

# 2

## Descriptive Statistics

# 统计描述

用图形和汇总统计量描述样本数据



统计学是科学的语法。

*Statistics is the grammar of science.*

—— 卡尔·皮尔逊 (Karl Pearson) | 英国数学家 | 1857 ~ 1936



- ▶ `joypy.joyplot()` 绘制山脊图
- ▶ `numpy.percentile()` 计算百分位
- ▶ `pandas.plotting.parallel_coordinates()` 绘制平行坐标图
- ▶ `seaborn.boxplot()` 绘制箱型图
- ▶ `seaborn.heatmap()` 绘制热图
- ▶ `seaborn.histplot()` 绘制频数/概率/概率密度直方图
- ▶ `seaborn.jointplot()` 绘制联合分布和边缘分布
- ▶ `seaborn.kdeplot()` 绘制 KDE 核概率密度估计曲线
- ▶ `seaborn.lineplot()` 绘制线图
- ▶ `seaborn.lmplot()` 绘制线性回归图像
- ▶ `seaborn.pairplot()` 绘制成对分析图
- ▶ `seaborn.swarmplot()` 绘制蜂群图
- ▶ `seaborn.violinplot()` 绘制小提琴图

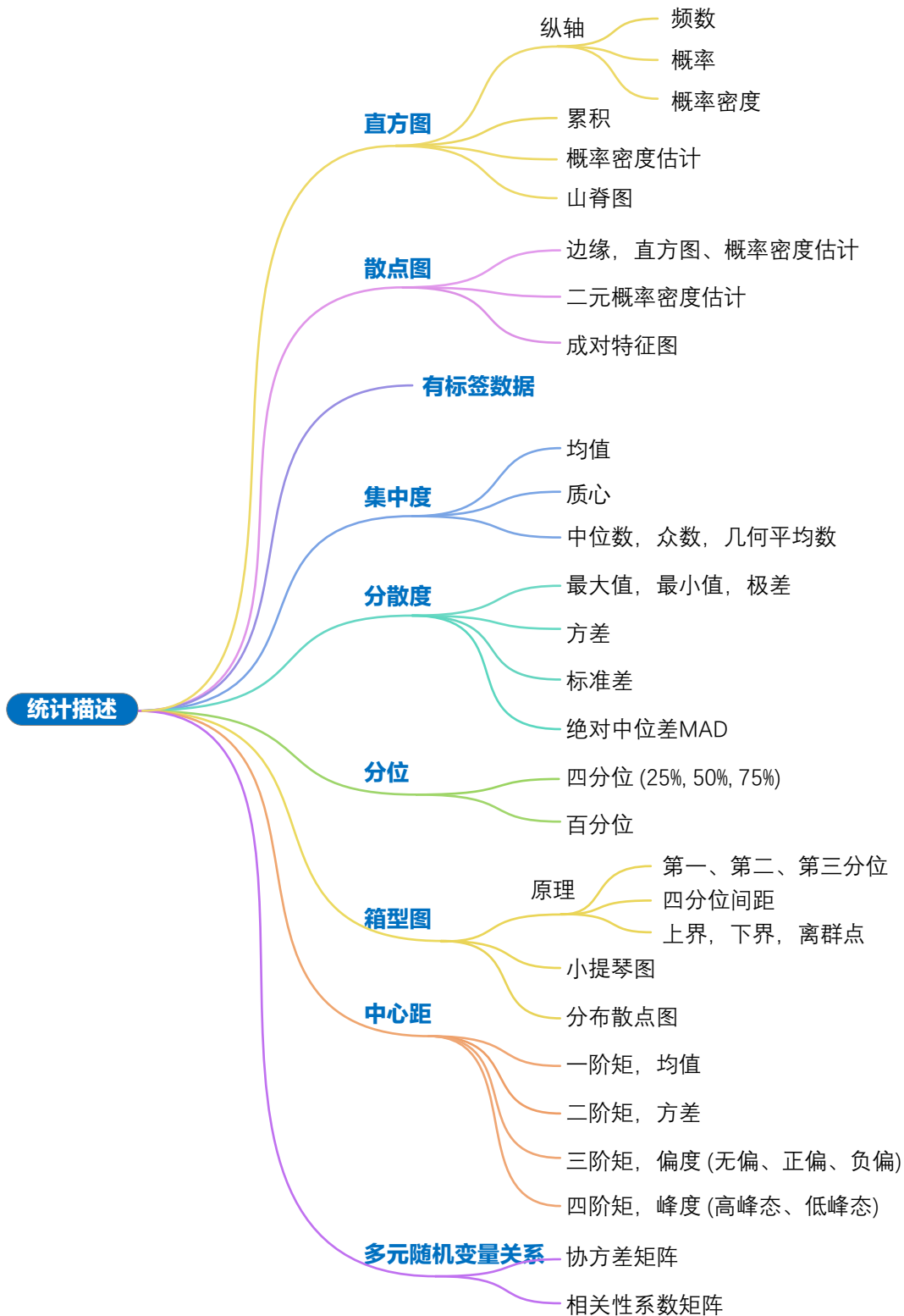
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

## 2.1 统计两大工具：描述、推断

如图 1 所示，本书中统计版图可以分为两大板块——描述、推断。

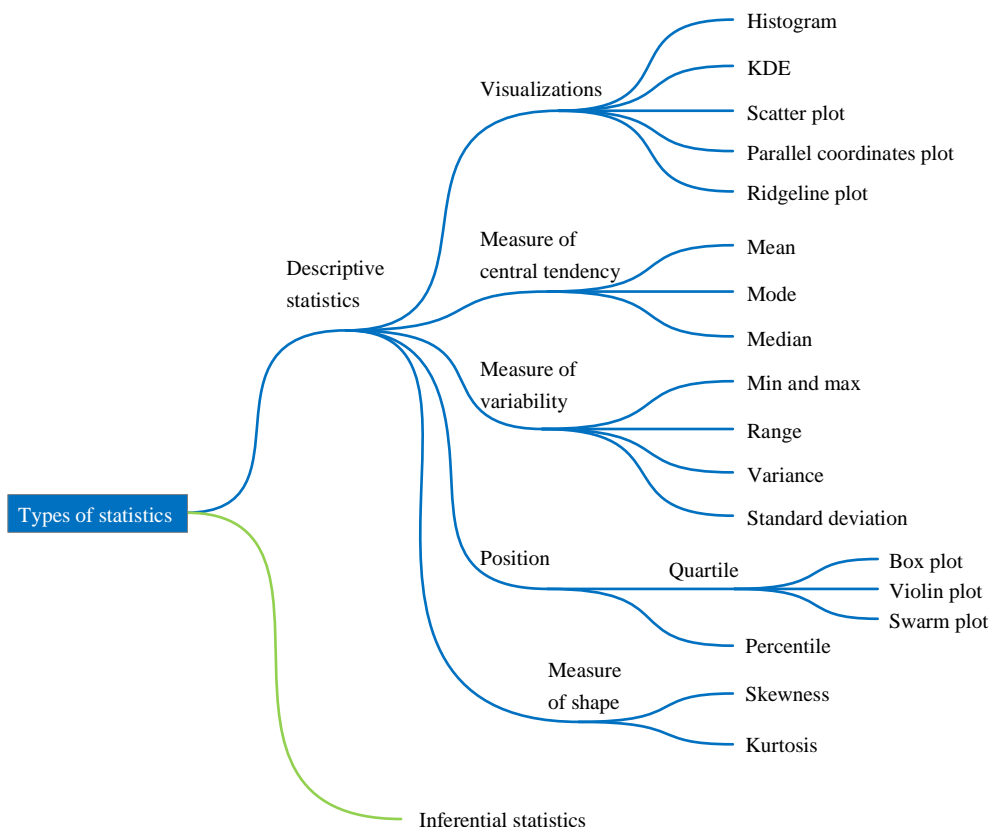


图 1. 两大类统计工具的分类

**统计描述** (descriptive statistics) 是指对数据进行整体性的描述和概括，以了解数据的特征和结构。统计描述旨在通过一些表格、图像、量化汇总来呈现数据的基本特征，比如中心趋势、离散程度、分布形态等。统计描述通常是数据分析的第一步，可以帮助我们了解数据的基本情况，判断数据的可靠性、准确性和有效性。

**统计推断** (statistical inference) 根据样本数据推断总体特征。统计推断是在对样本数据统计描述的基础上，对总体未知量化特征做出概率形式的推断。显然，统计推断的数学基础工具就是概率论。本书后续概率、高斯、随机这三个板块介绍概率论这个工具箱中常用工具。之后，我们将用频率派、贝叶斯派两个板块介绍统计推断。

➡ 请大家学习这一章时，重温《矩阵力量》第 22 章，回顾如何从线性代数视角看各种统计量。

本章主要介绍统计描述。常见的统计描述方法包括：

- ▶ 统计图表：可视化数据分布情况和异常值，比如直方图、箱线图、散点图等。
- ▶ 中心趋势：比如均值、中位数和众数，量化数据的集中程度。
- ▶ 离散程度：比如极差、方差、标准差、极差和四分位数，描述数据的分散程度。
- ▶ 分布形态：比如偏度和峰度，分析数据的分布形态。
- ▶ 协同关系：包括协方差矩阵、相关性系数矩阵，量化多元随机变量之间的关系。

下面，我们开始本章学习。

## 2.2 直方图：单特征数据分布

鸢尾花花萼长度的数据看上去杂乱无章，我们可以利用一些统计工具来分析这组数据，比如直方图。直方图 (histogram) 由一系列矩形组成，它的横轴为组距，纵轴可以为**频数** (frequency, count)、**概率** (probability)、**概率密度** (probability density 或 density)。直方图可视化样本分布情况，同时展示均值、众数、中位数的大致位置以及标准差宽度等。直方图也可以用来判断数据是否存在**离群值** (outlier)。



《数据有道》一册将专门讲解判断离群值的常用算法。

图 2 所示为鸢尾花花萼长度数据直方图。直方图通常将样本数据分成若干个连续的区间，也称为“箱子”或“组”。直方图中矩形的纵轴高度可以对应频数、概率或概率密度。

▲ 再次强调，一般情况，直方图的纵轴有三个选择——频数、概率和概率密度。

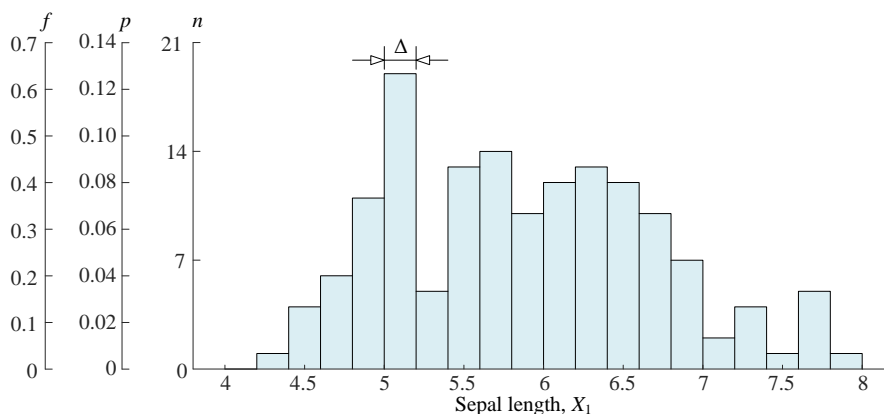


图 2. 鸢尾花花萼长度，频数、概率和概率密度的关系

下面聊聊频数、概率和概率密度分别是什么。

## 区间

花萼长度的最小值和最大值落在  $[4, 8]$  这个区间。如图 3 所示，将这个区间等分为 20 个区间。区间个数称为组数，记做  $M$ 。每个区间对应的宽度叫做组距，记做  $\Delta$ 。本例中组数  $M = 20$ ，组距  $\Delta = 0.2 \text{ cm} = 4 \text{ cm}/20$ 。

