ТАРТУСКИЙ
ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ТРУДЫ

ВЫЧИСЛИТЕЛЬНОГО ЦЕНТРА

51

ТАРТУ
1984

ТАРТУСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ НА ЕС ЭВМ

ТРУДЫ ВЫЧИСЛИТЕЛЬНОГО
ЦЕНТРА

ВЫПУСК 51

ТАРТУ 1984

Утверждено на заседании совета математического факультета ТГУ 19 октября 1984 года

РЕАЛИЗАЦИЯ ЯЗЫКА ОПИСАНИЯ ФАЙЛОВ

Ю. Каазик, К. Ээремаа

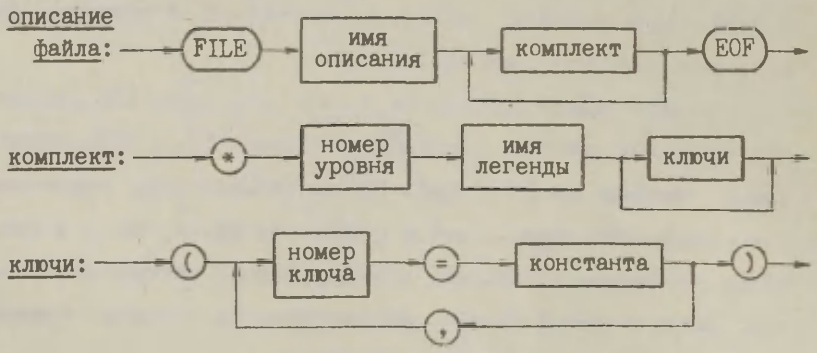
В системе РАМА каждый файл может содержать записи различной структуры, т.е. состоять из данных, описанных разными легендами (см. [1]). При этом, для облегчения обработки таких файлов, пользователь системы может еще требовать определенную последовательность записей в файле (не физически, а в смысле поиска записей по возрастанию их ключей доступа). Непосредственная реализация такой организации файлов средствами ОС не представляется возможным. Поэтому с каждым файлом следует связывать набор правил для преобразования ключей легенд в ключи доступа соответствующих комплектов записей, а также наоборот. В системе РАМА такие правила представляются в виде описания файла, задающего логическую структуру файлов, созданных по этому описанию.

Описания файлов пишутся на специальном языке FDL, основные концепции которого приведены в статье [2]. В ходе реализации системы эти концепции, однако, подверглись некоторым изменениям, относящимся как к синтаксису языка, так и к правилам построения по данному описанию ключей доступа комплектов. Далее в данной статье рассматриваются главным образом именно введенные изменения и уточнения.

Основное изменение самого языка FDL обусловлено тем, что доступ к записям любого конкретного файла осуществляется в системе РАМА теперь единым образом при помощи соответствующего каталога этого файла (см. [3]). Поэтому отпадает необходимость указать в описаниях файлов имя доступа - все файлы имеют доступ одного и того же типа.

Второе изменение синтаксиса языка связано с изменением принципа организации защиты файлов. Так как по одному описанию разрешается создать сколько угодно различных файлов, то не целесообразно заранее установить, что все эти файлы должны иметь (или не иметь) защиту от других пользователей. Естественнее фиксировать способ защиты каждого файла не в описании, а лишь одновременно с заданием конкретного пароля во время создания этого файла. Поэтому определение способа защиты файла полностью переносится из языка описания файлов в язык управления заданиями (см. [4]).

В результате указанных двух изменений несложный синтаксис языка описания файлов еще упрощается, принимая следующий весьма компактный вид:



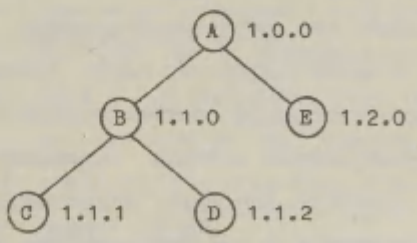
Оттранслированное описание файла будем называть макетом файла. Имя описания при этом по умолчанию принимается в качестве имени макета, хотя средствами языка управления заданиями макету можно присвоить любое другое имя (напр., когда обрабатываемая совокупность файлов, т.н. фонд [4] уже содержит макет с именем транслируемого описания).

Упомянутые в описании файла имена легенд должны быть там уникальными, т.е. в описании нельзя многократно использовать одно и то же имя легенды (хотя сами легенды могут по своей структуре совпадать). Кроме того, все эти легенды должны в оттранслированном виде находиться в том же фонде, в который по заданию на транслирование пишется получаемый макет.

Комплектом называется набор записей файла, структура которых описывается одной легендой. Номера уровней определяют иерархию комплектов в файле. В качестве номеров уровней следует использовать последовательные натуральные числа (начинающиеся с единицы), причем равные уровни отмечаются, конечно, одинаковыми номерами.

Иерархия комплектов естественным образом генерирует для каждого комплекта (или соответствующей легенды) координатный вектор, размерность которого равна максимальному из номеров уровня. Строение таких векторов видно по следующему примеру:

```
FILE ФАЙЛ1
  * 1 A
  * 2 B
  * 3 C
  * 3 D
  * 2 E
EOF
```



где рядом с описанием файла приведено соответствующее дерево иерархии, вершины которого снабжены координатными векторами (аналогично образуются и метки вершин записи в статье [5]). Следует отметить, что если в описании файла номер уровня 1 встречается более одного раза, то иерархия комплектов изображается в виде леса: первая компонента координатного вектора является ведь порядковым номером дерева в таком лесу.

Опираясь на иерархию комплектов строятся их ключи доступа, по которым происходит фактический доступ к записям файла. Учитывая требования операционной системы ОС ЕС все ключи доступа должны в пределах одного файла иметь одинаковую длину, но уникальное для каждой конкретной записи содержание.

Ключ доступа можно рассматривать состоящимся из полей, отведенных для конкретных значений ключей легенд и констант. Для каждого комплекта имеется свое правило разложения ключа доступа на поля и их заполнения.

Описанные в [2] правила образования ключей доступа комплектов подверглись в ходе реализации некоторым изменениям, хотя сохранились общие принципы. Так, например, остается в силе требование, по которому число ключей легенды комплекта-сына должно быть не меньше числа ключей легенды комплекта-отца. Если у сына ключей легенды больше, то остаток будем называть его собственными ключами.

В целях сокращения общей длины ключей доступа их костяк теперь составляют соответствующие координатные векторы, между компонентами которых размещаются поля для ключей легенды или констант. Методика образования ключа доступа комплекта основывается на следующих простых правилах.

Если уже имеется правило образования ключа доступа для комплекта-отца, то это правило переносится в состав соответствующего правила комплекта-сына, причем длины сопоставляемых ключей отца и несобственных ключей сына выравниваются по максимальным из них (длины собственных ключей братьев такому выравниванию не подлежат).

Последняя ненулевая компонента координатного вектора непосредственно предшествует всем собственным ключам рассматриваемого комплекта.

Ключи доступа всех комплектов при необходимости дополняются справа нулями до максимальной длины, а в конец добавляется еще один байт, содержащий порядковый номер комплекта в описании файла.

В приведенных правилах, а также в определении собственных ключей необходимо учитывать, что кроме обычных ключей легенды (т.е. описанных в соответствующей легенде) комплект может иметь еще т.н. мнимые ключи, принимающие во всех записях этого комплекта постоянные значения. Последние определяются в описании файла конструкцией "ключи" и предназначены для уточнения порядка записей в файле.

Порядковый номер создаваемого мнимого ключа в последовательности всех ключей данного комплекта определяется конструкцией "номер ключа", а его постоянное значение - конструкцией "константа". При этом требуется, чтобы связанные с одним комплектом номера ключей были в описании файла представлены в порядке возрастания и максимальный из этих номеров не превышал сумму числа ключей соответствующей легенды и числа вводимых для этого комплекта мнимых ключей.

Константа, задающая значение мнимого ключа, может быть либо текстовой константой, либо натуральным числом, либо одной звездочкой - последняя обозначает максимальное значение соответствующего поля. Текстовая константа при этом не может содержать кавычки, а саму эту константу следует представить в кавычках тогда, когда она содержит хотя бы один из следующих пяти символов: (|*|,|␣|) или же состоит только из цифр. Например, корректным является фрагмент описания файла

*3 C (2 = 7, 4 = '17', 6 = 'ЗАВОД␣НР.␣1', 7 = РАБОЧИЙ)

(если предполагать, что в легенде C определено не менее трех ключей), фиксирующий в качестве значения первого из мнимых ключей число 7, а для остальных трех мнимых ключей - соответствующие текстовые константы.

Если указываемая константа короче поля, отведенного для соответствующего ключа в результате выравниваний, то текстовая константа дополняется справа пробелами, числовая - слева нулями. Обратим еще внимание на то, что фрагменты 1 = '*' и 1 = * имеют в конструкции "ключи" разное значение: в первом случае звездочка является текстовой константой, а во втором - обозначением максимального возможного значения.

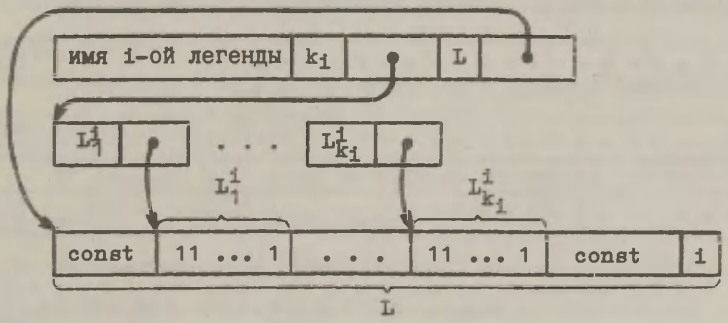
Получаемый в результате трансляции описания файла макет занимает одно связное поле памяти, которое по своему содержанию распадает на четыре подполя: заголовок, список легенд, список указателей ключей легенд и список т.н. шаблонов. Шаблон - это ключ доступа комплекта, в котором на поля, предназначенные для ключей легенды, написаны максимальные возможные значения (единицы во всех битах).

Заголовок макета содержит имя макета (8 байтов), число легенд в описании (1 байт) и общую длину макета (3 байта).

В списке легенд для каждой легенды отведено 16-байтовое поле, содержащее имя легенды (8 байтов), число ключей этой легенды (1 байт), ссылку на начало последовательности указателей ключей (3 байта), длину шаблона (1 байт) и ссылку на шаблон этого комплекта (3 байта). Все эти ссылки даются относительно начала макета.

Каждый указатель ключа легенды содержит длину отведенного для рассматриваемого ключа после всех выравниваний поля (2 байта), а также ссылку на начало этого поля в соответствующем шаблоне (2 байта, ссылка дается относительно начала шаблона). Указатели, относящиеся к одной легенде, расположены в макете подряд.

Таким образом, любой (i-ый) элемент списка легенд, указатели ключей этой легенды и шаблон соответствующего комплекта связаны между собой относительными ссылками, схематически представленными на следующем рисунке (где через const обозначены константы, образованные из компонент координатного вектора, заданных значений мнимых ключей и/или нулей):

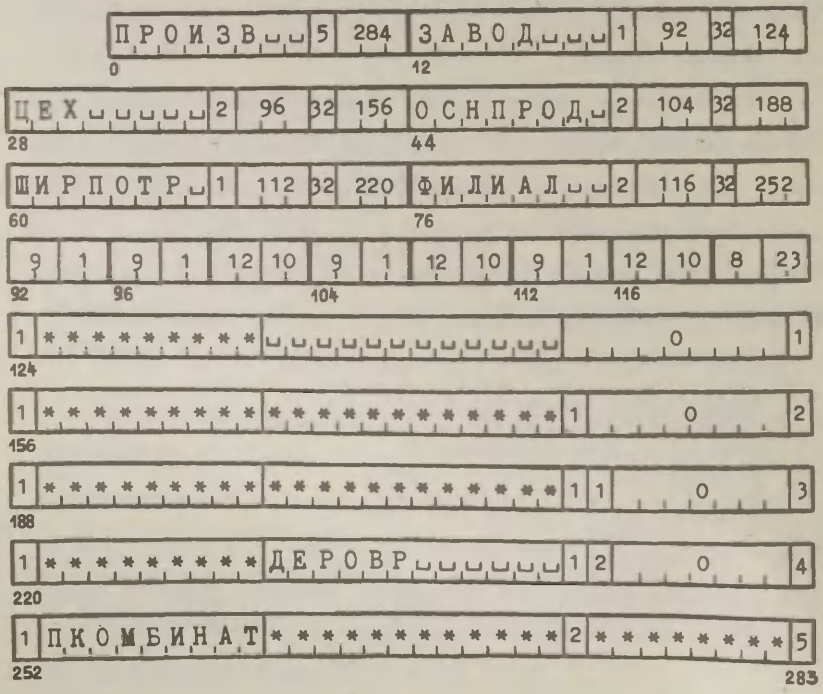


2

Рассмотрим в качестве примера следующее описание файла (где справа за вертикальной чертой для каждой легенды указано число ее ключей, а также их длины в байтах):

```
FILE ПРОИЗВ
  *1 ЗАВОД (2='␣')      | 1 ключ,  длина 6
  *2 ЦЕХ                | 2 ключа, длины 8 и 10
  *3 ОСНПРОД            | 2 ключа, длины 8 и 12
  *3 ШИРПОТР (2=ДЕРОБР) | 1 ключ,  длина 6
  *2 ФИЛИАЛ (1=ПКОМБИНАТ) | 2 ключа, длины 8 и 8
EOF
```

Получаемый в результате трансляции этого описания макет занимает 284 байта, которые заполнены следующим образом:



В целях улучшения читаемости этой схемы нумерация байтов на ней указана при помощи десятичных целых чисел. Такими же числами представлены все ссылки и прочие численные константы, встречающиеся в макете. Кроме того, максимальное значение байта обозначено здесь через * (вместо традиционного FF_{16}).

Даже не имея под рукой текстов используемых в этом описании легенд, можно угадать, что первый ключ каждой легенды (кроме легенды ФИЛИАЛ) является именем завода. При этом, по-видимому, только один завод имеет филиалы и имя этого завода (ПКОМБИНАТ) прибавлено ко всем филиалам в виде значения соответствующего мнимого ключа. Аналогично к комплекту ШИРПОТР добавлен второй ключ (имя цеха).

Следует также обратить внимание на то, что мнимый ключ (2='␣') добавляется к комплекту ЗАВОД лишь в целях выравнивания длин вторых ключей братьев ЦЕХ и ФИЛИАЛ.

Литература

1. Изотамм А., Каазик Ю., Томбак М., Язык определения записи. Труды ВЦ ТГУ, 1978, № 41, 7-64.

2. Каазик Ю., Образование файлов. Труды ВЦ ТГУ, 1978, № 41, 65-74.

3. Рауп А., Организация файлов в системе РАМА. Труды ВЦ ТГУ, 1982, № 49, 12-25.

4. Ээремаа К., Язык управления заданиями системы РАМА. Труды ВЦ ТГУ, 1983, № 50, 3-22.

5. Изотамм А., Дерево описания записи в системе РАМА. Труды ВЦ ТГУ, 1980, № 43, 3-35.

СРЕДСТВА АВТОМАТИЗАЦИИ ХРАНЕНИЯ И ВОССТАНОВЛЕНИЯ ФАЙЛОВ И ПРОГРАММ

Ю.Кяхрик, Х.Нярипя, А.Яэгер

Возрастающие мощности вычислительных ресурсов - объема оперативной памяти, скорости процессора и каналов, а также количества и объема устройств прямого доступа позволяют резко увеличить степень мультипрограммирования в ОС ЕС. Очень важной при этом становится роль оператора ЭВМ в управлении вычислительным процессом. Поэтому ясно, что без вспомогательных программных средств и автоматизации сервизных работ рациональное использование вычислительных ресурсов невозможно.

В вычислительных центрах научного и учебного наклона, где, как правило, много пользователей различной квалификации (занимающихся в основном составлением и отладкой программ), одним из коренных вопросов является обслуживание библиотек программ и данных. Основные принципы и некоторые программные средства, внедренные на ВЦ ТГУ в этой отрасли, приведены в [1]. За последние годы развитие системы шло по направлению автоматизации, повышения надежности и скорости обработки библиотек.

В настоящей статье рассматриваются принципы автоматизации использования архивного фонда программ и данных, а также описывается часть программных средств архивной системы, которая может представлять интерес более широкому кругу читателей. Ограниченный объем статьи не позволяет рассматривать всех подробностей описываемых средств.

Принципиально новый, внесистемный подход используется при работе с магнитными лентами. Это обусловлено тем, что в практической работе системные средства и методы обработки данных на магнитных лентах оказываются ненадежными и недостаточно гибкими.

Следует отметить, что существование архивного фонда и системы обслуживания являются лишь вспомогательным средством при управлении вычислительным процессом. Использование этих средств дает эффект в том случае, если разработана методика обслуживания библиотек и данных в целом, которая учитывает конкретные особенности и требования вычислительного центра. В статье коротко рассматриваются некоторые принципы, на которых основываются использование и обслуживание библиотек в ВЦ ТГУ.

1. Структура архивного фонда

Архивом или архивным фондом мы называем совокупность программ и данных, записанных в некотором определенном виде на магнитные ленты, откуда их при надобности можно восстановить в исходные библиотеки. Весь архивный фонд разделен на

подархивы так, что в каждом подархиве хранятся однотипные объекты (исходные тексты, модули, процедуры, макроопределения и т.д.). При записи некоторого библиотечного набора данных (или его части) в архив каждый раздел отмечается соответствующим признаком подархива как на ленте, так и в управляющем каталоге (см. п. 2). Раздел представляется в архиве физическими блоками, которые в свою очередь группируются в файлы. В одном файле содержится определенное число разделов, которые обычно принадлежат к одному подархиву.

Архивный том на магнитной ленте содержит стандартную метку тома VOL1, за которой следуют файлы. Ограничителями между файлами являются блоки ленточных марок (TM), за последним файлом на ленте записываются два блока ленточных марок. Таким образом архивный том имеет следующую структуру:

VOL1	файл 1	TM	файл 2	TM	...	TM	файл n	TM	TM

Такое расположение данных на магнитной ленте позволяет значительно ускорить поиск нужного раздела (за счет перемотки ленты через файлы) и уменьшить зависимость обработки от сбоев ленты.

Каждый физический блок файла представляется следующим образом

32	≤ 8000	32
KEY	данные	KEY

На поле KEY записывается начальный (или же идентичный с

ним конечный) ключ блока. Ключ используется при поиске блока и содержит следующую информацию:

2	2	4	8	1	1	1	1	1	3	4	4
L	T	I	NAME	FLAGS	S	BM	E	BNR	MADR	DATE	USER

где L - длина поля данных в байтах; T - номер тома (два последние символа имени тома); I - внутренний идентификатор системы; NAME - имя раздела; FLAGS - байт флажков (в настоящее время используется лишь нулевой бит этого байта, который устанавливается в единицу для последнего блока раздела); S - номер подархива; BM - порядковый номер блока в разделе; E - номер экстента; BNR - число блоков раздела (Ø, если эта величина не установлена); MADR - адрес раздела в томе (адрес представляется однобайтовым номером файла и двухбайтовым номером блока в файле). На поле DATE указывается дата создания раздела в библиотеке и USER предоставляется пользователю.

Экстентом является некоторая порция разделов по выбору пользователя. Обычно это разделы одного подархива, обработанные за один сеанс работы. Экстенты используются в процессе слияния подархивов.

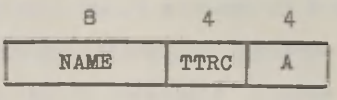
На поле данных в физическом блоке записывается очередная часть раздела. Внутренняя структура этого поля зависит от подархива и со стороны архивного фонда интереса не представляет.

2. Управляющий каталог архива

Управляющий каталог предназначен для быстрого нахождения из архива нужных разделов при восстановлении. С этой целью в каталоге строятся специальные элементы для архивных томов, а также для занесенных в архив разделов библиотек. Подробное описание таких элементов дано ниже.

В начале работы с некоторым архивным томом соответствующий элемент тома загружается в оперативную память в виде т.н. управляющего блока тома VCB (при использовании нового, некаталогизированного тома образуется новый блок VCB). Этот блок модифицируется в ходе каждой записи в том, а после завершения раздела или при закрытии тома записывается обратно в каталог. Элемент раздела содержит ссылки на разные версии раздела. Эти ссылки необходимы программам восстановления для поиска раздела. При записи в архив новой версии раздела ссылки обновляются.

Управляющий каталог архива находится на магнитном диске и представляется как библиотечный набор данных. Элемент справочника такой библиотеки имеет известную структуру



где поле NAME содержит имя раздела или префикс имени архивного тома (в виде ¤¤XXXX.. - имя тома получается добавлением к XXXX номера тома из блока VCB), а TTR (C=2) указывает адрес элемента каталога в библиотеке. На поле A каждому подархиву

выделено по два бита - ненулевое значение некоторой пары битов означает, что раздел представлен в соответствующем подархиве (номер подархива определяет положение битов, см. табл. на рис. 1). В элементе, ссылающем на блок архивного тома, поле A не используется.

В таблице на рис. 1 приведены коды и номера подархивов (определяется при генерации системы).

Элемент каталога (раздел библиотечного набора данных) состоит из блоков по 36 байтов. В случае элемента архивного тома первым является блок тома, за которым следуют блоки экстентов. В случае же элемента раздела каждый блок соответствует одному имеющемуся подархиву.

Блок тома имеет следующую структуру:

2	2	1	3	1	1	2	4	4	4	2	10
T	MNR	ST	BADR	CENR	ENR	BNRF	BNRT	D1	D2	SM	USER

На поле T записывается номер тома; MNR - количество разделов в томе; ST - байт состояния блока VCB; BADR - адрес первого свободного блока в томе (номер файла и номер блока в файле); CENR и ENR - соответственно количество блоков экстентов в каталоге и экстентов в томе (блоки экстентов резервируются с некоторым запасом); BNRF - количество блоков в последнем файле; BNRT - количество блоков в томе; D1 и D2 содержат соответственно дату первой и последней записи в томе; SM - маска подархивов (ненулевое значение некоторого бита маски определяет соответствующий подархив, см. табл. на рис. 1); USER - поле пользователя.

Код подархива	Номер подархива	Соответствующие биты на поле A	Соответствующий бит в маске SM
TEXT	1	0 и 1	0
MOD␣	2	2 и 3	1
LINK	3	4 и 5	2
PROC	4	6 и 7	3
MAC␣	5	8 и 9	4
....	...	.....	...
...	F	28 и 29	14

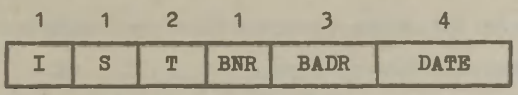
Рис. 1. Коды и номера подархивов.

Блок экстента содержит следующие поля:

2	2	1	3	4	4	1	4	4	10
SM	MNR	S	EADR1	BNRE	D1	E	EADR2	D2	USER

Здесь SM - маска подархивов экстента; MNR - количество разделов в экстенте; S - номер подархива; E - номер экстента; BNRE - количество блоков в экстенте. EADR1 и EADR2 - адреса первого и последнего блока экстента, а D1 и D2 - соответственно даты первой и последней записи в экстенте. Поле USER предоставляется пользователю.

Блок элемента раздела в каталоге (соответствующего одному подархиву) разделен на три сегмента по 12 байтов. Сегмент описывает одну версию раздела в архиве следующим образом:



где I - системная информация; S - номер подархива; T - номер тома; BNR - количество блоков раздела; BADR - адрес раздела в томе; DATE - дата составления этой версии раздела. Сегмент "новейшей" версии всегда первый в блоке, а сегмент самой "старой" версии - последний.

Как видно, описанная структура каталога накладывает на имена некоторые ограничения: в одном подархиве нельзя использовать одноименные разделы и имена архивных томов не должны совпадать с именами разделов.

3. Обслуживание архивного фонда

В систему обслуживания входят такие компоненты как архивация и восстановление библиотек исходных текстов, модулей или данных; автоматическое восстановление разделов в библиотеке и обслуживание архивных томов и каталога. Далее дается короткий обзор программных средств, используемых в настоящий момент в ВЦ ТГУ.

Библиотеки условно разделены на две категории - общие и пользовательские. Библиотеки пользователей предназначены для хранения исходных текстов, модулей и т.д. одной группы пользователей, общие же библиотеки - для всего ВЦ. Это разделение не очень строгое и производится лишь на уровне процедурных средств.

Архивация библиотек производится регулярно по мере заполнения библиотеки или по желанию пользователя. В настоящее время необходимость архивации определяет диспечерская служба по состоянию библиотеки. Внедряется монитор автоматического прослеживания библиотек, одной из задач которого является определение необходимости архивации библиотеки и запуск соответствующих программ архивной системы.

Архивация производится в следующих целях:

- для получения запасной копии библиотеки,
- для удаления неактивных разделов из библиотеки,
- для переноса данных с одной ЭВМ в другую.

При архивации разделы библиотеки свертываются и записываются на указанное место архивного фонда или на место, которую выбирает сама система. В каталоге ссылка на самую старую версию заменяется новой.

Библиотеки (или их части) восстанавливаются в случае разрушения имеющейся библиотеки или при активации удаленной библиотеки. Отдельные разделы восстанавливаются, как правило, автоматически.

Архивацией и восстановлением можно управлять выборочными предложениями, при помощи которых включаются или исключаются указанные разделы (можно также указать разделы, имена которых начинаются с указанным префиксом или разделы, дата составления которых или частота использования равно, больше или меньше указанной величины). Задание на архивацию или восстановление библиотеки оформляется аналогично рассмотренному в [1], за исключением описания архивных томов - тут указывается лишь параметр DUMMY, DYNAM или UNIT, конкретный

том определяется в ходе работы (см. п. 4).

Автоматически восстанавливаются исходные тексты или модули текстовым корректором IEBUPDTE или редактором связей IEWL. Для этого расширены возможности названных системных программ (добавлены средства восстановления и управления ресурсами; принципы таких расширений рассмотрены в [1]).

Подархив уточняется по имени библиотеки, вторая компонента составного имени которой зафиксирована (напр. TEXT, MAC, LINK и т.д.). Архивный том с нужной версией раздела определяется по данным каталога.

Следует отметить, что в ВЦ ТГУ оказывалось целесообразным использование промежуточных общих библиотек при архивации разделов из библиотеки пользователя. Таким образом, неактивные разделы сперва переносятся в эту общую библиотеку, а оттуда по необходимости в архив. В процедурах корректирования, трансляции и сборки эти общие библиотеки сцеплены с библиотеками пользователя. Автоматическое восстановление производится в библиотеку пользователя. Через общие библиотеки также удобно пересылать программы другим пользователям.

Программы обслуживания архивных томов и каталога позволяют соединить отдельные экстенты подархива, накапливать подархивы с новейшими версиями разделов в один том, сжать архивный том или каталог и т.п. Имеется и программа-конвертор для включения разделов из некаталогизированных архивных томов старой модификации в имеющийся архив.

4. Вспомогательные средства системы

Далее описываются макрокоманды, при помощи которых удобно обрабатывать каталог и архивные тома. Хотя эти макрокоманды созданы специально для работы с архивным фондом с вышеописанной структурой, они могут быть использованы и вне архивной системы с каталогами других типов (заменяя макрокоманды обслуживания каталога подходящими). Форматы макрокоманд не отличаются от принятых в языке АССЕМБЛЕР - каждой макрокоманде можно присвоить идентификатор, а на месте значения <адрес> указать по желанию номер регистра в скобках, содержащий нужный адрес.

Макрокоманды обслуживания каталога позволяют получить нужную информацию из элементов каталога, модифицировать их или добавить новые элементы в каталог.

Функционированием всей системы управляет т.н. блок управления архивом ACB, который составляется по макрокоманде

XAOPEN [PGM=<адрес>][,TAPE=<адрес>][,SA={<код> / <адрес>}]
[,OPT=(<режимы>)][,MSGL=(<режимы>)]

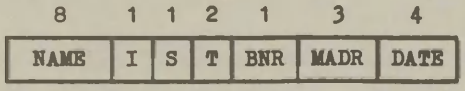
Здесь PGM определяет идентификатор пользователя, TAPE - адрес таблицы параметров для архивного тома, SA - код или адрес кода подархива. Параметрами OPT и MSGL уточняются соответственно режимы работы и вывода сообщений. При выполнении этой макрокоманды открывается каталог, проверяются и устанавливаются параметры для магнитной ленты (размер блока, ко-

личество блоков в файле и т.д.), составляется ACB, печатается заголовок и основные параметры системы.

При добавлении новых разделов в архив или при удалении ненужных соответствующие элементы каталога изменяются макрокомандой

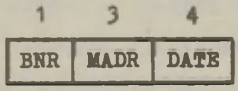
```
XASTOW SEEK=<адрес>[,TAPE=<адрес>][,DATE=<адрес>]
       [,ERR=<адрес>][,OPT=(<режимы>)]
```

где SEEK определяет адрес поля, содержащего данные для составления нового сегмента:



Здесь NAME - имя раздела, I - системная информация, S - номер подархива, T - номер тома, BNR - количество блоков, MADR - адрес первого блока раздела, а DATE - дата составления раздела.

С помощью параметров TAPE и DATE указываются, какие сегменты удаляются до введения в каталог нового сегмента: TAPE определяет номер тома, а DATE - следующее поле



Если оба эти параметра заданы, то удаляется один сегмент, определенный этими параметрами. При задании только TAPE

удаляются все сегменты данного тома, при задании только DATE - все сегменты старше указанной даты. Если оба параметра опущены, то удаляется последний сегмент.

Макрокоманда XASEEK предназначена для получения информации из каталога (на основе этой информации удобно переписывать нужные разделы из архива в соответствующую библиотеку):

XASEEK [TYPE=TAB|DIR|COM] [,REP=1|2|3|ALL] [,S=<адрес>] [,SLEN=<адрес>]
[,SNR=<адрес>] [,SA=<номер>|<адрес>] [,OPT=(<режимы>)]

Параметром TYPE уточняется операция: TAB указывает, что имена разделов уже имеются в таблице и их нужно дополнить информацией из элементов каталога; DIR указывает, что требуется создать таблицу имен из элементов справочника; COM включает возможности DIR и TAB. Параметр REP определяет порядковый номер сегмента, информация из которого нас интересует. S определяет начальный адрес таблицы; SLEN - число байтов таблицы, SNR - число элементов таблицы, SA - номер подархива или адрес маски подархивов (см. табл. на рис. 1), а OPT - режимы работы.

В конце работы с каталогом дается макрокоманда XACLOSE для выполнения заключительных действий и закрытия каталога.

Макрокоманды обработки архивного фонда предназначены для закрепления и освобождения устройств (XOPENT и XCLOSET), открытия и закрытия экстента (XOPENX и XCLOSEX), обработки раз-

дела (XOPENW, XWRITE и XCLOSEW или XOPENR, XREAD и XCLOSER), управления (XPOINT, XWAIT, XWIND), обработки физических блоков на ленте (XRBL, XWBL и XEOF). Обязательным для всех этих макрокоманд является параметр TB, задающий адрес 292-байтового блока управления томом. В параметре ERR можно задавать адрес собственной программы анализа ошибок.

В задании на использование магнитных лент должно быть для одновременной работы определено необходимое количество предложений DD с именами, начинающее префиксом TAPE и параметрами DUMMY, DYNAM или UNIT, определяющими ленточное устройство. Эти предложения используются в основном для освобождения устройств и в случае аварийного завершения задания.

Перед первым обращением к тому нужно выполнить макрокоманду XOPENT для закрепления устройства. По этой макрокоманде оператору выдается сообщение об установке тома и проверяется правильность установки.

XOPENT TB=<адрес>[,VOL=<адрес>][,GAPS=<адрес>]

[,OPER=([REP][,INPUT / ,OUTPUT / ,NEW][,REWIND][,TSW])]

[,STATUS=([NOFIXTSW][,NOCAT][,EXAUTO][,NEWF])]

[,ERR=<адрес>]

Параметром VOL задается адрес имени требуемого тома. По умолчанию имя тома формируется стандартно, используя номер тома из TB. Параметром GAPS указывается адрес таблицы макси-

мальных размеров физической записи, разделов в томе, блоков в файле и файлов в томе. Стандартные значения указываются при генерации или в макрокоманде XAOPEN.

