2021年数理统计上机课

–统计图绘制

黄启岳

北京师范大学统计学院

2021 年 3 月 24 日

目录

数据可视化

“The simple graph has brought more information to the data analyst’s mind than any other device.”—John Tukey

在数据时代，将数据“翻译”成大家喜闻乐见的形式是统计学的重要工作。眼睛是人类最重要的感觉器官，接近80%的外界信息由视觉获取。

统计很多时候是对数据进行直接“操作”，包括选择、运算等，但运算的结果相对于非专业人士并不直观。

想一想：这是什么？

400463	400462	0.9411765
400464	400463	0.9333333
400465	400464	0.9294118
400466	400465	0.9058824
400467	400466	0.9137255
400468	400467	0.9176471
400469	400468	0.9215686
400470	400469	0.9411765
400471	400470	0.945098
400472	400471	0.9333333
400473	400472	0.9098039
400474	400473	0.8745098
400475	400474	0.827451
400476	400475	0.7647059
400477	400476	0.6196078
400478	400477	0.4039216
400479	400478	0.254902
400480	400479	0.2039216
400481	400480	0.1568627
400482	400481	0.0862745
400483	400482	0.0470588
400484	400483	0.0431373
400485	400484	0.054902

图 1: What's this?

使用plot函数将它画出来(中间省略部分步骤)：

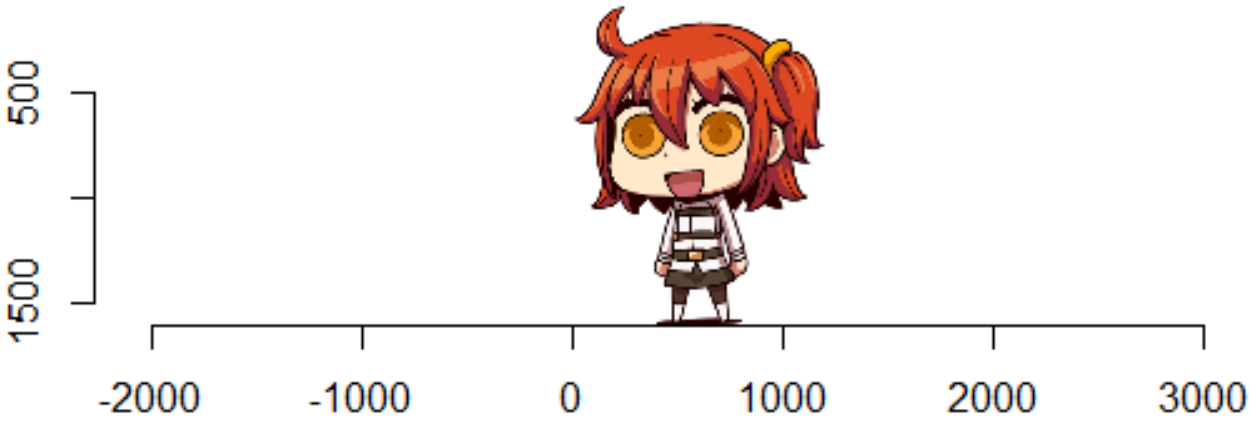

图 2: 数据还原

图片为1200×1600像素，每一个像素点上为RGB三色，向量总长度为5760000。这是人类看图与机器看图的对比。

准备

数据可视化的重要性不言而喻，随着大数据时代的到来，这方面需求与日俱增，相关工具也在不断更新。以下重点介绍常见统计图表。需要安装程序包ggplot2。也可以直接安装tidyverse，即直接加载系列包。

ggplot2是Hadley Wickham最成功的作品之一，极大改变了R语言绘图的模式。更多有关ggplot2的内容请参考R for Data Science与ggplot2: elegant graphics for data analysis两本书，均为作者本人编写。

目标

本节主要学习以下几种统计图的绘制：散点图、柱状图、折线图、直方图以及箱线图，这些图表常用于描述性统计，有利于初步掌握数据的基本信息。

绘图原理

ggplot()的绘图原理是图层叠加，按图层作图，将数据相关与数据无关的部分分离。函数格式为：以ggplot()作为开始，通过“+”连接，每一个语句对应一个独立的图层，最后这些图层按次序叠加，形成一张完整的图。

一张统计图形就是从数据到几何对象（geometric object，缩写geom）的图形属性（aesthetic attribute，缩写aes）的一个映射。此外，图形中还可能包含数据的统计变换（statistical transformation，缩写stats），最后绘制在某个特定的坐标系（coordinate system，缩写coord）中，而分面（facet）则可以用来生成数据不同子集的图形。

简要汇总要素如下：数据(data)、几何对象(geom)、统计变换(stats)、图形属性(aes)、标尺(scale)、图层(layer)、坐标系(coord)与分面(facet)。

接下来以具体图片类型为例说明使用方法。

散点图

散点图指由坐标确定位置的散乱的点组成的图表。一般用于直接呈现数据，进一步还可以初步判断数据的趋势。通常用来表述两个连续变量之间的关系，图中的每个点表示目标数据集中的每个样本。