图 3 第一列给出的是每个组距所在的区间。大家已经看到最后一个区间  $[7.8, 8.0]$  为闭区间，其他区间均为左闭右开。

区间	频数 $n$	累积频数 $\text{cumsum}(n)$	概率 $p$	累积概率 $\text{cumsum}(p)$	概率密度 $f$
[4.2, 4.4)	1	1	0.007	0.007	0.033
[4.4, 4.6)	4	5	0.027	0.033	0.133
[4.6, 4.8)	6	11	0.040	0.073	0.200
[4.8, 5.0)	11	22	0.073	0.147	0.367
[5.0, 5.2)	19	41	0.127	0.273	0.633
[5.2, 5.4)	5	46	0.033	0.307	0.167
[5.4, 5.6)	13	59	0.087	0.393	0.433
[5.6, 5.8)	14	73	0.093	0.487	0.467
[5.8, 6.0)	10	83	0.067	0.553	0.333
[6.0, 6.2)	12	95	0.080	0.633	0.400
[6.2, 6.4)	13	108	0.087	0.720	0.433
[6.4, 6.6)	12	120	0.080	0.800	0.400
[6.6, 6.8)	10	130	0.067	0.867	0.333
[6.8, 7.0)	7	137	0.047	0.913	0.233
[7.0, 7.2)	2	139	0.013	0.927	0.067
[7.2, 7.4)	4	143	0.027	0.953	0.133
[7.4, 7.6)	1	144	0.007	0.960	0.033
[7.6, 7.8)	5	149	0.033	0.993	0.167
[7.8, 8.0]	1	150	0.007	1.000	0.033

图 3. 鸢尾花花萼长度直方图数据

**⚠ 注意**，一般情况，除了最后一个区间之外，其他区间包含左侧端点，不含右侧端点，即左闭右开区间。最后一个区间为闭区间。

## 频数

频数，也叫次数，是指在一定范围内样本数据的数量。显然，频数为非负整数。如图 3 所示，落在  $4.2 \sim 4.4$  这个区间内的样本只有 1 个。而落在  $5 \sim 5.2$  这个区间内的样本多达 19 个。

数出落在第  $i$  个区间内的样本数量，定义为频数  $n_i$ 。图 3 第二列给出的就是频数。

显然，所有频数  $n_i$  之和为样本总数  $n$ ：

$$\sum_{i=1}^M n_i = n \quad (1)$$

## 概率

频数  $n_i$  除以样本总数  $n$  的结果做概率  $p_i$ :

$$p_i = \frac{n_i}{n} \quad (2)$$

图 3 第四列对应概率。容易知道概率值  $p_i$  的取值范围  $[0, 1]$ 。概率值代表“可能性”。

直方图的纵轴为概率时，直方图也叫归一化直方图。这是因为所有区间概率  $p_i$  之和为 1:

$$\sum_{i=1}^M p_i = \sum_{i=1}^M \frac{n_i}{n} = \frac{n_1 + n_2 + \cdots + n_M}{n} = 1 \quad (3)$$

## 概率密度

概率  $p_i$  除以组距  $\Delta$  得到的是**概率密度** (probability density)  $f_i$ :

$$f_i = \frac{p_i}{\Delta} = \frac{n_i}{n\Delta} \quad (4)$$

纵轴为概率密度的直方图，所有矩形面积之和为 1:

$$\sum_{i=1}^M f_i \Delta = \sum_{i=1}^M \frac{p_i}{\Delta} \Delta = \sum_{i=1}^M \frac{n_i}{n} = 1 \quad (5)$$

观察图 3，我们可以发现频数、概率、概率密度这三个值成正比关系。不同的是，看频数、概率时，我们在乎的是直方图矩形高度；而看概率密度时，我们关注的是矩形面积。

**▲ 注意**，概率密度不是概率；但是，概率密度本身也反映数据分布的疏密情况。

## 累积

图 3 中第三和第五列分别为**累积频数** (cumulative frequency) 和**累积概率** (cumulative probability)。累积频数就是将从小到大各区间的频数逐个累加起来，累积频数的最后一个值是样本总数。

类似地，我们可以得到累积概率，累积概率的最后一个值为 1。

## 绘制直方图

图 4 所示为利用 `seaborn.histplot()` 绘制的鸢尾花四个量化特征数据直方图，纵轴为频数。直方图的形状可以反映数据的分布情况，比如对称分布、左偏分布、右偏分布等。直方图可以通过调整箱子的数量和大小来改变分组的细度和粗细，以适应不同的数据特征。直方图也经常与其他统计图表一起使用，比如箱线图、散点图、概率密度估计曲线等，以便更深入地理解数据的特征和结构。

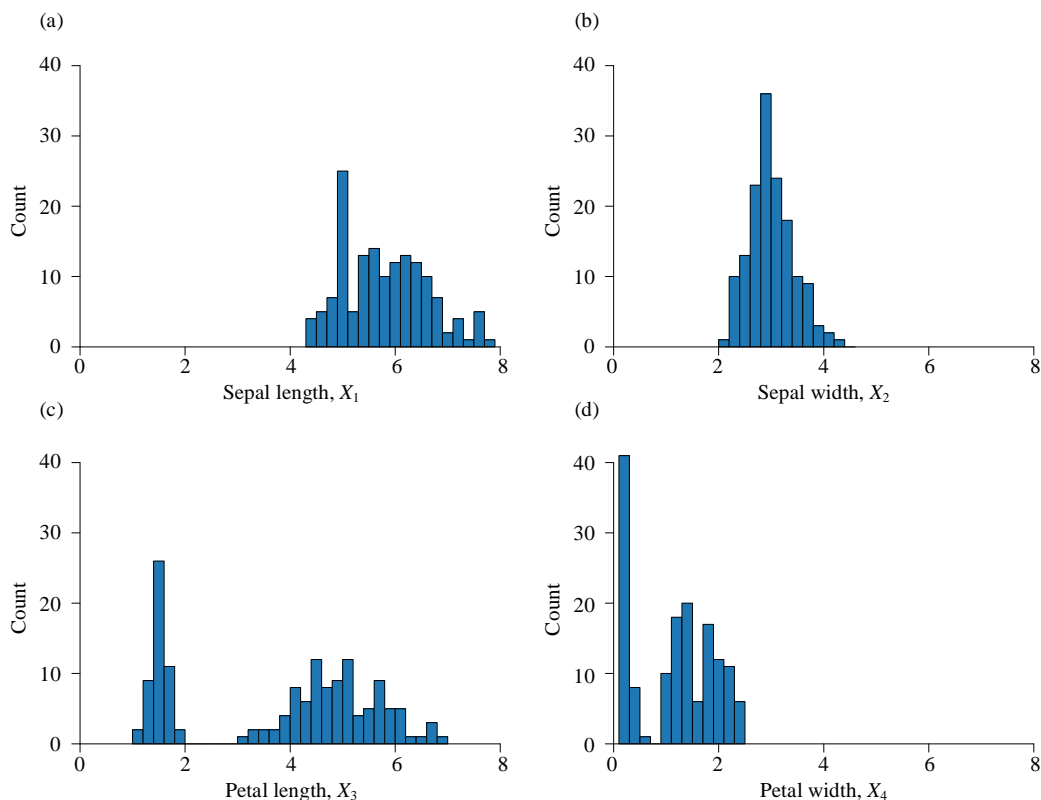


图 4. 鸢尾花四个特征数据的直方图，纵轴为频数

图 5 所示为同一个坐标系下对比鸢尾花四个特征数据直方图。图 5 (a) 纵轴为频数，图 5 (b) 纵轴为概率密度。

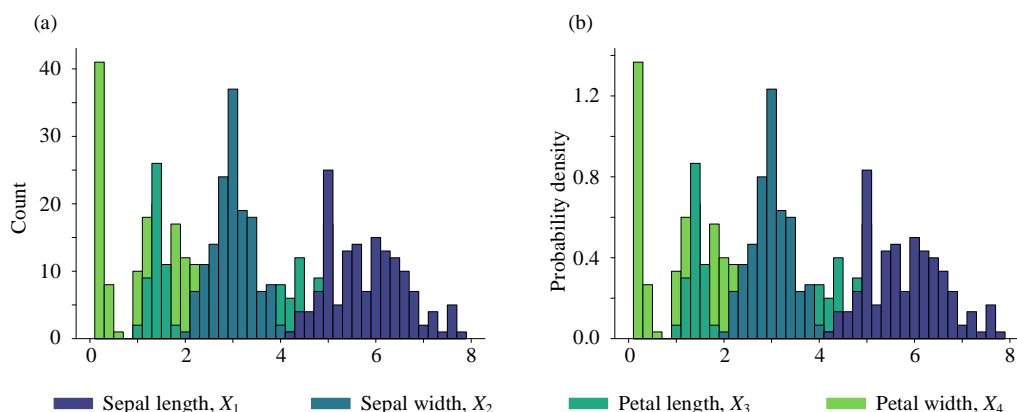


图 5. 直方图，比较频数和概率密度

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

## 累积频数、累积概率

图 6 对比四个鸢尾花特征样本数据的累积频数图、累积概率图。如图 6 (a) 所示，累积频数的最大值为 150，即鸢尾花数据集样本个数。如图 6 (b) 所示，累积概率的最大值为 1。

