

# 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



- ▶ `joyypy.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()` 绘制概率/概率直方图
- ▶ `seaborn.jointplot()` 绘制联合分布和边际分布
- ▶ `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)

## 2.1 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，我们要利用一些统计工具来分析这些数据，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)。直方图可视化数据分布情况，诸如众数、中位数的大致位置、数据是否存在异常值。

图 1 所示为鸢尾花花萼长度数据直方图，注意直方图的纵轴有三个选择——频数、概率和概率密度。下面聊聊频数、概率和概率密度分别是什么。

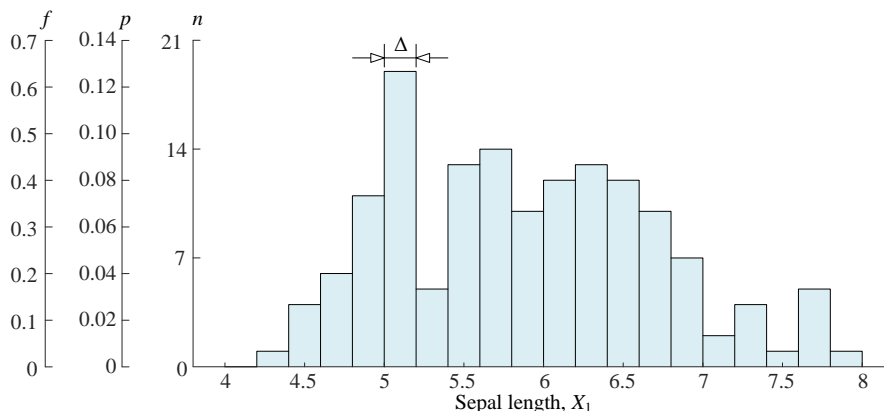


图 1. 鸢尾花花萼长度，频数、概率和概率密度的关系

### 频数、概率、概率密度

花萼长度的最小值和最大值落在  $[4, 8]$  这个区间。将这个区间等分为 20 个区间，每个区间对应的宽度叫做组距  $\Delta = 0.2$ 。全体样本分成区间个数称为组数  $M$ 。一般情况，每个区间包含左侧端点，不含右侧端点，即左闭右开区间。

频数是指在一定范围内样本数据的数量。比如，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。

图 2 第一列给出的是每个组距所在的区间，数出落在第  $i$  个区间内的样本数据的数量，这个数量定义为频数  $n_i$ 。频数  $n_i$  除以样本总数  $n$  叫做概率  $p_i$ ：

$$p_i = \frac{n_i}{n} \quad (1)$$

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 2. 鸢尾花花萼长度直方图数据

显然，所有频数  $n_i$  之和为样本总数  $n$ ：

$$\sum_{i=1}^M n_i = n \quad (2)$$

直方图的纵轴为概率时，直方图也叫归一化直方图。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  $p_i$  除以组距  $\Delta$  得到的是概率密度 (probability density)  $f_i$ ：

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

大家一定要注意，概率密度不是概率；但是，概率密度也反映数据分布的疏密情况。

纵轴为概率密度的直方图，所有矩形面积之和为 1：

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

图 2 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。

## 直方图

图 3 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。

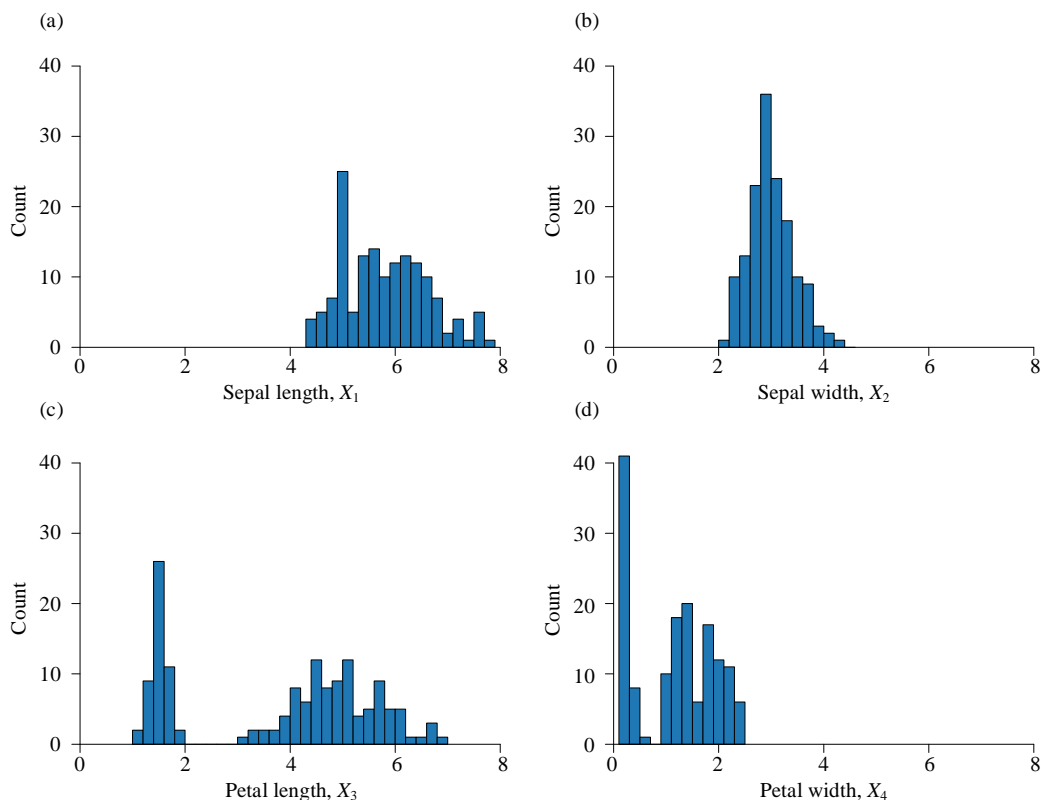


图 3. 鸢尾花四个特征数据的直方图，纵轴为频数

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。

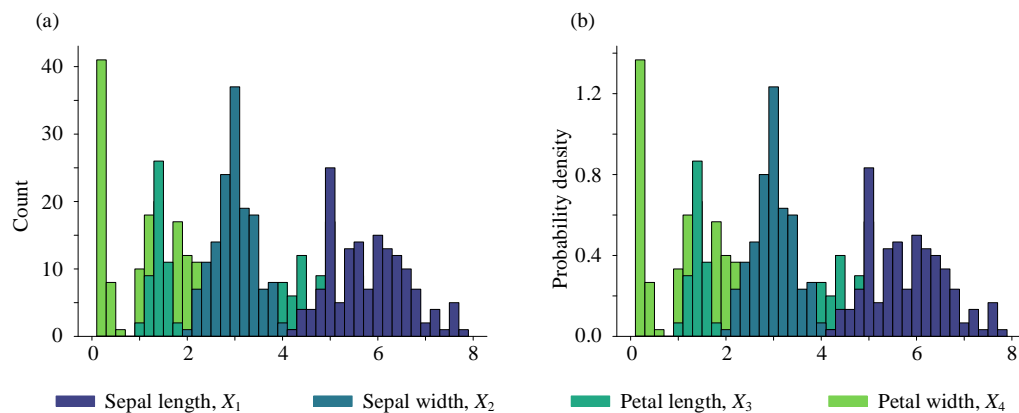


图 4. 直方图，比较频数和概率密度

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 5 对比四个鸢尾花特征数据的累积频数图、累积概率图。如图 5 (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 5 (b) 所示，累积概率的最大值为 1。

