## 5.6 整合与重构

R中提供了许多用来整合（aggregate）和重塑（reshape）数据的强大方法。在整合数据时，往往将多组观测替换为根据这些观测计算的描述性统计量。在重塑数据时，则会通过修改数据的结构（行和列）来决定数据的组织方式。本节描述了用来完成这些任务的多种方式。

在接下来的两个小节中，我们将使用已包含在R基本安装中的数据框mtcars。这个数据集是从*Motor Trend* 杂志（1974）提取的，它描述了34种车型的设计和性能特点（汽缸数、排量、马力、每加仑汽油行驶的英里数，等等）。要了解此数据集的更多信息，请参阅help(mtcars)。

### 5.6.1 转置

转置（反转行和列）也许是重塑数据集的众多方法中最简单的一个了。使用函数t()即可对一个矩阵或数据框进行转置。对于后者，行名将成为变量（列）名。代码清单5-9展示了一个例子。

**代码清单5-9 数据集的转置**

```
> cars <- mtcars[1:5,1:4]
> cars
                   mpg cyl disp  hp
Mazda RX4         21.0   6  160 110
Mazda RX4 Wag     21.0   6  160 110
Datsun 710        22.8   4  108  93
Hornet 4 Drive    21.4   6  258 110
Hornet Sportabout 18.7   8  360 175
> t(cars)
     Mazda RX4 Mazda RX4 Wag Datsun 710 Hornet 4 Drive Hornet Sportabout
mpg         21            21       22.8           21.4              18.7
cyl          6             6        4.0            6.0               8.0
disp       160           160      108.0          258.0             360.0
hp         110           110       93.0          110.0             175.0
```

为了节约空间，代码清单5-9仅使用了mtcars数据集的一个子集。在本节稍后讲解reshape包的时候，你将看到一种更为灵活的数据转置方式。

### 5.6.2 整合数据

在R中使用一个或多个by变量和一个预先定义好的函数来折叠（collapse）数据是比较容易的。调用格式为：

```
aggregate(x, by, FUN)
```

其中*x*是待折叠的数据对象，*by*是一个变量名组成的列表，这些变量将被去掉以形成新的观测，而*FUN*则是用来计算描述性统计量的标量函数，它将被用来计算新观测中的值。

作为一个示例，我们将根据汽缸数和挡位数整合mtcars数据，并返回各个数值型变量的均值（见代码清单5-10）。

**代码清单5-10　整合数据**

```
> options(digits=3)
> attach(mtcars)
> aggdata <-aggregate(mtcars, by=list(cyl,gear), FUN=mean, na.rm=TRUE)
> aggdata
  Group.1 Group.2  mpg cyl disp  hp drat   wt qsec  vs   am gear carb
1       4       3 21.5   4  120  97 3.70 2.46 20.0 1.0 0.00    3 1.00
2       6       3 19.8   6  242 108 2.92 3.34 19.8 1.0 0.00    3 1.00
3       8       3 15.1   8  358 194 3.12 4.10 17.1 0.0 0.00    3 3.08
4       4       4 26.9   4  103  76 4.11 2.38 19.6 1.0 0.75    4 1.50
5       6       4 19.8   6  164 116 3.91 3.09 17.7 0.5 0.50    4 4.00
6       4       5 28.2   4  108 102 4.10 1.83 16.8 0.5 1.00    5 2.00
7       6       5 19.7   6  145 175 3.62 2.77 15.5 0.0 1.00    5 6.00
8       8       5 15.4   8  326 300 3.88 3.37 14.6 0.0 1.00    5 6.00
```

在结果中，Group.1表示汽缸数量（4、6或8），Group.2代表挡位数（3、4或5）。举例来说，拥有4个汽缸和3个挡位车型的每加仑汽油行驶英里数（mpg）均值为21.5。

在使用aggregate()函数的时候，by中的变量必须在一个列表中（即使只有一个变量）。你可以在列表中为各组声明自定义的名称，例如by=list(Group.cyl=cyl, Group.gears=gear)。指定的函数可为任意的内建或自编函数，这就为整合命令赋予了强大的力量。但说到力量，没有什么可以比reshape包更强。