Параметром OPER уточняются начальные действия с лентой: REP разрешает оператору заменить том, INPUT, OUTPUT и NEW определяют тип обработки тома, REWIND требует перемотку в точку загрузки, TSW указывает, что VCB (первые 32 байта ТВ) уже загружен.

Параметром STATUS определяются режимы обработки - модифицированный VCB не записывается в каталог после каждой операции записи (NOFIXTSW), том вообще не каталогизирован (NOCAT), открытие и закрытие экстентов выполняется автоматически (EXAUTO), новый экстент начинается с нового файла (NEWF).

При закреплении проверяется, не установлен ли нужный том на некотором устройстве. Если нет, то оператору выдается сообщение

$$M\,␣\,(uuu)\,␣\,VVVVVV\,␣\,[(R)]\left[\begin{matrix}INPUT\\OUTPUT\\NEW\end{matrix}\right][CATAL][REWIND]$$

где uuu - адрес устройства и VVVVVV - имя тома. Оператор может назначить новое устройство, новое имя тома (OPER=REP) или отказаться от установки, ответив NO или END. Правильность установки проверяется путем ввода одного блока с ленты. В точку загрузки лента перемотается только по параметру REWIND.

После успешного закрепления можно выполнить любые макрокоманды, относящиеся к этому тому. Следует подчеркнуть, что если используются макрокоманды без ожидания завершения, то перед очередной операцией с этим томом следует использовать макрокоманду

XWAIT TB=<адрес>[,ERR=<адрес>]

После завершения обработки тома устройство освобождается макрокомандой

XCLOSET TB=<адрес>[,DISP={END / TOP / LEAV / <адрес>}]

[,STATUS=([FIXSA / MIXT][,NOWRITE][,FREE])]

[,OPER=([NOEOP][,NOSTEX][,NOSTTSW]

[,NOFREE][,EVEN])]

[,USER=<адрес>][,ERR=<адрес>]

Параметром DISP определяется конечная позиция ленты: END требует перемотки к концу данных, TOP в точку загрузки, LEAV сохраняет текущее положение, а адресом можно указать трехбайтовый адрес установки (номер файла и номер блока в файле). Параметром STATUS определяется дальнейшее использование архивного тома (имеет смысл только для каталогизированных

томов): FIXSA запрещает создание новых подархивов и NOWRITE запрещает запись на том вообще. MIXT и FREE отменяют эти режимы. По умолчанию сохраняется старый статус тома. Значениями параметра OPER задаются дополнительные действия: NOEOP не требует записи конечных ленточных марок, NOSTEX и NOSTTSW не требуют записи в каталог блока экстента и блока тома, NOFREE не освобождает устройства, EVEN не закрывает том и не освобождает устройства даже в случае ошибок. Параметр USER определяет адрес информации пользователя, которая записывается в блок тома и блок экстента.

Если в макрокоманде XOPENT не указан STATUS=EXAUTO, то пользователь сам должен открыть экстент перед обработкой и закрыть по завершению обработки соответственно макрокомандами

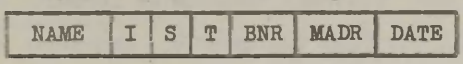
XOPENX TB=<адрес>[,SA={<номер>|<код>|<адрес>}][,OPER=CONT][,ERR=<адрес>]

XCLOSEX TB=<адрес>[,USER=<адрес>][,ERR=<адрес>]

Параметром SA определяется номер подархива, код подархива или соответствующий адрес. Код Ø соответствует т.н. "смешанному" экстенту и является значением по умолчанию. OPER=CONT означает, что добавленные разделы следует соединить с последним экстентом тома. Параметр USER определяет информацию пользователя для блока экстента.

Макрокоманды обработки раздела освобождают пользователя от многих подготовительных работ и позволяют сосредоточить

внимание только на обработку самих данных. Состояние раздела располагается в блоке раздела



который используется и в макрокомандах обслуживания каталога.

Для вывода раздела из библиотеки в архив используются макрокоманды XOPENW, XWRITE и XCLOSEW, соответственно для открытия, записи очередного блока и закрытия раздела. На одном томе можно одновременно обрабатывать только один раздел.

XOPENW TB=<адрес>,S=<адрес>[,IO=<адрес>][,ERR=<адрес>]

XWRITE TB=<адрес>[,IO=<адрес>][,L={<длина>|KEY}]

[,USER=<адрес>][,ERR=<адрес>][,WAIT=NO]

XCLOSEW TB=<адрес>[,S=<адрес>][,IO=<адрес>]

[,L={<длина>|KEY}][,USER=<адрес>][,ERR=<адрес>]

[,WAIT=NO][,TYPE=X'nn'][,OPER=DELETE]

Параметром S задается адрес блока раздела, где заранее заполнены поля NAME, S, BNR (если неизвестно, то Ø) и DATE. Параметром IO задается адрес буфера ввода. В этот буфер пользователь заносит данные перед выполнением макрокоманды

XWRITE. Ключ к данным добавляется автоматически. Если IO указан в XOPENW, то в дальнейших макрокомандах этот параметр не обязателен. Параметром L задается длина выводимого блока. По умолчанию это значение берется из XOPENT. Если задано L=KEY, то выводится блок без данных. Параметр USER задает адрес информации пользователя. WAIT=NO определяет режим вывода без ожидания - после запуска канальной программы управление возвращается в программу пользователя.

По макрокоманде XCLOSEW выводится последний блок и раздел закрывается. Параметр TYPE задает значение поля FLAGS в ключе последнего блока, причем нулевой бит устанавливается в единицу автоматически. Параметром OPER=DELETE уничтожаются все записанные блоки и обработка раздела завершается.

При выводе модифицируется элемент тома в каталоге после записи каждого раздела, если только не указываются режимы NOCAT или NOSTTSW.

Для ввода раздела из архива используется следующий набор макрокоманд:

```
XOPENR  TB=<адрес>,S=<адрес>[,IO=<адрес>]
        [,ERR=<адрес>][,WAIT=NO]
XREAD   TB=<адрес>[,IO=<адрес>][,ERR=<адрес>]
        [,WAIT=NO]
XCLOSER TB=<адрес>
```

Перед открытием раздела необходимо заполнить все поля

блока раздела. Вводимые данные помещаются в буфер с адресом IO, а ключ данных в буфер блока TB. Там располагается и количество введенных байтов.

Установку магнитной ленты в нужное положение можно выполнить макрокомандой

XPOINT TB=<адрес>[,P=<адрес>|END|NEXT][,ERR=<адрес>]

[,OPER=([TEST][,REPOINT][,BACK])][,WAIT=NO]

где параметром P указывается адрес трехбайтового поля адреса установки. P=END требует установку в конце данных, а P=NEXT перед первым блоком следующего или (при OPER=BACK) предыдущего раздела. Если этот параметр опущен, то макрокоманда определяет лишь место нахождения путем ввода очередного блока. Параметр OPER задает дополнительные режимы: TEST требует проверки правильности установки на основе данных ключа, REPOINT вторичную попытку установить том в случае ошибочной последовательности блоков, а BACK действий в обратном направлении.

Далее рассмотрим некоторые макрокоманды более низкого уровня, которые не требуют открытия раздела и из которых исключено проверка правильности раздела (ими используются, например, и в макрокомандах XREAD и XWRITE).

Ввод блока из архива осуществляется макрокомандой

XRBL TB=<адрес>[,IO=<адрес>][,L={<длина>|KEY}]

[,BRR=<адрес>][,WAIT=NO][,OPER=BACK]

Введенный блок располагается по адресу, указанному в параметре IO (это можно опустить в случае L=KEY). OPER=BACK требует ввода в обратном направлении и IO должен определить адрес последнего байта буфера.

По макрокоманде

XWBL TB=<адрес>[,IO=<адрес>][,L={<адрес>|KEY}]

[,KEY=<адрес>][,ERR=<адрес>][,WAIT=NO]

выводится блок с буфера IO. Если задан L=KEY, то параметр IO не используется. Параметром KEY указывается адрес, откуда берется подготовленный пользователем ключ (но поля L, T и MADR определяются во всех случаях системой).

Блоки ленточных марок выводятся по макрокоманде

XEOP TB=<адрес>[,OPER=([NOOAT][,TEND])]

[,ERR=<адрес>]

где OPER=TEND требует записи двух последовательных ленточных марок, а OPER=NOOAT блокирует обращение к каталогу. Операция XEOP не выполняется, если лента не установлена за последним блоком.

Для установки ленты в нужное положение можно пользовать-

ся макрокомандой

```
XWIND TB=<адрес>,OPER={TOP|BS|BB|PS|FB} [,L=<число>]
      [,ERR=<адрес>][,WAIT=NO]
```

где параметром OPER задается режим перемотки: по TOP в точку загрузки, по BS - на L блоков вперед, по BB на L блоков назад, по FS или FB - на L файлов вперед или назад.

Примером использования вышеописанных макрокоманд служит следующий фрагмент программы, где из архива восстанавливается в библиотеку исходных текстов раздел MEMBER.

```
         . . .
         XAOPEN  SA=SACODE
         . . .
         XASEEK  TYPE=TAB,S=SEEKT,SNR=TABL
         MVC     TAPET(2),SEEKT+1Ø НОМЕР ЛЕНТЫ
         XOPENT  TB=TAPET
         XOPENR  TB=TAPET,S=SEEKT,IO=BUF
         SR      3,3
         IC      3,SEEKT+12 ЧИСЛО БЛОКОВ
READ     XREAD   TB=TAPET
         LTR     15,15
         BNZ     ERROR ОШИБКА ВВОДА
         <разблокировка и запись в библиотеку. Длина
          блока находится в полуслове с адресом TAPET+76>
```

```
        . . .
        BCT     3,READ
        XCLOSET TB=TAPET
        . . .
        XACLOSE
SACODE  DC      C'TEXT'
TABL    DC      F'1'
SEEKT   DC      CL8'MEMBER'
        DS      CL12
TAPET   DS      CL292
BUF     DS      8ØØØC
```

Литература

I. Кяхрик Ю., Роомельди Р., Ээнма Т., Программные средства обслуживания общих библиотек программ. Труды ВЦ ТГУ, 1981, № 47, 3-36.

СИСТЕМА ДЛЯ ИНТЕГРАЦИИ ПАКЕТОВ

В. Лепинг

Процессу развития вычислительной техники и программирования в настоящее время характерно появление всё новых пакетов прикладных программ (ППП). Нередко разные фирмы, учреждения и лаборатории, которые разрабатывают математическое обеспечение для ЭВМ, составляют для одной и той же предметной области разные ППП. Как правило, у таких пакетов разная архитектура, и согласовать их можно только через данные, преобразовав заранее их структуру в соответствии с требованиями конкретных ППП.

В настоящей статье предлагается одна из возможностей для интеграции пакетов прикладных программ.

I. Введение

Специалисту, которому при решении своих проблем необходимо помощь ЭВМ, приходится выбирать соответствующее программное обеспечение:находить удовлетворяющий его пакет прикладных программ. В крайнем случае, если подходящий ППП не найден, он должен сам составлять нужные ему программы. Часто выясняется, что одного пакета мало - надо применять несколько ППП в неизменном или дополненном виде.

Все это заставляет его изучать ППП, выяснять их возмож-

ности, структуру, входный язык, средства ввода-вывода, а также редактирования и сохранения данных. Приходится прочитывать огромное количество документации, выбирать нужные ППП, при том не ошибиться, и учиться ими пользоваться - всё это заметно загромождает работу с пакетами. Дело обстоит ещё хуже в случае, если пользователю предметная область ППП не знакома. Например, во многих научно-исследовательских работах возникает необходимость статистической обработки данных. Но ученым других специальностей затруднительно самостоятельно работать с ППП по статистической обработке данных. Требуется посредник между пакетами и их пользователями. До сих пор посредниками, как правило, работают соответствующие специалисты (в нашем примере специалисты по статистической обработке данных). Но на самом деле, многие функции посредника можно поручить ЭВМ. Для этого необходимо создать диалоговую систему [1, 2, 3, 6] интеграции пакетов.

Очень важный показатель - удобство пользования ППП. Персональная ЭВМ для специалиста гораздо лучше большой ЭВМ, не говоря о цене машинного времени и о других хлопотах, связанных с большими машинами . Большинство пакетов, однако, реализованы на больших ЭВМ. И это обычно оправдано - многие пакеты в настоящее время персональным ЭВМ ещё не под силу. Тем не менее, персональная ЭВМ всё равно может быть применена как посредник. Ей можно поручать вспомогательные работы, связанные с изучением ППП, составлением заказов для его работы, подготовкой данных. Большой ЭВМ останутся действия с большими наборами данных.

Разработка ППП началась с середины 60-х годов, и в нас-

тоящее время их количество достигает нескольких тысяч. Это создает хорошую базу для интеграции программного обеспечения. Начиная с конца 70-х годов проводятся работы по объединению ППП в суперпакеты и системы [4, 5, 7].

Подобные системы обычно состоят из базы данных и ряда пакетов прикладных программ, которые соединены через мониторную программу. Применяются и другие решения: база данных или один из ППП принимается за центральный, а остальные приводятся в подчинение. В этом случае приходится модифицировать как доминирующий, так и подчинённые пакеты, а это, зачастую, приводит к осложнениям [8].

Так как появляются всё новые и новые пакеты, то трудно заранее фиксировать, какие из них включить в интегрированную систему. Другими словами, система интегрирования должна быть динамичной.

Рассматриваемая в данной статье система APIS (Applied Packages Integration System) предназначена для создания в интерактивном режиме интерактивных систем пакетов, притом пакеты могут находиться в разных ЭВМ. В системе предусмотрены средства защиты от неправильного использования включённых в пакеты методов, вызванного недостаточной подготовкой пользователя. Также имеются средства для повышения степени самообъяснения разных методов.

2. Описание системы

Каждый специалист - потенциальный пользователь ППП - решает свои проблемы в рамках некоторой предметной области. С другой стороны, для каждого ППП определена его область

применения. У разных ППП эти области могут быть разными или совпадать полностью или частично . При соединении ППП объединяются и их области применения. Исходя из интересов пользователя целесообразно соединять пакеты в систему, область применения которой покрывала бы определённую предметную область.

В системе APIS центральную роль играет т.н. DSR-модель, при помощи которой описывают желаемую предметную область, соединяемые пакеты и соотношения между пакетами и предметной областью. В действительности каждая DSR-модель является новой компонентой системы APIS и состоит из трёх моделей, которые назовем соответственно сценарием диалога, семантической моделью и моделью решения.

Пользователи системы APIS классифицируются следующим образом:

1. конечный пользователь, который общается с системой ППП на уровне DSR-модели в форме машинно-управляемого диалога;
2. конструктор - человек, который выясняет требования конечного пользователя и по этим требованиям выбирает нужные ППП и составляет DSR-модель;
3. диспетчер - человек, который координирует работы остальных пользователей и распределяет ресурсы ЭВМ.

Для конструирования системы ППП нужна также помощь системного программиста, свободно ориентирующегося в системе.

Дело в том, что для объединения пакетов обычно требуется составить некоторые программы, например, программы интерфейсов, программы тестирования данных и т.п. , а это не

входит в компетенцию конструктора.

2.I. Общая структура системы APIS

Система APIS состоит из следующих компонент:

- управляющая программа MONITOR;
- DSR-модели;
- банк данных пользователей;
- редактор данных пользователей;
- интерфейсные программы;
- пакеты прикладных программ.

Программа MONITOR управляет всей работой системы и является посредницей между пользователями и системой. Её средствами создаются DSR-модели, проводится управление редактором данных и вызываются интерфейсные программы и ППП.

DSR-модель состоит из сценария диалога, семантической модели и модели решения. Сценарий диалога является совокупностью т.н. меню, которые в свою очередь состоят из вопросов и перечней допустимых ответов. Кроме перечисленных ответов всегда дозволены специальные ответы "HELP" и "I DON'T KNOW".

Семантическая модель служит для определения предметных функций, т.е. соотношений между понятиями предметной области.

Модель решения для каждой предметной функции определяет некоторую "цену" её выполнения каждым пакетом, реализующим данную функцию. При необходимости "цены" можно присваивать и понятиям предметной области.

Данные всех пользователей хранятся в едином системном банке данных. Доступ к ним возможен лишь посредством систем-

ного редактора данных EDITOR. Основными функциями редактора являются ввод, редактирование и распечатка данных.

Интерфейсные программы реализуют обмен информацией между DSR-моделями и пакетами, а также обмен данными (включающий преобразование организации данных) между системным банком данных и пакетами.

Общую схему системы APIS объясняет схема (рис. I). На этой схеме как и на последующих схемах наборы данных изображены простыми прямоугольниками, программы - прямоугольниками с двойной рамкой, а движение информации или данных - стрелками, указывающими направление движения.

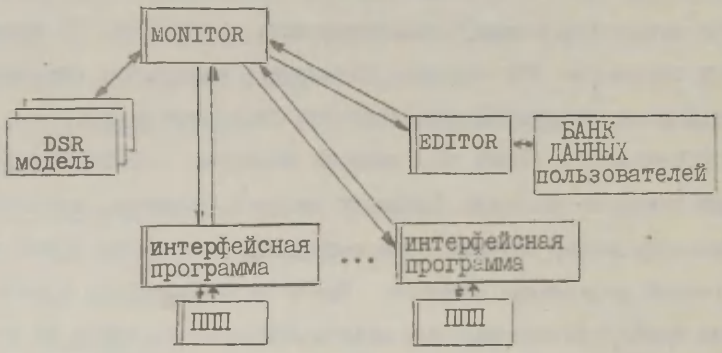


Рис. I

Основным пользователем системы APIS является конечный пользователь. Он общается с системой в рамках машинно-управляемого диалога, который определен DSR-моделью. В ходе диалога система выясняет желания пользователя и составляет по ним заказ для пакета или заказы для нескольких пакетов. Если это предусмотрено в используемой DSR-модели, система может составить последовательность заказов, в которой составление

очередного заказа зависит от результата выполнения предыдущего. В конце сеанса пользователь может получить распечатку проведенного диалога.

2.2. DSR -модель

2.2.I Сценарий диалога можно представить как ориентированный размеченный граф, где метками вершин служат вопросы, а метками дуг - допустимые ответы. Кроме того, из каждой вершины исходят еще две специальные дуги. Первая из них образует петлю, и её меткой является текст, выдаваемый пользователю в случае специального ответа "HELP". Вторая специальная дуга метки не имеет и соответствует ответу "I DON'T KNOW".

Единственное существенное ограничение при составлении сценария диалога следующее: без вышеупомянутых специальных дуг граф сценария является деревом.

Для составления сценария диалога конструктор имеет следующие системные средства:

I. Порождение нового сценария реализуется через порождение новой DSR-модели. После ввода имени новой DSR-модели, система генерирует корневую вершину графа.

2. Редактирование сценария диалога включает следующие операции:

- добавление дуги;
- удаление дуги;
- разметка дуги;
- добавление вершины;
- удаление вершины;

- разметка вершины;

- добавление специальных дуг.

Следует учесть, что меткой вершины может служить и оператор вызова подпрограммы. Такие программы оформляются по определенному стандарту и на выходе выдаётся сообщение о том, какая дуга выбирается в качестве ответа.

2.2.2 Семантическая модель представляется также в виде ориентированного размеченного графа, где метками вершин служат понятия и функции предметной области, а дуги описывают соотношения между понятиями и функциями. В системе прикладных пакетов понятиям соответствуют переменные или массивы, а функциям - модули или их части.

Вершины графа семантической модели порождаются в ходе конструирования диалога. Именно, некоторые дуги графа сценария соответствуют вершинам графа семантической модели, т. е. метки этих дуг являются понятиями и функциями предметной области. Дуги в графе семантической модели определяет конструктор, указывая входные и выходные понятия для каждой функции. Кроме того, каждой функцией определяется некоторое логическое условие, указывающее, при каких условиях допускается применение этой функции.

Роль конструктора при составлении семантической модели заключается в следующем:

1. Во время составления сценария диалога указывать, какие ответы являются понятиями предметной области и какие - функциями;
2. Для каждой функции указывать её входные и выходные понятия, а также логическое условие применения.

2.2.3 Модель решения будем рассматривать как ориентированный граф, который получается из графа семантической модели путем переориентирования всех его входных дуг. Кроме того, с вершинами, соответствующими функциям, связаны "цены", указывающие "стоимость" применения функций для всех пакетов. Аналогично, "цены" ставятся в соответствие с вершинами понятий. Заметим, что цены необязательно являются постоянными величинами.

2.2.4 Программы тестирования предназначены для проверки соответствия данных пользователя некоторым требованиям. Каждая программа тестирования выдаёт логическое значение, которое интерпретируется как результат теста. Управляющая программа MONITOR запускает все программы тестирования и объединяет их результаты в т.н. индикаторный вектор, который связывается с данными пользователя.

С каждой функцией семантической модели связано некоторое логическое условие. В действительности для всех логических условий единственным аргументом является индикаторный вектор. Такой подход осуществляет настройку семантической модели на данные пользователя.

3. Описание функций программы MONITOR

Программа MONITOR состоит из главной программы MANAGER, программы DIALM для создания и модифицирования DSR-модели, программы HANDLER для регистрации пользователей и проведения статистики использования системы, программы INFOR для осведомления пользователей и программы USER для управления решением при помощи пакетов прикладных программ (см.рис.2).

3.I Программа MANAGER проводит загрузку всей системы APIS и при необходимости обращается к одной из следующих программ: DIALM, HANDLER, INFOR или USER.

3.2 Программа DIALM общается с конструктором, осуществляет создание и модифицирование DSR-модели, а также подключение тестирующих и интерфейсных программ к системе.

3.3 Программа HANDLER общается с диспетчером, выдает статистику использования системы и ведет регистрацию пользователей и их ресурсов.

3.4 Программа USER общается с конечным пользователем. Она управляет процессом решения прикладных задач и вызывает редактор данных конечного пользователя (программа EDITOR) для ввода и модифицирования данных пользователя. Для управления процессом решения, программа USER пользуется DSR-моделью, тестами, программами интерфейса и пакетами прикладных программ.

3.5 У программы EDITOR три основных функции:

- ввод и модифицирование данных конечного пользователя;
- проведение логического и статистического контроля над данными во время ввода;
- подготовка подмножества данных конечного пользователя, необходимого для непосредственного решения прикладных задач.

3.6 DSR-модель состоит из сценария диалога, семантической модели и модели решения. Сценарий диалога - это совокупность меню. На рис. 3 показано одно из меню и соответствующий ему фрагмент из сценария диалога на логическом уровне.

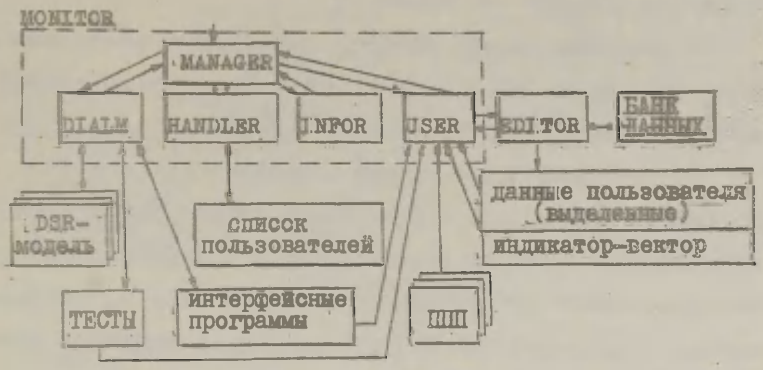


Рис.2

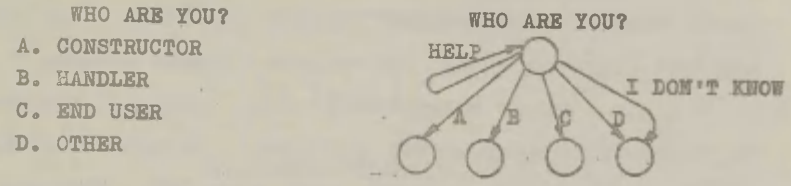


Рис.3

Для иллюстрации семантической модели и модели решения требуется пример гораздо длиннее, а это уже выходит за рамки данной статьи.

4. Программа USER

Основной частью системы с точки зрения пользователя является программа USER. Под управлением этой программы проводится решение прикладной задачи пакетами прикладных программ.

Программа USER состоит из следующих подпрограмм: DIAL,

DIAL2, PSP, STARTI и STARTP.

4.I Программа DIAL общается с конечным пользователем в форме диалога по DSR -модели и составляет первоначальный заказ решения прикладной задачи для ППП.

4.2 Программа DIAL2 также общается с конечным пользователем по DSR -модели и обрабатывает первоначальный заказ, используя при этом семантическую модель и индикатор-вектор, сгенерированный по результатам тестирования данных пользователя. Результатом работы программы DIAL2 является заказ на решение задачи для ППП, пока в терминах DSR -модели.

4.3 Программа PSP определяет, с помощью каких пакетов решать задачу, представленную заказом. Для этого она использует третью часть из DSR -модели - модели решения.

4.4 После выбора пакетов программа STARTI запускает соответствующие интерфейсные программы. На рис. 4 показан случай, когда весь заказ будет решён одним ППП . Интерфейсные программы переводят заказ из терминов DSR -модели в термины входного языка и синтаксиса соответствующих ППП, а также преобразуют системную организацию данных пользователя в организацию данных соответствующих пакетов.

4.5 Программа STARTP запускает конкретные пакеты, выбранные программой PSP. Результатами работы программ ППП могут быть листинги, наборы данных и сообщения для системы. Два последних, предназначенные для системы, пропускаются через интерфейсные программы, которые преобразуют организацию данных из конкретного пакетного вида в системный. Листинги работы пакетов идут к пользователю без изменений.

Схема функционирования программы USER показана на рис.4.

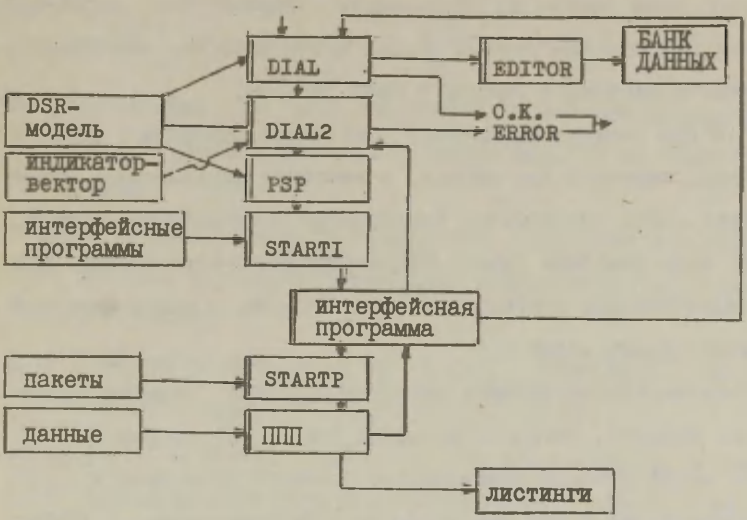


Рис.4

5. Использование системы

Человек соприкасающийся с системой APIS впервые и желающий решать содержательную задачу, а не просто что-либо спросить, должен обратиться к диспетчеру системы.

5.I Диспетчер выделит ему ресурсы ЭВМ и зарегистрирует его в списке пользователей, притом определяя уровень использования (уровень пользователя) системы. Пользователь имеет возможность поднять уровень, но только при условии удачной сдачи экзамена системе, или консультации у диспетчера или конструктора системы. После регистрации человек может пользоваться системой в пределах, определенных диспетчером.

5.2 Ввести данные пользователь может самостоятельно, общаясь непосредственно с системой. Вторая возможность - под-

готовить свои данные на перфолентах, перфокартах, магнетных лентах, кассетах или гибких дисках и передать их диспетчеру, который затем введет данные в банк системы.

5.3 Для решения конкретных задач надо сообщить системе свое имя, пароль и имя данных, с которыми пользователь будет работать. Если для решения понадобится подмножество данных, то его надо выделить сразу. После выделения подмножества данных, пользователь сообщает имя DSR-модели, в рамках которой он будет решать задачу.

Решение задачи система проводит в рамках машинно-управляемого диалога, сценарий которого дан в DSR-модели.

Во время диалога пользователь сообщает свои цели и система составляет заказ или сообщает о невозможности решить эту задачу в рамках данной DSR-модели. После составления заказа система выберет пакеты прикладных программ и решит задачу при помощи пакетов (см. рис. 4).

После решения задачи система выдает пользователю справку о том, сколько он потратил ресурсов ЭВМ и сколько ему ещё осталось.

По желанию пользователя система печатает стенограмму диалога в рамках DSR-модели.

5.4 Система APIS работает в операционных системах типа CP/M на персональной ЭВМ. Она создана для интеграции ППП, работающих в операционных системах типа CP/M и ОС-ЕС. Типы операционных систем и ЭВМ выбраны исходя из практических соображений.

Литература

I. В.Д.Алексеев, Принципы построения гибких диалоговых систем. Обработка символьной информации. ВЦ АН СССР, 1984, 29-39.

2. О.А.Гончаров, Адаптивный диалог - технология программирования. Обработка символьной информации. ВЦ АН СССР, 1984, 5-17.

3. Г.В.Сенин, О сетевой модели диалога. Обработка символьной информации. ВЦ АН СССР, 1984, 18-28.

4. Beutel, P., Heidelberg, D., New developments in integration of statistical systems. COMPSTAT 1980: Proceedings in Computational Statistics, p. 470-476.

5. Borzemski, L., Koszalka, L., Statistical data analysis using a data base for experiment design and systems identification. COMPSTAT 1984, Proceedings in Computational Statistics, p. 257-262.

6. Kaiser, P., Štetina, I., A dialogue generator Software - Practice and Experience, Vol. 12, 1982, p. 693-707.

7. Neumann, K., Metainformation as a tool for integration. COMPSTAT 1984, Proceedings in Computational Statistics, p. 273-278.

8. Stevens, F., Some cautionary aphorisms for user-oriented computer management. IFIP-80: Proceedings, p. 791-796.

ПРИМЕНЕНИЕ ПРИКЛАДНОЙ СТАТИСТИКИ В ЭМПИРИЧЕСКОЙ СОЦИОЛОГИИ

Л.М. Тоодинг

1. Постановка проблемы

Методы прикладной статистики твердо укоренились в исследовательской работе специалистов разных профессий. Предпосылки для дальнейшего роста роли статистического анализа данных создаются, в основном, двумя процессами:

1) интенсивным распространением вычислительной техники разных видов, причем создание и адаптация программного обеспечения ориентируется на пользователя, на достижение "товарного" вида программ;

2) повышением уровня формализованного анализа в конкретных областях приложения.

Предпосылкой для успешной реализации этих тенденций является тесная связь между статистиками и специалистами области приложения, связь, в которой заинтересованы и сами статистики-прикладники, чтобы следить за свойствами и качеством разработанной ими методики в реальных условиях массовой эксплуатации. Изучение восприятия пользователем сущности статистических методов создает основу для повышения практического эффекта прикладной статистики. Проблема результативнос-

ти анализа в высшей мере актуальна, так как появление персональных компьютеров делает практически без ограничений доступными самые разные методы анализа информации.

В данной статье рассматривается отношение к прикладной статистике в ходе социологических исследований. Из факторов, влияющих на успешность анализа данных, особо выделяется степень включенности пользователя в процесс анализа. Выбор социологии в качестве примера прикладной области связан, в основном, с тем, что социология выделяется, с одной стороны, размахом, весомостью общественной регулятивной роли своих результатов, а с другой стороны, ей свойственно, что ход измерения и анализа эмпирической информации сравнительно трудно репродуцируем. Это возлагает особую ответственность на исследователя, требуя безупречно корректного хода анализа.

Отправным пунктом статьи является желание проиллюстрировать на примере реальных социологических текстов недоверчивость социолога к разнообразию статистических методов (п.2). В качестве причин такой предосторожности выделяются принципиальные трудности математизации (п. 3.1), а также разные по результативности анализа подходы к построению стратегии машинной обработки (п. 3.2), присущие разным категориям пользователей (п. 3.3).

2. О выводе формализованного анализа в эмпирической социологии

Характеристика роли и эффективности статистического анализа в конкретной области применения является трудной задачей, так как предполагает изучение профессиональной творче-

ской деятельности, поддающейся анализу и оценке в ограниченной мере, в отношении к отдельным аспектам. В нашем рассмотрении изучается уровень реального, практического применения статистических методов в социологии. Выбранный для этого путь - анализ публикаций по эмпирической социологии - ведет к характеристике итогового вывода социологического исследования, оставляя процесс статистической обработки в стороне. Но, предположительно, уровень использования статистической аппаратуры в публикации создает определенную картину и о роли этой методики в исследовании в целом.