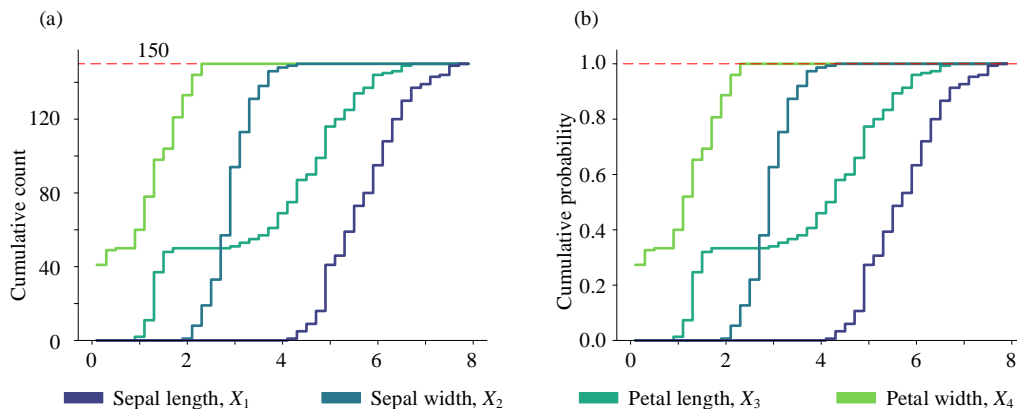


图 5. 累积频数图，累积概率图

## 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 6 (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 6 (a) 给出的纵轴为概率密度。

核密度估计 (Kernel Density Estimation, KDE) 是对直方图的扩展，如图 6 (b) 中曲线是通过核密度估计得到的概率密度图像。本书第 18 章将专门讲解概率密度估计。

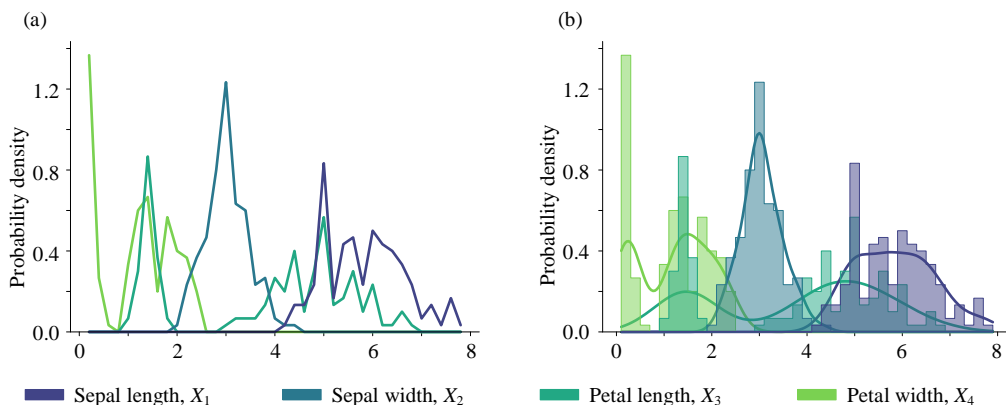


图 6. 比较多边形图和和概率密度估计曲线

## 山脊图

山脊图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 7 所示的山脊图 (ridgeline plot) 采用 joypy 绘制。

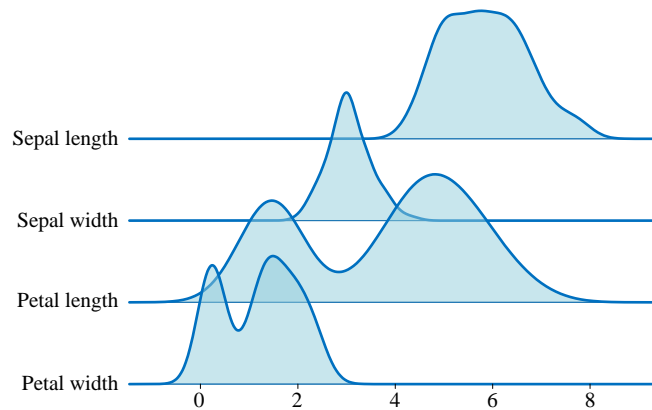


图 7. 鸢尾花山数据山脊图

## 2.2 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 8 (a) 所示。

在散点图的基础上，可以拓展得到一系列衍生图像。比如图 8 (a) 中，我们可以看到两幅直方图，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 8 (b) 增加了简单线性回归图像和 KDE 概率密度曲线。

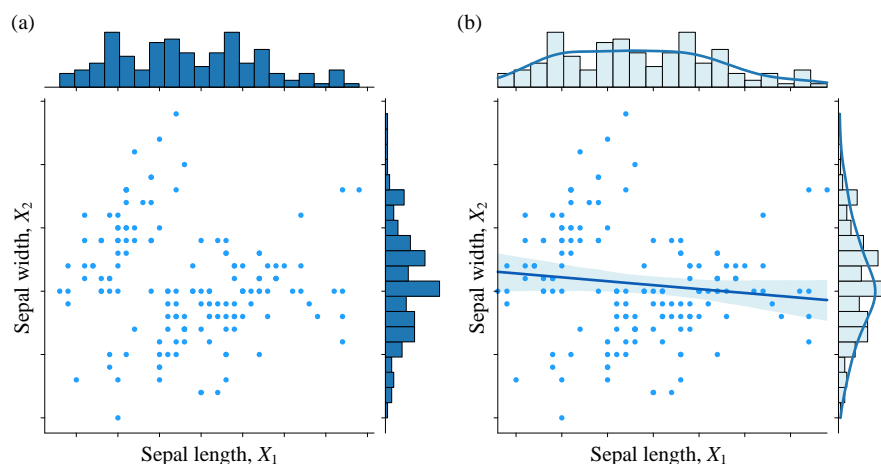


图 8. 二维数据散点图及扩展

### 二维概率密度

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)

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 9 (a) 所示为二维数据直方图热图，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 9 (b) 为二维 KDE 概率密度曲面等高线图。

