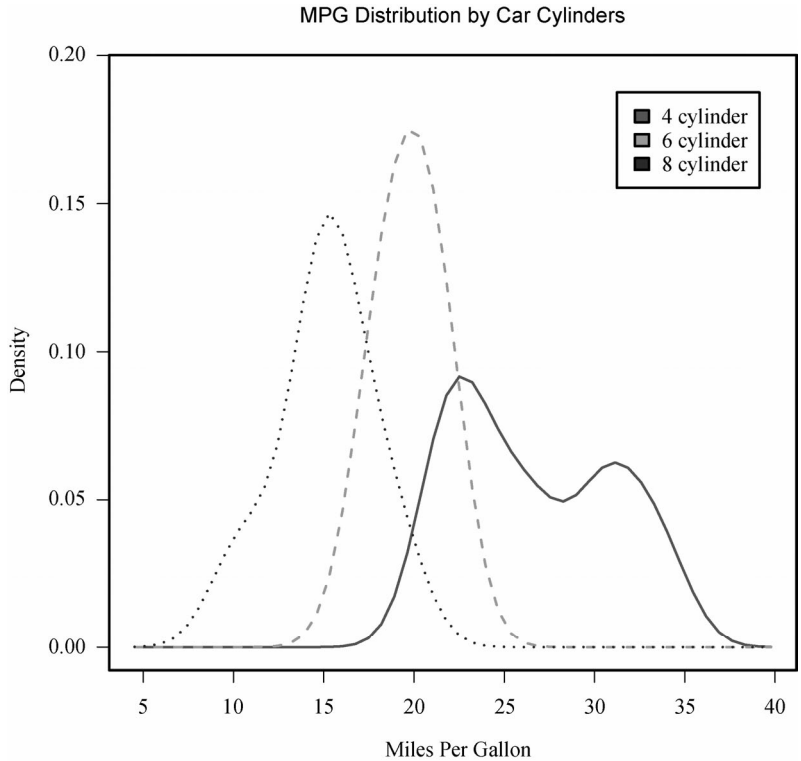

图6-10　按汽缸个数划分的各车型每加仑汽油行驶英里数的核密度图

箱线图同样是一项用来可视化分布和组间差异的绝佳图形手段（并且更常用），我们接下来讨论它。

6.5　箱线图

箱线图（又称盒须图）通过绘制连续型变量的五数总括，即最小值、下四分位数（第25百分位数）、中位数（第50百分位数）、上四分位数（第75百分位数）以及最大值，描述了连续型变量的分布。箱线图能够显示出可能为离群点（范围 ± 1.5*IQR以外的值，IQR表示四分位距，即上四分位数与下四分位数的差值）的观测。例如：

```
boxplot(mtcars$mpg, main="Box plot", ylab="Miles per Gallon")
```

生成了如图6-11所示的图形。为了图解各个组成部分，我手工添加了标注。

默认情况下，两条须的延伸极限不会超过盒型各端加1.5倍四分位距的范围。此范围以外的值将以点来表示（在这里没有画出）。

举例来说，在我们的车型样本中，每加仑汽油行驶英里数的中位数是19.2，50%的值都落在了15.3和22.8之间，最小值为10.4，最大值为33.9。我是如何从图中如此精确地读出了这些值呢？

执行boxplot.stats(mtcars$mpg)即可输出用于构建图形的统计量（换句话说，我作弊了）。图中似乎不存在离群点，而且略微正偏（上侧的须较下侧的须更长）。

