций, отделенные запятой. Возможные опции следующие:

1) обозначения пропуска используются в вычислениях как значения признака;

2) подавляется печать меток признаков;

3) совпадает с основной опцией;

4) вместе с таблицей частот выводится гистограмма;

5) выводится только гистограмма, печать таблицы частот подавляется;

6) для тех признаков, обычная выпечатка таблиц которых превышает страницу, печатается таблица с сокращенным интервалом строк.

Основная опция совпадает с опцией 3.

7.2.4. Статистики

На описательном поле карты STATISTICS перечисляются номера желаемых статистик, отделенные друг от друга запятой. Если вычисляются одновременно все статистики, вместо перечисления всех номеров, можно использовать клю-

чевое слово ALL.

Можно выбрать следующие статистики:

I) среднее значение (MEAN);

2) ошибка среднего (стандартное отклонение среднего) (STD ERROR);

3) медиана (MEDIAN);

4) мода (MODE);

5) стандартное отклонение (STD DEV);

6) дисперсия (VARIANCE);

7) эксцесс (KURTOSIS);

8) коэффициент асимметрии (SKEWNESS);

9) размах (RANGE);

I0) минимум (MIN);

II) максимум (MAX).

Без карты STATISTICS ни одной статистики не вычисляются, а печатаются только таблицы частот.

7.2.5. Пример

Предположим, что требуется найти таблицу частот для признака ВОЗКЛАС и печатать ее с указанием меток значений и гистограммы. Для признака ЧИСПРОВ нужно составить только таблицы частот и вычислить среднее, ошибку среднего и стандартное отклонение. Такой заказ представляется картами:

```
CODEBOOK    ВОЗКЛАС
OPTIONS     4
CODEBOOK    ЧИСПРОВ
STATISTICS  1,2,5
```

Два раза обращаются к процедуре потому, что для разных признаков требуются разные выводы.

7.3. Процедура FASTMARG

7.3.1. Цель и методические указания

Вычисления, совершаемые при помощи этой процедуры, совпадают с теми, которые выполняет процедура CODEBOOK. Различие между этими процедурами состоит в том, что FASTMARG обрабатывает только признаки с целочисленными значениями. Преимуществом этой процедуры является скорость работы и меньший требуемый объем оперативной памяти.

Распечатка процедуры FASTMARG похожа на распечатку процедуры CODEBOOK, отличаясь от последней отсутствием гистограммы в качестве дополнения к таблицам частот.

Обращаясь к процедуре, необходимо указать наибольшие и наименьшие значения обрабатываемых признаков. Для получения корректных распечаток требуется, чтобы символы пропусков также принадлежали указанному интервалу. Слишком широко заданный интервал не влияет отрицательно на результаты.

7.3.2. Процедурная карта

Новым моментом при обращении к процедуре FASTMARG является то, что вместе с именем признака нужно указать и его минимальное и максимальное значение. Общий вид процедурной карты следующий:

FASTMARG varlist $(n_1,m_1)/[.../\text{varlist } (n_k,m_k)/]$

В управляющем поле карты находится имя процедуры, в поле описаний - списки тех признаков, для которых требуется вычислить таблицы частот. В один список объединяют признаки с одинаковым минимальным (n_i) и максимальным (m_i) значениями. Соответствующие значения указываются в скобках после списка признаков. Затем следует наклонная черта и при необходимости новые списки.

7.3.3. Опции

Основная опция процедуры FASTMARG генерирует точно такую-же распечатку, как и основная опция процедуры CODEBOOK.

За картой FASTMARG может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций, отделенные запятой. Возможные опции следующие:

1) обозначение пропуска используется в вычислениях как значение признака;

2) не выводятся метки значений, а метки признаков выводятся;

3) метки признаков и значений не выводятся;

4) из выводимой информации образуют файл, адрес которого определяется оператором DD с именем FTØ9FØØ1; если этот файл определяется на печатающее устройство, то таблицы выводятся с сокращенным интервалом строки;

5) не выводится таблица частот, а только число присутствующих и отсутствующих значений.

Основная опция совпадает с опцией 3.

7.3.4. Статистики

Статистики процедуры FASTMARG совпадают со статистиками процедуры CODEBOOK.

7.3.5. Пример

Предположим, что для последовательных признаков T1,..., T20 нужно составить таблицы частот и все описательные статистики, таблицы желательно выводить с сокращенными интервалом строки. Все эти признаки целочисленные, с минимальным значением нуль и максимальным значением I0. Такой заказ представляется картами:

```
FASTMARG        T1 TO T2Ø (Ø,1Ø)/
OPTIONS         4
STATISTICS      ALL
```

7.4. Процедура MARGINALS

7.4.1. Цель и методические указания

Цель процедуры MARGINALS совпадает с целью процедуры CODEBOOK - это нахождение таблиц частот и эмпирических функций распределения для количественных и качественных признаков. При желании вычисляются и описательные статистики.

По сравнению с процедурами CODEBOOK и FASTMARG процедура MARGINALS имеет более компактную распечатку (примерно на 80% меньше, чем вышеуказанные процедуры). Поэтому применение процедуры MARGINALS является весьма целесообразным в тех случаях, когда признак имеет очень много разных значений, и представляет интерес нахождение вариационного ряда (с повторениями).

По сравнению с другими аналогичными процедурами различием является и то, что коды пропусков не носятся в таблицу частот, необходимая информация о пропусках прилагается в конце таблицы частот.

Процедура не выпускает гистограммы.

7.4.2. Процедурная карта

К процедуре обращаются при помощи процедурной карты, общий вид которой следующий:

MARGINALS {varlist | ALL}

В управляющем поле карты находится имя процедуры, в поле описаний - список тех признаков, для которых требуется составить таблицы частот. Если они нужны для всех признаков, список можно заменить ключевым словом ALL.

7.4.3. Опции

Основная опция процедуры не использует при вычислениях обозначений пропусков. Распечатка представляется не таблицей а строками. В каждой строке указывается под значением признака частота, относительная частота и значение эмпирической функции распределения. Указываются метки признаков. В конце распечатки находится информация о пропусках.

За картой MARGINALS может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций, отделенные запятой. Возможные опции следующие:

1) обозначения пропуска используются в вычислениях как значения признака;

2) не выводятся метки;

3) не выводятся эмпирическая функция распределения и процент отсутствующих значений;

4) выводятся метки признаков;

5) не выводятся частоты, относительные частоты и эмпирическая функция распределения;

6) отсутствует информация о пропусках.

Основная опция совпадает с опцией 4.

7.4.4. Статистики

Статистики процедуры MARGINALS совпадают со статистиками процедуры CODEBOOK.

7.4.5. Пример

Предположим, что для всех признаков нужно вычислить таблицы частот и - так как признаки численные - все описательные статистики. Такой заказ представляется при помощи карт:

```
MARGINALS      ALL
STATISTICS     ALL
```

VIII. ОПИСАНИЕ И СРАВНЕНИЕ ПОДВЫБОРОК

В настоящей главе мы рассмотрим возможности обработки массивов данных, разбитых на подвыборки, при этом описание и сравнение этих подвыборок (подмассивов) составляет цель изучения. Примером такой ситуации являются больные и здоровые в медицинских исследованиях, ученики специализированных школ, общеобразовательных школ и техникумов в педагогических задачах и т.д.

Одной возможностью решения поставленной задачи является образование соответствующих самостоятельных подфайлов и обработка их раздельно. В таком случае для описания подфайлов можно применить все процедуры, описанные в предыдущей главе.

Вторую возможность для раздельной обработки самостоятельных подфайлов предоставляет процедура AGGREGATE. При использовании этой процедуры объекты распределяют в подфайлы по значениям известных признаков и вычисляются описательные статистики для задаваемых признаков во всех подфайлах. При желании найденные значения статистик дублируются для всех объектов, принадлежащих рассматриваемому подфайлу и прибавляются к исходному массиву данных в качестве нового признака.

Третьей возможностью раздельной обработки подфайлов является использование специальных процедур системы САИСИ, предназначенных для описания и сравнения подвыборок. Такими являются процедуры BREAKDOWN, FASTBREAK, T-TEST и ONEWAY. Все эти процедуры обрабатывают только количественные признаки.

8.I. Процедура AGGREGATE

8.I.I. Цель и методические указания

В результате применения процедуры AGGREGATE образуются подвыборки, пользуясь группирующими признаками (их может быть не более четырех). Во всех выборках вычисляют для желаемых признаков некоторые описательные статистики, список которых задан в пункте 8.I.2. При желании эти статистики присоединяют (в качестве новых признаков) к исходному массиву данных. В таком случае значения статистик дублируются для всех объектов всех подфайлов так, чтобы их значения в пределах одного подфайла являлись постоянными.

На основании вычисленных статистик можно образовать и новый массив данных, где в качестве объекта (так сказать - "суперобъекта") выступает подфайл исходного массива, а значениями признаков являются вычисленные значения статистик.

Группирующие признаки, применяемые в процедуре AGGREGATE, должны быть численными. Кроме того требуется, чтобы все группирующие признаки были упорядочены в возрастающем порядке (процедура упорядочения заказывается с помощью карты SORT CASES в конце предыдущего прогона).

Статистические сравнения статистик в подфайлах с помощью процедуры AGGREGATE осуществить невозможно. Основным преимуществом этой процедуры является возможность образовать новые признаки (или новые массивы данных) на основании вычисленных статистик.

8.I.2. Процедурная карта

При обращении к процедуре нужно сказать, какие признаки используются для образования групп, для каких признаков вычислить описательные статистики и какие именно. Общий вид

процедурной карты следующий:

```
AGGREGATE     GROUPVAR = varlist/ VARIABLES=varlist/
              [SET = {YES | NO} /FORMAT = {STANDARD|BINARY}/]
              [MISS = значение/]
              ACTIONS = список ключевых слов/
              [... /VARIABLES = varlist /.../
              ACTIONS = список ключевых слов/]
```

В управляющем поле карты находится имя процедуры. Ключевые слова в поле описаний карты можно делить на обязательные и необязательные. Опишем сперва обязательные ключевые слова.

Первым в описательном поле находится ключевое слово GROUPVAR за которым, после знака равенства, следует список группирующих признаков. Группирующие признаки нужно перечислить точно в том-же порядке, как при заказе их упорядочения на карте SORT CASES, группирующих признаков не может быть больше четырех. Список кончается наклонной чертой. Ключевое слово GROUPVAR можно написать в поле описаний только один раз.

Следует ключевое слово VARIABLES, за которым после знака равенства следует список тех признаков, для которых вычисляются описательные статистики. И этот список кончается наклонной чертой.

Третье обязательное ключевое слово - это ACTIONS. После знака равенства за ним следует список желаемых статистик, которые идентифицируемы ключевыми словами. Возможные ключевые слова и соответствующие им статистики приведены в следующей таблице. Как из таблицы видно, за ключевым словом ACTIONS может следовать, например, список MEAN, SD или NS, SUM, PCTGT(20), PCTBTN(35,50) и т.д.

Ключевое слово	Статистика
NS	Число объектов в группе
SUM	Сумма значений признака в группе
MEAN	Среднее значение признака в группе
SD	Стандартное отклонение в группе
PCTGT (значение)	% тех объектов в группе, у которых значение признака больше значения в скобках
PCTLT (значение)	% тех объектов в группе, у которых значение признака меньше значения в скобках
PCTBTN(значение I, значение 2)	% тех объектов в группе, у которых значение признака находится между приведенными значениями

Если в данных встречаются пропуски, то может случиться, что в некоторой группе не определено ни одно значение признака. В этом случае невозможно вычислить статистики и, таким образом, в файле статистик возникает пропуск. Если файл статистик будет объединен с массивом данных, разумно определить обозначение пропуска. Это можно сделать при помощи ключевого слова RMISS - значение, следующее за знаком равенства будет считаться обозначением пропуска. Как всегда, обозначение пропуска должно различаться от возможных значений признака.

Ключевые слова SET и FORMAT используются, если из статистик образуют самостоятельный файл.

Если возникает потребность для разных признаков вычислить разные статистики, то ключевые слова VARIABLES и ACTIONS можно повторить до 25 раз. Один и тот же признак может в этом случае встречаться в разных списках.

8.I.3. Опции

Основная опция процедуры AGGREGATE выводит для каждой подвыборки заказанные статистики, дублирует их значения для каждого объекта в подвыборке и образует из них самостоятельный файл, адрес которого определяется оператором DD именем FTØ9FØØ1. Объекты с пропусками используются в вычислениях для тех признаков, значения которых измерены. За картой AGGREGATE может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Возможны следующие варианты:

I) обозначения пропусков будут использованы как значения признаков;

2) объекты, у которых отсутствует значение хоть одного признака, встречающегося в списке VARIABLES, не используются в вычислениях;

3) объекты, у которых отсутствует значение хоть одного признака, встречающегося в списке VARIABLES, не используются в вычислениях. Значения статистик не дублируются для всех объектов массива.

8.I.4. Статистики

При помощи карты STATISTICS можно заказать дополнительную информацию о подвыборках. Возможная информация и соответствующие ей номера следующие:

I) для каждой подвыборки печатаются порядковый номер и число объектов в подвыборке;

2) для каждой подвыборки печатаются порядковый номер, число объектов и значения заданных статистик;

3) печатается та же информация, что в опции 2, но для I0 первых подвыборок.

8.I.5. Пример

Предположим, что обрабатывается массив данных, где для жителей разных городов зарегистрирован возраст (признак ВОЗРАСТ), месячный доход (ДОХОД), объем жилплощади (ЖИЛПЛ) и численность семьи (ЧИССЕМ). Пусть для каждого города для этих признаков нужно вычислить среднее значение и стандартное отклонение. Кроме этого, для каждого города нужно выяснить % тех жителей в городе, чей доход меньше 70 рублей и % тех жителей, чья жилплощадь больше 40 $м^2$. Вычисленные статистики нужно дублировать и добавить к массиву данных. Группирующим признаком нужно в данном случае выбрать признак ГОРОД, значения которого должны быть кодированы числами. Объекты должны быть заранее упорядочены по значениям этого признака.

Заказ представляется следующими картами:

```
AGGREGATE     GROUPVARS = ГОРОД/
              VARIABLES = ДОХОД, ВОЗРАСТ, ЖИЛПЛ, ЧИССЕМ/
              ACTIONS = MEAN, SD/
              VARIABLES = ДОХОД /
              RMISS = Ø/ACTIONS = PCTLT(7Ø)
              VARIABLES = ЖИЛПЛ/
              RMISS = Ø/ACTIONS = PCGT(4Ø)
STATISTICS    1
```

8.2. Процедура BREAKDOWN

8.2.I. Цель и методические указания

Наиболее подходящей процедурой в системе САИСИ для вычисления описательных статистик в подвыборках является процедура BREAKDOWN. Эта процедура позволяет образовать подфайлы на основании одного или нескольких группирующих приз-

наков (максимальное допустимое число - 5). Относительно типа группирующих признаков нет никаких ограничений. Разумеется, что число разных значений группирующих признаков не должно быть чрезмерно большим, в особенности тогда, когда объем массива не очень большой.

Для всех исследуемых признаков вычисляются в каждой подвыборке суммы значений, средние, стандартные отклонения и дисперсии, также указываются объемы всех подвыборок. При желании можно проверять, являются ли средние отдельных подгрупп существенно различны. Заметим, что по существу проверяют различие средних в генеральных совокупностях, которые соответствуют рассматриваемым подвыборкам. Такое сравнение делается только для подвыборок, определенных с помощью первого группирующего признака. Для этого выпускается стандартная таблица однофакторного дисперсионного анализа, в последней строке которой указано значение F-статистики. Вероятности значимости (т.е. вероятность ошибки при принятии гипотезы о различии средних) при этой процедуре не вычисляют, поэтому необходимо пользоваться таблицами F -распределения.

Заметим, что таблицами пользоваться не надо в случаях, когда значение F-стстистики меньше единицы - тогда можно сразу утверждать, что различие между средними несущественно. Если значение статистики больше единицы, то из таблицы F-статистики необходимо найти значение, соответствующее выбранному уровню значимости и имеющимся степеням свободы. За уровень значимости обычно выбирается 0,05 или 0,01. Степени свободы получаем из распечатки: первым числом степеней свободы является число, полученное из столбца DEGREES OF FREE-

DOM и из строки BETWEEN GROUPS (это число k-1, где k - число групп). Второе число получается из строки WITHIN GROUPS - это равняется n-k, где n - число наблюдений. Если значение F-статистики, полученное в результате вычислений, больше соответствующего табличного значения, то различие между средними считается доказанным. В противном случае различия между средними не удастся доказать (но этим далеко не доказано, что различия нет). Часто при увеличении выборки различие между группами становится существенным.

В результате процедуры BREAKDOWN невозможно сказать, какие из средних существенно различаются друг от друга, какие нет. Для выяснения этого необходимо применить процедуру ONEWAY.

В задаче, где группирование осуществляется на основании одного количественного признака, можно выдвинуть гипотезу о значимости регрессионной зависимости между группирующими и исследуемым признаками. Для измерения такой зависимости вычисляется квадрат корреляционного отношения (подробнее см. пункт 9.I.I.), коэффициент линейной корреляции r и F-статистика, характеризующая линейность регрессионной функции. Если значение этой F-статистики больше табличного значения (при степенях свободы k-2 и n-k), то можно считать доказанной нелинейность регрессионной функции.

Заметим, что процедура BREAKDOWN не дает возможности нахождения регрессионной функции. В случае линейной зависимости для этого можно применить процедуру SCATTERGRAM, в нелинейном случае - процедуру REGRESSION. В случае зависимости от многих аргумент-признаков проблема решается методами регрессионного анализа (процедура REGRESSION).

Заметим, что в случае нескольких группирующих признаков возможно иерархическое разбиение исследуемого массива на подфайлы, т.е. разбиение по деревообразной схеме. В таком случае прежде всего образуются группы по значениям первого группирующего признака и вычисляются статистики. Затем эти группы разделяют по значениям второго группирующего признака и вычисляются статистики в новых подгруппах и т.д. пока не исчерпан список группирующих признаков.

8.2.2. Процедурная карта

При обращении к процедуре необходимо указать, какие признаки нужно изучать в подвыборках, и перечислить группирующие признаки. Общий вид процедурной карты следующий:

BREAKDOWN varlist BY varlist [BY ... BY varlist]

В управляющем поле карты находится имя процедуры. В поле описаний первым нужно написать список тех признаков, которые надо изучать в подвыборках. Следует ключевое слово BY, а за ним список группирующих признаков. Если для группировки используют несколько признаков, то снова следует ключевое слово BY и новый список признаков, пока все группирующие признаки не будут перечислены. Так с одним обращением можно заказать несколько задач образования подвыборок - в каждой задаче изучают один признак, приведенный в первом списке, а для образования групп используют столько признаков, сколько было разных списков группирующих признаков. При этом из каждого списка выбирают по одному признаку, а для каждого зависимого признака решают столько задач, сколько можно образовать различных комбинаций группирующих признаков.

Например, предположим, что перед нами управляющая карта:

BREAKDOWN УСПЕВ BY ШКОЛА, КЛАСС BY ПОЛ

При ее выполнении для признака УСПЕВ вычисляют статистики сперва в подвыбрках, полученных по признакам ШКОЛА и ПОЛ, а потом в подвыборках, полученных по признакам КЛАСС и ПОЛ.

8.2.3. Опции

Основная опция процедуры BREAKDOWN выводит для всех подвыборок следующие статистики: сумма значений признака (SUM), среднее значение (MEAN), стандартное отклонение (STD DEV) и дисперсия (VARIANCE). Перед статистиками подвыборки указаны значения группирующих признаков, определяющие данную подвыборку. Указано также численность каждой подвыборки, так как объекты, у которых пропущено значение изучаемого признака или одного из группирующих признаков, не участвуют в вычислениях. На распечатке указывают и метки.

За картой BREAKDOWN может следовать и карта OPTIONS, в описательном поле которой перечисляются номера желаемых опций, разделенные запятой. Возможные опции и их различия от основного варианта следующие:

1) обозначение пропуска используют в вычислениях как значение признака;

2) в вычислениях не используются объекты, у которых пропущено значение независимого признака, но из объектов, у которых пропущено значение некоторого группирующего признака, образуют самостоятельную подвыборку;

3) не выводят метки;

4) таблицы выводятся в более плотном виде;

5) результаты выводятся по деревообразной схеме.

8.2.4. Статистики

За картой BREAKDOWN может следовать карта STATISTICS, на которой указывают номера желаемых статистик. Возможные

дополнительные статистики и их номера следующие:

I) выводится стандартная таблица дисперсионного анализа (ANOVA TABLE), в последней строке которой находится значение F-статистики, а в предшествующем столбце степеней свободы (DEGREES OF FREEDOM) межгрупповые (BETWEEN GROUPS) и внутригрупповые (WITHIN GROUPS) степени свободы;

2) выводится таблица (TEST OF LINEARITY) с статистиками,описывающими регрессионную зависимость - квадрат регрессионного отношения (ETA SQRD) и линейный коэффициент корреляции (CORR COEFF), а также F-статистика для проверки значимости регрессионной связи (использование статистики описано в пункте 9.2.I.).

8.2.5. Пример

Предположим, что обрабатывается массив данных, включающий оценки контрольной работы, проведенной во всех параллельных классах различных школ. Нужно вычислить средние оценки по школам, по классам, а потом и для мальчиков и девочек одного класса. Желательно еще проверить различие среднего уровня оценки по разным школам. Очевидно, что для получения желаемых статистик, нужно данные вывести по деревообразной схеме.

Заказ представляется следующими картами:

```
BREAKDOWN     ОЦЕНКА ВУ ШКОЛА ВУ КЛАСС ВУ ПОЛ/
OPTIONS       5
STATISTICS    1
```

При выводе таблиц по деревообразной схеме, образуют подвыборки сперва по группирующему признаку ШКОЛА и выводятся соответствующие статистики. Затем делят полученные подвыборки по признаку КЛАСС и выводят соответствующие ста-

тистики. Наконец делят полученные подвыборки по признаку ПОЛ и выводят статистики.

При выводе таблицы дисперсионного анализа используют только признак ШКОЛА.

8.3. Процедура **FASTBREAK**

8.3.1. Цель и методические указания

Процедура **FASTBREAK** - это ускоренный вариант процедуры **BREAKDOWN**. Большая скорость достигается за счет дополнительного ограничения - для этой процедуры требуется, чтобы все группирующие признаки были целочисленными.

Заметим, что процедурой **FASTBREAK** не рекомендуется пользоваться, так как в распечатке значения двух статистик неправильны - это оценка общего стандартного отклонения (в части описательных статистик) и средняя вариабельность групп (в таблице дисперсионного анализа).

8.3.2. Процедурная карта

При обращении к процедуре используют карту, общий вид которой следующий:

FASTBREAK **VARIABLES = varlist** (n_1,m_1) [,...,
varlist (n_k,m_k)] /
TABLES = varlist BY varlist [**BY** ...**BY varlist**]/

В управляющем поле карты написано имя процедуры. В поле описаний на первом месте находится управляющее слово **VARIABLES**, за которой после знака равенства, следуют списки группирующих признаков. В один список объединены те группирующие признаки, которые имеют одинаковое минимальное (n_1) и максимальное (m_1) значение. Списки кончаются наклонной чертой. Следует ключевое слово **TABLES**, за которым после знака равенства следует список тех признаков, для которых нужно вычис-

лить статистики. За ключевым словом BY следует список группирующих признаков (или их списки, разделенные ключевым словом BY). Этому списку (этим спискам) следует наклонная черта.

8.3.3. Опции

Основная опция процедуры **FASTBREAK** выводит точно такую-же распечатку, как и основная опция процедуры **BREAKDOWN**. За картой **FASTBREAK** может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Возможные опции следующие:

1) обозначения пропусков используются как значения признака;

2) в вычислениях не используют объекты, у которых пропущено значение зависимого признака, но из объектов, у которых пропущено значение некоторого группирующего признака, образуют самостоятельную подвыборку;

3) не выводят метки;

4) результаты выводятся по деревообразной схеме.

8.3.4. Статистики

Статистики процедуры **FASTBREAK** совпадают со статистиками процедуры **BREAKDOWN**.

8.3.5. Пример

Пусть требуется изучать поведение признака TA в подвыборках, образованных тремя различными способами - по группирующим признакам T1 и TB, T2 и TB, T3 и TB. Нужно вычислить описательные статистики в таких группах. Признаки T1, T2, T3 и TB целочисленные, при этом для признаков T1, T2, T3 минимальное значение 0, максимальное - 8, а для признака TB минимальное значение 1, а максимальное - 3. Образуем заказ. Заказ представляется следующими картами:

```
FASTBREAK      VARIABLES = T1 TO T3 (Ø,8), TB(1,3)/
               TABLES = TA BY T1 TO T3 BY TB/
```

8.4. Процедура T-TEST

8.4.I. Цель и методические указания

Целью процедуры T-TEST является сравнение двух генеральных совокупностей по их средним значениям. Для этого имеются две возможности: случай зависимых и случай независимых выборок.

В случае независимых выборок все измеренные объекты разные, у всех объектов измерен один и тот же признак. Выборки (группы) определяются с помощью некоторого группирующего признака, причем для определения группы имеются следующие возможности:

I) задается критическое значение группирующего признака. В первую группу относятся все объекты, у которых значение группирующего признака не меньше критического значения, во вторую - все остальные;

2) задается два критических значения n_1 и n_2 группирующего признака. В первую (вторую) группу относятся все объекты, у которых группирующий признак имеет значение, равное n_1 (соответственно n_2);

3) задается числа k_1 и k_2 так, что k_1 первых объектов файла относят в первую группу, k_2 следующих - во вторую.

Заметим, что при выборе возможностей 2) и 3) не все объекты исследуемого файла участвуют в процедуре.

Возможность 3) образования группы хорошо пригодна для проверки однородности исследуемого материала.

Наиболее популярный вариант T-статистики для сравнения средних вычисляется при предположении, что дисперсии у обо-

их генеральных совокупностей равны. (На распечатке соответствующая дисперсия задается под заглавием POOLED T VALUE). Такое предположение можно сделать лишь в таком случае, когда отношение дисперсий (F VALUE) достаточно близко к единице и вероятность значимости P сравнительно большая (больше чем 0.05). Если же дисперсии в группах сильно различаются друг от друга (F VALUE большое, P меньше чем 0.05), то этим доказано, что дисперсии в генеральных совокупностях разные и для вычисление T-статистики необходимо применять предположение о разных дисперсиях (SEPARATE T VALUE).

Если решено, какой вариант T-статистики надо выбрать, то возможно установить на распечатке и значение T-статистики, число степеней свободы, и вероятность ошибочного решения при доказательстве различия средних. Если эта вероятность достаточно мала (меньше 0.05), то считается доказанным, что генеральные совокупности имеют разные средние значения.

Если же вероятность больше критического значения, то невозможно доказать, что совокупности разные. Причиной этого может быть либо недостаточные объемы выборок, либо фактическое совпадение средних. Ни в том, ни в другом случае доказать равенства средних с помощью T-теста невозможно.

Случай зависимых наблюдении имеет место тогда, когда одни и те же самые объекты измерены два раза. В таком случае признаки (первое и второе наблюдение) считаются разными.

В распечатке выдается разность признаков, стандартное отклонение и стандартная ошибка этой разности, а также значение T-статистики, число степеней свободы и вероятность ошибки. Если эта вероятность мала, то различие между средними доказано.

8.4.2. Процедурная карта

При обращении к процедуре указывается, являются ли изучаемые выборки зависимыми или независимыми (в последнем случае описывается, каким способом нужно образовать группы) и перечисляют, какие признаки нужно сравнить. Общий вид карты следующий:

```
T-TEST      {PAIRS = varlist/ | GROUPS = описание группы/
             VARIABLES = varlist/}
```

В управляющем поле карты находится имя процедуры, для выполнения поля описаний есть две возможности. Если нужно сравнить зависимые выборки, в поле описаний нужно написать ключевое слово PAIRS, за которым после знака равенства следует список тех признаков, которые нужно сравнить по парам. Если имеем дело с независимыми выборками, в поле описаний первым записывают ключевое слово GROUPS, за которым после знака равенства следует описание групп. Для оформления этого списка, по описанным в предыдущем пункте возможностям, можно использовать одну из трех следующих конструкций:

I) имя признака (значение);

2) имя признака (значение I, значение 2);

3) значение I, значение 2.

За описанием групп следует ключевое слово VARIABLES, а за ним, после знака равенства - список тех признаков, которые в разных группах нужно сравнить. На одной и той же карте может быть несколько заказов для нескольких групп.

8.4.3. Опции

Основная опция процедуры T-TEST выводит как для зависимых так и для независимых выборок для обеих групп численности, средние значения, стандартные отклонения и ошибки

средних значений. Для независимых выборок выводятся еще статистика для сравнения дисперсий, вероятность значимости этой статистики и две Т-статистики для сравнения средних вместе с их вероятностями значимости. Для зависимых выборок выводятся среднее значение и стандартное отклонение разностей признака, а также ошибка среднего значения; коэффициент корреляции между изучаемыми признаками вместе с вероятностью значимости; Т-статистика для сравнения средних вместе с ее вероятностью значимости. На распечатке указываются метки признаков. Объекты, у которых отсутствует значение группирующего признака, не будут использованы в вычислениях. Объекты, у которых отсутствует значение одного из признаков, не будут использованы в вычислениях только для этого признака.

За картой T-TEST может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Возможные опции и соответствующие им номера следующие:

1) обозначения пропуска используют как значения признака;

2) в вычислениях не используются объекты, у которых отсутствует значение хоть одного признака, встречающегося в списках VARIABLES или PAIRS;

3) не печатаются метки признаков.

8.4.4. Статистики

Процедура T-TEST не имеет свободно выбираемых статистик.

8.4.5. Пример

Предположим, что обрабатываются данные учеников разных школ одного города. Школы различаются по их номерам. Пусть требуется сравнить средние успеваемости и спортивную активность (оцененную обычной школьной оценкой) 3-ей и 5-ой шко-

лы. Здесь мы имеем дело с независимыми выборками, поэтому заказ нужно представить следующими картами:

```
T-TEST          GROUPS = ШКОЛА (3,5)/
                VARIABLES = УСПЕВАЕМ, СПОРТАКТ/
```

Приведем еще пример заказа для зависимой выборки, где нужно сравнить пары признаков:

X1 и X2, X1 и X3, X2 и X3, X1 и X4.

Такой заказ представляется картой

```
T-TEST          PAIRS = X1, X2, X3/ PAIRS = X1, X4/
```

8.5. Процедура ONEWAY

8.5.I. Цель и методические указания

8.5.I.I. Процедура ONEWAY выполняет однофакторный дисперсионный анализ. При этом предполагается, что массив данных состоит из k групп (выборок) из разных генеральных совокупностей, имеющих средние $\mu_i (i=1,\ldots, k)$. Целью дисперсионного анализа является проверка следующих гипотез:

H_1^1: не все средние μ_i равны друг другу (найдутся такие индексы i и j, при которых $\mu_i \neq \mu_j$);

H_0^1: Средние всех генеральных совокупностей равны ($\mu_1 = \mu_2 = \ldots = \mu_k$).

Дополнительно предполагается, что все генеральные совокупности имеют нормальное распределение, притом дисперсии равны. Кроме того естественно требуется, что выполнены обыкновенные предположения выборки: измерения независимы, нет повторяющихся объектов. Отдельные выборки (группы) могут иметь разные объемы, но в случае равных объемов выбор возможных процедур дополнительного анализа немного шире.

Для проверки гипотезы H_1^1 выпускается стандартная таблица дисперсионного анализа, где указано значение F-стати-

стики, число степеней свободы и соответствующая вероятность значимости. Если эта вероятность меньше выбранного исследователем уровня значимости (стандартным значением для него является 0.05), то гипотеза H_1^1 считается доказанной.

Если же вероятность значимости больше чем 0.05, то принимается гипотеза H_0^1. Это значит, что при данном материале не удастся доказать различия средних в разных совокупностях. Напомним, что этим не доказано, что средние в разных совокупностях совпадают (H_0^1). Этого с помощью дисперсионного анализа невозможно доказать в принципе.

8.5.I.2. Проверка равенства дисперсий. Если средние не отличаются существенно друг от друга, то имеется возможность, что дисперсии существенно различные. Для этого в процедуре ONEWAY выпускаются три разные тестовые статистики; из них тесты Кочрена (COCHRAN) и Хартли (HARTLEY) применяемы только при равных объемах групп, тест Бартлет-Бокса (BARTLETT-BOX) не предаолагает равенства групп, но его чувствительность меньше двух первых. Для тестов Кочрена и Бартлет-Бокса не выпускаются и вероятности значимости для проверки гипотез:
H_1^2: Существуют индексы i и j такие, что $\sigma_i^2 \neq \sigma_j^2$;
H_0^2: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$.

Дисперсии считаются существенно различными, если заданная в распечатке вероятность меньше уровня значимости. В обратном случае принимается (но не считается доказанной) нулевая гипотеза, что дисперсии не различны.

8.5.I.3. Дальнейшие возможности исследования средних. Если средние не оказались различными (по таблице дисперсионного анализа приняли нулевую гипотезу H_0^1), то на этом разумно закончить анализ, и можно планировать новый этап ра-

боты, объединяя некоторые из заданных k групп так, чтобы

I) дисперсии в них стали близки друг другу,

2) численности отдельных групп увеличились

и затем оформить новую задачу дисперсионного анализа.

Если по таблице дисперсионного анализа приняли гипотезу H_1^1 (средние различные), а предположения дисперсионного анализа выполнены (приняли H_0^2, т.е. дисперсии совпадают), то можно более детально исследовать поведение средних в группах.

8.5.I.4. Полиномиальная регрессия. Если группирующий признак X по содержанию есть количественный, то можно проверять гипотезу о том, существует ли полиномиальная регрессионная зависимость между X и исследуемым признаком У:

$$У \approx b_0 + b_1 x + \ldots + b_r x^r.$$

Для проверки этой гипотезы следует в заказе фиксировать степень r полинома (при помощи параметра POLYNOMIAL); имеется ограничение, что $r \leqslant 5$; кроме того разумеется, что r не может быть больше, чем число k разных значений группирующего признака X, значит $r \leqslant k$).

Статистики, необходимые для проверки адекватности полиномиальной регрессии, заданы дополнительно в таблице дисперсионного анализа (печатаны по столбцам на три позиции левее чем основные статистики дисперсионного анализа).

Одна из возможных схем анализа статистик, характеризующих полиномиальную зависимость, следующая:

Рассмотрим таблицу дисперсионного анализа и начнем с последней строки (DEV FROM R). На двух последних местах этой строки находится значение F-статистики (F-RATIO) предназначенной для проверки гипотез

H_1^1: Регрессионная функция порядка r не адекватная модель,

H_0^1: Регрессионная функция порядка r адекватна.

Последнее число в этой строке (F PROB) есть вероятность значимости F-статистики, по существу вероятность гипотезы $П_0^1$.

Если эта вероятность меньше уровня значимости (обычно 0.05), то полиномиальная модель порядка r неадекватна и рекомендуется сформить новый заказ, где выбирается значение r больше.

Если вероятность больше уровни значимости, то полученная модель адекватна, но возникает вопрос об уменьшении степени полинома. Для проверки этой гипотезы рассмотрим следующую строку таблицу (R TERM), где задана F-статистика для проверки гипотез:

H_1^2: Член $b_r x^r$ в модели существенный ($b_r \neq 0$)

H_0^2: Член $b_r x^r$ в модели несущественный ($b_r = 0$).

Если вероятность значимости (F PROB) меньше уровня значимости (0.05), то полиномиальная регрессия порядка r даст наиболее подходящую модель, и задача решена.

Если, наоборот, вероятность значимости больше уровня значимости, то можно в качестве модели рассмотреть полином (r-1)-й степени.

На первом шаге проверяется его адекватность.