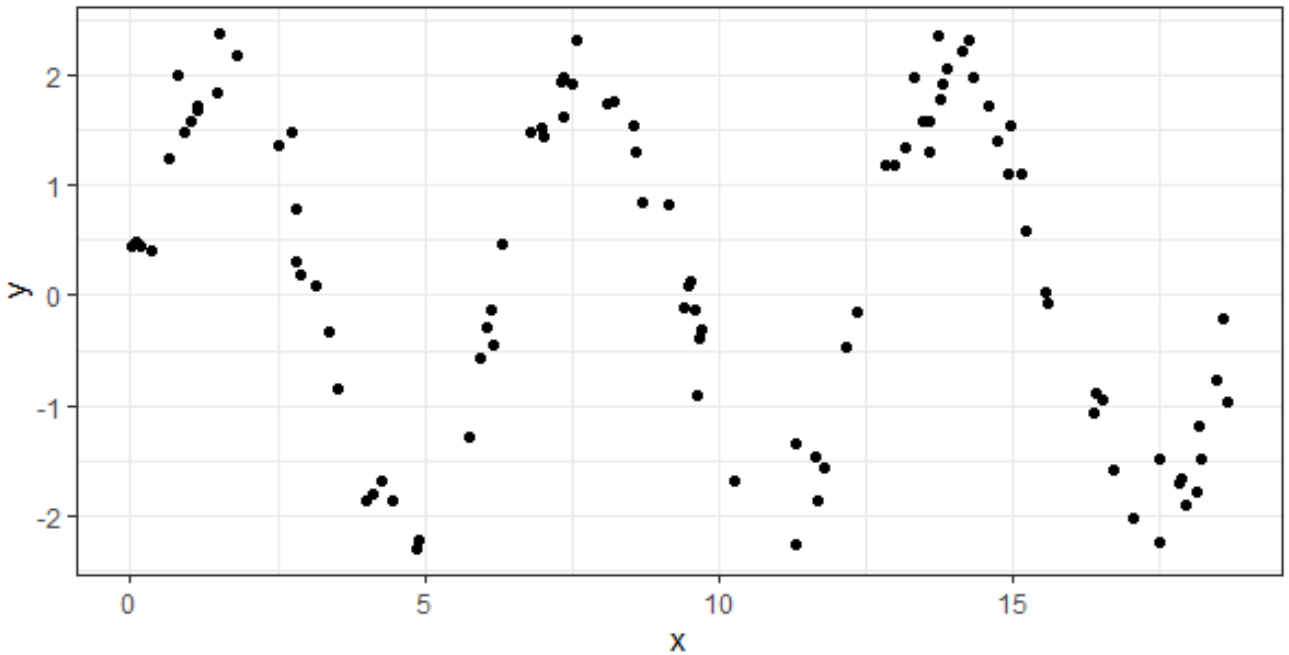

图 3: $y = 2\sin x + \epsilon$散点图

散点图的绘制可以直接使用plot()，也可以使用ggplot()，图形选择为geom_point()，或在layer()与stat()中设置geom = "point"。

格式如下：若数据集为data，其中自变量为x，因变量为y

(1)plot(x,y,type = "p")

(2)ggplot2::ggplot(data, mapping = aes(x = x,y = y))+geom_point()

其中(2)中的mapping也可以写在geom_point()中，效果是一致的。

就ggplot()的写法而言，ggplot(data)为数据集data创造了一个坐标系，之后可以向上叠加图层，但如果只有一个ggplot(data)则只会创建一个空的图片。

加号之后的geom_point()则创建了一个图层，图层为数据的散点图。实际上geom功能有很多，使用方法也是极为类似的，这里挑选的只是相对常用于统计学的。

mapping()则是任何一张ggplot()画出的图片中必不可少的元素。这个选项通常搭配aes()使用，用于定义数据集中的变量如何映射到视觉属性。aes(x = ,y =)中的x与y定义了映射到图片中x与y轴的变量。

Tips：散点图一般不适用样本量过大的情形。

最后介绍功能geom_smooth()，即在散点图上添加拟合的曲线。格式如下：

geom_smooth(mapping = NULL, data = NULL, method = NULL, formula = NULL,...)

例如图3，使用函数y~sin x拟合，结果如图4所示。

程序：ggplot(data = dataa,mapping = aes(x = x,y = y))+geom_point()+theme_bw()+geom_smooth(mapping = aes(x = x, y = y),formula = y~sin(x))

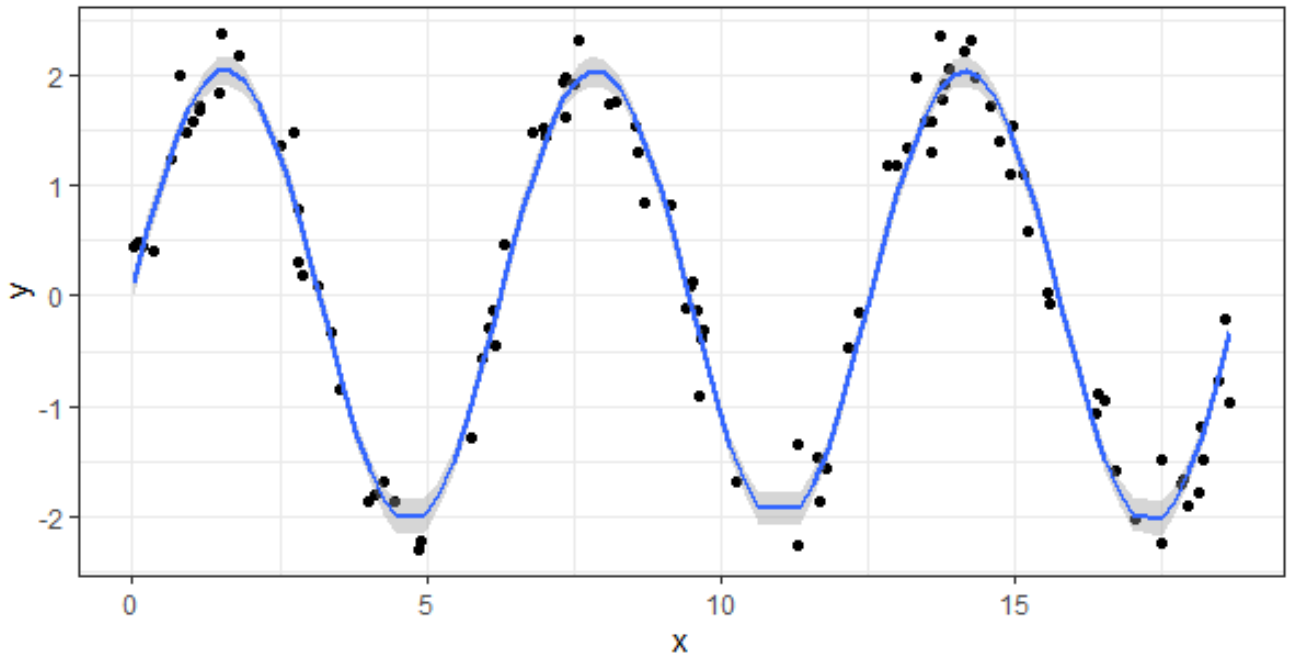

图 4: 带有拟合的散点图

柱状图

柱状图是一种使用矩形长度作为变量的统计图，用来比较两个或以上的值的差异，只有一个变量，常用于较小的数据集分析。除了R自带的barplot()外，还可以使用ggplot()中的geom_bar()，也可以在stat中指定geom = "bar"。其他语法与点状图类似。

