第 2 章

使用图形显示进行数据分析：美国和中国的恶性黑色素瘤家庭生活

2.1 介绍

Fisher 和 Belle (1993) 报告了美国大陆各州在 1950-1969 年间由于白人男性皮肤恶性黑色素瘤所致的死亡率。数据列于表 2.1，其中包括相应州因恶性黑色素瘤导致的死亡人数，各州地理中心的经度和纬度，以及指示与海洋相邻的二进制变量，即该州毗邻海洋之一。关于这些数据的有趣问题包括：海洋和非海洋国家的死亡率如何比较？纬度和经度如何影响死亡率？

表 2.1: 黑色素瘤数据美国因恶性黑色素瘤导致的白人男性死亡率。

	死亡	纬度	经度	海洋
阿拉巴马州	219	33.0	87.0	是
亚利桑那	160	34.5	112.0	没
阿肯色州	170	35.0	92.5	没
加利福尼亚州	182	37.5	119.5	是
科罗拉多州	149	39.0	105.5	没
康乃狄克州	159	41.8	72.8	是
特拉华州	200	39.0	75.5	是
哥伦比亚特区	177	39.0	77.0	没
佛罗里达	197	28.0	82.0	是
佐治亚州	214	33.0	83.5	是
爱达荷州	116	44.5	114.0	没
伊利诺伊州	124	40.0	89.5	没
印第安那州	128	40.2	86.2	没
爱荷华州	128	42.2	93.8	没
堪萨斯州	166	38.5	98.5	没
肯塔基州	147	37.8	85.0	没
路易斯安那州	190	31.2	91.8	是

表 2.1: 黑色素瘤数据 (续)

	死亡	纬度	经度	海洋
缅因州	117	45.2	69.0	是
马里兰州	162	39.0	76.5	是
马萨诸塞州	143	42.2	71.8	是
密西根州	117	43.5	84.5	没
明尼苏达州	116	46.0	94.5	没
密西西比州	207	32.8	90.0	是
密苏里州	131	38.5	92.0	没
蒙大拿	109	47.0	110.5	没
内布拉斯加	122	41.5	99.5	没
内华达州	191	39.0	117.0	没
新罕布什尔	129	43.8	71.5	是
新泽西州	159	40.2	74.5	是
新墨西哥	141	35.0	106.0	没
纽约	152	43.0	75.5	是
北卡罗来纳	199	35.5	79.5	是
北达科他州	115	47.5	100.5	没
俄亥俄	131	40.2	82.8	没
俄克拉荷马州	182	35.5	97.2	没
俄勒冈州	136	44.0	120.5	是
宾夕法尼亚州	132	40.8	77.8	没
罗德岛	137	41.8	71.5	是
南卡罗来纳	178	33.8	81.0	是
南达科他州	86	44.8	100.0	没
田纳西州	186	36.0	86.2	没
德州	229	31.5	98.0	是
犹他州	142	39.5	111.5	没
佛蒙特	153	44.0	72.5	是
维吉尼亚州	166	37.5	78.5	是
华盛顿州	117	47.5	121.0	是
西弗吉尼亚	136	38.8	80.8	没
威斯康星州	110	44.5	90.2	没
怀俄明州	134	43.0	107.5	没

*资*料来源: 摘自 LD 的 Fisher 和 GV 的 Belle，《生物统计》。卫生科学方法论，约翰●威利父子公司，英国奇切斯特，1993 年。获许可。

当代中国处于性革命的前沿，地区和代际差异巨大，为分析性行为的前因和后果提供了无与伦比的自然实验。作为芝加哥大学，北京大学和美国大学的一项合作研究项目，《中国健康与家庭生活调查》于 1999 年至 2000 年进行。

北卡罗莱纳州提供了一个基线，可以据此预测和跟踪未来的变化。具体来说，这项研究使用全国代表性的概率样本得出了有关性行为和疾病模式的基线结果集。《中国健康与家庭生活调查》对 60 个村庄和城市社区进行了抽样调查，这些村庄和城市社区的选择代表了不包括香港和西藏在内的当代中国的全部地理和社会经济范围。从 20 至 64 岁之间的成年人正式登记册中，随机选择每个地点的 83 个人，以总共 5000 个人为目标。在这里，我们将注意力集中在没有任何信息的现有男性伴侣的女性身上，从而对 1534 名具有以下变量的女性进行了抽样（数据表示例见表 2.2）：

回应女性的教育水平，

R_受访女性的月收入（元），R_去年该受访女性的健康状况，R_happy 去年该

受访女性的幸福程度，该女性伴侣的教育程度，

妇女伴侣的月收入（元）。

在上面的列表中，收入变量是连续的，其余变量是按有序类别分类的。收入变量基于（部分）估算的度量。所有信息，包括伴侣的收入，均来自仅由受访妇女回答的问卷。在这里，我们专注于图形显示，以检查异性恋女性及其伴侣的健康与社会经济变量之间的关系。

2.2 初始数据分析

根据钱伯斯等。（1983 年），“没有统计工具像选择好的图表那样强大”。当然，对大多数（可能是所有）数据集的分析应该首先尝试通过以希望的有用和有益的方式将其绘制成图表来理解数据的一般特征。Schmid（1954）总结了图形表示方法的可能优势。它们包括以下内容：