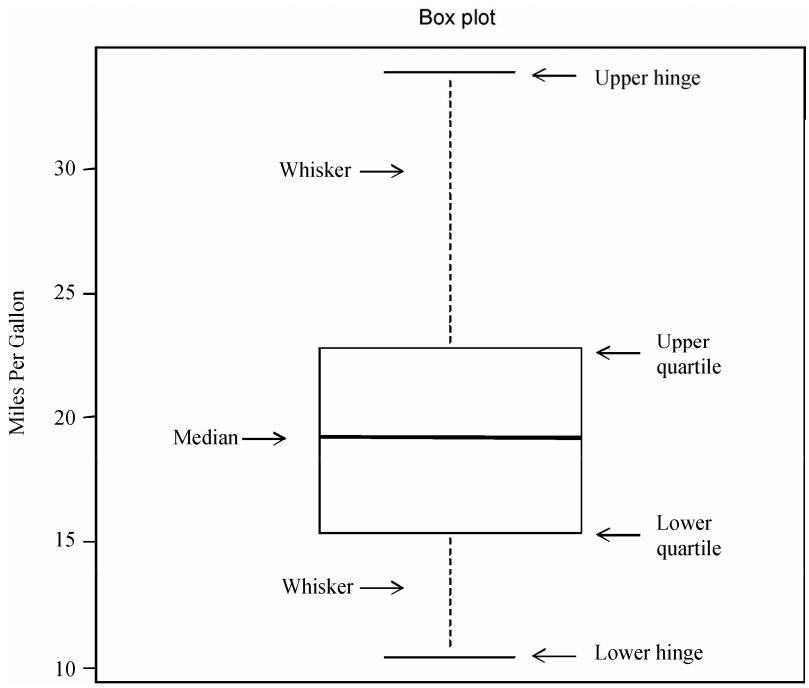

图6-11 含手工标注的箱线图

6.5.1 使用并列箱线图进行跨组比较

箱线图可以展示单个变量或分组变量。使用格式为：

```
boxplot(formula, data=dataframe)
```

其中的*formula*是一个公式，*dataframe*代表提供数据的数据框（或列表）。一个示例公式为y ~ A，这将为类别型变量A的每个值并列地生成数值型变量y的箱线图。公式y ~ A*B则将为类别型变量A和B所有水平的两两组合生成数值型变量y的箱线图。

添加参数varwidth=TRUE将使箱线图的宽度与其样本大小的平方根成正比。参数horizontal=TRUE可以反转坐标轴的方向。

在以下代码中，我们使用并列箱线图重新研究了四缸、六缸、八缸发动机对每加仑汽油行驶的英里数的影响。结果如图6-12所示。

```
boxplot(mpg ~ cyl, data=mtcars,
        main="Car Mileage Data",
        xlab="Number of Cylinders",
        ylab="Miles Per Gallon")
```

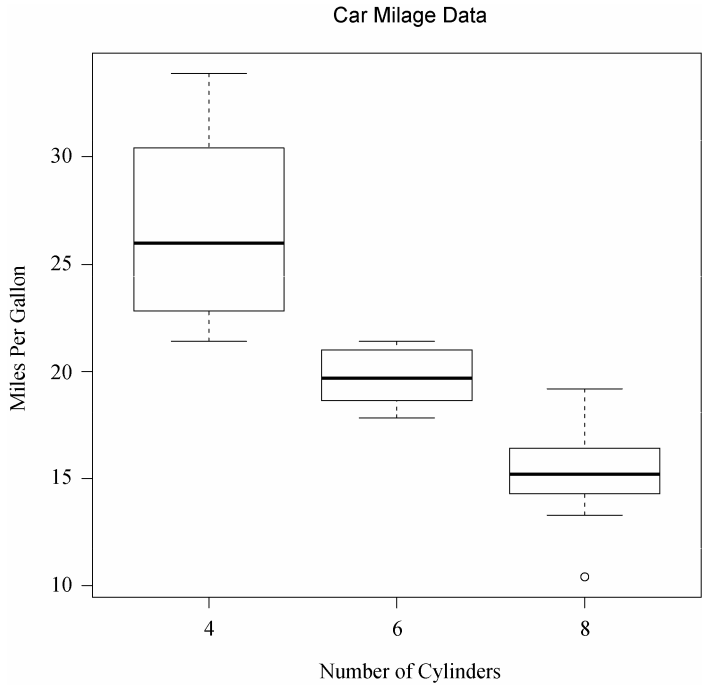

图6-12　不同汽缸数量车型油耗的箱线图

在图6-12中可以看到不同组间油耗的区别非常明显。同时也可以发现，六缸车型的每加仑汽油行驶的英里数分布较其他两类车型更为均匀。与六缸和八缸车型相比，四缸车型的每加仑汽油行驶的英里数散布最广（且正偏）。在八缸组还有一个离群点。