在图 9 基础上，图 10 (a) 在直方图热图上增加了边际直方图，图 10 (a) 在二维概率密度曲面等高线图上增加了边际概率密度曲线。

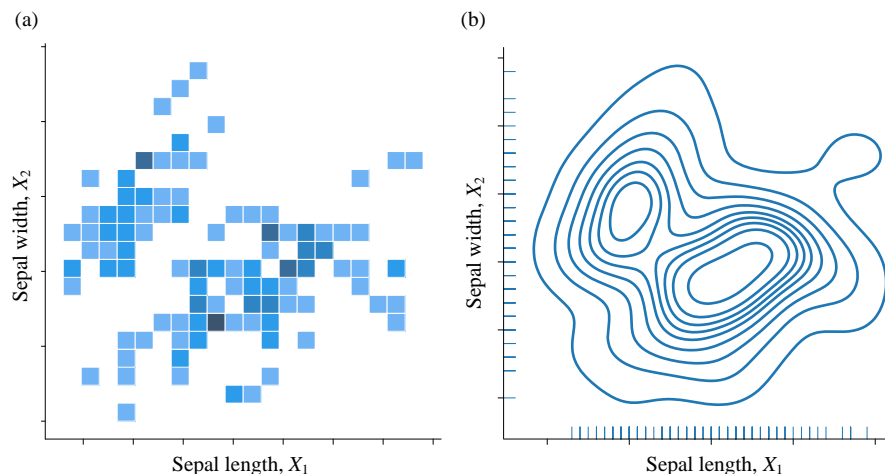


图 9. 二维数据直方图热图，二维 KDE 概率密度曲面等高线

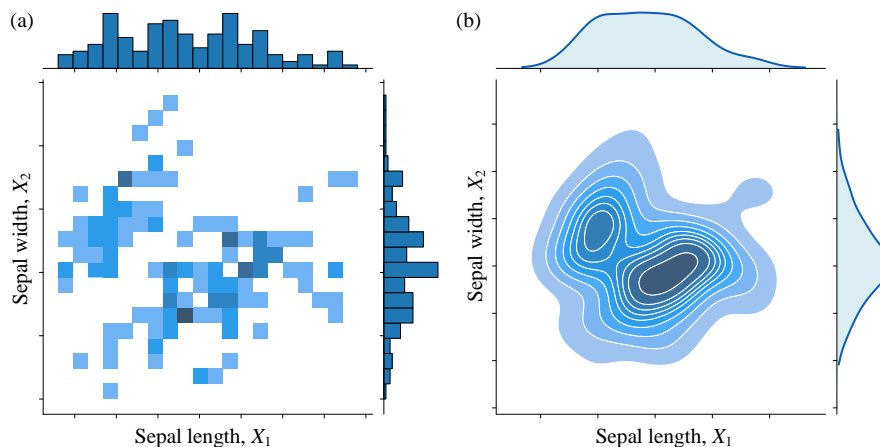


图 10. 直方图热图和概率密度曲面等高线拓展

## 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 11 所示为鸢尾花数据成对特征分析图。相信丛书读者对图 11 已经完全不陌生。

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。

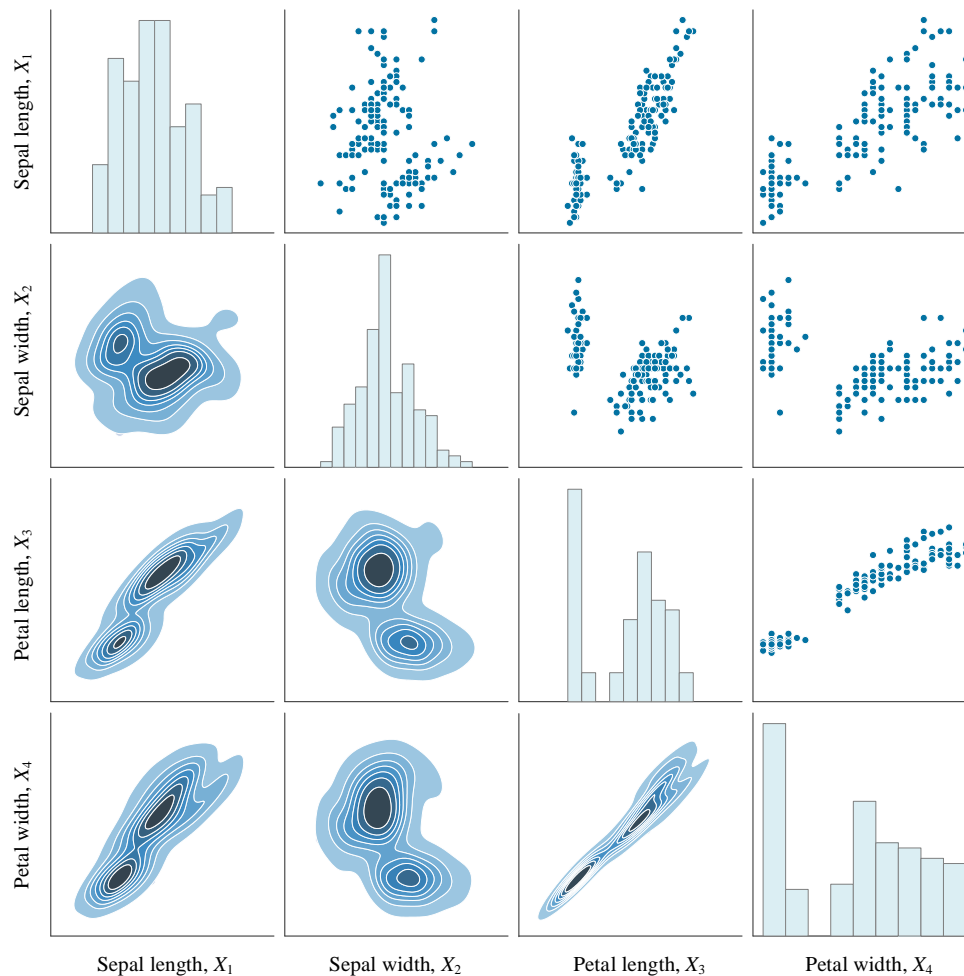


图 11. 鸢尾花数据成对特征分析图

## 2.3 有标签数据的统计可视化

《矩阵力量》专门区分过有标签 (labeled data) 和无标签数据，如图 12 所示。

鸢尾花数据就是典型的分类数据，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标签。



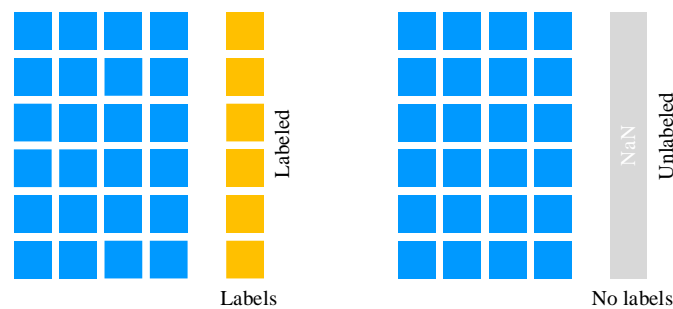


图 12. 根据有无标签分类数据

图 13 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 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 15 所示。图 16 所示为考虑标签的成对特征图。

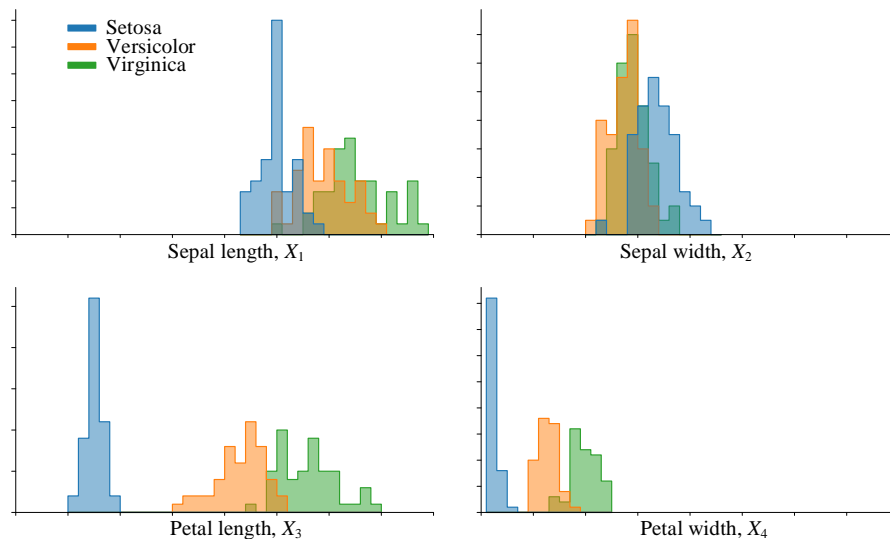


图 13. 标签分类的直方图

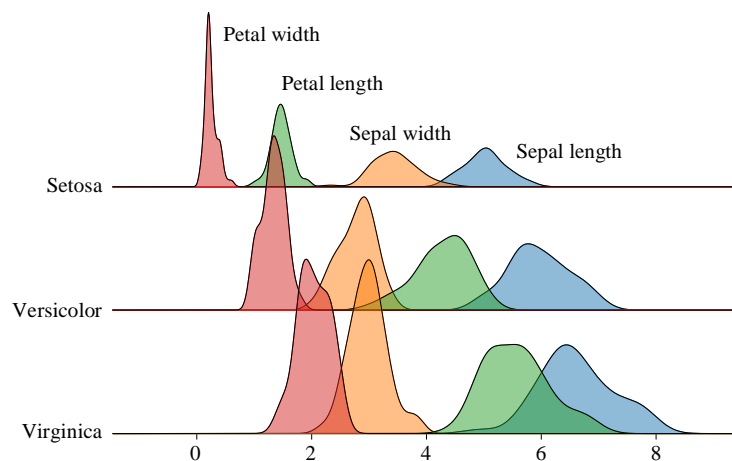


图 14. 鸢尾花山数据山脊图，特征分类

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)