如做出印度、巴西和美国新冠肺炎累计死亡人数，数据集命名为ncov2，程序如下：

ggplot(ncov2,mapping = aes(x = reorder(country, dead), y = dead, fill = country, label = dead))+geom_bar(stat = "identity", width = 0.5)+geom_text(aes(label = dead, vjust = -1))+xlab("country")+coord_cartesian(ylim = c(0,600000))

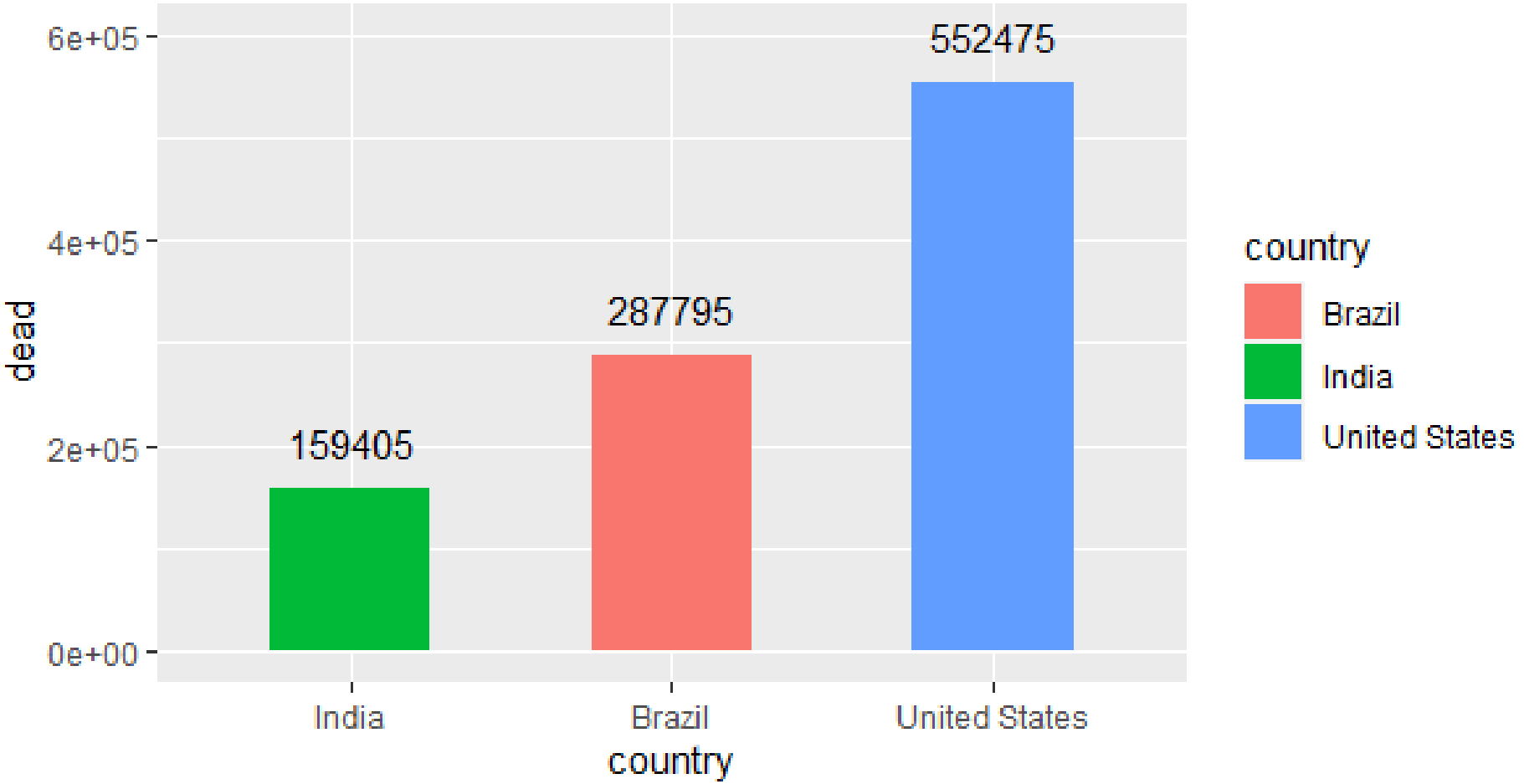
6e+05
4e+05
2e+05
0e+00
dead
552475
287795
159405
India
Brazil
United States
country
country
Brazil
India
United States