- 与其他类型的演示文稿相比，精心设计的图表可以更有效地引起人们的兴趣并吸引读者的注意力。
- 图表和图形所描绘的视觉关系更容易掌握和记忆。
- 使用图表可以节省时间，因为可以一目了然地看到大量统计数据的本质含义。
- 图表提供了问题的全面描述

表 2.2：CHFLS 数据中国健康与家庭生活调查。

	ReDeu	R_收入	健康	快乐的	阿伊杜	A_收入
2	高中	900	好	有点幸福	高中	500
3	高中	500	公平	有点幸福	高中	800
10	高中	800	好	有点幸福	初中	700
11	初中	300	公平	有点幸福	小学	700
22	初中	300	公平	有点幸福	初中	400
23	高中	500	优秀的	有点幸福	大专学院	900
24	初中	0	不好	很高兴	初中	300
25	初中	100	好	不太开心	高中	800
26	初中	200	公平	不太开心	大专学院	200
32	高中	400	好	有点幸福	高中	600
33	初中	300	不好	不太开心	初中	200
35	初中	0	公平	有点幸福	初中	400
36	初中	200	好	有点幸福	初中	500
37	高中	300	优秀的	有点幸福	高中	200
38	大专学院	3000	公平	有点幸福	大专学院	800
39	大专学院	0	公平	有点幸福	大学	500
40	高中	500	优秀的	有点幸福	高中	500
41	初中	0	不好	不太开心	初中	600
55	高中	0	优秀的	有点幸福	初中	0
56	初中	500	不好	很高兴	初中	200
57	⋮	⋮	⋮	⋮	⋮	⋮

比表格或文字形式的呈现方式更完整，更平衡。

- 图表可以显示出隐藏的事实和关系，可以激发并帮助分析性思维和调查。

图非常流行；据估计，每年打印的统计图形图像在 9000 亿张（9×10^{11}）和 2 万亿张（2×10^{12}）。如此受欢迎的主要原因之一可能是图形
数据表示通常为发现意外事件提供了手段；尽管应牢记已故的卡尔·萨根（在其著作 Contact 中）的以下警告，但人类的视觉系统在检测模式方面非常强大：

> 人类擅长辨别真正存在的细微图案，但是同样也可以想象它们完全不存在时的细微图案。

在过去的二十年中，已经开发了各种各样的用于以图形方式显示数据的新方法。这些将寻找数据中的特殊效果，指示异常值，识别模式，诊断模型并通常搜索新颖的，也许是意想不到的现象。可能需要大量的图，并且出于与它们用于数值分析的相同原因，通常需要计算机来提供它们，即它们快速且准确。

因此，由于机器正在工作，所以问题不再是“我们要绘图吗？”而是“我们该如何计划？”有许多激动人心的可能性，包括动态图形，但是数据的图形探索通常至少是从一些更简单，众所周知的方法开始的，例如，直方图，条形图，箱形图和散点图。本章将对每种方法以及更复杂的方法（如旋转图和网格图）进行说明。

2.3 使用 R 分析

2.3.1 恶性黑色素瘤

我们可能通过构建图 2.1 中所有死亡率的直方图或箱线图开始检查表 2.1 中的恶性黑色素瘤数据。第 1 章已经介绍了 plot，hist 和 boxplot 函数，我们想生成一个同时应用两种技术的图。布局功能可在一台绘图设备上组织两个独立的绘图，例如彼此重叠。使用这种相对简单的技术（稍后将介绍更高级的方法），我们必须确保两个图中的 x 轴相同。这可以通过计算可能的数据范围来完成，然后通过 xlim 参数在图中指定：

```
R> xr <-范围 (USmelanoma $ mortality) * c (0.9, 1.1)
R>
[1]  77.4 251.9
```

现在，要绘制直方图和箱线图，都需要将绘图设备的空间设置为相等，以便在彼此顶部放置两个独立的图。

```
R>布局 (矩阵 (1: 2, nrow = 2) )
r＝PAR (MAR＝PAR ( “MAR”) *C (0.8, 1, 1, 1) )
R> boxplot (USmelanoma $ mortality, ylim = xr, horizontal = TRUE,
+          xlab =“死亡率” )
RHIST (美国黑色素瘤$死亡率, XLIM= XR, XLAB＝ “, Ma= =””,
+       轴= FALSE, ylab =“”) R>
轴 (1)
```