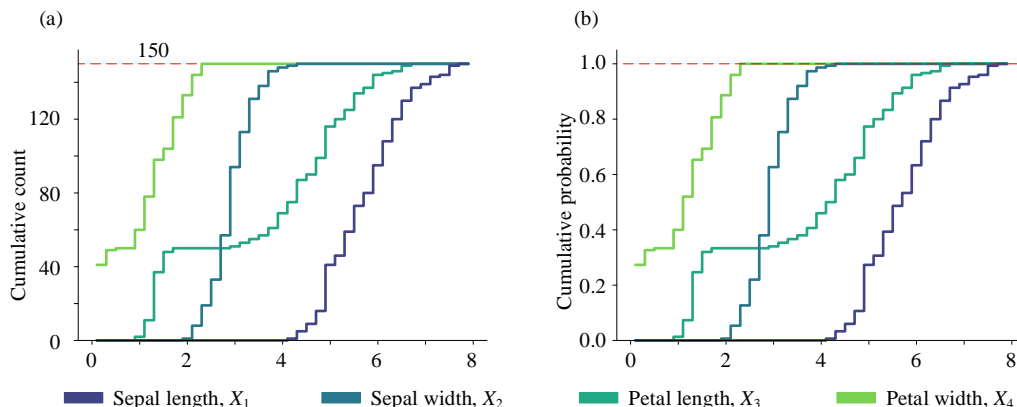


图 6. 累积频数图，累积概率图

## 多边形图、概率密度估计

**多边形图** (polygon) 将直方图矩形顶端中点连接，得到如图 7 (a) 所示线图。

**▲ 注意**，多边形图的纵轴和直方图一样有很多选择，图 7 (a) 给出的纵轴为概率密度。

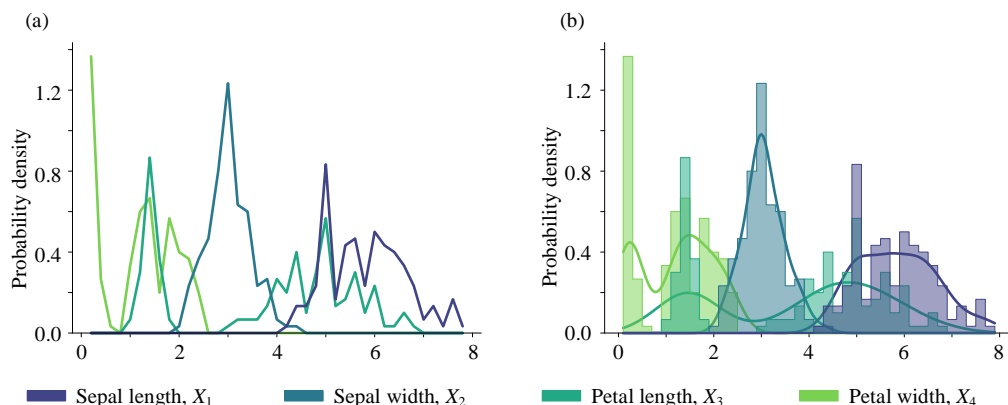


图 7. 比较多边形图和和概率密度估计曲线

**核密度估计** (Kernel Density Estimation, KDE) 是对直方图的扩展，如图 7 (b) 中曲线是通过核密度估计得到的概率密度函数图像。



概率密度函数描述的是随机变量在某个取值点的概率密度，是描述随机变量分布的基本函数之一。在实际问题中，往往无法直接获得概率密度函数，因此需要通过概率密度估计来估计概率密度函数。概率密度估计可以通过多种方法来实现，比如直方图法、核密度估计法、最大似然估计法等。其中，核密度估计法是最常用的方法之一，它假设数据的概率密度函数是由一些基本的核函数叠加而成，然后根据数据样本来确定核函数的带宽和数量，最终得到概率密度函数的估计值。



本书第 17 章将专门讲解概率密度估计。

## 山脊图

**山脊图** (ridgeline plot) 是由多个重叠的概率密度线图构成，这种可视化方案形式上紧凑。图 8 所示的山脊图采用 joypy 绘制。

山脊图的基本思想是，将数据沿着 y 轴的方向上的一条带状区间内进行展示，使得数据的分布曲线能够清晰地显示出来，并且不会重叠和遮挡。在山脊图中，每个变量的分布曲线通常用核密度估计法或直方图法进行估计，然后按照一定的顺序进行平移和叠加。

山脊图常用于探索多个变量之间的关系和相互作用，以及发现变量的共同分布特征和异常点。它可以用于可视化各种类型的数据，比如时间序列数据、连续变量数据、分类变量数据等。



本书第 20、21 章将利用山脊图可视化后验概率连续变化。

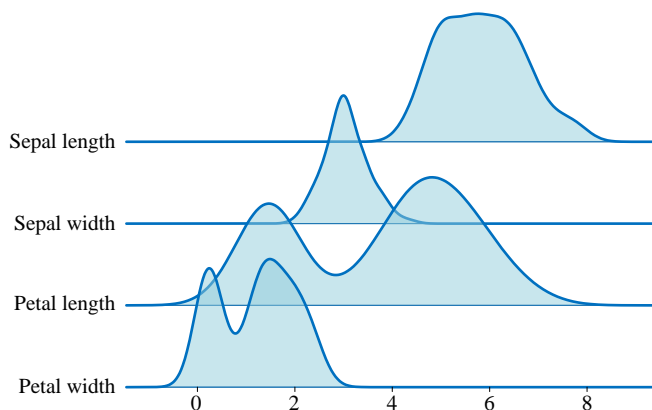


图 8. 鸢尾花数据山脊图

## 2.3 散点图：两特征数据分布

二维数据最基本的可视化方案是**散点图** (scatter plot)，如图 9 (a) 所示。散点图常用于展示两个变量之间的关系和相互作用。散点图将每个数据点表示为二维坐标系上的一个点，其中一个变量沿 x 轴方向表示，另一个变量沿 y 轴方向表示，每个点的位置反映了两个变量之间的数值关系。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

散点图可以用于研究两个变量之间的线性关系、非线性关系或者无关系。如果两个变量之间存在线性关系，那么散点图中的点会形成一条斜率为正或负的直线。如果两个变量之间存在非线性关系，那么散点图中的点会形成一条曲线或者散布在二维坐标系的不同区域。如果两个变量之间无关系，那么散点图中的点会均匀地分布在二维坐标系中。

散点图常用于探索数据中的异常值、趋势和模式，并且可以发现变量之间的相互作用和关联性。

在散点图的基础上，我们可以拓展得到一系列衍生图像。比如图 9 (a) 中，我们可以看到两幅**边缘直方图** (marginal histogram)，它们分别描绘花萼长度和花萼宽度这两个特征的分布状况。图 9 (b) 增加了简单线性回归图像和边缘 KDE 概率密度曲线。

**边缘概率** (marginal probability) 和**联合概率** (joint probability) 相对应。联合概率针对两个及以上随机变量的分布，边缘概率对应单个随机变量。图 9 中两幅图一方面展示两个随机变量的联合分布，同时展示每个随机变量的单独分布。大家会在本书后续经常看到类似的可视化方案。

➔ 本书第 24 章将介绍线性回归。此外，《数据有道》一册将专门讲解各种常见回归模型。

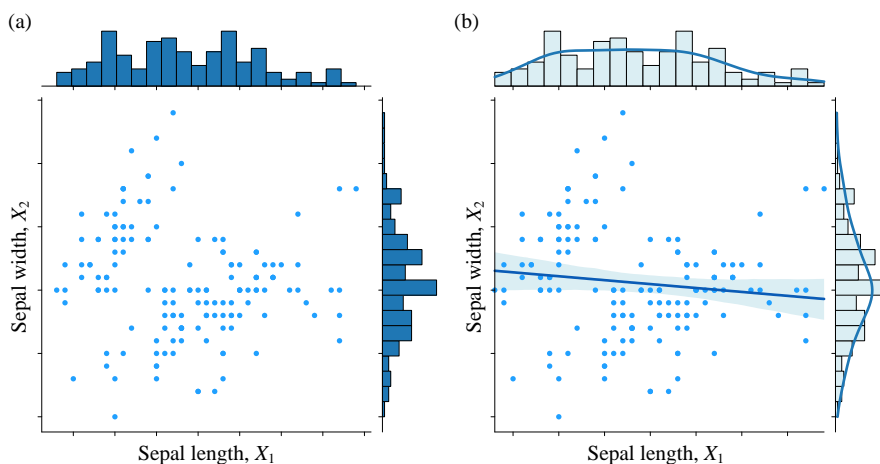


图 9. 二维数据散点图及扩展

## 二维概率密度

我们可以将上一节的直方图和 KDE 概率密度曲线，都拓展到二维数据。图 10 (a) 所示为二维直方图热图，热图每一个色块的颜色深浅代表该区域样本数据的频数。图 10 (b) 为二维 KDE 概率密度曲面等高线图。

图 11 (a) 在直方图热图上增加了边缘直方图，图 11 (b) 在二维联合概率密度曲面等高线图上增加了边缘概率密度曲线。

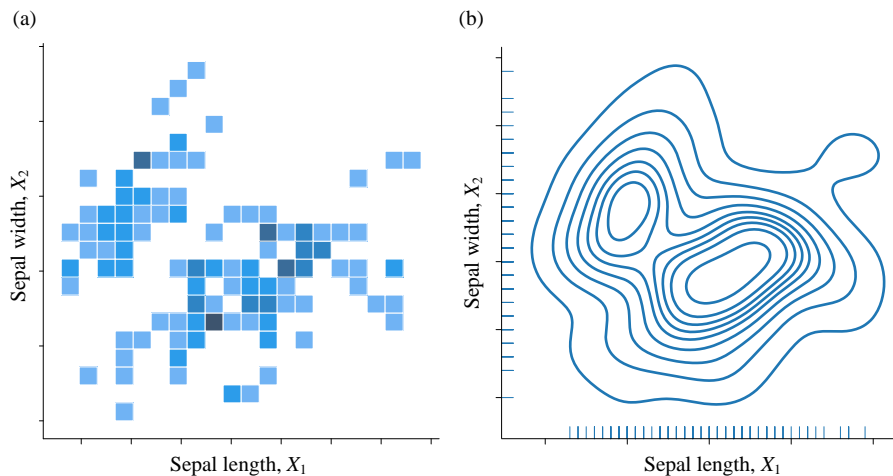


图 10. 二维数据直方图热图，二维 KDE 概率密度曲面等高线

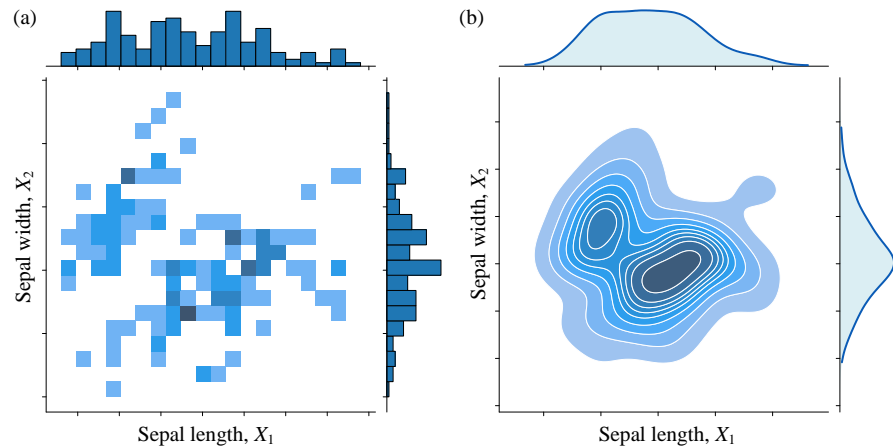


图 11. 直方图热图和概率密度曲面等高线拓展

### 成对特征图

本节介绍的几种二维数据统计分析可视化方案也可以拓展到多维数据，图 12 所示为鸢尾花数据成对特征分析图。相信大家对图 12 已经完全不陌生，我们在《数学要素》、《矩阵力量》都讲过成对特征分析图。

图 12 这幅图像有  $4 \times 4$  个子图，主对角线上的图像为鸢尾花单一特征数据直方图，右上角六幅子图为成对数据散点图，左下角六幅子图为概率密度曲面等高线图。