图 5: 新冠累计死亡人数

最后介绍并排与堆叠柱状图，这种柱状图相比于一般柱状图可以表现出更加高维的信息。

该参数设置在geom_bar(position =)中，其中“stack”表示堆叠柱状图，“dodge”表示并排柱状图。

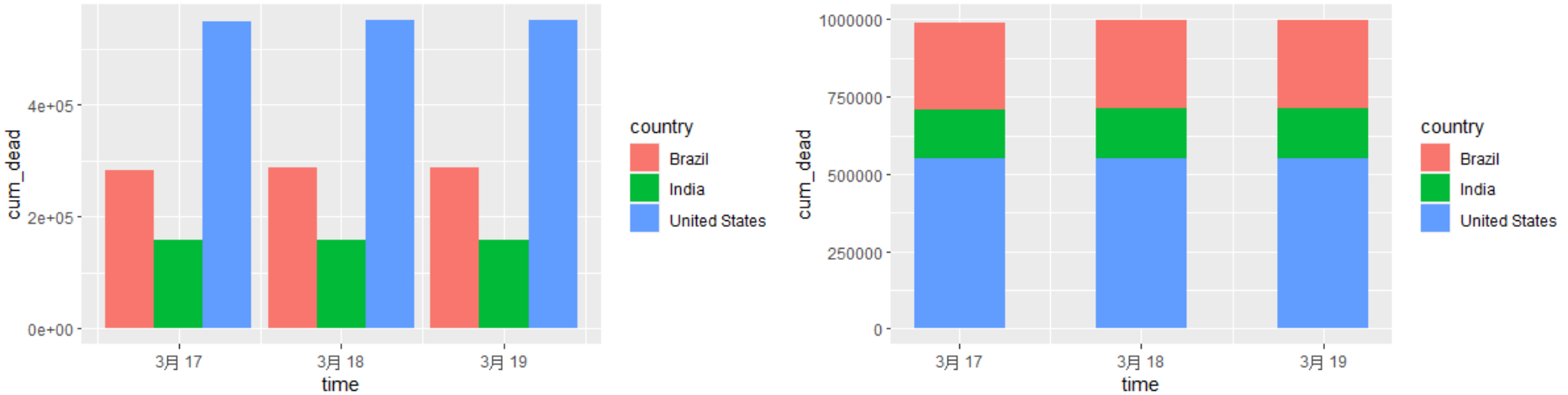

图 6: 并排(左)与堆叠(右)柱状图

折线图

折线图一般用于描述一维变量随着某一连续变量变化的情况，通常为时间，最适合描述时间序列数据的变化情况。若离散变量满足有序条件，则也可以使用折线图进行表示。调用功能geom_line()即可使用。

以美国和印度的新冠肺炎总确诊人数为例，数据集名称为data4，x轴为时间，y轴为累计确诊人数。程序如下：

```
ggplot(data4, aes(time,cum_confirm,colour =
country))+geom_line()+theme_bw()
```

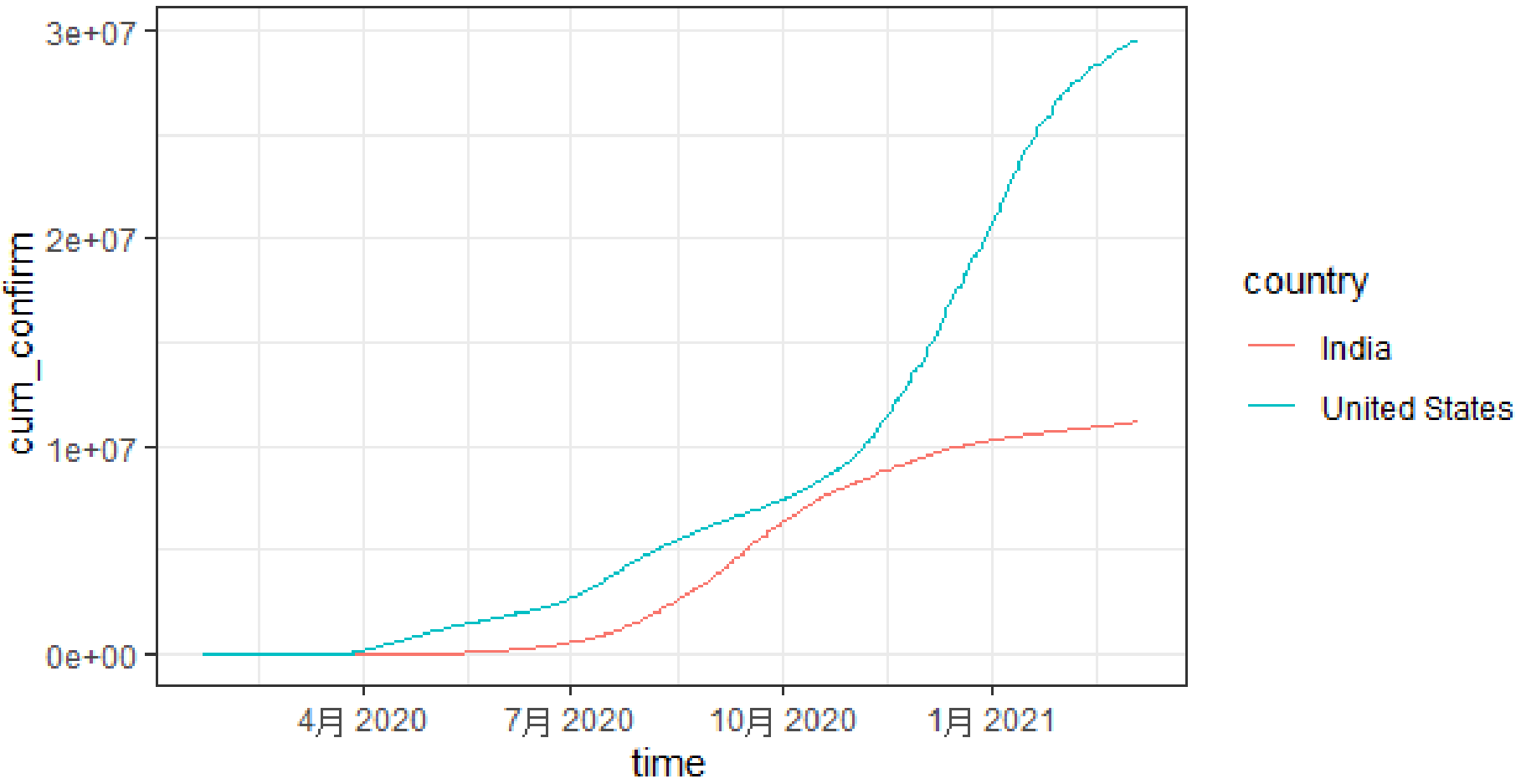
3e+07
2e+07
1e+07
0e+00
cum_confirm
4月 2020
7月 2020
10月 2020
1月 2021
time
country
India
United States