Ниже для анализа реального социологического текста выбраны тезисы докладов трех научных конференций по социологии молодежи и социологии семьи ([2], [4], [6]), всего 288 статей. Рассматривались следующие аспекты:

1) степень идентификации в тексте проведенного исследования, примененной методики,

2) логика изложения,

3) набор статистических понятий, представляющих эмпирические факты.

Хотя тезисы докладов отличаются специфической, сжатой формой и ограниченными возможностями, нет сомнений, что в тезисы автор включает самые выразительные, глубоко им осмысленные результаты изучения проблемы. Это относится в равной степени как к концептуальной, так и к эмпирической стороне. Другими словами, фиксируются те эмпирические факты, и в той форме, что по оценке социолога является самым точным отражением предметной сущности и, кроме того, соответствует самым наилучшим образом уровню среднего, типичного читателя, кото-

рого интересует данный эмпирический подход.

Рассматриваемая совокупность статей распределяется на три группы:

1) статьи на эмпирической основе, связанные с определенным набором социологических данных - 100 статей,

2) статьи, использующие статистические данные макроуровня (республика, регион) - 31 статья,

3) статьи концептуального или методологического характера - 157 статей.

Непосредственное изложение результатов анализа эмпирических данных, как видно, занимает значительное место в обмене текущей социологической научной информацией. Рассмотренные на концептуальном уровне вопросы также могут косвенно опираться на микроуровень, отличаясь большей обобщенностью. Далее предлагается характеристика первой группы статей.

Анализ показал, что идентификация исследования в рассматриваемой совокупности статей сравнительно скромная. Примерно треть из работ не содержит данных, достаточных для установления исследуемого контингента, цели изучения и т.п., в более 60% из работ не указывается время получения данных. Объем исследования отражается в объеме использованной статистической выборки (но лишь в половине работах), в то время как степень системности, комплексности измерения, выражаемая объемом набора исследованных показателей, не указывается. Статистическая значимость результатов, как правило, исследователями не указывается, так же, как и примененный комплекс методов. Поэтому более точное указание объема задачи всячески послужило бы большей состоятельности выводов.

По логике изложения материала выделяются два направления, которых условно можно называть описанием и аргументацией. В первом случае эмпирические факты слабо осмысляются, а во втором случае представление эмпирических данных непосредственно вытекает из логики предмета, факты служат для утверждения содержательного предположения. Описание и аргументация присутствуют в нашем материале в соотношении 2:3, что свидетельствует о заметной доле работ, опирающихся лишь на выразительность эмпирических фактов.

Базирующие на эмпирии выводы излагаются либо в чисто словесной формулировке (1/4 часть рассматриваемых работ), либо сопровождаются числами. В обоих случаях эмпирическими фактами в преобладающем большинстве являются средние значения или процентные распределения исследуемых показателей. Количество работ, где упоминаются еще, например, коэффициенты статистической связи, факторные признаки или другие показатели, интегрирующие информацию на основе определенной статистической процедуры, не превышает десятой доли.

Процент (процентное распределение) является, на самом деле, основой статистических процедур и несет в себе сущность явления. Однако, разнообразие процедур и получаемых их выводом статистических показателей создано именно для того, чтобы в каждом конкретном случае самым выразительным образом перейти с содержательного уровня на уровень эмпирических фактов.

В рассматриваемых работах широко применяется сравнительный анализ процентов и средних, имеющий явно интуитивный, а не количественный характер, не говоря уже о строгой вероят-

ностной обоснованности. Сравнение происходит в разрезе одной, в отдельных случаях и одновременно в разрезе 2-3 дискриминантных показателей. Это очередное доказательство того, что размерность анализа низкая, показатели рассматриваются в отдельности.

Возникает парадоксальная ситуация: социологам доступно и в социологической литературе апробировано множество современных, в том числе многомерных, методов статистики, но на "будничном" уровне самовыражения социологов они отражаются скромно. Не являясь сам убежденным в выразительности статистических методов менее традиционного характера, социолог не доверяет и профессиональной компетентности по прикладной статистике своих читателей, опасаясь быть непонятным.

В качестве замечания укажем на одну возможность конкретизации эмпирического вывода. Как известно ([1], стр. 19), в прикладной статистике выделяются статистический (основывающий на вероятностном характере данных) и нестатистический (не использующий вероятностные свойства) подходы к данным. Оба подхода могут опираться на один и те же формальные модели, но различаются друг от друга 1) при постановке задачи, 2) при интерпретации статистических показателей, которые в первом случае имеют прогностическую ценность для всей совокупности, а во втором осмысляются как описание исследованного, конкретного множества объектов. Вывод эмпирического анализа в социологическом тексте становится более четким и компактным, если исследователь в качестве методологического приема уясняет себе и читателю, какому подходу подлежит его задача.

Опираясь на все более возрастающий в последнее время интерес социологов к прикладной статистике и на интенсивную методологическую работу, проводимую рядом ведущих социологических центров ([3]), можно предполагать стремительный рост познавательной ценности статистики и в практической социологии.

3. Типология подходов к построению обработки

3.1. Трудности восприятия и усваивания результатов статистического анализа социологами выявляются в двояком смысле: с одной стороны - это принципиальные трудности математизации социологии, а другую создают субъективные факторы - профессиональные качества специалиста области приложения при разработке стратегии статистического анализа в сочетании с практической организацией обслуживания, введенной в конкретном центре обработки информации.

Сомнения и неуверенность в применении математики в социологии сочетаются с отрицательным отношением к математическим моделям вообще. Сущность аргументов против математизации на нынешнем этапе развития как математики (в особенности, математической статистики и анализа данных), так и прикладных по отношению к ней наук, в [1] (стр.66) отражается в следующих трех аспектах.

1) Изучаемые явления слишком сложны для адекватного отражения математическими средствами.

2) Математические модели, выражающие сущность поведения человека, лишают его индивидуальности.

3) В формализованном анализе процесс статистического мо-

делирования следует предпочитать разработке и применению математической модели в явном виде.

Ограничиваясь для опровержения приведенных против математизации возражений ссылкой на [1], где указываются пути их преодоления сквозь глаза математика, рассмотрим проблему статистического анализа на менее обобщенном уровне - с точки зрения конкретного пользователя статистическими методами.

В целях описания и анализа реальных задач прикладной статистики, на основе практики анализа данных, осуществляемого в Вычислительном центре Тартуского государственного университета, нами было проведено эмпирическое исследование в двух направлениях:

1) изучение реального потока задач [7] (две выборки с объемами N=670 и N=160, количество рассмотренных параметров M соответственно M=10 и M=60),

2) опрос пользователей [8] (N=90, M=140).

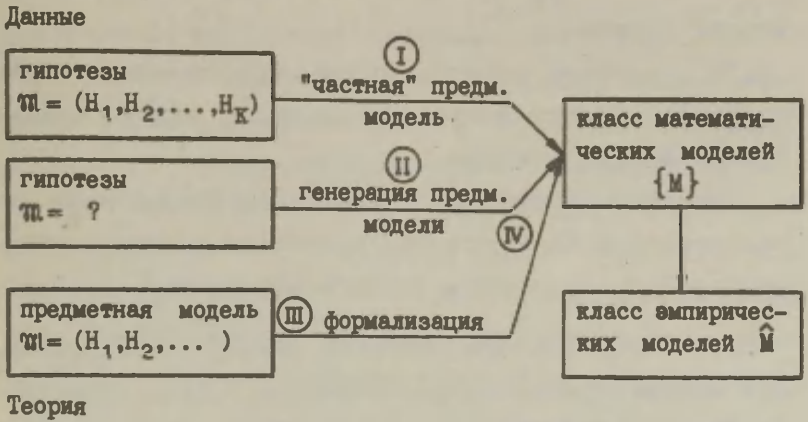
При решении исследованных нами реальных задач использовались как оригинальный, созданный на базе ВЦ ТГУ пакет статистического анализа данных, так и адаптированные пакеты.

Эти исследования показали, что влияющие положительно на результативность анализа факторы в конечном счете характеризуют меру включения пользователя ЭВМ в анализ: активность управления анализом, основательность обработки исходного массива и машинного вывода, компетентность в применяемых методах статистики. Другими словами, решающим оказался характер подхода к задаче. Рассмотрим некоторые возможности выделения типологии подходов к анализу данных, если применяется стандартная статистическая методика.

3.2. Объект социологического исследования представляется для статистического анализа в конечном виде признак-индивид матрицей. Ход обработки устанавливается через определенные утверждения - гипотезы - относительно признаков и/или объектов. Управление обработкой означает указания соответствия между совокупностью содержательных гипотез и наличных в применяемом статистическом пакете математических моделей, точнее, методов для их разработки. Различия в подходах разных пользователей к анализу возникают, с одной стороны, на базе разной способности идентификации составляемой математической модели, а с другой, тем, какую определяет исследователь структуру в совокупности содержательных гипотез.

Если набор гипотез об исследуемом объекте представляется в виде целостной системы, то есть основание говорить о предметной, содержательной модели исследования. Структура в совокупности гипотез может определяться 1) содержательно детерминированным порядком следования гипотез, 2) указанием решающих правил в альтернативных ситуациях, и др. способами, взаимосвязывающими гипотезы согласно теоретической базе предмета. Выбор математической модели означает формализацию предметной модели: фиксацию класса математических моделей, определение входных для соответствующего статистического метода параметров.

Структура предметной модели в ходе эмпирического анализа постепенно детализируется. Исходными для разных исследований бывают модели с разной степенью детализации, также различна степень формализации предметной модели. По этим двум основам выводим следующие типы подхода к данным (см. илл. 1).



Илл. 1.

I. Выдвигается конечный набор (перечень) гипотез $H_1, \ldots, H_K$. Предметная модель $\mathfrak{M} = (H_1, \ldots, H_K)$, на которой основывается анализ, в данном случае охватывает лишь часть системы содержательных гипотез, представляющей исследуемый объект. Решается узкая, отдельная проблема, исходная информация не используется в полной мере. Набором гипотез определяется, как правило, класс тривиальных статистических методик.

II. Предполагается равноправность всевозможных гипотез по наличным исходным данным, стратегия обработки ориентирована на генерацию предметной модели с помощью статистических методов. Класс математических моделей либо не задается, либо фиксируется, но формально, без содержательного обоснования, так как предметная модель отсутствует.

III. Пользователем прикладной статистики задается предметная модель исследования, $\mathfrak{M} = (H_1, H_2, \ldots)$, но он не способен точно указать формализованный эквивалент, применяемую

методику анализа.

IV. На основе предметной модели пользователь полностью определяет соответствующие математические модели, а также применяемые методы анализа.

Сущность статистического анализа после идентификации математической модели M состоит в статистической оценке параметров модели, в разработке эмпирической модели $\hat{M}$.

3.3. Выведенные типы подхода к построению статистического анализа отражают реально встречаемые способы обращения с данными, но, естественно, являются притом абстракциями. "Чистых" представителей одного или другого типа бывает относительно редко, чаще можно говорить об эволюции логики анализа. Поэтому трудно указать распределение реальных исследователей-прикладников по данной типологии, но можно дать характеристику тех, кто склонны к применению элементов конкретного типа. С точки зрения нашей проблемы - включенности социолога в анализ - укажем, прежде всего, на различия в деятельности во время анализа. Рассматриваемые подходы различаются по активности клиента-пользователя методикой вычислительного центра, притом упорядочены в порядке возрастания степени участия исследователя в анализе информационного массива.

В случае первого типа, где клиентом выделяется конкретный, конечный набор гипотез, он формально как будто бы полностью управляет обработкой, установив точный путь анализа. Но по сути дела в отношении к обработке как к процессу, он неактивен, выполняя свою роль - обдумывание запроса - до начала анализа.

К первому типу логики обработки склонны клиенты с консервативными в отношении к анализу взглядами, которые обращаются к ЭВМ, в первую очередь, как к средству, помогающему справиться с большими массивами данных. Ожидания насчет расширения познавательных возможностей социолога - второстепенные. Определяющую роль при выборе методов обработки играет имеющийся опыт или совет со стороны, пример, а не рассуждения о сущности данных и предмета. Новая методика принимается с недоверием. Профессиональная компетентность исследователей этой категории высокая, но для них существует и кажется непреодолимым разрыв между их предметом и математикой.

Клиенты первого типа непременно нуждаются в помощи со стороны консультанта-статистика, в сотрудничестве с которым на базе имеющейся высокой профессиональности клиента достигается хороший уровень обработки. Советы консультанта, в первую очередь, могут быть использованы в целях расширения диапазона методов анализа. Методы многомерной статистики, которые для этой группы пользователей относятся к менее привычным, при усвоении с помощью консультанта способствуют разработке системных представлений об объекте исследования. Наш опыт показывает, что именно путем обогащения методики можно находить общий язык с клиентом, направлять его на содержательную логику. Прямое указание на недостаточно полный анализ - а именно это главный результат ограниченности данного подхода -, остается, как правило, без отклика.

Клиенты первой группы - это часто отдельные исследователи, а не коллектив, исследовательская группа в целом. По времени, продолжительности анализа период обработки сравни-

тельно короткий, но часты "рецидивы" - постановка дополнительной задачи после длительной паузы, что предъявляет специальные требования к организации инфомассивов.

Клиенты второй категории, предпочитающие путь генерации гипотез, предусматривают систематический, формально полный просмотр всего материала, и, подобно первому типу, могут не вмешиваться в процесс обработки. Но в реальной обстановке они все-же вынуждены это делать, причем их участие осуществляется, скорее всего, в ограничивании объема анализа. Это происходит несистематично, а не на основе определенной целостной картины.

Среднему представителю этой категории типично ограничиваться анализом пар признаков, выбирая, например, пары для более детальной характеристики (совместное распределение и т.п.) по формальным критериям статистической связи. В такой ситуации регулятивная роль пользователя формальная, ее может выполнить и машина, так как решения не сопровождаются рассуждениями о сущности исследуемого явления.

Управляющая роль клиентов этой категории проявляется еще в том, что они склонны точно указать применяемый метод анализа, упорно настаивая на необходимости в конкретной, найденной ими технологии. Чаще такие требования возникают у лиц, недостаточно компетентных в технике анализа, которыми выбор методов мотивируется опять-таки посторонними причинами (мода, удача коллеги и т.п.). В предпочитании определенной технологии обработки обычно еще сильно преувеличиваются и ее познавательные возможности, фетишируется весь формальный анализ.

Представители второй категории, в отличие от первой, не обладают высокой профессиональностью. Мотивом обращения в центр обработки часто является абстрактное желание применять машинные методы. Как правило, это отдельные исследователи или лица, принимающие на себя роль лидера-математика в своей исследовательской группе. Часто они работают без консультанта, отказываясь от консультаций. Если все-же консультант ознакомится с работой, то повлиять на ее ход весьма трудно. Самооценка знаний у клиентов этой категории высокая, трудно объяснить несостоятельность стратегии. Часто возникают проблемы организационного порядка, так как объем обработки превышает реальные границы по всем параметрам.

Представителям данной категории характерно, что исследователь плохо "знает" свой информационный массив. Более того, набор данных часто носит случайный характер, ибо в основе сбора данных не лежит определенная система гипотез.

Если искать путей влияния на логику клиента этой категории, то можно опираться на анализ характера отдельных признаков, на структуризации совокупности показателей. При первом подходе роль консультанта мы видели, скорее всего, в направлении исследователя к рассмотрению совокупности признаков в целом, в преодолении изолированности отдельных предположений. В данном случае, наоборот, исследователя придется направить в сторону анализа содержательных деталей, показателей в отдельности, в специфике их информационных возможностей как со стороны статистики, так и предметной теории.

При третьей и четвертой категориях выделенных нами пользователей статистической методикой, где в стратегии обработ-

ки соединяются как теоретические рассуждения, так и решения по формальным критериям на основе уже выполненных шагов обработки, клиент сам определяет сущность и форму анализа. Он фиксирует порядок следования разработки отдельных блоков признаков, определяет действие в ситуациях с несколькими выходами, выбирает критерий решения и т.д.

Такая последовательная, поэтапная стратегия обработки пользователям машинного анализа хорошо известна и применяется во многих социологических исследованиях. Мы не намерены в данном рассмотрении оправдать этот, уже усвоенный социологами принцип обработки, а намерены подчеркнуть то, что единый, целостный подход должен быть основой в течение всей обработки, равномерно охватив всю исходную информацию, сложив в систему все показатели. Такое, казалось бы, крайне естественное требование на практике нередко игнорируется, что и приводит к тому, что процедуры машинного анализа выходят из-под контроля социолога, станут сами управлять его рассуждениями и логикой. Мы умышленно ввели для обозначения содержательной логики исследователя понятие предметной модели, чтобы подчеркнуть автономность, самостоятельность содержательной части, в противовес строгой математической форме заключений и рассуждений при машинном анализе. Предметная модель - это форма, помогающая исследователю ориентироваться во всей обилии информации, как исходной так и выводимой, это основа управления обработкой.

Что касается характерных свойств клиентов третьей и четвертой групп, то в них много сходного. Представители обоих типов отличаются высокой профессиональностью, часто они чле-

ны исследовательского коллектива, определенной исследовательской школы. Нередко исследование входит в более обширный комплекс исследований.

Разница между третьим и четвертым типами выявляется в различной способности идентификации соответствующего предметной модели математического эквивалента. На практике эта разница позволяет различать тех, кто не определяет применяемые методы, и тех, кем точно указывается путь анализа при данном пакете программ. Среди последних нередки случаи неудачного выбора. Таким образом, исследователи третьей категории при разработке стратегии анализа нуждаются в помощи для ориентации в разнообразии методов, а четвертой группы - для корректировки разработанного клиентом стратегии. Успешность работы третьей группы не ниже выделенной нами последней, казалось бы идеальной категории. Главное для социолога - это система теоретических знаний об исследуемом объекте, недостатки в остальных компонентах машинного анализа можно восполнить.

По словам Дж. Тьюки ([5], стр. 19): "Приложения математики всегда осложнены тем, что сущностью предмета исследования надо овладеть столь же хорошо, как и применяемой математикой." По нашему убеждению, эти осложнения могут быть успешно преодолены именно сотрудничеством, а не путем, например, профессиональной переспециализации математика в область социологии или наоборот. Уникальное социальное отношение - сотрудничество математика и специалиста соответствующей области исследования вызывает ряд серьезных проблем методики с

обоих сторон. Типология подходов социолога к построению анализа нами была введена именно с целью указания тех аспектов обработки, в отношении к которым можно найти общий язык между социологом и математиком. В [9] приводится некоторое обобщение принципов организации статистической обработки в массовом порядке, которые введены для сотрудничества с клиентами на базе ВЦ ТГУ. Отправным пунктом при их разработке послужило утверждение о наличии разной логики анализа, порождающей разную, часто недостаточную включенность клиента в ход анализа данных.

Литература

I. Айвазян С.А., Енюков И.С., Мешалкин Л.Д., Прикладная статистика. М., 1983.
2. Доходы и потребление семьи. Проблемы методологии изучения и моделирования. Мат. всес. конф., Ереван, 1983.
3. Комплексное применение математических методов в социологическом исследовании. М., 1983.
4. Молодежь в общественных отношениях развитого социализма. Мат. конф., Тарту, 1984.
5. Мостеллер Ф., Тьюки Дж., Анализ данных и регрессия. Вып. I, М., 1982.
6. Планирование и руководство коммунистическим воспитанием в высшем учебном заведении. Мат. респ. научн. конф., Рига, 1983.
7. Тоодинг Л.М., Описание структуры реального потока задач анализа данных. Труды ВЦ ТГУ, 1981, 48, 50-66.
8. Тоодинг Л.М., О результативности статистического анализа данных. Труды ВЦ ТГУ, 1982, 49, 96-111.
9. Тоодинг Л.М., Практические аспекты статистического анализа данных. Наст. сб., 67-82.

ПРАКТИЧЕСКИЕ АСПЕКТЫ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

Л.М. Тоодинг

1. Использование стандартного статистического пакета - реальная, широко доступная на практике альтернатива к применению статистического метода на ЭВМ, специально созданного для исследуемой проблемы. Однако, рутинный характер решения задач разной предметной сущности нередко приводит к ограничительной унификации методики исследования. Главным источником рутины в отрицательном смысле является стереотипность подхода исследователя к формализации содержательной проблемы. Факторами, определяющими выбор математических моделей, иногда являются опыт, традиция, совет со стороны и т.п., а не сущность изучаемого явления и измеренных параметров. С другой стороны, недостаточно полное знание возможностей наличного статистического пакета или статистической методики в целом часто приводит к удовлетворенности с тривиальным анализом. На практике эти аспекты характеризуются мерой включенности исследователя в ход анализа и отражаются в управляющей роли специалиста конкретной предметной области в построении стратегии машинной обработки. Но и в том идеальном случае, когда исследователь полностью способен определить нужный путь анализа, его управляющая функция реализуется

лишь тогда, когда в вычислительном центре созданы для этого необходимые условия. Гибкость стандартного подхода существенно повышается, если машинные методы прикладной статистики окажутся доступными в форме определенной службы обработки информации.

В данной статье рассматриваются некоторые положения практической организации статистического анализа при помощи стандартного программного обеспечения. В основу положен практический опыт службы обработки данных, накопленный в вычислительном центре Тартуского государственного университета в течение более чем 10 лет. Средний объем обработки за год составляет около 110-120 проблем-задач. Представлены основные традиционные области эмпирического анализа: медицина, биология, социология, экономика и управление, психология. К обобщению практического опыта присоединяются эмпирические выводы анализа реального потока задач статистического анализа [3] и результаты анкетного опроса пользователей статистическими пакетами [4].

2. В общении пользователя стандартным программным обеспечением по статистике (т.н. клиента) с представителями вычислительного центра условно можно выделить следующие этапы:

1) Формализация, ввод и редактирование данных.

2) Построение общей стратегии обработки, выбор класса математических моделей.

3) Фиксация задания для машинного анализа.

4) Интерпретация статистического вывода.

Данный перечень не отражает поочередность действий, так как

стратегия обработки в ходе анализа постоянно детализируется и уточняется, также могут быть введены дополнительные данные.

Действие клиента и обслуживающих пакета лиц имеет на разных этапах разную значимость, тяжесть управления анализом распределяется по-разному. Ниже рассматривается подробнее, кем реально выполняется конкретный шаг, какая помощь оказывается клиенту со стороны вычислительного центра и что ожидает клиент.

Общения с клиентами опирается на следующие положения.

1) Наличие у клиента при обращении в ВЦ содержательно полного набора данных и системы концептуальных предположений об исследуемом объекте. Как было указано в [5], в реальном потоке задач набор априорных содержательных гипотез характеризуется разнообразной степенью структуризации. Этим порождаются разные типы подхода к машинному анализу.

2) Правильность исходных данных гарантируется клиентом, он сам отвечает за качество входной информации.

3) На всех этапах обработки выделяются три уровня действия: а) содержательная, предметная сторона - выполняется клиентом; б) статистическая сторона, проблемы методики - выполняется совместно с консультантом (специалист по прикладной статистике) и клиентом; в) доступ к пакету - осуществляется администратором пакета (математик-программист).

Специализация функций обслуживания на практике оправдалась. На данном этапе развития эмпирических наук она предположительно, наилучшим образом отвечает уровню подготовки клиента, а также методологическим возможностям современных

пакетов массового использования. При нашем объеме обработки работают 2-3 администратора и 1-2 консультанта.

3. Клиенты, как правило, имеют различную установку в отношении к машинному статистическому анализу. По нашему общему положению различия в подходах следует учитывать и в сотрудничестве с клиентами. Поэтому существенно иметь представление о проблемах машинной обработки, учитывающее взгляды клиентов.

В ходе опроса выяснилось, что по значимости разные этапы анализа не считаются равноценными. Об этом свидетельствуют, хотя и косвенно, данные о том, в какой мере клиенты заинтересованы в просмотре разных аспектов технологии обработки в методологических пособиях по пакету (табл. I).

Таблица I.

	существенно %	может быть %	менее существенно %
Подготовка данных			
формализация	60	35	5
ввод	53	30	I7
редактирование	60	34	6
Построение стратегии обработки			
элементарный курс статистики	37	35	28
теоретические проблемы метода	22	25	53
вопросы технической реализации	I3	38	49
Интерпретация			
описание структуры выпечаток	65	25	I0
проблемы осмысления	88	9	3

Распределение оценок несомненно указывает на "потребительскую" установку клиента. Проблемы метода вызывают наименьший интерес, в то же время подчеркивается необходимость в помощи при осмыслении статистического вывода.

4. Первый этап формализованного анализа - подготовка данных - является одним из важнейших периодов в процессе обработки. Имеющиеся в исходных данных недостатки позднее трудно восполнить. Этап формирования массива данных содержит следующие моменты[1]:

1) Окончательная формализация информации с учетом требований используемых пакетов - выполняется клиентом.

2) Запись данных на инфоноситель ввода (перфорация) - клиент.

3) Ввод данных в ЭВМ - выполняется администратором пакета.

4) Анализ протокола ввода и составление корректуры для исправления ошибок - клиент.

5) Редактирование данных и оформление файла обработки - администратор.

6) Первичный анализ данных, составление эмпирических распределений - администратор.

Шаг 1, как правило, сопровождается консультацией статистика-прикладника, так как при этом фиксируется статистический характер данных. Шаги 2 и 4 консультирует администратор.

---

[1] При работе с дисплеем этап ввода имеет в некоторых аспектах иной характер.

Для ввода данных нами в большинстве случаев используется перфолента, но допускается ввод и магнитной ленты и перфокарт. Данные часто перфорируются вне вычислительного центра. Этап ввода унифицирован в отношении ко всем наличным в ВЦ пакетам. Генерируемый исходный массив можно обрабатывать при помощи любого из них.

По каким соображениям полная ответственность за правильность данных возложена на пользователя? Подготовка данных действительно содержит немало чисто технической работы, которую может эффективно выполнить посторонний по отношению к исследованию сотрудник ВЦ. На практике многих центров такой порядок и используется. Однако, на этапе ввода можно выделить моменты, где решающим является все-же слово самого исследователя. Примером, такой ситуации служит анализ выбросов, если при вводе происходит сглаживание исходных данных. Выбросы - часто это не ошибки, а информация, притом неожидаемая. Определение критерия сглаживания также основывается на содержательных свойствах исследуемого явления.

Вторая причина психологического порядка. Недостаточным включением на этапе ввода клиента в ход анализа нередко закладывается основа для своеобразного "отчуждения" исследователя от своей задачи. Самые простые приемы формализации (например, цифровое кодирование), выполненные без участия исследователя, могут ему казаться "математикой", порождать еще до начала анализа психологический барьер между предметом исследования и машинным анализом.

Перечислим в заключение возможности практической помощи, предлагаемые клиенту в течение подготовки данных.

Сперва проводится консультация по выбору схемы кодирования и фиксируется степень "ручной" кодировки. Дается руководство для перфорации и проводится обучение по применению перфоратора или другого устройства записи, также обеспечивают нужной техникой. Предлагается руководство по редактированию информации и составлению описания данных. Проводится формальный контроль результатов первичного анализа, устанавливается степень технического качества данных.

5. Этапы анализа следующие подготовке данных - построение стратегии обработки и интерпретация - тесно связаны с проблемами консультирования. Существуют разные мнения о результативности консультаций, которые среди статистиков иногда доходят и к экстремальному утверждению, что эта форма сотрудничества уже принципиально малоэффективна. Однако, повседневная нагрузка консультаций, а также данные опроса клиентов показывают, что имеется реальная потребность в консультационной службе.

Построение каждой консультации специфическое и зависит не только от сущности предмета, но и от уровня подготовки клиента. Отправным пунктом является то, насколько глубоко разработана клиентом структура набора исходных содержательных гипотез - либо это случайно собранные отдельные предположения, либо система гипотез, образует содержательную модель.

Главным правилом целесообразно взять простой принцип: нельзя предлагать клиенту методику, в надобности которой его не удается полностью убедить. Если у клиента нет никаких со-

ображений насчет применяемых математических моделей, то ему на основе его собственного эмпирического материала нужно продемонстрировать несколько решений, давая клиенту и в этом случае возможность участия в выборе стратегии. Если допускается малейший пробел в обосновании со стороны клиента плана обработки, если появляется технологический нажим, то этим уже создана возможность, что клиент потеряет контроль над задачей.

Разные области применения прикладной статистики отличаются разным уровнем традиций статистического анализа. Метод, хорошо усвоенный в одной области исследования, может в другой области иметь значение методологической оригинальности. Наша практика показывает, что роль новатора в таком смысле успешнее выполняют те клиенты, у которых выше профессиональная компетентность, а не те, кто сравнительно больше осведомлены о возможностях статистического анализа. Судя по результатам опроса клиентов, отношение к применению нового метода благосклонное, но осторожное. В альтернативной ситуации выбора между известным и новым методом, равноценными в смысле интерпретации, клиенты решали бы так:

использовать известный метод - 25%,
оба метода - 71%,
новый метод - 4%.

Адаптация новых статистических методик является перспективной формой сотрудничества статистика-консультанта с клиентом.

В связи с этим укажем на одну естественную особенность работы со стандартными статистическими методами. На практике

редко встречаются задачи, где первое решение дает хороший ответ. Часто один и тот же метод при одних и тех же исходных данных следует применять повторно. Методика, специально разработанная для решения данной проблемы, оптимизирована уже при разработке в отношении к конкретной задаче. При стандартных методах адаптация метода к задаче происходит во время решения задачи путем вариации параметров алгоритма. Нередко клиентами игнорируется вта особенность, анализ останавливается на полпути, возможности метода и информативность данных не исчерпываются. Именно в таких ситуациях имеется острая потребность в консультации, притом клиент сам может этого и не понимать.

Каков практический порядок проведения консультаций?

Как правило, каждый клиент имеет консультанта, которым визируются все содержательно новые шаги в анализе. Притом клиент обязательно познакомит консультанта с теми результатами уже проведенного анализа, которые нужны для разработки дальнейшей стратегии. Объем консультаций варьируется в широких пределах: опрошенные нами клиенты указали сотрудничество с консультантом ВЦ в объеме от 0.5 до 80 часов, в среднем на каадого клиента потратили 8.8 ($\pm$2.6) часов.

В деятельности консультанта можно выделить три типа тесно связанных функций.

5.1. Во-первых, это выявление общей линии задачи, требующее 1-2 вступительных консультаций. Именно при этом возникает с полной остротой проблема, до какой степени, до какой глубины должен вникнуть консультант в содержательную сторону задачи. Нами выбран путь избежания крайностей - профессио-

нальной переспециализации статистика в предметную область и наоборот - исследователя в область прикладной статистики. Также неуспешным оказывается противоположное, когда статистик останется на чисто формальном уровне, а клиент отказывается от вмешательства в "математику". На практике степень сотрудничества варьируется от соавторства в прямом смысле слова до ответов консультанта на конкретные вопросы клиента.

Что касается ожиданий клиента в отношении к роли консультанта, то абсолютно все опрошенные клиенты считают нужной помощь со стороны консультанта центра обработки информации. В отношении к степени углубления консультанта в содержательную сторону исследования были высказаны следующие мнения:

консультанту следовало бы ознакомиться с содержанием бегло, на уровне "обыденного" создания - 18%;

в отдельных частях необходимо более глубокое изучение содержания - 56%;

ожидается основательное знакомство с содержанием - 26%.

Итак, вышеуказанная линия воздержаться в сотрудничестве от крайностей специализации, согласуется с ожиданиями клиентов.

5.2. Вторым моментом деятельности консультанта является помощь при фиксации для ЭВМ запроса решения (заказа). Придерживаясь опять же распределения действий на содержательный, статистический и машинный уровни, нами используется следующая практика.

Окончательная формализация заказа проводится администратором пакета, который "переводит" на входный язык конкретного пакета составленное либо клиентом, либо при надобности консультантом полуформализованное задание. Задание в полу-

формализованном виде означает указание требуемого программного модуля, обязательной для него входной информации, а также значений параметров управления. Для общения консультанта с администратором на практике сложилась определенная терминология (жаргон), использование которой избавит консультанта от знания технических деталей входного языка разных пакетов, а администратора - от содержательной стороны задачи. Клиент, уже имеющий достаточный опыт работы в вычислительном центре, успешно справляется с необходимой для идентификации метода терминологией.