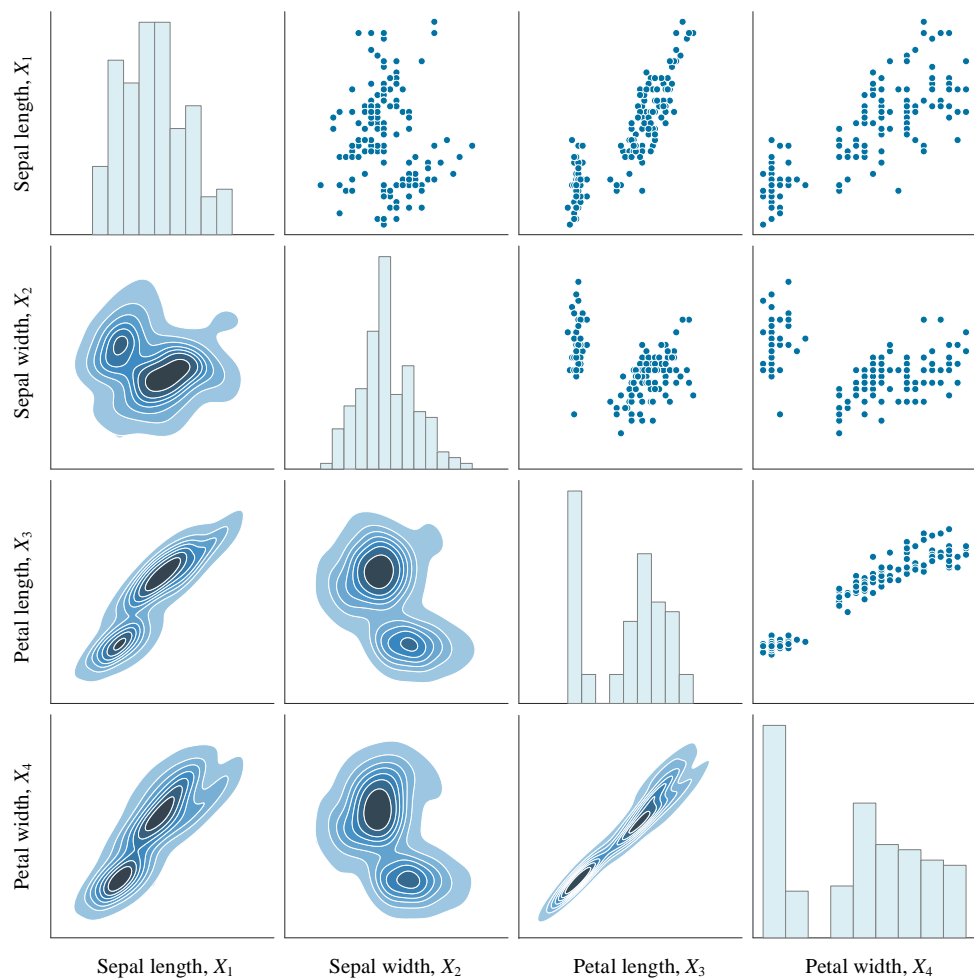


图 12. 鸢尾花数据成对特征分析图

## 2.4 有标签数据的统计可视化

《矩阵力量》专门区分过**有标签数据** (labeled data) 和**无标签数据** (unlabeled data)，如图 13 所示。

鸢尾花数据就是典型的有标签数据。鸢尾花数据有三个标签——**山鸢尾** (setosa)、**变色鸢尾** (versicolor) 和**维吉尼亚鸢尾** (virginica)。每一行样本点都对应一类鸢尾花分类。

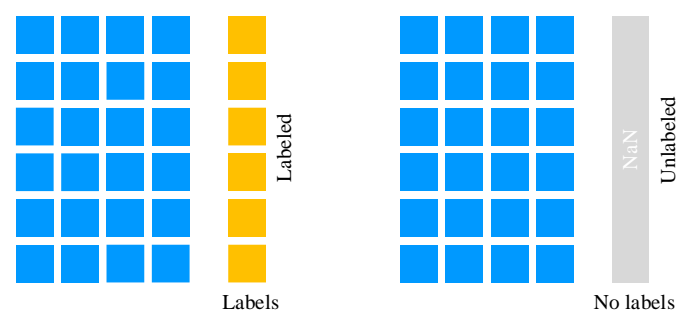


图 13. 根据有无标签分类数据

图 14 所示为含有标签分类的直方图。不同类别的鸢尾花数据采用不同颜色的直方图。图 14 的纵轴可以是频数、概率、概率密度。此外，考虑到分类标签，概率、概率密度也可以对应条件概率。举个例子，如果图 14 的纵轴对应“条件”概率密度的话，每幅子图中不同颜色的直方图面积均为 1。

“条件”听起来很迷惑，实际上大家在生活中经常用到。比如，高中二年 3 班男生的平均身高，“高中二年 3 班”和“男生”都是条件。大家可能已经发现，“条件”实际上就是限定讨论范围。

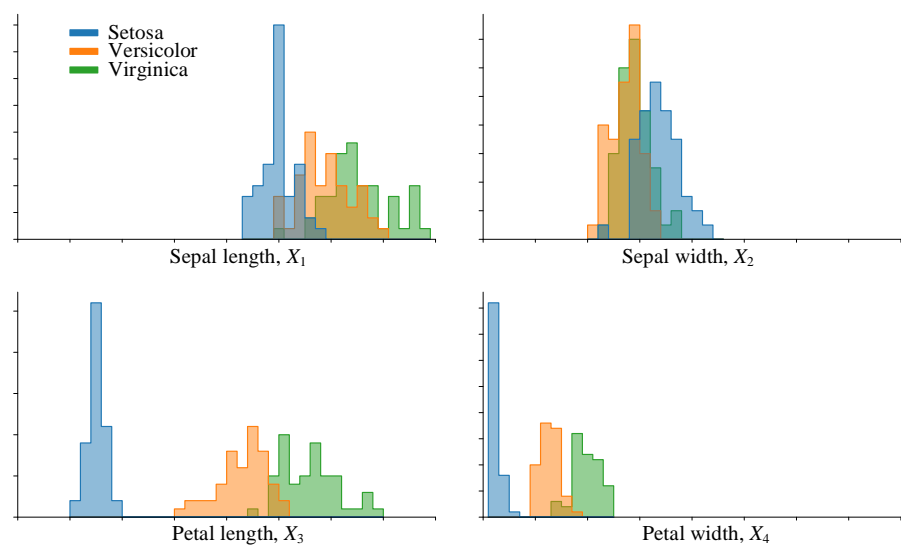


图 14. 直方图，考虑鸢尾花分类标签

图 15 所示为考虑分类的山脊图。我们也可以把这种可视化方案应用到二维数据可视化，如图 16 所示。图 17 所示为考虑标签的成对特征图。

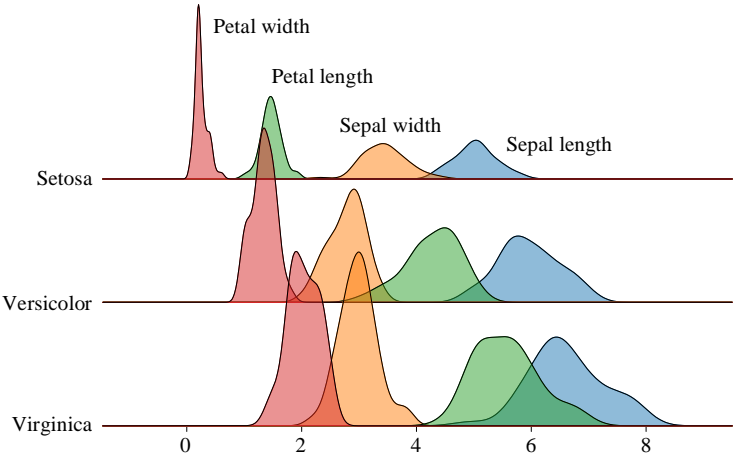


图 15. 鸢尾花山数据山脊图，特征分类

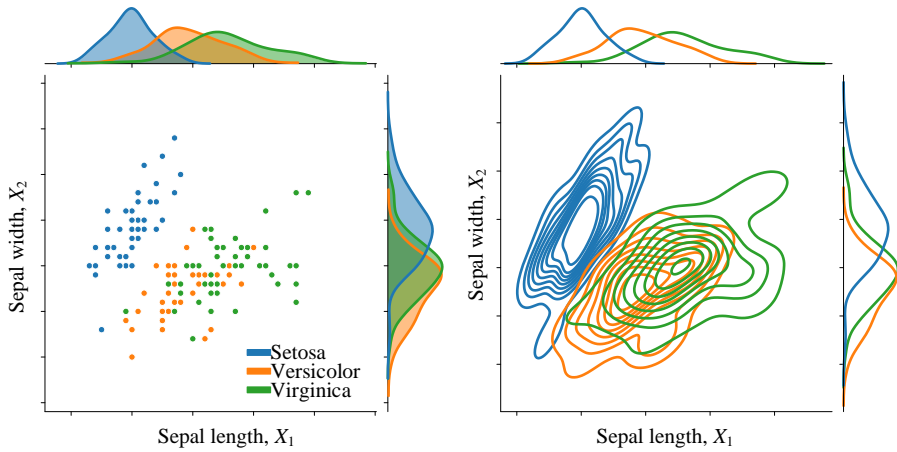


图 16. 二维数据散点图，KDE 概率密度曲面等高线，考虑鸢尾花分类标签

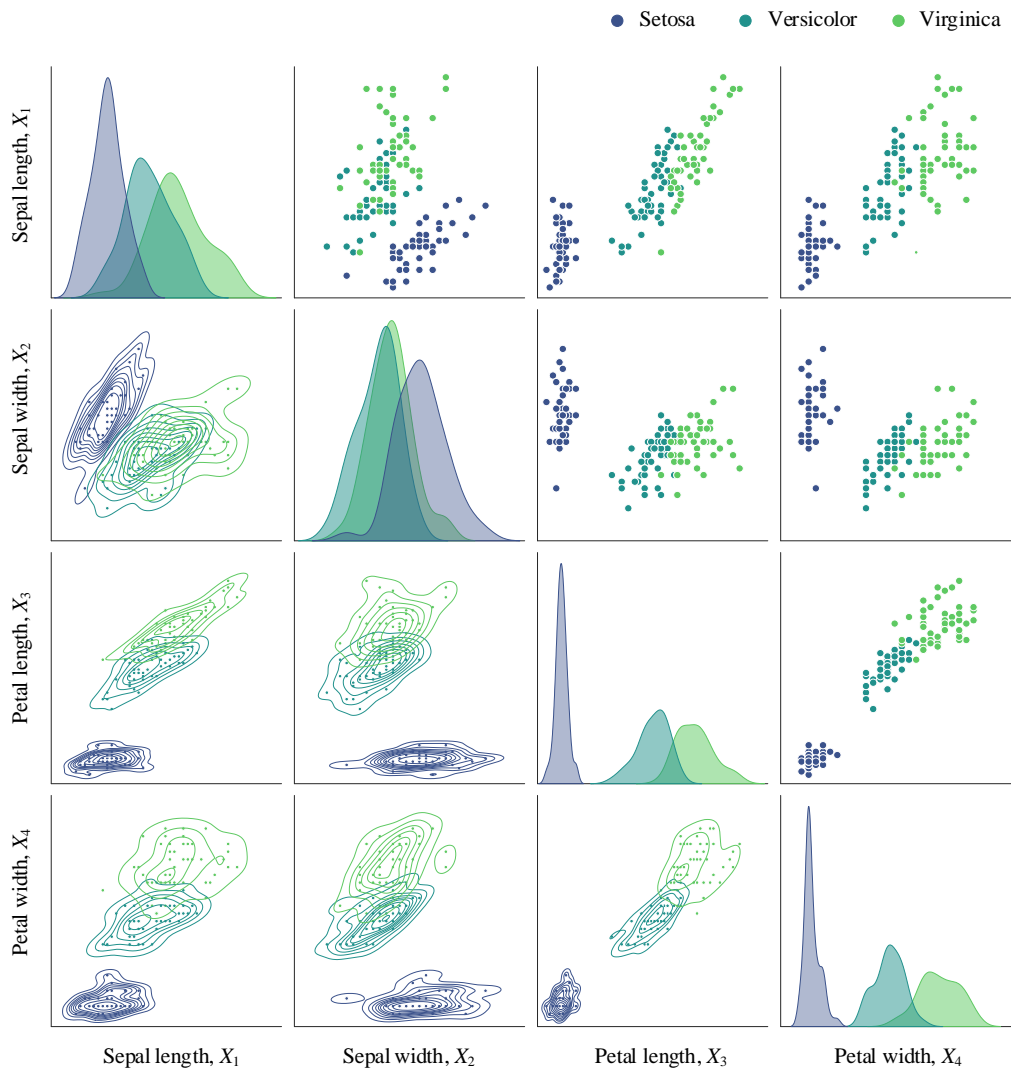


图 17. 鸢尾花数据成对特征分析图，考虑鸢尾花分类标签

平行坐标图

**平行坐标图** (Parallel Coordinate Plot, PCP) 能够在二维空间中呈现出多维数据。在平行坐标图中，每条折线代表一个样本点，图中每条竖线代表一个特征。折线的形状能够反映样本的若干特征。不同折线颜色代表不同分类标签，平行坐标图还可以不同特征对分类的影响。