图 7: 新冠累计确诊人数

折线图还有一些参数列举如下。

有时需要对多组数据进行绘图，可以在aes中指定group。还可以在aes中指定数据点，可以使用fill或shape，fill得出的点形状相同，shape则不同。

position_dodge函数可以将重叠的数据点区分开，如geom_line(position = position_dodge(0.2))，即可控制两点向项左右平移0.2。

linetype则可以设置线的类型，如solid, dashed, dotted等，如geom_line(linetype = "dotted")。

直方图

直方图使用一系列高度不等的长方形条纹或线段表示数据分布情况，通常用于刻画连续型数据。调用geom_histogram()即可绘制。

格式如下：ggplot(数据集,aes(统计频数的变量))+geom_histogram(fill="柱的颜色",color = "边的颜色",bins=直方数)+scale_x_continuous(trans = "变换")+xlab("x轴名称")

例如使用小龙虾销量数据集crawfish，绘制小龙虾销量的直方图，要求对销售数量(sale)做对数变换。绘制结果如图8：

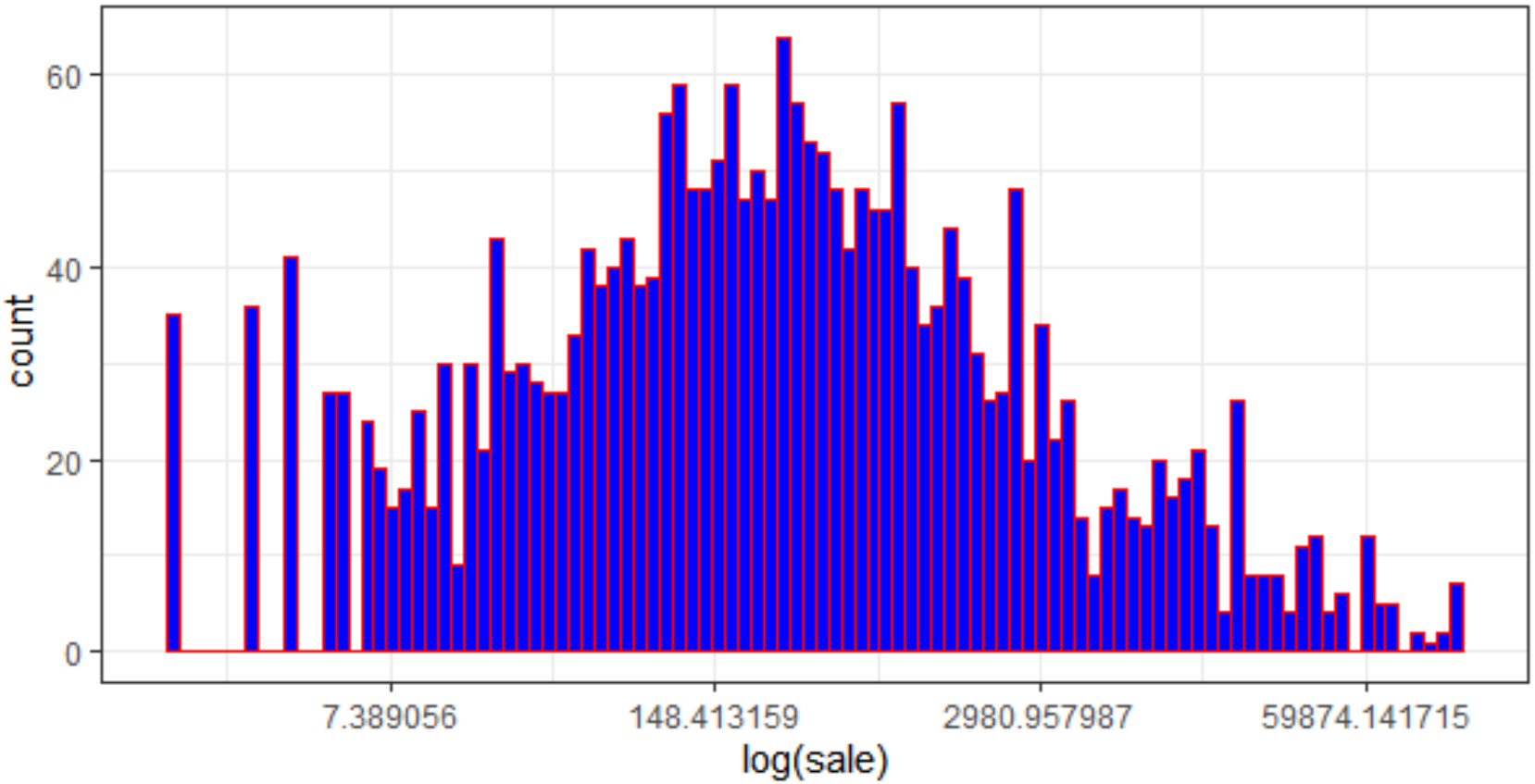

图 8: 对数变换后小龙虾销量直方图

有时数据出现离散程度较大或存在极端数据时，可以考虑采取Box-Cox变换，尤其以对数变换使用最为广泛。这里附上没有经过对数变换的直方图(图)，极端数据拉宽了x轴，使得直方图过于“拥挤”。