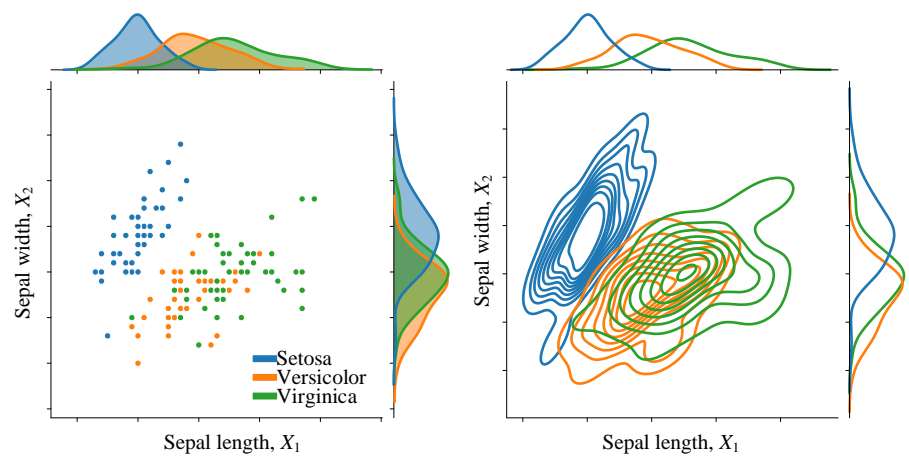


图 15. 二维数据散点图，KDE 概率密度曲面等高线，分类数据

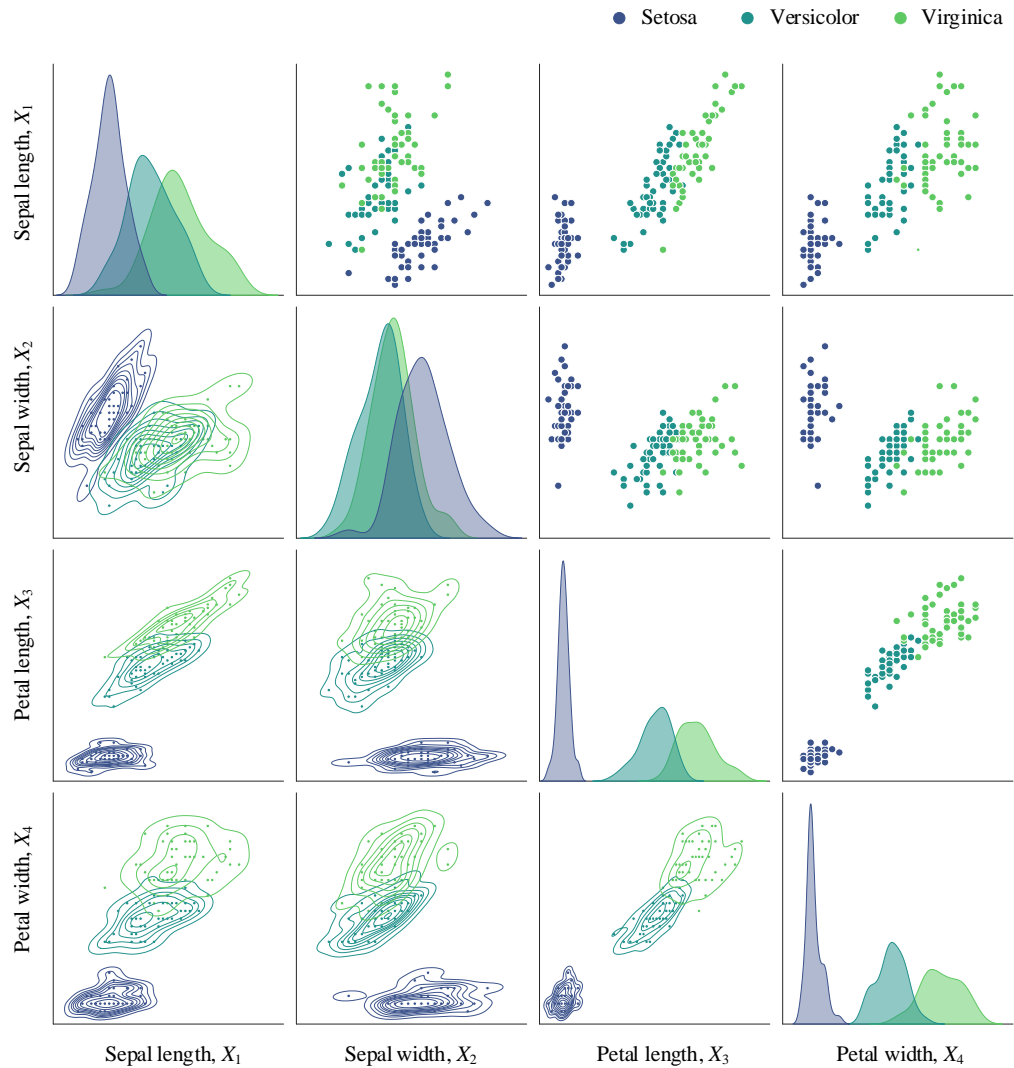


图 16. 鸢尾花数据成对特征分析图，分类数据

## 平行坐标图

平行坐标图 (Parallel Coordinate Plot, PCP) 是一个常见的可视化形式，能够在二维空间中呈现出多维数据。在平行坐标图中，每条折线代表一条数据，折线的形状能够反映数据的特征。配合不同颜色，平行坐标图还可以展示数据类别。图种每个竖线代表一个特征，上面的点代表该特征的值，每个样本表示出来就是一个贯穿所有竖线的折线图。一般来说，用不同的颜色代表不同的类别，这样可以方便的看出不同特征对分类的影响。

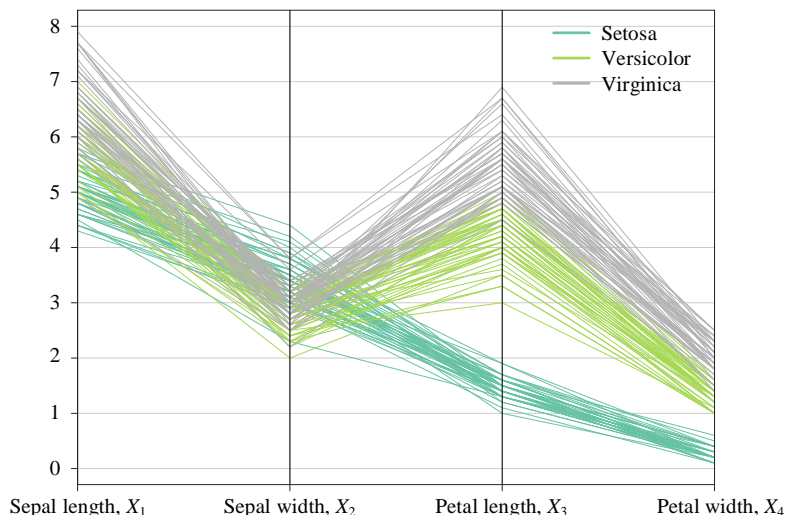


图 17. 鸢尾花数据的平行坐标图

## 2.4 集中度：均值、质心

本章前文通过可视化来展示数据的分布情况，本章后文介绍几种最基本的量化手段来描述样本数据。

估计总体集中趋向 (central tendency) 的最基本方法是算数平均数 (arithmetic mean):

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

如果数据是总体，算数平均数是总体平均数 (population mean)。如果数据是样本，算数平均数是样本平均数 (sample mean)。请大家回顾《矩阵力量》中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：

$$\mu_1 = 5.843, \mu_2 = 3.057, \mu_3 = 3.758, \mu_4 = 1.199$$

(6)