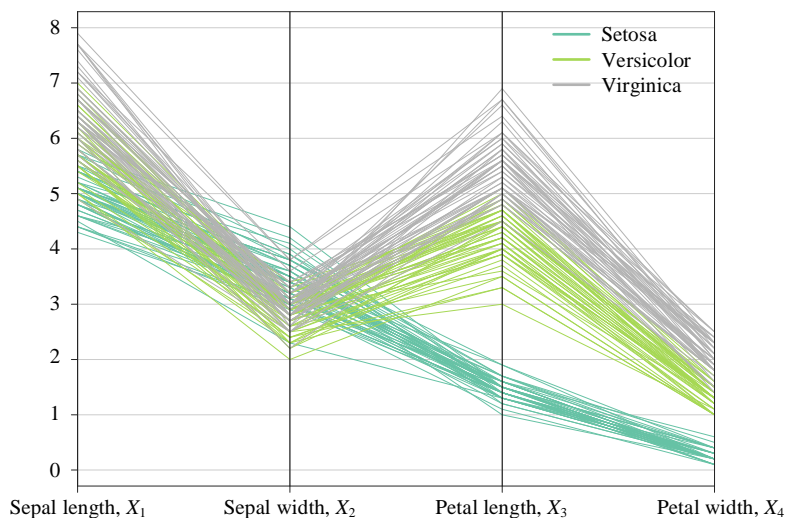


图 18. 鸢尾花数据的平行坐标图

## 2.5 集中度：均值、质心

本章前文通过图形可视化样本分布，本章后文介绍几种最基本的量化手段来描述样本数据。

量化样本数据集中度的最基本方法是**算数平均数** (arithmetic mean)：

$$\mu_X = \text{mean}(X) = \frac{1}{n} \left( \sum_{i=1}^n x^{(i)} \right) = \frac{x^{(1)} + x^{(2)} + x^{(3)} + \cdots + x^{(n)}}{n} \quad (6)$$

如果数据是总体，算数平均数为**总体平均值** (population mean)。如果数据是样本，算数平均数是**样本平均值** (sample mean)。



请大家回顾《矩阵力量》第 22 章讲过的均值的几何意义。

### 以鸢尾花数据集为例

鸢尾花四个量化特征——花萼长度 (sepal length)  $X_1$ 、花萼宽度 (sepal width)  $X_2$ 、花瓣长度 (petal length)  $X_3$  和花瓣宽度 (petal width)  $X_4$ ——均值分别为：

$$\mu_1 = 5.843, \mu_2 = 3.057, \mu_3 = 3.758, \mu_4 = 1.199 \quad (7)$$

图 4 所示为鸢尾花数据集四个特征均值在直方图位置。



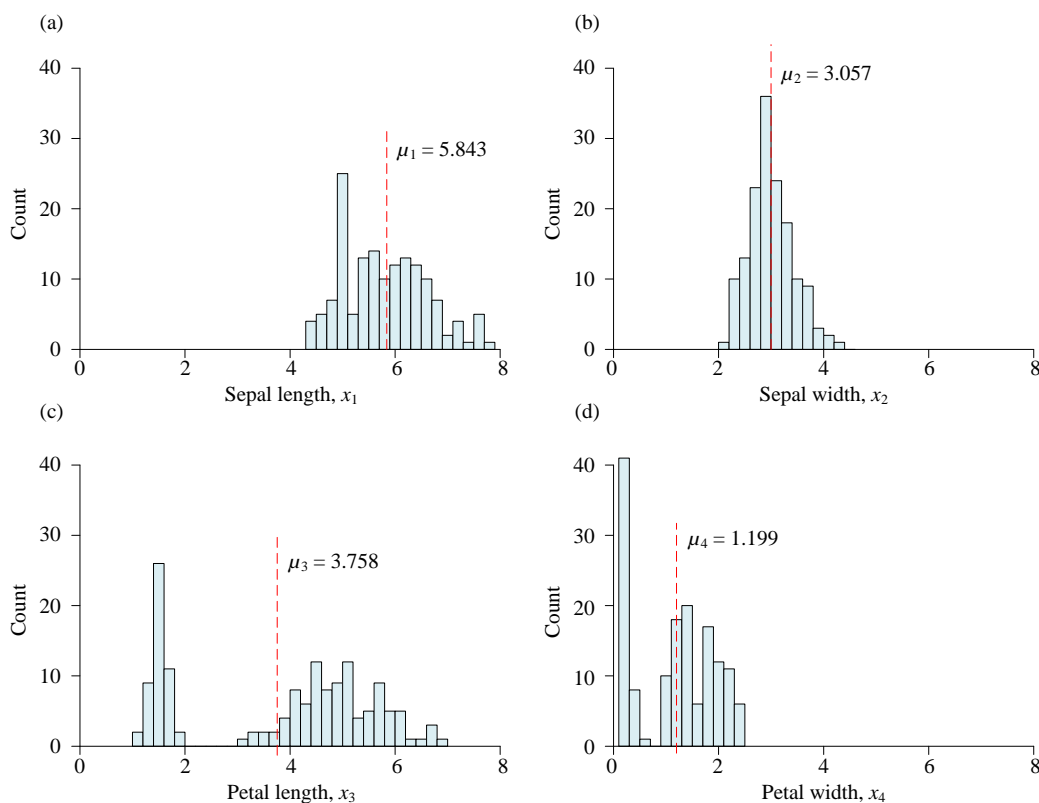


图 19. 鸢尾花四个特征数据均值在直方图位置

## 质心

当然，我们也可以把均值位置标注在散点图上。如图 20 所示，花萼长度、花萼宽度的均值相交于一点  $\times$ ，这一点常被称作数据的**质心** (centroid)。也就是说，有些场合，我们可以用质心这一个点代表一组样本数据。

比如，鸢尾花数据矩阵  $\mathbf{X}$  质心为：

$$\mathbf{E}(\mathbf{X}) = \boldsymbol{\mu}_X^T = \begin{bmatrix} 5.843 & 3.057 & 3.758 & 1.199 \\ \text{Sepal length, } x_1 & \text{Sepal width, } x_2 & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{bmatrix}^T \quad (8)$$

本书中， $\mathbf{E}(\mathbf{X})$  一般为行向量，而  $\boldsymbol{\mu}$  一般为列向量。本书一般不从符号上区别样本均值和总体均值 (期望值)，除非特别说明。

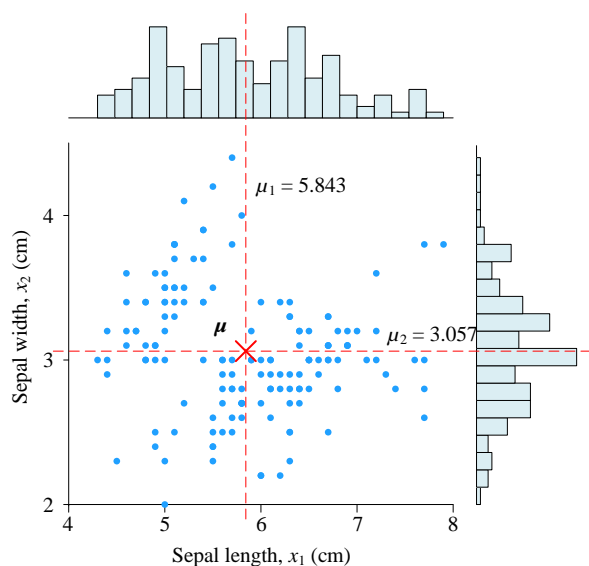


图 20. 均值在散点图的位置

### 考虑分类标签

分别计算鸢尾花不同分类标签 (setosa、versicolor、virginica) 花萼长度、花萼宽度平均值：

$$\begin{aligned}\mu_{1\_setosa} &= 5.006, & \mu_{2\_setosa} &= 3.428 \\ \mu_{1\_versicolor} &= 5.936, & \mu_{2\_versicolor} &= 2.770 \\ \mu_{1\_virginica} &= 6.588, & \mu_{2\_virginica} &= 2.974\end{aligned}\quad (9)$$

图 21 所示为不同分类标签的鸢尾花样本散点，以及各自的簇质心 (cluster centroid)。

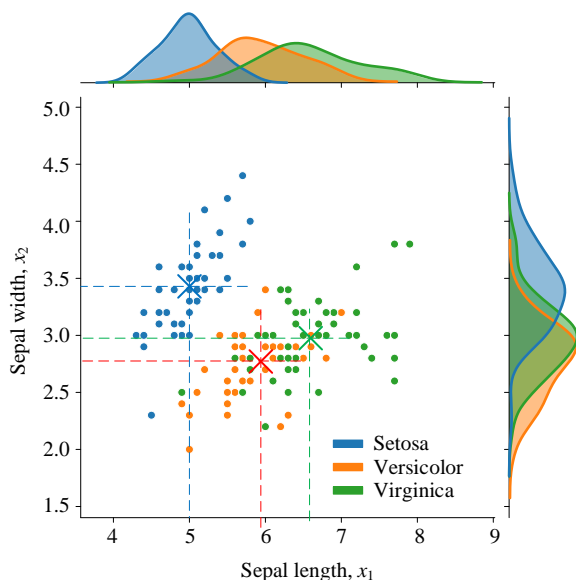


图 21. 均值在散点图的位置，考虑类别标签

## 中位数、众数、几何平均数

**中位数** (median) 又称中值，指的是按顺序排列的一组样本数据中居于中间位置的数。如果样本数量为奇数，从小到大排列居中的样本就是中位数；如果样本有偶数个，通常取最中间的两个数值的平均数作为中位数。本书后续将在贝叶斯推断中进一步比较均值、中位数。

**众数** (mode) 是一组数中出现最频繁的数值。众数通常用于描述离散型数据，因为这些数据中每个值只能出现整数次，而众数是出现次数最多的值。对于连续型数据，比如身高、体重，由于每个数值只有极小的概率出现，因此通常不会存在一个数值出现次数最多的情况，此时可以使用**区间众数** (interval mode) 来描述数据的分布形态。

众数的计算相对简单，只需要统计每个数值出现的次数，然后找到出现次数最多的数值即可。众数的优点是计算简单，易于理解和解释，但缺点是可能存在多个众数或者无众数的情况，而且受极端值的影响较大。

**几何平均数** (geometric mean) 的定义如下：

$$\left( \prod_{i=1}^n x^{(i)} \right)^{\frac{1}{n}} = \sqrt[n]{x^{(1)} \cdot x^{(2)} \cdot x^{(3)} \cdots x^{(n)}}$$