Итак, наша практика общения клиента с ЭВМ не отличается высоким уровнем формализации. В ходе опроса клиентов нас интересовал желаемый клиентами уровень общения. Распределение оценок по трем предложенным нами разным уровням формализации свидетельствует о стремлении оставаться в рамках терминологии предмета при постановке машинного задания (см. рис. 1).

Так как в ближайшем будущем станут широко доступными микро-ЭВМ, то это распределение заслуживает внимания. Возражение против формализованного интерфейса вызывает надобность хорошо изучить, при каких условиях персональная машина является наиболее эффективной для статистического анализа. Главная выгода получается, наверное, от того, что клиент берет на себя ряд технических функций, связанных с эксплуатацией пакета. В ускоренном темпе усваивается и применяемый метод, так как клиент может свободно "играть" с ним. Заодно возникает серьезный вопрос, как обеспечить корректность приложения статистического метода, так как клиент сам часто не замечает опасности злоупотребления методикой. Нередко он бес-

силен справляться с техническими проблемами машинного анализа. Поэтому можно предполагать, что и при широком непосредственном доступе к ЭВМ останутся актуальными задания, выделенные нами для консультанта.

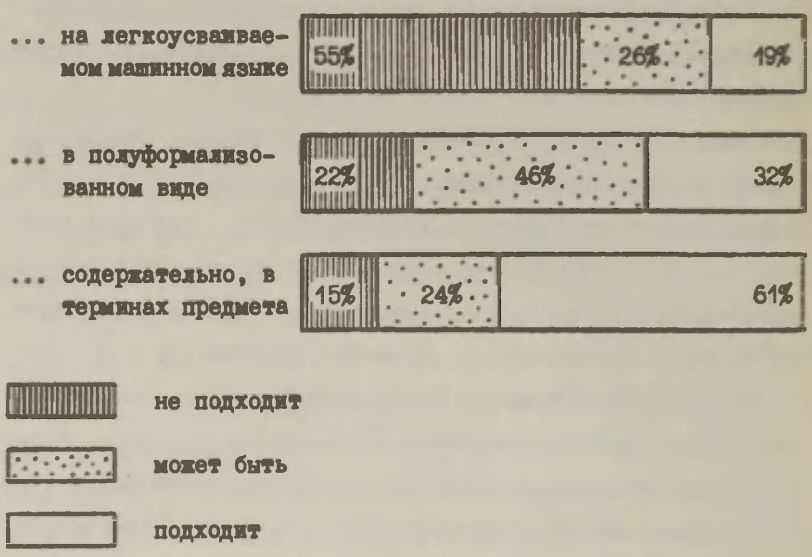


Рис. 1.

Естественно, что некоторые из задач в сотрудничестве статистика и клиента могла бы взять на себя ЭВМ, но пока отсутствует соответствующее программное обеспечение. К таким задачам относятся выбор пакета, если план обработки в общих чертах зафиксирован, указание значимости результатов, чтобы облегчить содержательную структуризацию машинного вывода. Также необходимо методику превращать в большей степени само-

объясняющей, увеличить объем информации по пакету, выводимой во время сеанса решения.

5.3. Третьим типом действий консультанта является помощь при интерпретации, самый сложный и своеобразный из всех актов общения с клиентом. В интерпретации можно выделить два более характерных уровня осмысления машинного вывода.

Во-первых, адаптация, конкретизация статистического понятия в контексте предмета исследования (например, не среднее значение, а средняя трудовая активность данного цеха, средняя успеваемость класса). Второй уровень - интерпретацией выявляются содержательные закономерности, дается их обоснование, объясняется смысл исследуемого (измеренного) показателя. Хотя реально эти два аспекта и тесно совмещены, их выделение в нашем рассмотрении является целесообразным, так как различает принципиально разные роли консультанта. Консультант может успешно содействовать при адаптации статистической модели в конкретном предмете исследования. Но занимаясь содержательным обоснованием статистического вывода, осмыслением связей, факторов влияния, различий и т.п., консультант, по сути дела, становится соавтором. Отношение соавторства, однако, требует от статистика-прикладника других качеств,чем роль консультанта пакета. Иногда это забывается как консультантом так и клиентом.

Вопросы интерпретации статистического вывода заслуживают постоянного внимания со стороны специалистов разных областей исследования. В качестве примера укажем на концептуальный рассмотр проблем интерпретации в эмпирической социологии [1]. Эти идеи могут быть приложены и к иным областям.

6. Остановимся в заключение на некоторых общих характеристиках задачи анализа данных.

Процесс обработки формально характеризуется, в основном, двумя параметрами: длительностью обработки инфомассива и частотой, интенсивностью обращения клиента в течение обработки в вычислительный центр. Нами придерживается практика, где эти характеристики определяются желанием и возможностями клиента, стилем его работы. По длительности обработка значительно варьируется от нескольких недель до несколько лет. Отметим, что по данным опроса желаемый клиентами период обработки короче реального (в среднем более чем 2 раза).

Длительность периода обработки частично определена тем, с какой частотой клиент обращается в вычислительный центр, какие паузы понадобятся ему для интерпретации результатов. По данным опроса ожидаются разнообразные ритмы работы:

короткая, интенсивная обработка - 19%;

начало более интенсивное - 22%;

середина более интенсивная - 6%;

конец более интенсивный - 4%;

равномерная, регулярная обработка - 49%.

Итак, желаемая клиентами стратегия обработки предусматривает примерно в трети случаях неравномерный режим работы, что следует учитывать при распределении машинного времени и консультационной нагрузки.

Длительный период обработки и варьирующая интенсивность анализа требуют создания архива информационных массивов. Архив статистики в ВЦ ТГУ сохраняется на магнитных лентах. Решение для включения массива в архив принимается согласно

указанию клиента или экспертов соответствующей предметной области (такой порядок установлен для социологических работ). Архив регулярно дополняется и очищается. При надобности массив данных актуализируется в рабочий архив пакетов и обработки продолжается.

В настоящее время, как правило, в каждом вычислительном центре для статистической обработки имеется не один, а несколько пакетов, которые по возможностям анализа частично покрываются. Поэтому при каждой задаче возникает проблема выбора пакета, а также необходимость совмещения нескольких пакетов. Один из современных направлений ремения этих трудностей состоит в разработке интегрирующей разные пакеты системы, но пока практических реализаций мало. Автоматизация управления несколькими пакетами является, заодно, ключевой проблемой организации процесса обработки, перераспределяющей общие соотношения "машинного" и человеческого факторов при построении обработки. Проводимая нами разработка системы интегрирования пакетов [2] нацелена, в первую очередь, на достижение статистической корректности анализа. Присоединение к этому методологически хорошо обоснованной практической организации статистического анализа данных создается реальная база для повышения результативности эмпирических исследований.

Литература

I. Батыгин Г.С., Обоснование интерпретационных схем. Социологические исследования, 2, 1984, 22-34.

2. Лепинг В., Система интеграции пакетов. Наст. сб., 35-49.

3. Тоодинг Л.М., Описание структуры реального потока задач анализа данных. Труды ВЦ ТГУ, 1981, 48, 50-66.

4. Тоодинг Л.М., О результативности статистического анализа данных. Труды ВЦ ТГУ, 1982, 49, 96-111.

5. Тоодинг Л.М., Применение прикладной статистики в эмпирической социологии. Наст. сб., 50-66.

ВЫРАБОТКА РАВНОМЕРНО РАСПРЕДЕЛЕННЫХ СЛУЧАЙНЫХ ЧИСЕЛ ДЛЯ ЭВМ ЕС-1022

Ю. Вилисмяэ

В данной работе мы рассматриваем созданный в вычислительном центре ТГУ датчик (генератор) случайных чисел, равномерно распределенных на отрезке $[0,1]$, и попробуем оценить качество случайных чисел, генерированных этим датчиком.

1. Датчик псевдослучайных чисел

Для ЭВМ ЕС-1022 реализован метод генерирования равномерно распределенных псевдослучайных чисел, изложенный в книге Д.Кнута [3].

Пусть нам задано какое-нибудь целое число $x_0 \in (0,2^{31})$, удовлетворяющее условию

$$x_0 \bmod 4 = 2 . \qquad (1)$$

Это число считается первым случайным числом в искомой последовательности случайных чисел и, исходя из него, следующие числа в последовательности вычисляются по формуле

$$x_{n+1} = x_n(x_n + 1) \bmod 2^{31} \quad (n=0,1,2,\dots). \qquad (2)$$

Таким образом, получается последовательность псевдослучайных (целых) чисел $x_0, x_1, x_2, \dots$, равномерно распределенных на отрезке $[0,2^{31}]$. Отметим, что в этом случае последо-

вательность $\left\{\frac{x_n}{2^{31}}\right\}$ $(n=0,1,2,\ldots)$ равномерно распределена на отрезке $[0,1]$.

В целях увеличения случайности генерируемой последовательности каждое создание новой последовательности начинается со случайного выбора начального значения x_0 при помощи машинных часов (учитывая, разумеется, условие (1)).

2. Статистическая проверка качества равномерно распределенных случайных чисел

Полученные программным способом псевдослучайные числа нуждаются в детальной статистической проверке, чтобы убедиться в их случайности и равномерности распределения. Для этого создано много статистических тестов. Ниже будут рассматриваться тесты, при помощи которых в вычислительном центре ТГУ были проведены анализы случайных чисел.

2.1. Проверка частот (критерий χ^2).

Пусть у нас генерирована последовательность длиной N случайных чисел: $x_1,\ldots,x_N$ $(x_i \in [0,1])$. Разделим отрезок $[0,1]$ на m частей (классов) равной длиной

$$A_1 = [0,\tfrac{1}{m}),\ A_2 = [\tfrac{1}{m},\tfrac{2}{m}),\ \ldots\ ,\ A_m = [\tfrac{m-1}{m},1).$$

Обозначим через ν_i количество случайных чисел из рассматриваемой последовательности, оказавших в классе A_i. Если предположить, что рассматриваемые случайные числа распределены равномерно на отрезке $[0,1]$, то в каждый класс A_i в среднем должно попасть N/m случайных чисел. В этом случае статистика

$$\chi^2_N = \frac{m}{N}\sum_{i=1}^{m}(\nu_i - N/m)^2 \qquad (3)$$

имеет χ^2-распределение с m-1 степенями свободы.

Учитывая вышесказанное, проверка частот проводится следующим образом.

1. Выбирается уровень значимости α (например, $\alpha = 0.05$).

2. Генерируется последовательность случайных чисел длиной N.

3. При заданном числе классов m находят эмпирические частоты ν_i (i=1,...,m) и вычисляется величина χ^2_N по формуле (3).

4. Если вычисленная нами величина χ^2_N принадлежит к интервалу $[\chi^2(\alpha,m-1),\ \chi^2(1-\alpha,m-1)]$, где $\chi^2(p,m-1)$ - p-квантиль χ^2-распределения с m-1 степенями свободы, то можно считать вероятным, что случайные числа распределены равномерно на отрезке [0,1].

Если $\chi^2_N > \chi^2(1-\alpha,m-1)$, то эмпирические частоты (ν_i) сильно отличаются от теоретических ($\frac{N}{m}$), и поэтому нельзя сказать, что случайные числа распределены равномерно. Если $\chi^2_N < \chi^2(\alpha,m-1)$, то эмпирические и теоретические частоты отличаются незначительно, но, поскольку вероятность появления такого события меньше чем α, то следует сомневаться в случайности рассматриваемой последовательности.

Пример 1. Мы провели проверку частот при различных последовательностях разной длиной (при m = 64). Результаты изложены в таблице 1.

Таблица 1.

N	500	2000	10000	50000	100000	150000
χ^2_N	56.80	53.38	56.63 74.92	81.04 79.58	95.19	82.60

Поскольку теоретические квантилы

$$\chi^2(0.05;63) = 45.74, \qquad \chi^2(0.95;63) = 82.53;$$
$$\chi^2(0.01;63) = 39.86, \qquad \chi^2(0.99;63) = 92.01,$$

то при $N < 100000$ наш генератор удовлетворяет критерию χ^2. При $N \geq 100000$ эмпирические значения χ^2_N слишком большие. Для выяснения степени случайности для этого события мы спользуемся двукратным критерием χ^2, который состоит в следующем.

2.2. Проверка частот при помощи двукратного критерия χ^2

Фиксируется длина N последовательности случайных чисел и число классов m, на которое при проверке частот разделяется отрезок [0,1]. После этого генерируются k последовательностей случайных чисел длиной N, причем каждый раз выбирается новое начальное значение. При этом каждый раз вычисляется величина $\chi^2_N(i)$ $(i=1,\ldots,k)$ по формуле (3).

Если предположить, что генерированные случайные числа распределены равномерно на отрезке [0,1], то числа $\chi^2_N(i)$ $(i=1,\ldots,k)$ представляют собой конкретные значения случайной величины с χ^2-распределением с m-1 степенями свободы. Соответствие между распределениями будет исследоваться снова при помощи критерия χ^2.

Для этого мы выбираем подходящее число s $(s < k/5)$ и разделим множество значений χ^2-статистики на s классов. Затем установим эмпирические частоты ν_i $(i=1,\ldots,s)$ для каждого класса и вычисляем величину

$$\chi^2 = \sum_{i=1}^{s} (\nu_i - kp_i)^2 / kp_i , \qquad (4)$$

где p_i - соответствующие теоретические вероятности для клас-

сов, и сравниваем ее с соответствующими квантилями χ^2-распределения с s-1 степенями свободы.

Пример 2. Генерировалось 120 последовательностей, в каждой 10000 случайных чисел. Каждый раз была проведена проверка частот (при $m = 64$) и вычислена величина $\chi^2_N(i)$. Пользуясь формулой (4) при $s = 20$, мы получили $\chi^2 = 13.05$. Так как имеет место неравенство $\chi^2(0.1;19) < 13.05 < \chi^2(0.25;19)$, то можно сказать, что наш генератор удовлетворяет критерию χ^2.

2.3. Проверка пар

Как известно, двоичное представление любого целого числа состоит из конечной последовательности нулей и единиц. При случайных числах, равномерно распределенных на отрезке [0,1], вероятность появления нуля (или единицы) на каждом разряде двоичного представления этих чисел равна 0.5 . Проверкой пар устанавливается, подчиняются ли рассматриваемые случайные числа этому закону.

При проверке пар все генерированные N целых случайных чисел представляются в двоичной системе. Затем фиксируется какой-то двоичный разряд и устанавливается число нулей (или единиц) на этом разряде во всех случайных числах. Математическое ожидание этого числа равно N/2, если предположить, что случайные числа распределены равномерно на отрезке [0,1]. Теперь для каждого двоичного разряда вычисляется χ^2-статистика

$$\chi^2_N(i) = \frac{4}{N}(n_i - N/2)^2,$$

где n_i - количество нулей на i-ом разряде, и они сравниваются с соответствующими квантилями χ^2-распределения с одной степенью свободы.

Пример 3. Результаты проверки пар для последовательности, содержащей 150000 случайных чисел, изложены в таблице 2.

Таблица 2.

разряд i	число нулей на этом разряде n_i	разность $n_i - \frac{N}{2}$	хи-квадрат
1	75184	184	0.9028
2	74827	-173	0.7981
3	74864	-136	0.4932
4	74990	- 10	0.0027
5	74983	- 17	0.0077
6	74780	-220	1.2907
7	75112	112	0.3345
8	75064	64	0.1092
9	75016	16	0.0068
10	74970	- 30	0.0240
11	75116	116	0.3588
12	75026	26	0.0180
13	74987	- 13	0.0045
14	75003	3	0.0002
15	74988	- 12	0.0038
16	74986	- 14	0.0052
17	74997	- 3	0.0002
18	74980	- 20	0.0107
19	75002	2	0.0001
20	74995	- 5	0.0007
21	75002	2	0.0001
22	74998	- 2	0.0001
23	75002	2	0.0001
24	75004	4	0.0004
25	75003	3	0.0002
26	75000	0	0
27	75000	0	0
28	75000	0	0

Надо заметить, что на низких двоичных разрядах числа нулей и единиц практически одинаковые. Этот слишком хороший результат показывает, что низкие разряды менее случайные чем высшие.

Теоретические квантилы χ^2-распределения:

$\chi^2(0.01;1) = 0.0002$, $\chi^2(0.05;1) = 0.0039$,

$\chi^2(0.5;1) = 0.455$, $\chi^2(0.95;1) = 3.84$,

$\chi^2(0.99;1) = 6.62$.

2.4. Проверка комбинаций

Проверка комбинаций основана на установлении числа нулей в двоичном представлении случайного числа. Тест осуществляется следующим образом.

Генерируется последовательность длиной N случайных чисел (рассматриваются целые числа). Затем случайные числа разделяются на классы: в один класс берутся те случайные числа, в двоичном представлении которых (на l высших разрядах) одинаковое число нулей (или единиц). Число разных классов l+1. Таким образом, для каждого класса устанавливаются эмпирические частоты ν_i. Соответствующие теоретические вероятности p_i выводятся из формулы

$$p_i = C_l^{i-1} 0.5^l \quad (i=1,\ldots,l+1),$$

поскольку число нулей в l-разрядном двоичном представлении случайных чисел, равномерно распределенных на отрезке [0,1], имеет биномиальное распределение B(l,0.5).

Для сравнения эмпирического и теоретического распределения следует спользоваться критерием χ^2. Для этого вычисляется величина χ^2, пользуясь формулой (4) при $s = l+1$ и $k = N$,

которая сравнивается с соответствующими квантилями χ^2-распределения с l степенями свободы.

Примечание. Чтобы корректно использовать критерий χ^2, надо учитывать, что теоретические частоты Np_i для каждого класса должны быть больше 5. Для достижения этого при классификации надо соединить те классы, для которых теоретические вероятности p_i маленькие.

Пример 4. Было генерировано 150000 случайных чисел и проведено для них проверка комбинаций (при $l = 28$). При классификации случайных чисел были соединены классы 1-5, 6-7, 23-24, 25-29, так осталось 19 разных классов. Результаты теста изложены в таблице 3. При помощи формулы (4) (при $s=19$ и $k=150000$) было получено $\chi^2=17.75$. Поскольку имеет место неравенство

$$\chi^2(0.5;18) < 17.75 < \chi^2(0.75;18),$$

то отсюда следует, что наш генератор удовлетворяет проверке комбинаций.

Чтобы повысить эффективность теста комбинаций, проведем двукратный контроль как и при проверке частот. Для этого генерируем k последовательностей случайных чисел длиной N и каждый раз вычисляем величины $\chi^2_N(i)$ $(i=1,\ldots,k)$ по формуле (4) (при $s=l+1$ и $k=N$). Затем, при помощи критерия χ^2, устанавливаем, являются ли числа $\chi^2_N(i)$ $(i=1,\ldots,k)$ конкретными значениями случайной величины с χ^2-распределением с l степенями свободы.

Таблица 3.

номер класса	число нулей в случайном числе	эмпирические частоты	теоретические частоты
1	0-4	12	13.50
2	5-6	265	265.44
3	7	666	661.63
4	8	1676	1736.79
5	9	3938	3859.53
6	10	7274	7333.11
7	11	11985	11999.63
8	12	16880	16999.48
9	13	21022	20922.44
10	14	22438	22416.90
11	15	20634	20922.44
12	16	17122	16999.48
13	17	12031	11999.63
14	18	7550	7333.11
15	19	3860	3859.53
16	20	1708	1736.79
17	21	662	661.63
18	22-23	265	265.44
19	24-28	12	13.50

Пример 5. Было генерировано 120 последовательностей, в каждой 10000 случайных чисел. Каждый раз была проведена проверка комбинаций. Значения статистики χ^2 разбивались на 17 классов. Из формулы (4) было получено итоговое значение $\chi^2=17.38$. Имеет место неравенство

$$\chi^2(0.5;16) < 17.38 < \chi^2(0.75;16) .$$

2.5. Критерий Колмогорова

При помощи критерия Колмогорова можно оценить различие между эмпирической и теоретической функциями распределения. Критерий основан на распределение величины

$$D_N = \max |F_N(x) - F(x)| ,$$

где $F_N(x) = \frac{m_x}{N}$ – эмпирическая функция распределения (N – объем выборки, m_x – количество объектов в выборке, не превышающих величины x), $F(x) = P\{X < x\}$ – теоретическая функция распределения. Для равномерного распределения на отрезке [0,1] $F(x) = x$.

Теорема Колмогорова утверждает, что какова бы ни была непрерывная функция распределения F(x), вероятность $P\{D_N < \frac{\lambda}{\sqrt{N}}\}$ при $N \to \infty$ стремится к пределу:

$$K(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2\lambda^2 k^2}.$$

Анализ генератора случайных чисел, равномерно распределенных на отрезке [0,1], при помощи критерия Колмогорова осуществляется следующим образом.

1. Генерируется последовательность случайных чисел длиной N: $x_1, x_2, \ldots, x_N$ ($x_i \in [0,1]$).

2. Упорядочивается эта последовательность в возрастающем* порядке: $x'_1, x'_2, \ldots, x'_N$ ($x'_i < x'_j$, если $i < j$).

3. Вычисляется величина D_N, которая в данном случае выражается формулой

$$D_N = \max_{i=1,\ldots,N} \left\{ \left| \frac{i-1}{N} - x'_i \right| , \left| \frac{i}{N} - x'_i \right| \right\} .$$

---

* Если N не слишком большой (N меньше периода), то все числа в последовательности разные.

Полученная величина сравнивается с соответствующими значениями функции $K(\lambda)$.

Генератор считается удовлетворительным, если величина $D_N\sqrt{N} \in [0.5,1.5]$ (теоретическая вероятность появления такого события 0.94). Генератор хороший, если $D_N\sqrt{N} \in [0.7,1.0]$.

Пример 6. Было генерировано несколько последовательностей случайных чисел разной длиной N и были вычислены соответствующие числа D_N и $D_N\sqrt{N}$. Результаты изложены в таблице 4.

Таблица 4.

N	D_N	$D_N\sqrt{N}$
1500	0.01496	0.58
	0.0137	0.53
5000	0.0144	1.02
	0.011	0.77
10000	0.007	0.67
15000	0.0048	0.58
	0.0074	0.91
20000	0.0035	0.50
30000	0.0035	0.61

Как видно, все числа $D_N\sqrt{N}$ расположены в интервале $[0.5,1.5]$.

Чтобы избежать случайности при однократной проверке генератора, надо, как и при проверке частот и комбинаций, генерировать несколько последовательностей случайных чисел, вычислять каждый раз величины D_N и устанавливать при помощи критерия χ^2, являются ли числа D_N конкретными значениями случайной величины с распределением Колмогорова.

Пример 7. Было генерировано 120 последовательностей, в каждой было 10000 случайных чисел. Каждый раз вычислялись величины D_N, которые были разделены на 19 классов. Полученное эмпирическое распределение сравнивалось с распределением Колмогорова. Соответствующий хи-квадрат был 16.68. Поскольку имеет место неравенство

$$\chi^2(0.25;18) < 16.68 < \chi^2(0.5;18),$$

то можно сказать, что наш генератор удовлетворяет критерию Колмогорова.

Заключение

Систематическое исследование генератора случайных чисел (1)-(2) показало, что характеристики случайности и статистики, измеряющие близость распределения полученных случайных чисел к равномерному, очень хорошо согласуются с теоретическими предположениями. Все-таки оказалось, что низкие разряды в двоичном представлении случайных чисел менее случайные чем высшие.

Следует отметить, что периодичность полученных случайных чисел не была специально исследована. Но так как случайность проверялась при последовательностях разной длины, то можно предположить, что рассматриваемый генератор не обладает заметной периодичностью при последовательностях длиной до 150000 единиц.

Резюмируя вышесказанное, можно сделать вывод, что датчик (2) при случайном исходном значении x_0, удовлетворяющем условию (1), генерирует последовательность случайных чисел,

равномерно распределенных на отрезке [0,1], и тем самым является применимым в решении практических задач статистического моделирования.

Литература

I. Голенко Д.И., Моделирование и статистический анализ псевдослучайных чисел на электронных вычислительных машинах. Москва, 1965.

2. Ермаков С.М., Метод Монте-Карло и смежные вопросы. Москва, 1971.

3. Кнут Д., Искусство программирования для ЭВМ, т.2, Москва, 1977.

МАТРИЧНАЯ ПРОИЗВОДНАЯ С ПРИМЕНЕНИЕМ ДЛЯ БЛОК-МАТРИЦ

Т.Колло, Т.Кинкар

В данной статье изучается понятие матричной производной, с целью ее дальнейшего применения в многомерном статистическом анализе. Хотя в литературе матричная производная определяется различными способами (см., например [3], [4], [5]), предложенная в данной статье возможность по мнению авторов имеет некоторые преимущества. Не все свойства рассматривались ранее в литературе или же имеют в данном случае более простой вид.

1. Понятия и обозначения

Пусть $M = (m_{ij})$ - $(p \times q)$-матрица. Тогда

$$\operatorname{vec} M = (m_{11}, m_{21}, \ldots, m_{p1}, m_{12}, m_{22}, \ldots, m_{p2}, \ldots, m_{1q}, m_{2q}, \ldots, m_{pq})'.$$

Единичная матрица порядка p обозначается через I_p.

Для указания элемента m_{ij} матрицы M применяется также обозначение $(M)_{ij}$. $(p \times q)$-матрица M является блок-матрицей, если матрица M разделена на uv подматрицы: $M=[M_{ij}]$ $(i=1,\ldots,u;$ $j=1,\ldots,v)$ так, что M_{ij} является $(p_i \times q_j)$-матрицей, где $\sum_{i=1}^{u} p_i = p$; $\sum_{j=1}^{v} q_j = q$. Для указания элемента блок-матрицы применяются двойные индексы: элемент находится в (k,l)-ой

строке и в (g,h)-ом столбце, если он находится в l-ой строке k-ой строки блоков и в h-ом столбце g-го столбца блоков, т.е.

$$m_{(k,l)(g,h)} = m_{\sum_{i=1}^{k-1} p_i + l,\ \sum_{j=1}^{g-1} q_j + h}.$$

Если в блок-матрице $u=1$ или $v=1$, то для указания строки и столбца соответственно применяется обычная индексация.

Для $(p \times q)$-матрицы M и $(r \times s)$-матрицы N прямое произведение $M \otimes N$ определяется равенством

$$M \otimes N = [m_{ij}N] \quad (i=1,\ldots,p;\ j=1,\ldots,q).$$

Перечислим основные свойства прямого произведения (1-6 см. [2], § 8.2), предполагая, что размерности матриц таковы, что все операции определены.

СВОЙСТВО 1. $(cM) \otimes N = M \otimes (cN) = c(M \otimes N), \quad c \in R.$

СВОЙСТВО 2. $(M \otimes N)' = M' \otimes N'.$

СВОЙСТВО 3. $(M+N) \otimes (U+V) = (M \otimes U) + (M \otimes V) + (N \otimes U) + (N \otimes V).$

СВОЙСТВО 4. $M \otimes (N \otimes U) = (M \otimes N) \otimes U.$

СВОЙСТВО 5. $(M \otimes N)(U \otimes V) = (MU) \otimes (NV).$

СВОЙСТВО 6. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$

Из определения прямого произведения непосредственно вытекает, что

СВОЙСТВО 7. $(M \otimes N)_{(i,j)(g,h)} = m_{ig} n_{jh}$ $(i=1,\ldots,p;\ g=1,\ldots,q;\ j=1,\ldots,r;\ h=1,\ldots,s).$

СВОЙСТВО 8. Пусть M - $(p \times q)$-матрица, N - $(q \times r)$-матрица и U - $(r \times s)$-матрица. Тогда

$$\operatorname{vec}(MNU) = (U' \otimes M) \operatorname{vec} N .$$

Доказательство. Обозначим j-ый столбец матрицы U через U_j, то j-ый столбец матрицы MNU равен MNU_j. Тогда

$$MNU_j = \sum_{i=1}^{r} MN_i u_{ij} = \sum_{i=1}^{r} u_{ij} MN_i = (U_j' \otimes M) \operatorname{vec} N .$$

Так как данная цепочка равенств имеет место для $j=1,\ldots,s$, то в матричном виде получим отсюда желаемое равенство.

Переставленной единичной матрицей $I_{p,q}$ называется следующая $(pq\times pq)$-блок-матрица, составленная из $(p\times q)$-блоков, где

$$(I_{p,q})_{(i,j)(g,h)} = \begin{cases} 1, & g=j \text{ и } h=i \\ & i,h=1,\ldots,p;\ j,g=1,\ldots,q; \\ 0, & \text{в противном случае.} \end{cases}$$

Из определения вытекают равенства:

$$I_{p,q}I_{q,p} = I_{pq};$$
$$I_{p,q} = I'_{q,p};$$
$$I_{p,1} = I_{1,p} = I_p .$$

СВОЙСТВО 9. Пусть M – $(p\times q)$-матрица и N – $(r\times s)$-матрица, тогда

$$N\otimes M = I_{p,r}(M\otimes N)I_{s,q}.$$

Доказательство. Из определения получим, что

$$\left[I_{p,r}(M\otimes N)\right]_{(i,j)(g,h)} = (M\otimes N)_{(j,i)(g,h)};$$
$$\left[(M\otimes N)I_{s,q}\right]_{(i,j)(g,h)} = (M\otimes N)_{(i,j)(h,g)}.$$

Тогда

$$\left[I_{p,r}(M\otimes N)I_{s,q}\right]_{(i,j)(g,h)} = (M\otimes N)_{(j,i)(h,g)} =$$
$$= m_{jh}n_{ig} = (N\otimes M)_{(i,j)(g,h)}.$$

2. Матричная производная

Предположим, что элементы $(r\times s)$-матрицы Y являются функциями элементов $(p\times q)$-матрицы X.

Определяем дифференциальный оператор $\frac{d}{dX}$ для $(p\times q)$-матрицы X через равенство:

$$\frac{d}{dX} = \frac{d}{d\ \mathrm{vec}\ X} = \left(\frac{\partial}{\partial x_{11}},\ldots,\frac{\partial}{\partial x_{p1}},\ldots,\frac{\partial}{\partial x_{1q}},\ldots,\frac{\partial}{\partial x_{pq}}\right)'.$$

ОПРЕДЕЛЕНИЕ 1. Пусть Y – $(r \times s)$-матрица, элементы которой являются функциями от $(p \times q)$-матрицы X. Назовем матричной производной от Y по X $(pq \times rs)$-матрицу

$$\frac{dY}{dX} = \frac{d}{dX} \otimes (\mathrm{vec}\ Y)' .$$

По определению i-ая строка $(i=1,\ldots,pq)$ матрицы $\frac{dY}{dX}$ состоит из частных производных координат вектора $(\mathrm{vec}\ Y)'$ по i-ой координате $\mathrm{vec}\ X$. Если $x_{ij} = \mathrm{const}$, то соответствующая строка в матрице $\frac{dY}{dX}$ является нулевой. Функциональные связи между элементами матрицы X не учитываются.

Приведем основные свойства матричной производной, указывая размерности матриц только в случае их несовпадения с размерностями в определении.

СВОЙСТВО 1. $\frac{dX}{dX} = I_{pq}$.

СВОЙСТВО 2. $\frac{dX'}{dX} = I_{p,q}$.

СВОЙСТВО 3. Если X_d – диагональная матрица, полученная из $(p \times p)$-матрицы X, то

$$\frac{dX_d}{dX} = H^p,$$

где $(H^p)_{(i,j)(g,h)} = \begin{cases} 1, & i=j=g=h,\ i,j,g,h=1,\ldots,p; \\ 0, & \text{в противном случае.} \end{cases}$

Свойства 1 – 3 вытекают непосредственно из определения производной.

СВОЙСТВО 4. Если элементы $(m \times n)$-матрицы Z являются функциями от $(r \times s)$-матрицы Y, элементы которой в свою очередь являются функциями от $(p \times q)$-матрицы X, то

$$\frac{dZ}{dX} = \frac{dY}{dX}\,\frac{dZ}{dY} .$$

Свойство доказано в [1].