图 3 所示为鸢尾花四个特征均值在直方图位置。

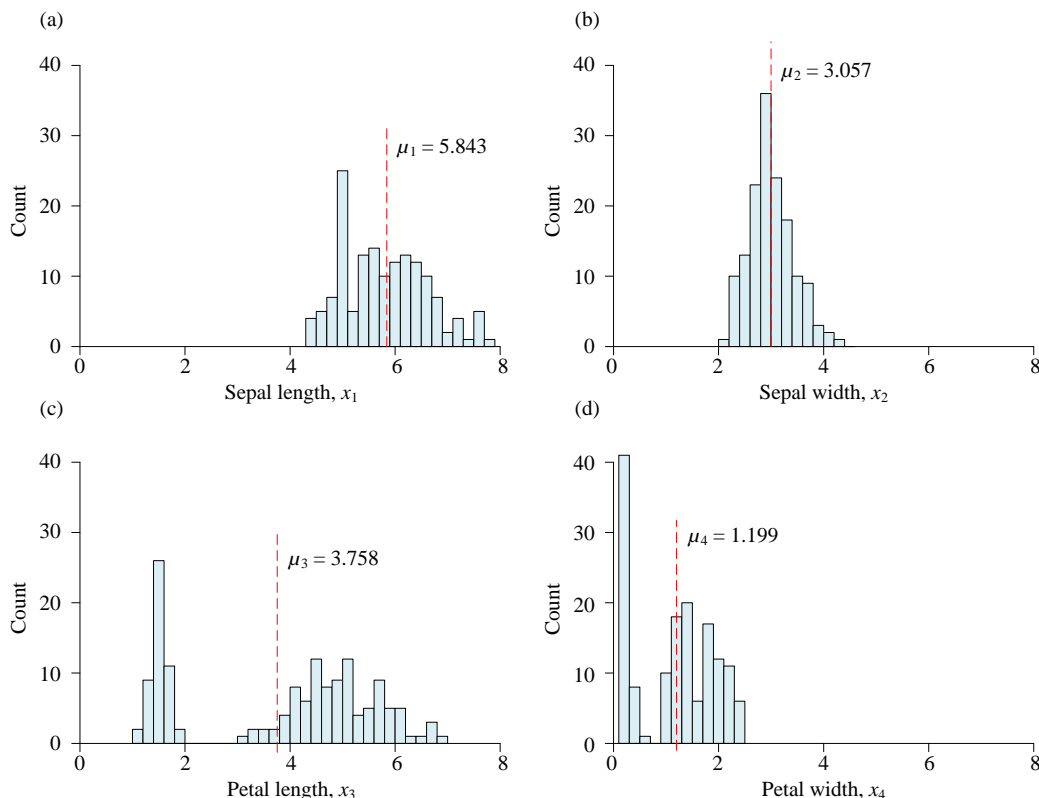


图 18. 鸢尾花四个特征数据均值在直方图位置

## 质心

当然，我们也可以把均值位置标注在散点图上。可以发现两个特征的均值相较于一点，这一点常被称作数据的质心 (centroid)。图 19 中红色 × 为花萼长度、花萼宽度的质心位置。

鸢尾花数据矩阵  $\mathbf{X}$  质心为：

$$\mathbf{E}(\mathbf{X}) = \boldsymbol{\mu}_X^T = \begin{bmatrix} 5.843 & 3.057 & 3.758 & 1.199 \\ \text{Sepal length, } x_1 & \text{Sepal width, } x_2 & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{bmatrix}^T \quad (7)$$

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)，除非有特别需求。

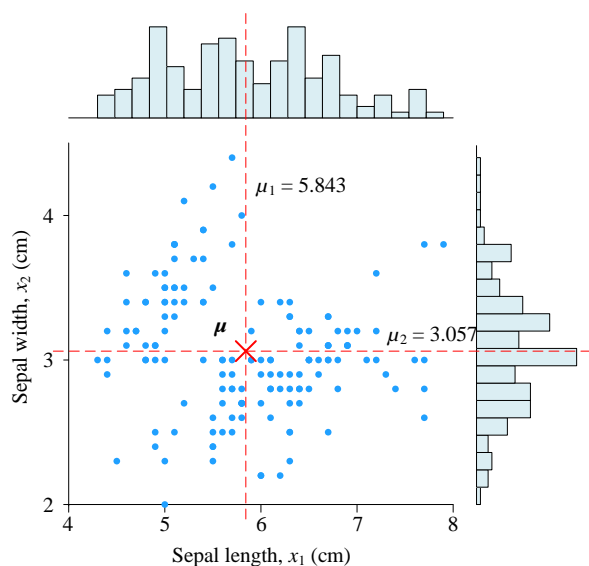


图 19. 均值在散点图的位置

## 分类标签

分别计算鸢尾花不同类别 (setosa、versicolor、virginica) 花萼长度、花萼宽度平均值：

$$\begin{aligned}
 \mu_{1\_setosa} &= 5.006, & \mu_{2\_setosa} &= 3.428 \\
 \mu_{1\_versicolor} &= 5.936, & \mu_{2\_versicolor} &= 2.770 \\
 \mu_{1\_virginica} &= 6.588, & \mu_{2\_virginica} &= 2.974
 \end{aligned} \tag{8}$$

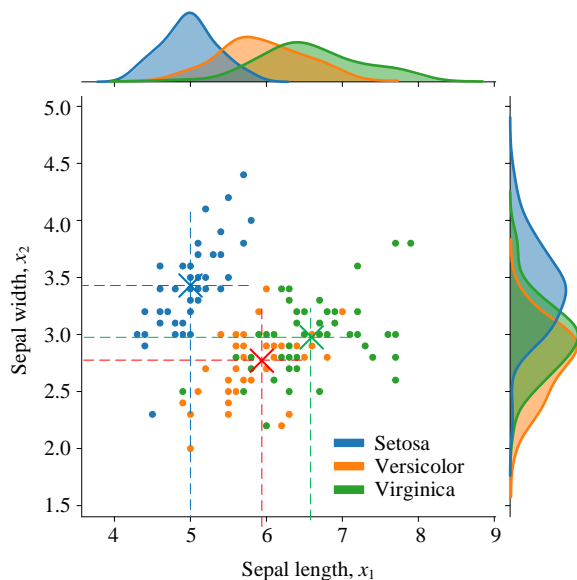


图 20. 均值在散点图的位置，考虑类别标签

## 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；如果观察值有偶数个，通常取最中间的两个数值的平均数作为中位数。本章后文在分位相关内容中，我们还会提到中位数。

众数 (mode) 是一组数中出现最频繁的数值。

几何平均数 (geometric mean) 的定义如下：

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

几何平均数有自身的应用局限性，它只适合正数。

## 2.5 分散度：极差、方差、标准差

### 极差

极差 (range)，又称全距，是指最大值与最小值之间的差距，即最大值减最小值的结果。极差可以用来度量分散度的最简单的指标：

$$\text{range}(X) = \max(X) - \min(X) \quad (9)$$

图 21 所示为最大值、最小值、极差、均值之间关系。

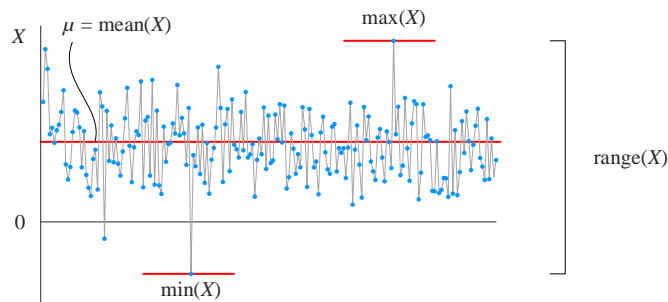


图 21. 最大值、最小值、极差、均值的关系

### 方差

方差 (variance) 衡量随机变量或样本数据离散程度。

样本的方差为：

$$\text{var}(X) = \sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)^2 \quad (10)$$

方差的单位是原始数据的平方，比如鸢尾花数据方差单位为  $\text{cm}^2$ 。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》中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：

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

鸢尾花四个量化特征的标准差：

$$\sigma_1 = 0.825, \sigma_2 = 0.434, \sigma_3 = 1.759, \sigma_4 = 0.759 \quad (12)$$

请读者格外注意，均方差和原始数据的单位是一致的。鸢尾花四个特征的量化数据单位均为厘米 (cm)。图 22 上，我们把  $\mu \pm \sigma$ 、 $\mu \pm 2\sigma$  位置也画在了直方图上。