箱线图灵活多变，通过添加`notch=TRUE`，可以得到含凹槽的箱线图。若两个箱的凹槽互不重叠，则表明它们的中位数有显著差异（Chambers et al., 1983, p. 62）。以下代码将为我们的车型油耗示例创建一幅含凹槽的箱线图：

```
boxplot(mpg ~ cyl, data=mtcars,
        notch=TRUE,
        varwidth=TRUE,
        col="red",
        main="Car Mileage Data",
        xlab="Number of Cylinders",
        ylab="Miles Per Gallon")
```

参数`col`以红色填充了箱线图，而`varwidth=TRUE`则使箱线图的宽度与它们各自的样本大小成正比。

在图6-13中可以看到，四缸、六缸、八缸车型的油耗中位数是不同的。随着汽缸数的减少，油耗明显降低。

最后，你可以为多个分组因子绘制箱线图。代码清单6-9为不同缸数和不同变速箱类型的车型绘制了每加仑汽油行驶英里数的箱线图。同样地，这里使用参数`col`为箱线图进行了着色。请

注意颜色的循环使用。在本例中，共有六幅箱线图和两种指定的颜色，所以颜色将重复使用三次。

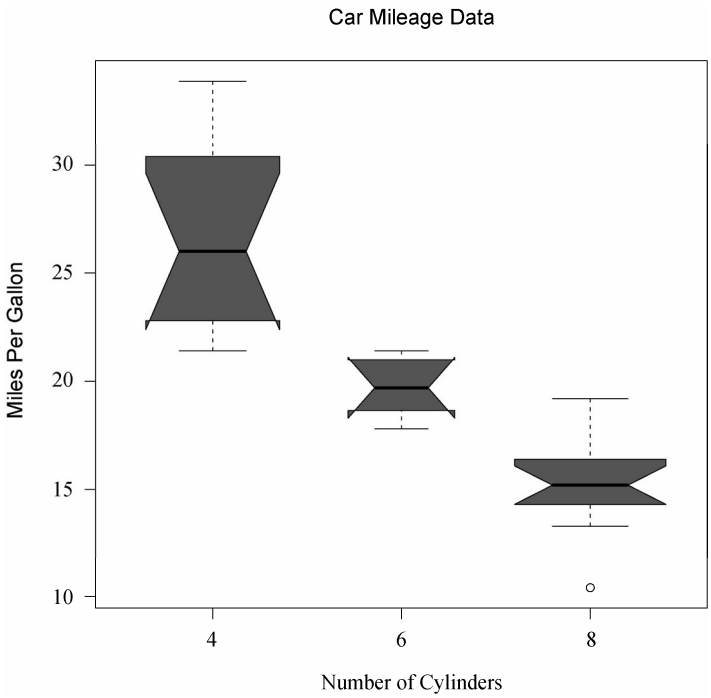

图6-13 不同汽缸数量车型油耗的含凹槽箱线图

代码清单6-9 两个交叉因子的箱线图

```
mtcars$cyl.f <- factor(mtcars$cyl,
                       levels=c(4,6,8),
                       labels=c("4","6","8"))

mtcars$am.f <- factor(mtcars$am,
                      levels=c(0,1),
                      labels=c("auto", "standard"))

boxplot(mpg ~ am.f *cyl.f,
        data=mtcars,
        varwidth=TRUE,
        col=c("gold","darkgreen"),
        main="MPG Distribution by Auto Type",
        xlab="Auto Type")
```