### 5.6.3　reshape包

reshape包[①]是一套重构和整合数据集的绝妙的万能工具。由于它的这种万能特性，可能学起来会有一点难度。我们将慢慢地梳理整个过程，并使用一个小型数据集作为示例，这样每一步发生了什么就很清晰了。由于reshape包并未包含在R的标准安装中，在第一次使用它之前需要使用install.packages("reshape")进行安装。

大致说来，你需要首先将数据“融合”（melt），以使每一行都是一个唯一的标识符-变量组合。然后将数据“重铸”（cast）为你想要的任何形状。在重铸过程中，你可以使用任何函数对数据进行整合。将使用的数据集如表5-8所示。

**表5-8　原始数据集（mydata）**

| ID | Time | X1 | X2 |
|---|---|---|---|
| 1 | 1 | 5 | 6 |
| 1 | 2 | 3 | 5 |
| 2 | 1 | 6 | 1 |
| 2 | 2 | 2 | 4 |

① 由同一作者开发的reshape2包是reshape的重新设计版本，功能更为强大。——译者注

在这个数据集中，测量（measurement）是指最后两列中的值（5、6、3、5、6、1、2、4）。每个测量都能够被标识符变量（在本例中，标识符是指ID、Time以及观测属于X1还是X2）唯一地确定。举例来说，在知道ID为1、Time为1，以及属于变量X1之后，即可确定测量值为第一行中的5。

**1. 融合**

数据集的融合是将它重构为这样一种格式：每个测量变量独占一行，行中带有要唯一确定这个测量所需的标识符变量。要融合表5-8中的数据，可使用以下代码：

```
library(reshape)
md <- melt(mydata, id=(c("id", "time")))
```

你将得到如表5-9所示的结构。

**表5-9 融合后的数据集**

| ID | Time | 变 量 | 值 |
|---|---|---|---|
| 1 | 1 | X1 | 5 |
| 1 | 2 | X1 | 3 |
| 2 | 1 | X1 | 6 |
| 2 | 2 | X1 | 2 |
| 1 | 1 | X2 | 6 |
| 1 | 2 | X2 | 5 |
| 2 | 1 | X2 | 1 |
| 2 | 2 | X2 | 4 |

注意，必须指定要唯一确定每个测量所需的变量（ID和Time），而表示测量变量名的变量（X1或X2）将由程序为你自动创建。

既然已经拥有了融合后的数据，现在就可以使用`cast()`函数将它重铸为任意形状了。

**2. 重铸**

`cast()`函数读取已融合的数据，并使用你提供的公式和一个（可选的）用于整合数据的函数将其重塑。调用格式为：

```
newdata <- cast(md, formula, FUN)
```

其中的*md*为已融合的数据，*formula*描述了想要的最后结果，而*FUN*是（可选的）数据整合函数。其接受的公式形如：

```
rowvar1 + rowvar2 + … ~ colvar1 + colvar2 + …
```

在这一公式中，*rowvar1 + rowvar2 + ...*定义了要划掉的变量集合，以确定各行的内容，而*colvar1 + colvar2 + ...*则定义了要划掉的、确定各列内容的变量集合。参见图5-1中的示例。