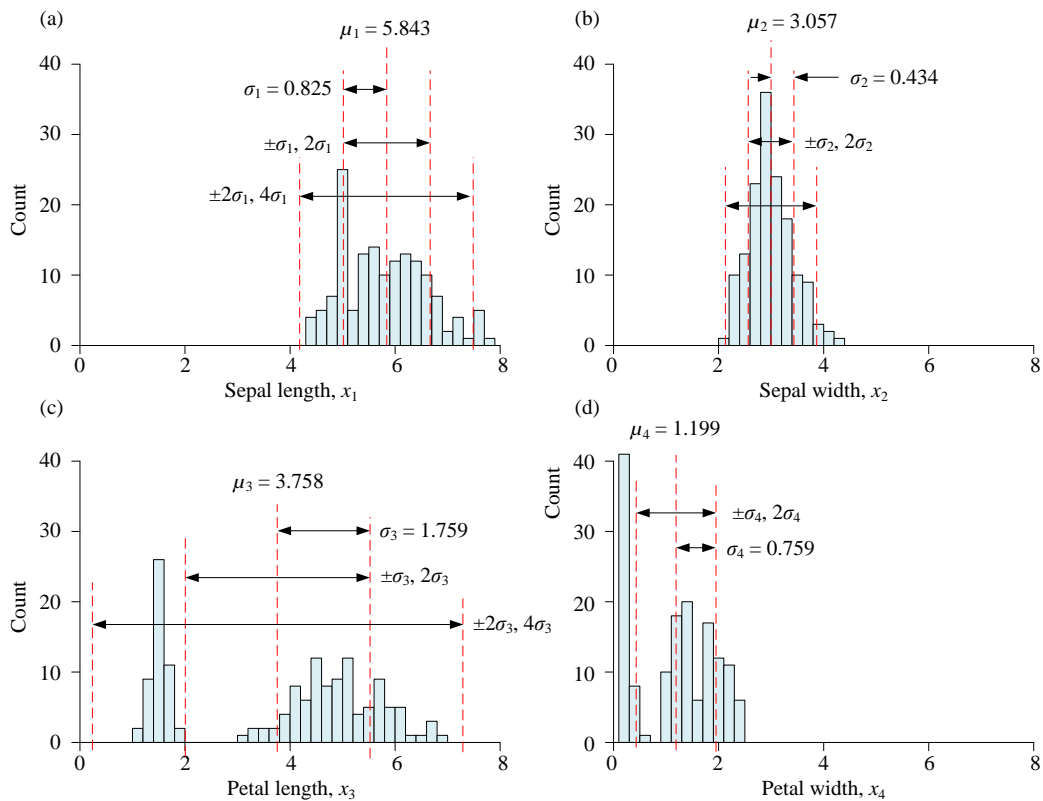


图 22. 鸢尾花四个特征数据均值、均方差所在位置在直方图位置

## 绝对中位差

绝对中位差 (mean absolute deviation, MAD) 的定义为：

$$\text{mad}(X) = \frac{1}{n} \sum_{i=1}^n |x^{(i)} - \mu_X| \quad (13)$$

MAD 的单位和原数据个。一般情况，方差、标准差比 MAD 更常用。

## 2.6 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。

本节主要介绍四分位和百分位。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将它们平分为四等份。而 50% 分位对应中位数。图 23 所示为将鸢尾花不同特征的四分位画在直方图上。

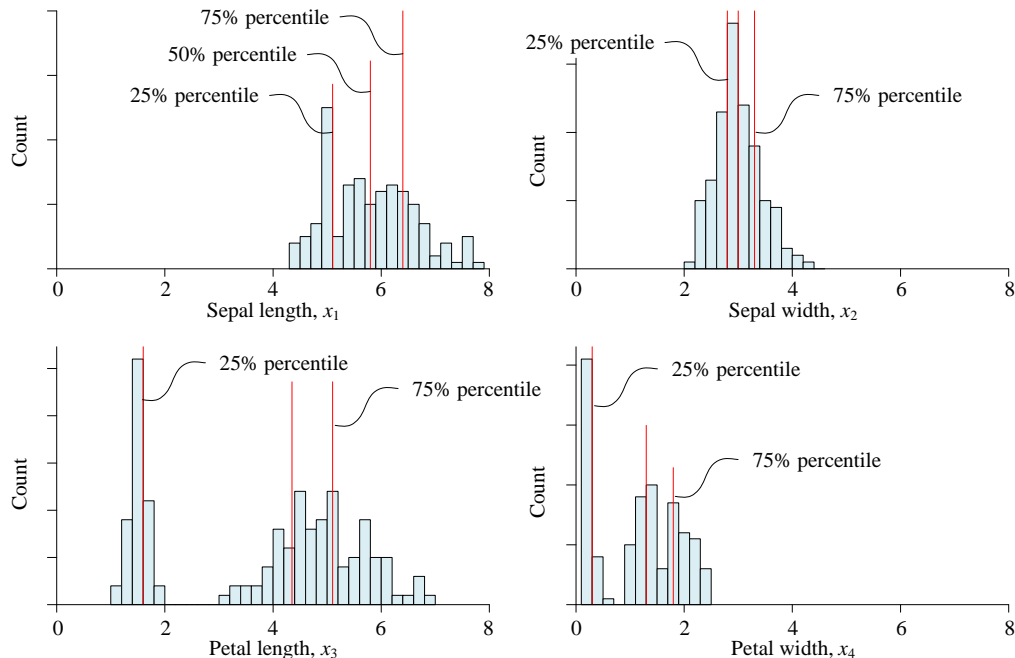


图 23. 鸢尾花数据直方图，以及 25%、50% 和 75% 百分位



图 24 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% 分别是数据的“左尾”、“右尾”。

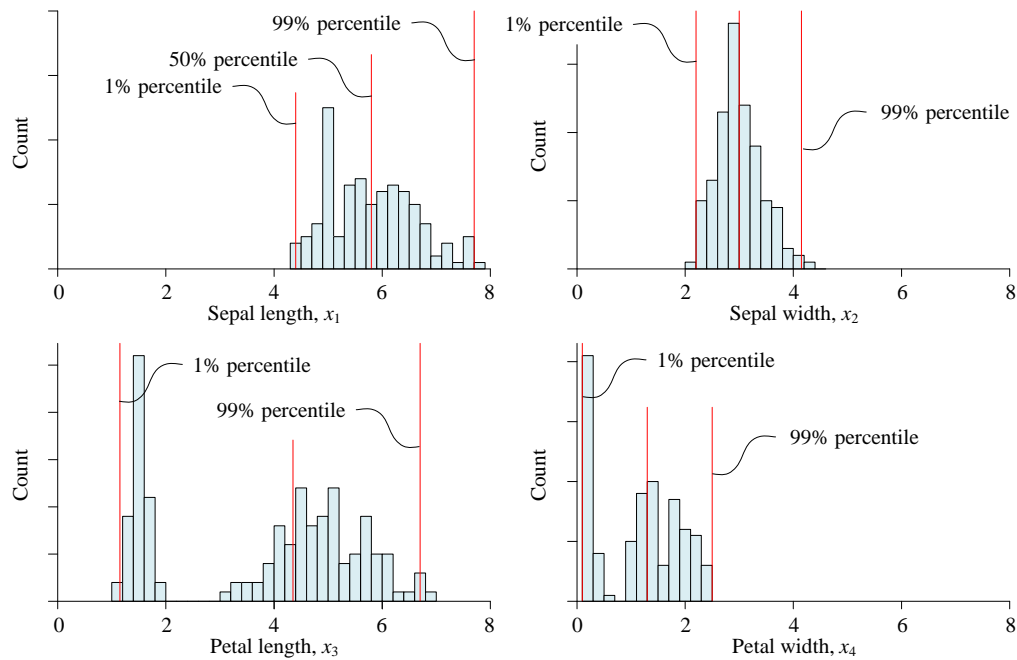


图 24. 鸢尾花数据直方图，以及 1%和 99%百分位

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 25 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 25. 鸢尾花数据帧统计总结

## 2.7 箱型图：小提琴图、分布散点图

**箱型图** (box plot)，图 26 所示为箱型图原理。箱型图箱型图利用第一 ( $Q_1$ )、第二 ( $Q_2$ ) 和第三 ( $Q_3$ ) 四分位数展示数据分散情况；箱型图也可以用来分析离群点。

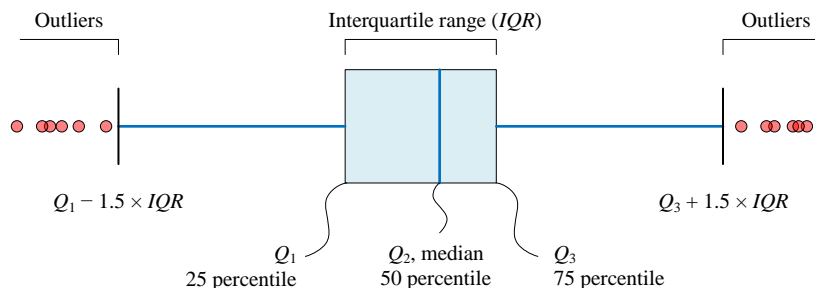


图 26. 箱型图原理

箱型图的**四分位间距** (interquartile range):

$$IQR = Q_3 - Q_1 \quad (14)$$

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。

图 27 所示为鸢尾花数据的箱型图。 $Q_1$  也叫下四分位， $Q_2$  也叫中位数， $Q_3$  也称上四分位。 $Q_3 + 1.5 \times IQR$  也称上界， $Q_1 - 1.5 \times IQR$  叫下界。