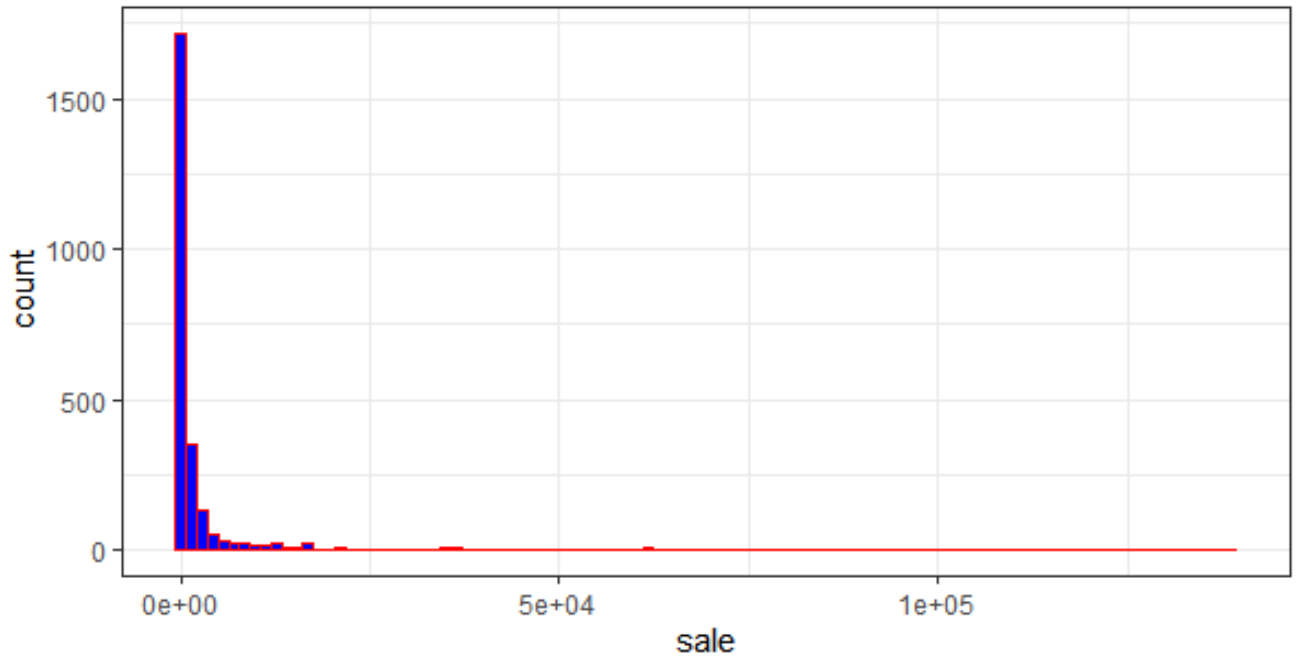

图 9: 未使用对数变换的直方图

箱线图

箱线图是一种显示数据离散情况的图表，通常包含分位数、上下边缘和异常值(如果存在)。箱线图结构如图10所示。

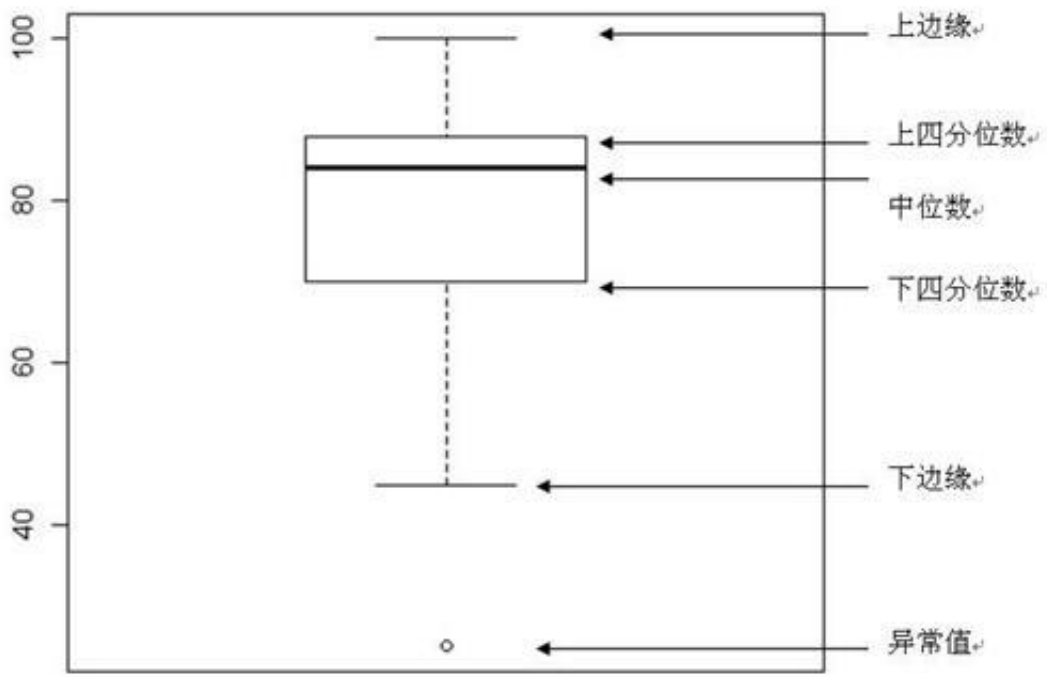

图 10: 箱线图结构

箱线图相对于柱状图包含的信息更多，不仅体现了变量的数量特征，还包含了数据离散程度的信息与异常值信息。

进一步，如果考虑箱的宽度，则箱线图又可以携带一个维度的信息，即箱所代表类别的样本量。

ggplot2中绘制箱线图的选项为geom_boxplot()，但直接使用无法绘制出上下边缘。绘制上下边缘的选项在stat_boxplot()中，且该项需要写在geom_boxplot()之前，否则会出现图层覆盖的问题。

具体格式为：ggplot(data = 数据集,mapping = aes(x = 自变量,y = 因变量, fill = 群组))+labs(x = "x轴" , y = "y轴")+stat_boxplot(geom = "errorbar")+geom_boxplot(varwidth = TRUE)

如果数据离散程度很大，可以对数据(指标)进行对数变换，方法与直方图中相关操作是完全一样的。

仍然以七座城市小龙虾销售数据集crawfish为例，绘制销量的箱线图，因为数据极端值较多采用了对数变换，按照销量高低降序排列。程序如下，输出如图11所示：

```
ggplot(data = crawfish,mapping = aes(x = reorder(city,-sale), y = sale, fill = city))+labs(x = "city" , y = "log(sale)")+stat_boxplot(geom = "errorbar")+geom_boxplot(varwidth = TRUE)+scale_y_continuous(trans = "log")+theme_update()
```