Случай, когда модель r-й степени не существенна, а (r-1)-й степени неадекватна, является противоречивым, и это может случиться в случае, когда предположения дисперсионного анализа не выполнены (в группах дисперсии сильно различаются друг от друга, распределения имеют некоторую "паталогическую" форму). В таком случае рекомендуется изменить

структуры групп или вообще отказаться от полиномиальной регрессии.

Нормальным является случай, когда в результате продолжения вышеизложенных шагов найдется такое значение r_1, при котором принимается нулевая гипотеза H_0^1 и содержательная гипотеза H_1^2, т.е. полиномиальная модель порядка r_1 адекватна и член $b_{x1}x^{r1}$ в ней существенный. На этом шагу эта процедура заканчивается.

Оценить параметры этой полиномиальной регрессионной модели с помощью процедуры ONEWAY невозможно, для этого необходимо пользоваться процедурой REGRESSION (в случае линейной регрессии применима и процедура SCATTERGRAM).

8.5.I.5. Метод линейных контраст. Для того, чтобы выяснить, какие из средних значений μ_1 существенно различаются друг от друга, можно применить метод линейных контрастов.

Метод линейных контрастов применим тогда, когда имеется достаточно априорной информации о группах, в которых определены средние. На основании этой информации можно выдвинуть гипотезы о средних, например:

$H_1^1: \mu_1 \neq \mu_4$,

или

H_1^2: среднее первых трех групп отличается от среднего всех следующих групп

и т.д.

Такие гипотезы можно записать в форме некоторых линейных выражений (условий), содержащих средние μ_i и постоянные c_i:

$$L = c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k . \qquad (8.I)$$

Для определенности считается, что сумма коэффициентов c_i равняется нулю:

$$c_1 + c_2 + \ldots + c_k = 0. \qquad (8.2)$$

Выражения (8.I) называются линейными контрастами. Легко переписать гипотезы H_1^1 и H_1^2 с помощью линейных контрастов

H_1^1: $\mu_1 - \mu_4 \neq 0$ $\quad (L_1 \neq 0)$

(значит, $c_1=1$, $c_4=-1$, все остальные $c_i = 0$),

H_1^2: $1/3\,\mu_1 + 1/3\,\mu_2 + 1/3\,\mu_3 - 1/5\,\mu_4 - \ldots - 1/5\,\mu_8 \neq 0$ $\quad (L_2 \neq 0)$

(предполагается, что $k = 8$).

Если фиксирована гипотеза H_1, то легко написать и соответствующую ей нулевую гипотезу H_0, которая является альтернативой H_1. В данном случае имеем:

H_0^1: $\mu_1 - \mu_4 = 0$ $\quad (L_1 = 0)$

H_0^2: $1/3\,\mu_1 + 1/3\,\mu_2 + 1/3\,\mu_3 - 1/5\,\mu_4 - \ldots - 1/5\,\mu_8 = 0.$ $(L_2=0)$

Для проверки гипотез

H_1: $L \neq 0$

H_0: $L = 0$

применима стандартная методика, базирующаяся на применении T-теста.

Для применения метода линейных контрастов необходимо задавать все c_i $(i=1,\ldots,k)$, в том числе и нулевые (требуется, чтобы было выполнено условие (8.2).

В распечатке выпускаются значение (оценка $\hat{L}$) контраста L, вычисленное на данном материале, и стандартное отклонение, вычисленное двумя способами: I) при предположении, что дисперсии все равны - POOLED VARIANCE ESTIMATE и 2) без этого предположения - SEPARATE VARIANCE ESTIMATE.

Из теста равенства дисперсий (см. п. 8.5.I.2) вытекает, какой из этих случаев выбрать (таким же образом, как в случае T-теста, см. 8.4.I). Окончательное решение делается на основании вероятности значимости T PROB: если оно меньше

уровня значимости (0.05), то принимается гипотеза H_1, которая утверждает, что средние в рассматриваемых нами группах разные. В противном случае принимается нулевая гипотеза: невозможно доказать существенного различия между средними групп, определяемых контрастом L.

8.5.I.6. Множественные сравнения. Если нет основания для преобразования групповой структуры, определенной группирующими признаком, то можно применить процедуру множественных сравнений. В результате применения этой процедуры группы упорядочивают по величине их средних μ_i, и проверяют для всех пар гипотезы

$$H_0: \mu_{i'} = \mu_{i'+1}, \qquad (8.3)$$
$$H_1: \mu_{i'} \neq \mu_{i'+1}$$

где $\mu_{i'}$ - среднее i'-ой группы в упорядоченной по величине $\bar{x}_i$ последовательности групп.

Если гипотеза H_0 принимается, то группы с номерами i' и i'+1 объединяют. Затем проверяют аналогичную пару гипотез с целью выяснения, следует ли этой группе присоединить и группу с индексом i'+2 и т.д. Заметим, что полученная в результате структура групп может быть частично покрывающая: например, одна группа состоит из исходных групп i_1 и i_1+1, другая из i_1+1 и i_1+2.

Уровень значимости для проверки гипотез (8.3) задается исследователем. Заметим, что здесь не рекомендуется выбирать очень маленький (0.0I) уровень значимости, так как при таком выборе объединяют в одну группу исходные группы с довольно сильно различающимися друг от друга средними.

Для проверки пары гипотез (8.3) можно применять 7 разных тестов, базирующихся на разных предположениях и имеющих

разную мощность.

Суммарная информация об этих тестах приведена в следующей таблице. Тесты приведены в порядке уменьшения мощности. Первые тесты, имеющие наибольшую мощность, требуют и более строгие ограничения.

Название теста	Обозначения	Требование о численности групп	Применяемые уровни значимости
I. Наименьшая существ. разность	LSD	равные группы	произвольная
2. Данкан	DUNCAN	равные группы	0.I;0.05;0.0I
3. Стьюдент-Ньюман Кеульс	SNK	равные группы	0.05
4. Тьюки альтернативный	TUKEYB	произвольные	0.05
5. Тьюки существ. разность	TUKEY	произвольные	0.05
6. Модифицированный тест наименьшей существ. разности	LSDMOD	произвольные	произвольная
7. Шеффе	SCHEFFE	произвольные	произвольная

8.5.I.7. Суммируя вышеизложенное можно сказать, что процедура ONEWAY применима для первичной обработки массива данных, состоящего из нескольких групп или содержащих мешающие признаки, имеющие несколько дискретных значений. Если рассмотреть все существенные для дальнейшей работы признаки зависящими от групповых признаков, и применять дисперсионный анализ (или сравнение средних), можно либо выделить некото-

рые гомогенные группы, в которых существенные явления имеют разный характер, или вообще игнорировать групповую структуру как несущественную.

8.5.2. Процедурная карта

При обращении к процедуре нужно указать способ образования групп и список тех признаков, для которых нужно провести дисперсионный анализ. Затем может следовать необязательная информация, нужная для выяснения конкретных различий средних значений. Общий вид карты следующий:

```
ONEWAY      GROUPS(k) = {имя группирующего признака|
            SUBFILES  | n1,n2, ... nk} /
            VARIABLES = varlist/ [POLYNOMIAL=n/]
            [CONTRAST = коэффициенты контраста/] [.../
            CONTRAST = коэффициенты контраста/]
            [RANGES = имя теста [(уровень значимости)]/.../
            RANGES = имя теста [(уровень значимости)]/]
```

В управляющем поле карты находится имя процедуры. Ключевые слова, написанные в поле описаний, можно разделить на обязательные и необязательные. Опишем сперва обязательную часть.

Первым запишется ключевое слово GROUPS, за которым в скобках следует число групп к. После знака равенства указывается, как образовать группы. Возможны три способа:

I) указывается имя группирующего признака. Одну группу образуют объекты с одинаковым значением группирующего признака; значениями группирующего признака должны быть натуральные числа I,2, ... ,к;

2) указывается ключевое слово SUBFILES . Одну группу образуют объекты одного подфайла;

3) указывается число последовательных объектов. Одну

группу образует указанное число последовательных объектов. Описание групп кончается наклонной чертой.

Следующим записывается в поле описаний ключевое слово VARIABLES, за которым после знака равенства следует список тех признаков, к которым при данной группировке нужно применить дисперсионный анализ.

Дальнейшее заполнение карты зависит от того, какой метод используется для сравнения средних. Если желательно использовать регрессионноый полином то нужно написать ключевое слово POLYNOMIAL, за которым после знака равенства следует степень используемого регрессионного полинома. Использованная степень не должна быть больше пяти. Следует наклонная черта. Ключевое слово POLYNOMIAL может на карте встречаться только один раз.

Если регрессионный полином заказан вместе с первым или вторым способом группирования, то значением группирующего признака является порядковой номер группы.

При использовании метода контрастов в поле описаний нужно записать ключевое слово CONTRAST, за которым после знака равенства следуют коэффициенты всех групповых средних, разделенные пробелом, или запятой. Коэффициенты средних, неиспользованных в данном контрасте, являются нулями. Различные контрасты разделяются наклонной чертой и каждому из них предшествует ключевое слово CONTRASTS. На одной карте не должно быть больше 10 контрастов.

При использовании множественного сравнения средних в поле описаний нужно записать ключевое слово RANGES, за которым, после знака равенства, указывается обозначение желаемого теста (эти обозначения приведены в предыдущем пункте).

За обозначением теста может следовать в скобках уровень значимости - вероятность ошибки, считая группы разными. Ограничения при выборе уровня значимости приведены в таблице, по умолчанию уровень значимости считается равным 0.05. Каждому использованному тесту предшествует ключевое слово RANGES. На одной карте это ключевое слово не должно встречаться более I0 раз.

8.5.3. Опции

Основная опция процедуры ONEWAY образует из объектов, у которых отсутствует значение группирующего признака, специальную группу. Объекты, у которых пропущено значение некоторого зависимого признака, не используются только при анализе этого признака, в анализе остальных признаков они присутствуют. На распечатке указаны метки признаков. Выводятся стандартные таблицы дисперсионного анализа, где кроме значения F-статистики приведена и ее вероятность значимости. Для характеризации регрессионной зависимости печатаются соответственные статистики. Для контрастов вычисляются оценки, значение T-статистики и вероятность значимости. При использовании рангового теста выводятся описания образуемых групп.

За картой ONEWAY может следовать карта OPTIONS, в поле описаний ее перечисляются номера желаемых опций, разделенные запятой. Возможные опции и их номера следующие:

I) обозначения пропуска используют как значения признака;

2) при вычислениях не используют объекты, у которых отсутствует значение хоть одного признака из списка VARIABLES;

3) не выводятся метки;

4) образуют файл промежуточных результатов (адрес ко-

торого определяется оператором DD с именем FT09F001, в который для каждой группы записывается число объектов, среднее значение и стандартное отклонение;

5) отсутствует;

6) для обозначения групп используют первые 8 символов из метки значения группирующего признака;

7) начальные данные представляются процедуре в нестандартном виде. Для каждой группы представляется число объектов, средние значения и стандартные отклонения. Если начальные данные вводятся с перфокарт, данные каждой группы начинаются с новой карты и численность представляется в формате F10.0, средние значения и стандартные отклонения в формате F10.4. Начальным данным предшествует карта READ MATRIX;

8) используют нестандартные начальные данные - частоты групп, средние значения, оценку общей дисперсии групп и ее степени свободы. Частоты и средние значения вводятся так же как при опции 7. Оценка дисперсии и ее степени свободы должны находиться на отдельной карте в формате 2F10.4. Если в столбцах перфокарты от 11 по 20 вместо степеней свободы находятся пробелы, число степеней свободы считают равным N-k, где N - число объектов, а k - число групп. По таким данным можно вычислить контрасты и провести множественное сравнение при помощи ранговых тестов.

8.5.4. Статистики

В описательном поле карты STATISTICS указываются номера желаемых статистик, разделяя их запятой. Если нужно одновременно вычислить все статистики, в поле описаний можно использовать ключевое слово ALL. Можно выбрать следующие статистики:

1) для каждой группы выводят численность объектов, среднее значение, стандартное отклонение, ошибка среднего, минимум, максимум, 95%-доверительные границы для средних;

2) оценка общей дисперсии как для модели с фиксированными уровнями фактора так и для модели со случайными уровнями фактора;

3) выводятся статистики для сравнения дисперсии групп и соответствующие им вероятности значимости.

9.5.5. Пример

Предположим, что обрабатываются данные учеников разных школ одного города. Пусть требуется сравнить средние уровни успеваемости и спортивной активности для всех школ в 8-ых классах. 1-, 2-, 5- и 6-ая школы специальные, куда учеников принимают по конкурсу, желательно проверить, является ли средний уровень успеваемости в этих школах более высоким, чем в остальных. Общее число школ в этом городе - 9.

При представлении заказа образуем сперва временный подмассив, который включает только данные учеников 8-ых классов. После этого обращаемся к процедуре ONEWAY:

```
*SELECT IF    (КЛАСС=8)
ONEWAY        GROUPS(9)=ШКОЛА/VARIABLES= УСПЕВАЕМ СПОРТАКТ/
              CONTRASTS=Ø.25,Ø.25,-Ø.2, -Ø.2,
              Ø.25, Ø.25, -Ø.2, -Ø.2, -Ø.2/
STATISTICS    3
```

При анализе распечатки нужно сперва изучить таблицу дисперсионного анализа. Если сравниваемые средние не различаются существенно на выбранном уровне значимости, то изучать контрасты уже не нужно.

IX. ХАРАКТЕРИЗАЦИЯ ЗАВИСИМОСТИ МЕЖДУ ДВУМЯ ПРИЗНАКАМИ

Характеризация связи между признаками - важнейшая задача на первом этапе анализа данных. Наиболее точное совместное распределение двух признаков представляет их таблица частот. Для образования этой таблицы все объекты классифицируются по значениям двух признаков X и У (пусть они имеют соответственно k и l значений). Элемент i-ой строки и j-ого столбца n_{ij} полученной таблицы указывает, какое количество объектов имели соответственно значения $X=x_i$ и $У=y_j$. В последней строке и в последнем столбце таблицы частот обычно приводятся т.н. маргинальные частоты соответствующих отдельных значений y_j и x_i (это есть суммы отдельных столбцов и строк).

Если при каждом значении одного признака поведение (распределение) второго признака одинаковое, то признаки независимы. В случае независимых признаков относительные частоты $\frac{n_{ij}}{n}$ приблизительно пропорциональны маргинальным частотам обоих признаков:

$$n_{ij} \approx \frac{n_i \cdot n_j}{n} .$$

Это значит, что совместное распределение независимых признаков определено распределениями отдельных признаков, и исследование совместного распределения и его характеристик не может дать никакой дополнительной информации.

Если разным значениям одного признака соответствуют различные распределения другого признака, то эти признаки статистически зависимые. Заметим, что если признак X зависит (в смысле статистической зависимости) от признака У, то

и признак У зависит от признака Х.

В связи с тем, что статистическая зависимость между разными парами признаков может быть различной, имеется много разных характеристик для измерения статистической зависимости между (двумя) признаками.

В случае номинальных признаков (качественные, без упорядоченности) исследование их зависимости ограничивается установлением статистической зависимости. Если признаки упорядочены, то можно точно характеризовать статистическую зависимость между ними, измеряя и их монотонную зависимость. Монотонная зависимость может быть либо возрастающей либо убывающей. В первом случае, как правило, "большим" значениям Х соответствуют "большие" значения У (в смысле введенной упорядоченности) и "малые" значения Х сопровождаются "малыми" значениями У. В случае убывающей зависимости "большим" значениям Х соответствуют "малые" значения У и наоборот.

Имеется ряд характеристик, т.н. ранговые коэффициенты корреляции, измеряющие монотонную зависимость между признаками. Монотонная зависимость является вообще говоря взаимной, но бывает, что Х зависит от У сильнее, чем У от Х.

Если один из исследуемых признаков есть количественный, то можно говорить о регрессионной зависимости.

По существу регрессионная зависимость есть односторонняя: (количественный) признак У зависит от Х в смысле регрессионной зависимости тогда, когда среднее значение У не является постоянным при всех значений Х. Это значит, что при регрессионной зависимости среднее значение У есть некоторая функция от Х.

Если Х также есть количественный признак, то можно ис-

следовать и регрессионную зависимость признака X от У. Регрессионная зависимость не является взаимной. Регрессионную зависимость характеризуют при помощи двух коэффициентов, показывающих зависимость X от У и У от X.

Регрессионную зависимость между двумя количественными признаками часто аппроксимируют посредством некоторой заданной функции - регрессионной функцией. Чаще всего в качестве регрессионной функции применяют линейную функцию. Если линейная функция между двумя признаками не постоянная, то говорят, что между этими признаками существует коррелятивная зависимость. Коррелятивная зависимость взаимная: если X коррелируется с У, то У коррелируется с X. Теснота коррелятивной зависимости характеризуется коэффициентом корреляции.

Коррелятивная зависимость - самый узкий тип статистической зависимости среди вышеуказанных. Из нее вытекает как регрессионная, так и монотонная, а также и статистическая зависимости. Однако из того, что между признаками нет коррелятивной зависимости, не вытекает отсутствие регрессионной, монотонной или статистической зависимостей. Практически, все-таки между некоррелированными признаками обычно нет и монотонной зависимости, а регрессионная зависимость между некоррелированными признаками весьма редко заметна (и то только односторонняя).

Для образования двумерных таблиц частот и вычисления разных коэффициентов зависимости в системе САИСИ имеется процедура CROSSTABS и ее модификация для целочисленных признаков FASTABS. Для исследования признаков с количественными значениями подходит процедура SCATTERGRAM. Для заказа корреляционных матриц имеются процедуры PEARSON CORR и

NONPAR CORR, а также и специальная процедура REGRESSION для нахождения многомерных функций регрессии.

9.I. Процедура CROSSTABS

9.I.I. Цель и методические указания

Процедура CROSSTABS предназначена для построения таблиц частот (дву- и многомерных), причем признаки могут быть количественными, порядковыми или номинальными. Многомерные таблицы (характеризующие совместное распределение трех, четырех или большего количества признаков) являются, по существу, таблицами условных распределений. Они образуются так, что в таблицу входят первые два признака, указанных на управляющей карте, а все остальные признаки, указанные на карте, определяют условия: для каждой возможности комбинации этих признаков получается одна двумерная таблица.

Разумеется, что чем больше признаков, определяющих условия, и чем больше у них значений, тем больше получается двумерных таблиц и тем меньше объектов попадет в каждую двумерную таблицу. Поскольку для того, чтобы сделать по таблицам содержательные выводы, необходимо, чтобы число наблюдений было достаточно большим, многомерными таблицами можно пользоваться лишь при исследовании очень больших массивов данных.

Для описания зависимости между признаками процедура имеет большой выбор характеристик зависимости. Каждый исследователь должен выбрать подходящие. Одновременное вычисление всех характеристик зависимости обычно не целесообразно.

В дальнейшем мы опишем вычисляемые процедурой CROSSTABS характеристики зависимости.

9.I.I.I. χ^2-статистика (Хи-квадрат статистика). Это ста-

тистика, которая характеризует тесноту статистической зависимости. Значения χ^2-статистики находятся на отрезке $[0, n(t-1)]$, где n объем выборки, а $t=\min(k, \ell)$, где k и ℓ - количества разных значений признаков X и У.

χ^2-статистика равняется нулю в случае, если признаки X и У независимы (наблюдаются конечные генеральные совокупности объема n), т.е. для всех i и j имеет место точное равенство $n_{ij}=n_i \cdot n_j/n$.

Статистической зависимости соответствуют сравнительно большие значения χ^2-статистики. При маленьких выборках ($n<50$) рекомендуется считать независимыми признаками X и У (в генеральной совокупности) тогда, когда значения статистики χ^2 не больше числа степени свободы $(k-1)(l-1)$.

Более корректное применение χ^2-статистики базируется на том факте, что при известных предположениях

$$\min(n_i, n_j) \geqslant c \qquad (9.I)$$

(для c применяются разные значения от $c=2$ до $c=5$) и для достаточно большого n χ^2-статистика имеет χ^2-распределение с числом степеней свободы $(k-1)(l-1)$.

В распечатке процедуры CROSSTABS наряду со значением статистики χ^2 задается и ее вероятность значимости. Если условие (9.I) выполнено и вероятность меньше чем уровень значимости (например, 0.05), то доказано, что исследуемые признаки статистически зависимы.

Если условие (9.I) выполнено, но вероятность значимости больше уровня значимости, то при данном материале невозможно доказать статистическую зависимость исследуемых признаков.

Если условие (9.I) не выполнено, то можно исходные признаки перекодировать при помощи карты RECODE, объединяя,

таким образом, слишком малочисленные классы.

Заметим, что процедурой всегда опускается вероятность значимости и проверка того, выполнены ли предположения использования χ^2-распределения, является делом исследователя.

В случае 2x2-таблицы вычисляется не χ^2-статистика, а т.н. "точная вероятность Фишера", т.е. вероятность получения такой таблицы при предположении, что признаки X и У, имеющие такие же маргинальные распределения, независимы. Практическое применение этой вероятности для проверки гипотезы о зависимости стандартное: признаки считаются зависимым, если эта вероятность меньше уровня значимости.

Недостаткой χ^2-статистики является тот факт, что ее значения зависят от размера таблицы и поэтому разные пары признаков трудно сравниваемы по тесноте статистической зависимости. В этом состоит причина, почему на основании χ^2-статистики вычисляются разные коэффициенты, характеризующие статистическую связь между парой признаков и имеющие значения всегда на промежутке $[0,1]$.

9.I.I.2. Коэффициент Крамера V (Ф-коэффициент) - это характеристика статистической зависимости, полученная путем нормирования χ^2-статистики. Для получения коэффициента Крамера χ^2-статистика нормирована таким образом, что ее максимальное возможное значение равняется единице, а минимальное - 0. Максимальное значение достигается в случае, когда значения одного признака однозначно определяют значения другого признака. Чем сильнее статистическая зависимость, тем больше значение коэффициента V, для проверки значимости коэффициента V (т.е. доказательство гипотезы о зависимости) применима χ^2-статистика; если предположения χ^2-распределе-

ния не выполнены, то и V не имеет содержания.

По традиции в случае 2x2-таблицы коэффициент Крамера называется Ф-коэффициентом.

9.I.I.3. Коэффициент контингентности или коэффициент Пирсона C - это также характеристика зависимости, полученная из χ^2-статистики путем нормирования. Максимальное значение этого коэффициента не равняется единице, а в случае $k \times \ell$ таблицы ($k<\ell$) равняется $\sqrt{(k-1)/k}$. Поэтому коэффициент Пирсона не применим для сравнения таблиц с разными измерениями.

Значимость коэффициента контингентности проверяется при помощи χ^2-статистики точно так, как и значимость коэффициента Крамера.

Следует сказать, что параллельное использование коэффициентами V и C лишено смысла. Если нет специальных рассуждений в пользу коэффициента C, то рекомендуется пользоваться коэффициентом V, учитывая его более удачную нормировку.

9.I.I.4. Следующие коэффициенты зависимости, вычисляемые процедурой CROSSTABS, характеризуют монотонную зависимость. Они имеют содержательный смысл только в том случае, когда оба признака количественные или упорядоченные, причем кодированы при помощи чисел, сохраняющих упорядоченность.

При вычислении характеристик монотонной зависимости основываются на рангах значений признаков - это порядковые номера упорядоченных значений. В отличие от характеристик, вычисленных по χ^2-статистике, ранговые корреляции работают тем лучше, чем больше у признаков разных значений, пустые клетки ($n_{ij}=0$) в таблице частот не мешают вычислению характеристик монотонной зависимости,

9.I.I.5. Коэффициенты Кэндалла τ_b и τ_c характеризуют тесноту монотонной зависимости. Их возможные значения расположены на отрезке $[-1,1]$.

Коэффициент τ_b подходит для исследования квадратной таблицы ($k=\ell$). Максимальное абсолютное значение τ_b достигает в случае взаимно-однозначной зависимости (таблица имеет одну из двух возможных диагональных форм; если все значения таблицы находятся по главной диагонали, то $\tau_b=1$, если на дополнительной диагонали, то $\tau_b=-1$. Чем меньше $|\tau_b|$, тем слабее монотонная зависимость между признаками.

Коэффициент τ_c применим для прямоугольных таблиц, т.е. при $k\neq\ell$. Максимальное абсолютное значение коэффициент τ_c достигает при однозначной монотонной зависимости.

Для проверки значимости монотонной зависимости в случае $n>10$ для коэффициентов τ_b и τ_c указывается и вероятность значимости.

9.I.I.6. γ-коэффициент Гудман-Краскала - это коэффициент, который характеризует взаимную монотонную зависимость и его рекомендуется использовать в случае, когда признаки в известном смысле равноправны. Коэффициент γ принимает значения на отрезке $[-1,1]$, причем значения 0 и -1,1 достигает в тех же случаях, как и коэффициент τ_b.

Заметим, что из того факта, что некоторый коэффициент монотонной зависимости равняется нулю, не вытекает, что признаки независимы, они могут быть и зависимыми, только форма зависимости отличается от монотонной.

9.I.I.7. D-коэффициент Сомерса характеризует также монотонную зависимость, но в данном случае эта зависимость направленная, не взаимная. Поэтому коэффициентом D можно

пользоваться в случае, когда имеется априорная информация о том, что признак X (первый признак в заказе) зависит от признака У (второй признак в заказе). Значения коэффициента D так же как и коэффициента τ, находятся в интервале $[-1,1]$.

9.I.2. Процедурная карта

К процедуре обращаются при помощи управляющей карты, общий вид которой следующий:

CROSSTABS **varlist BY varlist** [**BY ... BY varlist**]

В управляющем поле карты находится имя процедуры, в поле описаний перечисляются имена тех признаков, которые нужно использовать при классификации. При составлении таблицы частот используют только один признак из каждого списка, разделенного ключевым словом BY. Таблицы вычисляются для всевозможных комбинаций признаков, образованных таким способом. Так, например, при выполнении заказа

CROSSTABS **T1, T2, BY T3, T4**

образуют 4 таблицы для пар T1 и T3, T1 и T4, T2 и T3, T2 и T4.

9.I.3. Опции

Основная опция процедуры выводит двумерные таблицы для первой пары признаков при фиксированных значениях других. В j-ом столбце i-ой строки указывают кратность (COUNT) пары значений (x_i, y_j), относительная частота значения y_j в строке (ROW PCT), относительная частота значения x_i в столбце (COL PCT) и относительная частота пары (TOT PCT). На распечатке указываются метки значений и признаков, а также число объектов с пропусками - объекты, у которых пропущено значение одного используемого признака, не классифицируются.

За картой CROSSTABS может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций.

Возможные опции и их номера следующие:

I) обозначения пропусков используются как значения признака;

2) не выводятся метки;

3) не указывается относительная частота значения в строке;

4) не указывается относительная частота в столбце;

5) не указывается ни одной относительной частоты.

9.I.4. Статистики

На карте STATISTICS указывается, какую характеристику связи нужно вычислить. Соответствующие им номера следующие:

I) χ^2-статистика (CHI SQUARE);

2) коэффициент Крамера V или Ф-коэффициент (CRAMER'S V или PHI);

3) коэффициент контингентности (CONTINGENCY COEFF);

4) отсутствует;

5) отсутствует;

6) τ_b Кэндалла (KENDALL'S TAU B);

7) τ_c Кэндалла (KENDALL'S TAU C);

8) γ-коэффициент (GAMMA);

9) коэффициент Сомерса (SOMER'S D).

В поле описаний карты STATISTICS перечисляются номера желаемых статистик. Если нужно вычислить всех статистик, можно использовать ключевое слово ALL. Если карта STATISTICS отсутствует, ни один коэффициент не будет вычислен.

9.I.5. Пример

Предположим, что нужно образовать двумерные таблицы от признаков X 1 ,..., X6 с признаком FF, отдельно для мужчин и женщин. В таблице нужно указать только кратность и относительную частоту пары значений. Для оценки тесноты связи ме-

жду признаками выбран χ^2-статистика и V Крамера, а для оценки монотонности связи - τ_b Кэндалла (последний можно использовать, так как известно, что все применяемые признаки имеют одинаковое число значений). Заказ представляется следующими картами:

```
CROSSTABS      X1 TO X6 BY FF BY ПОЛ
OPTIONS        3, 4
STATISTICS     1, 2, 6
```

9.2. Процедура FASTABS

9.2.I. Цель и методические указания

Процедура FASTABS предназначена для образования таблиц в случае целочисленных признаков.

Для оценки зависимости между признаками можно использоваться всеми характеристиками, которые были описаны в пункте 9.I.I.; кроме того λ-коэффициенты, коэффициенты неопределенности, регрессионные (корреляционные) отношения η. Для коэффициента D вычисляют три разных варианта, а правило вычисления χ^2-статистики немного изменяется по сравнению с соответствующим правилом для процедуры CROSSTABS.

9.2.I.I. χ^2-статистика вычисляется так, что при построении таблиц учитываются и такие значения X и У, при которых частоты равняются нулю. В результате этого для всех таблиц, у которых $\min(n_{ij})=0$, χ^2-статистика, вычисленная при помощи процедуры FASTABS, имеет меньшее значение, тем эта же статистика, вычисленная при процедуры CROSSTABS. То же самое имеет место и для всех статистик (V,C), вычисленных по χ^2-статистике.

9.2.I.2. λ-коэффициенты (симметричный и несимметричные) характеризуют тесноту статистической зависимости в смы-

сле прогнозируемости одного признака по другому; полной прогнозируемости соответствует максимальное значение $\lambda=1$, полной непрогнозируемости (независимости) минимальное значение $\lambda=0$.

Так как для статистики λ вероятности значимости пакетом САИСИ не вычисляются, то они имеют только описательное значение и не пригодны для проверки гипотез о существовании зависимости; с этой целью можно использоваться χ^2-статитикой.

9.2.I.3. Коэффициенты неопределенности (симметричный и несимметричные) измеряют также, как и C, V и λ-коэффициенты, тесноту статистической зависимости, основываясь на понятии количества информации, введенной Кульбаком и Лейблером.

Количество информации I, которое содержит один признак о другом признаке, вычисляется по формуле

$$I = H(X) + H(Y) - H(X,Y),$$

где H(X), H(У) энтропия признаков X и У, а H(X,У) энтропия их совместного распределения. Все энтропии вычислены по формуле Шеннона. I является максимальным в случае взаимно однозначного соответствия между множествами значений признаков X и У, а при независимых X и У имеем I=0. Коэффициентом неопределенности называется значение I, нормированное так, что он изменяется на отрезке $[0,1]$.

При исследовании влияния одного признака (X) на другой признак (У) получается несимметричный коэффициент неопределенности I/H(У) (ASYMMETRIC, Y DEPENDENT). Аналогично получается и коэффициент, измеряющий влияние признака У на X, это I/H(X) (ASYMMETRIC, X DEPENDENT). Симметрический коэффициент неопределенности получается путем нормирования I на среднее (H(X) + H(У))/2 (SYMMETRIC).

Запомним, что коэффициенты неопределенности, так же как

и другие коэффициенты статистической зависимости (C, V, λ) не изменяются в результате перекодирования признаков.

9.2.I.4. Коэффициенты Сомерса вычисляются трех типов: два асимметрических (первый раз зависящим считается первый признак в списке, второй раз зависящим - второй признак). Кроме того, вычисляется еще симметричный коэффициент Сомерса (см. 9.I.I.7).

9.2.I.5. Регрессионное (корреляционное) отношение η^2 характеризует тесноту регрессионной зависимости и имеет смысл лишь в случае, когда зависимый признак является количественным. Так как система САИСИ вычисляет η^2 для всех пар признаков, то исследователь сам отвечает за корректность исползования этого показателя.

Значения регрессионного отношения также находятся на отрезке $[0,1]$, причем значение 1 соответствует случаю, когда зависимый признак полностью определен независимым; значение $\eta^2=0$ имеет место в случае, когда среднее значение зависимого признака не зависит от аргумента, а является постоянным при всех значениях аргумента. Заметим, что из этого в общем случае не следует независимость признаков.

Для η^2 процедура FASTABS не выдает вероятности значимости. Все же, значимость регрессионной зависимости легко проверять при помощи F-распределения. Для этого необходимо по найденному значению η^2 вычислить значение статистики $F=\eta^2(n-k)/\ ((1-\eta^2)(k-1))$ и сравнивать его с соответствующими данному уровню значимости табличным значением F-распределения со степенями свободы k-1 и n-k (здесь k - число разных значений аргумента, n - число используемых объектов). Если значение вычисленной F-статистики больше табличного

(при выбранном уровне значимости, например 0.05), то считается доказанной гипотеза о существовании регрессионной зависимости. Следует запомнить, что из того, что вычисленное значение F меньше табличного, вытекает, что невозможно доказать регрессионной зависимости (но не доказать отсутствия ее).

9.2.2. Процедурная карта

Обращаются к процедуре при помощи процедурной карты, общий вид которой следующий:

FASTABS VARIABLES=varlist (n_1,m_1)[,...,varlist (n_k,m_k)]/
TABLES= varlist BY varlist [BY ... BY varlist]/

В управляющем поле карты написано имя процедуры, но поле описаний заполняется иначе, чем в предыдущей процедуре. Сперва после ключевого слова VARIABLES нужно перечислить все признаки, которые используются при составлении таблиц. В один список объединяют признаки с одинаковым минимальным (n_i) и максимальным (m_i) значением. После перечисления признаков, за ключевым словом TABLES, заказывают таблицы. Так же, как и при процедуре CROSSTABS, при составлении таблицы используют только один признак из каждого списка, отделенный ключевым словом BY. Таблицы образуются для всевозможных комбинации пар признаков, полученных таким способом.

9.2.3. Опции

Основная опция процедуры FASTABS выводит точно такую же распечатку, как и основная опция процедуры CROSSTABS. За картой FASTABS может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Возможные опции и их номера следующие:

I) обозначения пропуска используются как значение признака;

2) не выводятся метки;

3) не указывается относительная частота значения в строке;

4) не указывается относительная частота в столбце;

5) не указывается относительная частота пары значений;

6) не выводятся метки значений, метки признаков выводятся;

7) обозначения пропуска используют как значения признака только при составлении таблиц, но их не используют при вычислении статистик.

9.2.4. Статистики

На карте STATISTICS указывают, какие характеристики связи нужно вычислить. Можно выбрать следующие:

I) χ^2-статистика (CHI SQUARE);

2) коэффициент Крамера V или Ф-коэффициент(CRAMER'S V или PHI);

3) коэффициент контингентности (CONTINGENCY COEFFICIENT);

4) λ-коэффициенты (LAMBDA SYMMETRIC, LAMBDA ASYMMETRIC);

5) коэффициент неопределенности (UNCERTAINTY COEFFICIENT (SYMMETRIC), (ASYMMETRIC));

6) τ_b Кэндалла (KENDALL'S TAU B);

7) τ_c Кэндалла (KENDALL'S TAU C);

8) γ коэффициент (GAMMA);

9) коэффициенты Сомерса (SOMER'S D(SYMMETRIC),(ASYMMETRIC));

I0) регрессионное отношение η^2 (ETA).

В описательном поле карты STATISTICS перечисляются номера желаемых статистик, отделяя их запятой. Если считается нужным вычислить все статистики (при данной процедуре это очень редкий случай), можно использовать ключевое слово ALL. Если при обращении к процедуре FASTMARG карта STATISTICS отсутствует, то ни одна характеристика не вычисляется.

9.2.5. Пример

Пусть по некоторой анкете требуется изучать совместное поведение признаков ИНДЕКС и СТАТ, отдельно для мужчин и женщин. Признаки ИНДЕКС и СТАТ целочисленные, значения первого признака меньше 20, второго - 10. Указать меток на распечатке не считается нужным. Для оценки тесноты связи выбрана τ_c Кэндалла. Заказ представляется следующими картами:

```
FASTABS     VARIABLES=ИНДЕКС(Ø,2Ø),СТАТ(Ø, Ø),ПОЛ(1,2)/
            TABLES=ИНДЕКС ВУ СТАТ ВУ ПОЛ/
OPTIONS     6
STATISTICS  7
```

9.3. Процедура SCATTERGRAM

9.3.1. Цель и методические указания

При помощи процедуры SCATTERGRAM можно задавать диаграмму рассеяния (корреляционное поле), описывающее совместное распределение двух количественных признаков. Для них можно вычислить и статистику, описывающую линейную регрессию.