创建汽缸数量的因子

创建变速箱类型的因子

生成箱线图

图形如图6-14所示。

图6-14再一次清晰地显示出油耗随着缸数的下降而减少。对于四缸和六缸车型，标准变速箱（standard）的油耗更高。但是对于八缸车型，油耗似乎没有差别。你也可以从箱线图的宽度看出，四缸标准变速箱的车型和八缸自动变速箱的车型在数据集中最常见。

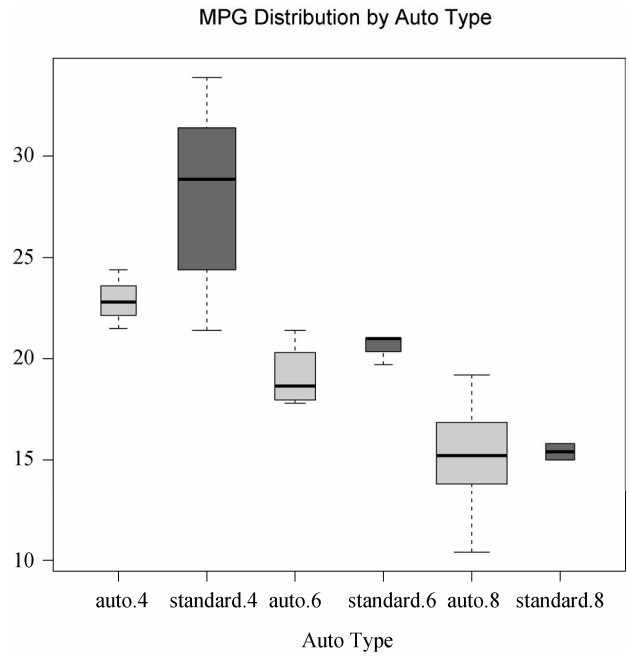

图6-14　不同变速箱类型和汽缸数量车型的箱线图（另见彩插图6-14）

6.5.2　小提琴图

在结束箱线图的讨论之前，有必要研究一种称为小提琴图（violin plot）的箱线图变种。小提琴图是箱线图与核密度图的结合。你可以使用vioplot包中的vioplot()函数绘制它。请在第一次使用之前先安装vioplot包。

vioplot()函数的使用格式为：

```
Vioplot(x1,x2,…,names=,col=)
```

其中*x1*，*x2*，…表示要绘制的一个或多个数值向量（将为每个向量绘制一幅小提琴图）。参数names是小提琴图中标签的字符向量，而col是一个为每幅小提琴图指定颜色的向量。

代码清单6-10中给出了一个示例。

代码清单6-10　小提琴图

```
library(vioplot)
x1 <- mtcars$mpg[mtcars$cyl==4]
x2 <- mtcars$mpg[mtcars$cyl==6]
x3 <- mtcars$mpg[mtcars$cyl==8]
vioplot(x1, x2, x3,
        names=c("4 cyl", "6 cyl", "8 cyl"),
        col="gold")
title("Violin Plots of Miles Per Gallon")
```