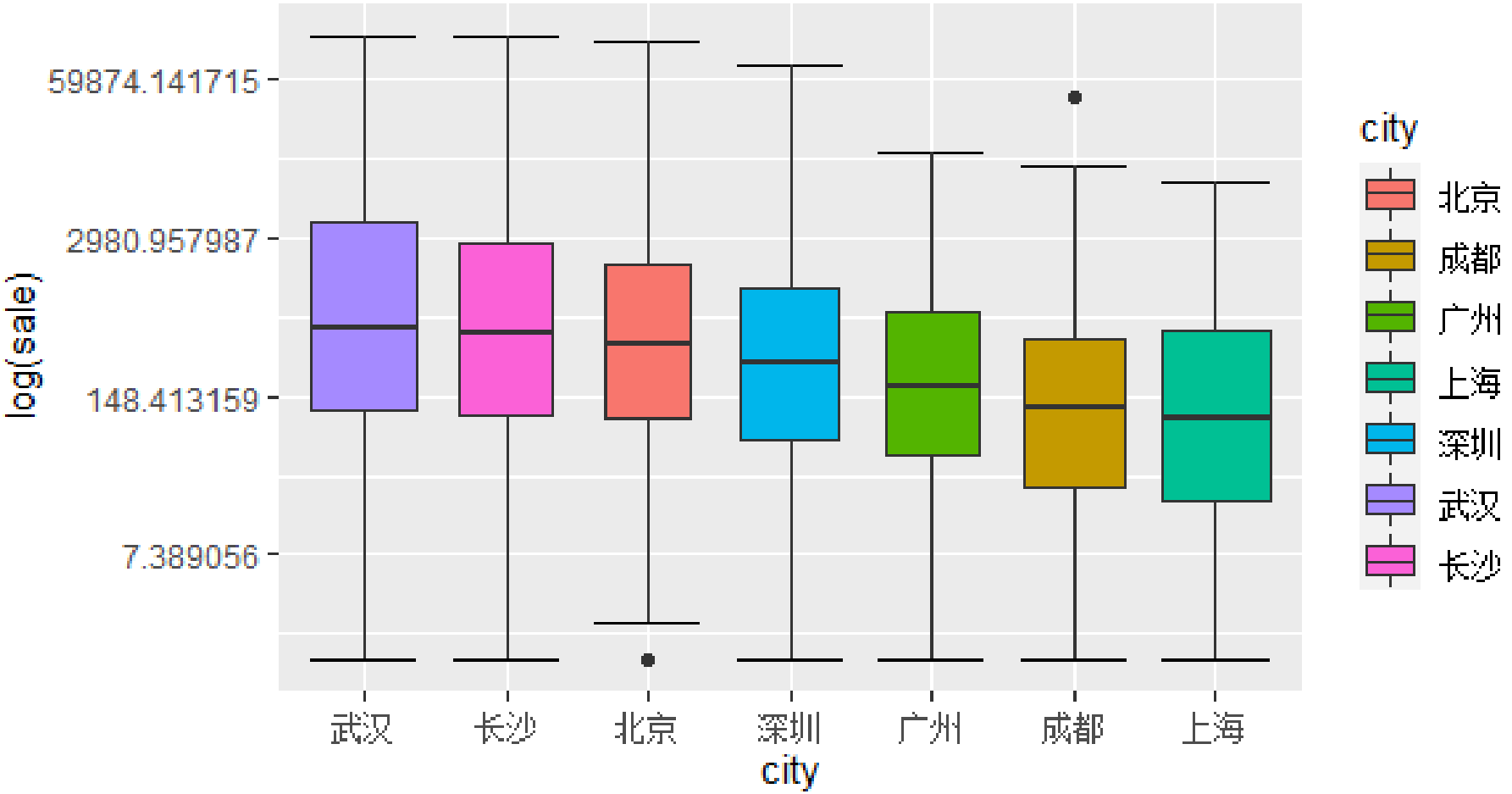

图 11: 七座城市小龙虾销量箱线图

如果不做对数变换，箱线图就会被挤压到底部，图片将不够美观。图12是未经过对数变换绘制的箱线图。

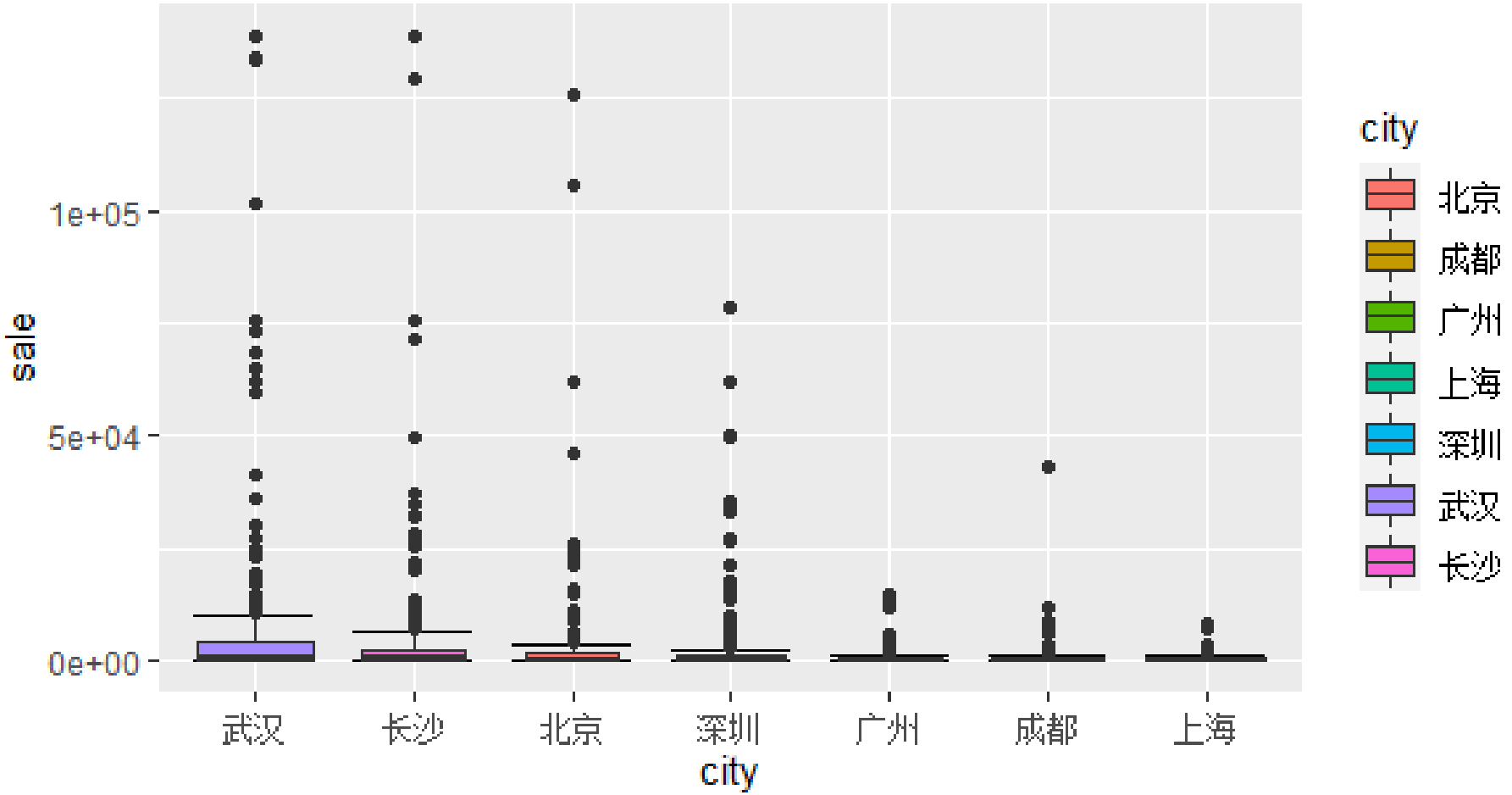

图 12: 未经过对数变换的箱线图

绘制地图

根据R for Data Science提供的代码使用ggplot2绘制地图，如图13所示：

```
world <- map_data("world2")
ggplot(world, aes(long, lat, group = group)) + geom_polygon(fill =