### 箱型图的变体

箱型图还有很多的“变体”。比如图 28 所示的小提琴图，图 29 所示的分布散点图。图 30 所示为箱型图叠加分布散点图。图 31 所示为考虑标签的箱型图。

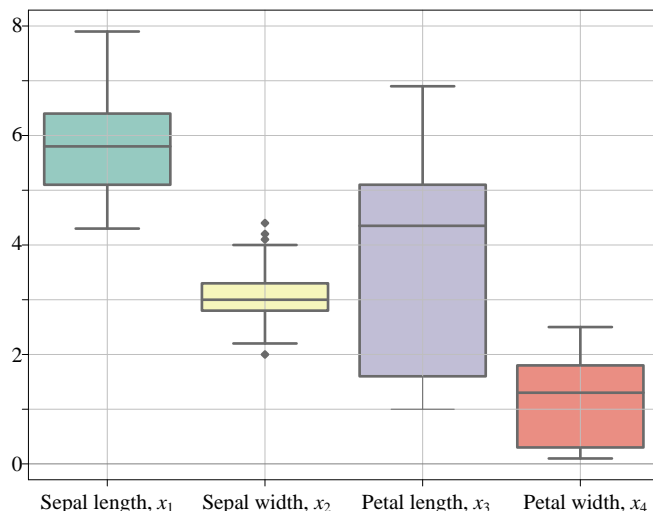


图 27. 鸢尾花数据箱型图

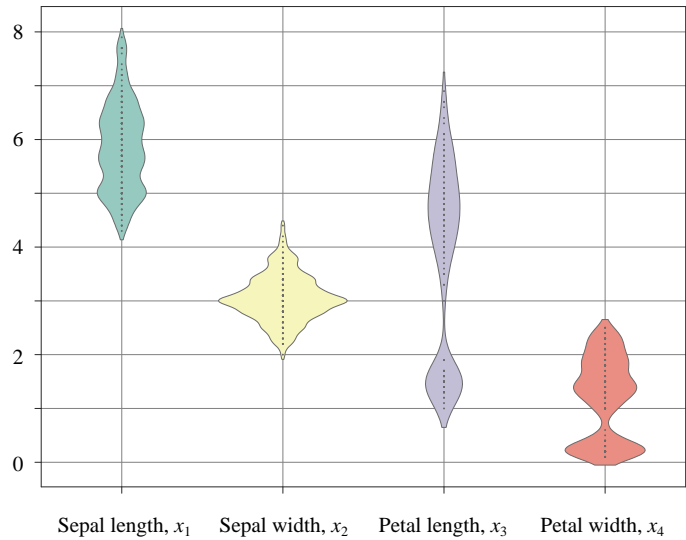


图 28. 鸢尾花数据小提琴图

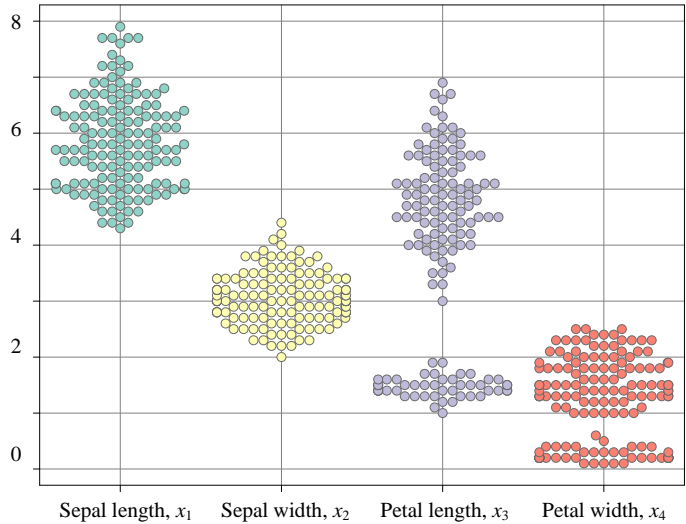


图 29. 分布散点图 (stripplot)