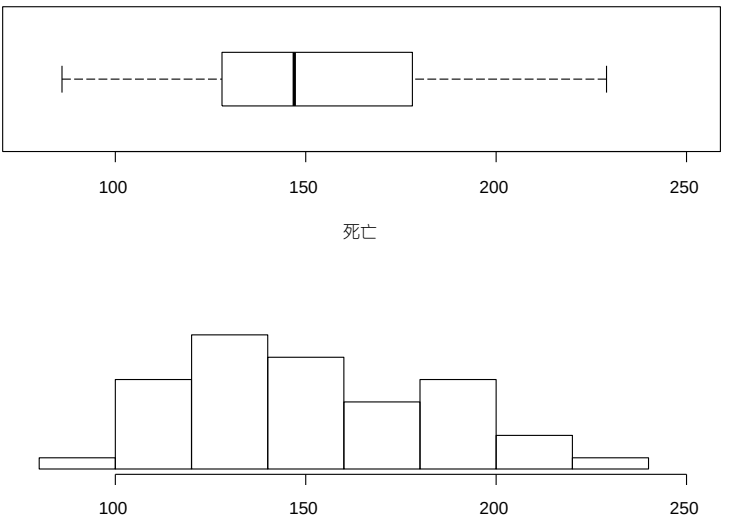

图 2.1 恶性黑色素瘤死亡率的直方图（上图）和箱形图（下图）。

在具有两行的两个单元格（包含数字 1 和 2）的矩阵上调用布局函数会导致这种划分。首先在死亡率数据上调用箱形图函数，然后在历史函数上调用该函数，其中两个图中 x 轴的范围由（77.4，251.9）定义。要解决的一个小问题是边距的大小。对于这样的绘图，其默认值太大。与其他许多图形参数一样，您可以针对
使用功能参数的特定图。R 代码和显示结果如图 2.1 所示。

图 2.1 中的直方图和箱线图均表明死亡率分布存在一定的偏斜。查看所有死亡率的特征是一个有用的开始，但是对于这些数据，我们可能更感兴趣于比较海洋州和非海洋州的死亡率。因此，我们可以构造两个直方图或两个箱形图。这样的平行箱线图

```
R>情节（死亡率～海洋，数据=美国黑色素瘤，
+          xlab =“与海洋接壤”，ylab =“死亡率”）
```

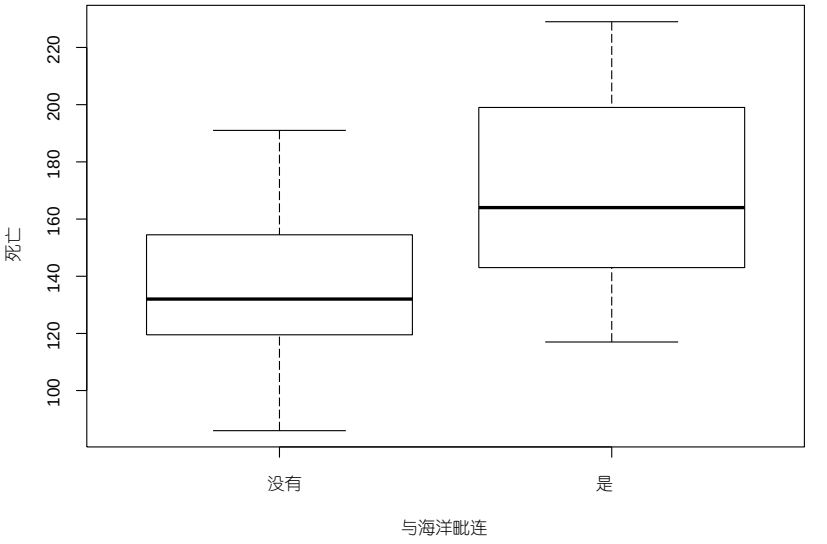

图 2.2 恶性黑色素瘤死亡率平行于海洋的箱线图。

使用 boxplot 函数可以很容易地计算出分类变量给定的成组数字变量的条件分布。连续响应变量和分类自变量是通过第 1 章中所述的公式指定的。图 2.2 显示了此类平行箱形图，默认情况下会生成此类数据的图函数，用于海洋和非海洋州和铅的死亡率给人的印象是，与该国其他地区相比，东部或西部沿海地区的死亡率增加了。

直方图通常用于两个目的：计数和显示变量的分布；以及根据威尔金森（Wilkinson，1992）的观点，“它们对两者都不起作用”。直方图通常会误导显示分布，因为直方图取决于所选类别的数量。另一种方法是正式估算变量的密度函数，然后绘制结果估算值。密度估算的详细信息在第 8 章中给出，但是对于海洋和非海洋国家，可以生成并绘制两个密度估算，如图 2.3 所示，这支持了图 2.2 的印象。有关这种密度估计的更多详细信息，请参阅第 8 章。

```
R>染料<-具有 (USmelanoma, 密度 (死亡率[海洋==“是”]]) ) R> dno
<-具有 (USmelanoma, 密度 (死亡率[海洋==“否”]) ) R>图 (染料,
lty = 1, xlim = xr, main =“”, ylim = c (0, 0.018) ,
+       xlab =“死亡率”) R>
行 (dno, lty = 2)
R>传奇 (ToPulft) , LTY＝1:2, 传说＝C (“沿海国家”) ,
+          “土地州”) , bty =“ n”)
```