СВОЙСТВО 5. Если $Y = AXB$, где A – постоянная $(r \times p)$-матрица и B – постоянная $(q \times s)$-матрица, то

$$\frac{dZ}{dX} = B \otimes A' .$$

Доказательство. По определению матричной производной

$$\frac{dY}{dX} = \frac{d}{d\ vec\ X} \otimes (vec(AXB))' .$$

По свойству 8 прямого произведения

$$vec(AXB) = (B' \otimes A)\ vec\ X .$$

Тогда

$$\frac{dY}{dX} = \frac{d}{d\ vec\ X} \otimes (vec\ X)'(B \otimes A') .$$

При помощи свойства 5 прямого произведения

$$\frac{dY}{dX} = \left(\frac{d}{d\ vec\ X} \times 1\right) \otimes (vec\ X)'(B \otimes A') =$$

$$= \left[\frac{d}{d\ vec\ X} \otimes (vec\ X)'\right](B \otimes A') = B \otimes A' .$$

СВОЙСТВО 6. Если $Z = AYB$, где $(r \times s)$-матрица Y является функцией $(p \times q)$-матрицы X, а A – постоянной $(m \times r)$-матрицей и B – постоянной $(s \times n)$-матрицей, то

$$\frac{dZ}{dX} = \frac{dY}{dX}(B \otimes A') .$$

Доказательство. Свойство вытекает из свойств 4 и 5 матричной производной. По свойству 4

$$\frac{dZ}{dX} = \frac{dY}{dX}\frac{d(AYB)}{dY} ,$$

а по свойству 5

$$\frac{d(AYB)}{dY} = B \otimes A'$$

и

$$\frac{dZ}{dX} = \frac{dY}{dX}(B \otimes A') .$$

СВОЙСТВО 7. Если элементы (m×n)-матрицы W являются функциями (r×s)-матрицы Y и (k×1)-матрицы Z, которые в свою очередь являются функциями (p×q)-матрицы X, то

$$\frac{dW}{dX} = \frac{dY}{dX}\frac{dW}{dY}\Big|_{Z=const} + \frac{dZ}{dX}\frac{dW}{dZ}\Big|_{Y=const}.$$

Доказательство. Для элемента w_{ij} $(i=1,\ldots,m;\ j=1,\ldots,n)$ матрицы W частная производная по элементу x_{gh} $(g=1,\ldots,p;\ h=1,\ldots,q)$ матрицы X равна

$$\frac{\partial w_{ij}}{\partial x_{gh}} = \sum_{u=1}^{r}\sum_{v=1}^{s}\frac{\partial y_{uv}}{\partial x_{gh}}\frac{\partial w_{ij}}{\partial y_{uv}}\Big|_{Z=const} + \sum_{u=1}^{r}\sum_{v=1}^{s}\frac{\partial z_{uv}}{\partial x_{gh}}\frac{\partial w_{ij}}{\partial z_{uv}}\Big|_{Y=const}.$$

По определению матричной производной

$$\frac{dw_{ij}}{dY} = \left(\frac{\partial w_{ij}}{\partial y_{11}},\ldots,\frac{\partial w_{ij}}{\partial y_{rs}}\right)'; \qquad \frac{dw_{ij}}{dZ} = \left(\frac{\partial w_{ij}}{\partial z_{11}},\ldots,\frac{\partial w_{ij}}{\partial z_{kl}}\right)';$$

$$\frac{dY}{dx_{gh}} = \left(\frac{\partial y_{11}}{\partial x_{gh}},\ldots,\frac{\partial y_{rs}}{\partial x_{gh}}\right); \qquad \frac{dZ}{dx_{gh}} = \left(\frac{\partial z_{11}}{\partial x_{gh}},\ldots,\frac{\partial z_{kl}}{\partial x_{gh}}\right).$$

Отсюда

$$\frac{\partial w_{ij}}{\partial x_{gh}} = \frac{dY}{dx_{gh}}\frac{dw_{ij}}{dY}\Big|_{Z=const} + \frac{dZ}{dx_{gh}}\frac{dw_{ij}}{dZ}\Big|_{Y=const}.$$

Учитывая, что равенство имеет место при любых значениях индексов i,j,g,h $(i=1,\ldots,m;\ j=1,\ldots,n;\ g=1,\ldots,p;\ h=1,\ldots,q)$, получим доказываемое равенство.

СВОЙСТВО 8. Если (m×r)-матрица Z и (r×s)-матрица Y являются функциями от элементов матрицы X, то

$$\frac{d(ZY)}{dX} = \frac{dZ}{dX}(Y\otimes I_m) + \frac{dY}{dX}(I_s\otimes Z').$$

Доказательство. По свойству 7 производной

$$\frac{d(ZY)}{dX} = \frac{dY}{dX}\frac{d(ZY)}{dY}\Big|_{Z=const} + \frac{dZ}{dX}\frac{d(ZY)}{dZ}\Big|_{Y=const}.$$

По свойству 6 матричной производной

$$\frac{d(ZY)}{dX} = \frac{dY}{dX}(I_s \otimes Z') + \frac{dZ}{dX}(Y \otimes I_m).$$

СВОЙСТВО 9. Если $(r \times r)$-матрица Y является функцией от элементов $(p \times q)$-матрицы X, то

$$\frac{dY^n}{dX} = \frac{dY}{dX} \sum_{\substack{i+j=n-1 \\ i,j\geq 0}} (Y^i \otimes (Y')^j), \quad \text{где} \quad Y^0 = I_r, \; n \geq 1.$$

Доказательство. Используем метод полной индукции. При $n = 1$

$$\frac{dY}{dX} = \frac{dY}{dX}(I_r \otimes I_r).$$

При $n = 2$ по свойству 8

$$\frac{dY^2}{dX} = \frac{dY}{dX}(I \otimes Y') + \frac{dY}{dX}(Y \otimes I) = \frac{dY}{dX} \sum_{\substack{i+j=1 \\ i,j\geq 0}} (Y^i \otimes (Y')^j).$$

При $n=3$, применяя свойство 8 и предыдущее равенство, получим

$$\frac{dY^3}{dX} = \frac{dY}{dX}(I \otimes (Y')^2) + \frac{dY}{dX}(Y \otimes Y') + \frac{dY}{dX}(Y^2 \otimes I) =$$

$$= \frac{dY}{dX} \sum_{\substack{i+j=2 \\ i,j\geq 0}} (Y^i \otimes (Y')^j).$$

Предположим, что утверждение имеет место при k. Тогда при $(k+1)$ получим:

$$\frac{dY^{k+1}}{dX} = \frac{dY \cdot Y^k}{dX} = \frac{dY^k}{dX}(I \otimes Y') + \frac{dY}{dX}(Y^k \otimes I) =$$

$$= \Big(\sum_{\substack{i+j=k-1 \\ i,j\geq 0}} \frac{dY}{dX}(Y^i \otimes (Y')^j)(I \otimes Y') + \frac{dY}{dX}(Y^k \otimes I) =$$

$$= \sum_{\substack{i+j=k-1 \\ i,j\geq 0}} \frac{dY}{dX}(Y^i \otimes (Y')^{j+1}) + \frac{dY}{dX}(Y^k \otimes I) =$$

$$= \sum_{\substack{m+l=k \\ m,l\geq 0}} \frac{dY}{dX}(Y^m \otimes (Y')^l) = \frac{dY}{dX} \sum_{\substack{m+l=k \\ m,l\geq 0}} (Y^m \otimes (Y')^l) .$$

Свойство доказано.

СВОЙСТВО 10. Если $(r \times r)$-матрица Y является функцией от элементов $(p \times q)$-матрицы X и матрица Y регулярна, то если $Y^{-n} = \underbrace{Y^{-1} \cdot \ldots \cdot Y^{-1}}_{n}$, $(n \geq 1)$, тогда

$$\frac{dY^{-n}}{dX} = \frac{dY^{-1}}{dX} \sum_{\substack{i+j=n-1 \\ i,j \geq 0}} (Y^{-i} \otimes (Y')^{-j}) = -\frac{dY}{dX} \sum_{\substack{i+j=n-1 \\ i,j \geq 0}} (Y^{-i-1} \otimes (Y')^{-j-1}).$$

Доказательство. Обозначим $Y^{-1} = Z$. Тогда используя свойство 9 получим

$$\frac{dY^{-n}}{dX} = \frac{dZ^{n}}{dX} = \frac{dZ}{dX} \sum_{\substack{i+j=n-1 \\ i,j \geq 0}} (Z^{i} \otimes (Z')^{j}) = \frac{dY^{-1}}{dX} \sum_{\substack{i+j=n-1 \\ i,j \geq 0}} (Y^{-i} \otimes ((Y^{-1})')^{j}).$$

Чтобы получить первое требуемое равенство используем соотношение

$$(Y^{-1})' = (Y')^{-1}.$$

Первое равенство доказано. Для доказательства второго равенства отметим, что $\frac{dX^{-1}}{dX} = X^{-1} \otimes (X')^{-1}$. Действительно, из соотношения $XX^{-1} = I$ получим, используя свойство 8, что

$$\frac{dX}{dX}(X^{-1} \otimes I) + \frac{dX^{-1}}{dX}(I \otimes X') = 0.$$

Отсюда при помощи свойства 5 прямого произведения

$$\frac{dX^{-1}}{dX} = -(X^{-1} \otimes I)(I \otimes (X')^{-1}) = -(X^{-1} \otimes (X')^{-1}).$$

По свойству 4 производной получим

$$\frac{dY^{-1}}{dX} = \frac{dY}{dX}\frac{dY^{-1}}{dY} = -\frac{dY}{dX}(Y^{-1} \otimes (Y')^{-1}).$$

Тогда

$$\frac{dY^{-n}}{dX} = -\frac{dY}{dX}(Y^{-1} \otimes (Y')^{-1}) \sum_{\substack{i+j=n-1 \\ i,j \geq 0}} (Y^{-i} \otimes (Y')^{-j}).$$

После применения свойства 5 прямого произведения получим желаемый результат.

СВОЙСТВО 11. Пусть X - $(p \times p)$-матрица и diag X - p-вектор, образованный из диагональных элементов матрицы X. Тогда

$$\frac{d(\operatorname{diag} X)}{dX} = K^{(p)},$$

где $K^{(p)}$ - $(p^2 \times p)$-матрица, состоящая из $(p \times p)$-блоков, а

$$(K^{(p)})_{(i,j)k} = \begin{cases} 1, & \text{если } i=j=k;\ i,j,k=1,\ldots,p; \\ 0, & \text{в противном случае.} \end{cases}$$

Свойство следует непосредственно из определения матричной производной.

СВОЙСТВО 12. Пусть Y - $(r \times r)$-матрица, элементы которой являются функциями $(p \times q)$-матрицы X. Тогда

$$\frac{d(\operatorname{diag} Y)}{dX} = \frac{dY}{dX} K^{(r)}.$$

Доказательство. Свойство вытекает из свойств 4 и 11 матричной производной. По свойству 4

$$\frac{d(\operatorname{diag} Y)}{dX} = \frac{dY}{dX} \frac{d(\operatorname{diag} Y)}{dY},$$

откуда по свойству 11

$$\frac{d(\operatorname{diag} Y)}{dX} = \frac{dY}{dX} K^{(r)}.$$

Следующие свойства производной касаются блок-матриц.

Пусть $(p \times q)$-матрица X является блок-матрицей, состоящей из uv блоков, которые являются $(p_i \times q_j)$-матрицами $(\sum_{i=1}^{u} p_i = p;\ \sum_{j=1}^{v} q_j = q)$:

$$X = [X_{ij}] \quad (i=1,\ldots,u;\ j=1,\ldots,v),$$

а $(r \times s)$-матрица Y составлена из $(r_i \times s_j)$-блоков, так что

$$Y = [Y_{ij}] \quad (i=1,\ldots,m;\ j=1,\ldots,n).$$

СВОЙСТВО 13. В вышеуказанных обозначениях

$$\frac{dY}{dX} = C_{ij} = \left[\frac{d\begin{pmatrix} Y_{1j} \\ \cdots \\ Y_{mj} \end{pmatrix}}{d\begin{pmatrix} X_{1i} \\ \cdots \\ X_{ui} \end{pmatrix}} \right] \qquad (i=1,\ldots,v;\ j=1,\ldots,n).$$

Доказательство. По определению

$$\frac{dY}{dX} = \frac{d}{d\ \mathrm{vec}\ X} \otimes (\mathrm{vec}\ Y)'.$$

Частная производная $\frac{\partial y_{ij}}{\partial x_{kl}}$ находится в этой матрице в $[p(l-1)+k]$-ой строке и $[r(j-1)+i]$-ом столбце. Пусть y_{ij} находится в i'-ой строке и j'-ом столбце блока Y_{cd} и x_{kl} - в k'-ой строке и l'-ом столбце блока X_{fg}.

Тогда

$$k = \sum_{t=1}^{f-1} p_t + k' \ ;$$

$$l = \sum_{t=1}^{g-1} q_t + l'$$

и

$$i = \sum_{t=1}^{c-1} r_t + i' \ ;$$

$$j = \sum_{t=1}^{d-1} s_t + j' \ .$$

Производная $\frac{\partial y_{ij}}{\partial x_{kl}}$ находится в строке с индексом

$$p(l-1)+k = p \sum_{t=1}^{g-1} q_t + p(l'-1) + \sum_{t=1}^{f-1} p_t + k'$$

и в столбце с индексом

$$r(j-1)+i = r\sum_{t=1}^{d-1} s_t + r(j'-1) + \sum_{t=1}^{c-1} r_t + i'.$$

В матрице

$$C = [c_{ij}] \quad (i=1,\ldots,v;\ j=1,\ldots,n)$$

производная $\frac{\partial y_{ij}}{\partial x_{kl}}$ находится в блоке

$$c_{gd} = \frac{d\begin{pmatrix} y_{1d} \\ \cdots \\ y_{md} \end{pmatrix}}{d\begin{pmatrix} x_{1g} \\ \cdots \\ x_{ug} \end{pmatrix}}$$

в $(p(l'-1)+\sum_{t=1}^{f-1} p_t + k')$-ой строке и $(r(j'-1)+\sum_{t=1}^{c-1} r_t + i')$-ом столбце. Тогда в матрице C индексы строки и столбца производной $\frac{\partial y_{ij}}{\partial x_{kl}}$ совпадают с соответствующими индексами в матрице $\frac{dY}{dX}$, откуда $\frac{dY}{dX} = [c_{ij}]$ $(i=1,\ldots,v;\ j=1,\ldots,n)$.

В многомерном статистическом анализе чаще всего встречается следующая блок-структура:

$$X = \left(\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right), \quad Y = \left(\begin{array}{c|c} Y_{11} & Y_{12} \\ \hline Y_{21} & Y_{22} \end{array}\right).$$

СВОЙСТВО 14. Пусть $X=[X_{ij}]$ $(i,j=1,2)$ и $Y=[Y_{ij}]$ $(i,j=1,2)$. Тогда

$$\frac{dY}{dX} = \left(\begin{array}{c|c} \dfrac{d\left(\frac{Y_{11}}{Y_{21}}\right)}{d\left(\frac{X_{11}}{X_{21}}\right)} & \dfrac{d\left(\frac{Y_{12}}{Y_{22}}\right)}{d\left(\frac{X_{11}}{X_{21}}\right)} \\ \hline \dfrac{d\left(\frac{Y_{11}}{Y_{21}}\right)}{d\left(\frac{X_{12}}{X_{22}}\right)} & \dfrac{d\left(\frac{Y_{12}}{Y_{22}}\right)}{d\left(\frac{X_{12}}{X_{22}}\right)} \end{array}\right).$$

Это равенство получим непосредственно из предыдущего свойства при $m,n,u,v=2$.

СВОЙСТВО 15. Пусть $X=[X_{ij}]$ $(i,j=1,2)$, где X_{11} - $(r \times s)$-матрица и X_{22} - $(p \times q)$-матрица. Тогда

$$\frac{dX_{11}}{dX} = \begin{pmatrix} I_s \\ 0_{q,s} \end{pmatrix} \otimes \begin{pmatrix} I_r \\ 0_{p,r} \end{pmatrix} ;$$

$$\frac{dX_{12}}{dX} = \begin{pmatrix} 0_{s,q} \\ I_q \end{pmatrix} \otimes \begin{pmatrix} I_r \\ 0_{p,r} \end{pmatrix} ;$$

$$\frac{dX_{21}}{dX} = \begin{pmatrix} I_s \\ 0_{q,s} \end{pmatrix} \otimes \begin{pmatrix} 0_{r,p} \\ I_p \end{pmatrix} ;$$

$$\frac{dX_{22}}{dX} = \begin{pmatrix} 0_{s,q} \\ I_q \end{pmatrix} \otimes \begin{pmatrix} 0_{r,p} \\ I_p \end{pmatrix} ,$$

где $0_{m,n}$ - нулевая $(m \times n)$-матрица.

Равенства свойства 15 проверяются непосредственно как равенства матриц, с применением на левой стороне определения матричной производной, а на правой - определения прямого произведения.

Литература

I. Колло Т., Некоторые понятия матричного исчисления с применением в математической статистике. Труды ВЦ ТГУ, Тарту, 1977, 40, 30-51.

2. Ланкастер П., Теория матриц. М., 1982.

3. Macrae, E.G., Matrix derivatives with an application to an adaptive linear decision problem. Ann. Statist., 1974, 2, 337-346.

4. McDonald, R.R., Swaminathan, H., A simple matrix calculus with applications to multivariate analysis. Gen. Syst., 1973, 18, 37-54.

5. Neudecker, H., Some theorems on matrix differentiation with special reference to Kronecker matrix products. J. Amer. Statist. Assoc., 1969, 64, 953-963.

МОМЕНТЫ ВЫБОРОЧНОЙ КОВАРИАЦИОННОЙ МАТРИЦЫ

И. Траат

При выведении асимптотических распределений случайных векторов, являющихся функциями выборочной ковариационной матрицы, необходимо знать выражения моментов ковариационной матрицы (см., например [10]). В двумерном случае формулы для моментов отдельных элементов ковариационной матрицы были получены в 1951 г. Куком [7]. В тензорных обозначениях Каплана [8] эти формулы представляются в компактном виде для p-мерного случая. Однако, как заметил сам Каплан, при выводе из них формул для частных случаев необходима осторожность. Например, тензорная формула для вторых моментов элементов ковариационной матрицы включает в себя семь различных выражений: дисперсия дисперсии, дисперсия ковариации, ковариация двух дисперсий и т.д. (см. [1], стр. 440). Частных случаев для третих моментов гораздо больше.

В настоящей статье, используя определения моментов для случайного вектора Колло [3], и векторное представление ковариационной матрицы, выводятся в матричном виде второй момент и главные члены третьего и четвертого моментов ковариационной матрицы. Полученные матричные выражения, являющиеся функциями моментов исходного случайного вектора, содержат в

себе все выражения моментов отдельных элементов ковариационной матрицы. Требуется существование конечных моментов исходного вектора до восьмого порядка.

Матричное представление момента особенно удобно в вычислениях с помощью ЭВМ.

1. Некоторые понятия матричного исчисления

Основным средством при выведении результатов настоящей статьи является прямое (кронекеровское) произведение матриц [2], [5], [9]. Приведем здесь определение и нужные нам свойства.

Пусть $M = (m_{ij})$ - $(p \times q)$-матрица, $N = (n_{kl})$ - $(r \times s)$-матрица, а размерности матриц U и V таковы, что все операции определены. Прямым произведением матриц M и N называется следующая $(pr \times qs)$-блок-матрица:

$$M \otimes N = [m_{ij}N] .$$

При установлении связи между элементами матриц M, N и $M \otimes N$, заметим что элементы матрицы M фиксируют позицию блока (i,j) в матрице $M \otimes N$, а элементы матрицы N фиксируют позицию элемента (k,l) в этом блоке. Следовательно,

$$m_{ij}n_{kl} = (M \otimes N)_{(i-1)r+k,(j-1)s+l} .$$

Важной матрицей в связи с операциями прямого произведения является переставленная единичная матрица $I_{p,q}$. Матрица $I_{p,q}$ определяется как $(pq \times pq)$-матрица, составленная из $(p \times q)$-блоков так, что в (i,j)-ом блоке (j,i)-ый элемент равен единице, а все остальные нулям. Умножение матрицы U слева на $I_{p,q}$ переставляет ряды матрицы U, а умножение U справа

на $I_{p,q}$ переставляет столбцы матрицы U. Имеет место равенство:

$$I_{p,q}I_{q,p} = I_{pq} .$$

В частности,

$$I_{p,1} = I_{1,p} = I_p .$$

В матричных выражениях часто приходится использовать вместе с матрицей M ее векторное представление vM (в литературе и vec M):

$$vM = (m_{11},\dots,m_{p1},m_{12},\dots,m_{p2},\dots,m_{1q},\dots,m_{pq})' .$$

Справедливы свойства:

$$v(cM) = c \cdot vM;$$

$$v(M+U) = vM + vU.$$

Прямое произведение обладает следующими свойствами:

1) если c константа, то $(cM)\otimes N = M\otimes(cN) = c(M\otimes N)$;

2) $(M\otimes N)(U\otimes V) = (MU)\otimes(NV)$;

3) $(M\otimes N)' = M'\otimes N'$;

4) $(M+U)\otimes(N+V) = (M\otimes N)+(M\otimes V)+(U\otimes N)+(U\otimes V)$;

5) $M\otimes(N\otimes U) = (M\otimes N)\otimes U$;

6) $M\otimes N = I_{r,p}(N\otimes M)I_{q,s}$.

Если X и Y p-векторы, тогда:

7) $XY' = X\otimes Y'$;

8) $X\otimes Y' = Y'\otimes X$;

9) $v(X\otimes Y') = Y\otimes X$.

2. Моменты и кумулянты случайного вектора

Пусть X случайный p-вектор. Следуя [3] моменты $M_i(X)$ определяются равенствами:

$$\begin{cases} M_1(X) = EX = \mu; \\ M_2(X) = E(X \otimes X'); \\ M_3(X) = E(X \otimes X' \otimes X); \\ \cdots \end{cases} \tag{1}$$

Аналогичным образом определяются центральные моменты $\bar{M}_i(X)$:

$$\begin{cases} \bar{M}_1(X) = 0; \\ \bar{M}_2(X) = E[(X-\mu) \otimes (X-\mu)'] = \Sigma; \\ \bar{M}_3(X) = E[(X-\mu) \otimes (X-\mu)' \otimes (X-\mu)]; \\ \cdots \\ \bar{M}_6(X) = E[(X-\mu) \otimes (X-\mu)' \otimes (X-\mu) \otimes (X-\mu)' \otimes (X-\mu) \otimes (X-\mu)']; \\ \cdots \end{cases} \tag{2}$$

Ниже для центральных моментов исходного вектора X используем обозначение без аргумента ($\bar{M}_i(X) = \bar{M}_i$).

Для представления в матричном виде всех маргинальных и совместных моментов i-го порядка вектора X много эквивалентных возможностей. Так замена одного или нескольких векторов $X-\mu$ на $(X-\mu)'$ или, наоборот, в выражениях $\bar{M}_i(X)$, не изменяет элементов, а только структуру $\bar{M}_i(X)$. Для наших целей полезно определить шестой центральный момент как следующую $(p^4 \times p^2)$-матрицу:

$$\bar{M}_6^*(X) = E[(X-\mu) \otimes (X-\mu)' \otimes (X-\mu) \otimes (X-\mu)' \otimes (X-\mu) \otimes (X-\mu)]. \tag{3}$$

Часто характеристиками распределения вместо моментов используют кумулянты. Определяем матричное представление i-ых совместных кумулянтов вектора X - $K_i(X)$, имея в виду струк-

туру момента $M_i(X)$. Так соответствующие элементы матриц $K_i(X)$ и $M_i(X)$ представляют собой совместный кумулянт и совместный момент i-ого порядка некоторой подгруппы вектора X. Связи между первыми моментами и кумулянтами $K_i(X)$ простые:

$$K_1(X) = M_1(X) = E(X);$$

$$K_2(X) = \bar{M}_2(X) = \Sigma; \tag{4}$$

$$K_3(X) = \bar{M}_3(X); \tag{5}$$

$$K_4(X) = \bar{M}_4(X) - \Sigma \otimes \Sigma - I_{p,p}(\Sigma \otimes \Sigma) - v\Sigma(v\Sigma)'. \tag{6}$$

Кумулянт $K_5(X)$ представляется через центральные моменты уже с помощью 11 слагаемых, $K_6(X)$ с помощью 56 слагаемых. Выражения для отдельных элементов матриц $K_i(X)$ через элементы $\bar{M}_i(X)$ в одномерном и двумерном случаях даны, например, в [1]. Матричное равенство дает одновременно все связи для p-мерного случая.

Если случайный p-вектор Y зависит линейно от X:

$$Y = UX ,$$

то с помощью свойств 2, 3, 7, 8 прямого произведения можно выводить соотношения между моментами $M_i(Y)$ и $M_i(X)$:

$$\begin{cases} M_1(Y) = UM_1(X); \\ M_2(Y) = UM_2(X)U'; \\ M_3(Y) = (U \otimes U)M_3(X)U'; \\ M_4(Y) = (U \otimes U)M_4(X)(U \otimes U)'; \\ \cdots \end{cases} \tag{7}$$

Соотношения (7) справедливы и для центральных моментов $\bar{M}_i(X)$.

3. Методика выведения моментов ковариационной матрицы

Выведение матричного выражения для моментов выборочной ковариационной матрицы происходит в общем аналогично выведению выражения момента для одного элемента ковариационной матрицы (например, в [1], стр. 381-382 выведено выражение дисперсии для дисперсии).

Пусть $X_1,\ldots,X_n$ выборка объема n вектора X. Каждый X_i можно рассматривать как независимый от X_j, $i \neq j$, p-вектор с распределением вектора X, моменты которого определены равенствами (1), (2). Формула для выборочной ковариационной матрицы представляется в виде

$$S = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i X_i' - n\bar{X}\bar{X}'\right], \qquad (8)$$

где

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i . \qquad (9)$$

Центральные выборочные моменты, в том числе и матрица S, не зависят от сдвига исходного случайного вектора X. Поэтому, не ограничивая общности, можно при выведении моментов для S предполагать, что

$$EX = 0 . \qquad (10)$$

Полученные при этом формулы будут справедливыми независимо от значения EX.

Образуем случайный p^2-вектор

$$vS = \frac{1}{n-1}\left[\sum_{i=1}^{n} v(X_i X_i') - nv(\bar{X}\bar{X}')\right] . \qquad (11)$$

Учитывая что,

$$E\,vS = v\Sigma , \qquad (12)$$

применяем определения моментов (2) для νS. Подставляем в полученные выражения равенство (11), заменяя $\bar{X}$ равенством (9). После применения свойств 4, 7, 9, остается найти математические ожидания от сумм вида

$$\sum_{i,j,k,\ldots,s=1}^{n} X_i \otimes X_j' \otimes X_k \otimes \ldots X_s ,$$

где число сомножителей X_i равняется r. Эта сумма разлагается на частные суммы по составу индексов у сомножителей. Рассматриваются всевозможные разбиения набора индексов i,j,k,...,s на подмножества равных индексов по следующей таблице.

Количество подмножеств в наборе индексов	Численности равных индексов в подмножествах	Разбиение	C_1	C_2
1	r	ii...i	1	n
2	r-1 ⋮ 1	ii...ij	C_r^1	n(n-1)
2	r-2 ⋮ 2	ii...ijj	C_r^2	n(n-1)
2	r/2 ⋮ r/2	i...ij...j	$C_r^{r/2}/2$	n(n-1)
3	r-2 ⋮ 1 ⋮ 1	ii...ijk	$C_r^1 \cdot C_{r-1}^1/2$	n(n-1)(n-2)
r	1 ⋮ ... ⋮ 1	ij...s	r!/r!	n(n-1)...(n-r+1)

Для проверки $\Sigma C_1 \cdot C_2 = n^r$, где суммирование ведется через строки таблицы.

---

* При нечетном r в отмеченной строке численности индексов (r+1)/2, (r-1)/2, а $C_1 = C_r^{(r-1)/2}$.

При каждом рассматриваемом разбиении существует C_1 различных размещений индексов. Каждому из таких размещений соответствует частная сумма, число слагаемых в которой C_2, $(i,j,\ldots,s=1,\ldots,n)$. Математические ожидания вычисляются отдельно от всех таких частных сумм, учитывая предположения независимости векторов X_i и условие $EX_i = 0$. Притом слагаемые $X_i \otimes X_j' \otimes \ldots X_s$ преобразуются с помощью свойств 1-9 для представления соответствующих математических ожиданий в терминах моментов исходного вектора.

Например, в случае разбиения iijj $(r = 4)$ существуют $C_4^2/2 = 3$ различных размещений индексов: ijij, ijji, iijj. Математическое ожидание от суммы с индексами ijij имеет следующий вид:

$$E\sum_{i,j=1}^{n} X_i \otimes X_j' \otimes X_i \otimes X_j' = \sum_{i,j=1}^{n} E\left[X_i \otimes X_i \otimes X_j' \otimes X_j'\right] =$$

$$= \sum_{i,j=1}^{n} E\left[v(X_i \otimes X_i')\right] \otimes E\left[v(X_j \otimes X_j')\right]' = n(n-1)v\Sigma \cdot (v\Sigma)'.$$

Аналогично вычисляются и математические ожидания других слагаемых.

4. Второй момент ковариационной матрицы

Согласно определениям (2) имеется

$$\tilde{M}_2(vS) = E\left[(vS - v\Sigma) \otimes (vS - v\Sigma)'\right] =$$
$$= E\left[vS \cdot (vS)'\right] - v\Sigma \cdot (v\Sigma)' . \qquad (13)$$

Используя равенство (11) получим:

$$E\left[vS \cdot (vS)'\right] = \frac{1}{(n-1)^2}\left[EA_1 - EA_2 - EA_2' + EA_3\right],$$

где

$$A_1 = v\Big(\sum_i X_i X_i'\Big)\Big[v\Big(\sum_i X_i X_i'\Big)\Big]',$$

$$A_2 = nv\Big(\sum_i X_i X_i'\Big)\big[v(\bar{X}\bar{X}')\big]',$$

$$A_3 = n^2 v(\bar{X}\bar{X}')\big[v(\bar{X}\bar{X}')\big]'.$$

Отсюда

$$EA_1 = E\sum_i\sum_j v(X_iX_i')\,[v(X_jX_j')]' =$$

$$= E\Big\{\sum_i X_i\otimes X_i'\otimes X_i\otimes X_i' + \sum_{i\neq j}\sum v(X_iX_i')\,[v(X_jX_j')]'\Big\} =$$

$$= n\bar{M}_4 + n(n-1)v\Sigma\cdot(v\Sigma)'.$$

С помощью (9) для A_2 получим:

$$EA_2 = \frac{1}{n}E\sum_i\sum_j\sum_k v(X_iX_i')\,[v(X_jX_k')]' =$$

$$= \frac{1}{n}\Big\{\sum_i E(X_i\otimes X_i'\otimes X_i\otimes X_i') + \sum_{i\neq j}\sum E\big[v(X_iX_i')\big[v(X_jX_j')\big]'\big]\Big\} =$$

$$= \bar{M}_4 + (n-1)v\Sigma\cdot(v\Sigma)'.$$

Заметим, что сумма $\Sigma E(X_i\otimes X_j'\otimes X_k\otimes X_l') = 0$, если три или более индексов различны. $EA_2' = EA_2$ по симметричности $\bar{M}_4(X)$.

$$EA_3 = \frac{1}{n^2}E\sum_i\sum_j\sum_k\sum_l v(X_iX_j')\,[v(X_kX_l')]' =$$

$$= \frac{1}{n^2}E\sum_i\sum_j\sum_k\sum_l X_i\otimes X_j\otimes X_k'\otimes X_l'.$$

Отсюда ненулевое математическое ожидание получим в случаях 1) равных индексов и 2) двух пар равных индексов:

$$EA_3 = \frac{1}{n^2}E\Big\{\sum_i X_i\otimes X_i'\otimes X_i\otimes X_i' + \sum_{i\neq j}\sum\big[X_i\otimes X_i\otimes X_j'\otimes X_j' +$$

$$+ X_i\otimes X_j\otimes X_j'\otimes X_i' + X_i\otimes X_j\otimes X_i'\otimes X_j'\big]\Big\}.$$

Поскольку по свойствам 5, 6, 8 и 2

$$X_i\otimes(X_j\otimes X_j'\otimes X_i') = I_{p,p}\big[(X_j\otimes X_j'\otimes X_i')\otimes X_i\big]I_{1,p^2} =$$

$$= I_{p,p}(X_j\otimes X_j'\otimes X_i\otimes X_i'),$$

то

$$E A_3 = \frac{1}{n}\bar{M}_4 + \frac{n-1}{n}\left[\nu\Sigma(\nu\Sigma)' + I_{p,p}(\Sigma\otimes\Sigma) + \Sigma\otimes\Sigma\right].$$

В результате

$$E\left[\nu S(\nu S)'\right] = \frac{1}{n}\bar{M}_4 + \frac{n-1}{n}\nu\Sigma(\nu\Sigma)' + \frac{1}{n(n-1)}\left[\Sigma\otimes\Sigma + I_{p,p}(\Sigma\otimes\Sigma)\right]. \quad (14)$$

По определению (13) второй центральный момент выборочной ковариационной матрицы представляется в виде:

$$\bar{M}_2(\nu S) = \frac{1}{n}\left[\bar{M}_4 - \nu\Sigma(\nu\Sigma)' + \frac{1}{n-1}(\Sigma\otimes\Sigma + I_{p,p}(\Sigma\otimes\Sigma))\right]. \quad (15)$$

С помощью (4) и (6, равенство (15) выражается в терминах кумулянтов:

$$\bar{M}_2(\nu S) = K_2(\nu S) = \frac{1}{n}K_4(X) + \frac{1}{n-1}\left[\Sigma\otimes\Sigma + I_{p,p}(\Sigma\otimes\Sigma)\right].$$

В случае нормального вектора X, $K_4(X) = 0$. и

$$\bar{M}_2(\nu S) = \frac{1}{n-1}\left[\Sigma\otimes\Sigma + I_{p,p}(\Sigma\otimes\Sigma)\right].$$

5. Третий момент ковариационной матрицы

Выражение третьего центрального момента выборочной ковариационной матрицы S получается из определений (2):

$$\begin{aligned}\bar{M}_3(\nu S) &= E\left[(\nu S - \nu\Sigma)\otimes(\nu S - \nu\Sigma)'\otimes(\nu S - \nu\Sigma)\right] = \\ &= E\left[\nu S\otimes(\nu S)'\otimes\nu S\right] - E\left[\nu S\otimes(\nu\Sigma)'\otimes\nu S\right] - E\left[\nu\Sigma\otimes(\nu S)'\otimes\nu S\right] - \\ &- E\left[\nu S\otimes(\nu S)'\otimes\nu\Sigma\right] + 2\nu\Sigma\otimes(\nu\Sigma)'\otimes\nu\Sigma. \quad (16)\end{aligned}$$

Математическое ожидание $E\left[\nu S\otimes(\nu S)'\otimes\nu S\right]$ может быть получено непосредственно в результате довольно громоздких вычислений. В асимптотических исследованиях, например, в разложениях типа Эджворта, необходим лишь главный член $\bar{M}_3(\nu S)$, выражение которого мы здесь и выводим.