注意，几何平均数只适合正数。

## 2.6 分散度：极差、方差、标准差

本节介绍度量分散度的常见统计量。

### 极差

**极差** (range)，又称全距，是指样本最大值与最小值之间的差距：

$$\text{range}(X) = \max(X) - \min(X) \quad (10)$$

极差是度量分散度最简单的指标。图 22 所示为最大值、最小值、极差、均值之间关系。注意，极差很容易受到离群值影响。

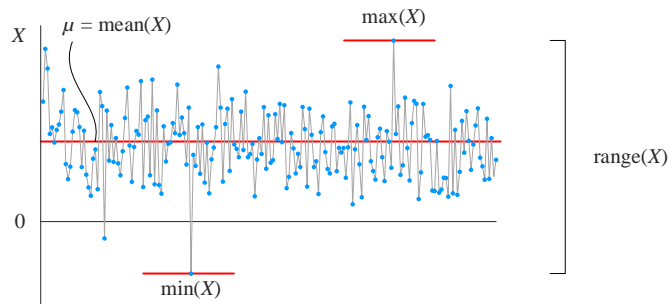


图 22. 最大值、最小值、极差、均值的关系

## 方差

**方差** (variance) 衡量随机变量或样本数据离散程度。方差越大，数据的分布就越分散；方差越小，数据的分布就越集中。样本的方差为：

$$\text{var}(X) = \sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)^2 \quad (11)$$

简单来说，方差是各观察值与数据集平均值的差的平方的平均值。

方差的单位是样本单位的平方，比如鸢尾花数据方差单位为  $\text{cm}^2$ 。请大家注意，本书中样本方差、总体方差符号上完全一致，不做特别区分。此外，请大家回顾《矩阵力量》第 22 章讲过的方差的几何意义。

## 标准差

样本的**标准差** (standard deviation) 为样本方差的平方根：

$$\sigma_X = \text{std}(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)^2} \quad (12)$$

同样，标准差越大，数据的分布就越分散；标准差越小，数据的分布就越集中。鸢尾花样本数据四个量化特征的标准差分别为：

$$\sigma_1 = 0.825, \quad \sigma_2 = 0.434, \quad \sigma_3 = 1.759, \quad \sigma_4 = 0.759 \quad (13)$$


 注意，标准差和原始数据单位一致。比如，鸢尾花四个特征的量化数据单位均为厘米 (cm)。

图 23 上，我们把  $\mu \pm \sigma$ 、 $\mu \pm 2\sigma$  对应的位置也画在直方图上。



68-95-99.7 法则和  $\mu \pm \sigma$ 、 $\mu \pm 2\sigma$ 、 $\mu \pm 3\sigma$  有关，本书第 9 章将介绍 68-95-99.7 法则。

其实，大家在生活中经常用到“均值”、“标准差”这两个概念，只不过大家没有注意到而已。举个例子，想要提高考试成绩，大家平时练习时会尽量提高平均分，并减小各种因素对分数带来的负面波动。这就是在增大均值，减小标准差 (波动)。

再举个例子，一个教练在选择哪个选手上场的时候，也会看“均值”、“标准差”。“均值”代表一个选手的绝对实力，“标准差”则代表选手成绩的波动幅度。

教练求稳的时候，会派出均值相对高、标准差 (波动) 小的选手。在大比分落后情况下，教练可能会派出临场发挥型选手。发挥型选手成绩均值可能不是最高，但是有能力“冲一冲”。

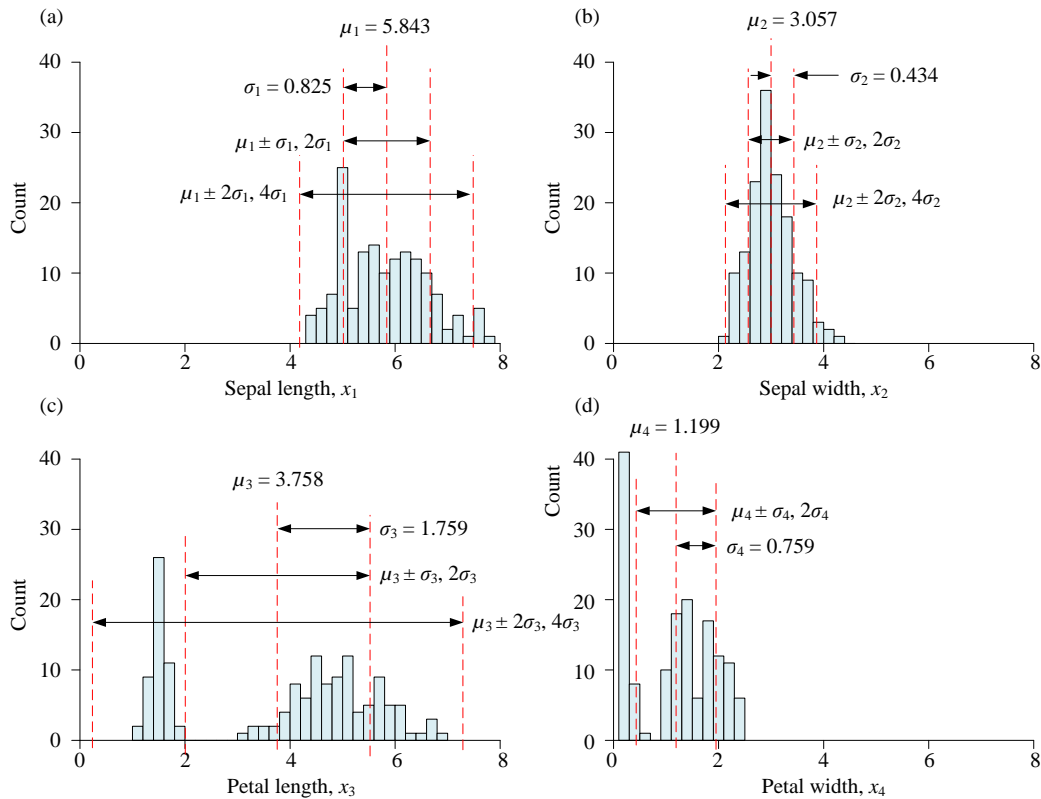


图 23. 鸢尾花四个特征数据均值、标准差所在位置在直方图位置

## 2.7 分位：四分位、百分位等

**分位数** (quantile), 亦称分位点, 是指将一个随机变量的概率分布范围分为几个等份的数值点。常用的分位数有**二分位点** (2-quantile, median)、**四分位点** (4-quantiles, quartiles)、**五分位点** (5-quantiles, quintiles)、**八分位点** (8-quantiles, octiles)、**十分位点** (10-quantiles, deciles)、**二十分位点** (20-quantiles, vigintiles)、**百分位点** (100-quantiles, percentile) 等。

实践中, 四分位和百分位最常用。以百分位为例, 把一组从小到大排列的样本数据分为 100 等份后, 每一个分点就是一个百分位数。同理, 将所有样本数据从小到大排列, 四分位数对应三个分割位置 (25%、50%、75%)。这三个分割位置将样本平分为四等份。50%分位对应中位数。图 24 所示为将鸢尾花不同特征的四分位画在直方图上。

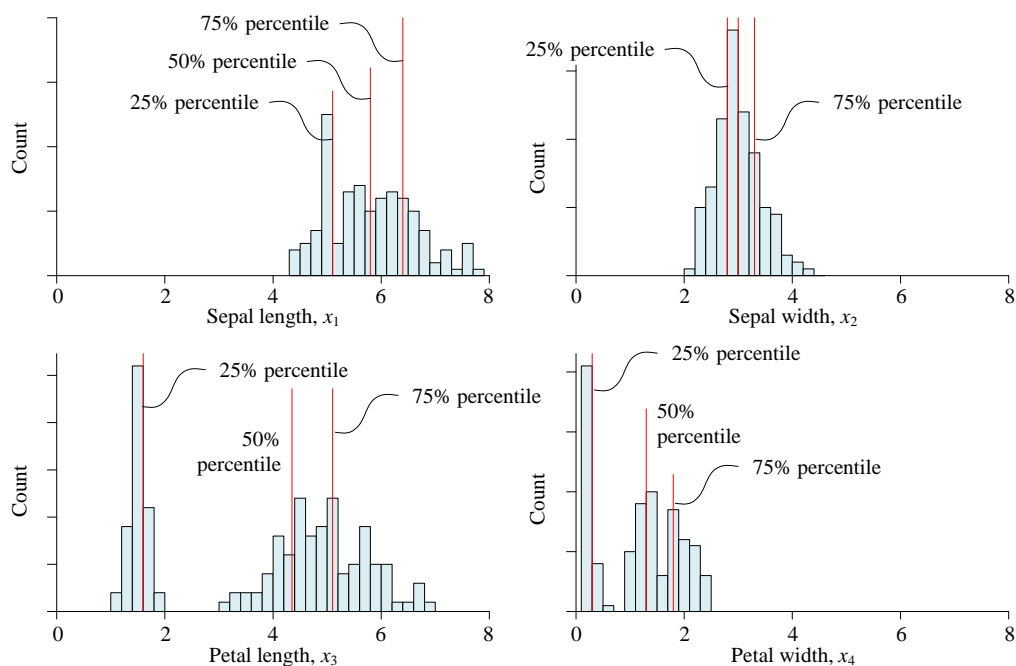


图 24. 鸢尾花数据直方图，以及 25%、50% 和 75% 百分位

图 25 所示为鸢尾花四个特征数据 1%、50%、99% 两个百分位分位位置，1%、99% 可以用来描述样本分布的“左尾”、“右尾”。

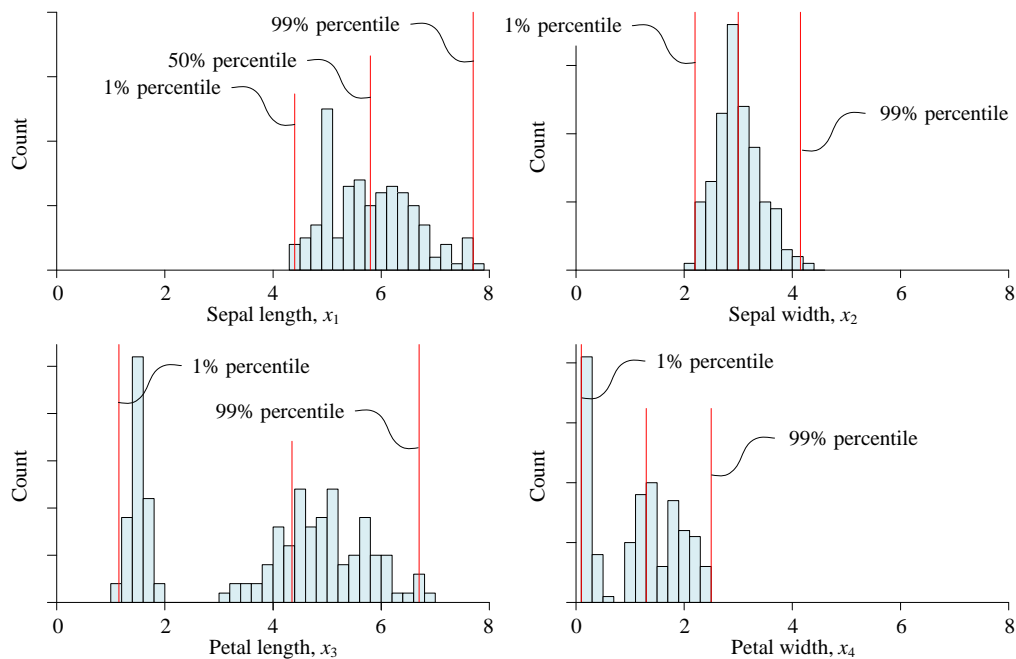


图 25. 鸢尾花数据直方图，以及 1% 和 99% 百分位

对于 Pandas 数据帧 df，df.describe() 默认输出数据的样本总数、均值、标准差、最小值、25% 分位、50%分位 (中位数)、75%分位。图 26 所示鸢尾花数据帧的总结，其中还给出 1%百分位、99%分位。