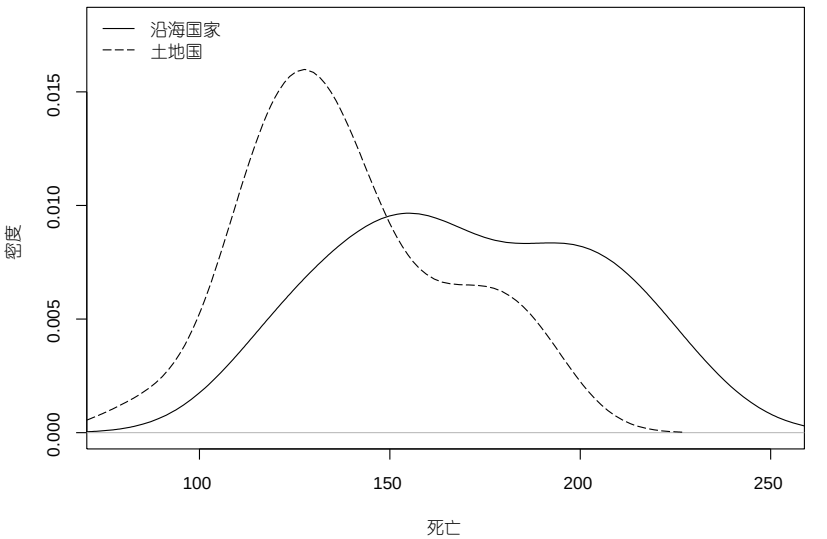

图 2.3 恶性黑色素瘤死亡率的估计值（按邻近海洋而定）。

现在，我们可能继续研究死亡率与州中心地理位置所代表的地理位置之间的关系。这里的主要图形将是散点图。至少从 18 世纪开始，简单的 xy 散点图就已经使用，并且具有许多优点-实际上，根据 Tufte (1983) :

> 关系图形（以散点图及其变体的最佳形式）是所有图形设计中最大的。它链接了至少两个变量，从而鼓励甚至劝告观看者评估所绘制变量之间的可能因果关系。它面对因果理论，即 x 导致 y 的 x 和 y 之间的实际关系的经验证据。

让我们从死亡率对经度和死亡率对纬度的简单散点图开始，这些散点图可以通过图 2.4 之前的代码生成。同样，版图功能用于对绘图进行分区

```
R>布局 (矩阵 (1: 2, ncol = 2) )
R>情节 (死亡率∽经度，数据=美国黑色素瘤，
+         ylab =“死亡率”，xlab =“经度”) R> plot
 (mortality∽latitude, data = USmelanoma,
+         ylab =“死亡率”，xlab =“纬度”)
```

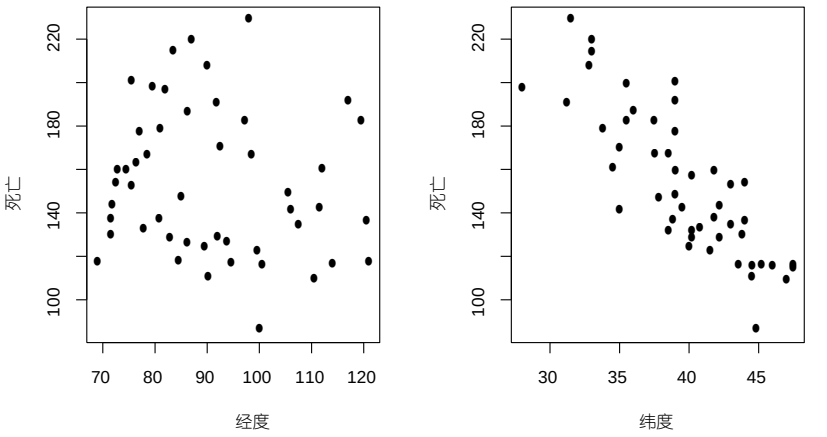

图 2.4 恶性黑色素瘤死亡率分布图（按地理位置）。

设备，现在产生两个并排图。现在，layout 的参数是一个只有一行但两列包含数字 1 和 2 的矩阵。在每个单元格中，调用 plot 函数以生成公式中给出的变量的散点图。

由于死亡率显然仅与纬度相关，因此我们现在可以分别针对海洋州和非海洋州产生针对纬度的死亡率散点图。除了可以产生两个显示之外，还可以为两种状态选择不同的绘图符号。这可以通过指定向量来实现
pch 的整数或字符，其中此向量的第 i 个元素定义要绘制的数据中第 i 个观测值的绘制符号。为了简单起见，我们将海洋因子转换为整数向量，该向量包含陆地国家的数字 1 和海洋国家的数字 2。作为结果，
陆地状态可以用点符号标识，海洋状态可以用三角形标识。向图例中添加图例非常有用，最方便的方法是使用图例功能。该函数采用三个参数：一个指示图例在图中的位置的字符串，要打印的标签的字符向量以及相应的绘图符号（以整数表示）。另外，预计将显示边界框（bty =“ n”）。散点图中

```
R>情节（死亡率～纬度，数据=美国黑色素瘤，
+        pch = (1: 2)[ocean], ylab =“死亡率”，
+        xlab =“纬度”）
R>传奇（TopTrand），传说＝C（“陆地状态”，“海岸国”），
+          pch = 1:2, bty = "n")
```