На корреляционном поле объекты изображаются в виде точек, (ж), первая координата которой определяется первым признаком, указанным в заказе, вторая координата - вторым. Если одной точке графика соответствует больше объектов, то на распечатке изображается вместо ж их число (от 2 до 9); 9 обозначает и те точки, которым соответствует 10 и больше объектов.

Размеры корреляционного поля всегда одинаковые: 51 символов х 100 символов. Масштаб осей выбирается заказчиком, а если он соответствующих параметров не указал, то масштаб выбирается автоматически так, что минимальное и максимальное значения признаков определяют крайние точки поля.

С целью большей наглядности корреляционное поле разби-

вается прямыми на 9 равных четырехугольников. На корреляционное поле можно нанести и диагонали.

С целью характеризации корреляционной зависимости с помощью этой процедуры вычисляют среднее квадратичное отклонение (STD ERR OF EST) - это статистика, характеризующая рассеяние зависимого признака относительно прямой регрессии, свободный член a и коэффициент регрессии b для функции линейной регрессионной зависимости $y=a+bx$. Для проверки корреляционной зависимости (т.е. для проверки гипотезы $H_1: b<0$ ($r<0$) или $H_1': b>0$ ($r>0$) при нулевой гипотезе $H_0: b=0$ (или $r=0$))
опускается коэффициент корреляции r и его вероятность значимости (вероятность нулевой гипотезы). Если эта вероятность меньше выбранного уровня значимости, то существование корреляционной зависимости считается доказанным.

В противном случае признаки не существенно коррелированы (принимается, но не доказывается нулевая гипотеза).

Если отсутствует предварительная информация о возможном знаке корреляционного коэффициента, правильно использовать содержательную гипотезу:
$H_1'': b\neq0 (r\neq0)$.

Знак коэффициента корреляции указывает направление зависимости: в случае возрастающей зависимости корреляция положительна, а в случае убывающейся зависимости корреляция отрицательна. При функциональной зависимости (значение зависимости признака полностью определено прямой регрессии, все точки находятся на этой прямой) абсолютное значение коэффициента равняется единице. Чем слабее корреляционная зависимость, тем ближе к нулю значение коэффициента корреляции.

При желании вычисляется и т.н. коэффициент детерминации - это квадрат коэффициента корреляции. Значение коэффициента детерминации имеет наглядную интерпретацию - он показывает, какая доля (сколько %) из изменяемости одного признака описывается через линейную регрессию (по другому признаку).

9.3.2. Процедурная карта

При обращении к процедуре нужно указать пары, для которых надо составить диаграмму рассеивания. Общий вид карты следующий:

SCATTERGRAM **varlist** $[(n_1,m_1),\ldots,$ **varlist** $(n_k,m_k)]$
$[$**WITH varlist** $[(s_1,t_1),\ldots,$**varlist** $(s_1,t_1)]]$

В поле управления находится имя процедуры. Так как для идентификации управляющего слова используются только первые восемь символов, то ее можно представить и в виде **SCATTERG**. Поле описаний можно заполнить двумя способами. Во-первых, можно туда написать только список переменных. В таком случае диаграммы рассеивания образуют для всевозможных различных пар признаков, которые можно образовать по данному списку. Во-вторых, можно в поле описаний написать список переменных, за ним ключевое слово **WITH**, а за ним новый список переменных. В таком случае диаграммы рассеивания составляются для всех пар признаков, которые можно образовать, использовав один признак из первого списка, а другой - из второго.

Если желательно определить масштаб диаграммы рассеивания, то списки делят на несколько подсписков. За каждым из них указывается в скобках минимальное и максимальное значение признака, который будет указан на диаграмме. В один подсписок объединяются признаки, у которых эти значения сов-

падают. Можно также передать только одно из этих значений, а вместо другого использовать ключевое слово LOWEST или HIGHEST.

9.3.3. Опции

Основная опция процедуры выводит диаграмму рассеивания признаков. Указывают метки признаков.

Объекты с отсутствующими значениями не используются. Диаграмма делится на девять равных прямоугольников. Вероятность значимости коэффициента корреляции вычисляется для одностороннего теста. Если объектов слишком много для данного объема оперативной памяти, то диаграмма не выводится.

За картой SCATTERGRAM может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Возможные опции и соответствующие им номера следующие:

1) обозначения пропуска используются как значения признака;

2) не используются объекты, у которых отсутствует значение хоть одного признака, перечисляемого на процедурной карте;

3) не выводятся метки;

4) диаграмма рассеивания не делится на прямоугольники;

5) на диаграмме рассеивания печатаются диагонали;

6) вероятность значимости коэффициента корреляции вычисляется для двухстороннего теста; если заказана соответствующая статистика;

7) на диаграмме рассеивания используются целочисленные метки;

8) если объектов слишком много для данного объема оперативной памяти, на графике выводятся только n первых объектов, где n - число объектов, допустимое при данном объе-

ме памяти.

9.3.4. Статистики

На карте STATISTICS указывают, какие дополнительные статистики вместе с диаграммой рассеивания нужно вывести. Можно выбрать следующие статистики:

1) коэффициент корреляции (CORRELATION);

2) коэффициент детерминации (R SQUARED);

3) вероятность значимости коэффициента корреляции (SIGNIFICANCE);

4) среднеквадратная ошибка (STD ERR OF EST);

5) свободный член регрессионного уравнения (INTERCEPT);

6) коэффициент регрессии (SLOPE).

Если нужно вычислить все статистики, можно использовать ключевое слово ALL.

9.3.5. Пример

Изучается массив антропометрических данных, где вместе с другими признаками измерено рост (признак РОСТ), обхват груди (ГРУДЬ), вес (ВЕС) и обхват бедер (БЕДРА). Нужно изучать связь признаков РОСТ и ГРУДЬ с признаками ВЕС и БЕДРА. Для этого заказывают четыре диаграммы рассеивания вместе со всеми описательными статистиками. Выделить прямоугольники на диаграмме рассеивания не считается нужным, но требуется вывести диагонали. Заказ представляется следующими картами:

```
SCATTERGRAM   РОСТ(150,180), ГРУДЬ(75,100)WITH ВЕС(45,80),
              БЕДРА(80,110)
OPTIONS       4, 5
STATISTICS    ALL
```

X. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

При массиве данных, содержащем много признаков, первоначальной работой обычно является описание общего характера связей между признаками и выяснение структуры этих связей. Основой для такой работы является корреляционная матрица всех признаков массива. Для вычисления корреляционной матрицы в пакете САИСИ имеются три процедуры: PEARSON CORR - для вычисления (линейных) коэффициентов корреляции, NONPAR CORR - для вычисления непараметрических коэффициентов корреляции между упорядоченными или численными признаками и PARTIAL CORR - для вычисления частных коэффициентов корреляции. Кроме того, для численных признаков корреляционную матрицу можно вычислить и в ходе работы многих других процедур - FACTOR, CANCORR, DISCRIMINANT, REGRESSION.

IO.I. Процедура PEARSON CORR

IO.I.I. Цель и методические указания

Процедура PEARSON CORR вычисляет для численных признаков корреляционную матрицу - таблицу, в которой для данного списка признаков находятся коэффициенты корреляции всевозможных пар признаков. (При желании можно вычислить только часть этой таблицы.) Вместе со значением коэффициента корреляции указывается число объектов, которые были использованы при вычислении и вероятность значимости. Хотя вероятность значимости вычисляется всегда, ее использование оправдано только при предположении, что признак имеет нормальное распределение. По умолчанию вероятности значимости вычисляются для односторонних тестов, т.е. исходя из пар гипотез:
H_0: признаки независимые ($\rho = 0$);

H_1: признаки положительно ($\rho > 0$) коррелированы (или признаки отрицательно ($\rho < 0$) коррелированы).

Такую одностороннюю гипотезу можно выдвинуть при наличии дополнительной информации о характере связи между признаками. Если такая информация отсутствует, нужно использовать двухстороннюю гипотезу, т.е. содержательную гипотезу представить в виде:

H_1: признаки коррелированы ($\rho \neq 0$).

В обоих случаях содержательную гипотезу считают доказанной, если вероятность значимости не превышает выбранного уровня значимости (например, 0.05).

Напомним еще, что коэффициент корреляции описывает только линейную связь между признаками. При численных признаках с достаточно широкой шкалой значений достаточно часто возможна ситуация, когда связь между признаками характеризуется нелинейной функцией (тесноту связи измеряет корреляционное отношение); а связь между некоторыми функциями этих признаков оказывается линейной. В таком случае можно вычислить (картой COMPUTE) подходящие функции первоначальных признаков (например, x^2, $\ln x$, e^x и т.д.) и добавить их к первоначальным данным. Если окажется, что новый признак $f(x)$ действительно связан с остальными более сильно, чем первоначальный признак x, то нужно его включить в дальнейшую обработку. В факторном и каноническом анализе при использовании нового признака $f(x)$ признак x нужно обязательно исключить из массива данных, но при подходящем варианте регрессионного или дискриминантного анализа этого не нужно делать - ненужный признак исключается в ходе вычислений.

Первый шаг при анализе корреляционной матрицы - это ор-

ганизация распечатки в виде квадратной таблицы. Распечатки пакета САИСИ выводят в основном четырехугольные матрицы, которые по техническим причинам (при большом количестве признаков) разделены на отдельные куски. Эти куски нужно вырезать и так склеивать, чтобы образовалась квадратная симметричная матрица, столбцы и строки которой определены одним и тем же списком признаков.

На следующем шаге полезно дать общую оценку о характере связей - являются ли они "сильными" или "слабыми". Естественно, что такая оценка относительна и зависит от области исследования - связь, которая для характеризации технологического процесса является "слабой", в социологии может оказаться "сильной". При выработке общей оценки полезно выделить (например, разноцветными карандашами) коэффициенты корреляции, абсолютные значения которых находятся в разных интервалах. Наиболее часто используемые интервалы следующие: меньше 0.3-х, от 0.3 до 0.5; от 0.5 до 0.7; от 0.7 до 0.9; больше 0.9-и. В качестве приблизительного правила можно теперь использовать следующую классификацию:

I) максимальные корреляции всех признаков удовлетворяют условию $|r_{ij}| \leqslant 0.3$ - очень слабые связи;

2) большинство коэффициентов корреляции удовлетворяют условию $|r_{ij}| \leqslant 0.3$, но встречаются и признаки, наибольшие коэффициенты корреляции которых удовлетворяют условию $0.3 < |r_{ij}| \leqslant 0.5$ - слабые связи;

3) большинство коэффициентов корреляции удовлетворяют условию $0.3 < |r_{ij}| \leqslant 0.5$, а некоторые - условию $0.5 < |r_{ij}| \leqslant 0.7$ - средние связи;

4) большинство коэффициентов корреляции удовлетворяют

условию $0.5 < |r_{ij}| \leqslant 0.7$, а некоторые - условию $0.7 < |r_{ij}| \leqslant 0.9$ - сильные связи;

5) большинство коэффициентов корреляции удовлетворяют условию $0.7 < |r_{ij}| \leqslant 0.9$, а некоторые - условию $0.9 < |r_{ij}|$ - очень сильные связи.

Конечно, можно найти матрицы корреляций, для которых эта классификация неприменима. Так, например, могут почти все корреляции быть слабыми, но встречаться некоторые очень сильные связи ($|r_{ij}| > 0.9$). В этом случае мы имеем дело с парами дублирующих друг друга признаков и обычно в дальнейшей обработке правильным будет использовать из каждой такой пары только один признак. То, что в корреляционной матрице много слабых связей, не означает еще, что связи вообще слабые. В этом случае могут встречаться несколько групп сильно связанных признаков, которые между собой коррелированы.

Для описания корреляционных матриц можно использовать еще графы корреляции.

При анализе корреляционной матрицы можно выдвинуть следующие вопросы:

I) какие признаки имеют наибольшее значение, т.е. наиболее сильно связанные с остальными?

2) какие сильно связанные группы признаков встречаются в обрабатываемым массиве?

3) какие дублирующие пары (тройки) признаков встречаются в обрабатываемом массиве?

4) какие признаки практически не зависят от остальных?

5) встречаются ли сильно связанные группы признаков, которые между собой практически независимы?

Такой анализ может оказаться очень полезным при даль-

нейшей обработке.

Нужно обращать внимание также и на взаимоотношение числа признаков и числа объектов. "Правилом кулака" в многомерном анализе называется условие, при котором число изучаемых признаков должно быть хотя бы в два-три раза меньше числа изучаемых объектов. В общем случае имеет место то же требование и для корреляционной матрицы - при n объектах, в принципе, может быть только n независимых признаков, остальные обязательно являются линейными комбинациями этих признаков. Этот факт очень часто не принимается заказчиком - все признаки будто-бы совсем самостоятельные. Дело здесь в том, что при большом количестве признаков и маленьком числе объектов в данной конкретной выборке возникают разные псевдосвязи - по содержанию бессмысленные, но характеризующие конкретные наблюдаемые объекты. Например, в изучаемой выборке могут все 35-летние женщины носить очки или все мужчины с высшим образованием быть бородатыми.

Естественно возникает вопрос, можно ли учитывая такой эффект, вообще изучать корреляционные матрицы, где число объектов не больше числа признаков? Хотя разные авторы выражают здесь разные мнения, можно все-таки (учитывая опыт практической работы) утверждать, что для общей характеризации материала можно использовать и сверхбольшие корреляционные матрицы, где число признаков превышает число объектов. Причина здесь в том, что в большинстве случаев псевдосвязи можно выяснить при помощи содержательного анализа связей. Такие связи будут мешать тогда, когда их формально используют в выражениях, связывающих несколько признаков. Таким образом, всякие статистические процедуры, которые исходят

из корреляционной матрицы (факторный, канонический, регрессионный и дискриминантный анализ) требуют обязательно, чтобы число объектов превышало число признаков. При описании связей между признаками это требование может и не выполняться.

Если корреляционную матрицу выводят в самостоятельный файл для дальнейшего ее использования в многомерном анализе, то ее нужно вычислить по полной выборки. В противном случае она может не быть неотрицательно определенной и дальнейший анализ может в этом случае привести к некорректным результатам.

IQ.I.2. Процедурная карта

При обращении к процедуре сообщают, корреляционные коэффициенты каких признаков нужно вычислить. Общий вид карты является следующим:

PEARSON CORR varlist [WITH varlist/varlist...varliat/]

В управляющем поле карты находится имя процедуры. Как обычно, можно указать только первые восемь символов управляющего слова, т.е. PEARSON. Если в поле описаний написан только список признаков (ключевое слово WITH не используется), то для признаков, перечисленных в списке, вычисляется квадратная корреляционная матрица. Если в поле описаний написано два списка признаков, разделенных ключевым словом WITH, то коэффициенты корреляции выводятся только для таких пар признаков, где первый признак взят из первого списка, а второй - из второго списка. После наклонной черты могут следовать новые списки или пары списков, разделенные ключевым словом WITH. Таким образом, одной процедурной картой можно заказать несколько корреляционных матриц.

IO.I.3. Опции

Основная опция процедуры выводит заказанные корреляционные коэффициенты, число объектов, используемых при вычислении данного коэффициента и вероятности значимости для одностороннего теста. При вычислении коэффициента корреляции используются те объекты, у которых имеются значения обоих используемых признаков, а значения остальных признаков могут отсутствовать. В каждой строке (и столбце) матрицы размещены коэффициенты корреляции некоторого определенного признака с остальными признаками, перед строкой (над столбцом) указано имя этого признака. За картой PEARSON CORR может следовать карта OPTIONS, в поле описаний которой перечислены номера желаемых опций, разделенные запятой. Возможны следующие опции:

I) обозначение пропуска используется как значение признака;

2) для вычисления коэффициентов корреляции используются только те объекты, у которых имеются значения всех признаков данного заказа (обработка полной выборки);

3) вероятности значимости выводятся для двухстороннего теста;

4) от квадратной корреляционной матрицы образуют самостоятельный файл, который можно записывать; адрес выводимого файла определяют оператором DD с именем FT09F00;

5) отсутствует;

6) выводятся только те элементы корреляционной матрицы, которые находятся под главной диагонали (трехугольная форма корреляционной матрицы).

10.1.4. Статистики

За картой PEARSON CORR может следовать карта STATISTICS, в поле описании которой перечисляют номера желаемых дополнительных статистик, разделяя их запятой. Возможные дополнительные статистики следующие:

1) для каждого признака выводят среднее значение и стандартное отклонение;

2) для каждой пары признаков печатают сумму произведений отклонений от среднего значения и ковариацию.

Если нужно вычислить все дополнительные статистики, в поле описаний можно написать ключевое слово ALL.

10.1.5. Пример

Предположим, что в заказе нужно вычислить корреляционную матрицу признаков X1,..., X9. Нужно использовать полную выборку. Так как отсутствует информация, с каким знаком могут быть коэффициенты корреляции, вероятности значимости вычисляют для двухстороннего теста. Требуется еще вывод всех средних значений и стандартных отклонений. Кроме этой корреляционной матрицы нужно вычислить еще коэффициенты корреляции признаков ВОЗРАСТ, ДОХОД, ЖИЛПЛ с признаками ИНДЕКС1, ИНДЕКС2, ... , ИНДЕКС12. Дополнительные условия точно такие же, как и для предыдущей корреляционной матрицы. Заказ представляется картами:

```
PEARSON CORR  X1 TO X9/ ВОЗРАСТ, ДОХОД, ЖИЛПЛ WITH ИНДЕКС1
              TO ИНДЕКС12/
OPTIONS       2, 3
STATISTICS    1
```

I0.2. Процедура NONPAR CORR

I0.2.I. Цель и методические указания

Процедура NONPAR CORR вычисляет для упорядоченных или численных признаков корреляционную матрицу (или часть ее), элементами которой являются коэффициенты корреляции Спирмэна или Кэндалла. Указывают число объектов, используемых в вычислениях и вероятность значимости. Для работы процедуры нужно, чтобы упорядоченные признаки были закодированы числами.

При вычислении коэффициента корреляции Спирмэна вместо значениями признаков используются их ранги - номера порядка значений в упорядоченной выборке. При вычислении вероятности значимости не нужно предположение о нормальном распределении признаков; именно это обстоятельство оправдывает иногда использование этого коэффициента для численных признаков. Использование вероятности значимости точно такое же, как это было описано в пункте I0.I.I.

Монотонное преобразование признаков не может изменить значения коэффициента корреляции Спирмэна, поэтому не имеет смысла одновременно изучать некоторый признак и его монотонную функцию ($\ln x$, e^x и т.д.). Возможные значения коэффициента находятся между значениями -1 и 1. Значение 1 коэффициент приобретает тогда, когда все значения изучаемой пары признаков имеют одинаковые ранги. Если ранги значений упорядочены в противоположном порядке, то коэффициент приобретает значение -1. При независимых признаках значение коэффициента корреляции равно нулю.

Коэффициент корреляции Кэндалла и его свойства описаны в пункте 9.I.I.

Коэффициенты корреляции, вычисленные процедурой NONPAR CORR, можно использовать в описательных задачах точно так же, как линейные коэффициенты корреляции (см. пункт I0.I.I), но образованную корреляционную матрицу нельзя использовать в качестве начальных данных для остальных процедур многомерного анализа. Хотя в литературе встречаются редкие примеры применения факторного анализа к матрице непараметрических коэффициентов корреляций, этот прием не является математически обоснованным и поэтому считается некорректным.

I0.2.2. Процедурная карта

При обращении к процедуре сообщают, какие коэффициенты корреляции нужно вычислить. Общий вид процедурной карты следующий:

NONPAR CORR varlist [WITH varlist/ varlist...varlist/]

В поле управления карты находится имя процедуры. Как обычно, можно указать только первые 8 символов, т.е. NONPAR C. Если в поле описаний находится только список признаков - ключевое слово WITH не используется - вычисляют коэффициенты корреляции для всевозможных пар, которые можно образовать из перечисленных признаков. Если в поле описаний находятся два списка признаков, разделенных ключевым словом WITH, вычисляются коэффициенты корреляции только для таких пар, где первый признак взят из первого списка, а второй - из второго. После наклонной черты могут следовать новые списки признаков или пары списков, разделенные ключевым словом WITH. Таким образом, одним обращением к процедуре можно заказать несколько различных корреляционных матриц.

I0.2.3. Опции

Основная опция процедуры NONPAR CORR выводит коэф-

фициенты корреляции Спирмэна. Они выводятся не таблицей, а строками, в которых находятся 6 коэффициентов корреляции. Для каждой пары указаны имена соответствующих признаков, значение коэффициента, число использованных объектов (N) и вероятность значимости (SIG). За картой NONPAR CORR может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций, разделенные запятой. Возможные опции следующие:

1) обозначение пропусков используются как значения признаков;

2) при вычислении коэффициента корреляции используются только те объекты, у которых имеются значения всех признаков данного заказа (обработка полной выборки);

3) вероятности значимости выводятся для двухстороннего теста;

4) из корреляционной матрицы (при ее заказе нельзя использовать ключевое слово WITH) образуют самостоятельный файл, адрес которого определяется оператором DD с именем FT09F001;

5) вычисляются коэффициенты корреляции Кендалла;

6) вычисляются как коэффициенты корреляции Кендалла, так и коэффициенты корреляции Спирмэна. Оба выводятся по обычной схеме, в двух разных таблицах.

10.2.4. Статистики

Дополнительно выбираемых статистик эта процедура не имеет.

10.2.5. Пример

Предположим, что нужно вычислить корреляционную матрицу для упорядоченных признаков X1, X2,..., X30, причем ис-

пользуются коэффициенты корреляции Кэндалла. Вероятности значимости нужно найти для двухстороннего теста. Заказ представляется следующими картами:

```
NONPAR CORR     X1 TO X30/
OPTIONS         3, 5
```

10.3. Процедура **PARTIAL CORR**

10.3.1. Цель и методические указания

Процедура **PARTIAL CORR** предназначена для вычисления такой корреляционной матрицы, элементами которой являются частные коэффициенты корреляции для численных признаков.

Объясним кратко содержательное значение коэффициента частной корреляции. Дело в том, что корреляция между двумя признаками часто может быть обусловленной влиянием некоторого третьего признака (или влиянием группы некоторых признаков). Анализ таких связей может дать странные результаты, непосредственная интерпретация которых приводит к неправильным заключениям. Пусть, например, для нескольких лет измерено количество солнечных дней и средняя урожайность. Вычислен коэффициент корреляции этих признаков, который оказался довольно большим по абсолютному значению, но отрицательным. Сделать из этого вывод, что солнце влияет на урожайность отрицательно, было бы все-таки преждевременно - годы несравнимы между собой по другим условиям. Годы могут существенно различаться по количеству осадков, что очень сильно влияет на урожайность. Так в общем урожайность положительно коррелирована с количеством осадков, но количество осадков отрицательно коррелировано с количеством солнечных дней. Таким образом, сильное влияние количества осадков с одной стороны на количество урожайность, а с другой стороны на

количество солнечных дней, обусловливает отрицательную корреляцию этих признаков. Чтобы элиминировать влияние количества осадков, нужно изучать годы с одинаковым значением этого признака. Если поступать так, окажется, что урожайность положительно коррелирована с количеством солнечных дней.

Для исключения влияния некоторых факторов или мешающих признаков очень хорошо подходят частные коэффициенты корреляции $r(xy|z)$ между признаками X и Y в отношении признака z (влияние признака z исключено) - это коэффициент корреляции между остатками прогноза $\hat{x}$ и $\hat{y}$, где использованы линейные прогнозы при помощи признака z ($\hat{x} = X - a - bz$, $\hat{y} = Y - c - dz$). Таким же способом можно вычислить частные коэффициенты корреляции, исключая влияние нескольких признаков одновременно.

При использовании частных коэффициентов корреляции часто используемые цели следующие:

I) исключение влияния мешающих признаков. Возникает в случае, если изучаемый материал может быть неоднородным - его собрали в разных годах, пациенты имеют различный вес и возраст и т.д. В случае, если признаки год, возраст, вес не представляют самостоятельного интереса, было бы правильно при описании связей между признаками использовать матрицу частных коэффициентов корреляции, где влияние этих признаков исключено;

2) анализ структуры связей между признаками. Возникает в случае, если изучают влияние некоторой группы признаков на признак-функцию. Исключая признаки один за другим, можно при помощи частных коэффициентов корреляции выяснить непосредственные влияния признаков, совместные влияния нескольких

признаков и "косвенные" влияния некоторых признаков через другие признаки.

Вместе с частными коэффициентами корреляции печатают вероятности значимости. Но в отличие от двух предыдущих процедур, выводятся не число объектов, которые были использованы при вычислении, а число степеней свободы (число использованных признаков - 2 - порядок частной корреляции). При использовании вероятностей значимости нужно учитывать то же ограничение, как в случае обычных линейных корреляций: признаки должны иметь нормальное распределение. Использование вероятностей значимости происходит так же, как это описано в пункте I0.I.I.

I0.3.2. Процедурная карта

При обращении к процедуре сообщают, для каких признаков частные коэффициенты корреляции нужно вычислить и влияние каких признаков нужно исключить. Общий вид процедурной карты следующий:

```
PARTIAL CORR   varlist [WITH varlist] BY varlist (n1,...,nk)/
               [.../ varlist [WITH varlist] BY varlist
                (m1,...,mk)/]
```

В поле управлений карты записывают имя процедуры. Как обычно, можно ограничиться только первыми восемью символами управляющего слова PARTIAL. Списки признаков в поле описаний имеют две различных цели-списки, предшествующие ключевому слову BY определяют, для каких пар признаков нужно вычислить корреляции; список, следующий за ключевым словом BY определяет, влияние каких признаков нужно исключить. Как в предыдущих двух процедурах вычисляемые корреляции можно определить двумя способами - используя ключевое слово WITH и без

его использования. Если ключевому слову BY предшествует только один список признаков, частные коэффициенты корреляции вычисляются для всевозможных пар этого списка. Если ключевому слову BY предшествуют два списка разделенных ключевым словом WITH, то коэффициенты частных корреляции вычисляются только для таких пар, где один признак берется из первого списка, а другой - из второго. За ключевым словом BY следует список признаков, влияние которых нужно исключить, за которым в скобках следует целое число или список целых чисел. Этим числом (или числами) определяется порядок исключения; можно использовать до пяти разных чисел. Числа в скобках уточняют, каким способом нужно исключать влияние признаков. Если, например, порядок исключения единица, вычисляют столько матриц коэффициентов частных корреляций, сколько признаков было перечислено после ключевого слова BY. При этом в каждой из них исключено влияние одного признака. Если, например, порядок исключения два, то вычисляются частные коэффициенты корреляции, исключая влияние всевозможных пар исключаемых признаков и т.д. Так при выполнении заказа

PARTIAL CORR X1,X2,X3 BY T1,T2,T3 (1,2)/

выводят шесть разных матриц частных коэффициентов корреляции - во-первых, исключая по одному влияние признаков T1, T2, T3, а потом исключая по парам влияние признаков T1 и T2, T1 и T3, T2 и T3.

Заказ кончается наклонной чертой. После нее могут следовать новые заказы.

I0.3.3. Опции

Основная опция процедуры использует в вычислениях только те объекты, у которых имеются значения всех признаков

используемых в данном заказе, т.е. опция работает с полной выборкой. Значения частных коэффициентов корреляции выводятся в виде таблицы, указываются вероятности значимости, вычисленные для одностороннего теста и степени свободы. Перед таблицей печатается список исключаемых признаков.

За картой PARTIAL CORR может следовать карта OPTIONS, в поле описаний которой перечисляют номера желаемых опции, разделяя их запятой. Возможные опции следующие:

1) обозначения пропусков используются как значения признака;

2) при вычислении частного коэффициента корреляции используются все объекты, которые имеют значения используемых признаков, значения других признаков, встречающихся в заказе, могут отсутствовать;

3) вероятность значимости вычисляется для двустороннего теста;

4) в качестве начальных данных используется корреляционная матрица, адрес которой определяется оператором DD с именем FTØ8FØØ1;

5) для каждого заказа вводится самостоятельная корреляционная матрица;

6) для всех заказов используется одна корреляционная матрица, порядок признаков в которой перечисляется картой VARIABLE LIST;

7) совпадает с основной опцией;

8) выводятся только коэффициенты частных корреляций, находящие над главной диагонали; они представляются не таблицей, а строками, включающими шесть коэффициентов.

Как мы увидим, в качестве начальных данных процедуры PARTIAL CORR кроме обычной объект-признак-матрицы можно использовать и заранее вычисленные корреляционные матрицы. Такая возможность существенно уменьшает объем работы, нужный для вычисления частных корреляции. При использовании корреляционной матрицы возможны два варианта действия. Для опции 5 нужно для каждого заказа, который записан на процедурной карте, ввести отдельную корреляционную матрицу, в которой порядок признаков должен соответствовать порядку их перечисления в списке корреляции и исключения. Для опции 6 нужно для всех заказов на данной процедурной карте ввести одну корреляционную матрицу, имена и порядок признаков в ней представляются картой VARIABLE LIST. При этой опции нужно в заказе использовать еще следующие управляющие карты:

1) VARIABLE LIST - на карте перечисляются имена всех используемых признаков, порядок их перечисления должен совпадать с порядком признаков в корреляционной матрице;

2) INPUT MEDIUM - на карте указывают, на каком носителе информации находится вводимая корреляционная матрица. Адрес матрицы, находящейся на магнитной ленте или магнитном диске уточняют оператором DD с именем FTØ8FØØ1;

3) # OF CASES - на карте указывают число объектов, которые были использованы для вычисления данной корреляционной матрицы (эти числа нужны при вычислении вероятностей значимости);

4) INPUT FORMAT - на карте уточняют формат вводимой матрицы. Коэффициенты корреляции нужно представить в стандартном формате;

5) OPTIONS - на карте обязательно должны быть указаны опции 4 и 6;

6) READ MATRIX - карта, которая информирует систему о надобности ввести начальные данные. Должна обязательно следовать за картой STATISTICS (или при ее отсутствии - за картой OPTIONS);

7) PROCESS SERIES - карта необязательна; ее используют, если подфайлы обрабатываются параллельно, она тогда должна предшествовать карте PARTIAL CORR. В этом случае и корреляционные матрицы должны быть вычислены для соответствующих групп подфайлов.

I0.3.4. Статистики

За картой PARTIAL CORR может следовать карта STATISTICS, на которой перечисляются номера желаемых дополнительных статистик, разделяя их запятой. Возможные статистики следующие:

I) выводится корреляционная матрица для всех используемых объектов и вероятность значимости;

2) для каждого признака выводятся средние значения, стандартные отклонения и числа присутствующих значений;

3) корреляционная матрица выводится только в том случае, если некоторый коэффициент корреляции невозможно вычислить (нет объектов, у которых значения признаков были бы измерены одновременно или один из признаков был постоянный).

Отсутствующие коэффициенты обозначаются символом 99.

Если считается нужным вычислить все статистики одновременно, то в поле описаний карты STATISTICS можно написать слово ALL.

I0.3.5. Пример

Предположим, что нужно вычислить частные корреляции для признаков X1, X2 и X3, исключая сперва влияние одного из признаков T1 или T2, а потом влияние обоих. Корреляционная матрица всех используемых признаков вводится с перфокарт. Заказ представляется следующими картами:

```
RUN NAME        ВЫЧИСЛЕНИЕ ЧАСТНЫХ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ
VARIABLE LIST   X1 TO X3, T1, T2/
INPUT MEDIUM    CARD
# OF CASES      211
INPUT FORMAT    FIXED(5F1Ø.7)
TASK NAME       ПРИМЕР
PARTIAL CORR    X1 TO X3 BY T1, T2(1,2)/
OPTIONS         4, 6, 8
STATISTICS      3
READ MATRIX
     1.Ø      Ø.6774   Ø.4112   Ø.4568  -Ø.2
     Ø.6774   1.Ø      Ø.1696   Ø.7628   Ø.Ø5
     Ø.4112   Ø.1696   1.Ø      Ø.3173  -Ø.3
     Ø.4568   Ø.7628   Ø.3173   1.Ø      Ø.165
   - Ø.2      Ø.Ø5     -Ø.3     Ø.165    1.Ø
FINISH
```

XI. ОПИСАТЕЛЬНЫЕ МЕТОДЫ МНОГОМЕРНОГО АНАЛИЗА

К числу описательных методов многомерного анализа относится факторный анализ (процедура FACTOR), канонический анализ (процедура CANCORR), группировка (процедура MULTIVARIANT) и шкалирование по Гуттману (процедура GUTTMAN SCALE).

II.I. Процедура FACTOR

II.I.I. Цель и методические указания

II.I.I.I. Общая модель факторного анализа. Целью факторного анализа является построение новых признаков, т.н. факторов для одновременного описания заданных исходных признаков. При помощи факторов часто возможно в сжатом и обозримом виде описать структуру зависимостей исходных признаков.

При построении факторов предполагается, что каждый исходный признак представим в форме суммы известной линейной комбинации факторов и независимого от них слагаемого - особенности, т.е.

$$X_i = W_{i1}F_1 + W_{i2}F_2 + \ldots + W_{ik}F_k + e_i, \qquad (II.I)$$

где W_{ij} - вес фактора F_j в линейной комбинации, определяющей признак X_i, а e_i - особенность признака X_i. Как правило, число факторов k меньше исходного числа признаков m.

В ходе факторного анализа находится матрица факторных весов W(FACTOR MATRIX) - это таблица, в i-ой строке j-ом столбце которой находится элемент W_{ij}. Кроме того, находится общность (коммуналитет) каждого признака h_i^2(COMMUNALITY) - это та часть вариабельности признака, которая описывается при помощи факторов.

При проведении факторного анализа все исходные признаки

нормируют и центрируют. Тогда общая вариабельность признаков - (сумма их дисперсий) равняется числу признаков.

При нахождении факторов можно пользоваться разными критериями и разными математическими методами, при этом получаются разные факторные модели. В процедуре FACTOR реализованы пять методов факторного анализа (за названием метода в скобках указано его символическое обозначение в процедуре FACTOR):

I) метод главных компонент (PA1);

2) классический факторный анализ (PA2);

3) каноническая факторизация Рао (RAO);

4) альфа-факторизация (ALPHA);

5) имедж-факторизация (IMAGE).

II.I.I.2. Метод главных компонент. При методе главных компонент выбирают главные компоненты - известные линейные комбинации исходных признаков, удовлетворяющие условию $w_{ij}^2 = 1$, $(j=1,2,\ldots,m)$. Первая главная компонента - это такая линейная комбинация исходных признаков, дисперсия которой имеет максимальное возможное значение (говорят, что она описывает по возможности большую часть из общей вариабельности признаков). Вторая главная компонента некоррелирована с первой и определяет максимальную возможную часть из остаточной вариабельности (т.е. имеет максимальную дисперсию) и т.д. В принципе число главных компонент равняется числу исходных признаков.

Математически нахождение главных компонент сводится к нахождению собственных векторов и собственных значений корреляционной матрицы, причем дисперсия k-ой главной компоненты равняется k-му собственному значению (предполагается,

что собственные значения упорядочены в убывающую последовательность).

В качестве факторов выбираются k первых главных компонент остальные образуют особенности. Для описания q-ой главной компоненты опускаются q-ые собственные значения (EIGENVALUE) корреляционной матрицы, отношение q-го собственного значения к сумме собственных значений, выраженной в %-ах (PCT OF VAR) и отношение суммы первых q собственных значений к сумме всех собственных значений (CUM PCT OF VAR) - это есть процент описания всех исходных признаков через q первых факторов.