Математические ожидания (кроме первого) выражаются в фор-

муле (16) через второй момент $E[vS(vS)']$. Поскольку главный член $\bar{M}_3(vS)$ имеет порядок n^{-2}, то формула (14) для $E[vS(vS)']$ используется в виде:

$$E[vS\otimes(vS)'] = \frac{1}{n}\bar{M}_4 + \frac{n-1}{n}v\Sigma(v\Sigma)' + \frac{1}{n^2}[\Sigma\otimes\Sigma + I_{p,p}(\Sigma\otimes\Sigma)] + O(n^{-3}). \quad (17)$$

Учитывая, что

$$vS\otimes(v\Sigma)'\otimes vS = vS\otimes vS\otimes(v\Sigma)' = v[vS\otimes(vS)']\otimes(v\Sigma)',$$

то используя (17), получим от (16):

$$\bar{M}_3(vS) = E[vS\otimes(vS)'\otimes vS] - \frac{1}{n}[v\bar{M}_4\otimes(v\Sigma)' + v\Sigma\otimes\bar{M}_4 + \bar{M}_4\otimes v\Sigma] -$$

$$- \frac{n-3}{n}v\Sigma\otimes(v\Sigma)'\otimes v\Sigma - \frac{1}{n^2}[v(\Sigma\otimes\Sigma)\otimes(v\Sigma)' + v\Sigma\otimes\Sigma\otimes\Sigma +$$

$$+ \Sigma\otimes\Sigma\otimes v\Sigma + v[I_{p,p}(\Sigma\otimes\Sigma)]\otimes(v\Sigma)' + v\Sigma\otimes[I_{p,p}(\Sigma\otimes\Sigma)] +$$

$$+ [I_{p,p}(\Sigma\otimes\Sigma)]\otimes v\Sigma] + O(n^{-3}). \quad (18)$$

Для $E[vS\otimes(vS)'\otimes vS]$ выводим только те члены, которые порядка 1, $\frac{1}{n}$ или $\frac{1}{n^2}$.

С помощью (11) получим после умножения:

$$E[vS\otimes(vS)'\otimes vS] = E[A_1 - A_2 - A_3 - A_4 + A_5 + A_6 + A_7 - A_8], \quad (19)$$

где

$$A_1 = \frac{1}{(n-1)^3}\Big[\sum_i v(X_iX_i')\Big]\otimes\Big[\sum_i v(X_iX_i')\Big]'\otimes\Big[\sum_i v(X_iX_i')\Big];$$

$$A_2 = \frac{n}{(n-1)^3}\Big[\sum_i v(X_iX_i')\Big]\otimes\Big[\sum_i v(X_iX_i')\Big]'\otimes v(\bar{X}\bar{X}');$$

$$A_3 = \frac{n}{(n-1)^3}\Big[\sum_i v(X_iX_i')\Big]\otimes[v(\bar{X}\bar{X}')]'\otimes\Big[\sum_i v(X_iX_i')\Big];$$

$$A_4 = \frac{n}{(n-1)^3}v(\bar{X}\bar{X}')\otimes\Big[\sum_i v(X_iX_i')\Big]'\otimes\Big[\sum_i v(X_iX_i')\Big];$$

$$A_5 = \frac{n^2}{(n-1)^3}\Big[\sum_i v(X_iX_i')\Big]\otimes[v(\bar{X}\bar{X}')]'\otimes v(\bar{X}\bar{X}');$$

$$A_6 = \frac{n^2}{(n-1)^3}v(\bar{X}\bar{X}')\otimes\Big[\sum_i v(X_iX_i')\Big]'\otimes v(\bar{X}\bar{X}');$$

$$A_7 = \frac{n^2}{(n-1)^3}\, v(\bar{X}\bar{X}') \otimes \left[v(\bar{X}\bar{X}')\right]' \otimes \left[\sum_i v(X_i X_i')\right];$$

$$A_8 = \frac{n^3}{(n-1)^3}\, v(\bar{X}\bar{X}') \otimes \left[v(\bar{X}\bar{X}')\right]' \otimes v(\bar{X}\bar{X}').$$

Отсюда

$$EA_1 = \frac{1}{(n-1)^3}\, E \sum_i \sum_j \sum_k v(X_i X_i') \otimes \left[v(X_j X_j')\right]' \otimes v(X_k X_k') .$$

Обозначим $v(X_i X_i') = v_i$. Тогда

$$EA_1 = \frac{1}{(n-1)^3}\, E\Big\{\sum_i v_i \otimes v_i' \otimes v_i + \sum_{i\neq j}\sum (v_i \otimes v_i' \otimes v_j +$$

$$+ v_i \otimes v_j' \otimes v_i + v_j \otimes v_i' \otimes v_i) + \sum_{i\neq j\neq k}\sum\sum v_i \otimes v_j' \otimes v_k\Big\}.$$

Применяя равенства

$$v_i = X_i \otimes X_i, \quad v_i \otimes v_i = X_i \otimes X_i \otimes X_i \otimes X_i = v(X_i \otimes X_i \otimes X_i' \otimes X_i'),$$

получим

$$EA_1 = \frac{1}{(n-1)^3}\Big\{n\bar{M}_6^* + n(n-1)\left[\bar{M}_4 \otimes v\Sigma + v\bar{M}_4 \otimes (v\Sigma)' + v\Sigma \otimes \bar{M}_4\right] +$$

$$+ n(n-1)(n-2)v\Sigma \otimes (v\Sigma)' \otimes v\Sigma\Big\}.$$

Переписываем последнее равенство, рассматривая коэффициенты как многочлены $\frac{1}{n}$ с точностью $O(n^{-3})$, например, $\frac{n(n-1)}{(n-1)^3} = \frac{1}{n} + \frac{2}{n^2} + O(n^{-3})$:

$$EA_1 = \frac{1}{n^2}\bar{M}_6^* + \left(\frac{1}{n} + \frac{2}{n^2}\right)\left[\bar{M}_4 \otimes v\Sigma + v\bar{M}_4 \otimes (v\Sigma)' + v\Sigma \otimes \bar{M}_4\right] +$$

$$+ \left(1 - \frac{1}{n^2}\right)v\Sigma \otimes (v\Sigma)' \otimes v\Sigma + O(n^{-3}) .$$

Учитывая, что от 4 возможных комбинаций индексов iiij и от 6 возможных комбинаций индексов iijk, только размещения ijii, jiii и jkii дадут ненулевые математические ожидания, получим для A_2:

$$EA_2 = \frac{1}{n(n-1)^3} E \sum_i \sum_j \sum_k \sum_l v_i \otimes v_j' \otimes X_k \otimes X_l =$$

$$= \frac{1}{n(n-1)^3} E \Big\{ \sum_i v_i \otimes v_i' \otimes X_i \otimes X_i + \sum_{i \neq j} \sum (v_i \otimes v_j' \otimes X_i \otimes X_i +$$

$$+ v_j \otimes v_i' \otimes X_i \otimes X_i) + \sum_{i \neq j} \sum (v_i \otimes v_i' \otimes X_j \otimes X_j + v_i \otimes v_j' \otimes I_i \otimes X_j +$$

$$+ v_i \otimes v_j' \otimes X_j \otimes X_i) + \sum_{i \neq j \neq k} \sum \sum v_i \otimes v_j' \otimes X_k \otimes X_k \Big\}.$$

Окончательно

$$EA_2 = \frac{1}{n^2} \left[v\bar{M}_4 \otimes (v\Sigma)' + v\Sigma \otimes \bar{M}_4 + \bar{M}_4 \otimes v\Sigma + \right.$$

$$\left. + v\bar{M}_3 \otimes \bar{M}_3' + I_{p^2,p^2}(\bar{M}_3' \otimes v\bar{M}_3) \right] + \frac{1}{n} v\Sigma \otimes (v\Sigma)' \otimes v\Sigma + O(n^{-3}) .$$

Подобные выражения выводятся для EA_3 и EA_4, различными оказываются лишь слагаемые, содержащие третие моменты.

$$EA_3 = \frac{1}{n^2} \left[v\bar{M}_4 \otimes (v\Sigma)' + v\Sigma \otimes \bar{M}_4 + \bar{M}_4 \otimes v\Sigma + \right.$$

$$\left. + \bar{M}_3 \otimes \bar{M}_3 + I_{p^2,p^2}(\bar{M}_3 \otimes \bar{M}_3) \right] + \frac{1}{n} v\Sigma \otimes (v\Sigma)' \otimes v\Sigma + O(n^{-3}) ,$$

$$EA_4 = \frac{1}{n^2} \left[v\bar{M}_4 \otimes (v\Sigma)' + v\Sigma \otimes \bar{M}_4 + \bar{M}_4 \otimes v\Sigma + \right.$$

$$\left. + \bar{M}_3' \otimes v\bar{M}_3 + I_{p^2,p^2}(v\bar{M}_3 \otimes \bar{M}_3') \right] + \frac{1}{n} v\Sigma \otimes (v\Sigma)' \otimes v\Sigma + O(n^{-3}) .$$

Для A_5 получим:

$$EA_5 = \frac{1}{n^2(n-1)^3} E \sum_i \sum_j \sum_k \sum_l \sum_s v_i \otimes X_j' \otimes X_k' \otimes X_l \otimes X_s =$$

$$= \frac{1}{n^2(n-1)^3} E \Big\{ \sum_{i \neq j \neq k} \sum (v_i \otimes X_j' \otimes X_j' \otimes X_k \otimes X_k +$$

$$+ v_i \otimes X_j' \otimes X_k' \otimes X_j \otimes X_k + v_i \otimes X_j' \otimes X_k' \otimes X_k \otimes X_j) \Big\} + O(n^{-3}) =$$

$$= \frac{1}{n^2} \left[v\Sigma \otimes (v\Sigma)' \otimes v\Sigma + v\Sigma \otimes \Sigma \otimes \Sigma + v\Sigma \otimes [I_{p,p}(\Sigma \otimes \Sigma)] \right] + O(n^{-3}).$$

Аналогично

$$EA_6 = \frac{1}{n^2(n-1)^3} E\Big\{\sum_{i\neq j\neq k}\sum\sum (v_i'\otimes X_j\otimes X_j\otimes X_k\otimes X_k +$$

$$+ v_i'\otimes X_j\otimes X_k\otimes X_j\otimes X_k + v_i'\otimes X_j\otimes X_k\otimes X_k\otimes X_j)\Big\} + O(n^{-3}) =$$

$$= \frac{1}{n^2}\big[v\Sigma\otimes(v\Sigma)'\otimes v\Sigma + (v\Sigma)'\otimes v(\Sigma\otimes\Sigma) +$$

$$+ (v\Sigma)'\otimes v[I_{p,p}(\Sigma\otimes\Sigma)]\big] + O(n^{-3}),$$

$$EA_7 = \frac{1}{n^2(n-1)^3} E\Big\{\sum_{i\neq j\neq k}\sum\sum (X_i\otimes X_i\otimes X_j'\otimes X_j'\otimes v_k +$$

$$+ X_i\otimes X_j\otimes X_i'\otimes X_j'\otimes v_k + X_i\otimes X_j\otimes X_j'\otimes X_i'\otimes v_k)\Big\} + O(n^{-3}) =$$

$$= \frac{1}{n^2}\big[v\Sigma\otimes(v\Sigma)'\otimes v\Sigma + \Sigma\otimes\Sigma\otimes v\Sigma + [I_{p,p}(\Sigma\otimes\Sigma)]\otimes v\Sigma\big] + O(n^{-3}).$$

Математическое ожидание EA_8 в главный член $\bar{M}_3(vS)$ своего вклада не вносит.

Суммируя $EA_1 - EA_7$, согласно (19) получим:

$$E\big[vS\otimes(vS)'\otimes vS\big] = v\Sigma\otimes(v\Sigma)'\otimes v\Sigma + \frac{1}{n}\big[\bar{M}_4\otimes v\Sigma + v\bar{M}_4\otimes(v\Sigma)' +$$

$$+ v\Sigma\otimes\bar{M}_4 - 3v\Sigma\otimes(v\Sigma)'\otimes v\Sigma\big] + \frac{1}{n^2}\big[\bar{M}_6^* - \bar{M}_4\otimes v\Sigma -$$

$$- v\bar{M}_4\otimes(v\Sigma)' - v\Sigma\otimes\bar{M}_4 - v\bar{M}_3\otimes\bar{M}_3' - \bar{M}_3\otimes\bar{M}_3 - \bar{M}_3'\otimes v\bar{M}_3 -$$

$$- I_{p^2,p^2}(v\bar{M}_3\otimes\bar{M}_3' + \bar{M}_3\otimes\bar{M}_3 + \bar{M}_3'\otimes v\bar{M}_3) + v\Sigma\otimes\Sigma\otimes\Sigma +$$

$$+ v\Sigma\otimes[I_{p,p}(\Sigma\otimes\Sigma)] + (v\Sigma)'\otimes v(\Sigma\otimes\Sigma) + (v\Sigma)'\otimes v[I_{p,p}(\Sigma\otimes\Sigma)] +$$

$$+ \Sigma\otimes\Sigma\otimes v\Sigma + [I_{p,p}(\Sigma\otimes\Sigma)]\otimes v\Sigma + 3v\Sigma\otimes(v\Sigma)'\otimes v\Sigma\big] + O(n^{-3}).$$

Подставляя последний результат в равенство (18), получим для главного члена $\bar{M}_3(vS)$ сравнительно несложное выражение:

$$\bar{M}_3(vS) = \frac{1}{n^2}\big[\bar{M}_6^* - v\bar{M}_4\otimes(v\Sigma)' - W - I_{p^2,p^2}W\big] + O(n^{-3}), \qquad (20)$$

где

$$W = \bar{M}_4\otimes v\Sigma + v\bar{M}_3\otimes\bar{M}_3' + \bar{M}_3\otimes\bar{M}_3 + \bar{M}_3'\otimes v\bar{M}_3 - v\Sigma\otimes(v\Sigma)'\otimes v\Sigma .$$

6. Четвертый момент ковариационной матрицы

Главный член четвертого момента выборочной ковариационной матрицы получим из соотношения (6) между четвертым моментом и кумулянтом.

$$\bar{M}_4(vS) = K_4(vS) + (I_{p^4} + I_{p^2,p^2})[\bar{M}_2(vS) \otimes \bar{M}_2(vS)] + \\ + v\bar{M}_2(vS)\,[v\bar{M}_2(vS)]' . \qquad (21)$$

Кумулянты порядка r элементов vS выражаются в терминах кумулянтов исходного вектора X, притом полученное выражение относительно n порядка $O(n^{-(r-1)})$ (см.[7],[8]). Следовательно, $K_4(vS) = O(n^{-3})$ и главный член $\bar{M}_4(vS)$, который порядка n^{-2}, является функцией только от $\bar{M}_2(vS)$.

Используя (15), найдем:

$$\bar{M}_2(vS) \otimes \bar{M}_2(vS) = \frac{1}{n^2}[\bar{M}_4 \otimes \bar{M}_4 - \bar{M}_4 \otimes v\Sigma \otimes (v\Sigma)' - v\Sigma \otimes (v\Sigma)' \otimes \bar{M}_4 + \\ + v\Sigma \otimes (v\Sigma)' \otimes v\Sigma \otimes (v\Sigma)'] + O(n^{-3}),$$

$$v\bar{M}_2(vS) = \frac{1}{n}\,[v\bar{M}_4 - v[v\Sigma \otimes (v\Sigma)']] + O(n^{-2}) = \\ = \frac{1}{n}\,[v\bar{M}_4 - v\Sigma \otimes v\Sigma] + O(n^{-2}),$$

$$v\bar{M}_2(vS) \otimes [v\bar{M}_2(vS)]' = \frac{1}{n^2}\,[v\bar{M}_4 \otimes (v\bar{M}_4)' - v\bar{M}_4 \otimes (v\Sigma \otimes v\Sigma)' - \\ - (v\Sigma \otimes v\Sigma) \otimes (v\bar{M}_4)' + v\Sigma \otimes (v\Sigma)' \otimes v\Sigma \otimes (v\Sigma)'] + O(n^{-3}).$$

В результате получим:

$$\bar{M}_4(vS) = \frac{1}{n^2}\,\gamma + O(n^{-3}), \qquad (22)$$

где

$$\gamma = (I_{p^4} + I_{p^2,p^2})(\bar{M}_4 \otimes \bar{M}_4 - \bar{M}_4 \otimes v\Sigma \otimes (v\Sigma)' - \\ - v\Sigma \otimes (v\Sigma)' \otimes \bar{M}_4 + v\Sigma \otimes (v\Sigma)' \otimes v\Sigma \otimes (v\Sigma)') + \\ + v\bar{M}_4 \otimes (v\bar{M}_4)' - v\bar{M}_4 \otimes (v\Sigma \otimes v\Sigma)' - \\ - v\Sigma \otimes v\Sigma \otimes (v\bar{M}_4)' + v\Sigma \otimes (v\Sigma)' \otimes v\Sigma \otimes (v\Sigma)' . \qquad (23)$$

7. Асимптотические результаты для ковариационной матрицы

Известно, что элементы выборочной ковариационной матрицы распределены асимптотически нормально. Для νS этот результат представляется (см. [3], [6]) в виде:

$$\sqrt{n}\,\nu(S-\Sigma) \xrightarrow{\mathcal{D}} N(0,\beta)\,, \tag{24}$$

где

$$\beta = \bar{M}_4 - \nu\Sigma(\nu\Sigma)'\,. \tag{25}$$

Из равенств для моментов νS (15), (20), (22) получим предельные значения для моментов случайного вектора $\xi = \sqrt{n}\,\nu(S-\Sigma)$ при $n \to \infty$:

$$\bar{M}_2(\xi) \to \beta\,, \quad \bar{M}_3(\xi) \to 0\,, \quad \bar{M}_4(\xi) \to \Psi\,.$$

При нормальном векторе ξ между Ψ и β должно иметь место равенство

$$\Psi = (I_{p^4} + I_{p^2,p^2})\beta \otimes \beta + \nu\beta(\nu\beta)'\,.$$

Соотношение (21) согласуется с этим требованием.

Распределение (24) вырожденное нормальное распределение. Для получения невырожденного распределения следует смотреть треугольную часть от S, обозначим ее ΔS. Тогда $\frac{1}{2}p(p+1)$-вектор

$$\nu(\Delta S) = J_p \cdot \nu S\,, \tag{26}$$

где J_p ($\frac{1}{2}\,p(p+1) \times p^2$)-матрица от нулей и единиц:

$$J_p = \left(\begin{array}{c|c|c|c}
\begin{matrix}1&0&\dots&0\\0&1&\dots&0\\ \dots&\dots&\dots&\dots\\0&0&\dots&1\end{matrix} & \begin{matrix}0&0&\dots&0\\0&0&\dots&0\\ \dots&\dots&\dots&\dots\\0&0&\dots&0\end{matrix} & \dots & \begin{matrix}0&0&\dots&0\\0&0&\dots&0\\ \dots&\dots&\dots&\dots\\0&0&\dots&0\end{matrix} \\ \hline
\begin{matrix}0&0&\dots&0\\0&0&\dots&0\\ \dots&\dots&\dots&\dots\\0&0&\dots&0\end{matrix} & \begin{matrix}0&1&0\dots&0\\0&0&1\dots&0\\ \dots&\dots&\dots&\dots\\0&0&0\dots&1\end{matrix} & \dots & \begin{matrix}0&0&\dots&0\\0&0&\dots&0\\ \dots&\dots&\dots&\dots\\0&0&\dots&0\end{matrix} \\ \hline
\dots & \dots & \dots & \dots \\ \hline
\begin{matrix}0&0&\dots&0\end{matrix} & \begin{matrix}0&0&\dots&0\end{matrix} & \dots & \begin{matrix}0&0&\dots&1\end{matrix}
\end{array}\right)$$

p строк; $(p-1)$ строк; одна строка

p столбцов $\quad$ p столбцов $\quad$ p столбцов

Моменты для $v(\Delta S)$ следуют от формул (7). Для предельного распределения $v(\Delta S)$ получим:

$$\sqrt{n}\, v(\Delta S - \Delta\Sigma) \xrightarrow{\mathcal{L}} N(0, \beta_J),$$

где $(\frac{1}{2}p(p+1) \times \frac{1}{2}p(p+1))$-матрица

$$\beta_J = J_p \beta J_p' .$$

Автор выражает свою признательность доценту Э.-М. Тийт и к.ф.-м.н. Т.Колло за полезные замечания в ходе подготовки статьи.

Литература

I. Кендалл М., Стьюарт А., Теория распределений. М., 1966.

2. Колло Т., Некоторые понятия матричного исчисления с применением в математической статистике. Труды ВЦ ТГУ, 1977, 40, 30-51.

3. Колло Т., О предельном распределении выборочной ковариационной матрицы. Труды ВЦ ТГУ, 1978, 42, 3-19.

4. Крамер Г., Математические методы статистики. М., 1975.

5. Ланкастер П., Теория матриц. М., 1978.

6. Парринг А.-М., Вычисление асимптотических характеристик функции выборки. Уч. зап. ТГУ, 1979, 492, 86-90.

7. Cook, M.B., Bi-variate k-statistics and cumulants of their joint sampling distribution. Biometrika, 1951, 38, 179-195.

8. Kaplan, E.L., Tensor notation and the sampling cumulante of k-statistics. Biometrika, 1952, 39, 319-323.

9. Neudecker, H., Some theorems on matrix differentiation with special reference to Kronecker matrix products. J. Amer. Statist. Assoc., 1969, 64, 953-963.

10. Traat, I., The asymptotic normal distribution of the sample roots for a nonnormal population. Уч. зап. ТГУ, 1984, 685, 14-20.

ПРЕДСТАВЛЕНИЕ НЕИЗВЕСТНЫХ РАСПРЕДЕЛЕНИЙ СТАТИСТИК С ПОМОЩЬЮ СМЕСИ НОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

И. Траат

Многие выборочные функции - статистики - распределены асимптотически нормально со сравнительно простым выражением дисперсии. Использование предельного распределения для проверки гипотез или для построения доверительных интервалов было бы удобно, но, к сожалению, распределение статистики сходится медленно к своему предельному распределению.

В последнее время очень развивалось представление функции распределения статистики в виде ряда Эджворта, который состоит из функции предельного нормального распределения и ее производных, умноженных на кумулянты рассматриваемой статистики. Ряд упорядочен по степеням объема выборки - n, так что при $n \to \infty$, ряд сходится к нормальной функции распределения. Исследования показывают, что использование членов до порядка $n^{-1/2}$ (включительно) в ряде Эджворта улучшает точность оценивания истинного распределения статистики. Для дальнейшего повышения точности надо в ряде Эджворта учитывать члены со степенями n^{-1}, но это ведет к слишком сложным вычислениям.

В настоящей статье предлагается приблизительное распределение для статистик в виде смеси двух нормальных распре-

делений. При образовании смеси требуется, чтобы ее первые четыре кумулянта совпадали с соответствующими кумулянтами (или только главными членами) рассматриваемой статистики. Преимуществом такого приближения является простота функции распределения, значения которой вычислимы только с помощью таблицы стандартного нормального распределения. Исследуется и сходимость смеси к нормальному закону при увеличении объема выборки.

1. Определение смеси нормальных распределений с заданными первыми кумулянтами

Пусть X,Y случайные величины с соответственными распределениями $N(\mu_1,\sigma^2)$, $N(\mu_2,\sigma^2)$. Тогда функция распределения их смеси Z выражается формулой:

$$F_Z(x) = \gamma\Phi\left(\frac{x-\mu_1}{\sigma}\right) + (1-\gamma)\Phi\left(\frac{x-\mu_2}{\sigma}\right) =$$

$$= \gamma\Phi\left(\frac{x-\mu_1}{\sigma}\right) + (1-\gamma)\Phi\left(\frac{x-\mu_1}{\sigma} + \frac{\mu_1-\mu_2}{\sigma}\right), \qquad (1)$$

где $0 \leqslant \gamma \leqslant 1$ и $\Phi(x)$ - функция стандартного нормального распределения. Проблема состоит в определении параметров γ, μ_1, μ_2, σ, исходя из известных кумулянтов $\varkappa_r$, $r = 1,\ldots,4$, случайной величины Z.

Из (1) следует выражение для моментов Z:

$$EZ^r = \gamma EX^r + (1-\gamma)EY^r,$$

откуда с написанием $EZ^r = E[(Z-EZ) + EZ]^r$ выводятся формулы для центральных моментов. Учитывая выражения моментов нормальных случайных величин X,Y и связи между моментами и

кумулянтами

$$EZ = \varkappa_1; \; E(Z-\varkappa_1)^r = \varkappa_r, \; r = 2,3;$$

$$E(Z-\varkappa_1)^4 = \varkappa_4 + 3\varkappa_2^2$$

после элементарных преобразований получаются уравнения:

$$\varkappa_1 = \gamma(\mu_1 - \mu_2) + \mu_2 \tag{2}$$

$$\varkappa_2 = \gamma(1-\gamma)(\mu_1 - \mu_2)^2 + \sigma^2 \tag{3}$$

$$\varkappa_3 = \gamma(1-\gamma)(1-2\gamma)(\mu_1 - \mu_2)^3 \tag{4}$$

$$\varkappa_4 = 6\gamma(1-\gamma)\left(\frac{3+\sqrt{3}}{6} - \gamma\right)\left(\frac{3-\sqrt{3}}{6} - \gamma\right)(\mu_1 - \mu_2)^4 \tag{5}$$

Из уравнений (2)-(5) видно, что если z имеет нормальное распределение ($\varkappa_3 = 0$, $\varkappa_4 = 0$), то $\gamma = 0$ или $\gamma = 1$ или $\mu_1 = \mu_2$. В последнем случае γ произвольный. Если z имеет симметрическое распределение, отличающее от нормального ($\varkappa_3 = 0$, $\varkappa_4 \neq 0$), то $\gamma = \frac{1}{2}$ [1]. Если z имеет несимметрическое распределение с $\varkappa_4 = 0$, то $\gamma = \frac{3+\sqrt{3}}{6}$ или $\gamma = \frac{3-\sqrt{3}}{6}$. В случае $\varkappa_3 \neq 0$, $\varkappa_4 \neq 0$ значение γ (отличающееся в этом случае от $0, \frac{3-\sqrt{3}}{6}, \frac{1}{2}, \frac{3+\sqrt{3}}{6}, 1$) определяется из уравнений (4),(5). Используя обозначение

$$\frac{\varkappa_4^3}{\varkappa_3^4} = a < \infty, \tag{6}$$

получается равенство

---

[1] Из (5) следует, что такое распределение возможно лишь с $\varkappa_4 < 0$. Для получения приблизительного распределения для симметрических случайных величин лучше использовать функцию распределения смеси, где средние совпадают $\mu_1 = \mu_2$, а дисперсии различны (см. [3]).

$$\frac{\left[6\left(\frac{3+\sqrt{3}}{6}-\gamma\right)\left(\frac{3-\sqrt{3}}{6}-\gamma\right)\right]^3}{\gamma(1-\gamma)(1-2\gamma)^4}=\frac{\alpha_4^3}{\alpha_3^4}=a, \qquad (7)$$

откуда следует уравнение 6-й степени для определения γ:

$$f(\gamma)=(216+16a)\gamma^6-(648+48a)\gamma^5+(756+56a)\gamma^4-$$
$$-(432+32a)\gamma^3+(126+9a)\gamma^2-(18+a)\gamma+1=0. \qquad (8)$$

Поскольку $f(0)=1$, $f(\frac{1}{2})=-\frac{1}{8}$, $f(1)=1$, то уравнение (8) имеет по крайней мере 2 решения в интервале $(0,1)$. Видно, что если γ является решением, то и $1-\gamma$ будет решением.

Остальные параметры смеси определяются из (2), (3) с помощью обозначения

$$\bar{\mu}=\mu_1-\mu_2 \qquad (9)$$

по следующим формулам:

$$\mu_1=\alpha_1+(1-\gamma)\bar{\mu}, \qquad (10)$$

$$\mu_2=\alpha_1-\gamma\bar{\mu}, \qquad (11)$$

$$\sigma^2=\alpha_2-\gamma(1-\gamma)\bar{\mu}^2. \qquad (12)$$

Разность $\bar{\mu}$ определяется из

$$\bar{\mu}=\sqrt[3]{\frac{\alpha_3}{\gamma(1-\gamma)(1-2\gamma)}}, \qquad (13)$$

или при $\alpha_3=0$, $\gamma=\frac{1}{2}$ из

$$\bar{\mu}=\sqrt[4]{\frac{\alpha_4}{6\gamma(1-\gamma)\left(\frac{3+\sqrt{3}}{6}-\gamma\right)\left(\frac{3-\sqrt{3}}{6}-\gamma\right)}} \qquad (14)$$

При $\gamma=0$ или $\gamma=1$ разность $\bar{\mu}$ может быть произвольной, $\mu_1=\mu_2$ определяются непосредственно из (10), (11).

В дальнейшем выводится условие для кумулянтов $\varkappa_2$, $\varkappa_3$, $\varkappa_4$, достаточное для того, чтобы $\sigma^2 \in (0, \varkappa_2]$. Обозначим

$$\varkappa_2 - \sigma^2 = y , \tag{15}$$

то из уравнения (3) следует

$$\bar{\mu}^2 = \frac{y}{\gamma(1-\gamma)} . \tag{16}$$

Из (4) и (16) получается равенство

$$\gamma(1-\gamma) = \frac{y^3}{4y^3 + \varkappa_3^2} . \tag{17}$$

Переписывая (5) в виде

$$\varkappa_4 = \gamma(1-\gamma)\left[1 - 6\gamma(1-\gamma)\right]\bar{\mu}^4 ,$$

можем с помощью (16), (17) получить кубическое уравнение относительно y:

$$f(y) = 2y^3 + \varkappa_4 y - \varkappa_3^2 = 0 . \tag{18}$$

Поскольку $f(0) = -\varkappa_3^2 \leqslant 0$, для существования решения $y \in [0, \varkappa_2)$ должно иметь место неравенство $f(\varkappa_2) = 2\varkappa_2^3 + \varkappa_4\varkappa_2 - \varkappa_3^2 > 0$. Полученное условие

$$2 + \frac{\varkappa_4}{\varkappa_2^2} - \frac{\varkappa_3^2}{\varkappa_2^3} > 0 \tag{19}$$

является (см., например, [1], стр. 61) необходимым условием для кумулянтов произвольного распределения.