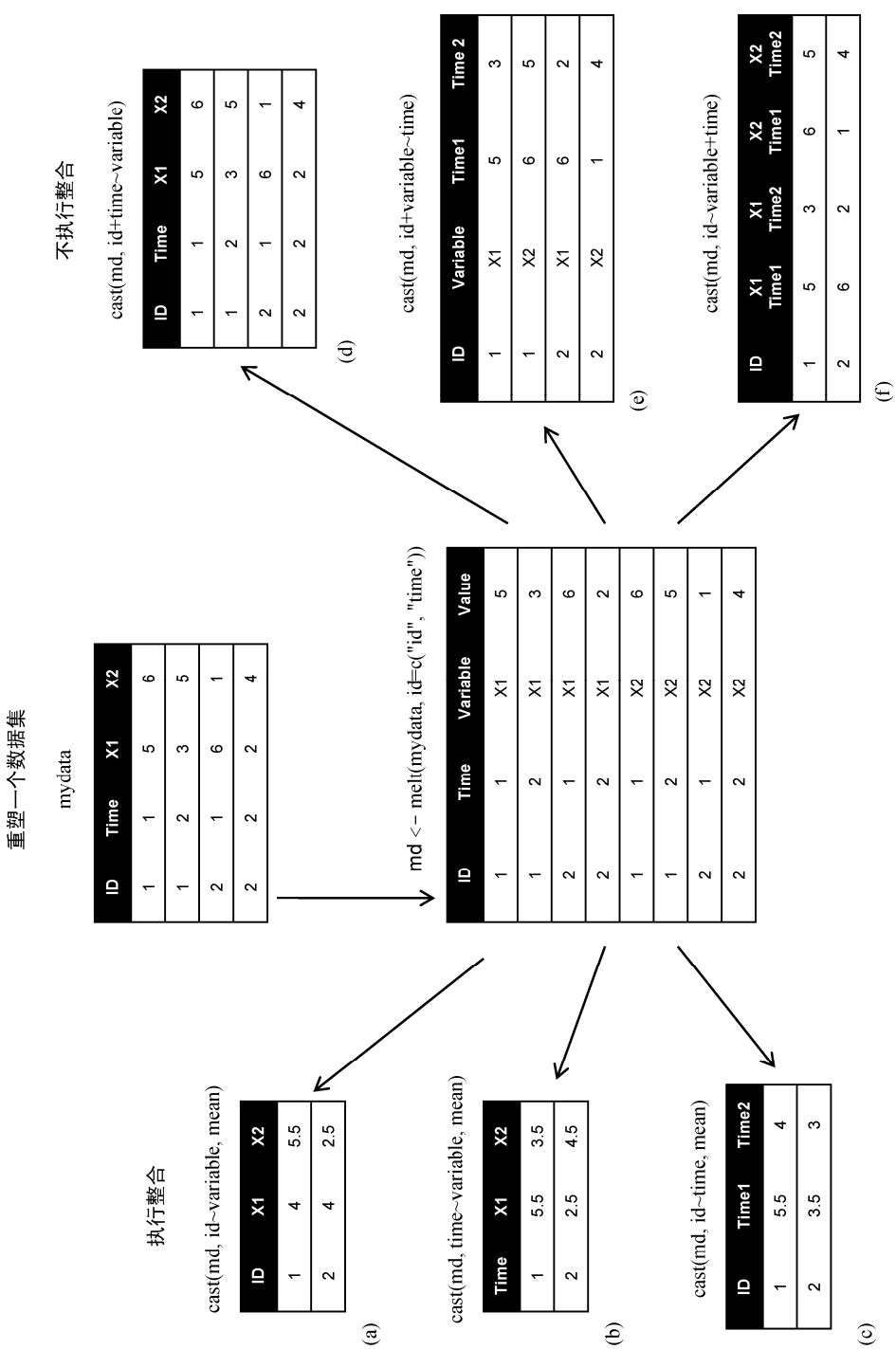


图5-1　使用函数`melt()`和`cast()`重塑数据

由于右侧（d、e和f）的公式中并未包括某个函数，所以数据仅被重塑了。反之，左侧的示例（a、b和c）中指定了mean作为整合函数，从而就对数据同时进行了重塑与整合。例如，(a)中给出了每个观测所有时刻中在X1和X2上的均值；示例(b)则给出了X1和X2在时刻1和时刻2的均值，对不同的观测进行了平均；在(c)中则是每个观测在时刻1和时刻2的均值，对不同的X1和X2进行了平均。

如你所见，函数`melt()`和`cast()`提供了令人惊叹的灵活性。很多时候，你不得不在进行分析之前重塑或整合数据。举例来说，在分析重复测量数据（为每个观测记录了多个测量的数据）时，你通常需要将数据转化为类似于表5-9中所谓的“长格式”。示例参见9.6节。

## 5.7 小结



本章总结了数十种用于处理数据的数学、统计和概率函数。我们看到了如何将这些函数应用到范围广泛的数据对象上，其中包括向量、矩阵和数据框。我们学习了控制流结构的使用方法：用循环重复执行某些语句，或用分支在满足某些特定条件时执行另外的语句。然后你编写了自己的函数，并将它们应用到了数据上。最后，我们探索了折叠、整合以及重构数据的多种方法。

既然已经集齐了数据塑形（没有别的意思）所需的工具，我们就准备好告别第一部分并进入激动人心的数据分析世界了！在接下来的几章中，我们将探索多种将数据转化为信息的统计方法和图形方法。