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
1%	4.400000	2.200000	1.149000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
99%	7.700000	4.151000	6.700000	2.500000
max	7.900000	4.400000	6.900000	2.500000

图 26. 鸢尾花数据帧统计总结

## 2.8 箱型图：小提琴图、分布散点图

图 27 所示为**箱型图** (box plot) 原理。箱型图利用第一 (25%,  $Q_1$ )、第二 (50%,  $Q_2$ ) 和第三 (75%,  $Q_3$ ) 四分位数展示数据分散情况。 $Q_1$  也叫下四分位， $Q_2$  也叫中位数， $Q_3$  也称上四分位。

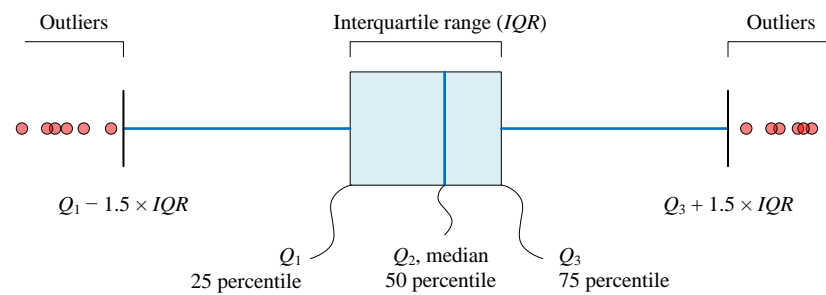


图 27. 箱型图原理

箱型图的**四分位间距** (interquartile range) 定义为：

$$IQR = Q_3 - Q_1$$

(14)

箱型图也常用来分析样本中可能存在的离群点，图 27 中两侧的红点。 $Q_3 + 1.5 \times IQR$  也称上界， $Q_1 - 1.5 \times IQR$  叫下界。而在  $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$  之外的样本数据则被视作离群点。

数据分析中，四分位间距  $IQR$  也常常用来度量样本数据的分散程度。相比标准差，四分位间距  $IQR$  不受厚尾影响，受离群值影响小得多。

图 28 所示为鸢尾花数据四个特征上的箱型图。

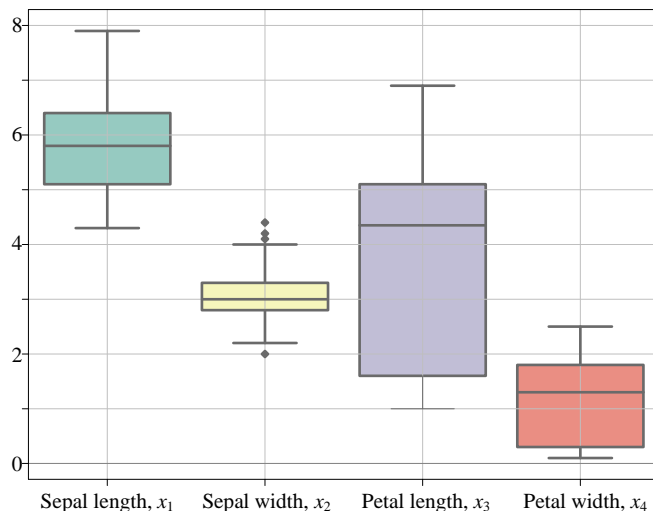


图 28. 鸢尾花数据箱型图

### 箱型图的变体

箱型图还有很多的“变体”。比如图 29 所示的小提琴图，图 30 所示的分布散点图。图 31 所示为箱型图叠加分布散点图。图 32 所示为考虑标签的箱型图。箱型图的优点是简单易懂，可以同时展示数据的中心趋势、离散程度和离群值等信息。因此，箱型图经常被用来比较多组数据的分布情况，或者发现异常值。

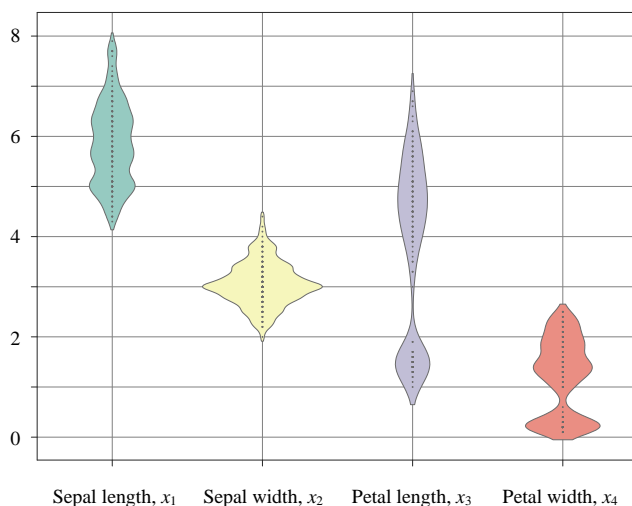


图 29. 鸢尾花数据小提琴图



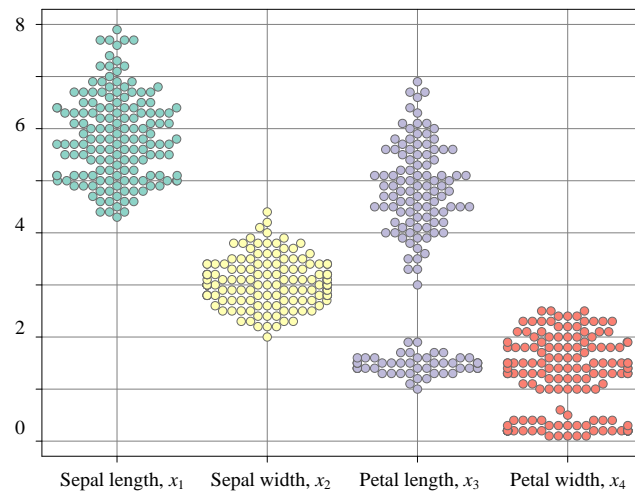


图 30. 分布散点图 (stripplot)

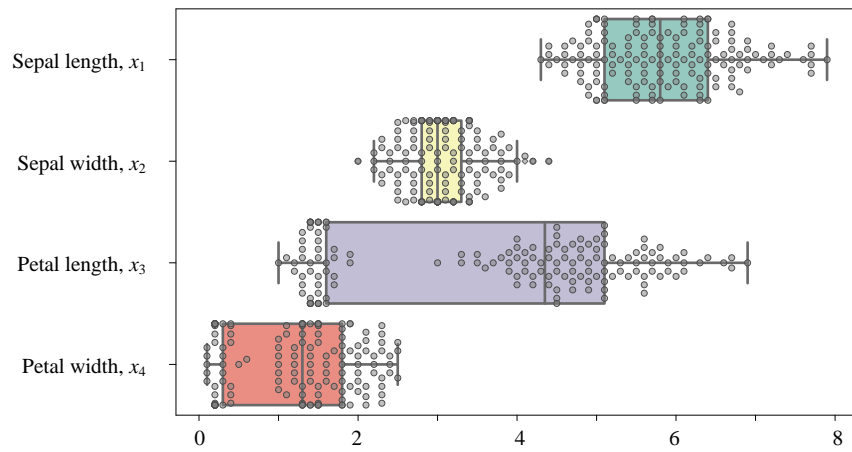


图 31. 鸢尾花箱型图，叠加分布散点图 swarmplot

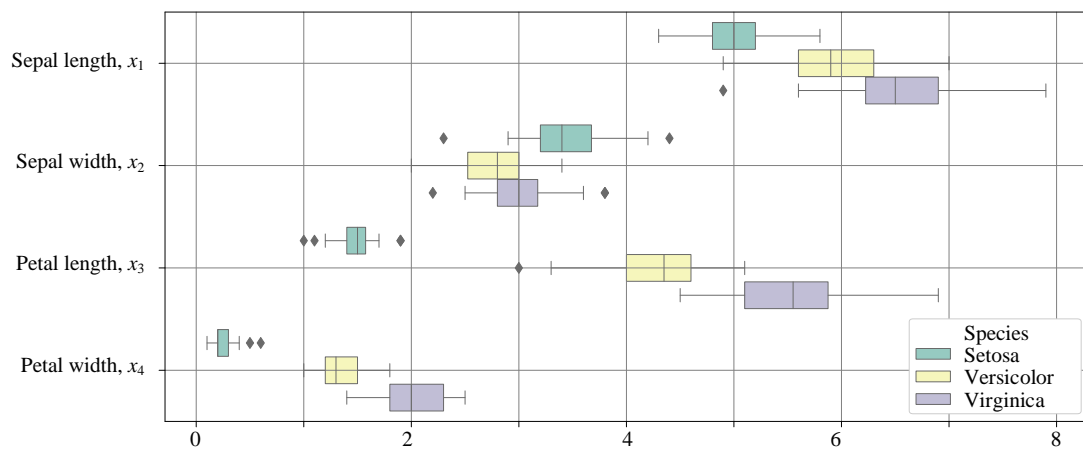


图 32. 鸢尾花箱型图，考虑分类标签

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：[jiang.visualize.ml@gmail.com](mailto:jiang.visualize.ml@gmail.com)

## 2.9 中心矩：均值、方差、偏度、峰度

统计学中的**矩** (moment)，又称为**中心矩** (central moment)，是对变量分布和形态特点进行度量的一组量，其概念借鉴物理学中的“矩”。在物理学中，矩是描述物理性状特点的物理量。

零阶矩表示随机变量的总概率，也就是 1。具体而言，常用的中心矩为一至四阶矩，分别表示数据分布的位置、分散度、偏斜程度和峰度程度。

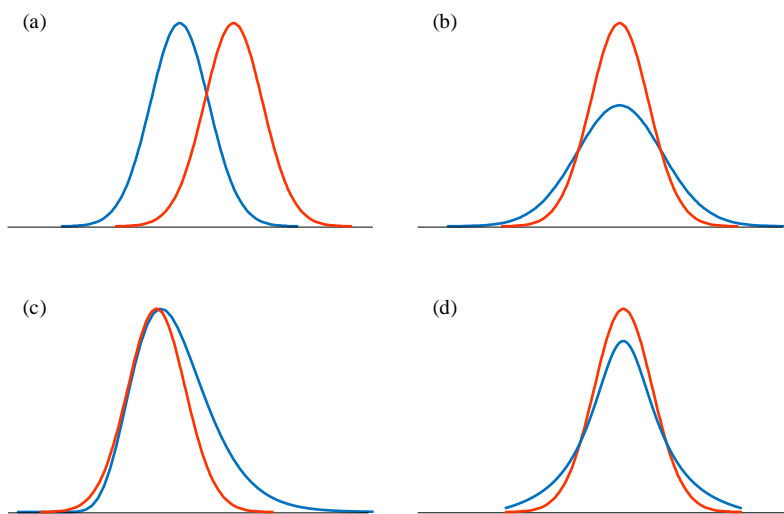


图 33. 期望 (一阶矩)、方差 (二阶矩)、偏斜度 (三阶矩)、峰度 (四阶矩)