Следовательно, для несимметрической случайной величины Z всегда существует распределение в виде (1), которое является приближением настоящего распределения Z с точностью совпадения первых четырех кумулянтов.

2. Представление распределений статистик с помощью смеси и асимптотическое поведение смеси

Пусть $X_1,\dots,X_n$ - выборка объема n случайной величины X. Тогда X_i можно рассматривать как независимые, одинаково распределенные случайные величины с кумулянтами

$$\varkappa_r(X_i) = \varkappa_r(X) = k_r, \quad r = 1,2,\dots .$$

Если Z - некоторая функция выборки, тогда ее кумулянты $\varkappa_r(Z)$ определяемы функциями от исходных кумулянтов k_r. Исходя из $\varkappa_r(Z)$, $r = 1,\dots,4$, можно построить приблизительную функцию распределения для Z в виде (1).

Если Z сходится по распределению к нормальному закону, то соответственно (1) это может случиться путем $\gamma \to 0$, $\gamma \to 1$ или $\mu_1 - \mu_2 \to 0$. При выборочных функциях реализуется последний способ, поскольку их кумулянты зависят от n следующим образом:

$$\varkappa_3 = O(n^{-1/2}), \quad \varkappa_4 = O(n^{-1}) .$$

Отсюда $a = O(n^{-1}) \to 0$, при $n \to \infty$. Тогда соответственно (7) $\gamma \to \frac{3+\sqrt{3}}{6}$ $(1-\gamma \to \frac{3-\sqrt{3}}{6})$ и соответственно (13) $\mu_1 - \mu_2 \to 0$.

В качестве примера образуем функцию распределения (1) для статистики

$$Z = \sum_{i=1}^{n} \frac{X_i - k_1}{\sqrt{nk_2}} , \qquad (20)$$

где предполагается, что X_i несимметричны, $k_3 \neq 0$.

С помощью свойств кумулянтов получаются:

$$\varkappa_1(Z) = 0, \quad \varkappa_2(Z) = 1,$$

$$\varkappa_3(Z) = n^{-1/2} k_3/k_2^{3/2}, \quad \varkappa_4(Z) = n^{-1} k_4/k_2^2.$$

Следовательно, используя

$$a = n^{-1} k_4^3 / k_3^4 ,$$

из (8) вычисляется параметр смеси $\gamma = \gamma_0$. Остальные параметры для распределения (1) выражаются с помощью (10)-(13) следующим образом:

$$\bar{\mu} = n^{-1/6} \left[\frac{k_3}{k_2^{3/2} \gamma_0 (1-\gamma_0)(1-2\gamma_0)} \right]^{1/3} = n^{-1/6} c , \quad c < \infty ,$$

$$\mu_1 = n^{-1/6} (1-\gamma_0) c , \tag{21}$$

$$\mu_2 = n^{-1/6} \gamma_0 c , \tag{22}$$

$$\sigma^2 = 1 - n^{-1/3} \gamma_0 (1-\gamma_0) c^2 . \tag{23}$$

Известно, что распределение статистики (20) сходится к $\Phi(x)$, при $n \to \infty$. Из (21), (22) следует, что при $n \to \infty$, $\mu_1, \mu_2 \to 0$ со скоростью $n^{-1/6}$; $\sigma^2 \to 1$, со скоростью $n^{-1/3}$ и следовательно распределение (1) тоже сходится к $\Phi(x)$.

3. Распределение собственных значений выборочной ковариационной матрицы

Пусть $\mathbf{X}$ p-мерный случайный вектор с $EX = \mu$, $DX = \Sigma$. Выборочной ковариационной матрицей является

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})' ,$$

где X_i - вектор наблюдений, $\bar{X}$ - выборочное среднее. Обозначаем $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$ - собственные значения матрицы Σ и $l_1 > l_2 > \ldots > l_p$ - соответствующие собственные значения мат-

рицы S. Тогда распределение статистики

$$z_r = \sqrt{n}\ (l_r - \lambda_r),\ r = 1,\dots,p\ ,$$

представимо формулой (1), с точностью совпадения главных членов первых четырех кумулянтов. Параметры распределения (1) определяются формулами (7)-(13), где выражения для кумулянтов $\varkappa_1,\dots,\varkappa_4$, соответственно [2] следующие:

$$\varkappa_1 = n^{-1/2} \sum_{j\neq r} \lambda_{rj}(k_{22}^{rj} + \lambda_r\lambda_j) + O(n^{-3/2}),$$

$$\varkappa_2 = k_4^r + 2\lambda_r^2 + 2n^{-1}b_2 + O(n^{-2})\ ,$$

$$\varkappa_3 = n^{-1/2}\left[k_6^r+12\lambda_r k_4^r+4(k_3^r)^2+8\lambda_r^3\right] + 6\sum_{j\neq r} \lambda_{rj}(k_{31}^{rj})^2 + O(n^{-3/2}),$$

$$\varkappa_4 = 24n^{-1}b_4 + O(n^{-2}),$$

$$b_2 = -\tfrac{1}{2}k_4^r + \sum_{j\neq r} \lambda_{rj}\left[k_{42}^{rj}+\lambda_j k_4^r+5\lambda_r k_{22}^{rj}+2k_3^r k_{12}^{rj}+2(k_{21}^{rj})^2+2\lambda_r^2\lambda_j\right] -$$

$$- \sum_{j\neq r} \lambda_{rj}^2\left[(k_{22}^{rj}+\lambda_r\lambda_j)(k_4^r+2\lambda_r^2-2k_{22}^{rj}-\lambda_r\lambda_j)+2(k_{31}^{rj})^2-2k_{31}^{rj}k_{13}^{rj}\right] +$$

$$+ \sum_{j\neq k\neq r} \lambda_{rj}\lambda_{rk}\left[k_{31}^{rj}k_{112}^{rjk}+k_{31}^{rk}k_{121}^{rjk}+2(k_{211}^{rjk})^2\right],$$

$$b_4 = \tfrac{1}{24}\left[k_8^r+24\lambda_r k_6^r+32k_5^r k_3^r+32(k_4^r)^2+144\lambda_r^2 k_4^r+96\lambda_r(k_3^r)^2+48\lambda_r^4\right] +$$

$$+ \sum_{j\neq r} \lambda_{rj}k_{31}^{rj}(k_{51}^{rj}+8\lambda_r k_{31}^{rj}+4k_3^r k_{21}^{rj}) -$$

$$- \sum_{j\neq r} \lambda_{rj}^2(k_{31}^{rj})^2(k_4^r+2\lambda_r^2-3k_{22}^{rj}-2\lambda_r\lambda_j) +$$

$$+ 3\sum_{j\neq k\neq r} \lambda_{rj}\lambda_{rk}k_{31}^{rj}k_{31}^{rk}k_{211}^{rjk}\ .$$

Здесь $\lambda_{rj} = (\lambda_r - \lambda_j)^{-1}$, а $k_{s\dots t}^{i\dots j} = k_{s\dots t}(u_i,\dots,u_j)$ - кумулянт главных компонент: $u_r = \xi_r'(X-\mu)$, где ξ_r собственный вектор, соответствующий значению λ_r.

Литература

1. Рао С.Р., Линейные статистические методы и их применения. М., 1968.
2. Fujikoshi,Y., Asymptotic expansions for the distributions of the sample roots under nonnormality. Biometrika, 1980, 67, 1, 45-51.
3. Tiit,E., Traat,I., Experimental designing for Monte-Carlo study in multivariate statistics. Уч.зап.ТГУ, 1984, 685, 39-55.

ПОВЕДЕНИЕ МНОЖЕСТВЕННОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ В ЗАВИСИМОСТИ ОТ КОРРЕЛЯЦИЙ МЕЖДУ РЕГРЕССОРАМИ

Э. Силлат, Э.-М. Тийт

В анализе данных для оценивания качества регрессионной модели обычно применяется множественный коэффициент корреляции К. Для исследования поведения К при некоторых трансформациях элементов r_{ij} матрицы R мы исходим из таких форм матрицы, для которых теоретическое вычисление значения К сравнительно просто (см. напр. [1,2]).

В настоящей заметке

1) доказывается, что некоторое "циклическое" изменение всех коэффициентов корреляции не изменяет значения К,

2) демонстрируется путем статистического моделирования, что сложение "равномерного" шума всем коэффициентам корреляции заметно изменяет К, притом возможно и положительное смещение.

1. Пусть корреляционная матрица регрессоров X_i ($i=1,\dots,k$) имеет следующую "циклическую" форму:

$$R = \begin{pmatrix} 1 & \alpha_1 & \dots & \alpha_{k-1} \\ \alpha_{k-1} & 1 & \dots & \alpha_{k-2} \\ \dots & \dots & \dots & \dots \\ \alpha_1 & \alpha_2 & \dots & 1 \end{pmatrix}, \qquad (1)$$

где $\alpha_i = \alpha_{k-i}$. Предположим, что все регрессоры одинаково коррелированы с регрессандом Y:

$$r(X_i, Y) = \beta; \quad B = \underbrace{(\beta, \ldots, \beta)}_{k}' . \tag{2}$$

Если дополнительно

$$\alpha_i = \alpha \quad (i=1, \ldots, k-1), \tag{3}$$

то имеет место равенство[1] (см. [1])

$$K^2 = \frac{k\beta^2}{1 + (k-1)\alpha} . \tag{4}$$

Докажем, что формула (4) сохраняется и в случае, когда (3) не выполнено. Обозначим

$$\alpha = \frac{1}{k-1} \sum_{i=1}^{k-1} \alpha_i . \tag{5}$$

В таком случае обратная матрица R^{-1} матрицы (1) имеет также циклическую форму:

$$R^{-1} = \frac{1}{D} \begin{pmatrix} A_0 & A_1 & \cdots & A_{k-1} \\ A_{k-1} & A_0 & \cdots & A_{k-2} \\ \cdots & \cdots & \cdots & \cdots \\ A_1 & A_2 & \cdots & A_0 \end{pmatrix},$$

$A_i = A_{k-i}$.

Отсюда вытекает равенство

$$\Big(\sum_{i=0}^{k-1} \alpha_i\Big)\Big(\sum_{i=0}^{k-1} A_i\Big) = \sum_{i=0}^{k-1} \alpha_i A_i = D , \tag{6}$$

где $\alpha_0 = 1$. Учитывая определение множественного коэффициента корреляции K и предположение (2) имеем:

$$K^2 = B'R^{-1}B = \frac{k\beta^2}{D} \sum A_i ,$$

---

[1] Здесь и в дальнейшем предполагается, что K существует и выполняется неравенство $|K| \leq 1$, т.е. R неотрицательно определенная.

откуда при помощи равенств (5) и (6) вытекает (4), что и требовалось доказать.

Заметим, что приведенный результат сохраняется и для более общего случая, когда корреляционная матрица имеет структуру латинского квадрата (число разных внедиагональных элементов равняется $k-1$; каждый из них встречается точно в раз в каждой строке и в каждом столбце).

2. Далее рассмотрим случай, когда матрица R имеет блочную структуру:

$$\left\{\begin{array}{l} R = \left(\begin{array}{c|c} R(\alpha_1,k_1) & \Gamma(k_1,k_2) \\ \hline \Gamma(k_2,k_1) & R(\alpha_2,k_2) \end{array}\right); \\ R(\alpha_i,k_i) = \underbrace{\left(\begin{array}{cccc} 1 & \alpha_i & \cdots & \alpha_i \\ \alpha_i & 1 & \cdots & \alpha_i \\ \cdots & \cdots & \cdots & \cdots \\ \alpha_i & \alpha_i & \cdots & 1 \end{array}\right)}_{k_i}; \\ \Gamma(k_i,k_j) = \left.\left(\begin{array}{cccc} \gamma & \gamma & \cdots & \gamma \\ \gamma & \gamma & \cdots & \gamma \\ \cdots & \cdots & \cdots & \cdots \\ \gamma & \gamma & \cdots & \gamma \end{array}\right)\right\} k_i \\ \qquad\qquad \underbrace{\qquad\qquad}_{k_j} \end{array}\right. . \tag{7}$$

Не предполагается и выполненным условие (2).

В таком случае для K^2 имеется равенство (см. [2])

$$K^2 = \frac{\overline{\beta}_1^2\{[1+(k_1-1)\alpha_1]k_2\} + \overline{\beta}_2^2\{[1+(k_2-1)\alpha_2]k_1\} - 2\overline{\beta}_1\overline{\beta}_2 k_1 k_2 \gamma}{(1+(k_1-1)\alpha_1)(1+(k_2-1)\alpha_2) - k_1 k_2 \gamma^2} +$$

$$+ \frac{k_1 DB_1}{1-\alpha_1} + \frac{k_2 DB_2}{1-\alpha_2}, \tag{8}$$

где $B = (\underbrace{B_1}_{k_1} : \underbrace{B_2}_{k_2}) = (\beta_1^1,\ldots,\beta_{k_1}^1 : \beta_1^2,\ldots,\beta_{k_2}^2)$; $\bar\beta_j = \frac{1}{k_j}\sum_{i=1}^{k_j}\beta_i^1$;

$DB_j = \frac{1}{k_j}\sum_{i=1}^{k_j}(\beta_i^j - \bar\beta_j)^2$, $j=1,2$.

Мы рассмотрели в качестве примера случай со следующими параметрами:

$$\begin{aligned} k &= 10 \quad \alpha_1 = 0.35 \quad \beta_1=\beta_2=\beta_3=\beta_4=0.4 \\ k_1 &= 7 \quad \alpha_2 = 0.25 \quad \beta_5=\beta_6=\beta_7=0 \\ k_2 &= 3 \quad \gamma = 0.15 \quad \beta_8=\beta_9=\beta_{10}=0.4 \ . \end{aligned} \tag{9}$$

По формуле (8) найдем, что в таком случае

$K^2 = 0.771606$, $K = 0.878411$.

Заметим, что такой выбор параметров определяет "почти наилучшую модель" из всех потенциальных моделей, полученных путем выбора 10 регрессоров из 60 регрессоров, корреляции которых определяются формулами (7) при параметрах:

$$\begin{aligned} k &=60 \quad \alpha_1=0.35 \quad \beta_1=\ldots=\beta_{15}=0.4 \quad \beta_{31}=\ldots=\beta_{45}=0.4 \\ k_1=k_2&=30 \quad \alpha_2=0.25 \quad \beta_{16}=\ldots=\beta_{30}=0 \quad\quad \beta_{46}=\ldots=\beta_{60}=0 \ . \\ &\quad\ \gamma=0.15 \end{aligned}$$

Наилучшая модель определяется комплектом параметров

$k_1=8$, $k_2=2$, $\beta_1=\ldots=\beta_5=0.4$; $\beta_6=\beta_7=\beta_8=0$; $\beta_9=\beta_{10}=0.4$,

в таком случае $K = 0.881783$.

На ЭВМ были определены генератор случайных чисел z_1,

$$z_1 \sim U(-0.05\ ,\ 0.05), \tag{10}$$

и последовательность корреляционных матриц R_h $(h=1,\ldots,N)$. $R_h = (r_{ij}^h)$ (симметрические) по формуле:

$$r_{ij}^h = r_{ij} + z_1 \quad (i=2,\ldots,k;\ j=1,\ldots,i-1),$$

где r_{ij} – элемент матрицы (7) при комплекте параметров (9), а z_1 случайное число (10).

Для каждой матрицы R_n вычислился коэффициент корреляции K_h.

Результаты моделирования в случае N = 100 представлены на рис. 1.

Отсюда видно, что при данном примере оценка K по среднему моделированных оценок имеет положительное смещение с величиной 0.8938 – 0.8784 = 0.0154 .

Вторым неожиданным результатом является заметная положительная асимметрия распределения R_h. Заметим, что выборочный коэффициент корреляции при положительном теоретическом коэффициенте корреляции имеет левую асимметрию.

Из последнего результата вытекает и неожиданно большое значение максимума K_h. Заметим, что по данному примеру

$$\max_h (K_h) = 0.9639 \approx \bar{K} + 3.4\,\sigma(K_h),$$

где $\bar{K}$ – среднее всех эмпирических оценок и $\sigma(K_h)$ – их стандартное отклонение. Заметим, что этот эффект может вызвать переоценку качества модели при определении модели путем выбора "наилучшей эмпирической модели".

Заметим, что приближенное моделирование выборочных корреляционных матриц путем прибавления теоретической корреляционной матрице независимого равномерно распределенного шума, является, конечно, весьма грубым приближением. Однако найденных эффектов (несимметричность распределения K и неожиданно большой максимум) трудно объяснить особенностями моделирования, так как полученное распределение коэффициентов корреляции является симметричным и ограниченным.

Распределение множественного коэффициента корреляции $\tilde{R}$ при объеме выборки $n = 100$.

$\min \tilde{R} = 0.732$

$\max \tilde{R} = 0.929$

$E\tilde{R} = 0.799$; $D\tilde{R} = 0.00106$; $\sqrt{D\tilde{R}} = 0.0326$

Таблица частот:

i	Границы классов a_i	Частоты классов $[a_i, a_{i+1})$
0	0.732	1
1	0.752	2
2	0.771	17
3	0.791	25
4	0.811	24
5	0.831	13
6	0.850	14
7	0.870	1
8	0.890	1
9	0.909	1
10	0.929	1

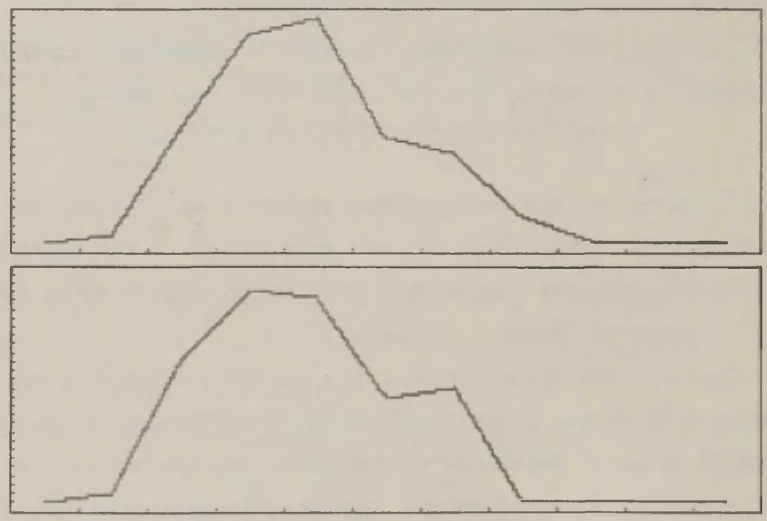


Рис. 1. Графики эмпирической плотности $\tilde{R}^2$ (наверху) и $\tilde{R}$ (внизу).

Литература

1. Тийт Э., Выбор моделей в линейном регрессионном анализе. Труды ВЦ ТГУ, 1981, 46, 60-84.
2. Тийт Э., Класс допустимых эмпирических моделей в линейном регрессионном анализе. Вычислительные системы 88, Новосибирск, 1981, 99-107.

ПОСТРОЕНИЕ ДИСКРЕТНЫХ МНОГОМЕРНЫХ РАСПРЕДЕЛЕНИЙ С ЗАДАННЫМИ МОМЕНТАМИ. ДИСКРЕТНЫЙ АНАЛОГ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Э.-М. Тийт

§1. Постановка проблемы

1.1. В настоящее время в прикладной и вычислительной статистике проблема тестирования программ и алгоритмов является весьма актуальной. Для эффективного решения этой проблемы необходимо иметь достаточно богатый комплект "пробных распределений", удовлетворяющих следующим условиям:

1° Распределения определены (нетривиальным образом) для любой размерности m.

2° Для этих распределений существует большое количество свободно задаваемых параметров (напр. маргинальные и смешанные моменты, коэффициенты корреляции и т.д.).

3° На основании этих распределений легко найти точные значения разных статистик, характеризующих исследуемые статистические процедуры.

4° Существуют эффективные алгоритмы для генерирования выборок реализаций случайных векторов, имеющих определенное распределение.

1.2. Одним из возможных решений поставленной проблемы является применение в качестве "пробных распределений" дискретных распределений, имеющих заданный набор моментов. Такое семейство дискретных распределений имеет некоторые преимущества перед непрерывными распределениями:

1^{0} В случае дискретного распределения возможно конструировать "генеральную совокупность", имеющую точно заданное распределение. Это целесообразно для вычисления точных (теоретических) значений исследуемых характеристик (статистик).

2^{0} Пользуясь дискретными аналогами непрерывных распределений возможно исследовать влияние дискретности исходного распределения. Так как изучаемые в анализе данных признаки часто дискретны, то этот вопрос имеет большое значение.

3^{0} Симулирование дискретных распределений является простым и эффективным.

1.3. В настоящей статье излагается методика определения одномерных дискретных распределений по заданным моментам (§2), многомерных дискретных распределений по заданным маргинальным распределениям и корреляционной матрице (§3) и многомерных дискретных распределений по заданным смешанным моментам (§4–5). В качестве иллюстрации изложения конструируется дискретный аналог нормального вектора, т.е. дискретное распределение, имеющее все моменты до четвертого порядка (включая) равные соответствующим моментам нормального распределения (§6). Важнейшие конструкции иллюстрированы примерами.

§2. Определение дискретного одномерного распределения по заданным моментам

2.1. В настоящем параграфе рассматривается класс $\mathcal{P}$ одномерных конечных распределений P, определенных функцией вероятности

$$\{p_i, a_i;\ i=1,\dots,n\}, \tag{1}$$

где $A=\{a_i\}$ - носитель меры P, $\varkappa(A)=n$. Точки $a_1,\dots,a_n$ и вероятности $p_1,\dots,p_n$ называются определяющими параметрами распределения P. Степенью сложности $\tau(P)$ распределения P называется число свободно меняющихся (выбираемых) определяющих параметров этого распределения. Ввиду условия

$$\sum_{i=1}^{n} p_i = 1 \tag{2}$$

всегда $\tau(P) \leq 2n-1$. В случае дополнительных условий, фиксирующих часть определяющих параметров, например, точки $a_1,\dots,a_l$ $(l \leq n)$, степень сложности уменьшается и $\tau(P) \leq 2n-l-1$. В случае симметричного распределения $\tau(P) \leq n-1$ (тогда $a_i = -a_{n-i}$, $p_i = p_{n-i}$, $i=1,\dots,n$).

Если задано некоторое семейство $\mathcal{P}^*$, $\mathcal{P}^* \subset \mathcal{P}$, то наиболее простым распределением P^* из этого семейства считается тот, который имеет минимальную степень сложности:

$$\tau(P^*) = \min_{P \in \mathcal{P}^*} \tau(P).$$

Наиболее простое распределение семейства не всегда определяется однозначно.

Нашей целью является нахождение наиболее простых распределений, моменты которых имеют заданные значения.

2.2. У каждого распределения P из семейства $\mathcal{P}$ существуют все моменты M_h, h=1,2,... , и они выражаются следующей суммой:

$$M_h = \sum_{i=1}^{n} p_i a_i^h \,. \qquad (3)$$

Напомним, что центральные моменты $\bar{M}_h = E(X-M_1)^h$ распределения P связаны соотношением (см. [1], стр. 61)

$$\begin{pmatrix} 1 & 0 & \dots & \bar{M}_h \\ 0 & \bar{M}_2 & \dots & \bar{M}_{h+1} \\ \dots & \dots & \dots & \dots \\ \bar{M}_h & \bar{M}_{h+1} & \dots & \bar{M}_{2h} \end{pmatrix} \geq 0, \qquad (h=2,3,\dots) \qquad (4)$$

а смешанные моменты вектора удовлетворяют неравенству Коши-Буняковского

$$(EXY)^2 \leq EX^2 \cdot EY^2 \,. \qquad (5)$$

В случае симметричного распределения все нечетные моменты равняются нулю:

$$M_{2h+1} = 0 \,. \qquad (h=0,1,\dots) \qquad (6)$$

2.3. Пусть заданы постоянные $K_1,\dots,K_k$, удовлетворяющие условию (4), и требуется построить такое распределение P, чтобы имелись место равенства

$$M_h = K_h \quad (h=1,\dots,k).$$

Учитывая равенство (3), мы получим систему уравнений

$$\sum_{i=1}^{n} p_i a_i^h = K_h \quad (h=1,\dots,k), \qquad (7)$$

где неизвестными являются точки a_i и вероятности p_i. Решение системы (7) определяет распределение P тогда и только тогда, когда выполнено равенство (2) и

$$p_i \in [0,1], \quad a_i \in \mathbb{R}_1 \quad (i=1,\dots,n). \qquad (8)$$

Наиболее простое распределение, определенное условиями (7), имеет в общем случае степень сложности $\tau(P) = k$ и мощность $\varkappa(\mathcal{A}) = n = \left[\frac{k+2}{2}\right]$ (здесь $[x]$ - целая часть числа x).

В симметричном случае из k равенств системы (7) $\left[\frac{k+1}{2}\right]$ вытекают непосредственно из равенств (6). Оставшиеся $\left[\frac{k}{2}\right]$ равенств определяют такое же количество параметров и $\tau(P)=\left[\frac{k}{2}\right]$.

В общем случае возможно найти решение системы (7) и (2) в области (8), пользуясь методом, изложенным в [2].

Для практически интересных случаев $k \le 4$ выведено решение системы (7), выражаемое через заданные константы K_j. Этими доказывается существование наиболее простых дискретных распределений для всех удовлетворяющих условию (4) комплектов заданных моментов K_j.

2.4. В дальнейшем будем пользоваться хорошо известной возможностью стандартизации любого распределения P путем линейного преобразования множества $\mathcal{A}$, т.е. построением распределения P_0 с нулевым математическим ожиданием и единичной дисперсией:

$$P_0 = \{p_1, \alpha+\beta a_1;\ i=1,\dots,n\},\quad M_1(P_0) = 0,\ M_2(P_0) = 1, \qquad (9)$$

где $\alpha = -M_1(P)/\sqrt{M_2(P)-M_1^2(P)}$, $\beta = 1/\sqrt{M_2(P)-M_1^2(P)}$.

Аналогично возможно из P_0 получить и распределение $P(K_1,K_2)$ с заданными моментами K_1 и K_2, а высшие моменты распределений P и P_0 (см. формулы (1) и (9)) связаны соотношением

$$M_h(P_0) = \sum_{l=0}^{h} C_l^h \alpha^{h-l}\beta^l M_l(P).$$

Учитывая вышесказанное в дальнейшем без ограничения общности будем считать, что $K_1=0$ и $K_2=1$, значит, будем определять желаемые дискретные распределения P в стандартизованной форме.

ТАБЛИЦА 1

Определение наиболее простого дискретного распределения по заданным моментам K_1, K_2, K_3 и K_4.

№	Заданные моменты	Тип распределения: n	симметричность	$0 \in \Lambda$	$\tau(P)$	Выражения определяющих параметров*
1	$K_1=0$	1	+	+	0	$p_1=1, \quad a_1=0$
2	$K_1\neq0$	1	–	–	1	$p_1=1, \quad a_1=K_1$
3	$K_1=0$ K_2	2	+	–	1	$p_1=0.5, \quad a_1=-\sqrt{K_2}$ $p_2=0.5, \quad a_2=\sqrt{K_2}$
4	$K_1\neq0$ K_2 $(K_1^2\leq K_2)$	2	–	+	2	$p_1=K_1^2/K_2$ $a_1=K_2/K_1$ $p_2=1-K_1^2/K_2, \quad a_2=0$
5	$K_1=0$ $K_2=1$ K_3	2	–	–	3	$p_1=0.5(1-\lvert K_3\rvert/A)$ $a_1=\operatorname{sgn}(K_3)\sqrt{(A+\lvert K_3\rvert)/(A-\lvert K_3\rvert)}$ $p_2=0.5(1+\lvert K_3\rvert/A)$ $a_2=-\operatorname{sgn}(K_3)\sqrt{(A-\lvert K_3\rvert)/(A+\lvert K_3\rvert)}$
6	$K_1=0$ $K_2=1$ $K_3=0$ K_4 $(K_4\geq1)$	3	+	+	2	$p_1=1/2K_4$ $a_1=-\sqrt{K_4}$ $p_2=1/2K_4$ $a_2=\sqrt{K_4}$ $p_3=1-1/K_4, \quad a_3=0$
7	$K_1=0$ $K_2=1$ K_3 K_4 $(K_4\geq K_3^2+1)$	3	–	+	4	$p_1=0.25(2B-\lvert K_3\rvert)/(B\cdot\lvert K_4-K_3^2\rvert)$ $a_1=\operatorname{sgn}(K_3)(0.5\lvert K_3\rvert+B)$ $p_2=0.25(2B+\lvert K_3\rvert)/(B\cdot\lvert K_4-K_3^2\rvert)$ $a_2=-\operatorname{sgn}(K_3)(0.5\lvert K_3\rvert-B)$ $p_3=1-1/(K_4-K_3^2), \quad a_3=0$

* Здесь $A=\sqrt{4+K_3^2}$, $B=\sqrt{K_4-0.75K_3^2}$.

2.5. Пусть заданы $K_1=0$, $K_2=1$, K_3 и K_4. Система уравнений (7) позволяет в данном случае определить две точки a_1 и a_2 и их вероятности p_1 и p_2 при условии, что точка a_3 фиксирована. Ввиду условия $K_1=0$ целесообразно выбрать $a_3=0$, тогда система (7) имеет вид:

$$\begin{cases} p_1a_1 + p_2a_2 = 0 \\ p_1a_1^2 + p_2a_2^2 = 1 \\ p_1a_1^3 + p_2a_2^3 = K_3 \\ p_1a_1^4 + p_2a_2^4 = K_4 \end{cases} \qquad (7')$$

Эта система имеет решение всегда, когда выполнено условие

$$K_4 \geq K_3^2 + 1,$$

вытекающее из (4). Решение системы (7') приведено в таблице 1 (случай №7). Кроме того в таблице изложены практически интересные частные случаи системы (7') (№5 и 6) и всевозможные типы наиболее простых распределений при $n \leq 2$ (№1-4).

ПРИМЕР 1

Вычислим числовые значения определяющих параметров распределений 3-7 из таблицы 1 по заданным конкретным значениям параметров K_1.

ТАБЛИЦА 2

Конкретные значения определяющих параметров наиболее простых распределений, вычисленных по заданным моментам.

№	Заданные моменты	Определяющие параметры	№	Заданные моменты	Определяющие параметры
3	$K_1=0$ $K_2=3$	$p_1=0.5$ $a_1=-1.73205$ $p_2=0.5$ $a_2=1.73205$	6	$K_1=0$ $K_2=1$ $K_3=0$ $K_4=6$	$p_1=0.08333$ $a_1=-2.44949$ $p_2=0.08333$ $a_2=2.44949$ $p_3=0.83333$ $a_3=0$
4	$K_1=1$ $K_2=3$	$p_1=0.33333$ $a_1=3$ $p_2=0.66667$ $a_2=0$	7	$K_1=0$ $K_2=1$ $K_3=-2$ $K_4=6$	$p_1=0.10566$ $a_1=-2.73205$ $p_2=0.39434$ $a_2=0.73205$ $p_3=0.5$ $a_3=0$
5	$K_1=0$ $K_2=1$ $K_3=-2$	$p_1=0.85355$ $a_1=0.41421$ $p_2=0.14645$ $a_2=-2.41421$			

§3. Определение дискретного многомерного распределения с заданными маргинальными распределениями и корреляционной матрицей

3.1. Рассмотрим класс $\mathcal{P}^m$ m-мерных дискретных распределений P, все маргинальные распределения P_i (i=1,...,m) которого равны. Пусть R - заданная m×m-матрица, имеющая свойства корреляционной матрицы.

В таком случае с помощью алгоритма, изложенной в [3], возможно проверять, принадлежит ли R в класс C корреляционных матриц, конструируемых при помощи смесей векторов, и если это так, то найти и соответствующую конструкцию. Маргинальные распределения возможно определить по результатам §2.

Если наиболее простое одномерное распределение P_1 ($P_1 \in \mathcal{P}^*$) имеет носитель A_1, $\varkappa(A_1)=n$, то наиболее простое невырожденное m-мерное распределение P с маргинальными распределениями P_i имеет носитель A, $\varkappa(A) \leq n^m$.