По умолчанию факторами выбираются те главные компоненты, которым соответствуют собственные значения больше единицы. Не всегда это правило дает хорошо интерпретируемую модель. По желанию число факторов, вычисляемых в модель, можно регулировать двумя разными способами - либо определяя параметром NFACTORS число включаемых в модель факторов, либо задавая параметром MINEIGEN минимальное допустимое для данной модели значение дисперсии фактора.

Факторную модель строят через матрицы W факторных весов и через общности признаков. По своему содержанию общность в этой модели равняется множественному коэффициенту корреляции при прогнозировании данного признака по имеющемуся комплекту признаков, показывая таким образом, какая часть из вариабельности признака описывается через факторы. Факторный вес w_{ij} по своему содержанию есть коэффициент корреляции между признаком X_i и фактором F_j.

II.I.I.3. Классический факторный анализ. При классическом факторном анализе исходят из основной формулы (II.I) и

из предположения, что корреляции между исходными признаками обусловлены $k (k<m)$ факторами, т.е.

$$r_{ij} = w_{i1}w_{j1} + w_{i2}w_{j2} + \dots + w_{ik}w_{jk}, \qquad (II.2)$$

дисперсии нормированных исходных признаков являются суммами двух слагаемых: общности h_i^2, определенной через факторы.

$$h_i^2 = w_{i1}^2 + w_{i2}^2 + \dots + w_{ik}^2 \qquad (II.3)$$

и дисперсии d_i^2 особенности e_i:

$$d_i^2 = 1 - h_i^2.$$

Для определения факторных весов пользуются т.н. редуцированной корреляционной матрицей, отличающейся от обыкновенной корреляционной матрицы тем, что на ее главной диагонали вместо единиц расположены общности $h_i^2 (h_i^2 \leqslant 1)$, (см. [3], стр. II6).

Классический факторный анализ методом главных компонент в том и состоит, что метод главных компонент применяется к редуцированной матрице. Значения общностей можно получить из некоторой априорной информацией, их можно оценить на основании корреляций исходных признаков (например, выбирая $h_i = \max_j r_{ij}$) или пользуясь квадратом коэффициента множественного коэффициента корреляции R_i^2 (определенного при прогнозе признака X_i по всем остальным). Величину R_i^2 можно определить по формуле

$$R_i^2 = 1 - 1/r^{ii}, \qquad (II.4)$$

где r^{ii} – i-ый диагональный элемент обратной корреляционной матрицы $R^{-1} = (r^{ij})$, выпускаемой в виде дополнительной статистики.

Общности, находящиеся на главной диагонали, задаются

посредством параметра DIAGONAL, причем это возможно при методе PA1.

Следует однако заметить, что обычно вычисленные таким образом факторные веса не удовлетворяют условию (II.3) при общностях, выбранных любым методом. Поэтому обычно при выполнении классического факторного анализа применяется итеративная процедура, в ходе которой приближение общностей после вычислении факторных весов методом компонентного анализа уточняется при помощи равенства (II.3). Найденное новое приближение общностей выбирается за основу нового шага компонентного анализа и т.д. Такой итеративный процесс реализован в процедуре PA2, где по умолчанию общностями берутся величины R_i^2(I2.4). Число итерации и точность полученной общности определяются параметрами ITERATE (молчанием 25) и STOPFACT (молчанием 0.00I).

При методе PA2 исходные приближения параметром DIAGONAL задавать невозможно, они автоматически выбираются равным величинам R_i^2.

II.I.I.4. Интерпретация результатов факторного анализа. После нахождения факторной модели исследователю придется приступить к интерпретации найденной факторной модели. Следует сказать, что это - довольно интуитивная деятельность. При интерпретации факторов исходят из матрицы факторных весов - это коэффициенты корреляции между признаками и факторами. Для этого можно задавать некоторую (субъективную) категоризацию коэффициентов корреляции, которая, разумеется, в некоторой мере зависит от объема выборки. При достаточно большом объеме (не менее нескольких сотен наблюдений) рекомендуется следующая категоризация:

1) $|w_{ij}| < 0.3$ - незначимая зависимость,

2) $0.3 \leqslant |w_{ij}| < 0.5$ - значимая, но слабая зависимость,

3) $0.5 \leqslant |w_{ij}| < 0.7$ - зависимость среднего уровня,

4) $|w_{ij}| \geqslant 0.7$ - сильная зависимость.

При меньшем объеме выборки слабые зависимости становятся незначимыми, при весьма больших объемах значимыми могут быть и меньшие зависимости (но они будут весьма слабыми).

На следующем шаге для каждого фактора составляется список коррелированных с ним признаков, причем обязательно учитывается и знак корреляции w_{ij}. На основании этого списка выдвигается гипотеза о содержании фактора, обычно это сделается посредством присваивания фактору соответствующего имени. При именовании факторов обязательно учитывается, что они некоррелированы.

II.I.I.5. Вращение факторной модели. Полученная таким образом факторная модель может оказаться трудно интерпретируемой, так как факторы обычно коррелируются с многими различными признаками. Поэтому обыкновенно найденную путем факторного анализа модель еще не рассматривают как конечный результат, а она заменяется на новый комплекс факторов, который с математической точки зрения является эквивалентным с первоначальным (сохраняются общности и процент описания общей вариабельности), но лучше интерпретируется. Такое преобразование называется вращением факторов.

В процедуре реализованы три метода ортогонального вращения (сохраняется ортогональность факторов в пространстве признаков) - это VARIMAX, QUARTIMAX и EQUIMAX, из которых наиболее быстрым и часто применяемым является VARIMAX. Остальные два более трудоемки, но как правило, по результату

от первого не отличаются.

Если число факторов, включенных в модель, слишком мало, то факторы трудно интерпретируются и после вращения - все же с каждым фактором коррелируются многие различные признаки. При увеличении числа факторов признаки разбиваются на более компактные группы, связанные с отдельными факторами. В случае слишком большого числа факторов появляются специальные факторы отдельных признаков - факторы, связанные только с одним признаком, которые уже никакой дополнительной информации не дадут.

Для нахождения подходящей модели целесообразно заказать подрядь много факторных моделей с разным числом факторов и выбрать среди них такой, который интерпретируется наилучшим образом. Никогда нельзя из вращенной модели некоторые факторы просто выбросить - это некорректно.

После вращения методом VARIMAX поведение модели по некоторой мере зависит от объема выборки. При большой выборке обычно наступает момент, когда прибавление нового фактора мало влияет на имеющуюся факторную структуру. При маленьких выборках такой стабильности может и не быть - прибавлением нового фактора пока хорошо интерпретируемая факторная модель становится бессмысленной - это результат индивидуальных особенностей отдельных объектов выборки.

Иногда по априорной (содержательной) информации можно предполагать, что факторы, обуславливающие зависимость между исследуемыми признаками, не ортогональны. В таком случае следует пользоваться неортогональным вращением (OBLIQUE). При неортогональном вращении факторные веса w_{ij} уже не являются коэффициентами корреляции между признаком и фактором.

Поэтому в случае неортогонального поворота кроме матрицы факторных весов (FACTOR PATTERN) выпускается и матрица коэффициентов корреляции между признаками и факторами (FACTOR STRUCTURE). При интерпретации факторов необходимо взять за основу последнюю матрицу.

В качестве дополнительной статистики, облегчающей интерпретацию факторов (или вращенных факторов) можно заказать графическое представление признаков в k-мерной системе координат, определенной факторами (если их число есть k) в форме проекций на всевозможные координатные плоскости; таким образом, на координатной плоскости F_i/F_j признак X_h изображается в виде точки (w_{hi}, w_{hj}). Разумеется, что число таких графиков для k факторов равняется $k(k+1)/2$.

II.I.I.6. Индивидуальные факторные веса. Факторные веса определяют выражение исходных признаков через факторы. В то же время можно выдвинуть и задачу определения факторов через исходные признаки.

В случае главных компонент эта задача решается точно - ведь каждая главная компонента определяется как линейная комбинация исходных признаков.

В случае классического факторного анализа точного выражения факторов через исходные признаки невозможно найти, так как неизвестны особенности e_i (в ходе факторного анализа оцениваются только их дисперсии). Поэтому здесь для выражения факторов через исходные признаки пользуются их оценками методом наименьших квадратов. Таким же методом надо пользоваться и в случае компонентного анализа, если число компонент k меньше числа факторов, причем факторы могут быть и повернутыми.

Коэффициенты, определяющие факторы как линейные комбинации исходных признаков, выпускаются в таблице с заглавием FACTOR SOORE COEFFICIENTS. При помощи этих соотношений для всех объектов можно вычислить значения (или их оценки) всех факторов. Их можно выпускать в качестве самостоятельного файла или присоединить к системному файлу в качестве значений новых признаков, которыми можно пользоваться на следующих этапах анализа.

Коэффициентами линейных комбинаций, определяющими факторы, можно пользоваться и при интерпретации факторов, поскольку в данном случае мы имеем дело с регрессионной зависимостью для прогнозирования фактора по исходным признакам (см. [4] , стр. 218-220).

II.I.I.7. Другие методы факторного анализа. Кроме основных методов факторного анализа PA1 и PA2 в процедуре FACTOR реализовано еще три метода факторного анализа. Они значительно реже применяются и для сознательного пользования ими необходимо познакомиться с ними при помощи специальной литературы (см. [7] , [5] и [6]). Здесь мы очень бегло охарактеризуем особенности этих методов.

Метод канонической факторизации Рао определяет факторы как такие линейные комбинации исходных признаков, которые имеют максимальные канонические корреляции с исходными признаками. При этом методе существует тест, позволяющий решать статистической значимости каждого фактора.

При методе альфа-факторизации имеется такой статистический подход, при котором исходные признаки рассматривают как выборку из генеральной совокупности признаков. Факторы определяются таким образом, чтобы они были максимально кор-

релированы с факторами, характеризующими генеральную совокупность признаков.

Методом имедж-факторизации факторы определяются так, чтобы они определили максимальную часть из той части каждого признака, которая прогнозируема через остальные признаки. Это получается по существу таким образом, что общности постулируются равными величинами R_i^2. В случае этого метода допущены коррелированные факторы.

Классический факторный анализ (метод PA2), и рассмотренные здесь его модификации (методы RAO, ALPHA и IMAGE) могут не иметь действительного приближения (при заданном числе факторов), т.е. в результате итерации получается модель, в которой некоторые факторные веса W_{ij} или некая общность h_i^2 по абсолютной величине больше единицы. В таком случае либо следует изменить число факторов, либо пользоваться методом компонентного анализа (PA1) с единицами на диагонали, при котором такого обстоятельства не возникает.

II.I.2. Процедурная карта

При обращении к процедуре FACTOR указывают используемые признаки, методы факторизации и параметры, уточняющие работу этих методов. Общий вид управляющей карты следующий:

FACTOR VARIABLES=varList/[TYPE={PA1|PA2|RAO|ALPHA|IMAGE|BYPASS}/]
[DIAGONAL=список значений/]
[NFACTORS=целое число/] [MINEIGEN=значение/]
[ITERATE=целое число/] [STOPFACT=значение/]
[ROTATE={VARIMAX|QUARTIMAX|EQUIMAX|OBLIQUE|NOROTATE}/]
[DELTA=начальное значение[,приращение,конечное значение]/]
[FACSCORE/]

В управляющем поле карты находится имя процедуры. В поле

описаний первым написано единственное обязательное ключевое слово VARIABLES, за которому после знака равенства следует список признаков, используемых при выделении факторов. За этим списком следует наклонная черта. Наклонная черта разделяет и все следующие необязательные параметры.

За ключевым словом TYPE указывают метод, используемый для построения факторов. Последним в списке выбираемых методов находится ключевое слово BYPASS. При использовании этого ключевого слова факторы не выделяются, а вычисляется только корреляционная матрица для используемых признаков. Вместе с этим ключевым словом нужно обязательно заказать статистику 3, иначе корреляционная матрица не будет выведена. При отсутствии ключевого слова TYPE используют метод факторизации PA2.

Необязательный параметр DIAGONAL используют для указания первоначальных оценок общностей при методе PA1 . После знака равенства следуют оценки общностей, перечисленных значений должно быть столько, сколько было использовано признаков; разделяются они или пробелом, или запятой. Например, при наличии шести признаков за параметром DIAGONAL может следовать список DIAGONAL = Ø.8 Ø.75 Ø.83 Ø.6 Ø.6 Ø.6/. Если в списке признаков встречаются подряд несколько одинаковых значений, то для их представления можно использовать конструкцию nx значение, где звездочке предшествует число повторений, а следует - повторяемое значение. Таким образом можем предыдущий список представить в виде DIAGONAL = Ø.8 Ø.75 Ø.83 3 * Ø.6/ .

Следующие два параметра дают возможность определить число выделяемых факторов при методах PA1 и PA2. Параметр

NFACTORS непосредственно указывает число факторов, параметр MINEIGEN определяет минимальное значение собственного значения, при котором соответствующий фактор еще включается в модель. Эти параметры могут встретиться, например, в виде NFACTORS = 3 или MINEIGEN = Ø.8. При одновременном использовании обоих параметров предпочитают первый – NFACTORS.

Параметры ITERATE и STOPFACT используются при итеративных методах факторизации (PA2, RAO, ALPHA). Параметр ITERATE определяет максимальное число итераций. Если точность, достигаемая числом итерации по умолчанию (25), не устраивает заказчика, то он может этим параметром определить другое максимальное число итераций. Например, ITERATE = 99/ или ITERATE = 10/. Параметром STOPFACT определяется точность, нужная для окончания итераций, если определенное по умолчанию Ø.ØØI не устраивает заказчика. Например, STOPFACT=Ø.ØØØI/ или STOPFACT = Ø.5/. Максимальное значение, которое здесь можно использовать – единица, для представления точности после запятой нельзя использовать больше семи мест. Если параметры ITERATE и STOPFACT используются одновременно, то процесс итерации заканчивается только тогда, когда совершено определяемое количество итераций, независимо от того, была ли требуемая точнос ь достигнута или нет.

Параметром ROTATE определяется метод для вращения факторов. При одном обращении к процедуре FACTOR можно использовать только один метод вращения. При отсутствии параметра ROTATE используют метод VARIMAX. Если за параметром ROTATE находится ключевое слово NOROTATE, то факторы не вращаются. Факторы не вращаются и в случае, если при первоначальном решении был выведен только один фактор. Неортогональным враще-

нием можно управлять параметром DELTA, по умолчанию значением которого является нуль. При желании изменить степень неортогональности нужно изменить значение параметра DELTA. Ему можно передать только одно значение, например, DELTA= .5/ или интервал значений, указывая начальное значение, приращение и конечное значение, например, DELTA = .5 -.I -.2/.

Для вывода факторных весов для вращенных факторов нужно заказать статистику 6, иначе они не будут выведены. Для вывода факторных весов для первоначальных факторов, нужно заказать статистику 5.

Ключевое слово FACSCORE нужно использовать при вычислении значений факторов для всех объектов, т.н. индивидуальных факторных весов. Вместе с этим ключевым словом нужно заказать и статистику 7, иначе результаты не будут выведены. Значения факторов образуют самостоятельный файл, адрес которого определяется оператором DD с именем FTØ9FØØ1, их можно использовать в дальнейшей обработке данных.

На одной карте FACTOR может быть несколько обращений к проуедуре, т.е. ключевые слова VARIABLES и TYPE вместе со следующими за ними параметрами можно повторить.

II.I.3. Опции

Основная опция процедуры FACTOR работает с полной выборкой, т.е. использует только те объекты, у которых измерены все признаки из списка VARIABLES.

Особенностью процедуры FACTOR является отсутствие распечатки основного варианта. За картой FACTOR должна всегда следовать карта STATISTICS, на которой перечисляются номера желаемых статистик.

За картой FACTOR может следовать карта OPTIONS, в поле

описаний которой перечисляются номера желаемых опций, разделяя их запятой. Возможные опции следующие:

I) обозначения пропусков используются как значения признаков;

2) объект не используют при вычислении коэффициента корреляции, если у него отсутствует значение используемого признака, при вычислении остальных коэффициентов корреляции он может участвовать. Использование этой опции является нежелательным - полученная корреляционная матрица может не сказаться положительно определенной, а тогда выделенная факторная модель может быть некорректной;

3) в качестве начальных данных используется корреляционная матрица, элементы которой вводятся в формате F10.7. Адрес корреляционной матрицы определяется оператором DD с именем FT08F001. Для ввода корреляционной матрицы нужно использовать управляющие карты, описанные в пункте I0.3.3. Корреляционную матрицу можно ввести и с перфокарт;

4) отсутствует;

5) из корреляционной матрицы образуют самостоятельный файл, адрес которого определяется оператором DD с именем FT09F001. Вместе с этой опцией нужно заказать и статистику 2;

6) из матрицы факторных весов и общностей образуют самостоятельный файл, адрес которого определяется оператором DD с именем FT09F001; формат для факторных нагрузок при этом является F10.7, а для общностей - F10.5. Вместе с этой опцией нужно заказать статистики 5 и 6;

7) отсутствует;

8) из средних значений и стандартных отклонений образу-

ют самостоятельный файл, адрес которого определяется с оператором DD именем FT09F001, форматом вывода является F10.4. Если опции 5, 6, 7 и 8 используются одновременно, результаты выводятся в порядке:

- средние значения;
- стандартные отклонения;
- корреляционная матрица;
- матрица факторных весов к общности.

Вместе с опцией 8 нужно заказать и статистику I;

9) в качестве начальных данных используют корреляционную матрицу, при этом для всех списков VARIABLES вводится одна корреляционная матрица; используется формат F10.7;

I0) из коэффициентов определения индивидуальных факторных весов образуют самостоятельный файл, адрес которого определяется оператором DD с именем FT09F001; используется формат F10.7;

II) из индивидуальных факторных весов образуют самостоятельный файл, где для каждого объекта записывают его порядковый номер в подфайле, номер записи, имя подфайла и значения индивидуальных факторных весов в формате 6F10.6 (по умолчанию они выводятся в формате 8F10.6.

II.I.4. Статистики

За картой FACTOR должна обязательно следовать карта STATISTICS, в поле описаний которой перечисляются номера желаемых статистик. Если нужно заказать все статистики, то в поле описаний карты можно написать ключевое слово ALL.

Можно выбирать следующие статистики:

I) средние значения и стандартные отклонения используемых признаков;

2) корреляционная матрица;

3) обратная матрица для корреляционной матрицы и детерминант корреляционной матрицы (для регулярной матрицы);

4) общности, собственные значения корреляционной матрицы, процент общей вариации, определенный фактором и суммарный процент общей вариации, описанный последовательными факторами;

5) матрица первоначальных факторных весов;

6) при ортогональных вращениях матрица факторных нагрузок вращенных факторов и матрица преобразования, а при косоугольных вращениях матрица факторных весов вращенных факторов, матрица переобразования факторов и корреляционная матрица для признаков и факторов;

7) индивидуальные факторные веса и коэффициенты для их определения;

8) графическое представление признаков в пространстве факторов.

II.I.5. Пример

Пусть в качестве начальных данных для факторного анализа можно использовать корреляционную матрицу, ее нужно ввести с перфокарт. Желательно использовать классический факторный анализ, при этом для общностей определены следующие оценки: 0.8; 0.85; 0.3; 0.75; 0.9. Для вращения факторов используют по умолчанию метод VARIMAX. Нужно вывести общности и матрицу факторных весов для вращенных факторов, а для проверки начальных данных и корреляционную матрицу. Такой заказ представляется картами:

```
VARIABLE LIST    П1, П2, П3, A, B
INPUT MEDIUM     CARD
# OF CASES       9Ø
INPUT FORMAT     FIXED (5F1Ø.7)
FACTOR           VARIABLES= П1, П2, П3, A, B/
                 TYPE=PA1/
                 DIAGONAL= .8 .85 .3 .75 .9/
OPTIONS          3
STATISTICS       2, 4, 6
READ MATRIX
      I.      - 824    2Ø2Ø5    8ØI49    - 6838
    - 824       I.      424ØI     3I49     885Ø6
    2Ø2Ø5     424ØI       I.    -I3363     3Ø292
    8ØI49       3I49   -I3363      I.       I56I3
  - 6838      885Ø6     3Ø292    I56I3       I.
FINISH
```

II.2. Процедура CANCORR

II.2.I. Цель и методические указания

II.2.I.I. Общая модель канонического анализа. Канонический анализ (нахождение канонических компонент) принадлежит к числу сравнительно редко используемых процедур многомерного статистического анализа, хотя он применим для решения как описательных, так и прогностических задач. Целью канонического анализа является нахождение зависимости между двумя группами признаков. Эта зависимость реализуется при помощи новых признаков, т.н. канонических компонент (канонических факторов), вычисленных по исходным признакам первой и второй группы соответственно. Притом канонические компоненты выбя-

раются так, чтобы они удовлетворяли следующим условиям:

1) они являются линейными комбинациями соответствующих групп признаков;

2) канонические компоненты одной группы взаимно некоррелированы;

3) канонические компоненты выбраны таким образом, что корреляции между соответствующими компонентами (т.н. канонические корреляции) разных групп были максимальны;

4) канонические компоненты упорядочены по мере убывания соответствующих канонических корреляций;

5) число канонических корреляций не больше числа признаков в меньшей группе.

Канонические корреляции всегда неотрицательны, причем их основные свойства совпадают со свойствами обыкновенных коэффициентов корреляций. Чем больше канонические корреляции, тем сильнее связаны рассматриваемые группы признаков между собой.

Канонический анализ, точно так же как и факторный анализ, реализуется в форме задачи нахождения собственных значений и собственных векторов на основании некоторой функции корреляционной матрицы исходных признаков. Эта есть $m_2 \times m_2$ матрица $B=R_{21}R_{11}^{-}R_{12}R_{22}^{-}$, где R_{11} и R_{22} - корреляционные матрицы первой и второй группы (их размерности соответственно $m_1 \times m_1$ и $m_2 \times m_2$), R_{12} - матрица взаимных корреляций первой и второй группы, а R_{11}^{-} и R_{22}^{-} - обобщенные обратные матрицы матриц R_{11} и R_{22}.

Собственные значения матрицы B, упорядоченные в монотонно убывающем порядке, равняются квадратам канонических корреляций, левые и правые собственные векторы - соответст-

венно каноническим компонентам первой и второй группы признаков.

С точки зрения канонического анализа обе группы признаков равноценны. Для разрешимости задачи требуется, чтобы по крайней мере одна из корреляционных матриц R_{11} и R_{22} была положительно определена, т.е. в соответствующей группе не должны существовать линейно зависимые признаки (в смысле функциональной зависимости). В противном случае следует один или несколько признаков из группы выбросить.

Хотя канонические компоненты всегда являются ортогональными, не существует однозначной договоренности относительно их нормирования. В принципе можно здесь пользоваться несколькими разными способами, дающими различные, но приводимые в принципе друг к другу решения.

II.2.I.2. Применение канонического анализа и интерпретация результатов. В больших задачах анализа данных часто имеется дело со многими самостоятельными блоками признаков, которые все выступают в качестве аргументов для некоторой группы функциональных признаков или которые образуют некоторую иерархию основании причиненности. Например, в педагогических исследованиях такими блоками могут быть социально-экономические показатели домов учащихся, характеристики их физического развития, признаки, характеризующие их черты личности, показатели общественной активности и успеваемости в учебном процессе. Каждый из названных блоков может включать в себя несколько десятков или более сотни признаков, а целью исследования является нахождение зависимостей между отдельными блоками признаков. На основании отдельных признаков эта работа имела бы чрезмерно большой объем, причем, как правило,

большинство корреляций между признаками из разных групп являются очень низкими. Использование метода канонического анализа в некотором смысле "усиливает" эти корреляции: между группами из нескольких десятков признаков возникает, как правило, значимые канонические корреляции.

Для полного решения поставленной задачи приходилось бы сделать $r(r+1)/2=15$ канонических анализов (r - число групп признаков, в данном примере $r=5$).

Разумеется, что необходимо следить, чтобы число исследуемых объектов было достаточно большим. Грубым правилом является здесь требование, чтобы число объектов, у которых все признаки измерены, не было менее $2(m_1+m_2)$.

При интерпретации результатов канонического анализа можно использовать ту же схему, которой мы пользовались при интерпретации результатов факторного анализа. При рассмотрении содержания канонических компонент необходимо учитывать и правило нормировки коэффициентов канонических компонент.

II.2.I.3. Проверка значимости канонических компонент. Значимость канонических компонент (или, что тоже самое: отличие от нуля канонических коэффициентов) можно проверять при помощи χ^2-теста. Если вычислено k канонических корреляций, то для каждого $m(m=1,\ldots,k)$ следует проверять пару гипотез:

H_0^m: канонические корреляции начиная с ρ_m равняются нулю:

$$\rho_m = \rho_{m+1} = \ldots = \rho_k = 0$$

H_1^m: по крайней мере ρ_m отличается от нуля

$$\rho_m \neq 0$$

Решение получается на основании значения χ^2-статистики, которое печатается вслед за каждой канонической корреляцией

в столбце с заглавием CHI-SQUARE. Если значение статистики больше критического (табличного) значения при выбранном уровне значимости (например, 0.05) и имеющимся числе степеней свободы (выдается в столбце DEGREES OF FREEDOM), то принимается содержательная гипотеза H_1^m. Процедура повторяется при следующей ($m=m+1$) канонической корреляции. Если значение выпускаемой χ^2-статистики меньше соответствующего табличного значения, то принимается нулевая гипотеза H_0^m, т.е. - зависимость между группами исследуемых признаков исчерпывающим образом уже описана каноническими компонентами с индексами до $m-1$.

Понятно, что если при некотором значении m_* принимается нулевая гипотеза, то и для всех следующих значений $m_*+1,\ldots,k$ придется принимать H_0.

Интерпретировать следует только такие канонические компоненты, которым соответствуют существенные канонические корреляции.

II.2.I.4. Применение канонического анализа для решения прогностических задач. Кроме решения описательных задач по примеру пункта II.2.I.2. можно каноническим анализом пользоваться и при решении задач прогностического типа. Это целесообразно в таких случаях, когда потенциальных функциональных признаков много и исследователю не ясно, какие из них следует выбрать в модель. Образуя из потенциальных функциональных признаков одну группу, из аргументов - другую группу, можем в ходе канонического анализа образовать новые, в некоторым смысле оптимальные функциональные признаки - это канонические компоненты первой группы. Регрессионным уравнением (моделью) для них через предполагаемые ар-

гументы и являются соответствующие канонические компоненты второй группы – линейные комбинации исходных аргумент-признаков. Качество прогноза характеризуется соответствующей канонической корреляцией. Заметим, что первая каноническая корреляция определяет супремум множественного коэффициента корреляции при прогнозировании всех линейных комбинаций функциональных признаков по всевозможным линейным функциям аргументов.

При интерпретации прогнозов как регрессионных уравнений следует учитывать тот факт, что при проведении канонического анализа все исходные признаки нормируют и центрируют; поэтому регрессионные уравнения и не содержат свободного члена.

Если массив данных содержит пропуски (неизмеренные признаки), то с точки зрения сохранения информации целесообразно все отдельные задачи канонического анализа оформить как самостоятельные заказы не содержащие в исходном списке лишних признаков.

II.2.2. Процедурная карта

При обращении к процедуре указывают, какие признаки нужно использовать и как образовать группы признаков. Общий вид процедурной карты следующий:

```
CANCORR     VARIABLES = varlist/
            RELATE = имя TO имя  WITH имя TO имя/ [.../
            RELATE = имя TO имя  WITH имя TO имя/]
```

В поле управлений карты находится имя процедуры. В поле описаний первым находятся ключевое слово VARIABLES, за которым следует список используемых в каноническом

анализе признаков, список кончается наклонной чертой. Затем следует ключевое слово RELATE, за которым после знака равенства следует описание первой группы, ключевое слово WITH и описание второй группы. Так как в одну группу можно объединить только признаки, находящиеся подряд в списке VARIABLES, то описание группы должно быть представлено при помощи TO-соглашения, т.е. указывается только имя первого и последнего признака группы, между которыми, разделенное пропусками, находится ключевое слово TO. Ключевое слово RELATE вместе с следующими описаниями групп можно повторить, заказывая таким способом одним обращением к процедуре несколько канонических анализов.

II.2.3. Опции

Основная опция процедуры использует только те объекты, у которых измерены значения всех признаков из списка VARIABLES (работа с полной выборкой). Выводятся средние значения, стандартные отклонения, корреляционная матрица. Для заказанных канонических анализов выводятся собственные значения, определяющие квадраты коэффициентов канонических корреляций, величины Λ (WILK'S LAMBDA), используемые при вычислении χ^2-статистик, χ^2-статистики (CHI-SQUARE), описывающие значимость коэффициентов канонической корреляции и их степени свободы (DEGREES OF FREEDOM), нужные для нахождения критических значений из таблицы χ^2-распределения. После этого печатаются коэффициенты канонических компонентов для обоих групп.

За картой CANCORR может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Можно выбрать следующие опции:

I) обозначения пропусков используются как значения признаков;

2) объект не используется при вычислении коэффициента корреляции, если у него отсутствует значение используемого признака, при вычислениях остальных коэффициентов корреляции он может участвовать. Использование этой опции нежелательно;

3) в качестве начальных данных используют корреляционную матрицу, элементы которой вводятся в формате F10.7. Адрес корреляционной матрицы определяется оператором DD с именем FT08F001. Для ввода корреляционной матрицы нужно использовать управляющие карты, описанные ниже; для каждого заказа RELATE нужно ввести отдельно корреляционную матрицу;

4) из корреляционной матрицы образуют самостоятельный файл, адрес которого определяется оператором DD с именем FT09F001. Элементы корреляционной матрицы выводятся в формате F10.7;

5) отсутствует;

6) отсутствует;

7) отсутствует;

8) отсутствует;

9) в качестве начальных данных используют корреляционную матрицу всех признаков из списка VARIABLES (т.е. для каждого списка RELATE используют нужную часть этой корреляционной матрицы).

В отличие от процедур, описанных выше, при вводе корреляционной матрицы нужно использовать только три управляющие карты - VARIABLE LIST, INPUT MEDIUM и READ MATRIX.

II.2.4. Статистики

За картой CANCORR может следовать карта STATISTICS, в поле описаний которой перечисляются номера желаемых статис-

тик. В отличие от предыдущих процедур, этой картой можно только уменьшить число выводимых статистик, так как основная опция выводит как средние значения и стандартные отклонения используемых признаков, так и их корреляционную матрицу. Можно заказать следующие статистики:

I) средние значения и стандартные отклонения;

2) корреляционную матрицу;

3) корреляционную матрицу, где указаны невычисленные из-за нехватки объектов коэффициенты корреляции (используется только опцией 2).

II.2.5. Пример

Пусть требуется провести два разных канонических анализа, причем в первой задаче первую группу образуют признаки T1, T2, T3, а во второй - T3, T4 и T5. Вторую группу образуют в обоих задачах признаки A, B, C. Общая корреляционная матрица этих признаков записана на магнитный диск. Заказ можно представить следующими управляющими картами:

```
VARIABLE LIST   T1, T2, T3, T4, T5, A, B, C/
INPUT MEDIUM    DISK
CANCORR         VARIABLES = T1 TO C/
                RELATE = T1 TO T3 WITH A TO C/
                RELATE = T3 TO T5 WITH A TO C/
 OPTIONS         3, 9
READ MATRIX
FINISH
```

II.3. Процедура MULTIVARIANT

II.3.I. Цель и методические указания

II.3.I.I. Целью процедуры MULTIVARIANT является группировка объектов, т.е. распределение объектов, описанных при

помощи некоторых признаков, в классы так, чтобы в одном классе находились в некотором смысле близкие объекты. Проверить значимость полученных групп нет возможности, единственным критерием, оценивающим качество решения, является интерпретируемость полученных групп.

Во многом результат группировки зависит от выбора признаков - чем лучше они характеризуют тот аспект материала, на основе которого требуется группировать объекты, тем лучше результат. Для всех процедур группировки в системе САИСИ требуется, чтобы признаки были количественными. Не рекомендуется выбирать группирующие признаки сильно зависимыми.

После того, как группы уже образованы, можно исследовать дискриминирующего способность каждого признака с помощью процедуры ONEWAY или DISCRIMINANT.

В зависимости от априорной информации рассматриваются две категории методов группировки.

Если число классов известно, и известно несколько объектов из каждого класса (т.н. обучающая выборка), то имеется дело с задачей классификации (CLASSIFY).

Если число классов не известно, и нет уже классифицированных объектов, то придется решать задачу кластер-анализа (CLUSTER).

II.3.I.2. Меры растояния между объектами. Для применения методов группировки необходимо иметь некоторую меру, характеризующую "близость" или "подобие" объектов. В процедуре MULTIVARIANT имеется возможность вычисления четыре разных обобщенных расстояний, определяемых общей формулой

$$D^2(x,y) = (x-y)'M(x-y),$$

где x и y - (px1)-векторы значений группирующих признаков у

двух объектов x и y, а M некоторая неотрицательно определенная матрица, определяющая тип расстояния. Для выбора матрицы M есть следующие возможности:

I) M = I, т.е. единичная матрица. В таком случае получается евклидовое расстояние D_1^2:

$$D_1^2(x,y) = \sum_{i=1}^{p} (x_j - y_j)^2,$$

где x_j и y_j - значения j-го признака у объектов x и y соответственно. Если $p \leqslant 3$, то D(x,y) есть обычное расстояние между точками x и y (длина отрезка $[x,y]$);

2) $M = S_i^{-1}$, где S_i - ковариационная матрица i-ой группы объектов,

$$D_2^2(x,y) = (x-y)' S_i^{-2}(x-y).$$

Заметим, что такое расстояние зависит от априорной информации о том, в какие группы принадлежат объекты;

3) M = A'A, где A' - матрица преобразования объектов в "пространство оптимальной дискриминации",

$$D_3^2(x,y) = (x-y)' A'A(x-y);$$

4) $M = W^{-1}$, где W - ковариационная матрица, общая для всех групп:

$$D_4^2(x,y) = (x-y)' W^{-1}(x-y).$$

В случае, когда $S_1 = S_2 = \ldots = S_K$ для всех групп, расстояния D_2^2 и D_4^2 совпадают.

II.3.I.3. Многомерная классификация (CLASSIFY). Метод многомерной классификации реализован следующим образом. Для каждой группы находят среднее $\bar{y}_i$ по обучающей выборке (это не реальный, а фиктивный объект, который считается центром группы). Затем для каждого измеренного объекта x находят

его расстояние до всех средних $D(x,\bar{y}_i)$. Объект считают принадлежащим группе i_*, если он находится ближе к центру этой группы $\bar{y}_{i_*}$:

$$D(x,\bar{y}_{i_*}) = \min_{1\le j\le k} D(x,y_i).$$

Тип расстояния $D(x,y)$ (метрики) выбирается исследователем. Для этого можно сделать следующие рекомендации:

I) метрика D_1^2 применима лишь в случае, если все группирующие признаки имеют приблизительно одинаковую вариабельность (измерены на равных шкалах) и не сильно коррелированы. Если некоторый признак имеет намного большую дисперсию чем другие, то расстояние зависит, в основном, от этого признака. Таким же образом группа сильно коррелированных признаков чрезмерно усиливает влияние каждого из них по сравнению с другими признаками, не входящими в эту группу;

2) в случае, когда шкалы признаков и их дисперсии различные и признаки коррелированы, то следует пользоваться метрикой D_2^2 или D_4^2 (это т.н. расстояние Махаланобиса). Метрикой D_2^2 следует пользоваться в случаях, когда предполагается , что ковариационные матрицы в группах заметно отличаются друг от друга. (Об этом можно судить, пользуясь возможностью распечатки ковариационных матриц процедурой MULTIVARIANT). Заметим все-таки, что для получения статистически корректных результатов необходимо, чтобы число объектов обучающей выборки, принадлежащих в каждую группу, было достаточно большое - не менее, чем $p(p+1)/2$, где p число группирующих признаков .