注意 vioplot()函数要求你将要绘制的不同组分离到不同的变量中。结果如图6-15所示。

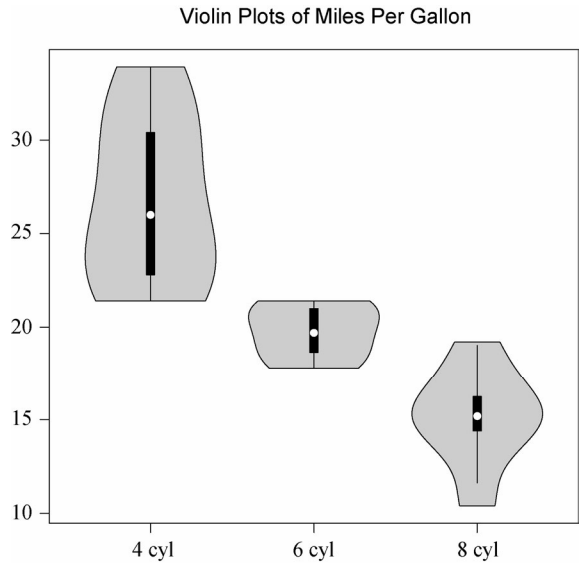

图6-15 汽缸数量和每加仑汽油行驶英里数的小提琴图

小提琴图基本上是核密度图以镜像方式在箱线图上的叠加。在图中，白点是中位数，黑色盒型的范围是下四分位点到上四分位点，细黑线表示须。外部形状即为核密度估计。小提琴图还没有真正地流行起来。同样，这可能也是由于普遍缺乏方便好用的软件导致的。时间会证明一切。

我们将以点图结束本章。与之前看到的图形不同，点图绘制变量中的所有值。

6.6 点图

点图提供了一种在简单水平刻度上绘制大量有标签值的方法。你可以使用dotchart()函数创建点图，格式为：

```
dotchart(x,labels=)
```

其中的x是一个数值向量，而labels则是由每个点的标签组成的向量。你可以通过添加参数groups来选定一个因子，用以指定x中元素的分组方式。如果这样做，则参数gcolor可以控制不同组标签的颜色，cex可控制标签的大小。这里是mtcars数据集的一个示例：

```
dotchart(mtcars$mpg, labels=row.names(mtcars), cex=.7,
         main="Gas Mileage for Car Models",
         xlab="Miles Per Gallon")
```

绘图结果已在图6-16中给出。

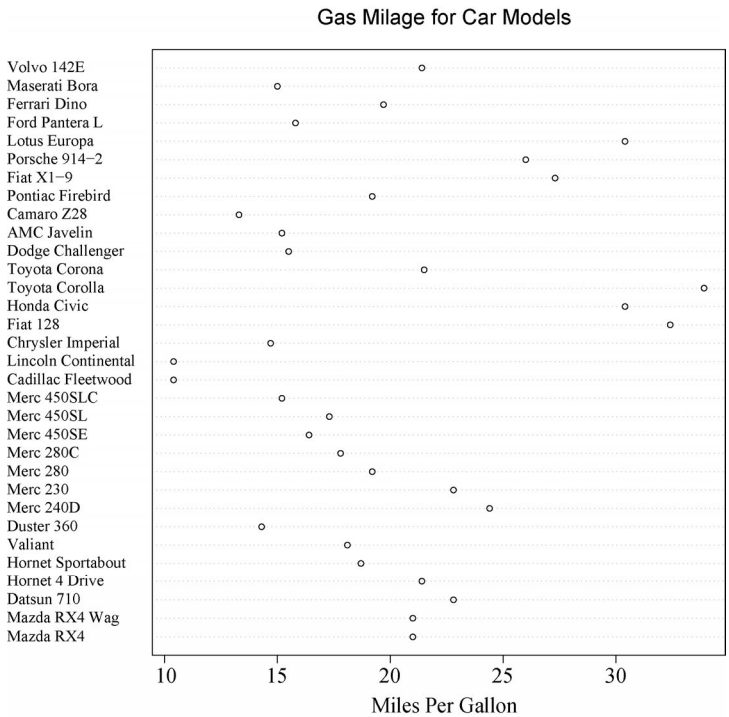

图6-16　每种车型每加仑汽油行驶英里数的点图

图6-16可以让你在同一个水平轴上观察每种车型的每加仑汽油行驶英里数。通常来说，点图在经过排序并且分组变量被不同的符号和颜色区分开的时候最有用。代码清单6-11给出了一个示例。

代码清单6-11　分组、排序、着色后的点图

```
x <- mtcars[order(mtcars$mpg),]
x$cyl <- factor(x$cyl)
x$color[x$cyl==4] <- "red"
x$color[x$cyl==6] <- "blue"
x$color[x$cyl==8] <- "darkgreen"
dotchart(x$mpg,
         labels = row.names(x),
         cex=.7,
         groups = x$cyl,
         gcolor = "black",
         color = x$color,
         pch=19,
         main = "Gas Mileage for Car Models\ngrouped by cylinder",
         xlab = "Miles Per Gallon")
```