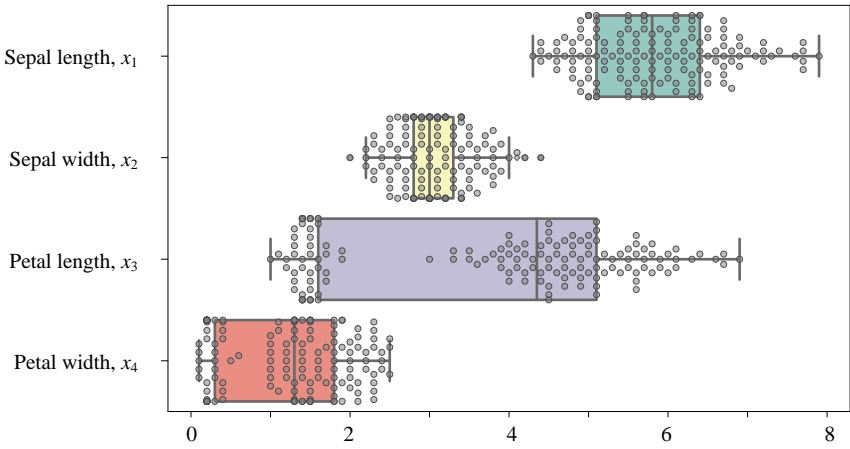


图 30. 鸢尾花箱型图，叠加分布散点图 swarmplot

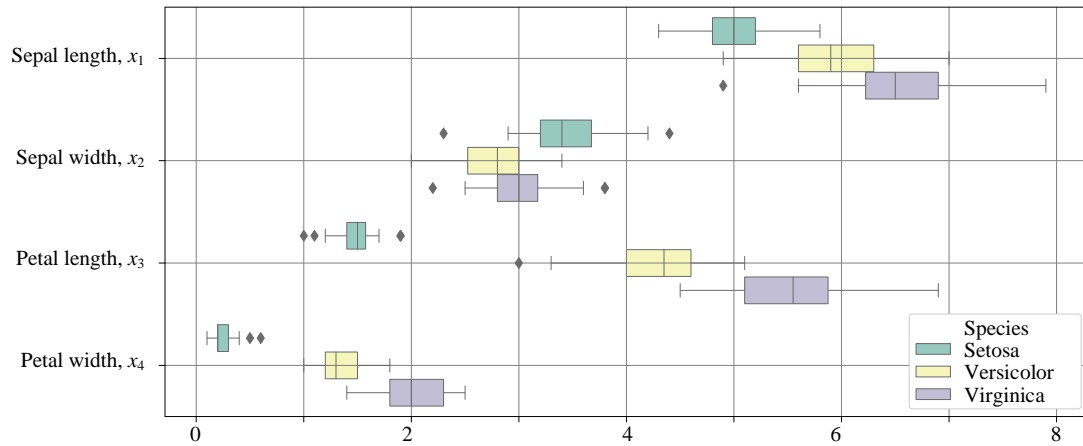


图 31. 鸢尾花箱型图，考虑分类标签

## 2.8 中心矩：均值、方差、偏度、峰度

统计学中的矩 (moment)，是对变量分布和形态特点进行度量的一组量，其概念来自于物理学中的“矩”。在物理学中，矩是描述物理性状特点的物理量。

在统计学中，矩，又称为中心矩 (central moment) 同样用于描述随机变量分布的特点。零阶矩表示这些点的总概率 (也就是 1)。

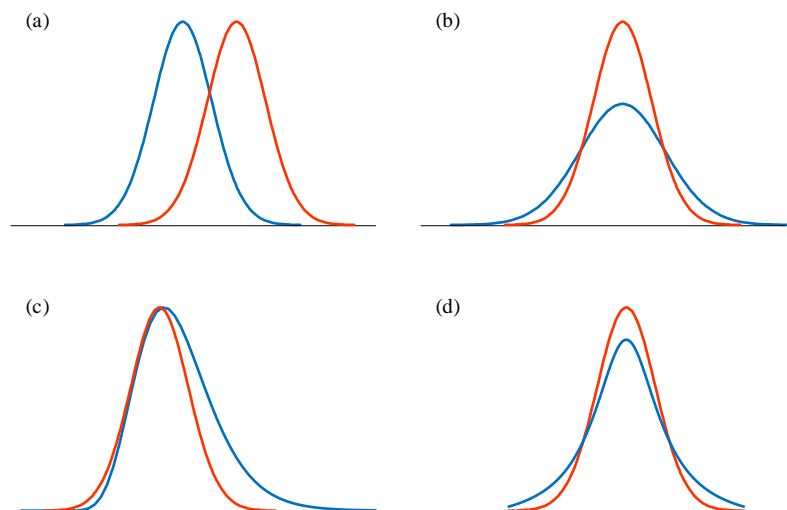


图 32. 期望 (一阶矩)、方差 (二阶矩)、偏斜度 (三阶矩)、峰度 (四阶矩)

### 一阶矩、二阶矩

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)

一阶矩为均值，即期望 (expectation)，用来描述分布中心位置，如图 15 (a) 所示。二阶矩为方差 (variance)，描述分布分散情况，如图 15 (b) 所示。

虽然一元高斯分布的参数仅需要均值和方差。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，这就需要偏度和峰度。

### 三阶矩

如图 15 (c) 所示，三阶矩为偏度 (skewness)  $S$  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、中位数和众数，处在同一位置，图 33 (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 33 (b) 所示，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，这三个数值的关系。如果在样本中引入几个极大的离群值的话，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)，但是众数不变。

负偏 (negative skew, negatively skewed)，又称左偏 (left-skewed, left-tailed, skewed to the left)，如图 33 (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。

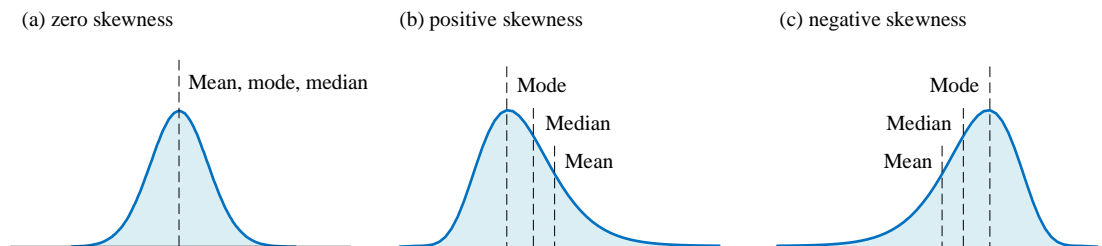


图 33. 无偏、正偏和负偏

值得注意的是，偏度为零不一定意味着分布对称。如图 34 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。

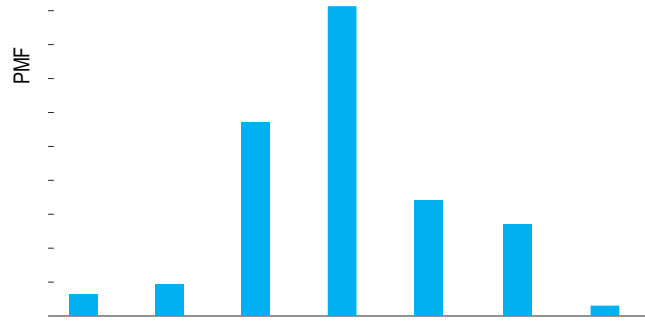


图 34. 偏度为 0，但是不对称的分布

## 四阶矩

图 15 (d) 所示，四阶矩表示峰度 (kurtosis)  $K$  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。

图 35 展示两种峰态：高峰态 (leptokurtic) 和低峰态 (platykurtic)。高峰度的峰度值大于 3。如图 35 (a) 所示，和正态分布相比，高峰态分布有明显的尖峰，两侧稍后，两侧尾端有肥尾 (fat tail)。

图 35 (b) 展示的是低峰态。相比正态分布，低峰态明显稍扁。但是有意思的是低峰态尾部更薄，这是因为概率密度函数和横轴构成的图形面积为 1。

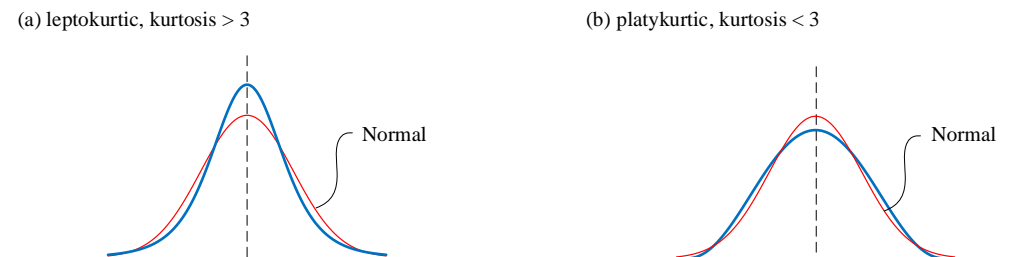


图 35. 高峰态和低峰态


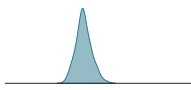

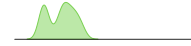
实践中，一般采用超值峰度 (excess kurtosis)，即 (16) 减去 3：

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

“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。

表 1. 鸢尾花四阶矩

				
均值 (cm)	5.843	3.057	3.758	1.199
均方差 (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

## 2.9 协方差矩阵、相关性系数矩阵

对于样本数据，随机变量  $X$  和  $Y$  的协方差为：

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

对于样本数据，随机变量  $X$  和  $Y$  的相关性系数：

$$\rho_{x,y} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \quad (19)$$

本系列丛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。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。

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 36 所示为协方差矩阵和相关性系数矩阵热图。

本书第 12 章将专门讲解协方差矩阵。

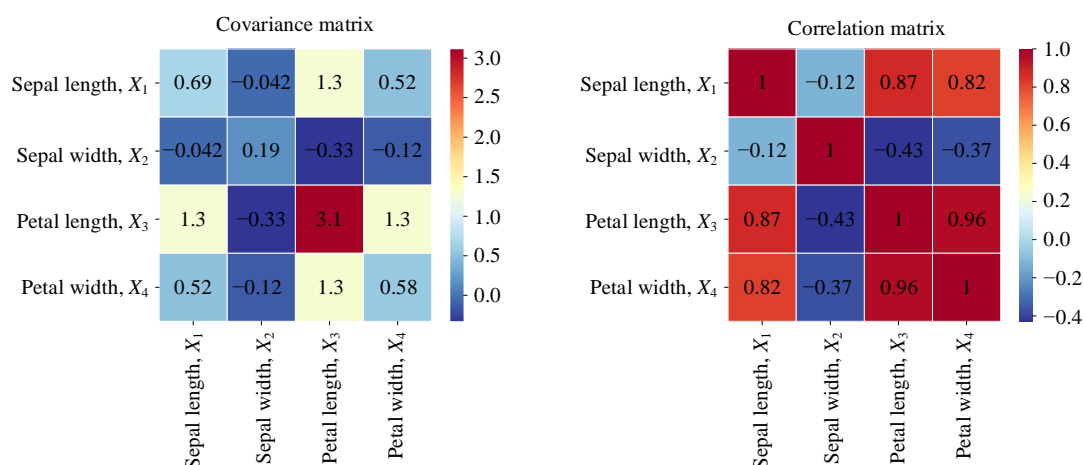


图 36. 协方差矩阵、相关性系数矩阵热图



代码文件 Bk5\_Ch02\_01.py 绘制本章几乎所有图像。