Метрика D_4^2 является наиболее стандартной для решения задач классификации по количественным признакам: она применима в случае коррелированных признаков, имеющих разные дис-

персии. Для корректного применения этой метрики требуется, чтобы объем обучающейся выборки был больше числа $p(p+1)/2$.

Заметим, что классификация по метрике D_4^2 соответствует процедуре линейного, а классификация по метрике D_3^2 - нелинейного дискриминантного анализа;

3) метрика D_2^2 соответствует случаю, когда вместо исходных признаков в основу классификации выбраны новые, которые выражаются как линейные комбинации исходных, и определенные условием, что они должны разделять классы наилучшим образом. Такие новые признаки определяются собственными значениями матрицы $W^{-1}B$, где B - межгрупповая матрица ковариации. Процедура аналогична компонентному анализу (см. FACTOR), причем число новых классифицирующих признаков $r = \min(p,k-1)$ и определяется исследователем через % описания дискриминирующей способности исходных признаков через новые. В качестве дополнительных статистик можно распечатать и собственные значения и матрицы преобразования (собственные векторы) A; их интерпретация и содержательный анализ имеет иногда самостоятельное познавательное значение.

II.3.I.3. Кластерный анализ (CLUSTER) . В процедуре MULTIVARIANT реализованы два метода кластерного анализа, но на основании системного файла работает из них только один, т.н. метод лидера (LEADER).

Методы кластерного анализа реализованы только для евклидового расстояния. Это значит, что если желательно при применении этих методов пользоваться признаками, имеющими существенно разные дисперсии или которые сильно коррелированы, следует их заранее перекодировать или нужным образом переобразовать.

II.3.I.4. Методом лидера объекты распределяются в непересекающиеся сферические кластеры, исходя из заданного исследователем значения d порога (THRESOLD). Лидером первого кластера y_1 выбирают первый объект файла: x_1, -а в первый кластер определяют все объекты x_i, для которых выполняется условие

$$D_1(y_j, x_i) < d \qquad (II.5)$$

при j=1.

Первый объект, при котором это условие не выполнено, выбирается лидером y_2 второго кластера, и второй кластер образуют из всех объектов, которые не попали в первый кластер и которые удовлетворяют условию (II.5) при j=2. Такая процедура продолжается до тех пор пока не достигнута максимальное число кластеров k, или не сгруппированы все объекты. Если некоторые объекты не удовлетворяют условию (II.5) при всех j(j=1,...,k), то эти объекты вообще не войдут в кластеры.

Обычно при первом заказе не удастся хорошо выбрать значение порога d. Поскольку процедура CLUSTER очень быстрая, то целесообразно повторять заказ, руководствуясь следующими очевидными правилами:

I) если кластеры слишком большие, то следует уменьшать порог;

2) если кластеров слишком много, то следует увеличить порог.

Заметим, что кластеры, полученные этим методом, зависят от порядка элементов в файле. В среднем первый кластер содержит больше объектов, чем второй, и т.д.

II.3.I.5. Метод k-средних (K MEANS) работает только в

случае данных, непосредственно вводимых в ЭВМ, и не работает на основании системного файла.

При этом методе объекты разбивают в кластеры шаг за шагом, причем число кластеров k выбирается исследователем.

На первом этапе все объекты образуют один кластер.

На каждом шаге для каждого кластера определяют среднее и суммарную внутриклассовую вариабельность, а затем выбирают кластер, для которого вариабельность наибольшая, и распределяют его на два. Причем здесь используется признак, имеющий самую большую дисперсию.

После того, как образовался новый кластер, происходит переклассификация всех объектов: вычисляют новые центры всех кластеров, а каждый объект определяется в тот кластер, центр которого находится для него ближе всех остальных. На каждом этапе процедуру вычисления параметров кластеров повторяют несколько раз, пока они не станут стабильными (число итераций определяется соответствующим параметром). Так как на каждом шагу объекты переклассифицируются, то получаемая кластер-структура не является иерархической.

II.3.2. Процедурная карта

Так как при использовании многомерной классификации и кластерного анализа общий вид процедурной карты различен, то приведен ее описание для обоих методов отдельно. При многомерной классификации общий вид процедурной карты следующий:

```
MULTIVARIANT   ANALYSIS=CLASSIFY,METHOD={1|2|3|4},
               [ML={YES|NO}], [TP=процент используемых
               собственных значений,]
               MAXGROUP =целое,GROUPVAR =имя
               VARLIST=varlist
```

В управляющем поле карты написано имя процедуры; так как для его идентификации хватит и восьми символов, то можно его написать в виде MULTIVAR. Параметры в поле описаний являются обязательными и необязательными. Опишем сперва обязательные параметры.

Первым в поле описаний находится ключевое слово ANALYSIS, за которым после знака равенства следует ключевое слово, определяющее тип используемого метода группировки. При многомерной классификации здесь нужно использовать ключевое слово CLASSIFY. Номер, следующий за ключевым словом METHOD, определяет, какую метрику нужно при группировке использовать. Ключевым словом MAXGROUP определяется число классов. Для каждого объекта номер класса определяется признаком, имя которого следует за ключевым словом GROUPVAR. Для объектов, не принадлежащих к учебной выборке, значение этого признака нужно закодировать обозначением пропуска. За ключевым словом VARLIST следует список тех признаков, которые используют для описания объектов.

Из необязательных параметров параметр ML можно использовать только с метрикой 2 - в случае, где для группировки желательно использовать критерий отношения правдоподобия. Параметр TP можно использовать только вместе с метрикой 3, он определяет число дискриминирующих координат. При кластерном анализе общий вид процедурной карты следующий:

MULTIVARIANT ANALYSIS = TAXONOMY, CLUSTER= максимальное число кластеров, OBSERV=число объектов, [THRESOLD= значение,]
[ITRATION =число повторений ,] METHOD= {2|3},
VARLIST=varlist

В поле описаний за ключевым словом ANALYSIS следует этот раз ключевое слово TAXONOMY, которое сообщает о желании использовать методы кластерного анализа. Следом за ключевым словом CLUSTER сообщают максимальное количество кластеров. Если это число окажется меньшим действительного числа кластеров, возникших в ходе работы алгоритма лидера, классификация заканчивается, выводится соответственное сообщение, а также описание уже образованных кластеров. За ключевым словом OBSERV следует число классифицируемых объектов. Следующие параметры в некотором смысле "условно" необязательные. Вместе с методом 2 (алгоритм лидера) обязательно нужно использовать параметр THRESOLD, за которым после знака равенства следует значение порога, используемого при образовании кластеров, а параметр ITRATION использовать нельзя. Вместе с методом 3 (алгоритм k-средних) нужно обязательно использовать параметр ITRATION, которым определяют число перевычислений центров кластеров и переформирования кластеров, но использовать параметр THRESOLD нельзя.

Параметром METHOD определяется, какой метод кластерного анализа нужно использовать: 2 - алгоритм лидера, 3 - метод k-средних. За ключевым словом VARLIST следует список признаков, используемых для описания объектов.

Обращаем внимание на то, что в отличие от всех предыдущих процедурных карт на этой карте параметры разделяются запятой, а не наклонной чертой.

При использовании процедуры MULTIVARIANT имеется еще одна нерегулярность - для работы процедуры нужен еще вспомогательный файл, адрес которого определяется оператором DD с именем FT19F001. Этот оператор можно представить в виде:

```
//FT19FØØ1 DD DSN=SAISI.MLTVB,UNIT=SYSSQ,
// DSB=(RECFM=VS,BLKSIZE=8Ø),SPACE=(TRK,(2,2)),
// VOL=SER=OSØØØ1,DISP=(NEW,DELETE)
```

II.3.3. Опции

Вывод процедуры зависит от использованного метода. При методах многомерной классификации выводятся порядковой номер объекта, значение группирующего признака и номер группы, определяемой при классификации. Неправильно классифицированные объекты отмечаются тремя звездочками. Дополнительно выбираемые опции при методах многомерной классификации отсутствуют.

При методах кластерного анализа выводятся описания всех образуемых кластеров: число кластеров, число объектов в кластере и порядковые номера объектов, принадлежащих данному кластеру. Кроме того, для каждого кластера еще выводятся максимальное и минимальное значение, среднее значение, стандартное отклонение всех признаков, используемых для описания объектов, а также значения всех признаков у объекта, выбранного лидером.

При кластерном анализе имеются следующие дополнительно выбираемые опции:

I) выводятся начальные данные;

2) (используется только при методе k-средних) выводится деревообразная схема образования кластеров.

II.3.4. Статистики

Число дополнительно выбираемых статистик зависит от использованного метода. При многомерной классификации можно выбрать следующие дополнительные статистики:

I) число классов, число используемых признаков, число объектов в каждом классе;

2) средние значения всех признаков во всех классах и общие средние значения по всем классам;

3) ковариационные матрицы всех классов, общая внутриклассовая ковариационная матрица;

4) (используемый только при метрике 3) собственные значения и собственные векторы матрицы $W^{-1}B$;

5) (используемый только при метрике 2) - матрицы, обратные ковариационным матрицам классов и их детерминанты.

При кластерном анализе дополнительно выбираемых статистик не предусмотрено.

II.4. Процедура GUTTMAN SCALE

II.4.I. Цель и методические указания

Целью процедуры GUTTMAN SCALE является шкалирование группы порядковых признаков, т.е. образование нового количественного признака, который содержал бы максимальную возможную информацию, передаваемой исходной группой признаков (см. [2] , стр. 227-288).

В ходе процедуры GUTTMAN SCALE все исходные признаки кодируются в дихотомические путем разделения их шкал на две части с помощью заданных исследователем точек разреза (CUTTING POINTS). В результате каждому исходному признаку X_i соответствует кодированный $У_i$, имеющий значения 0 и 1.

Для того, чтобы метод шкалирования Гуттмана по возможности полностью сохранил информацию исходных признаков, предполагается, что для этих признаков выполняется условие кумулятивности. Это условие состоит в том, что найдется такая перестановка $У_{i1},\ldots, У_{ir}$ признаков $У_1,\ldots,У_r$, что они с достаточно большой вероятностью являются монотонными, т.е. выполняется условие

$y^j_{i_1} \geqslant y^j_{i_k}$, если $i_1 < i_k$,

где y^j_i - значение i-ого (кодированного) признака у объекта j(j=1,...,n).

Процедурой GUTTMAN SCALE по заданному списку признаков и заданным правилам кодирования определяется перестановка признаков $y_{i_1},\ldots,y_{i_r}$, вычисляют новый признак z, имеющий значения от 0 до r (число признаков в группе). В идеальном случае, когда условие кумулятивности выполнено, объекту j присваивают значение $z^j = \sum_{i=1}^{r} y^j_i$, но в случае, когда условие кумулятивности не выполнено, значение z^j оценивается при помощи специальной процедуры, но таким образом, что z^j всегда целое число на отрезке $[0, r]$.

Заметим, что в случае, когда условие кумулятивности выполнено точно, процедура GUTTMAN SCALE дает результаты, эквивалентные с теми, которые получаются при помощи процедуры COUNT (при тех же условиях кодирования).

Процедурой GUTTMAN SCALE даются еще разные характеристики, описывающие степень сохранения исходной информации в результате шкалирования. Это - коэффициент воспроизводимости группы признаков, а также коэффициент воспроизводимости того признака, который воспроизводится менее всех остальных. Дается еще матрица коэффициентов корреляции Юла между исходными признаками. Коэффициенты воспроизводимости характеризуют вероятность правильных результатов при обратном вычислении признаков y_{i_k} по шкале z. Если этот процент меньше 90, то результат шкалирования считается неудовлетворительным.

Так как результат шкалирования существенно зависит от кодирования, то имеется возможность задавать сразу для как-

дого признака несколько (до трех) точек разреза, а тогда образуется шкала для всех возможных комбинаций получаемых правил кодирования. Из числа этих различных шкал (их число C_1, $C_2 \ldots C_r$, где C_r - число правил кодирования 1-го признака) можно найти наиболее подходящую шкалу по величине коэффициента воспроизводимости.

Наибольшим недостатком процедуры GUTTMAN SCALE есть тот факт, что образованный новый признак z не сохраняется для дальнейшей работы, он имеет только познавательное значение для анализа исследуемой группы признаков.

Единственная возможность нахождения признака, близкого к найденному z, есть применение процедуры COUNT при выборе шкал рассматриваемых признаков, соответствующих наибольшему значению коэффициента воспроизводимости. Хотя полученый новый признак не совпадает точно со шкалой z, он является достаточно близким для него (тем ближе, чем выше коэффициент воспроизводимости).

II.4.2. Процедурная карта

При обращении к процедуре нужно перечислить имена тех признаков, которые используются для образования шкалы и определить имя шкалы. Общий вид карты следующий:

GUTTMAN SCALE имя=имя (точки разреза)[...имя (точки разреза)]/

В управляющем поле находится имя процедуры.

В поле описаний первым записывается имя образуемой шкалы. За именем шкалы следует после знака равенства список тех признаков, которые будут использованы при образовании шкалы. В этом списке не может быть больше I2 признаков. После каждого имени признака в скобках указаны точки разреза, их не может быть больше трех. Список признаков заканчивается нак-

лонной чертой. При надобности могут следовать новые заказы до 50 разных шкал.

II.4.3. Опции

Основная опция процедуры выводит описание шкалы - таблицу, каждая строка которой соответствует некоторому значению шкалы, каждый столбец - некоторому значению некоторого признака. Недопустимое значение признака, соответствующее данному значению шкалы, обозначаются через ERR. В таблице указывают частоты данного признака в случае данного значения шкалы, на краях указаны маргинальные частоты значений шкалы и значений признаков. Для значений признаков указывается и частота ошибок при их восстановлении по значениям шкалы. В вычислениях участвуют только объекты со всеми значениями признаков, используемых в данном заказе (работа с полной выборкой). В распечатке указываются и метки.

За картой GUTTMAN SCALE может следовать и карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Возможные опции следующие:

1) обозначение пропуска используется как значение признака;

2) не выводятся метки;

3) не происходит автоматическое упорядочение признаков, при образовании шкалы их используют в порядке перечисления на процедурной карте.

II.4.4. Статистики

За картой GUTTMAN SCALE может следовать карта STATISTICS, в поле описаний которой перечисляются номера желаемых статистик. Можно заказать следующие статистики:

I) коэффициенты корреляции Пирсона;

2) коэффициент воспроизводимости;

3) коэффициент минимальной маргинальной воспроизводимости;

4) процент достижимого улучшения шкалой Гуттмана;

5) коэффициент шкалируемости.

Если одновременно нужно заказать все признаки, в поле описаний карты STATISTICS можно написать ключевое слово ALL.

II.4.5. Пример

Предположим, что нужно образовать шкалу Гуттмана для признаков ИНДЕКС1, ИНДЕКС2, ИНДЕКС3, при этом точками разреза для всех признаков выбраны значения 4 и 6. Желательно напечатать статистики от двух до пяти. Заказ представляется картами:

```
GUTTMAN SCALE  ШКАЛА=ИНДЕКС (4,6),ИНДЕКС2(4,6),ИНДЕКС3(4,6)/
STATISTICS     2, 3, 4, 5
```

В результате заказа обрабатываются шесть разных шкал Гуттмана, при образовании которых используют всевозможные комбинации точек разреза.

XII. ПРОГНОЗИРУЮЩИЕ МЕТОДЫ МНОГОМЕРНОГО АНАЛИЗА

К числу прогнозирующих методов многомерного анализа относятся регрессионный анализ (процедура REGRESSION), дискриминантный анализ (процедура DISCRIMINANT) и анализ временных рядов (процедура TIME SERIES). Оба последних метода содержат возможности и для решения некоторых задач описательного характера.

I2.I. Процедура REGRESSION

I2.I.I. Цель и методические указания

I2.I.I.I. Регрессионная модель. Предпосылки регрессионного анализа. Целью процедуры REGRESSION является построение линейной модели и оценка ее параметров. Точнее, эта задача формулируется следующим образом: предполагается, что зависимый признак (функция) У выражается в форме линейной комбинации аргументов (т.н. независимых признаков) $x_1, \ldots, x_p$ с прогрешностью ε:

$$y = \bar{y} + \varepsilon,$$

где $$\bar{y} = b_0 + b_1 x_1 + \ldots + b_p x_p, \qquad (I2.I)$$

а $b_0, b_1, \ldots, b_p$ – неизвестные коэффициенты и ε – случайная ошибка. Пользуясь измеренными значениями признаков (выборкой), оцениваются параметры регрессии (регрессионные коэффициенты) $b_0, \ldots, b_p$, а также характеристики точности модели и тесноты регрессионной зависимости.

Зная регрессионную модель (I2.I), можно определить прогноз $\hat{y}_j$ для объекта, у которого известны наблюдения аргумент-признаков $x_{j1}, \ldots, x_{jp}$:

$$\hat{y}_j = b_0 + b_1 x_{j1} + \ldots + b_k x_{jk}. \qquad (I2.2)$$

При выполнении регрессионного анализа предполагается,

что среди аргументов нет чрезмерно сильно коррелированных признаков и что среднее значение ошибки ε равняется нулю. При проверке гипотез предполагается дополнительно, что

I) случайная ошибка ε имеет нормальное распределение;

2) наблюдения независимы.

I2.I.2. Характеристики точности регрессионной модели и тесноты регрессионной зависимости. Анализ регрессионной зависимости начинается, как правило, с анализа статистик, характеризующих тесноту регрессионной зависимости и точность модели. Таких статистик процедура REGRESSION выдает три - это множественный коэффициент корреляции R(MULTIPLE R), его квадрат R^2, который имеет самостоятельное значение и называется коэффициентом детерминации (R SQUARE) и средняя стандартная ошибка (STANDARD ERROR). Множественный коэффициент корреляции - это по существу обыкновенный коэффициент корреляции между зависимым признаком У и его прогнозом $\hat{y}$ (см.I2.2). Множественный коэффициент корреляции изменяется в пределах от нуля до единицы, причем он равен нулю лишь в том случае, когда все аргументы некоррелированы с функцией, по ним невозможно прогнозировать У и, следовательно, все коэффициенты регрессии $b_1,\ldots b_p$ равняются нулю.

Множественный коэффициент корреляции равняется единице тогда, когда регрессионная функция (I2.I) определяет зависимый признак точно, т.е. когда $\varepsilon = 0$.

Коэффициент детерминации показывает, какая часть из дисперсии зависимого признака У описывается при помощи регрессионной функции (I2.I), часто коэффициент детерминации выражается в %-ах.

Средняя стандартная ошибка оценивает стандартное откло-

нение ошибки ε, т.е. показывает среднее отклонение наблюдения У от его прогноза $\hat{У}$.

I2.I.I.3. Стандартизированная модель. Наряду с регрессионной модели (I2.I) иногда рассматривается и т.н. стандартизированная модель

$$У_z = \beta_1 z_1 + \ldots + \beta_p z_p, \qquad (I2.3)$$

где все признаки стандартизированы, т.е. центрированы и нормированы.

Заметим, что ввиду центрированности модель (I2.3) не содержит свободного члена. По математическим характеристикам модели (I2.I) и (I2.2) эквивалентны: для них R общее, а по регрессионным коэффициентам одной всегда однозначно определяются и регрессионные коэффициенты второй, однако с точки зрения интерпретации их преимущества различны.

I2.I.I.4. Проверка гипотез при помощи F-статистик. Так как в данной главе гипотезы проверяются при помощи F-статистик, напомним здесь общие правила проверки гипотез при помощи F-статистик.

I) Прежде всего определяется уровень значимости α - это есть максимальная допустимая вероятность ошибки при доказательстве гипотезы H_1. Обычно выбирается $\alpha = 0.05$, иногда выбираются значения $\alpha = 0.0I$, $\alpha = 0.I$ или $\alpha = 0.025$. Больше чем $0.5\,\alpha$ никогда не выбирается, это дало бы бессмысленные по содержанию результаты.

2) Находится значение F-статистики, выпускаемое программой. Если это значение меньше единицы, то сразу принимается нулевая гипотеза: не удастся доказать гипотезы H_1 - и анализ (этап анализа) этим заканчивается.

3) Если значение F-статистики больше единицы, то про-

веряется, напечатана ли вероятность значимости этой статистики: (F PROBABILITY). Если эта вероятность P имеется, то сравниваем её с уровнем значимости:

если $P \leq \alpha$, то H_1 считается доказанной,

если $P > \alpha$, то принимается H_0.

4) Если вероятность P не выдана программой, то придется найти степени свободы статистики F (которые либо имеются в распечатке, либо следует их вычислять по определенным правилам в зависимости от конкретной задачи) и затем пользоваться таблицами статистики F. В этих таблицах разыскивается значение F_t (критическое значение F-теста), соответствующее заданным в распечатке степеням свободы (первый - степень свободы числителя - обычно в таблицах указаны на первом месте) и уровню значимости α.

Если $F \geq F_t$, то H_1 считается доказанной,

если $F < F_t$, то принимается H_0.

I2.I.I.5. Проверка значимости регрессионной модели. Значимость регрессионной модели (существование регрессионной зависимости) проверяется при помощи следующих гипотез:

H_1: найдется хотя бы один коэффициент регрессии b_i (i>1) так, что $b_i \neq 0$

H_0: $b_1 = \ldots = b_p = 0$

Для проверки этой гипотезы принимается F-статистика, расположенная в таблице дисперсионного анализа (ANALYSIS OF VARIANCE).

I2.I.I.6. Анализ регрессионной модели. Если удастся принять содержательную гипотезу H_1, то необходимо проанализировать регрессионную модель (I2.I) подробнее.

Коэффициенты регрессии b_i (точнее, их оценки по выбор-

ке) находятся в столбце B таблицы VARIABLES IN THE EQUATION, в последней строке этого столбца находится свободный член b_0 (CONSTANT).

При интерпретации коэффициентов регрессии полезно иметь в виду, что b_1 показывает, как сильно изменяется значение У, если X_1 увеличивается на одну единицу, а все остальные аргументы постоянны. Отсюда и видно, что b_1 зависят от единиц измерения соответствующих признаков X_i, и поэтому они, как правило, не сравнимы между собой.

Все коэффициенты B_1 стандартизированной модели (I2.3), (они находятся в столбце BETA) наоборот, сравнимы между собой. Если признаки $X_1,\ldots, X_p$ независимы, то можно точно сказать, что чем больше коэффициент β_i, тем больше влияние аргумента X_i на функцию У, причем $R^2 = B_1^2 +\ldots+ B^2_p$. При слабо зависимых аргументах можно оценить влияние аргументов X_1 на У по величине коэффициентов B_i, но следует иметь в виду, что все коэффициенты B_1 зависят не только от влияния X_1 на У, но и от взаимных влияний аргументов. При сильно зависимых аргументах исследование коэффициентов B_1 практически никакой дополнительной информации не дает.

В третьем столбце (STD ERROR B) находится стандартные отклонения (оценок) коэффициентов регрессии. Следует отметить, что чем меньше объем выборки, или чем больше зависимость между аргументами, тем больше стандартные отклонения коэффициентов регрессии, значит, тем меньше точность модели.

В последней столбце рассматриваемой таблицы находятся значения F-статистик, характеризующие значимость всех аргументов X_1.

Напомним, что из этого факта, что регрессионная зависи-

мость существенная (см. I2.I.I.5) отнюдь не вытекает, что все аргументы в регрессионной модели являются необходимыми. Следовательно возникает проблема проверки значимости отдельных аргументов в модели. F-статистики, выпускаемые в рассматриваемой таблице, характеризуют значимость каждого признака x_i на фоне всех остальных, т.е. проверяют гипотезу:

H_1: Модель (I2.I) описывает признак У существенно лучше, чем модель

$$b'_1x_1 +\ldots+ b'_{i-1}x_1 + b'_{i+1}x_1 +\ldots+ b'_px_p \qquad (I2.4)$$

H_0: Модель (I2.4) описывает признак У столь же хорошо, как и модель (I2.I).

I2.I.I.7. Упрощение регрессионной модели. Если после анализа (при помощи F-статистик) выясняется, что в модель включено слишком много аргументов, то следующим шагом будет упрощение модели путем удаления лишних аргументов.

Если в таблице имеется только один аргумент x_i, при котором принимается нулевая гипотеза о его несущественности в модели, то задача решается просто: следует оформить новый заказ без этого аргумента и полученная модель уже может быть удовлетворительной.

Если же таких аргументов существует больше, то удалить их всех из модели нельзя - в противном случае модель может существенно ухудшаться. В случае, когда предполагается, что в модели (I2.I) есть k лишних признаков, можно заказать соответствующую модель и при помощи F-статистики

$$F = (R^2_p - R^2_{p-k})(n - p - 1)/(1 - R^2_p)k, \qquad (I2.5)$$

которая вычисляется вручную по результатам распечаток (R^2_p и R^2_{p-k} соответственно коэффициенты детерминации полной и уп-

рощенной модели, p - число аргументов в полной модели и n - число наблюдений), проверяется гипотеза:

H_1: Полная модель (I2.I) описывает признак У существенно лучше, чем упрощенная модель, содержащая p-к аргументов.

H_0: Упрощенная модель описывает признак У столь же хорошо, как и модель (I2.I).

Обычно в ходе анализа конкретного материала приходится не только создать разные регрессионные модели, но и выбрать среди них наилучшие в математическом смысле, и имеющие наилучшим образом интерпретируемый содержательный смысл.

Для выбора подходящей модели целесообразно пользоваться той суммарной информацией, которая располагается в конце распечатки в форме суммарной таблицы (SUMMARY TABLE).

В процедуре REGRESSION имеется гибкая аппаратура для пользования разной априорной информации об ожидаемой модели, например, о признаках, которые надо обязательно включить в модель, о предполагаемом порядке включения признаков в модель и т.д.

I2.I.I.8. Пошаговая процедура нахождения регрессионной модели. Если исследователь имеет большую группу потенциальных аргументов $x_1,\ldots,x_q$ для регрессионной модели, причем нет существенной дополнительной информации для выбора аргументов, то наиболее подходящим методом для нахождения наиболее оптимальной модели (или нескольких допустимых моделей) является пошаговой регрессионный анализ.

При пошаговой процедуре аргументы включаются в модель подряд в таком порядке, что всегда прибавляемый аргумент описывает наибольшую часть дисперсии У (или, что то же самое - обуславливает наибольшее увеличение коэффициента де-

терминации).

Пошаговая процедура управляется при помощи трех необязательных параметров - это максимальное число аргументов в модели, критическое значение F при прибавлении аргумента (F вычисляется по формуле (I2.5) в случае k=1) и критическое значение толерантности. Толерантность - это та часть дисперсии прибавляемого признака X_i, которая не описывается через признаки $X_1,\ldots,X_{i-1}$ при помощи линейной модели типа (I2.I). Понятно, что при сильной зависимости аргументов толерантность мала, и ограничением толерантности запрещается включение в модель таких аргументов, которые очень сильно зависят от других и тем самым ухудшают качество и точность прогноза (в результате увеличения ошибок коэффициентов регрессии).

На каждом шагу выпускается информация как об аргументах уже входящих в модель, соответствующих регрессионных коэффициентах, характеризующих точность модели и т.д., но также и информация о тех признаках, которые еще не включены в модель - указывается соответствующий стандартизованный коэффициент регрессии и толерантность.

Вся эта информация применима при создании окончательной модели (которой может служить как модель, полученная на некотором шаге, так и совсем другая модель в зависимости об имеющейся априорной информации).

Хотя на каждом шагу пошаговая процедура прибавляет наиболее подходящий аргумент модели, этим еще не гарантируется, что полученная на k-ом шаге модель является оптимальной k-аргументной моделью. Как правило, она все же является довольно близкой к оптимальной модели.

Эта ситуация типична для задач анализа данных, при ко-

торых следует учитывать следующие общие обстоятельства:

I) полученная модель, как правило, не является оптимальной, а одной из многих, практически эквивалентных моделей, которые могут включать в себе разные подгруппы множества потенциальных аргументов;

2) часто можно получить более высокий коэффициент детерминации включением в модель вместо аргументы x_1 некоторые их функций. Такое усложнение модели все-таки допустимо лишь при предположении достаточно большого объема выборки и достаточно высокой точности измерения признаков x_1 и У.

I2.I.I.9. Анализ остатков прогноза. На основании регрессионной модели можно заказать в качестве дополнительных статистик значения прогнозов $\hat{y}_j$ по формуле (I2.2) и остатки прогнозов $\hat{\varepsilon}_j = \hat{y}_j - y_j$ (y_j - измеренное значение признака У у объекта с номером j).

Анализ поведения остатков прогноза позволяет проверять

I) выполнение предположений регрессионного анализа;

2) линейность функции регрессии.

В основном для анализа остатков пользуются графиками (диаграммами рассеяния).

Диаграммы рассеяния, где остатки прогноза $\hat{\varepsilon}_j$ рассматриваются зависимыми от их порядковых номеров j, позволяет визуально проверять гипотезы о нормальности и независимости ошибки ε: при таких предположениях семейство точек образует полосу приблизительно равной ширины, причем в средней части (около оси $\hat{\varepsilon}_j = 0$) точки сгущены плотнее чем на краях полосы. Если ширина полосы заметно изменяется, то дисперсия распределения изменяется со временем. Всякие другие отклонения от описанной формы указывают, как правило, на ненор-

мальности распределения ε.

Второй дополнительный график - это совместное распределение стандартизированных прогнозов $\hat{y}_j/\hat{\sigma}_y$ и $\hat{\varepsilon}_j$. И здесь расположение остатков прогноза $\hat{\varepsilon}_j$ должно быть, в основном, таким, как и в случае первого графика. Некоторые специальные сгущения точек на краях графика или в общем криволейная форма фигуры тесно сгущенных точек указывают на неадекватность регрессионной модели или на криволинейную зависимость функции У от аргументов $x_1,\ldots,x_p$. В последнем случае может оказаться целесообразным построение некоторой нелинейной регрессионной модели: заменить У или некоторые из аргументов на их функции, ввод в модель произведений $x_i x_j$ и т.д. (см. [I], стр. 95-I09).

Если возникают сомнения, что наблюдения, находящиеся подряд, коррелированы (наблюдения типа временного ряда), то можно заказать статистику Дубрин-Уатсона для проверки этой гипотезы.

I2.I.2. Процедурная карта

При обращении к процедуре нужно перечислить, какие признаки нужно использовать при составлении регрессионных уравнений и указать, какие регрессионные уравнения надо образовать. Общий вид процедурной карты следующий:

REGRESSION VAHIABLES=varlist/REGRESSION=имя[(n,F,T)] WITH varlist (порядок включения) [,...,varlist(порядок включения)] [RESID=Ø] / [.../REGRESSION=имя [(n,F,T)] WITH varlist (порядок включения) [,...,varlist (порядок включения)] [RESID=Ø]/

В поле управления карты записывается имя процедуры. Как обычно, при этом можно ограничиться первыми восемью символами,

т.е. REGRESSI. В поле описаний первым находится ключевое слово VARIABLES, за которым следует список всех тех признаков, которые нужно использовать при конструировании следующих функций регрессии. Затем следует ключевое слово REGRESSION, за которым после знака равенства указывается вид желаемой функции регрессии и способ ее конструирования. Первым при описании функции регрессии указывается имя зависимого признака. За этим именем при использовании пошаговой регрессии может следовать в скобках до трех необязательных параметров: n - число включаемых аргументов, F - критическое значение для проверки существенности включаемого аргумента и T - критическое значение толерантности. Если значения этих параметров при пошаговой регрессии не указывается, то по умолчанию используют следующие значения: n = 80, F = = 0.01 и T = 0.001. При надобности можно определить значение только первого или первого и второго параметра, а значение остальных выбирать по умолчанию.

За именем зависимого признака должно следовать ключевое слово WITH, а за ним - список аргументов регрессионной функции. Список аргументов может быть разделен на подсписки, за каждым подсписком обязательно указывается порядок его включения - целое число, которое определяет очередь и способ включения перечисленных аргументов. Если номером порядка включения выбрано четное число, то включаются все аргументы из этого подсписка одновременно. Но это будет сделано только после того, как все аргументы или списки аргументов с меньшими номерами порядка включения уже включены. Если номером порядка включения выбрано нечетное число, то аргументы этого списка включаются шаг за шагом, при этом включается толь-

ко те аргументы, которые удовлетворяют критерию, определенному параметрами n, F и T. Такое пошаговое включение начинается только после того, как все группы аргументов, имеющих четный номер включения, уже включены, а также включены шаг за шагом группы аргументов с меньшим нечетным номером порядка включения. Объясним вышесказанное примером. Пусть, например, описание регрессионной функции следующее: REGRESSION=Y WITH XZQ(6) ABD (1) W (4)/. Конструирование регрессионнной функции начинается с функции $У = b_0 + b_1W$. После нахождения и вывода параметров этой функции, включается группа аргументов, следующий по величине номером уровня включения, т.е. образуется функция $y = b_0+b_1W+b_2X+b_3Z+b_4Q$. После нахождения и вывода параметров этой функции в регрессионное уравнение включаются шаг за шагом те аргументы, которые находятся в группе аргументов с нечетным номером и удовлетворяют критерию, определенному значениями параметров n, F и T по умолчанию. После включения каждого нового аргумента, выводятся параметры для полученной функции регрессии.

Последним в определении уравнения регрессии можно включить необязательное ключевое слово RESID, за которым после знака равенства следует нуль. Этим словом заказывается вычисление остатков прогноза, которые можно по-разному использовать при вычислении некоторых дополнительных статистик.

Списки VARIABLES вместе со следующими описаниями функции регрессии можно повторить, и таким образом, при одном обращении к процедуре REGRESSION заказать несколько различных уравнений регрессии.

I2.I.3. Опции

Основная опция процедуры работает с полной выборкой. Для каждой заказанной регрессионной функции выводятся характеристики тесноты и точности связи, таблица дисперсионного анализа, а также таблица со значениями всех коэффициентов регрессии и их характеристики (VARIABLES IN THE EQUATION). При использовании пошаговой регрессии выводится определенная информация для признаков, не включенных в регрессионное уравнение. Последней выводится сводная таблица (SUMMARY TABLE), в которой для каждого признака указывается достигнутое при его включении значение коэффициента детерминации и соответствующий аргументу коэффициент в конечном уравнении.

За картой REGRESSION может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Возможные опции следующие:

I) обозначения пропуска используются как значения признака;

2)объект не используется при вычислении коэффициента корреляции между двумя признаками, если у него отсутствует значение одного используемого признака, при выключении других коэффициентов корреляции он может участвовать. Нужно заметить, что полученное решение может оказаться статистически некорректным, поэтому пользоваться опцией нежелательно;

3) на распечатке не указываются метки;

4) в качестве начальных данных используется корреляционная матрица, адрес которой определяется оператором DD с именем FT08F001. При введении этой матрицы нужно использовать дополнительные управляющие карты, описанные в пункте I0.3.3;

5) перед корреляционной матрицей вводятся средние значения и стандартные отклонения признаков в формате 8F1Ø.Ø. Так как средние значения и стандартные отклонения не являются целыми числами, при их представлении обязательно нужно использовать десятичную точку;

6) выводится только сводная таблица, а поэтапный вывод подавляется;

7) печатается только поэтапная часть вывода, вывод сводной таблицы подавляется;

8) отсутствует;

9) в качестве начальных данных используется корреляционная матрица, при этом вводится единая корреляционная матрица для целого списка VARIABLES, адрес которого определяется оператором DD с именем FTØ8FØØ1. Элементы корреляционной матрицы представляются в формате F1Ø.7. Опцию можно использовать только вместе с опцией 4;

I0) из прогнозов образуется самостоятельный файл, адрес которого определяется оператором DD с именем FTØ9FØØ1. В файл выводятся еще порядковый номер объекта в подфайле, номер записи и имя подфайла;

II) выводятся и стандартизированные прогнозы;

I2) запрещается вывод стандартизированных остатков в выходном файле;

I3) отсутствует;

I4) на графике стандартизированных прогнозов и стандартизированных остатков (статистика 6) печатаются осевые линии;

I5) средние значения и стандартные отклонения признаков выводят в самостоятельный файл в формате F1Ø.4.