在本例中，根据每加仑汽油行驶英里数（从最低到最高）对数据框mtcars进行排序，结果保存为数据框x。数值向量cyl被转换为一个因子。一个字符型向量（color）被添加到了数据框x中，根据cyl的值，它所含的值为"red"、"blue"或"darkgreen"。另外，各数据点的标签取

自数据框的行名（车辆型号）。数据点根据汽缸数量分组。数字4、6和8以黑色显示。点和标签的颜色来自向量color，点以填充的圆圈表示。代码清单6-11绘图的结果如图6-17所示。

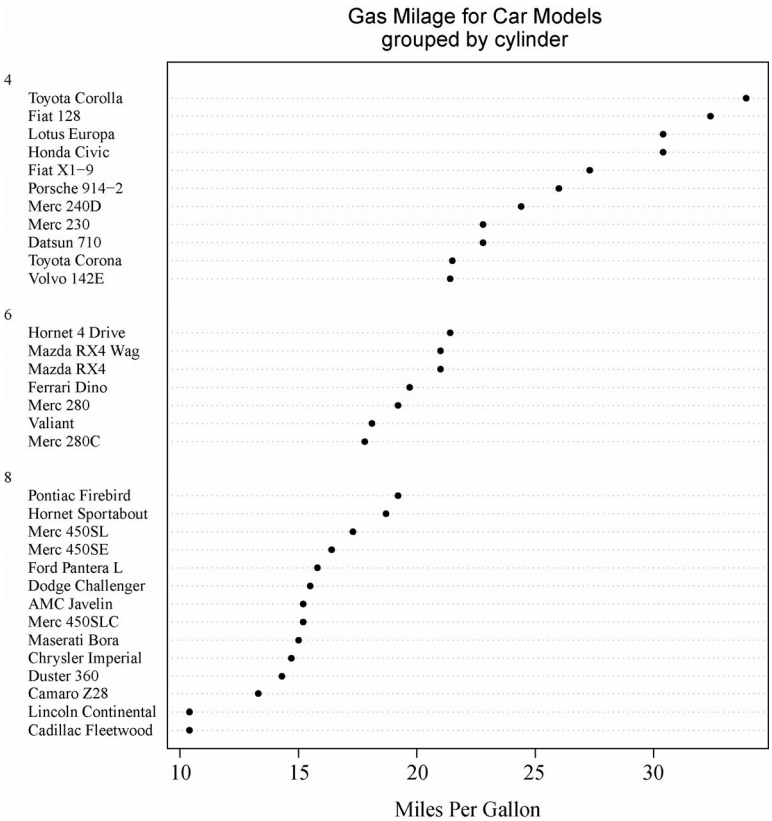

图6-17 各车型依汽缸数量分组的每加仑汽油行驶英里数点图

在图6-17中，许多特征第一次明显起来。你再次看到，随着汽缸数的减少，每加仑汽油行驶的英里数有了增加。但你同时也看到了例外。例如，Pontiac Firebird有8个汽缸，但较六缸的Mercury 280C和Valiant的行驶英里数更多。六缸的Hornet 4 Drive与四缸的Volvo 142E的每加仑汽油行驶英里数相同。同样明显的是，Toyota Corolla的油耗最低，而Lincoln Continental和Cadillac Fleetwood是英里数较低一端的离群点。

在本例中，你可以从点图中获得显著的洞察力，因为每个点都有标签，每个点的值都有其内在含义，并且这些点是以一种能够促进比较的方式排布的。但是随着数据点的增多，点图的实用性随之下降。

注意 点图有许多变种。Jacoby（2006）对点图进行了非常有意义的讨论，并且提供了创新型应用的R代码。此外，Hmisc包也提供了一个带有许多附加功能的点图函数（恰如其分地叫做dotchart2）。