### 一阶矩、二阶矩

一阶矩为均值，即**期望** (expectation)，用来描述分布中心位置，如图 33 (a) 所示。前文提过，均值的量纲 (单位) 和原始数据相同。

**⚠ 注意**，量纲和单位虽然混用，但是两者还是有区别。从量纲的角度来看，m、cm、mm 都是长度度量单位，含义相同。但是，m、cm、mm 的单位不同，它们之间存在一定换算关系。

二阶矩为**方差** (variance)，描述分布分散情况，如图 33 (b) 所示。方差的量纲为原始数据量纲的平方。

图 33 中的一元分布都是高斯分布。虽然一元高斯分布的参数仅为均值和方差，但是真实的样本数据分布不可能仅仅用均值和方差来刻画，有时还需要偏度 (三阶矩) 和峰度 (四阶矩)。

### 三阶矩

三阶矩为**偏度** (skewness)  $S$ 。如图 33 (c) 所示，偏度描述分布的左右倾斜程度：

$$S = \text{skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^3}{\left( \frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^2 \right)^{\frac{3}{2}}} \quad (15)$$

与期望和标准差不同，偏度没有单位，是无量纲的量。偏度的绝对值越大，表明样本数据分布的偏斜程度越大。

对于完全对称的单峰分布，平均数、中位数、众数，处在同一位置，图 34 (a) 所示。这种分布的偏度为零。如果样本数服从一元高斯分布，则偏度为 0，即均值 = 中位数 = 众数。

**正偏** (positive skew, positively skewed)，又称**右偏** (right-skewed, right-tailed, skewed to the right)。如图 34 (b) 所示，正偏分布的右侧尾部更长，分布的主体集中在图像的左侧。正偏 (右偏) 时，均值 > 中位数 > 众数。

大家可以这样理解平均数、中位数、众数这三个数值的关系。如果在样本中引入少数几个特别大的离群值的话，均值肯定增大 (向右移动)，中位数微微受到影响 (样本数量增加)，但是众数 (出现次数最多) 不变。

**负偏** (negative skew, negatively skewed)，又称**左偏** (left-skewed, left-tailed, skewed to the left)，如图 34 (c) 所示，特点是分布的左侧尾部更长，分布的主体集中在右侧。负偏 (左偏) 时，众数 > 中位数 > 均值。

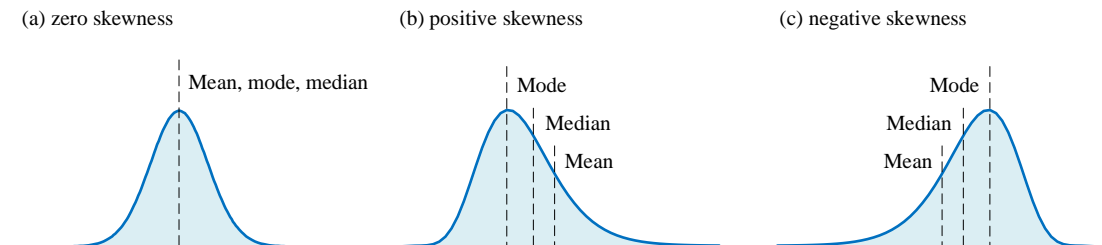


图 34. 无偏、正偏和负偏

⚠ 值得注意的是，偏度为零不一定意味着分布对称。如图 35 所示，这个离散分布的偏度计算出来为 0，但是很明显这个分布不对称。

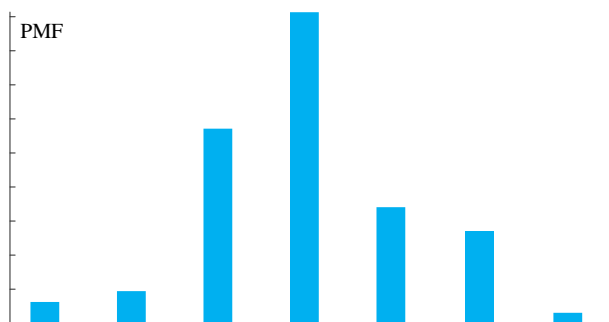


图 35. 偏度为 0，但是不对称的分布

## 四阶矩

四阶矩表示**峰度** (kurtosis)  $K$ 。图 33 (d) 所示，峰度描述分布与正态分布相比的陡峭或扁平程度：

$$K = \text{kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_X)^4}{\left( \frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_X)^2 \right)^2} \quad (16)$$

和偏度一样，峰度也没有单位，是无量纲的量。

▲ 注意，用 (16) 计算的话，正态分布的峰度为 3。

图 36 展示两种峰态：**高峰态** (leptokurtic)、**低峰态** (platykurtic)。高峰度的峰度值大于 3。如图 36 (a) 所示，和正态分布相比，高峰态分布有明显的尖峰，两侧尾端有**肥尾** (fat tail)。

图 36 (b) 展示的是低峰态。相比正态分布，低峰态明显稍扁。

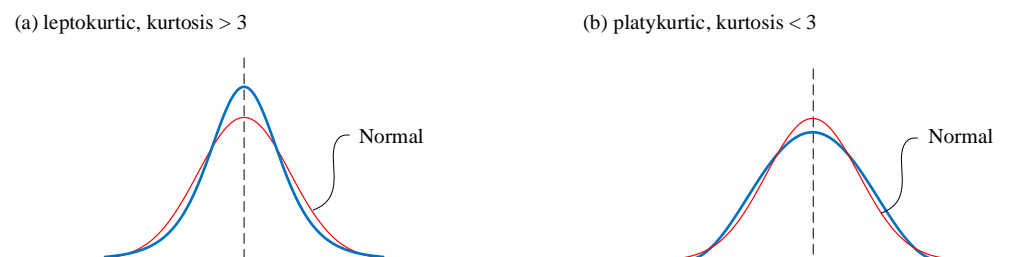


图 36. 高峰态和低峰态

实践中，一般采用**超值峰度** (excess kurtosis)，即 (16) 减去 3：


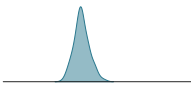

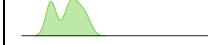
$$\text{Excess kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^4}{\left( \frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^2 \right)^2} - 3 \quad (17)$$

“减去 3”是为了让正态分布的峰度为 0，方便其他分布和正态分布比较。

**表 1** 总结鸢尾花数据的四阶矩。花萼长度、花萼宽度上，样本数据都存在正偏。花萼长度分布存在低峰态，花萼宽度上出现高峰态。

对比表中样本数据分布和四阶矩的具体值，不难发现即便使用四阶矩也未必能够准确描述真实分布。比如，花瓣长度、花瓣宽度上，样本数据分布存在明显的双峰态。

表 1. 鸢尾花四阶矩

				
	花萼长度	花萼宽度	花瓣长度	花瓣宽度
均值 (cm)	5.843	3.057	3.758	1.199
标准差 (cm)	0.825	0.434	1.759	0.759
偏度	0.314	0.318	-0.274	-0.102
超值峰度	-0.552	0.228	-1.402	-1.340

## 2.10 多元随机变量关系：协方差矩阵、相关性系数矩阵

**协方差** (covariance) 是用来度量两个变量之间的线性关系强度和方向的统计量。当两个变量的协方差为正时，说明它们的变化趋势同向，即当一个变量增加时，另一个变量也倾向于增加；当协方差为负时，说明它们的变化趋势是相反的，即当一个变量增加时，另一个变量倾向于减少。协方差为 0 时，则表明两个变量之间没有线性关系。

对于样本数据，随机变量  $X$  和  $Y$  的协方差为：

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_x)(y^{(i)} - \mu_y) \quad (18)$$

线性相关性系数 (linear correlation coefficient)，也叫皮尔逊相关系数 (Pearson correlation coefficient)，是一种用来度量两个变量之间线性相关程度的统计量。它的取值范围在 -1 到 1 之间，数值越接近 -1 或 1，表示两个变量之间的线性关系越强；数值接近 0，则表示两个变量之间没有线性关系。

对于样本数据，随机变量  $X$  和  $Y$  的线性相关性系数为：

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (19)$$

“鸢尾花书”读者对**协方差矩阵** (covariance matrix)、**相关性系数矩阵** (correlation matrix) 应该非常熟悉。协方差矩阵和相关性系数矩阵都是描述多维随机变量之间关系的矩阵。

➡ 建议大家回顾《矩阵力量》中 Cholesky 分解和特征值分解协方差矩阵会产生怎样的结果。此外，也请大家回顾协方差矩阵和格拉姆矩阵的关系。

以鸢尾花四个特征为例，它的协方差矩阵为  $4 \times 4$  矩阵：

$$\Sigma = \begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \text{cov}(X_1, X_3) & \text{cov}(X_1, X_4) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \text{cov}(X_2, X_3) & \text{cov}(X_2, X_4) \\ \text{cov}(X_3, X_1) & \text{cov}(X_3, X_2) & \text{cov}(X_3, X_3) & \text{cov}(X_3, X_4) \\ \text{cov}(X_4, X_1) & \text{cov}(X_4, X_2) & \text{cov}(X_4, X_3) & \text{cov}(X_4, X_4) \end{bmatrix} \quad (20)$$

其相关性系数矩阵为  $4 \times 4$ ：

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \rho_{1,4} \\ \rho_{2,1} & 1 & \rho_{2,3} & \rho_{2,4} \\ \rho_{3,1} & \rho_{3,2} & 1 & \rho_{3,4} \\ \rho_{4,1} & \rho_{4,2} & \rho_{4,3} & 1 \end{bmatrix} \quad (21)$$

图 37 所示为协方差矩阵和相关性系数矩阵热图。

➡ 本书第 13 章将专门讲解协方差矩阵。

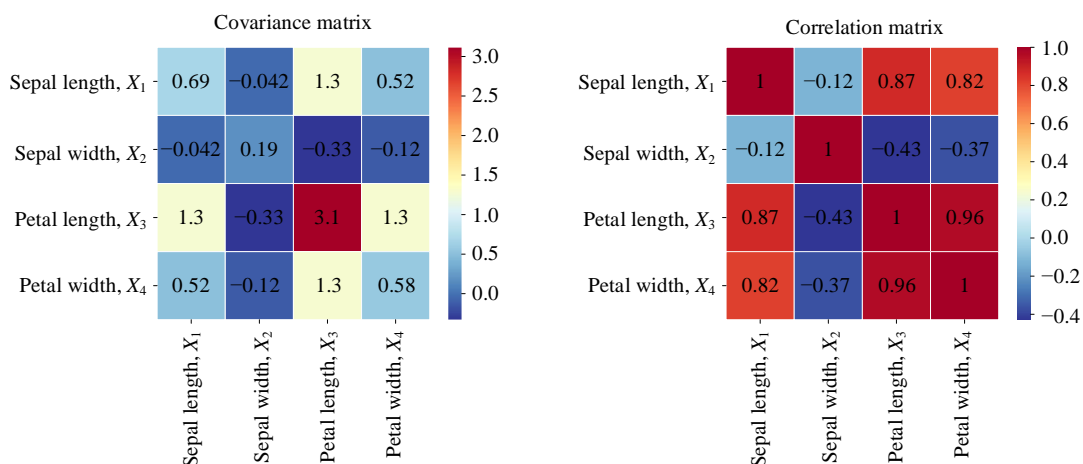


图 37. 协方差矩阵、相关性系数矩阵热图



代码文件 Bk5\_Ch02\_01.py 绘制本章几乎所有图像。



描述、推断是统计的两个重要板块。本章介绍了常见的统计描述工具。统计分析中，可视化和量化分析都很重要。本章介绍的重要的统计可视化工具有直方图、散点图、箱型图、热图等。此外，也需要大家数量掌握样本数据的均值、方差、标准差、协方差、协方差矩阵、相关性系数矩阵等等。

统计描述、统计推断之间的桥梁正是概率。从下一章开始，我们正式进入概率板块学习。