I2.I.4. Статистики

За картой REGRESSION может следовать карта STATISTICS, в описательном поле которой перечисляются номера желаемых дополнительных статистик. Возможны следующие статистики:

I) для каждого списка VARIABLES печатается корреляционная матрица;

2) для каждого списка VARIABLES печатаются средние значения, стандартные отклонения признаков и численности измеренных объектов;

3) печатается корреляционная матрица, где отдельно указываются те корреляции, вычисление которых невозможно из-за отсутствии объектов или постоянства признаков. Такая информация обращает внимание использователя на признаки, которые в следующем заказе желательно исключить;

4) выводится диаграмма рассеяния остатков прогноза и номеров объектов, а также значения прогноза и остатков прогноза; но последние выводятся только в том случае, когда они запрошены при помощи ключевого слова RESID при описании уравнении регрессии;

5) вычисляется статистика Дурбин-Уатсона;

6) выводится диаграмма рассеяния для стандартизированных прогнозов и остатков прогнозов;

7) печатается корреляционная матрица, где кроме коэффициентов корреляции указываются численности объектов, использованных для их подсчета (можно заказать вместе с опцией 2).

Если используются все статистики, то в описательном поле карты STATISTICS можно написать ключевое слово ALL.

I2.I.5. Пример

Пусть требуется получить уравнение регрессии между зависимым признаком УУ и аргументами А1, А2,...,А5. При этом необходимо получить остатки прогноза и распечатку диаграммы совместного рассеяния стандартизированных прогнозов и остатков прогноза. Кроме этого требуется построить и такое уравнение регрессии, в которое обязательно вошли бы аргументы А1 и А2, а остальные аргументы выбирались среди признаков Т1, Т2,...,Т20.

Получаемое уравнение регрессии не должно содержать больше чем шесть аргументов. Такой заказ может быть представлен с помощью следующих карт:

```
REGRESSION   VARIABLES=YY A1 TO A5, T1 TO T2Ø/
             REGRESSION=YY WITH A1 A2 A3 A4 A5(2) RESID=Ø/
             REGRESSION=YY(6) WITH A1 A2(2) T1 TO T2Ø(1)/
STATISTICS   6
```

I2.2. Процедура DISCRIMINANT

I2.2.I. Цель и методические указания

I2.2.I.I. Общая модель дискриминантного анализа. Дискриминантный анализ - это методика многомерного статистического анализа, предназначенная для описания известных подсовокупностей (групп) данной совокупности объектов и для построения правил (дискриминантных функций) для группировки негруппированных объектов. Для решения этой задачи в системе САИСИ имеется процедура DISCRIMINANT. Заметим, что аналогичную задачу (но с меньшим объемом выпускаемой информации) можно решать и при помощи процедуры MULTIVARIANT, описанной в главе II.

Для решения задачи дискриминантного анализа предполага-

ется, что существует группирующий признак (имеющий k значений), разделяющий исследуемый массив данных на k подсовокупностей. Кроме того, измерены (количественные) признаки $x_1,...,x_p$, которые каким-то образом характеризуют имеющиеся группы. Признаки x_i будем в дальнейшем называть аргументами.

В ходе дискриминантного анализа определяются дискриминант-функции (дискриминирующие функции), при помощи которых можно решать задачу о принадлежности конкретного объекта в ту или другую группу. Число, форма и конкретное значение этих дискриминант-функций зависит от конкретного метода дискриминантного анализа.

В процедуре DISCRIMINANT рассматривается линейный дискриминантный анализ, при котором дискриминантные функции имеют вид

$$w_1x_1 + w_2x_2 + \ldots + w_px_p \ . \qquad (I2.6)$$

Процедура дает средства для решения следующих задач:

I) описание имеющихся групп, выделение из множества аргумент-признаков $x_1,...,x_p$ те, которые наилучшим образом описывают различия между группами;

2) проверка гипотезы о том, что имеющиеся группы существенно различаются друг от друга (по средним значениям аргумент-признаков);

3) построение линейного решающего правила для классификации и переклассификации объектов по заданным k группам.

Для решения описательной части задач I) и 3) никаких дополнительных предположений делать не надо, но для проверки всех гипотез предполагается, что аргумент-признаки $x_1,...,x_p$ имеют многомерное нормальное распределение.

I2.2.I.2. Метод пошаговой дискриминации. Основанием

процедуры DISCRIMINANT является идея пошаговой дискриминации (по существу аналогична пошаговой регрессии, описанной в предыдущем параграфе), при которой в дискриминант-функции аргументы x_i включаются один за другим. На каждом шагу выбирается тот аргумент, который имеет максимальную дискриминирующую способность.

Основной характеристикой расположения точек $x=(x_1,\ldots,x_q)$ и $y=(y_1,\ldots,y_q)$ в q-мерном подпространстве всех аргументов-признаков $X_1,\ldots,X_q$ является их расстояние Махаланобиса

$$D^2(x,y) = (x-y)'W^{-1}(x-y), \qquad (12.7)$$

где W - дисперсионная матрица, характеризующая распределение точек x и y (предполагаются, что они имеют одинаковое распределение). В процедуре MULTIVARIANT это расстояние обозначалось символом $D^2_4(x,y)$.

Суммарная вариабельность массива данных определяется как сумма квадратов расстояний Махаланобиса всех объектов $x_j=(x_{j1},\ldots,x_{jq})$ до их общего среднего $\bar{x}=(x_1,\ldots,x_q)$. Эта суммарная вариабельность разделяется на две части (таким же образом как и в регрессионном и дисперсионном анализе, см. процедуры REGRESSION и ONEWAY). Первая часть - эта вариабельность между группами, измеряемая посредством расстояний средних групп $\bar{x}_i$ до общего среднего. Вторая часть - внутригрупповая вариабельность, которая равняется сумме всех квадратов расстояний отдельных объектов x_j до средних соответствующих групп. Отношение этих частей вариабельности является характеристикой удачности выбора групп или (что то же самое) - дискриминирующей способности аргумент-признаков: чем меньше доля внутригрупповой вариабельности сравнению

с общей вариабельности, тем больше дискриминирующая способность аргументов $x_1,\ldots,x_q$. Это отношение называется U-статистикой (она известна и под названием Λ Уилкса).

I2.2.I.3. Проверка гипотез о дискриминирующей способности аргументов. На каждом (q-м) шаге, когда в функцию (I2.6) включается признак $x_{q\,м}$ выпускается ряд статистик, позволяющих проверять гипотезы о дискриминирующей способности признаков и о существенности различий между группами. Первым является нулевой шаг, когда в модель не включен ни один признак. Выдаются следующие статистики:

I) F-удаления (F'S TO REMOVE) для всех q признаков $x_1,\ldots,x_q$, включенных в функцию (I2.6). Они характеризуют дополнительную дискриминирующую способность признака x_r $(r=1,\ldots,q)$ на фоне остальных включенных в функцию признаков $x_1,\ldots,x_{r-1}$, x_{r+1} ,$\ldots,x_q$ и применяются для проверки гипотез:

H_1^r: дискриминант-функции

$$w_1x_1+ \ldots + w_rx_r+ \ldots + w_qx_q \qquad (I2.8)$$

дискриминируют имеющиеся группы существенно лучше, чем дискриминант-функции

$$w_1x_1+\ldots+ w_{r-1}x_{r-1}+ w_{r+1}x_{r+1} +\ldots+ w_qx_q \qquad (I2.9)$$

при

H_0^r: дискриминирующие функции (I2.8) и (I2.9) дискриминируют группы одинаковым образом; r = 1,...,q.

Процедура проверки гипотез при помощи выпускаемого значения F-статистики происходит стандартным образом, см. пункт I2.I.I.

2) F-включения (F'S TO ENTER) для всех p-q признаков, $x_{q+1},\ldots,x_p$, не включенных в функцию (I2.6). Они характери-

зуют дискриминирующую способность признаков x_s ($s=q+1,\dots,p$) на фоне уже включенных в функцию признаков $x_1,\dots,x_q$, и применяются для проверки гипотез:

H_1^s: дискриминант-функции

$$w_1'x_1+\dots+ w_q''x_q + w_s''x_s \qquad (12.10)$$

дискриминируют имеющиеся группы существенно лучше, чем дискриминант-функции (12.8).

H_0^s: дискриминант-функции (12.8) и (12.10) дискриминируют группы одинаковым образом.

По значениям F-включения выбирается аргумент, включаемый в модель на следующем, $q+1$ шаге: это именно тот x_s, при котором значение F-статистики наибольшее; $s = q+1,\dots,p$.

Заметим еще, что высказанные утверждения необязательно должны быть правильными относительно одной дискриминант-функции, характеризующей некоторую конкретную группу с номером l ($1\leqslant l\leqslant k$), но они касаются совокупности всех дискриминирующих функций, позволяющих дискриминировать все k группы. Кроме того, разумеется, даже при фиксированной функции, как правило, $w_i \neq w_i''$.

12.2.1.4. Проверка гипотез о различии между группами. Кроме описанных F-статистик на каждом шаге выпускается и значение U-статистики для проверки гипотез:

H_1: группы различаются друг от друга: найдутся индексы l и h ($1\leqslant l, h\leqslant k$) так, что

$$\mu_l \neq \mu_h;$$

H_0: группы не различаются друг от друга:

$$\mu_1 = \mu_2 = \dots = \mu_k;$$

здесь $\mu_i = (\mu_1^i,\dots,\mu_q^i)$ - средняя признак-вектора $x_1,\dots,x_q$

в группе с индексом 1.

Ввиду отсутствия таблиц критических значений U-статистики, обычно для эффективного выполнения поставленной задачи пользуются известной функцией U-статистики, которая имеет асимптотически F-распределение со специально вычисленными (обычно не целочисленными) значениями степеней свободы. Эти значения даются под названием приближенная F (F APPROXIMATE) и прилагаются соответствующие степени свободы (DEGREES OF FREEDOM). Для проверки гипотез можно пользоваться целой частью дробных степеней свободы (или пользоваться формулой линейной интерполации, если это необходимо).

Для выяснения различий между парами отдельных групп (с индексами 1 и h; 1,h =1,...,k) на каждом шагу выпускается F-матрица (F-MATRIX), причем число степеней свободы для всех статистик F в этой матрице равны и заданы под заглавием (DEGREES OF FREEDOM). F-статистика, находящаяся в I-м столбце и h-ой строке дает возможность для проверки пар гипотез:

$H_1^{1,h}: \mu_1 \neq \mu_h$

$H_0^{1,h}: \mu_1 = \mu_h$

пользуясь стандартной методикой.

I2.2.I.5. Исследование дискриминантных функций. Если после p-го (последнего) шага все средние μ_1 оказались существенно различными друг от друга, то следует этап анализа дискриминантных функций. Определение последних базируется на расстоянии Махаланобиса (I2.6): естественным образом точка x считается принадлежащей в ту группу с индексом 1, среднее (центр) $\bar{x}_1$ которой находится к точке x ближе, чем цент-

ры других групп.

На практике обычно вместо того, чтобы вычислить для каждой точки x все k расстояний до центров групп, вычисляют значения эквивалентных расстояниям линейных функций - т.н. информантов S_l, определяемых следующей формулой:

$$S_l(x) = x\,W^{-1}\bar{x}_l - 0.5\bar{x}_l W^{-1}\bar{x}_l \qquad (l=1,\ldots k) \qquad (I.I2.$$

где l - индекс группы; объект x считается принадлежащим в группу l, если $s_l(x)$ имеет максимальное значение. Так как по существу информант имеет общую форму линейной дискриминант-функции (I2.6):

$$S_l(x) = w_{l1}x_1 + \ldots + w_{lp}x_p + w_{l0},$$

то соответствующие коэффициенты даются в виде (p+1)xk матрицы под названием коэффициентов дискриминантных функций (DISCRIMINANT SCORES).

Если исследователя интересует вопрос - находится ли объект x ближе к группе l, или к группе h, то можно рассмотреть дискриминирующую функцию $\lambda_{lh}(x)$, разделяющую группы с индексами l и h. Эта функция определяется как разность функций S_l и S_h: $\lambda_{lh}(x) = S_l(x) - S_h(x)$ (при $l<h$).

О близости объекта к этим группам решают просто по знаку дискриминирующей функции:

если $\lambda_{lh}(x) > 0$, то x ближе к группе l,

если $\lambda_{lh}(x) < 0$, то x ближе к группе h.

При нулевом значении $\lambda_{lh}(x)$, разумеется, x находится на равном расстоянии от обеих групп.

При варианте 2 в качестве дополнительных статистик для каждого объекта дается:

I) его значение группирующего признака;

2) группа, к которой он причисляется по правилу дискриминации;

3) расстояния Махаланобиса до центра каждой группы;

4) апостериорные вероятности принадлежности в каждую группу.

Эти апостериорные вероятности вычисляются в предположении, что признаки $x_1,\ldots,x_p$ во всех группах имеют p-мерное нормальное распределение с равной ковариационной матрицей. Дискриминация по расстоянию Махаланобиса эквивалентна дискриминации по принципу максимального правдоподобия (каждый объект относится в ту группу, куда он попадает с наибольшей вероятностью). Разумеется, что сумма таких апостериорных вероятностей может быть как больше единицы (объект находится близко к некоторым группам), так и меньше единицы (объект находится далеко от всех групп).

I2.2.I.6. Методика проведения дискриминантного анализа. Распечатка процедуры DISCRIMINANT пакета САИСИ отличается очень большим объемом информации, причем существует сравнительно мало возможностей управлять этим количеством. При пошаговой процедуре значения управляющих статистик (F-включения, F-удаления, толерантность и задаваемое число групп) определяются программой автоматически таким образом (например, значения F-статистик равняются 0.0I и 0.005), что процесс практически никогда не прекращается до того, как в дискриминант-функцию не будут включены все потенциальные аргументы $x_1,\ldots,x_p$. Учитывая этот факт, можно рекомендовать при весьма больших совокупностях потенциальных аргументов заказать на первом этапе вариант 6 дискриминантного анализа, который выдает только сводные данные пошаго-

вой процедуры дискриминации.

Выбирая из этой таблицы те первые включенные аргументы, при которых значения статистики F-включения были достаточно большими (существенными), можно сделать следующий заказ, результаты которого следует исследовать уже более подробно (например, проверять, не надо ли удалить некоторые аргументы по значению F-удаления). Кроме того, на этом шагу можно проверять и то, различимы ли все группы, или, может быть, следует некоторые из них объединить.

После некоторых попыток обычно удастся найти удовлетворяющий исследователя комплекс групп и дискриминирующих признаков. В таком случае целесообразно заказать и все дополнительные статистики (вариант 2), позволяющие описать результаты в полном объеме.

I2.2.2. Процедурная карта

При обращении к процедуре указывают имя группирующего признака и имена признаков, описывающих группы. Общий вид карты следующий:

DISCRIMINANT имя WITH varlist [LEVELS(n_1,...,n_k)]

В управляющем поле находится имя процедуры; как обычно, можно его представить при помощи первых 8 символов, т.е. в виде DISCRIMI. В поле описаний первым находится имя группирующего признака, за которым следует ключевое слово WITH а за ним написаны имена описательных признаков. Необязательным является ключевое слово LEVELS, за которым в скобках следует одно или несколько целых чисел. Этими числами определяются номера тех шагов, на которых желают вывести коэффициенты вычисления информантов и коэффициенты вычисления дискриминантных функций.

Особенностью процедурной карты является то обстоятельство, что на ней нельзя использовать наклонную черту, при ее использовании выводится немедленно сообщение об ошибке.

I2.2.3. Опции

Основная опция процедуры работает с полной выборкой. На каждом шагу для всех используемых признаков выводятся F-статистики, описывающие влияние выбрасывания данного признака (F'S TO REMOVE) и для всех пока не используемых признаков F-статистики, описывающие влияние включения данного признака (F'S TO ENTER). Кроме того, выводятся U- и F-статистики, описывающие суммарное различие между групповыми средними и F-матрица, описывающая различия между разными парами групп. После включения всех описательных признаков выводятся коэффициенты информантов (DISCRIMINANT SCORE) и разности коэффициентов информантов для всевозможных разных пар групп. Распечатка заканчивается сводным протоколом выбора признаков, где указано порядок включения признаков, значения F-статистик, описывающие влияние включения данного признака и U-статистики. На распечатке указываются и метки признаков. За картой DISCRIMINANT может следовать карта OPTIONS, в поле описаний которой перечисляются номера желаемых опций. Можно выбирать следующие опции:

I) не выводятся метки;

2) для всех объектов выводятся номер группы, номер группы с наибольшей апостериорной вероятностью, расстояния Махаланобиса и апостериорные вероятности для всех групп;

3) отсутствует;

4) обозначения пропуска используются как значения признака;

5) объект не используется при вычислении ковариации между двумя признаками, если у него отсутствует значение одного используемого признака, при вычислении остальных ковариаций он может участвовать. F-матрица парных сравнений групповых средних не выдается;

6) выводятся только коэффициенты информантов, разности коэффициентов информантов для всех разных пар групп и сводный протокол выбора признаков;

7) не выводятся коэффициенты информантов и их разности.

I2.2.4. Статистики

За картой DISCRIMINANT может следовать карта STATISTICS, в поле описаний которой перечисляются номера желаемых дополнительных статистик. Можно выбирать следующие статистики:

I) средние значения для всех признаков во всех группах, всеобщие средние значения признаков;

2) стандартные отклонения для всех признаков во всех группах;

3) общая внутригрупповая ковариационная матрица;

4) общая внутригрупповая корреляционная матрица.

Если одновременно желают получить значения всех дополнительных статистик, то в поле описаний карты STATISTICS можно написать ключевое слово ALL.

I2.2. 5. Пример

Пусть группирующим признаком является признак с именем СПЕЦ, а описательными - признаки X1, X2,...,X7. В ходе предшествующего анализа выдвинута гипотеза, что для описания групп существенны только пять из этих признаков, поэтому заказывают коэффициенты информантов и их разности после включения пятого признака. Кроме обычных результатов нужно выво-

дить и результаты перегруппировки при помощи информантов и апостериорные вероятности. Заказ представляется следующими картами:

```
DISCRIMINANT  СПЕЦ  WITH X1 TO X7 LEVELS(5)
OPTIONS       2
STATISTICS    ALL
```

I2.3. Процедура TIME SERIES

I2.3.I. Цель и методические указания

I2.3.I. Процедура TIME SERIES предназначена для обработки временных рядов. В некотором смысле эта процедура отличается от всех остальных процедур пакета САИСИ - именно, здесь используются начальные данные иного типа, чем во всех предыдущих процедурах, так как обрабатывается временный ряд. Временным рядом называется последовательность значений одного признака , измеренных в различные моменты времени. Предполагается, что измерения проделаны через равные промежутки времени, в этом случае без ограничения общности можно момент измерения считать равным порядковому номеру данного измерения.

Целью анализа временного ряда является выбор математической модели для описания временного ряда, интерпретация особенностей ряда при помощи этой модели и прогнозирование будущих значений временного ряда.

Процедура TIME SERIES дает возможность использовать три разные модели для временного ряда - регрессионную модель, модель скользящей средней и модель периодограммы.

I2.3.I.2. При составлении регрессионной модели предполагается, что между результатом измерения x_t и моментом измерения существует обыкновенная линейная регрессионная связь

$$x_t = a + bt + \varepsilon_t,$$

где a и b неизвестные константы, а ε_t - случайное слагаемое. Слагаемое a+bt, линейно зависящее от времени, описывает постоянную тенденцию развития временного ряда, называемой трендом.

Процедура выводит оценки параметров a и b, средняя стандартная ошибка (оценка стандартного отклонения случайного слагаемого) и требуемое количество прогнозов на будущее $x_{T+1},\ldots,x_{T+K}$, определенных оцененой регрессионной функцией.

I2.3.I.3. Постоянная тенденция развития временного ряда может быть завуалирована случайными и циклическими колебаниями. Наиболее распространенным и простым путем выявления тенденции развития является сглаживание временного ряда. Суть сглаживания сводится к замене фактических значений временного ряда расчетными, имеющие значительно меньшие колебания, чем исходные данные. Уменьшение колебаемостк позволяет тенденции развития появить себя более наглядно.

Одним из самых простых методов сглаживания является реализуемый в процедуре метод скользящих средних. При этом методе фактическое значение временного ряда заменяется скользящим средним - для каждых k последовательных значений временного ряда подсчитывается среднее значение. Первое среднее значение вычисляется для $x_1,\ldots,x_k$, следующее для $x_2,\ldots,x_{k+1}$ и т.д. Таким образом интервал сглаживания, т.е. интервал, для которого подсчитывается среднее, как-бы скользит по временному ряду с шагом, равным единице.

Число элементов временного ряда, используемых для вычисления скользящей средней (т.е. длину "скользящего" интер-

вала) определяет заказчик. Чаще всего сглаживание производят по трем, пяти или семи элементам; тем выше колебаемость ряда, тем шире берется интервал сглаживания. Если можно предполагать, что изучаемому временному ряду свойственны циклические колебания, повторяющиеся через l моментов времени, то разумно вычислять скользящее среднее по l элементам.

Процедура вычисляет скользящие средние по данному числу элементов (скользящие средние первого порядка) и полученный ряд скользящих средних еще раз сглаживают скользящими средними при помощи скользящего интервала той-же длины (скользящие средние второго порядка). Можно заказать требуемое количество последовательных прогнозов $x_{T+1},\ldots,x_{T+k}$ на будущее, они определяются при помощи скользящих средних первого и второго порядка.

I2.3.I.4. Модель периодограммы предполагает, что результаты измерения представимы в виде мультипликативной модели, т.е.

$$x_t = S(t)\,T(t)C(t)\varepsilon_t,$$

где $S(t)$ - сезонная компонента, $T(t)$ - трендовая компонента, $C(t)$ - циклическая компонента, ε_t - случайная компонента.

Основные шаги при определении модели периодограммы следующие:

I) удаление трендовой и циклической компонент, вычисление грубых оценок сезонной компоненты;

2) сглаживание полученных грубых оценок при помощи анализа Фурье, проверка значимости коэффициентов Фурье;

3) замена выделяющих элементов сглаженного временного ряда;

4) проверка адекватности сезонной модели;

5) вычисление периодограммы.

Процедура выводит результаты каждого этапа образования модели.

I2.3.2. Процедурная карта

При обращении к процедуре нужно указать используемую модель и некоторые параметры этой модели. Общий вид процедурной карты следующий:

```
TIME SERIES    {REGRSSN|MOVAVRGE|PERIODGR},
               OBSERV=значение,[PERIOD=значение.]
               [SIZE=значение]
```

В управляющем поле карты находится имя процедуры. В поле описаний первым нужно указать ключевое слово, определяющее тип используемой модели. В случае регрессионной модели используют ключевое слово REGRESSN, в случае модели скользящей средней ключевое слово - MOVAVRGE, в случае модели периодограммы - ключевое слово - PERIODGR. Затем следует ключевое слово OBSERV, за котором после знака равенства, указывается длина обрабатываемого временного ряда. Параметром PERIOD определяется число последовательных прогнозов, которые требуется вычислить. Параметр SIZE имеет при разных моделях разное значение. При регрессионной модели этот параметр не используются, при модели скользящей средней он определяет длину скользящего интервала , а при модели периодограммы - период сезонных колебаний. Прогнозы при этой модели вычисляются всегда для всего периода.

На процедурной карте не используют наклонной черты, параметры разделяются запятой.

Чтобы не дать возможности заказчику забыть, что процедура использует начальные данные по своему характеру отлич-

ные от начальных данных всех остальных процедур, требования представления начальных данных здесь весьма своеобразные. Процедура не использует данных, записанных в системный файл, в каждом прогоне нужно снова с перфокарт ввести значения используемого временного ряда. Перфокарты с начальными данными размещают среди перфокарт операционной системы, они могут непосредственно предшествовать оператору DD с именем FT05F001. Перед начальными данными находятся всегда две стандартных карты:

```
//FT20F001 DD *
FORMAT          (5F10.0)
```

Начальные данные перфорируются по пять чисел на одной карте. Первое число начинается с нулевой, второе число - с десятой, третье число - с двадцатой, четвертое число - с тридцатой и пятое число - с сороковой колонки.

I2.3.3. Опции

Возможные опции этой процедуры связаны с конкретными моделями. Так при регрессионной модели возможна одна опция:

I) печатаются значения временного ряда и прогноз для всех моментов времени.

При модели периодограммы возможны следующие три опции:

I) печатаются значения временного ряда, скользящие средние и отношения элемента временного ряда на скользящее среднее;

2) печатаются отношения, усредненные по всем периодам;

3) печатаются значения временного ряда.

I2.3.4. Статистики

Свободно выбираемых статистик процедура TIME SERIES не имеет.

I2.3.5. Пример

Предположим, что временной ряд, содержащий I8 элементов, нужно сглаживать регрессионной моделью и вычислить прогноз для I9-го момента времени. Для представления этого заказа нужно приготовить следующую колоду карт:

```
//SAISIT JOB K-Ø2Ø,L.TELLIJA
//JOBLIB DD DSN=SAISI.LOAD,DISP=SHR
//S1 EXEC PGM=SAISI,PARM=2ØØK,REGION=5ØØK
//FTØ1FØØ1 DD UNIT=SYSDA,SPACE=(8ØØ,(1ØØ,5Ø))
//FTØ2FØØ1 DD UNIT=SYSDA,SPACE=(2Ø12,(2ØØ,4Ø))
//FTØ6FØØ1 DD SYSOUT=A
//FT2ØFØØ1 DD *
FORMAT          (5F1Ø.Ø)
55.3    64.7    79.5    62.6    5I.3
5I.Ø    54.Ø    45.2    44.9    42.4
4I.4    49.Ø    5Ø.8    65.8    85.9
7Ø.6    55.8    49.I
//FTØ5FØØ1 DD *
RUN NAME        АНАЛИЗ ВРЕМЕННОГО РЯДА
INPUT MEDIUM    CARD
TIME SERIES     REGRESSN,OBSERV=18,PERIOD=1
OPTIONS         1
FINISH
```

XIII. ПРИМЕР СОЗДАНИЯ ПЛАНА ОБРАБОТКИ ДАННЫХ И РЕАЛИЗАЦИИ ПЛАНА В ФОРМЕ ЗАКАЗОВ

Впервые пакет САИСИ был применен с целью систематического анализа данных во втором семестре 1983/84 учебного года в ходе спецкурса "АНАЛИЗ ДАННЫХ", излагаемого студентам IУ курса математического факультета ТГУ. Слушателями этого спецкурса были 51 студент по специальности "математика" и "прикладная математика" (эстонский и русский потоки).

Целью спецкурса являлось:

1) Усвоение основных понятий и методов анализа данных, базируясь на курс теории вероятностей и математической статистики, пройденный ранее.

2) Изучение самостоятельного использования математического обеспечения по анализу данных посредством специального языка управления (с этой целью был выбран пакет САИСИ).

3) Изучение корректного понимания статистического содержания результатов вычислений и их интерпретации в терминах рассматриваемой конкретной науки.

4) Изучение изложения результатов вычислений в форме обобщающего отчета.

Необходимый простой и хорошо интерпретируемый массив данных был получен в результате анкетного опроса всех студентов, проходивших рассматриваемый курс. Для этого все студенты заполнили опросный лист (см. приложение 1).

В некотором смысле типичный ход обработки этих анкет и излагается в качестве примера.

I3.I. Создание массива данных

Первым шагом обработки было создание правила кодирования и само кодирование. При создании правила кодирования учитывались:

I) содержательная упорядоченность ответов,

2) специфика массива опрошенных,

3) однотипные признаки.

В результате было получено следующее правило кодирования (излагается только часть):

Признак №2 - "Специальность": I - математика; 2 - прикладная математика, эстонский поток; 3 - прикладная математика, русский поток.

Признак №4 (место рождения) и №6 (место учебы в школе): I - деревня; 2 - маленький город, поселок; 3 - средний город; 4 - большой город.

Признак №8 (пол): I - мужской; 2 - женский.

Признак № 34 (любимое занятие): спорт (I), литература (2), рукоделье (3),музыка (4), походы (5), кино,театр (6), общение с друзями (7)

Значения номинальных признаков № 2 и № 34 заданы в произвольном порядке, также как значения дихотомического признака № 8. Зато значения признаков № 4 и № 6 упорядочены в порядке увеличения соответствующих населенных пунктов.

Для большинства остальных признаков кодами являются номера вариантов ответа, или, по существу, они количественные признаки, не требующие кодирования (например - год рождения, количество общественных обязанностей, средняя оценка). Первым признаком было ИМЯ, притом каждый студент выбрал себе "имя", длина которого была не более 4-х букв (АНДР, ДАНА,

ЭГЕ). Соответствующий признак - качественный.

Для всех численных признаков была определена точность, в данном случае это было для большинства признаков 0 (десятичных разрядов), для всех средних оценок I (десятичный разряд).

Кроме того для всех признаков были выбраны символы, обозначающие пропуск (пропущенное значение). Для некоторых признаков (год рождения, пол) символом пропуска можно было выбрать Ø, так как нуль не принадлежит области возможных значений этих признаков. Для таких признаков, как "число общественных обязанностей" Ø не пригоден в качестве обозначения пропуска, этим можно было выбрать, например, 99.

I3.2. Определение формата признаков

Формат первого признака ИМЯ - A4.

Все остальные признаки числовые и имеют F-формат.

Формат признаков № 2, № 4, № 6, № 8 и № 34 есть F1.Ø. т.е. они все кодированы целыми числами в интервале [Ø,9]. Признак № 3 "Год рождения", № 5 "Год окончания средней школы" и № 7 "Год вступления" имеют формат F2.Ø (указывают только две последние цифры, например 6I и 79), а признаки 22-28 "Средняя оценка"первой сессии" до "Средняя оценка седьмой сессии") требуют формат F2.1 (например 4.2).

Особенностью выбора формата в данном примере является то, что все признаки (кроме первого) считаются количественными и им приписывается F-формат. В результате этого для всех признаков все статистические процедуры формально допущены, и ответственность за корректность результатов возлагается полностью на исследователя.

Для менее опытных исследователей рекомендуется присвои-

вать номинальным признакам А-формат, например, для признака №2 А3, значениями являлись бы МАТ, ПМЭ, ПМР и т.д.

I3.3. Создание файла данных

Ввод данных в ЭВМ обычно состоит из двух шагов. На первом - данные пишут на бланку, а на втором - на носитель ЭВМ - перфокарту, магнитную ленту и т.д. Если имеется хорошая возможность визуального контроля (работа с дисплеем), то можно данные ввести сразу с опросных листов в память ЭВМ.

Во всяком случае этот этап работы требует тщательного соблюдения заданного формата и порядка признаков на опросном листе.

В настоящем примере все данные перфорировались на перфокарты.

Следующим шагом было создание описания данных. Для этого выбирались имена для файла, для признаков и для некоторых значений признаков. Для некоторых признаков прибавили еще объясняющие метки - "длинные имена".

Именем файла выбрали МАТИК.

Первый заказ - образование системного файла - выглядит следующим образом:

```
RUN NAME        ОБРАЗОВАНИЕ СИСТЕМНОГО ФАЙЛА
FILE NAME       МАТИК
VARIABLE LIST   ИМЯ,СПЕЦ,ГРОЖД,МРОЖД,ГОКОНЧ,МШКОЛ,ГВСТУНГ,ПОЛ,
                ...,СЕСС1 TO СЕСС 7
INPUT MEDIUM    CARD
# OF CASES      49
INPUT FORMAT    FIXED(A4,F1.Ø,F2.Ø,F1.Ø,F2.Ø,... )
VAR LABELS      ГОКОНЧ ГОД ОКОНЧАНИЯ ШКОЛЫ/МШКОЛ НАСЕЛ ПУНКТ,
                ГДЕ НАХОДИТСЯ ШКОЛА/
```

```
VALUE LABELS     МРОЖД,МШКОЛ (1)ДЕРЕВНЯ (2)МАЛГОРОД (3)СРЕГО-
                 РОД (4)БОЛГОРОД/
MISSING VALUES   СПЕЦ ТО ЛЮБЗАН (Ø)/ОБЩОБЪЯЗ,ПРОВЭКЗ,СНО,(99)/
LIST CASES       CASES=49/VARIABLES=ALL/
TASK NAME        ПРОВЕРКА
CONDESCRI        ALL
STATISTICS       ALL
READ INPUT DATA
Карты с данными
SAVE FILE
FINISH
```

Напомним коротко правила, которыми мы здесь пользовались.

Имя признака не содержит пробелов МШКОЛ (не М ШКОЛ), не начинается числом (СЕСС1 , не 1СЕСС), не содержит более 8 символов. То-соглашением можно объединить признаки, находящиеся в списке признаков подряд. Между именем и ключевым словом должен находиться не менее чем I разделитель. Разбивать имена на части, находящиеся на различных картах, нельзя.

Заказывают некоторые контролирующие выводы:

I) вывод всех данных для всех объектов (с помощью карты LIST CASES),

2) описательные статистики всех признаков (с помощью карты CONDESCRI, вместе с картой STATISTICS).

В конце заказа - требование сохранить системный файл в памяти ЭВМ.

I3.4. Анализ результатов первого этапа обработки

Целью анализа результатов первого заказа является обнаружение грубых ошибок и неточностей в массиве данных. Наиболее простой метод для этого - рассмотрение прежде всего

минимальных и максимальных значений всех признаков.

В данном случае мы нашли три грубых ошибки - максимальное значение признака РОСТ было 884 (в сантиметрах) максимальное значение признака ОБРМАТ (образование матери) было 93 (классов школы или курсов вуза) и минимальное значение признака МАТШКО (оценка математики в школе) было 0.5. Так как все значения всех объектов были распечатаны, сразу стало известным, при каких объектах эти ошибки имелись и какими должны быть правильные значения.

Кроме того, в данном случае (массив данных был небольшой) оказалось возможным проверять материал полностью путем визуального сравнения распечаток с исходными документами (опросные листы). Тогда оказалось, что у одного студента оценка первой сессии оказалось равной 3 вместо 5.

Кроме того, появились два студента, которые отсутствовали на первом занятии и их данные надо было прибавить к имеющемуся массиву данных.

В ходе проверки выяснилось еще, что признак "ВСТУПБЛЛ" (вступительный балл) был кодирован нецелесообразно: у медалистов это -10, у тех, которые поступили "по эксперименту" -13,5 до 15 у остальных от 15 до 24. Вместо года рождения желательно ввести более наглядный признак - возраст.

13.5. Редактирование системного файла

Для ввода всех исправлении, перечисленных в предыдущем пункте, образуется следующий заказ:

```
RUN NAME        РЕДАКТИРОВАНИЕ СИСТЕМНОГО ФАЙЛА
GET FILE        МАТИК
ADD SUBFILE     ОПОЗД
# OF CASES      2
```

```
INPUT FORMAT     FIXED(4A,F1.Ø,F2.Ø,...
RECODE           РОСТ(884=I84)/ОБРМАТ(93=I3)/МАТШКО(Ø.5=5)/
IF               (SEQNUM EQ 25)СЕСС1=5
COMPUTE          ВОЗРАСТ=84-ГРОЖД
IF               (ВСТУПБАЛ EQ 1Ø)ВСТУПБАЛ=25
IF               ((ВСТУПБАЛ GT 13) AND (ВСТУПБАЛ LE 15))
                 ВСТУПБАЛ =ВСТУПБАЛ+1Ø
ASSIGE MISSING   ВОЗРАСТ(Ø)
TASK NAME        ВТОРИЧНАЯ ПРОВЕРКА
CORRESCRI        РОСТ ОБРМАТ МАТШКО ВОЗРАСТ ВСТУПБАЛ
STATISTICS       ALL
HEAD INPOT DATA
```

(Следуют карты с данными опоздавших)

```
PROCESS SEFILES  ОПОЗД
LIST CASES       CASES=2
CORRESCRI        РОСТ
SAVE FILE
FINISH
```

Прокомментируем заказ.