terrain.colors(99385), colour = "black") +
coord_quickmap()+theme_minimal()
```

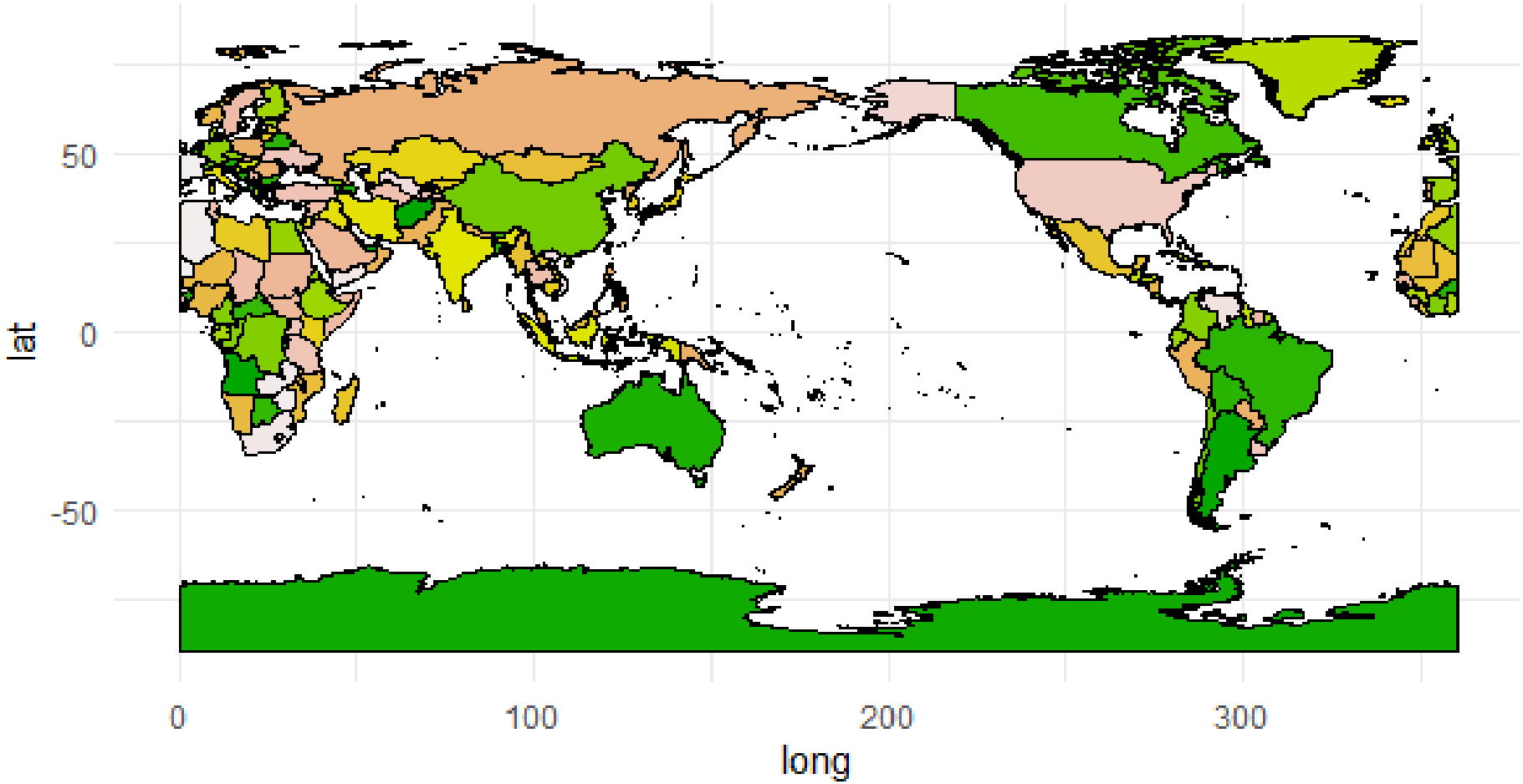

图 13: 世界地图

总结

本节主要讲述了ggplot的绘图原理以及几种常见统计图的绘制。事实上统计图绘制有很多“讲究”，这里很难一一列举，只能概述几种基本统计图的绘制方法，更多的知识与经验还需要在实践中不断积累。

更多有关作图与数据可视化的内容可以参考狗熊会《**丑图百讲**》系列推文。