3.2. Продемонстрируем определение наиболее простого дискретного многомерного распределения с заданными маргинальными распределениями и заданной корреляционной матрицей на конкретном примере.

ПРИМЕР 2

Пусть требуется построить наиболее простое 4-мерное распределение, имеющее фиксированные маргинальные моменты $K_1=K_3=0$ и $K_2=K_4=1$ и корреляционную матрицу R:

$$R = \begin{pmatrix} 1 & 0.3 & 0.1 & 0.5 \\ 0.3 & 1 & 0.4 & 0.2 \\ 0.1 & 0.4 & 1 & 0.1 \\ 0.5 & 0.2 & 0.1 & 1 \end{pmatrix}.$$

Из таблицы 1 (случай №6) получим маргинальное распределение P_1:

a_1	-1	1
p_1	0.5	0.5

Следовательно, $\varkappa(\mathcal{A}) \leq 2^4 = 16$, и каждая точка множества $\mathcal{A}$ имеет форму $a=(a_1,a_2,a_3,a_4)$, где a_i равняется либо 1, либо -1. Вероятности этих точек вычисляются по корреляционной матрице R.

Следуя алгоритму, изложенному в [3], по R определяются составители Y_j смеси векторов и их вероятности γ_j $(j=1,\ldots,q)$, где $q \leq 0.5m(m-1)+1$. В данном случае $q \leq 7$.

Каждый составитель Y_j - т.н. связка - 4-мерный вектор, компоненты Y_{ji} $(i=1,\ldots,4)$ которого либо совпадают, либо являются независимыми. Для наглядности будем совпадающиеся компоненты обозначать равными значениями второго индекса, тогда имеется

$$r(Y_{ji},Y_{jh}) = \begin{cases} 1, & \text{если } i=h, \\ 0, & \text{если } i\neq h, \end{cases} \qquad (10)$$

(здесь $r(U,V)$ обозначает коэффициент корреляции между случайными величинами U и V). Представим найденную смесь в таблице 3.

ТАБЛИЦА 3

Составители Y_j и вероятности γ_j смеси, определяющей корреляционную матрицу R.

	№ составителя					
	1	2	3	4	5	6
Вероятность	0.1	0.3	0.2	0.1	0.1	0.2
1. компонент	Y_{11}	Y_{21}	Y_{31}	Y_{41}	Y_{51}	Y_{61}
2. компонент	Y_{11}	Y_{22}	Y_{31}	Y_{42}	Y_{52}	Y_{62}
3. компонент	Y_{11}	Y_{22}	Y_{32}	Y_{43}	Y_{53}	Y_{63}
4. компонент	Y_{11}	Y_{21}	Y_{33}	Y_{43}	Y_{51}	Y_{64}

Так как $\sum_{i=1}^{q-1} \psi_i = 0.8 < 1$, то $R \in C$ и соответствующее распределение существует.

Для проверки возможно вычислить коэффициенты корреляции r_{ij} полученной смеси, пользуясь аналогом формулы (3) для моментов вектора и соотношение (10), например,

$$r_{12} = 0.1r(Y_{11},Y_{11}) + 0.3r(Y_{21},Y_{22}) + 0.2r(Y_{31},Y_{31}) +$$
$$+ 0.1r(Y_{41},Y_{42}) + 0.1r(Y_{51},Y_{52}) + 0.2r(Y_{61},Y_{62}) =$$
$$= 0.1 + 0.2 = 0.3 .$$

Учитывая определение маргинального распределения P легко выписать носители составителей Y_j $(j=1,\ldots,6)$ и их распределения. Например, составитель Y_1, имеющий 4 совпадающих компонентов, имеет следующее распределение:

a_1	$(-1,-1,-1,-1)$	$(1,1,1,1)$
p_1	0.5	0.5

,

а составитель Y_2, имеющий 2 пары совпадающих компонентов, которые взаимно ортогональны, имеет распределение:

a_1	$(1,1,1,1)$	$(1,-1,-1,1)$	$(-1,1,1,-1)$	$(-1,-1,-1,-1)$
p_1	0.25	0.25	0.25	0.25

.

Представим все распределения составителей и распределение смеси в таблице 4. В последнем столбце указаны частоты и точек A для конструирования генеральной совокупности, имеющей определенное распределение.

ТАБЛИЦА 4

Распределение составителей Y_j и смеси, имеющей корреляционную матрицу R и маргинальные моменты $K_1=K_3=0$, $K_2=K_4=1$.

№	Точки a_i	Распределения составителей						Распределение смеси	N
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6		
	вероятности τ_j	0.1	0.3	0.2	0.1	0.1	0.2		
1	1 1 1 1	0.5	0.25	0.125	0.125	0.125	0.0625	0.1875	15
2	1 1 1-1			0.125			0.0625	0.0375	3
3	1 1-1 1			0.125		0.125	0.0625	0.05	4
4	1 1-1-1			0.125	0.125		0.0625	0.05	4
5	1-1 1 1				0.125	0.125	0.0625	0.0375	3
6	1-1 1-1						0.0625	0.0125	1
7	1-1-1 1		0.25			0.125	0.0625	0.1000	8
8	1-1-1-1				0.125		0.0625	0.0250	2
9	-1 1 1 1				0.125		0.0625	0.0250	2
10	-1 1 1-1		0.25			0.125	0.0625	0.1000	8
11	-1 1-1 1						0.0625	0.0125	1
12	-1 1-1-1				0.125	0.125	0.0625	0.0375	3
13	-1-1 1 1			0.125	0.125		0.0625	0.05	4
14	-1-1 1-1			0.125		0.125	0.0625	0.05	4
15	-1-1-1 1			0.125			0.0625	0.0375	3
16	-1-1-1-1	0.5	0.25	0.125	0.125	0.125	0.0625	0.1875	15
							Σ	1.0000	80

§4. Определение дискретных многомерных распределений с заданными маргинальными распределениями и смешанными моментами

4.1. Пусть заданы дискретные одномерные распределения P_j $(j=1,\ldots,m)$ и набор постоянных $\{K^{(l)}_{j_1\ldots j_m}\}$ $(l=1,\ldots,L)$, удовлетворяющих условиям (4) и (5).

Требуется построить дискретное многомерное распределение P так, чтобы

1^0 маргинальными распределениями P были заданные распределения P_j $(j=1,\ldots,m)$,

2^0 смешанными моментами $M_{j_1\dots j_m}(P)$ распределения P были заданные постоянные $(l=1,\dots,L)$.

Для решения поставленной задачи необходимо найти носитель $\mathcal{A}$ распределения P, а затем выписать соотношения, определяющие вероятности точек a_i, $a_i \in \mathcal{A}$.

Если носитель распределения P_j есть $\mathcal{A}_j$, то $\mathcal{A}$ есть прямое произведение

$$\mathcal{A} = \mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_m.$$

Пусть $\mathcal{A}_j = \{a_i^{(j)}; i=1,\dots,n_j\}$. Обозначаем $P(a_{i_1}^{(1)},\dots,a_{i_m}^{(m)}) = p_{i_1\dots i_m}$.

4.2. Для того, чтобы было выполнено условие 1^0, где $P_j = (p_i^{(j)}, a_i^{(j)}; i=1,\dots,m_j)$, необходимо и достаточно, чтобы были выполнены следующие условия:

$$\sum_{i_1} \dots \sum_{\substack{i_m \\ i_h \neq i_j}} p_{i_1\dots i_m} = p_{i_j}^{(j)} \quad (i_j=1,\dots,n_j;\ j=1,\dots,m). \qquad (11)$$

Найти такие вероятности $\{p_{i_1\dots i_m}\}_m$, которые удовлетворяют системе (11), состоящей из N, $N = \sum_{j=1} n_j$ уравнений, всегда возможно.

4.3. Для выполнения условий 2^0 необходимо, чтобы вероятности $\{p_{i_1\dots i_m}\}$ удовлетворяли условиям

$$\sum_{i_1} \dots \sum_{i_m} a_{i_1}^{j_1} \dots a_{i_m}^{j_m} \cdot p_{i_1\dots i_m} = K_{j_1\dots j_m}^{(l)} \quad (l=1,\dots,L) \qquad (12)$$

притом требуется еще, чтобы

$$\begin{cases} \sum_{i_1} \dots \sum_{i_m} p_{i_1\dots i_m} = 1, \\ p_{i_1\dots i_m} \in [0,1]. \end{cases} \qquad (13)$$

Нахождение решения системы (11), (12) в области (13) возможно с помощью методики математического программирования. В общем случае не гарантировано ни существование решения, ни его однозначность.

4.4. Для практических задач тестирования алгоритмов и исследования свойств статистик целесообразно пользоваться распределениями, имеющими некоторые фиксированные свойства, гарантирующие их простоту. С этой целью будем вводить семейство дискретных симметричных распределений.

§5. Дискретные многомерные симметричные распределения

5.1. Пусть P – m-мерное дискретное распределение, $a = (a_1,\ldots,a_m)$ – произвольная точка его носителя $\mathcal{A}$.

Распределение P называется симметричным, если выполнены следующие условия:

1^0 $P(a) = P(-a)$,

2^0 $P(a_1,\ldots,a_m) = P(a_{i_1},\ldots,a_{i_m})$,

где $(i_1,\ldots,i_m) = Q(1,\ldots,m)$ – некоторая перестановка последовательности индексов $(1,\ldots,m)$.

Выводим некоторые простые свойства симметричного распределения в форме следствий из определения.

С.1. Все маргинальные распределения P_i симметричного распределения P равны.

С.2. Все маргинальные распределения P_i симметричного распределения P симметричны.

С.3. Все нечетные (маргинальные и смешанные) моменты симметричного распределения равняются нулю.

5.4. Все смешанные моменты симметричного распределения инвариантны относительно перестановки компонентов вектора.

5.2. Из определения и следствий вытекает, что носитель $\mathcal{A}$ симметричного распределения разлагается на множества A_i равновероятных точек,

$$\mathcal{A} = A_1 \cup \ldots \cup A_q, \quad A_i \cap A_j = \emptyset, \; i \neq j, \; i,j=1,\ldots,q ,$$

которые располагаются некоторыми свойствами симметрии.

5.3. В классе симметричных распределений возможно задавать все маргинальные и смешанные моменты до некоторого фиксированного порядка, получив таким образом практически решаемую систему (12), так как число различных моментов такого распределения не растет очень быстро (см. таблицу 5).

ТАБЛИЦА 5

Число различных моментов k-го порядка у m-мерного симметричного распределения.

Порядок момента k	Размерность вектора m							
	1	2	3	4	5	6	7	8
2	1	2	2	2	2	2	2	2
4	1	3	4	5	5	5	5	5
6	1	4	7	9	10	11	11	11
8	1	5	10	14	17	19	20	21

5.4. Определим равновероятные множества A_i для случая, когда носитель $\mathcal{A}_1$ маргинального распределения содержит три точки: $\mathcal{A}_1 = \{-C, 0, C\}$. В таком случае для m-мерного распределения $\varkappa(\mathcal{A}) = 3^m$.

Пусть $a = (a_1,\dots,a_m)$, $a \in \mathcal{A}$. Обозначаем число координат a_i, равняющих нулю, через s, а число координат, равных $-C$, через r. Остальные координаты точки a равны C. Множество всех таких точек, у которых числа координат, равных 0 и $-C$ соответственно s и r, обозначается через $A(s,r)$, где ввиду условия 1° $r := \min(r, m-s-r)$. Элементарным вычислением получается, что

$$\begin{cases} \varkappa(A(s,r)) = C_m^s \cdot C_{m-s}^r \cdot C_o, \\ \text{где} \\ C_o = \begin{cases} 1, & \text{если } 2r = m-s, \\ 2, & \text{если } 2r < m-s. \end{cases} \end{cases} \tag{14}$$

Так как при фиксированном s число разных классов $A(s,r)$ есть $\left[1 + \frac{m-s}{2}\right]$, то общее число q классов $A(s,r)$ вычисляется простым суммированием:

$$q = \sum_{s=0}^{m} \left[1 + \frac{m-s}{2}\right] = \begin{cases} 0.25(m+1)(m+3), & \text{если } m \text{ нечетное}, \\ (0.5m+1)^2, & \text{если } m \text{ четное}. \end{cases} \tag{15}$$

Значит, для определения распределения P необходимо определить вероятности $P(s,r) = P(a)$ при $a \in A(s,r)$ так, чтобы было выполнено условие

$$\begin{cases} \sum_{s=0}^{m} \sum_{r=0}^{w} P(s,r) \cdot \varkappa(A(s,r)) = 1, \\ \text{где} \quad w = [0.5(m-s)]. \end{cases} \tag{16}$$

5.5. Для того, чтобы в случае симметричного распределения выписать уравнения (11), связывающие маргинальные вероятности с вероятностями многомерного распределения, необходимо их выписать через вероятности точек a_i равновероятных подмножеств A_i. В случае $\varkappa(\mathcal{A}_1) = 3$ мы имеем следующее простое соотношение между вероятностями $P(s,r)$ в случае $(m+1)$-

мерного вектора и вероятностями $P^*(s,r)$ его m-мерного подвектора:

$$P^*(s,r)=P(s,r+1)+P(s+1,r)+P(s,r) \quad (s=1,\dots,m;r=1,\dots,w). \quad (17)$$

Повторным применением соотношения (17) имеем и равенство, связывающее вероятности (m+g)-мерного распределения P(s,r) с вероятностями $P^*(s,r)$ его m-мерного подвектора:

$$P^*(s,r) = \sum_{\nu=0}^{g} \sum_{\mu=0}^{g-\nu} P((s+\mu),(r+g-\nu-\mu)) \cdot C_g^{\nu} C_{g-\nu}^{\mu}, \quad (17')$$
$$(s=1,\dots,m;r=1,\dots,w).$$

§6. Построение дискретного аналога нормального распределения

6.1. Пользуясь методикой, изложенной в §5, определим многомерное распределение, имеющее все моменты до четвертого порядка равными соответствующим моментам нормального распределения. Значения моментов стандартизированного нормального распределения, удовлетворяющего условиям симметричности 1^0 и 2^0, заданы в таблице 6.

ТАБЛИЦА 6

Значения моментов стандартизированного нормального распределения, удовлетворяющего условиям 1^0 и 2^0.

Порядок моментов k	Размерность вектора m			
	1	2	3	4
2	$M_2 = 1$	$M_2 = 1$ $M_{11} = \varrho$	$M_2 = 1$ $M_{11} = \varrho$	$M_2 = 1$ $M_{11} = \varrho$
4	$M_4 = 3$	$M_4 = 3$ $M_{31} = 3\varrho$ $M_{22} = 1 + 2\varrho^2$	$M_4 = 3$ $M_{31} = 3\varrho$ $M_{22} = 1 + 2\varrho^2$ $M_{211} = \varrho + 2\varrho^2$	$M_4 = 3$ $M_{31} = 3\varrho$ $M_{22} = 1 + 2\varrho^2$ $M_{211} = \varrho + 2\varrho^2$ $M_{1111} = 3\varrho^2$

6.2. Так как поставленная задача имеет наибольшее практическое значение в случаях, когда размерность $m \leq 4$, то для этих случаев выводятся явные выражения для вероятностей $P(s,r)$.

Из таблицы 1 (случай №6) мы получим значение для C

$$C = \sqrt{3},$$

и найдем, что в одномерном случае распределение P_1 имеет вид:

a_i	$-\sqrt{3}$	0	$\sqrt{3}$
p_i	1/6	2/3	1/6

(18)

6.3. В двумерном случае мы имеем 9 точек, которые разбиты на 4 класса $A(s,r)$ (см. рис. 1 и табл. 7).

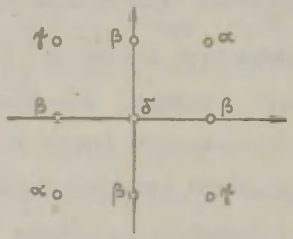


Рис. 1. Дискретный аналог нормального распределения на плоскости: α, β, γ и δ - вероятности.

Классы равновероятных точек $A(s,r)$, их типичные представители, мощности и сокращенные обозначения вероятностей приведены в таблице 7.

ТАБЛИЦА 7

Дискретный аналог нормального распределения на плоскости. Характеристика множеств $A(s,r)$ $(C = \sqrt{3})$.

№	Класс	Представитель	$\varkappa(A)$	Обозначение вероятности		Значение вероятности $\varrho \in [-0.5;0.5]$
1	A(0,0)	(C,C)	2	P(0,0)	α	$(2\varrho^2+3\varrho+1)/36$
2	A(0,1)	(-C,C)	2	P(0,1)	γ	$(2\varrho^2-3\varrho+1)/36$
3	A(1,0)	(0,C)	4	P(1,0)	β	$(-\varrho^2+1)/9$
4	A(2,0)	(0,0)	1	P(2,0)	δ	$(2\varrho^2+4)/9$

Для определения вероятностей $P(s,r)$ выпишем по формуле (17) уравнения, связывающие их с маргинальными вероятностями (18):

$$\begin{cases} P(0,0) + P(1,0) + P(0,1) = 1/6 \\ 2\cdot P(1,0) + P(0,0) = 2/3 \end{cases} \qquad (19)$$

и из условия (12) соотношения, определяющие моменты, значения которых заданы в столбце 2 таблицы 6:

$$\begin{cases} 2(P(0,0) - P(0,1))c^2 = \varrho \\ 2(P(0,0) + P(0,1))c^4 = 1 + 2\varrho^2 \\ 2(P(0,0) - P(0,1))c^4 = 3\varrho \end{cases} \qquad (19')$$

Решение системы (19) и (19') при дополнительном требовании (8) дает значения вероятностей, изложенные в последнем столбце таблицы 7. Условие (8) удовлетворяется лишь при $\varrho \in [-0.5;0.5]$ или $|\varrho| = 1$, в последнем случае распределение вырождено. При $|\varrho| > 0.5$ не существует 9-точечного дискретного аналога нормальному распределению (но может существовать в случае $\varkappa(A) > 9$).

6.4. В случае $m=3$ мы имеем $\varkappa(A)=27$, см. рис.2 и табл.8.

К соотношениям, связывающим вероятности $P(s,r)$ с маргинальными вероятностями, выражения которых уже известны (см. (19)), прибавляется еще одно уравнение, фиксирующее значение момента M_{211} (см. табл. 6, столбец 3):

$$2(P(0,0) - 3P(0,1))c^4 = 2\varrho^2 + \varrho .$$

Полученная система содержит 5 уравнений для 6 неизвестных, значит - распределение не определяется однозначно. В таблице 8 излагается два возможных решения, притом одна из вероятностей выбирается равной нулю.

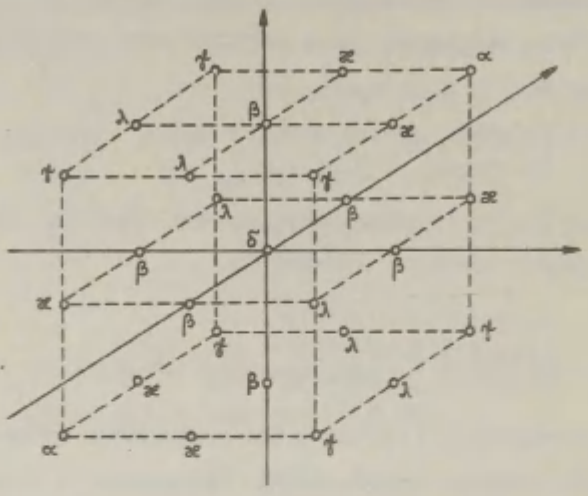


Рис. 2. Дискретный аналог нормального распределения в пространстве.

ТАБЛИЦА 8

Дискретный аналог нормального распределения в 3-мерном пространстве. Характеристики множеств A(s,r).

№	Класс	Типичный представитель	$\varkappa(A)$	Обозначение вероятности		Значение вероятности	
						$\varrho \in [0;0.5]$	$\varrho \in [-0.5;0]$
1	A(0,0)	(C,C,C)	2	P(0,0)	α	$(2\varrho^2+\varrho)/18$	0
2	A(0,1)	(-C,C,C)	6	P(0,1)	γ	0	$-(2\varrho^2+\varrho)/54$
3	A(1,0)	(0,C,C)	6	P(1,0)	$\varkappa$	$(-2\varrho^2+\varrho+1)/36$	$(10\varrho^2+11\varrho+3)/108$
4	A(1,1)	(0,-C,C)	6	P(1,1)	λ	$(2\varrho^2-3\varrho+1)/36$	$(14\varrho^2-5\varrho+3)/108$
5	A(2,0)	(0,0,C)	6	P(2,0)	β	$(-2\varrho^2+\varrho+1)/18$	$(-6\varrho^2-\varrho+1)/18$
6	A(3,0)	(0,0,0)	1	P(3,0)	δ	$(8\varrho^2-2\varrho+6)/18$	$(8\varrho^2+\varrho+3)/9$

6.5. В случае m=4 к 6 уравнениям, связывающим вероятности 4-мерного распределения, прибавляется еще одно, фиксирующее значение момента $M_{1111}(P)$ (см. табл. 6, столбец 4)

$$(2P(0,0) + 6P(0,2) - 8P(0,1))c^4 = 3\varrho^2.$$

Таким образом, мы получим семейство решений для значений $\varrho \in [-0.2;0.5]$; в таблице 9 указано два сравнительно простых решения (соответственно для положительных и отрицательных значений ϱ).

ТАБЛИЦА 9

Дискретный аналог нормального распределения в 4-мерном пространстве.

№	Класс	Типичный представитель	ж(A)	ВЕРОЯТНОСТЬ		
				Обозначение	$\varrho \in [0;0.5]$	$\varrho \in [-0.2;0]$
1	A(0,0)	(C,C,C,C)	2	P(0,0)	$\varrho^2/6$	0
2	A(0,1)	(-C,C,C,C)	8	P(0,1)	0	0
3	A(0,2)	(-C,-C,C,C)	6	P(0,2)	0	$\varrho^2/18$
4	A(1,0)	(0,C,C,C)	8	P(1,0)	$(-\varrho^2+\varrho)/18$	0
5	A(1,1)	(0,-C,C,C)	24	P(1,1)	0	$-(5\varrho^2+\varrho)/54$
6	A(2,0)	(0,0,C,C)	12	P(2,0)	$(-\varrho+1)/36$	$(20\varrho^2+13\varrho+3)/108$
7	A(2,1)	(0,0,-C,C)	12	P(2,1)	$(2\varrho^2-3\varrho+1)/36$	$(34\varrho^2-\varrho+3)/108$
8	A(3,0)	(0,0,0,C)	8	P(3,0)	$(-\varrho^2+\varrho)/16$	$-(5\varrho^2+\varrho)/6$
9	A(4,0)	(0,0,0,0)	1	P(4,0)	$(7\varrho^2-4\varrho+3)/9$	$(23\varrho^2+4\varrho+3)/9$

Продемонстрируем изложенные результаты и для некоторых конкретных значений ϱ.

21

ПРИМЕР 3

Рассмотрим построение аналога нормального распределения в 2-, 3- и 4-мерном пространстве при значениях $\varrho=0.4$ и $\varrho=-0.4$ (в 4-мерном случае второй вариант: $\varrho=-0.1$). Результаты изложены в таблице 10.

ТАБЛИЦА 10

Значения вероятностей для дискретного аналога многомерного нормального распределения.

Размерность m	ВЕРОЯТНОСТИ	
	$\varrho = 0.4$	$\varrho = -0.4$
2	P(0,0)=0.07 P(0,1)=0.0033333 P(1,0)=0.0933333 P(2,0)=0.48	P(0,0)=0.0033333 P(0,1)=0.07 P(1,0)=0.0933333 P(2,0)=0.48
3	P(0,0)=0.04 P(0,1)=0 P(1,0)=0.03 P(1,1)=0.0033333 P(2,0)=0.06 P(3,0)=0.36	P(0,0)=0 P(0,1)=0.0014815 P(1,0)=0.0018519 P(1,1)=0.067037 P(2,0)=0.0244444 P(3,0)=0.4311111
	$\varrho = 0.4$	$\varrho = -0.1$
4	P(0,0)=0.026667 P(0,1)=0 P(0,2)=0 P(1,0)=0.0133333 P(1,1)=0 P(2,0)=0.0166667 P(2,1)=0.0033333 P(3,0)=0.04 P(4,0)=0.28	P(0,0)=0 P(0,1)=0 P(0,2)=0.0005556 P(1,0)=0 P(1,1)=0.0009259 P(2,0)=0.0175926 P(2,1)=0.0318519 P(3,0)=0.0083333 P(4,0)=0.3144444

На основании этих данных легко построить генеральные совокупности, имеющие заданные распределения (см. табл. 11).

ТАБЛИЦА 11

Дискретные аналоги 3-мерного нормального распределения для $\varrho=0.4$, $\varrho=0.5$ и $\varrho=-0.5$.

i	a_i	Вероятности и частоты					
		$\varrho=0.4$		$\varrho=0.5$		$\varrho=-0.5$	
1	$\sqrt{3}, \sqrt{3}, \sqrt{3}$	0.04	12	0.05556	2	0	0
2	$-\sqrt{3},-\sqrt{3},-\sqrt{3}$	0.04	12	0.05556	2	0	0
3	$-\sqrt{3}, \sqrt{3}, \sqrt{3}$	0	0	0	0	0	0
4	$\sqrt{3},-\sqrt{3}, \sqrt{3}$	0	0	0	0	0	0
5	$\sqrt{3}, \sqrt{3},-\sqrt{3}$	0	0	0	0	0	0
6	$\sqrt{3}, \sqrt{3},-\sqrt{3}$	0	0	0	0	0	0
7	$-\sqrt{3}, \sqrt{3},-\sqrt{3}$	0	0	0	0	0	0
8	$-\sqrt{3},-\sqrt{3}, \sqrt{3}$	0	0	0	0	0	0
9	$0, \sqrt{3}, \sqrt{3}$	0.03	9	0.02778	1	0	0
10	$\sqrt{3}, 0, \sqrt{3}$	0.03	9	0.02778	1	0	0
11	$\sqrt{3}, \sqrt{3}, 0$	0.03	9	0.02778	1	0	0
12	$0,-\sqrt{3},-\sqrt{3}$	0.03	9	0.02778	1	0	0
13	$-\sqrt{3}, 0,-\sqrt{3}$	0.03	9	0.02778	1	0	0
14	$-\sqrt{3},-\sqrt{3}, 0$	0.03	9	0.02778	1	0	0
15	$0,-\sqrt{3}, \sqrt{3}$	0.00333	1	0	0	0.08333	1
16	$0, \sqrt{3},-\sqrt{3}$	0.00333	1	0	0	0.08333	1
17	$\sqrt{3}, 0,-\sqrt{3}$	0.00333	1	0	0	0.08333	1
18	$-\sqrt{3}, 0, \sqrt{3}$	0.00333	1	0	0	0.08333	1
19	$\sqrt{3},-\sqrt{3}, 0$	0.00333	1	0	0	0.08333	1
20	$-\sqrt{3}, \sqrt{3}, 0$	0.00333	1	0	0	0.08333	1
21	$0, 0, \sqrt{3}$	0.06	18	0.05556	2	0	0
22	$0, 0,-\sqrt{3}$	0.06	18	0.05556	2	0	0
23	$0, \sqrt{3}, 0$	0.06	18	0.05556	2	0	0
24	$0,-\sqrt{3}, 0$	0.06	18	0.05556	2	0	0
25	$\sqrt{3}, 0, 0$	0.06	18	0.05556	2	0	0
26	$-\sqrt{3}, 0, 0$	0.06	18	0.05556	2	0	0
27	$0, 0, 0$	0.36	108	0.38889	14	0.5	6
Σ		1.00	300	1.00000	36	1.000	12

ТАБЛИЦА 12

Наиболее простые дискретные аналоги нормального распределения при $\varrho=0.5$

Размерность	Класс A(s,r)	æ(A)	P(s,r)
1	A(0,0)	2	0.16667
	A(1,0)	1	0.66667
2	A(0,0)	2	0.08333
	A(1,0)	4	0.08333
	A(2,0)	1	0.5
3	A(0,0)	2	0.05556
	A(1,0)	6	0.02778
	A(2,0)	6	0.05556
	A(3,0)	1	0.38889
4	A(0,0)	2	0.04167
	A(1,0)	8	0.01389
	A(2,0)	12	0.01389
	A(3,0)	8	0.04167
	A(4,0)	1	0.30556
5	A(0,0)	2	0.03333
	A(1,0)	10	0.00833
	A(2,0)	20	0.00556
	A(3,0)	20	0.00833
	A(4,0)	10	0.03333
	A(5,0)	1	0.23889
6	A(0,0)	2	0.02778
	A(1,0)	12	0.02556
	A(2,0)	30	0.00278
	A(3,0)	40	0.00278
	A(4,0)	30	0.00556
	A(5,0)	12	0.02778
	A(6,0)	1	0.18333
7	A(0,0)	2	0.02381
	A(1,0)	14	0.00397
	A(2,0)	42	0.00159
	A(3,0)	70	0.00119
	A(4,0)	70	0.00159
	A(5,0)	42	0.00397
	A(6,0)	14	0.02381
	A(7,0)	1	0.13571
8	A(0,0)	2	0.02083
	A(1,0)	16	0.00298
	A(2,0)	56	0.00099
	A(3,0)	112	0.00060
	A(4,0)	140	0.00060
	A(5,0)	112	0.00099
	A(6,0)	56	0.00298
	A(7,0)	16	0.02083
	A(8,0)	1	0.09405
9	A(0,0)	2	0.01852
	A(1,0)	18	0.00231
	A(2,0)	72	0.00067
	A(3,0)	168	0.00033
	A(4,0)	252	0.00026
	A(5,0)	252	0.00033
	A(6,0)	168	0.00067
	A(7,0)	72	0.00231
	A(8,0)	18	0.01852
	A(9,0)	1	0.0570
10	A(0,0)	2	0.01667
	A(1,0)	20	0.00185
	A(2,0)	90	0.00046
	A(3,0)	240	0.00020
	A(4,0)	420	0.00013
	A(5,0)	504	0.00013
	A(6,0)	420	0.00020
	A(7,0)	240	0.00046
	A(8,0)	90	0.00185
	A(9,0)	20	0.01667
	A(10,0)	1	0.02368

ПРИМЕР 4

При предельных допустимых значениях ϱ искомые распределения получаются наиболее простыми. Например, при $\varrho = 0.5$ для каждого m имеется только $m+1$ различающихся от нуля вероятностей $P(s,0)$, $s=0,\ldots,m$, притом рекуррентным пользованием системы (17) мы имеем:

$$\begin{aligned} &æ(A(m-1,0))\cdot P(m-1,0) = 1/3,\\ &æ(A(m-2,0))\cdot P(m-2,0) = 1/6,\\ &\cdots\cdots\cdots\cdots\\ &æ(A(0,0))\cdot P(0,0) = 1/3m\ , \end{aligned}$$

откуда вытекает:

$$P(m,0) = 1 - 1/3\left(1 + \frac{1}{2} + \ldots + \frac{1}{m}\right), \qquad (20)$$

значит, наиболее простой дискретный аналог нормального распределения существует только для таких значений m, при которых левая сторона равенства (20) неотрицательна, значит для $m \leq 10$. Характеристики этих распределений заданы в табл. 12.

Литература

1. Рао С.Р., Линейные статистические методы и их применения. М., 1968.
2. Miller, A.C., Rice, T.R., Discrete approximations of probability distributions. Management Science, 1983, 29, 3, 352-362.
3. Tiit, E.-M., Definition of random vectors with given marginal distributions and given correlation matrix. Уч. зап. ТГУ, 1984, 685, 21-39.

Содержание

СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ НА ЕС ЭВМ.
Труды вычислительного центра. Выпуск 51.
На русском языке.
Тартуский государственный университет.
ЭССР, 202400. г.Тарту, ул.Юликооли, 18.
Ответственный редактор М. Вяхи.
Подписано к печати 19.11. 1984.
МВ 10534.
Формат 60x84/16.
Бумага писчая.
Машинопись. Ротапринт.
Условно-печатных листов 9,76.
Учетно-издательских листов 8,4.
Печатных листов 10,5.
Тираж 200.
Заказ № 1126.
Цена 1 руб. 30 коп.
Типография ТГУ. ЭССР, 202400, г.Тарту, ул.Пялсона, 14.

1 руб. 30 коп.