Для устранения грубых ошибок проще всего применять процедуру RECODE. В результате этого все значения признака РОСТ, равняющиеся 884, заменяют на I84. Для устранения же случайных ошибок процедурой RECODE пользоваться нельзя - ведь преобразование всех оценок 3 в 5 недопустимо, так надо сделать только у одного объекта (именно того, который имеет порядковый номер 25). Поэтому случайные ошибки устраняются при помощи процедуры IF, так же как перекодируется частично неудачный признак ВСТУПБАЛ.

Заметим, что на описательном поле карты в скобках нахо-

дится логическое выражение, после скобок - имя признака, знак равенства и тогда арифметическое выражение; обе выражения могут содержать все уже имеющиеся признаки (но не новые).

В конце заказа проверяют все исправленные, преобразованные и добавленные признаки.

Напомним еще, что карты ввода данных находятся за картами первой статистической процедуры.

Так как в результате второго заказа в исходный файл введено много исправлений, то необходимо сохранить и этот файл.

Заметим, что если правильные значения признаков, имеющие грубые ошибки, не были известны, то их надо было бы считать пропущенными. Наиболее просто было бы это сделать картой MISSING VALUES:

```
MISSING VALUES    POCT(884) ОБРМАТ(93) МАТШКО(Ø.5)
```

Если имеется опасность, что грубых ошибок больше, то можно считать грубыми все "слишком большие" и "слишком маленькие" значения, например:

```
RECODE                POCT(210 THRU HIGHEST = Ø)
MISSING VALUES    POCT(Ø)
```

I3.6. План обработки

Когда проверка и редактирование массива данных завершена, уместно уточнить план обработки, так как наполненная в ходе первого этапа обработки дополнительная информация поможет исследовательно оформить содержательные гипотезы.

Целью настоящего исследования является описание результатов учебы и характеристик общественной активности студентов-старшекурсников математического факультета, а также выяснить факторов, влияющих на успеваемость в вузе. Выдвинуты следующие гипотезы:

I) успеваемость в вузе зависит от успеваемости в средней школе, от образования родителей;

2) успеваемость имеет положительную корреляцию с общественной активностью.

Представляет интерес найти некоторую кластеризацию (выявить типичные группы) среди исследуемых студентов.

I3.7. Образование новых признаков. Первичная обработка.

Для более подробного исследования успеваемости желательно определить некоторые новые признаки, такие как средняя оценка по всем сессиям, характеристики динамики результатов, а также признак, показывающий, сколько раз каждый студент был отличником.

Кроме того, для первичного анализа номинальных признаков необходимо применять такие процедуры, которые выпускают таблицы частот (первичный анализ количественных признаков сделан уже в ходе проверки данных - в результате процедуры CONDESCRI известны все описательные статистики всех количественных признаков).

Оформлен следующий заказ:

```
RUN NAME        ПЕРВИЧНАЯ ОБРАБОТКА НОВЫЕ ПРИЗНАКИ
GET FILE        МАТИК
COMPUTE         СРЕДНЕЕ = (СЕСС1+СЕСС2+СЕСС3+СЕСС4+СЕСС5+
                СЕСС6+СЕСС7)/7
COUNT           ОТЛИЧНИК=СЕСС1 ТО СЕСС7 (5)/
DO              Х=СЕСС1 ТО СЕСС6
                У=СЕСС2 ТО СЕСС7
                Z=РАЗН1 ТО РАЗН6
COMPUTE         Z=У-Х
DOEND
```

```
COMPUTE          ДИСКРБАЛ=ВСТУПБАЛ
RECODE           ДИСКРБАЛ(LO THRU 18=1)(18 THRU 20=2)(20
                 THRU 22=3)(22 THRU 24=4)(24 THRU 25=5)/
ASSIGN MISSING   СРЕДНЕЕ ТО ДИСКРБАЛ(99)
TASK NAME        ПРОВЕРКА НОВЫХ ПРИЗНАКОВ
CONDESCRI        СРЕДНЕЕ ТО ДИСКРБАЛ
STATISTICS       ALL
TASK NAME        РАСПРЕДЕЛЕНИЯ
CODEBOOK         СПЕЦ ТО ДИСКРБАЛ
OPTIONS          4
STATISTICS       3, 4
SORT CASES       СПЕЦ ПОЛ(A)/
SAVE FILE
FINISH
```

Прокомментируем заказ.

Для образования шести разностей применяется более экономическая DO...DOEND-процедура. Чтобы получить наглядные таблицы частот признака ВСТУПБАЛ, имеющего много различных значений, его дискретизировали, сохраняя исходной вариант (поэтому признак дублировали с помощью процедуры COMPUTE). Для всех новых признаков надо предвидеть возможность пропуска, их надо опять проверять.

Таблицы частот (с помощью процедуры CODEBOOK) образуются в основном для номинальных признаков. Для наглядности заказывают и гистограммы (OPTION 4), а описательные статистики (STATISTICS 1, 2) заказывать не надо, поскольку они уже найдены.

Для подготовки следующего этапа обработки сортируют объекты по специальности и полу (начиная с маленьких значе-

ний обоих признаков.

I3.8. Проверка влияния мешающих и группирующих признаков.

Целью следующего шага (второй этап первичной обработки) является проверка влияния известных мешающих и группирующих признаков на исследуемые признаки. От результатов этого этапа зависит, будет ли возможно в дальнейшем рассмотреть весь массив как один целый или придется исследовать самостоятельные подмассивы отдельно, ввиду их больших различий между собой.

Прежде всего нас интересует описание студентов разных специальностей в зависимости от их пола. Этого можно добиться, пользуясь процедурой AGGREGATE; для этого в конце предыдущего этапа работы объекты были соответственно упорядочены (при помощи процедуры SORT CASES).

При помощи процедуры AGGREGATE образуются новые признаки, значениями которых являются значения некоторых статистик, найденных для отдельных групп. Процедура AGGREGATE не дает возможности проверять гипотезы о различии групп, для этого применимы процедуры T-TEST, BREAKDOWN (FASTBREAK) и ONEWAY.

Составим заказ:

```
RUN NAME      ПРОВЕРКА ГОМОГЕННОСТИ
GET FILE      МАТИК
TASK NAME     ОБРАЗОВАНИЕ ГРУПП ПО ПОЛУ И СПЕЦИАЛЬНОСТИ
AGGREGATE     GROUPVARS = СПЕЦ, ПОЛ/VARIABLES = СРЕДНЕЕ/
              ACTIONS = NS, MEAN, SD, PCTGT(4)/
STATISTICS    2
```

```
TASK NAME      ВЛИЯНИЕ СПЕЦИАЛЬНОСТИ НА УСПЕВАЕМОСТЬ И
               УДОВЛЕТВОРЕННОСТЬ
ONEWAY         GROUPS (3) = СПЕЦ /VARIABLES = УДОВСПЕЦ,
               СРЕДНЕЕ/ RANGES=TUKEY/
STATISTICS     ALL
TASK NAME      ЗАВИСИМОСТЬ УСПЕВАЕМОСТИ ОТ СТАЖА
T-TEST         GROUPS = ГОКОНЧ(8Ø)/VARIABLES = СРЕДНЕЕ,
               ОТЛИЧНИК/
TASK NAME      ВЛИЯНИЕ МЕСТА УЧЕБЫ НА ВСТУПЛЕНИЕ В ВУЗ
BREAK DOWN     ВСТУПБАЛ ВУ МШКОЛ, МРОЖД, ПОЛ/
STATISTICS     ALL
FINISH
```

Прокомментируем заказ и проанализируем результат обработки. В результате процедуры **AGGREGATE** видно, что средняя успеваемость (признак СРЕДНЕЕ) является в рассматриваемых группах довольно различной (см. таблицу; в каждой клетке в скобках указаны численности).

СПЕЦ / ПОЛ	МАТ	ПРИКЛЭСТ	ПРИКЛРУС
МУЖ	3.45 (3)	3.58 (5)	4.03 (3)
ЖЕН	3.79(24)	4.I4 (9)	4.I3 (7)

Довольно большие различия есть между группами и по дисперсии данного признака, а также по удельному весу хорошо учащихся (СРЕДНЕЕ больше чем 4). Об этом, являются ли эти различия статистически существенны, можно судить на основании резуль-

татов процедуры ONEWAY.

Прежде всего, проверяем гипотезы о различии среднего значения признака СРЕДНЕЕ в разных группах специальности. На распечатке для всех групп задаются средние, их стандартные отклонения и 95% доверительные интервалы. Сразу видно, что доверительные интервалы заметно пересекаются - по ним невозможно доказать различие средних. Нам придется пользоваться значением F-статистики, которое может быть найдено из таблицы дисперсионного анализа. Так как это значение в данном случае равняется 2.408, а числа степеней свободы есть 2 (число групп - I) и 47 (число измеренных объектов - число групп), то вероятность значимости есть 0.099. Мы считаем везде уровень значимости равным 0.05, а так как 0.099> > 0.05, нам приходится признаться, что доказать существенного различия в средних признака СРЕДНЕЕ нам не удалось.

Также и тест гомогенности дисперсий (F Бартлета-Бокса) дает результат, соответствующий нулевой гипотезе: невозможно доказать, чтобы группы имели разные дисперсии признака СРЕДНЕЕ (для этого достаточно заметить, что вероятность P - больше уровня значимости 0.05).

Так как различия между группами не существенны, то, разумеется, и тест Тьюки оставляет все наблюдения в одной группе.

Также оказалось, что специальность не имеет существенного влияния на удовлетворенность выбранной специальностью.

Следующей задачей было исследование влияния года окончания средней школы на признак СРЕДНЕЕ. Чтобы получить достаточно многочисленные группы, мы разделили весь материал на две группы: окончившие школу в 80-м году и окончившие ее

раньше. Поэтому удалось использовать процедуру T-TEST.

Во-первых, оказалось, что окончившие школу до 80-го года учились пока лучше (их средняя оценка была 4.02) чем окончившие в 80 году (оценка 3.83). Для проверки гипотезы нам пришлось прежде всего рассмотреть значение F-отношения для проверки равенства дисперсии. Оказалось, что здесь вероятность значимости P является заметно большей, чем 0.05, и поэтому при проверке равенства средних следует пользоваться объединенной оценкой дисперсии (POOLED). В таком случае значение T-статистики равняется I.I4 и при числе степеней свободы 48 (число объектов - I) вероятность значимости - 0.26. Значит, разное время окончания средней школы не вызывает существенного различия в успеваемости в университете. Такой же результат мы получили и относительно средней частоты по получению повышенной стипендии.

Последняя задача проверки в настоящем заказе выяснение зависимости результатов учебы от разных социальных факторов: пола, места рождения, места учебы. Рассмотрим последнюю зависимость, исследуемую при помощи процедуры BREAKDOWN. Как и в случае процедуры ONEWAY, здесь выдаются средние и стандартные отклонения исследуемых признаков. Средний балл поступления зависит от места школы следующим образом:

большой город	2I.47
средний город	20.38
малый город	2I.46
село	2I.00

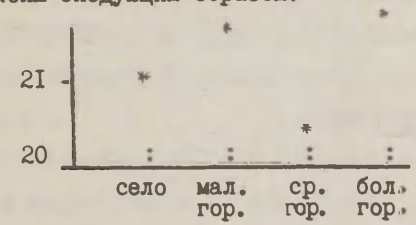

Схема I.

Из таблицы дисперсионного анализа видно, что значение F-статистики есть 0.555. Так как $0.555 < 1$, то следует принять нулевую гипотезу: средние не различаются друг от друга существенно.

В некотором условном смысле мы можем говорить и о линейности зависимости (по содержанию это ближе к монотонности). Проверка линейности осуществляется при помощи процедуры BREAKDOWN (статистика 2, при условии, что число групп $k > 2$) с помощью F-теста. В настоящем случае значение F-статистики есть 0.809, значит - принимается нулевая гипотеза - нет существенной нелинейной зависимости (хотя на первый взгляд график на схеме I может вызвать обратное мнение). Заметим, что значение корреляционного (регрессионного) отношения η в данном примере 0.184, коэффициент корреляции $R = -0.31$, и поэтому характеризующая нелинейность зависимости статистика $\eta^2 - R^2 = 0.034 - 0.001 = 0.033$ является достаточно малой.

Резюмируя настоящий этап следует сказать, что нам не удалось доказать ни одной содержательной гипотезы. Эта ситуация (впрочем, довольно типична на начальной стадии исследования) в данном случае благоприятна: наш материал не является явно негомогенным и в дальнейшем мы можем его рассмотреть как один целостный массив.

13.9. Анализ статистических зависимостей по группам признаков

Следующим этапом работы, относящимся условно к первичной обработке, является анализ статистических зависимостей по более крупным группам признаков. Для этого в основном годятся процедуры PEARSON CORR, NONPAR CORR и PARTIAL CORR.

Образуем заказ:

```
RUN NAME        ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ ВНУТРИ ГРУПП ПРИЗ-
                НАКОВ
GET FILE        МАТИК
TASK NAME       БОЛЬШАЯ КОРРЕЛЯЦИОННАЯ МАТРИЦА
PEARSON CORR    МШКОЛ,ОБРОТЦ,ОБРМАТ,СРОЦШКО ТО ССТРОЙ,
                ЛИТ,СПРТАКТ,МУЗ,СЕМ,СТАЖ/
TASK NAME       АНАЛИЗ ЗАВИСИМОСТЕЙ ОЦЕНОК
NONPAR CORR     СЕСС1 ТО СЕСС7/
OPTIONS         4,6
PEARSON CORR    РАЗН11 ТО РАЗН6
PARTIAL CORR    СЕСС1 ТО СЕСС7 ВУ СРОЦШКО  ПОЛ(1,2)/
FINISH
```

Прокомментируем заказ и полученные результаты.

В задаче "Большая корреляционная матрица" заказывается корреляционная матрица для большинства исследуемых признаков, которые желают включить в единую модель. В модель не вошли признаки, касающиеся большинства сторон семейной жизни студентов, а также их внешности, так как эти стороны материалы будут исследованы отдельно от общей проблематики.

Также и те мешающие и группирующие признаки, которые практически не влияют на исследуемые явления, не были включены в заказ. Кроме того, для вычисления матрицы обыкновенных коэффициентов корреляции, следует вычеркнуть из списка признаков все номинальные признаки (ЛЮБЗАН - любимое занятие). В списке останутся дихотомические (ПОЛ), а также порядковые, характеризующие, например, интенсивность занятий (ЛИТ, СПРТАКТ).

Заметим, что такой нестрогий подход к предположениям

вычислении корреляций допустим на данном этапе потому, что результаты первичного этапа имеют только описательный характер, никакие строго доказательные выводы (модели) сейчас еще не строятся.

Ввиду отсутствия карты STATISTICS выводится квадратичная корреляционная матрица; по техническим причинам она печатается по отдельным кускам, которые рекомендуется для облегчения анализа склеивать по схеме 2. Так как корреляционная матрица симметрична, то достаточно анализировать только часть, находящуюся под диагональю. Следует однако отметить, что некоторые методы анализа (построение графиков или дендрограмм) более удобно провести на основании квадратичной матрицы.

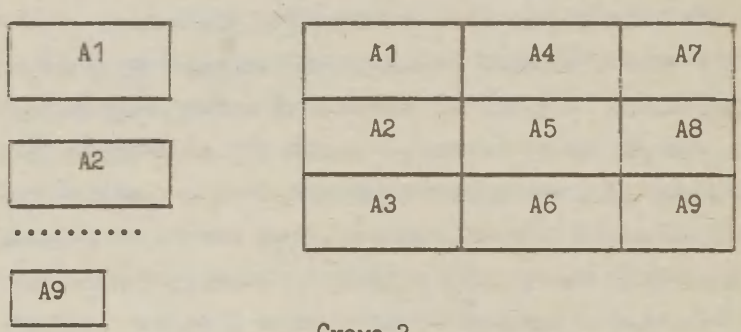

Схема 2

Заметим еще, что под каждым коэффициентом корреляции напечатаны соответственно объем (число объектов, у которых эти две признаки измерены) и вероятность значимости P (если $P<\alpha$, то коррелятивная зависимость значима).

Чтобы сделать информацию, содержащуюся в большой корреляционной матрицы, обозримой, рекомендуется пользоваться

разными графическими приемами. На первом шагу можно разными цветами отметить коэффициенты разной тесноты, а затем построить некоторые графы корреляции (путь наибольшей корреляции, $\wedge$-графы, зависящие от заданного семейства пороговых уровней и т.д.). Одной возможностью интерпретации корреляционной матрицы является построение дендрограммы.

В данном примере ядро первой группы образовалось из признаков, характеризующих успеваемость в учебе: прежде всего объединились в одну группу оценки первых сессий, наиболее тесно связанные между собой, затем оценки 4-7 сессий, а затем первая группа присоединилась ко второй. К группе вузовских оценок присоединились еще число провалов на экзаменах (корреляция с оценками, разумеется, отрицательная, но достаточно большая по абсолютной величине), а затем характеристики успеваемости в школе и на вступительных экзаменах.

Вторая группа образовалась из характеристик общественной активности. Эти признаки заметно слабее коррелированы друг с другом; отрицательной оказалась корреляция между числом общественных обязанностей и стажем супружества. Третья, еще слабее связанная группа, образовалась из признаков, связанных с образованием родителей и интересом к музыке и литературе. На уровне 0.35 объединяются первая и третья группа, присоединившие к себе еще удовлетворенность специальностью и признак, характеризующий желаемую специальность (1 - тот, на которую поступили и 0 - не тот).

На уровне 0.3 объединяются все группы, на этом уровне с признаками второй группы связывается и активность в спорте. Несвязанным с группами останется только признак "семейное состояние". Последние факты показывают, что по общим

характеристикам семейные студенты не отличаются от несемейных, а также активность в спорте не зависит от результативности в учебе и т.д.Полученная дендрограмма изображена на схеме 3.

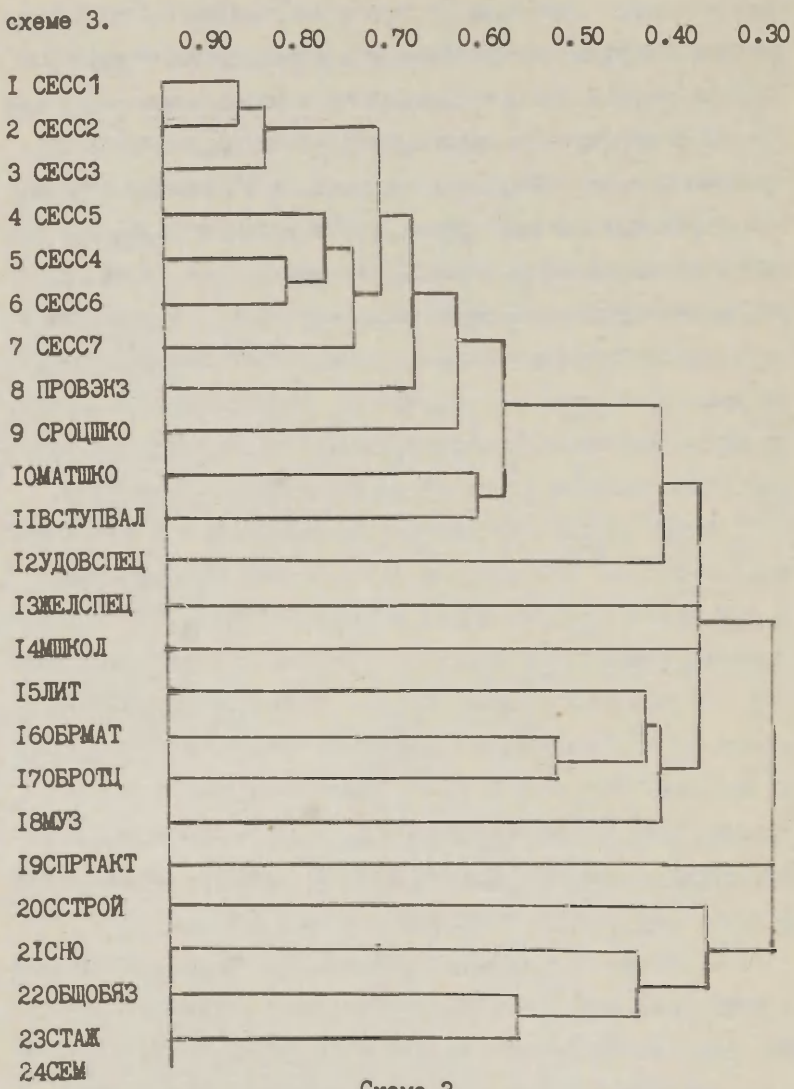

Схема 3

Следующим заказом было специальное исследование оценок, причем учитывался и тот факт, что оценки по своему содержанию - порядковые признаки и поэтому для них корректно пользоваться ранговыми коэффициентами связи (такие как коэффициенты корреляции Кендалла и Спирмена). Они заказываются при помощи процедуры NONPAR CORR.

Сравнивая полученные коэффициенты между собой можно утверждать, что все типы коэффициентов зависимости показывают существенную связь между всеми оценками (даже при уровне значимости 0.01). По абсолютным величинам коэффициенты Кендалла и Спирмена немного отличаются от линейных коэффициентов корреляции (на 0.05 - 0.1 и 0.1 - 0.2 соответственно). Структура зависимостей между отдельными оценками совпадает во всех случаях - сильнее связаны группы оценок на 1 - 3 сессиях и на 4 - 7 сессиях, их взаимная связь слабее.

Этот факт показывает, что в дальнейшем можно при анализе оценок пользоваться и коэффициентами линейной корреляции, не опасаясь неправильных результатов.

Целью следующего заказа было изучение того, в какой мере зависимость между результатами отдельных экзаменационных сессий зависит от школьных оценок (общее умение учиться) и от пола. Это можно выяснить при помощи частных корреляций (заказываются в проуедуре PARTIAL CORR).

Если элиминируется влияние признака СРОЦШКС, то по существу можно рассмотреть коэффициенты корреляции между вузовскими оценками студентов, имевших равные школьные оценки. Если выдвинуть гипотезу, что школьные оценки полностью описывают успеваемость в вузе, то получилась бы корреляционная матрица, где все коэффициенты близки к нулю (заметим, что

частные коэффициенты корреляции по своим свойствам совпадают с обычными коэффициентами корреляции). По распечаткам процедуры PARTIAL CORR выясняется, что, наоборот, общая структура корреляционной матрицы сохранилась, только по абсолютной величине коэффициенты уменьшились примерно на 0.15. Таким образом, выясняется, что для студентов, имеющих любой уровень школьной подготовки, результаты первых трех сессий похожи друг на друга, также как и результаты следующих четырех сессии.

При элиминировании влияния пола также сохраняется структура корреляционной матрицы, причем величина коэффициентов корреляции практически не изменяется. Структура сохраняется и после элиминирования влияния двух признаков ПОЛ и СРОЦШКО (коэффициенты частных корреляций второго порядка).

Проанализируем еще и корреляционную матрицу разностей между результатами сессий (признаки РАЗН1 - РАЗН6). Здесь все зависимости слабые, несущественные. Это показывает, что колебания в результатах сессий носят случайный харктер, нет существенного тренда ни в сторону заметного улучшения, ни в сторону ухудшения результатов. Единственное, что можно заметить, как правило, результаты весенних сессий немного хуже. чем результаты зимних сессий.

После анализа большой корреляционной матрицы можно более подробно исследовать некоторые пары или группы признаков, представляющие специальный интерес. В случаях, когда зависимость оказалась заметно ниже ожидаемой, можно проверять существование нелинейной зависимости.

I3.I0. Анализ зависимостей пар признаков

Очередной этап первичного анализа состоит из анализа совместного распределения и характеристик зависимости пар признаков. Образуется следующий заказ:

```
RUN NAME          ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ ПАР ПРИЗНАКОВ
GET FILE          МАТИК
RECODE            ЛЮБЗАН (5=I) (7=6)/
MISSING VALUES    ЛЮБЗАН (Ø,4)/
VALUE LABELS      ЛЮБЗАН (6) РАЗВЛЕЧ/
TASK NAME         ТАБЛИЦЫ КАЧЕСТВЕННЫХ ПРИЗНАКОВ
CROSSTABS         ЛЮБЗАН ВУ МРОЖД, МШКОЛ
STATISTICS        I, 2, 3
FASTABS           VARIABLE=ЛЮБЗАН(Ø,6), ПОЛ(I,2),РЕСП(I,2)/
                  TABLES=ЛЮБЗАН ВУ ПОЛ ВУ РЕСП/
STATISTICS        I, 2, 8
TASK NAME         КОРРЕЛЯЦИОННОЕ ПОЛЕ ЧИСЛОВЫХ ПРИЗНАКОВ
SCATTERGRAM       СРОЦШКО, СРЕДНЕЕ, WITH ОБРОТЦ, ОБРМАТ/
STATISTICS        ALL
FINISH
```

Прокомментируем заказ и проанализируем распечатку.

Для качественных (в особенности номинальных) признаков наиболее подходящей формой представления зависимости между признаками является таблица двумерных частот: причем теснота зависимости характеризуется при помощи некоторых характеристик (коэффициентов) зависимости. Единственное, что надо соблюдать - это, чтобы во всех клетках было (в среднем) достаточно много наблюдений.

Для того, чтобы выяснить зависимость любимых занятий от пола и места жительства (до вступления в университет),

применяем процедуры CROSSTABS (для анализа двумерной зависимости) и FASTABS (для анализа трехмерной зависимости - включается и республика, в которой родился студент). Предварительно признак ЛЮБЗАН перекодируется с целью уменьшения числа значений: объединяют значения "походы" и "спорт", также "кино" и "общение с друзьями". Так как значение "музыка" появилось только I раз, это вообще исключили из таблицы. В результате получили 4x4 таблицу, из которой выяснилось, что спорт любимое занятие студентов, приехавших из средних городов, рукоделие - для большинства бывших жителей малых городов, но и многих из тех, кто жил в большом городе. Чтением и разными развлечениями занимаются больше всего в большом городе, а в селе менее всего. Все эти зависимости сравнительно слабые, значение χ^2-статистики I5.44; при числе степеней свободы 9 =(4-I)x(4-I) соответствующая вероятность значимости - 0.08; поэтому невозможно считать доказанной зависимость занятий от происхождения студентов. Наверное такая зависимость существует, но для этого надо исследовать не 50, а по крайней мере несколько сотен студентов. Значение коэффициента Крамера есть 0.32, и, естественно, это несущественно.

При одновременном анализе трех признаков тем более невозможно доказать существенных зависимостей, а получаются лишь некоторые описательные характеристики о господствующих тенденциях. Например, выяснились "ряды популярности" любимых занятий для всех исследуемых групп:

девушки ЭССС - рукоделие, чтение, спорт, кино
девушки других республик - кино, чтение, рукоделие
юноши ЭССР - чтение, спорт, кино
юноши других республик - (чтение, кино, рукоделие)(равные)

Для более подробного исследования зависимостей между количественными признаками применима процедура SCATTERGRAM. Для более наглядного представления зависимостей к распечатке этой процедуры прилагается и корреляционное поле (диаграмма рассеяния). Заказывались и все статистики, характеризующие коррелятивную связь между признаками (коэффициент корреляции R, его квадрат R^2 - т.н. коэффициент детерминации, вероятность значимости P). Также были заказаны статистики, позволяющие выписать зависимость между признаками в виде линейной регрессии У = А + ВХ (коэффициенты А и В, среднее квадратичное отклонение). На диаграмме видно, что зависимость между образованием отца и результатами сессии действительно существует: у пяти студентов, имеющих высокие показатели успеваемости (СРЕДНЕЕ > 4,5), отцы имеют высшее образование, а у двух отличников отцы не имеют высшего образования. У половины студентов, отцы которых имеют образование 8 классов и ниже, средняя оценка ниже чем 3,7. Кроме того, из диаграммы видно, что нет оснований для предположения нелинейной зависимости.

Все же зависимость сравнительно слабая: R = 0.299, но при α = 0.05 существенная. Регрессионная модель, прогнозирующая среднюю оценку студента у по образованию отца х выглядит следующим образом:

$$y = 3.40 + 0.043x,$$

где х измеряется в числе законченных классов (курсов).

I3.II. Заказ факторного анализа

В течение первичного анализа признаки исследовались по группам, состоящим из I-2-3 признаков (заметим, что даже в большой корреляционной матрице все числовые характеристики

зависят только от двух признаков). После завершения этого этапа обработки характерным является исследование групп, состоящих из большого количества признаков.

Если целью данного этапа является описание некоторого множества признаков и их применение с описательной целью, то почти всегда целесообразно пользоваться факторным анализом. Образуем следующий заказ:

```
RUN NAME      ОПИСАНИЕ ПРИЗНАКОВ ПРИ ПОМОЩИ ФАКТОРОВ
TASK NAME     ФАКТОРНЫЙ АНАЛИЗ
FACTOR        VARIABLES=МШКОЛ, ОБРОТЦ, ОБРМАТ, СРОЦШКО ТО
              ССТРОЙ, ЛИТ, СПРТАКТ, МУЗ, СЕМ, ЧИСДЕТ/TYPE=
              PA1/
STATISTICS    4, 5, 6
FINISH
```

Обратим внимание на то, что в заказе факторного анализа необходимо исключить такие признаки, которые не измерены у многих объектов. Дело в том, что для получения корректных результатов факторного анализа необходимо ограничиться такими объектами, у которых измерены все признаки, входящие в модель факторного анализа. Признаки, содержащие много пробелов, выясняются на предыдущих этапах работы (например, PEARSON CORR). В настоящем заказе по этой причине исключен признак СТАЖ, а вместо него включили в список признак "число детей"(ЧИСДЕТ), измеренный практически у всех (хотя у большинства он равняется нулю).

На практике факторный анализ состоит из многих заказов, и первым выбирается сравнительно короткий вариант - быстрый метод главных компонент PA1, сравнительно небольшое количество дополнительной информации.

Проанализируем полученную факторную модель. На основании 24x24 корреляционной матрицы выделили 7 факторов (число собственных значений, которые больше единицы), степень описания - 73%. Так как последние три главных компонента оказались трудно интерпретируемыми, то в дальнейшем будем ограничиваться только 4 первыми факторами. Для получения такой модели, повторим заказ, добавляя к нему карту

NFACTORS = 4/

после карты TYPE = PA1. Теперь определяются только четыре первых фактора, и вращается комплект, состоящий из 4 факторов. Для исследования отдельных объектов при помощи образованных нами факторов закажем еще и индивидуальные факторные веса при помощи карты

FACSCORE/,

находящейся в конце заказа факторного анализа, и потребуем их распечатку при помощи карты

STATISTICS 4, 5, 6, 7.

Полученная факторная матрица является после поворота методом VARIMAX (делается по умолчанию) уже довольно хорошо интерпретируемой.

Рассмотрим матрицу факторных весов и ее интерпретацию. Для этого сделаем для каждого фактора таблицу, которая содержит признаки, коррелированные с ним сильно и умеренно. Обычно целесообразно иметь в таблице два столбца, соответствующих положительному и отрицательному направлению фактора, причем необходимо учитывать и направление шкал исследуемых признаков.

Матрица повороченных факторных весов

Признак	F1	F2	F3	F4	Общность
МШКОЛ	0,09	-0,38	-0,53	0,22	0,48
ОБРОТЦ	0,30	0,06	0,52	-0,55	0,66
ОБРМАТ	0,24	0,53	0,27	-0,29	0,49
СРОЦШКО	0,78	0,I9	0,05	-0,I3	0,60
МАТШКО	0,46	0,3I	-0,39	0,I3	0,47
ВСТУПБАЛ	0,70	-0,04	0,32	0,08	0,59
ЖЕЛСПЕЦ	0,02	-0,I6	-0,46	-0,II	0,25
СЕССI	0,85	0,07	0,23	0,II	0,80
СЕСС2	0,83	0,03	0,2I	0,2I	0,75
СЕСС3	0,85	0,06	0,I4	-0,00	0,74
СЕСС4	0,87	0,04	-0,02	-0,04	0,76
СЕСС5	0,80	-0,I6	0,04	-0,I2	0,68
СЕСС6	0,88	0,06	0,04	-0,08	0,79
СЕСС7	0,75	-0,04	-0,I7	-0,I3	0,6I
ПРОВЭКЗ	-0,74	0,I5	0,28	0,04	0,65
УДОВСПЕЦ	0,05	-0,02	0,66	0,07	0,44
СНО	0,23	0,59	-0,I4	-0,04	0,42
ОБЩОБЯЗ	0,I0	0,58	0,0I	-0,23	0,40
ССТРОЙ	-0,I2	0,64	0,34	-0,03	0,54
ЛИТ	0,29	-0,I3	0,52	-0,03	0,37
СПРТАКТ	-0,0I	0,29	0,0I	0,09	0,09
МУЗ	-0,I9	0,6I	-0,02	0,02	0,4I
СЕМ	-0,03	0,04	0,I7	0,73	0,64
ЧИСДЕТ	0,09	-0,I6	-0,04	0,83	0,73

Таблица интерпретации первого фактора следующая:

F1

+		-	
СРОЦШКО СЕССI - СЕСС7 ВСТУПБАЛ	как правило выше среднего	СРОЦШКО СЕССI -СЕСС7 ВСТУПБАЛ	как правило ниже среднего
ПРОВЭКЗ	заметно ниже среднего	ПРОВЭКЗ	выше среднего
МАТШКО ОБРОТЦ	часто выше среднего	МАТШКО ОБРОТЦ	часто ниже среднего

Ясно, что первый фактор - фактор успеваемости в вузе, причем интересно заметить, что он сильно зависит от успеваемости в средней школе и от результатов вступительных экзаменов, и умеренно от оценки математики в школе и образования отца.

F2

+		-	
ОБРМАТ СНО ОБЩОБЯЗ ССТРОЙ МУЗ	как правило значения выше среднего	ОБРМАТ СНО ОБЩОБЯЗ ССТРОЙ МУЗ	как правило значения ниже среднего
МШКОЛ	чаще большой город	МШКОЛ	чаще село или малый город
МАТШКО	часто значения выше среднего	МАТШКО	часто значения ниже среднего

Второй фактор является фактором общественной активности, причем более высокая активность связывается с более высоким образованием матери и городским образом жизни.

F3

+		−	
ОБРОТЦ УДОВСПЕЦ	как правило значения	ОБРОТЦ УДОВСПЕЦ	как правило значения
ЛИТ	выше среднего	ЛИТ	ниже среднего
МШКОЛ	как правило большой или средний город	МШКОЛ	как правило малый город или село
ВСТУПБАЛ ССТРОЙ	часто выше среднего	ВСТУПБАЛ ССТРОЙ	часто ниже среднего
ЖЕЛСПЕЦ	часто ниже среднего: больше поступивших учиться по желаемой специальности (код I)	ЖЕЛСПЕЦ	часто выше среднего: больше поступивших учиться не по желаемой специальности (код 2)
МАТШКО	часто ниже среднего	МАТШКО	часто выше среднего

Третий фактор имеет довольно сложную структуру - здесь связываются удовлетворенность выбранной специальностью, образование отца, в некотором смысле ориентация на умственные ценности. Положительное направление фактора связывается с городским местом жительства, а также и сознательным выбором профессии, хотя их оценки по математике в средней школе были ниже среднего (для рассматриваемого контингента).

Четвертый фактор характеризует семейное положение студентов; зависимость от образования отца непонятно, и может

быть вызвана случайностью.

Из столбца общностей видно, что кроме активности в спорте все остальные признаки, характеризующие исследуемый контингент студентов, описываются 4 факторами.