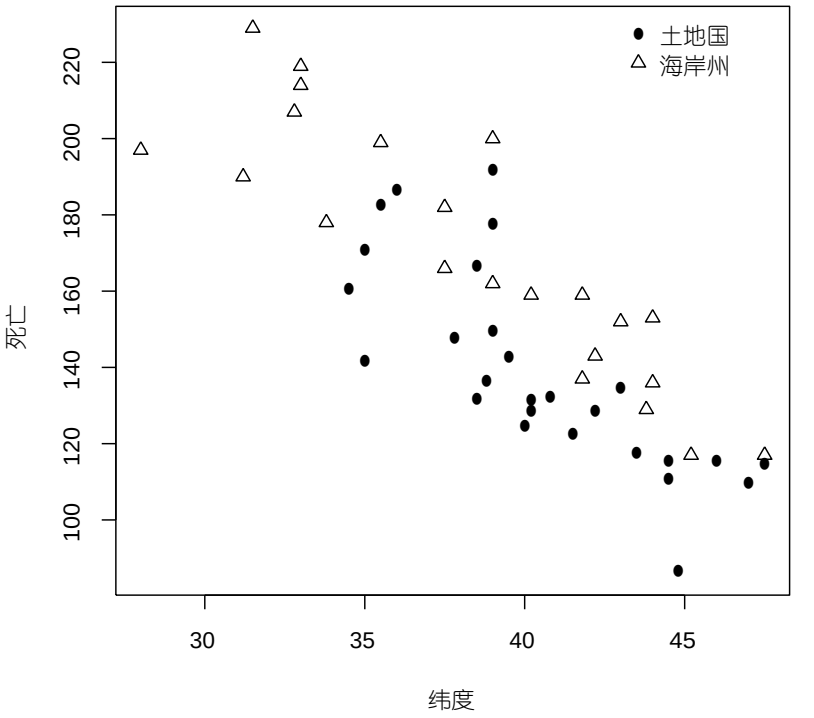

图 2.5 恶性黑色素瘤死亡率相对于纬度的散点图。

图 2.5 突出显示了北部土地州的死亡率最低。在大致相同的纬度下，沿海州的死亡率高于陆上州。在南部沿海州，纬度小于 32°的死亡率最高。

```
R>子集（美国黑色素瘤，纬度<32）
```

	死亡	*纬度*	*经度*	*海洋*
佛罗里达	*197*	*28.0*	*82.0*	*是*
路易斯安那	*190*	*31.2*	*91.8*	*是*
德州	*229*	*31.5*	*98.0*	*是*

或者，我们也可能只是想看一看美国的彩色地图，其中每个州都以与其死亡率相对应的颜色绘制。使用 sp 系列软件包建立这样的地块非常简单（Pebesma 和 Bivand，2013）。我们从加载大陆州的地图开始，基本上是一些多边形：

```
R>库（“ sp”）R>库
（“地图”）
R>库（“ maptools”）
R>状态<-map（“ state”，plot = FALSE，fill = TRUE）
```

使死亡率与相应状态相匹配当然很重要。因此，我们为多边形和死亡率数据均以小写字母创建了州的唯一名称

```
R> ID <-sapply（strsplit（states $ names，“：”），function（x）
x [1]）R>行名（USmelanoma）<-降低（rownames（USmelanoma））
```

现在我们准备将这两个对象合并为一个所谓的 SpatialPolygons-DataFrame 对象。我们首先在正确的参考系统（本例中为 WGS84）中根据地图创建一个 SpatialPolygons 对象，然后将这些多边形与数据合并

```
R> U1＜-MAP2 空间多边形（状态，IDS＝IDS）
+      Prj4Stult= CRS（“+ PROJ= LoLAT+DATAM= WGS84”）R>
US2 < SpatialPolygonsDataFrame（US1，美国黑色素瘤）
```

现在可以使用 spplot 函数绘制结果对象 us2，请参见图 2.6。颜色对应于死亡率，如地图右侧的颜色图例所示。我们看到，在东部和西部沿海的南部沿海州，出现了对应于较高死亡率的较暗的灰度值，这与我们之前的结果吻合良好。

到目前为止，我们主要集中于连续变量的可视化。现在，我们将重点扩展到分类变量的可视化。

2.3.2 中国健康与家庭生活

中国健康和家庭生活调查所关注的问卷的一部分是自我报告的健康状况。两个问题对我们很有趣。第一个是“通常来说，您认为您的健康状况是好，好，一般，不好还是差？”。第二个问题是“通常来说，在过去的十二个月中，您有多高兴？”。通常使用条形图来可视化此类变量的分布，其中针对每个类别显示观察值的总数或相对数。通过将 barplot 函数应用于数据列表，可以方便地生成这种条形图。变量 R_happy 的经验密度是通过 xtabs 函数计算来生成（列）表的；结果条形图如图 2.7 所示。

两个类别变量的可视化可以通过条件条形图完成，即，第二个变量类别中的第一个变量的条形图。显示此类双向表的一种有吸引力的替代方法是

RSP> (2)，“死亡率”，Cel.地区＝Rev（灰色，颜色（100））

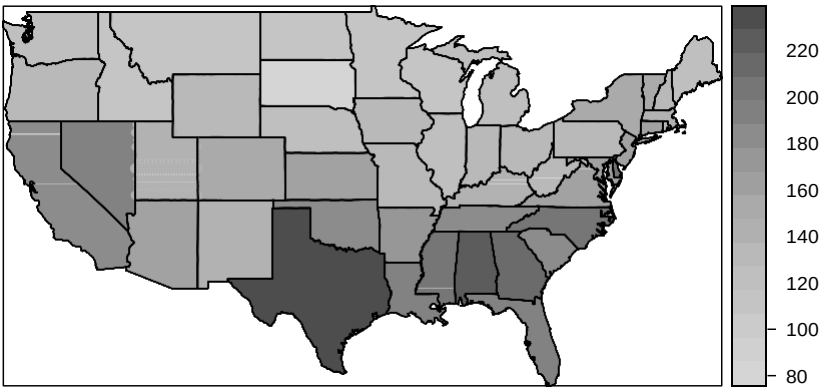

图 2.6。美国地图显示恶性黑色素瘤死亡率。

*自*旋印迹（Friendly，1994；Hofmann and Theus，2005；Chen et al。，2008）；当查看图 2.8 中的图时，名称的含义将变得清楚。

在构建这样的图之前，我们使用 xtabs 函数生成一个关于健康状况和自我报告的幸福感的双向表：

R> xtabs (∽R_happy + R_health，数据= CHFLS)

	健康				
快乐的	*较差*	*不 好*	*公平*	*好*	*优秀的*
非常不开心	*2*	*7*	*4*	*1*	*0*
不太开心	*4*	*46*	*67*	*42*	*26*
有点幸福	*3*	*77*	*350*	*459*	*166*
很高兴	*1*	*9*	*40*	*80*	*150*

自旋图是一组矩形，每个矩形代表双向列联表中的一个单元。矩形的面积与单元中的观察数量成比例。在这里，我们在图 2.8 中绘制了健康状况和幸福感的镶嵌图。

考虑图 2.8 中的右上方单元格，即 150 名健康状况非常好的非常幸福的女性。最右边的条的宽度对应于具有良好健康状况的女性的频率。鉴于女性的健康状况非常好，右上方矩形的长度对应于非常幸福的女性的条件频率。将这两个数量相乘得出

```
R> barplot (xtabs (~R_happy, data = CHFLS) )
```

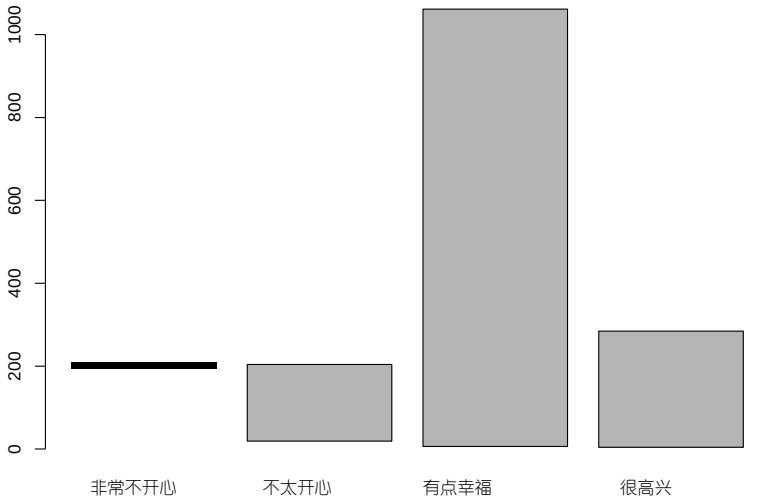

图 2.7 快乐条形图。

该单元的面积对应于既非常快乐又享有良好健康状况的女性的频率。非常幸福的女人的条件频率随着健康状况的增加而增加，而非常不幸福或不太幸福的女人的条件频率降低。

当关注分类变量和连续变量的关联时，比如说月收入和自我报告的幸福感，可以使用平行箱形图可视化取决于幸福感的收入分配。但是，如果我们研究自我报告的幸福作为反应，收入作为自变量，则可以表示给定幸福的收入的条件分布，但是我们对给定收入的幸福的条件分布感兴趣。产生更合适的图的一种可能性称为自旋图。在此，连续的 x 变量首先被分类。在这些类别的每一个类别中，响应变量的条件频率由堆叠的条形图给出，其方式类似于自旋图。对于幸福取决于对数收入（由于收入自然偏斜，我们使用收入的对数转换），似乎不幸福和不太幸福的妇女的比例随着收入的增加而减少，而非常幸福的妇女的比例则保持恒定。与旋转图相反，如直方图所示，在 x 轴上给出了 bin，

```
R> plot (R_happy∽R_health, data = CHFLS, ylab =“ Happiness”,
+        xlab =“健康”)
```