F4

+		-	
ЧИСДЕТ СЕМ	как правило, выше среднего, т.е. студенты брачные, имеют детей	ЧИСДЕТ СЕМ	как правило, ниже среднего, т.е. студенты одинокие, не имеют детей
ОБРОТЦ	чаще ниже среднего	ОБРОТЦ	чаще выше среднего

13.12. Заказ канонического анализа

После того, как мы проанализировали зависимость внутри группы описательных признаков, попытаемся выяснить зависимость функциональных признаков (успеваемость в вузе и общественная активность) от аргументов (характеристика семьи и дома, успеваемость в средней школе). Так как корреляционный анализ показал довольно слабые зависимости между этими группами признаков, то мы попытаемся их "усиливать" при помощи канонического анализа. Образуем заказ:

```
RUN NAME     КАНОНИЧЕСКИЙ АНАЛИЗ
CANCORR      VARIABLES =ОБРОТЦ TO СРОЦШКО, СЕСС1 TO ССТРОЙ/
             RELATE =ОБРОТЦ TO СРОЦШКО WITH СЕСС1 TO ССТРОЙ/
FINISH
```

Обратим внимание на то обстоятельство, что в заказе канонического анализа обязательно признаки обоих групп должны

находиться подряд, списки, следующие за ключевым словом RELATE, должны быть заданными при помощи ТО-соглашения.

Анализ распечатки начинается с проверки значимости канонических корреляций. Имеется следующая таблица:

Каноническая корреляция	χ^2-статистика	Число степеней свободы
0.78	84.63	48
0.77	49.27	33
0.47	I5.90	20
0.4I	6.7I	9

Выберем уровень значимости 0.0I. Сравнивая значения χ^2-статистик с соответствующими критическими значениями из таблиц, можно убедиться, что только первые две канонические корреляции существенно отличаются от нуля, и поэтому необходимо интерпретировать только две пары канонических переменных (компонент). Их коэффициенты заданы в следующей таблице:

Группа аргумент-признаков			Группа функциональных признаков		
	I КАН.КОМП.	II КАН.КОМП.		I КАН.КОМП.	II КАН.КОМП.
ОБРОТЦ	-0.02	0.04	СЕССI	0.48	-0.54
ОБРМАТ	0.27	0.II	СЕСС2	0.33	-0.0I
ПОЛ	-0.34	0.96	СЕСС3	-0.22	0.70
СРОЦШКО	0.86	0.20	СЕСС4	0.55	-0.78
			СЕСС5	-0.24	0.75
			СЕСС6	0.09	-0.9I
			СЕСС7	0.I3	I.I0
			ПРОВЭКЗ	0.28	-0.48
			УДОВСПЕЦ	-0.I3	-0.04
			СНО	0.06	-0.05
			ОБЩОБЯЗ	0.24	0.02
			ССТРОЙ	0.I8	-0.02

При интерпретации канонических переменных исходим из заданных коэффициентов и получаем первую пару канонических компонент:

z_1 = -0.02 ОБРОТЦ + 0.27 ОБРМАТ - 0.34 ПОЛ + 0.86 СРОЦШКО

w_1 = 0.48 СЕССІ + 0.33 СЕСС2 - 0.22 СЕСС3 + 0.55 СЕСС4 - - 0.24 СЕСС5 + 0.09 СЕСС6 + 0.I3 СЕСС7 + 0.28 ПРОВЭКЗ - 0.I3 УДОВСПЕЦ + 0.06 СНО + 0.24 ОБЩОБЯЗ + + 0.I8 ССТРОЙ.

Аналогично можно выписать и вторую пару канонических компонент z_2 и w_2.

Из коэффициентов видно, что z_1, зависящий от средней школьной оценки и образования матери, влияет сильно на результаты первой и четвертой сессии. Вторая каноническая компонента, которая в основном зависит от пола студента, распределяет сессии довольно интересно на "удачные для девушек" - это третья, пятая и седьмая и "удачные для юношей" - первая, четвертая и шестая (здесь учитывается правило кодирования признака ПОЛ: I - мужчина, 2 - женщина). Кроме того видно, что и провалов у юношей в среднем больше.

Заметим здесь, что хотя канонические компоненты нормированы ($Dz_i = Dw_i = 1$), коэффициенты канонических переменных зависят от дисперсий исходных признаков и не являются обязательно по абсолютной величине меньше единицы (как факторные веса).

I3.I3. Заказ регрессионного анализа

Этим заказом мы попытаемся решить одну из основных проблем, выдвинутых нами - именно выяснить, как хорошо мож-

но прогнозировать успеваемость студента в вузе по признакам, характеризующим его до вступления в вуз. В качестве модели, определяющий прогноз, используем линейную регрессию.

Прогнозируемым признаком выбираем среднюю успеваемость в вузе, т.е. признак СРЕДНЕЕ. Потенциальными аргументами являются МШКОЛ, ГОКОНЧ, ГВСТУП, ПОЛ, ОБРОТЦ, ОБРМАТ, СРОЦШКО, МАТШКО, ЖЕЛСПЕЦ. Для выбора наилучшей модели пользуемся пошаговой процедурой. Таким образом, составим следующий заказ:

```
RUN NAME      РЕГРЕССИОННЫЙ АНАЛИЗ
REGRESSION    VARIABLES = МШКОЛ ТО ОБРМАТ, СРОЦШКО ТО
              ЖЕЛСПЕЦ, СРЕДНЕЕ /
              REGRESSION  = СРЕДНЕЕ WITH  МШКОЛ ТО ОБРМАТ,
              СРОЦШКО ТО ЖЕЛСПЕЦ (I) RESID= Ø/
 STATISTICS   ALL
 FINISH
```

Число (I) в конце списка аргументов информирует систему о желании пользоваться пошаговой процедурой. Чтобы проверить, не будет ли нелинейная модель лучше линейной, мы закажем график остатков прогноза (RRSID=Ø). В конце учитывается и тот факт, что в ходе анализа и уточнения первичной модели, наверное, необходимо много промежуточных результатов, поэтому и заказываются все дополнительные статистики.

Для создания регрессионной модели используется полная подвыборка, т.е. только такие объекты, у которых измерены все признаки, указанные в списке "VARIABLES". Так как в случае неполных данных такая подвыборка, как правило, заметно меньше полной выборки, необходимо исследовать ее репрезентативность: не будут ли средние, стандартные отклонения и корреляции в этой подвыборке существенно отличаться от соответ-

ствующих характеристик для полной выборки.

Если различия маленькие, то достаточно визуальной проверки, при более существенных отклонений необходимо пользоваться соответствующими тестами, например, Т-тестом. При необходимости придется исключить из модели некоторые аргументы.

В настоящем примере в подвыборку входит 47 объектов, притом подсовокупность является представительной.

В ходе пошаговой регрессии на первом шагу в модель включается среднее школьных оценок и получается модель в виде линейного уравнения:

$$У = 0.887\ x_1 - 0.089\ , \qquad (I3.I)$$

где У - признак СРЕДНЕЕ, а x_1 - СРОЩШКО. Тесноту регрессионной зависимости характеризует множественный коэффициент корреляции, который в данном случае равняется обычному коэффициенту корреляции между у и x_1, R = 0.695. Его квадрат - коэффициент детерминации - равняется 0.483; таким образом, из средней успеваемости в вузе 48% описывается средней успеваемостью в средней школе.

Значимость регрессионной зависимости (учитывая объем выборки) характеризует F-статистика, значением которой в настоящем случае есть 42.0. При числах степеней свободы I (число коэффициентов регрессионной модели - I) и 45 (число объектов - число коэффициентов модели) и уровне значимости 0.05 критическое значение F-статистики, получаемое из таблиц, есть 4.06. Ввиду неравенства $42.0 > 4.06$ регрессионная модель (I3.I) является существенной.

На следующем шаге в уравнение включается аргумент x_2 - образование матери (ОБРМАТ). Модель выглядит следующим:

$$y = 0.826x_1 + 0.029x_2 - 0.123, \qquad (13.2)$$

коэффициент множественной корреляции приобретает значение 0.718 ($R^2 = 0.516$, значит степень описания - 52%). Значение F-статистики равняется 23.3 и при числах степеней свободы 2 и 44 модель (13.2) существенна.

Сравнивая модели (13.1) и (13.2) видно, что прибавлением нового аргумента изменяются коэффициенты b_1 и b_2 уравнения (13.2). Возникает вопрос - является ли модель (13.2) существенно лучше чем модель (13.1)? Ответ на этот вопрос можно получить, пользуясь значением F-статистики, характеризующей признак x_2 - это есть 2.09. Так как при степенях свободы 1 и 44 уровню значимости $\alpha = 0.05$ соответствует критическое значение F-статистики 4.06, то из неравенства $2.09 < 4.06$ вытекает, что модель (13.2) не существенно лучше модели (14.1).

Математически эквивалентным вышеизложенному является и следующее, может быть, более наглядное рассуждение. Для проверки гипотез

H_1: $b_2 \neq 0$ (т.е. модель (13.2) лучше, чем модель (13.1)

H_0: $b_2 = 0$ (модели 13.2) и (13.1) статистически эквивалентны)

рассматривается отношение b_2/s_2, где s_2 - стандартное отклонение коэффициента регрессии b_2. Это отношение имеет T-распределение с числом степеней свободы $n-k$ (k - число коэффициентов модели). В данном случае $b_2/s_2 = 1.446$ (заметим, что эта величина $\sqrt{F} = \sqrt{2.09}$), а так как при уровне значимости 0.05 и $f = 44$ критическое значение T-теста есть 2.02, то из неравенства $1.446 < 2.02$ вытекает несущественность улучшения модели (13.1) включением аргумента x_2.

Уже сейчас можно сказать, учитывая правило пошаговой процедуры, что включение следующих аргументов в модель не может существенно улучшить модель (I3.I). Все же дальнейшее исследование поведения регрессионного уравнения представляет некоторый интерес, в особенности в случае, когда мы имеем дело с т.н. пилотажным исследованием, которое в будущем повторяется на основе значительно большего объема выборки.

В результате прибавления каждого нового аргумента множественный коэффициент корреляции R увеличивается (но не намного), значение F-статистики, наоборот, уменьшается. После включения в модель последнего, восьмого аргумента имеется R = 0.76, степень описания 59% и значение F-статистики 6.7. Так как числа степеней свободы при данной модели 8 и 38, то критическое значение F-статистики при $\alpha = 0.05$ есть 2.I9, и ввиду неравенства 6.7> 2.I9 модель существенна.

Можно сказать, что пользоваться моделью, содержащей все 8 аргументов, было бы правильно лишь тогда, когда объем выборки был весьма большой (например, несколько тысяч), так как удельный вес большинства аргументов в модели маленькое.

В конце распечатки для каждого объекта (студента) заданы истинное значение y_j, его прогноз $\hat{y}_j$, определенный при помощи регрессионной модели и их разность (остаток прогноза) $\hat{\varepsilon}_j$.

О тех студентах, для которых $\hat{y}_j > y_j$, можно сказать, что их успеваемость в университете слабее, чем предполагается, наоборот, если $\hat{y}_j < y_j$, то результаты в вузе выше ожидаемого; изучение причин того или другого явления относится к предмету педагогики высшей школы.

Для проверки адекватности модели необходимо изучать распределение остатков прогноза. Если оно является приблизи-

тельно нормальными и на втором графике нет явных криволинейных тенденций, то модель можно считать адекватной.

Зависимость объектов измерения от порядкового номера (эффект временного ряда) можно проверять при помощи статистики Дурбина-Уатсона. На распечатке выдается значение этой статистики; если оно находится между заданными критическими значениями d_L и d_U, то принимается гипотеза о зависимости наблюдений от их порядка. В данном примере $d < d_L$ (при уровне зависимости $\alpha = 0.05$), значит, явной зависимости от порядкового номера в массиве данных нет. Также изучение графиков остатков прогноза ведет к убеждению, что нет основания предполагать нелинейность зависимости или неадекватность модели. Учитывая этот факт, сделаем окончательный заказ регрессионного анализа, при котором в модель включаются только нужные нам (при уровне 0.I) аргументы, а прогнозы $\hat{y}_i$ и остатки $\hat{\varepsilon}_i$ вычисляются уже по выбранной нами оптимальной модели:

```
RUN NAME      ОКОНЧАТЕЛЬНАЯ МОДЕЛЬ РЕГРЕССИИ
REGRESSION    VARIABLES = ОБРМАТ, СРОЩКО, СРЕДНЕЕ/
              REGRESSION = СРЕДНЕЕ WITH ОБРМАТ
              СРОЩКО (2) RESID = Ø/
STATISTICS    4, 5, 6
FINISH
```

Так как выбор аргументов уже сделан, то мы не будем пользоваться пошаговой процедурой (это показывает номер (2)), сокращается и набор заказываемых статистик.

I3.I4. Заказ дискриминантного анализа

Поставим цель: найти различия между студентами трех потоков (специальностей) - математики, эстонский и руский потоки прикладной математики, выяснить признаки, наилучше ди-

скриминирующие рассматриваемые группы и определить правило, по которому возможно объекты (студенты) распределять в указанные три группы. Необходимо еще проверять, есть ли такая дискриминация вообще возможна, т.е. являются ли эти группы вообще различными друг от друга по признакам, измеренным нами.

Для решения поставленной задачи пользуемся процедурой DISCRIMINANT, в качестве группирующего признака выбираем СПЕЦ, а за дискриминирующие: СЕССІ до СЕСС7, СНО, ССТРОЙ и УДОВСПЕЦ. Таким образом, оформляем заказ:

```
RUN NAME          ДИСКРИМИНАНТНЫЙ АНАЛИЗ
DISCRIMINANT      СПЕЦ WITH СЕССІ TO СЕСС7, СНО, ССТРОЙ,
                  УДОВСПЕЦ
STATISTICS        1, 2, 4
FINISH
```

В качестве дополнительных статистик процедуры DISCHIMINANT прежде всего выступают характеристики данной выборки, позволяющие проверять ее репрезентативность (также как и в случае регрессивного анализа), а также и некоторые статистики для более подробного описания групп.

Распечатка дискриминантного анализа начинается описательными статистиками, описывающими как изучаемую подвыборку, так и все группы. Средние всех признаков для каждой группы описывают центр или типичный элемент данной группы, при помощи которого можно интерпретировать имеющиеся группы на содержательном уровне.

Выбор аргументов в дискриминант-функций происходит шаг за шагом, причем на каждом шаге в функцию включается один аргумент - именно тот, который наилучшим образом дискриминирует имеющиеся группы.

На нулевом шаге для всех признаков вычисляются только значения статистики F-включения и выделяются соответствующие значения степеней свободы.

В данном примере мы имеем степени свободы 2 и 46, критическое значение F-статистики при $\alpha = 0.05$ есть 3.20 и мы сразу видим, что для всех признаков значение статистики включения меньше критического. Максимальное значение статистики F имеет признак CECCI (F = 2.82).

На первом шаге в модель включается признак CECCI. Критические значения для разных статистик F при уровне значимости 0.05 мы получим из таблиц (в скобках степени свободы):

F-включения F(2,45) = 3.21;

F-удаления F(2,46) = 3.20;

F-приближенный F(2,46) = 3.20;

F между отдельными группами F(1,46) = 4.05.

Сравним с этими значениями результаты нашей распечатки. Так как приближенный F имеет значение 2.8I6 (2.8I6 < 3.20), дискриминант-функция, содержащая признак x_1 = CECCI, не различает группы существенным образом. Это утверждает и тот факт, что аргумент x_1 в дискриминантной функции несущественный (F-удаления признака x_1 имеет значение 2.8I6 < 3.20). Не улучшает дело и включение в модель следующего аргумента x_2 (CCTPOЙ), так как F-включения для него имеет значение 3.I3 (3.I3 < 3.2I). Единственный положительный результат, наблюдаемый нами, это выяснение факта, что дискриминант-функция, полученная на первом шаге, существенно дискриминирует две группы: математиков (группа I) и прикладников русского потока (группа 3). Для них F имеет значение 4.27 (4.27 > 4.05).

На следующем, втором шаге найдем из таблиц критические

значения:

F-включения: (2,44) = 3.21

F-удаления: (2,45) = 3.21

F-приближенный: (4,90) = 2.47

F между группами: (2,45) = 3.21.

На этом шаге мы из распечатки имеем значение F-приближенного, равное 2.94, значит, модель можно считать значимой. Значимой оказалось теперь различие между группами I и 2 (прикладники эстонского потока), значение статистики F равняется 4.79; зато различия между остальными группами несущественные.

Дальнейшее изучение протокола пошаговой дискриминации показывает, что включение в модель дополнительных аргументов не улучшает дискриминирующую способность модели, а, наоборот, начиная с пятого шага становится несущественной. Ни одна из полученных моделей не дает возможности для эффективной дискриминации всех групп. В конце распечатки дискриминантного анализа выдаются коэффициенты w_i линейных функций - информантов - для всех групп (с индексом i, $i = 1,\ldots,k$). Если мы имеем объект x_0, о котором мы не знаем, какой из этих групп он принадлежит, то придется для него вычислить значение $w_i x_0 = \sum_{h=1}^{p} w_{ih} x_{0h}$ и считать его принадлежащим той группе, для которой информант имеет максимальное значение:

$$w_{i_*} x_0 > w_i x_0, i = 1,\ldots,k, i \neq i_* .$$

Для различения двух групп i и j применимы разности информантов, дискриминирующие функций

$$\gamma_{ij}(x_0) = \sum_{h=1}^{p} \gamma_h^{ij} x_{0h} .$$

Заметим, что все эти функции следует интерпретировать

на том этапе работы, когда в модель выбраны самые эффективные аргументы и среди них нет лишних. На этом этапе работы целесообразно изучать и дополнительные данные о всех объектах, выдаваемые при помощи статистики 2: их расстояния Махаланобиса до всех центров групп, апостериорные вероятности и принадлежности в группы (по значениям исходного группирующего признака и по полученному группирующему правилу).

По этой таблице можно и группировать негруппируемые объекты, которые в начале исследования могут образовать формальную группу, а в ходе дискриминации их относят в некоторую из имеющихся групп.

В случае, когда выбранные признаки не позволяют дискриминировать все группы, целесообразно ограничивать задачу в том смысле, что попытаться найти правило для дискриминации более крупных групп; в нашем примере можно объединить в одну группу прикладников русского и эстонского потока.

Учитывая это замечание, сделаем следующий заказ:

```
RUN NAME        ВТОРИЧНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ
RECODE          СПЕЦ (3=2)
VALUE LABELS    СПЕЦ (I) МАТИК (2) ПРИКЛАД
DISCRIMINANT    СПЕЦ WITH СЕССI TO СЕСС4,ССТРОЙ,УДОВСПЕЦ
STATISTICS      I, 2
```

В ходе исследования распечатки этого заказа выяснилось, что для дискриминации двух групп достаточно было двух признаков - x_1 = СЕССI и x_2 = ССТРОЙ, остальные признаки существенную дополнительную информацию не прилагают. Учитывая этот факт, заказ повторяется еще раз, причем карта DISCRIMINANT приобретает форму:

```
DISCRIMINANT    СПЕЦ WITH СЕССI, ССТРОЙ.
```

I3.I5. Заказ кластерного анализа

Поставим себе цель группировки исследуемого контингента, беря за основу успеваемость и общественную активность студентов. Не имея надежную априорную информацию об ожидаемой групповой структуре исследуемого контингента, применим методику таксономии (кластер-анализа) и представим следующий заказ:

```
GET FILE        МАТИК
MULTIVARIANT    ANALYSIS=TAXONOMY, CLUSTER=7, OBSERV=51,
                THRESHOLD=3, METHOD=2, VARLIST=ПОЛ,
                СЕСС1 TO СЕСС7, СНО, ОБЩОБЯЗ, СРОЦЕНКО
FINISH
```

Напомним, что при процедуре MULTIVARIANT наклонная черта в качестве разделителя не допускается, вместо этого применяется запятая.

Параметры - максимальное число кластеров 7 и порог 3 выбраны интуитивно без серьезного обоснования; метод 2 - это метод лидера, который выбирается потому, что массив образует системный файл.

В результате обработки получаем 3 кластера, характеризующие параметры которых изложены в следующей таблице:

№ кластера	I	2	3
Число объектов	I0	38	2
Лидер (№ объекта)	I	3	22
Средние признаков			
ПОЛ	I.89	I.79	I.50
СЕССI	4.58	3.4I	4.35
СЕСС2	4.62	3.59	4.25

СЕСС3	4.62	3.67	4.45
СЕСС4	4.60	3.64	4.25
СЕСС5	4.62	3.83	4.00
СЕСС6	4.56	3.7I	4.30
СЕСС7	4.65	3.8I	4.30
СНО	0	0	I.50
ОБЩОБЯЗ	0.70	0.92	3.00
СРОШКО	4.87	4.84	4.90

Как видно, первый кластер состоит из отличников по учебе, которые не очень активны в общественной жизни. Второй кластер - средние студенты, третий - активисты СНО.

Так как второй кластер требует дальнейшего исследования, оформим новый заказ, прибавляя один признак - участие в студенческих стройотрядах - и уменьшим порог до 2.5.

В результате получим 7 кластеров (см. схему 4) - первый и второй кластер разделились на части в основном по общественной активности и участии в студенческих строительных отрядах.

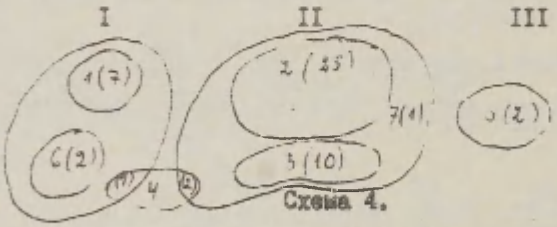

Схема 4.

Двухэтапная кластеризация. Кластеры, образованные на первом этапе, обозначены римскими цифрами, на втором этапе - арабскими цифрами. В скобках указаны численности объектов, входящих в кластеры второго этапа.

ОПРОСНЫЙ ЛИСТ

I. Имя

2. Специальность

3. Год вступления в университет

4. Какую среднюю школу вы закончили?

5. В каком году?

6. Место рождения

7. Год рождения

8. Пол

9. Профессия отца

I0. Образование отца

II. Профессия матери

I2. Образование матери

I3. Живут ли родители в настоящее время вместе?

I. Да
2. Нет, отец умер
3. Нет, отец ушел
4. Нет, мать умерла
5. Нет, мать ушла
6. Обе умерли

I4. Число сестер

I5. Число братьев

I6. Республика, в которой Вы родились

I7. Домашний язык в семье Ваших родителей

I8. Средняя оценка аттестата зрелости

I9. Оценка математики в средней школе

20. Вступительный бал

2I. На какую специальность Вы хотели поступить?

Средняя оценка по сессиям:

22. I сессия 23. II сессия 24. III сессия

25. IУ сессия 26. У сессия 27. УI сессия

28. УII сессия

29. На скольких экзаменах Вы провалились в течение учебы в университете?

30. Удовлетворенность со своей специальностью:

5 - очень доволен
4 - доволен
3 - более-менее
2 - не доволен
I - совсем не доволен

3I. Сколько раз Вы выступили на конференции СНО?

32. Сколько у Вас общественных обязанностей?

33. Сколько раз Вы участвовали в студенческом стройотряде?

34. Любимое занятие в свободное время

35. Заинтересованы ли Вы литературой? (Ответы см. вопрос 30)

36. Занимаете ли Вы активно спортом?

37. Ваш любимый вид спорта?

38. Заинтересованы ли Вы музыкой?

39. Ваша специальность по музыке

40. Вес 4I. Рост 42. Номер обуви

43. Семейное положение

44. Стаж супружества

45. Возраст супруга 46. Профессия супруга

47. Домашний язык в Вашей семье

48. Число детей

49. Возраст старшего ребенка

ЛИТЕРАТУРА

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ, Москва, 1973.
2. Сб. Математические методы в современной буржуазной социологии, Москва, 1966.
3. Харман Г. Современный факторный анализ, Москва, 1972.
4. Harris, R.A. A primier of Multivariate Analysis. Academic Press, New-York e.a., 1975, XI + 332 pp.
5. Kaiser, H.F. Image Analysis In C.W. Harris, Ed. Problems in Measuring Change, Madison, Wisconsin, Univ. of Wisconsin Press, 1963.
6. Kraiser, H.F., Gaffery, J. A l p h a Factor Analysis. Psychometrika, 1965, 30, 1 - 14.
7. Rao, C.R. Estimations and Tests of Significance in Factor Analysis. Psychometrika, 1955, 20, 93 - 111.

СОДЕРЖАНИЕ

I. ОБЩАЯ ХАРАКТЕРИСТИКА ПАКЕТА САИСИ 4

II. ПОДГОТОВКА К АНАЛИЗУ ДАННЫХ 6

2.I. Подготовка данных 6

2.2. Создание плана обработки данных 7

2.3. Представление данных в пакете САИСИ II

III. УПРАВЛЯЮЩИЙ ЯЗЫК ПАКЕТА САИСИ I5

3.I. Общая характеристика управляющего языка I5

3.2. Общие правила оформления управляющих карт I5

3.2.I. Имя I6

3.2.2. Значение I6

3.2.3. Ключевое слово I7

3.2.4. Метка I7

3.2.5. Разделитель I7

3.2.6. Арифметическое выражение I8

3.2.7. Логическое выражение I9

3.3. Классификация управляющих карт 20

3.4. Обозначения, используемые для управляющих карт 20

3.4.I. Необязательный элемент 20

3.4.2. Выбираемый элемент 2I

3.4.3. Список признаков 2I

IУ. ОПИСАНИЯ УПРАВЛЯЮЩИХ КАРТ 22

4.I. Карты описания данных 22

4.I.I. Карта FILE NAME 22

4.I.2. Карта VARIABLE LIST 22

4.I.3. Карта SUBFILE LIST 23

4.I.4. Карта INPUT MEDIUM 24

4.I.5. Карта # OF CASES 24
4.I.6. Карта INPUT FORMAT 24
4.I.7. Карта PRINT FORMATS 25
4.I.8. Карта VALUE LABELS 26
4.I.9. Карта VAR LABELS 26
4.I.I0. Карта MISSING VALUES 26
4.2. Карты преобразования данных 27
4.2.I. Карта RECODE 27
4.2.2. Карта COMPUTE 28
4.2.3. Карта IF 29
4.2.4. Карта COUET 29
4.2.5. Карта ASSIGN MISSING 30
4.2.6. Карты DO и DOEED 3I
4.2.7. Карта SELECT IF 32
4.2.8. Карта SAMPLE 32
4.2.9. Карта WEIGHT 32
4.3. Карты описания задачи и прогона 33
4.3.I. Карта RUN NAME 33
4.3.2. Карта TASK NAME 34
4.3.3. Карта COMMENT 34
4.3.4. Карта FINISH 34
4.3.5. Карта NUMBERED 34
4.3,6. Карты обращения к статистическим процедурам 35
4.3.7. Карта PROCESS SBFILES 35
4.3.8. Временные версии карт преобразования данных 36
4.3.9. Карта READ INPUT DATA 37
4.3.I0. Карты SAVE FILE и GET FILE 37
4.4. Карты редактирования данных 37
4.4.I. Карта DELETE VARS 38

4.4.2. Карта KEEP VARS 38
4.4.3. Карта ADD VARIABLES 39
4.4.4. Карта REORDER VARS 39
4.4.5. Карта NEW SUBFILE 39
4.4.6. Карта DELETE SUBFILES 40
4.4.7. Карта ADD SUBFILES 40
4.4.8. Карта SORT CASES 4I
4.5. Карты образования контрольной печати 4I
4.5.I. Карта DUMP 4I
4.5.2. Карта LIST CASE 43
У. ПРЕДСТАВЛЕНИЕ ЗАКАЗА И ПОРЯДОК УПРАВЛЯЮЩИХ КАРТ 44
5.I. Прогон, в ходе которого образуется системный файл 44
5.I.I. Начало прогона 44
5.I.2. Описание данных 44
5.I.3. Обращение к статистической процедуре 47
5.I.4. Карта OPTIONS 47
5.I.5. Карта STATISTICS 50
5.I.6. Введение начальных данных 50
5.I.7. Завершение прогона 5I
5.I.8. Пример 5I
5.I.9. Порядок управляющих карт 53
5.2. Прогон с преобразованием данных 55
5.2.I. Преобразование признаков 55
5.2.2. Размещение карт преобразования данных 57
5.2.3. Пример 58
5.2.4. Порядок управляющих карт 60
5.3. Прогон редактирования 6I
5.3.I. Возможности для редактирования 6I
5.3.2. Размещение карт редактирования 62

5.3.3. Пример 62
5.3.4. Порядок управляющих карт 64
5.4. Прогон, включающий контрольную печать 65
5.4.I. Возможности контрольной печати 65
5.4.2. Пример 65
УI. УПРАВЛЯЮЩИЕ КАРТЫ ОПЕРАЦИОННОЙ СИСТЕМЫ 67
6.I. Представление заказа САИСИ 67
6.2. Операторы определения места для вводной и выводной информации 68
6.2.I. Оператор DD с именем FTØ1FØØ1 68
6.2.2. Оператор DD с именем FTØ2FØØ1 69
6.2.3. Оператор DD с именем FTØ5FØØ1 69
6.2.4. Оператор DD с именем FTØ6FØØ1 70
6.2.5. Оператор DD с именем FTØ9FØØ1 70
6.2.6. Оператор DD с именем FTØ8FØØ1 70
6.2.7. Ввод системного файла 7I
6.2.8. Вывод системного файла 7I
6.3. Операторы сортирования 72
6.3.I. Операторы DD с именами SORTLIB и SYSOUT 72
6.3.2. Операторы DD с именами SORTWKØ1, SORTWKØ2, SORTWKØ3 72
6.4. Примеры 73
6.4.I. Заказ, в ходе которого образуется системный файл 73
6.4.2. Заказ, в который используют системный файл 73
УII. ПЕРВИЧНЫЙ АНАЛИЗ ПРИЗНАКОВ 75
7.I. Процедура CONDESCRIPTIVE 75
7.I.I. Цель и методические указания 75
7.I.2. Процедурная карта 76

7.I.3. Опции 76
7.I.4. Статистики 77
7.I.5. Пример 78
7.2. Процедура CODEBOOK 78
7.2.I. Цель и методические указания 78
7.2.2. Процедурная карта 79
7.2.3. Опции 80
7.2.4. Статистики 80
7.2.5. Пример 8I
7.3. Процедура FASTMARG 82
7.3.I. Цель и методические указания 82
7.3.2. Процедурная карта 82
7.3.3. Опции 83
7.3.4. Статистики 83
7.3.5. Пример 83
7.4. Процедура MARGINALS 84
7.4.I. Цель и методические указания 84
7.4.2. Процедурная карта 84
7.4.3. Опции 85
7.4.4. Статистики 85
7.4.5. Пример 85
VIII. ОПИСАНИЕ И СРАВНЕНИЕ ПОДВЫБОРОК 86
8.I. Процедура AGGREGATE 87
8.I.I. Цель и методические указания 87
8.I.2. Процедурная карта 87
8.I.3. Опции 90
8.I.4. Статистики 90
8.I.5. Пример 9I
8.2. Процедура BREAKDOWN 9I

8.2.I. Цель и методические указания 9I
8.2.2. Процедурная карта 94
8.2.3. Опции 95
8.2.4. Статистики 95
8.2.5. Пример 96
8.3. Процедура FASTBREAK 97
8.3.I. Цель и методические указания 97
8.3.2. Процедурная карта 97
8.3.3. Опции 98
8.3.4. Статистики 98
8.3.5. Пример 98
8.4. Процедура T-TEST 99
8.4.I. Цель и методические указания 99
8.4.2. Процедурная карта I0I
8.4.3. Опции I0I
8.4.4. Статистики I02
8.4.5. Пример I02
8.5. Процедура ONEWAY I03
8.5.I. Цель и методические указания I03
8.5.2. Процедурная карта III
8.5.3. Опции II3
8.5.4. Статистики II4
8.5.5. Пример II5
IX. ХАРАКТЕРИСТИКА ЗАВИСИМОСТИ МЕЖДУ ДВУМЯ ПРИЗНАКАМИ II6
9.I. Процедура CROSSTABS II9
9.I.I. Цель и методические указания II9
9.I.2. Процедурная карта I24
9.I.3. Опции I24
9.I.4. Статистики I25

9.I.5. Пример I25
9.2. Процедура PASTABS I26
9.2.I. Цель и методические указания I26
9.2.2. Процедурная карта I29
9.2.3. Опции I29
9.2.4. Статистики I30
9.2.5. Пример I3I
9.3. Процедура SCATTERGRAMM I3I
9.3.I. Цель и методические указания I3I
9.3.2. Процедурная карта I33
9.3.3. Опции I34
9.3.4. Статистики I35
9.3.5. Пример I35
X. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ I36
I0.I. Процедура PEARSON CORR I36
I0.I.I. Цель и методические указания I36
I0.I.2. Процедурная карта I4I
I0.I.3. Опции I42
I0.I.4. Статистики I43
I0.I.5. Пример I43
I0.2. Процедура NONPAR CORR I44
I0.2.I. Цель и методические указания I44
I0.2.2. Процедурная карта I45
I0.2.3. Опции I45
I0.2.4. Статистики I46
I0.2.5. Пример I46
I0.3. Процедура PARTIAL CORR I47
I0.3.I. Цель и методические указания I47
I0.3.2. Процедурная карта I49

10.3.3. Опции 150
10.3.4. Статистики 153
10.3.5. Пример 154
XI. ОПИСАТЕЛЬНЫЕ МЕТОДЫ МНОГОМЕРНОГО АНАЛИЗА 155
11.1. Процедура FACTOR 155
11.1.1. Цель и методические указания 155
11.1.2. Процедурная карта 164
11.1.3. Опции 167
11.1.4. Статистики 169
11.1.5. Пример 170
11.2. Процедура CANCORR 171
11.2.1. Цель и методические указания 171
11.2.2. Процедурная карта 176
11.2.3. Опции 177
11.2.4. Статистики 178
11.2.5. Пример 179
11.3. Процедура MULTIVARIANT 179
11.3.1. Цель и методические указания 179
11.3.2. Процедурная карта 185
11.3.3. Опции 188
11.3.4. Статистики 188
11.4. Процедура GUTTMAN SCALE 189
11.4.1. Цель и методические указания 189
11.4.2. Процедурная карта 191
11.4.3. Опции 192
11.4.4. Статистики 192
11.4.5. Пример 193
XII. ПРОГНОЗИРУЮЩИЕ МЕТОДЫ МНОГОМЕРНОГО АНАЛИЗА 194
12.1. Процедура REGRESSION 194

12.1.1. Цель и методические указания 194
12.1.2. Процедурная карта 203
12.1.3. Опции 206
12.1.4. Статистики 208
12.1.5. Пример 209
12.2. Процедура DISCRIMINANT 209
12.2.1. Цель и методические указания 209
12.2.2. Процедурная карта 217
12.2.3. Опции 218
12.2.4. Статистики 219
12.2.5. Пример 219
12.3. Процедура TIME SERIES 220
12.3.1. Цель и методические указания 220
12.3.2. Процедурная карта 223
12.3.3. Опции 224
12.3.4. Статистики 224
12.3.5. Пример 225
XIII. ПРИМЕР СОЗДАНИЯ ПЛАНА ОБРАБОТКИ ДАННЫХ И РЕАЛИЗАЦИЯ ПЛАНА В ФОРМЕ ЗАКАЗОВ 226
13.1. Создание массива данных 227
13.2. Определение формата признаков 228
13.3. Создание файла данных 229
13.4. Анализ результатов первого этапа обработки 230
13.5. Редактирование системного файла 231
13.6. План обработки 233
13.7. Образование новых признаков. Первичная обработка 234
13.8. Проверка влияния меняющих и группирующих признаков 236
13.9. Анализ статистических зависимостей по группам признаков 240
13.10. Анализ зависимостей пар признаков 247
13.11. Заказ факторного анализа 249
13.12. Заказ канонического анализа 253
13.13. Заказ регрессионного анализа 257
13.14. Заказ дискриминантного анализа 262
13.15. Заказ кластерного анализа 267
ПРИЛОЖЕНИЕ. ОПРОСНЫЙ ЛИСТ 269
ЛИТЕРАТУРА 271

40 коп.