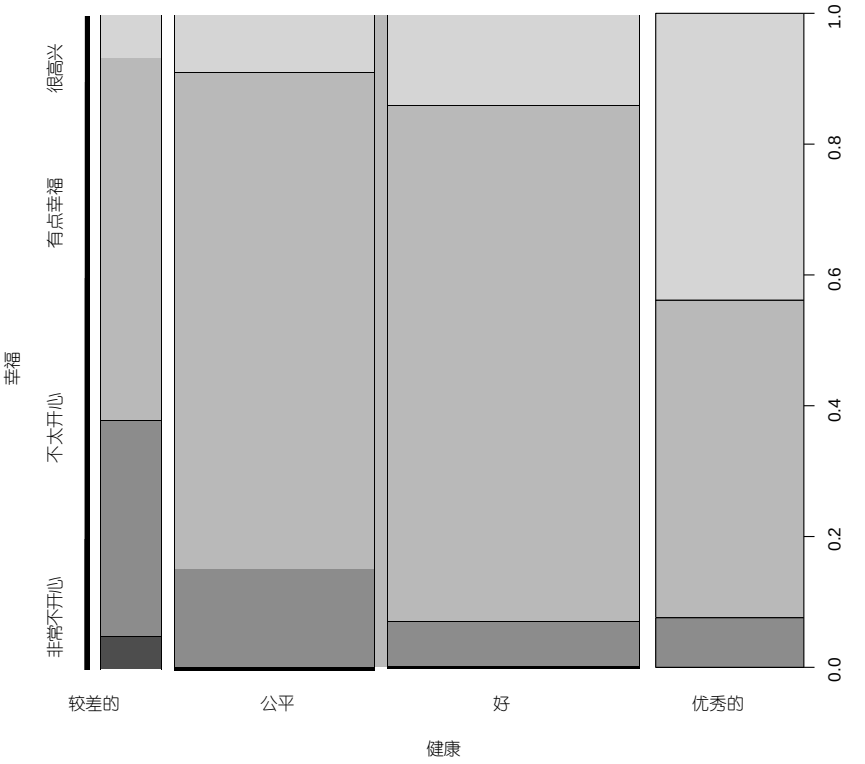

图 2.8 健康状况和幸福感的自转图。

条件密度图使用原始 x 轴显示给定独立变量的分类响应的条件密度。

对于我们的最后一个示例，我们返回散点图，检查一个女人的月收入与其伴侣的收入之间的关系。这两个收入变量均已计算，并从其他自我报告的变量中部分推算得出，它们只是对实际收入的粗略评估。此外，数据本身是数字的，但联系紧密，由于点将重叠，因此很难生成“正确的”散点图。一个相对简单的技巧是通过在每个点上添加一个小的随机噪声来抖动观察结果，以避免重叠的绘图符号。另外，我们想研究两个月收入之间的关系，这些条件取决于妇女的受教育程度。这种条件图被称为网格图，并在包装格子中实现（Sarkar，2014，2008）。我们利用 xyplot 函数

```
R>布局 (矩阵 (1: 2, ncol = 2) )
R>图 (R_happy∽log (R_income +1) , 数据= CHFLS,
+        ylab =“幸福”, xlab =“ log (收入+ 1) ”) R>
cdplot (R_happy∽log (R_income +1) , 数据= CHFLS,
+          ylab =“幸福”, xlab =“ log (收入+ 1) ”) R>
```

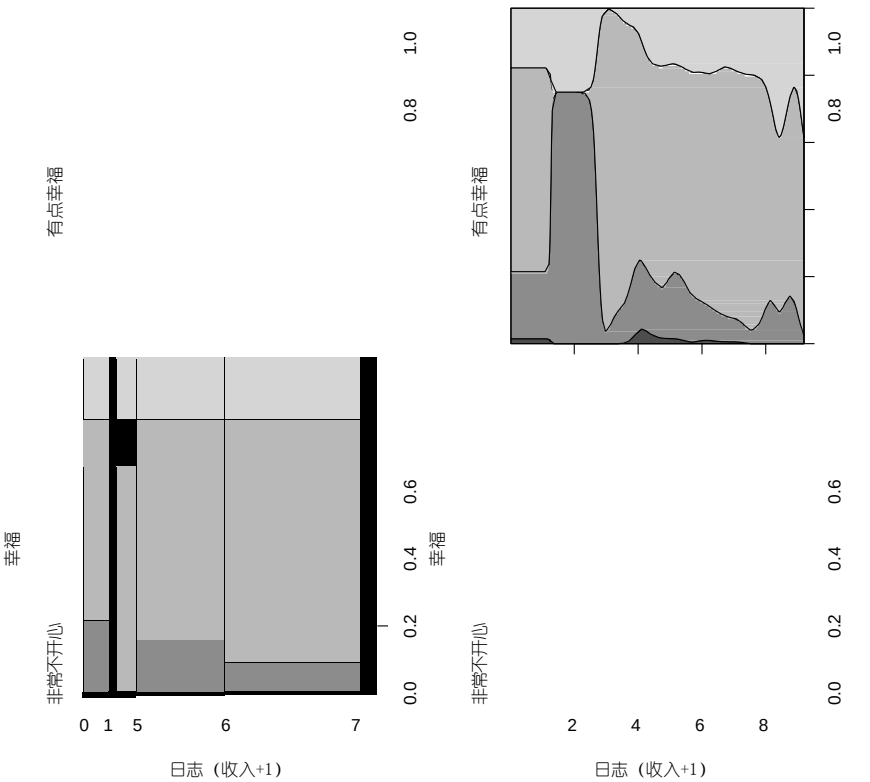

图 2.9 取决于对数收入的幸福感的心电图（左）和条件密度图（右）。

从包装晶格生成散点图。公式的读取方式已经解释过了，只是存在第三个条件变量 R_edu。对于每个教育水平，将生成一个单独的散点图。这些图可直接比较，因为所有图的轴均相同。

图 2.10 所示的图揭示了几个有趣的问题。一些观测值位于斜率为 1 的直线上，最有可能是线性模型对缺失值进行插补的结果（如数据字典中所述，请参见文档？CHFLS）。可以确定四个星座：双方都没有收入，对方没有收入，女人没有收入，或者双方都有正收入。

对于拥有女性大学学历的夫妻来说，双方的收入都相对较高（只有两对夫妻只有女性有收入）。少数以前的大专学生生活在只有男人有收入的恋爱关系中，对于剩余的夫妻来说，伴侣的收入似乎只是正相关。对于较低的教育水平，所有四个星座都存在。只有男人有一些收入的夫妻的频率似乎比其他人高。忽略

```
R>库（“晶格”）
R> xyplot（抖动（log（R_income + 0.5））～
+          抖动（log（A_income + 0.5））|R_edu，数据= CHFLS，
+          PCH = 19， COL = RGB （1， 1， 1，
1） ，
+          ylab =“ log（妻子的收入+ .5）”，
+          xlab =“ log（丈夫的收入+ .5）”）
```

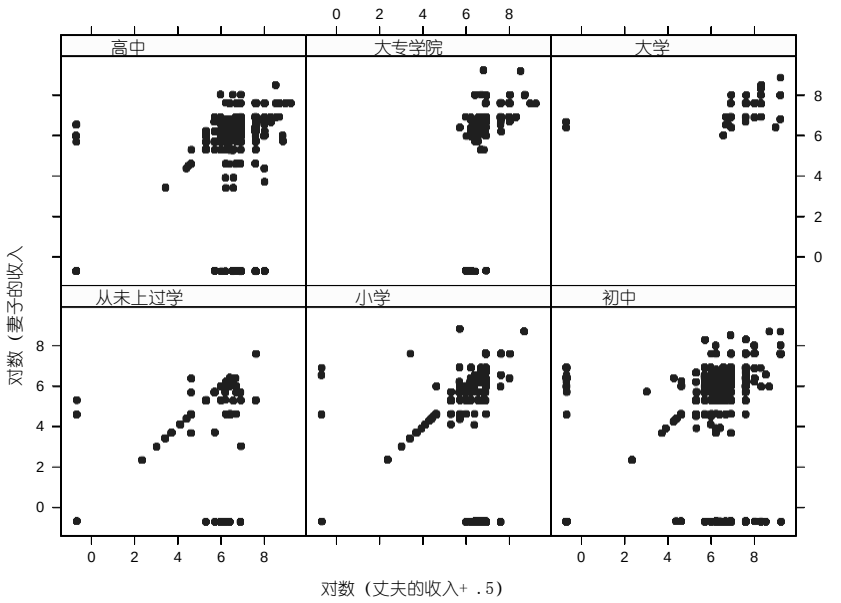

图 2.10 夫妻的对数收入抖动的散点图，取决于妻子的受教育程度。

在直线上的观察结果中，双方的收入之间几乎没有关联。

2.4 调查结果摘要

仅在本章考虑的两组数据上使用相对简单的图形技术，我们便能够发现每个数据集的许多重要特征。

黑色素瘤死亡率死亡率仅与一个州的纬度有关，而与该州的经度无关，沿海国家的死亡率高于陆上国家，在南部沿海国家的纬度小于 32 度时，死亡率最高。

健康与家庭生活我们看到幸福取决于健康状况。妇女报告称自己经常更快乐时，她们也会

良好或良好的健康状况。幸福对妇女收入的依赖性似乎不太明确，但是我们得出结论，在受教育的条件下，妻子及其丈夫的收入高度相关。

2.5 最后评论

制作具有出版物质量的图形是 R 系统的主要优势之一，由于图形可以在 R 中进行编程，因此几乎可以做任何事情。自然地，本章仅是对一些常用显示器的非常简短的介绍，读者可以参考专业知识。书籍，大部分
重要的 Murrell（2005），Sarkar（2008）和 Chen 等。（2008）。交互式 3D 图形可从 rgl 软件包获得（Adler 和 Murdoch，2014 年）。

练习题

例如 2.1 表 2.3 中的数据是从家庭支出调查中收集的数据集的一部分，给出了在四个商品组上的 20 位单身男性和 20 位单身女性的支出。支出单位是港元，四个商品组是

住房，包括燃料和照明，
包括烟酒在内的食品
其他商品，包括服装，鞋类和耐用品，
服务服务，包括运输和车辆。

该调查的目的是调查四个商品组之间家庭支出的分配如何取决于总支出，并查明男女之间的这种关系是否不同。使用适当的图形方法回答这些问题并陈述您的结论。

表 2.3：家庭数据单身男女的家庭支出。

住房	餐饮	产品	服务	性别
820	114	183	154	女
184	74	6	20	女
921	66	1686	455	女
488	80	103	115	女
721	83	176	104	女
614	55	441	193	女
801	56	357	214	女
396	59	61	80	女
864	65	1618	352	女
845	64	1935	414	女
404	97	33	47	女

表 2.3: 家庭数据 (续)

住房	餐饮	产品	服务	性别
781	47	1906	452	女
457	103	136	108	女
1029	71	244	189	女
1047	90	653	298	女
552	91	185	158	女
718	104	583	304	女
495	114	65	74	女
382	77	230	147	女
1090	59	313	177	女
497	591	153	291	男
839	942	302	365	男
798	1308	668	584	男
892	842	287	395	男
1585	781	2476	1740	男
755	764	428	438	男
388	655	153	233	男
617	879	757	719	男
248	438	22	65	男
1641	440	6471	2063	男
1180	1243	768	813	男
619	684	99	204	男
253	422	15	48	男
661	739	71	188	男
1981	869	1489	1032	男
1746	746	2662	1594	男
1865	915	5184	1767	男
238	522	29	75	男
1199	1095	261	344	男
1524	964	1739	1410	男

例如 2.2 表 2.4 中显示的数据集包含美国十个州的七个变量的值。七个变量是

人口总数除以 1000，人均收入，文盲率（人口百分比），寿命，预期寿命（年），凶杀凶杀率（每千人），
高中毕业生的毕业生比例，
每次冻结以下的平均冻结天数。