

10

2D Scatter Plots

平面散点

通过散点颜色、大小呈现更多特征



理论上可以用科学描述一切，但这没有意义；这就像把贝多芬的交响乐描述为一组声波，毫无意义。

It would be possible to describe everything scientifically, but it would make no sense; it would be without meaning, as if you described a Beethoven symphony as a variation of wave pressure.

—— 阿尔伯特·爱因斯坦 (Albert Einstein) | 理论物理学家 | 1879 ~ 1955



- ▶ `matplotlib.patches.Circle()` 创建正圆图形
- ▶ `matplotlib.pyplot.scatter()` 绘制散点图
- ▶ `numpy.exp()` 计算括号中元素的自然指数
- ▶ `numpy.linspace()` 在指定的间隔内, 返回固定步长的数据
- ▶ `numpy.meshgrid()` 创建网格化数据
- ▶ `numpy.random.rand()` 生成满足均匀分布的随机数
- ▶ `numpy.random.randn()` 生成满足标准正态分布的随机数
- ▶ `seaborn.scatterplot()` 绘制散点图
- ▶ `sklearn.neighbors.KernelDensity()` 概率密度估计函数

10.1 平面散点图

点动成线，线动成面，面动成体。本章介绍如何在平面上绘制最基本的散点图。本书后中，大家会发现，线图也是散点的连线；等高线、曲面也离不开点。

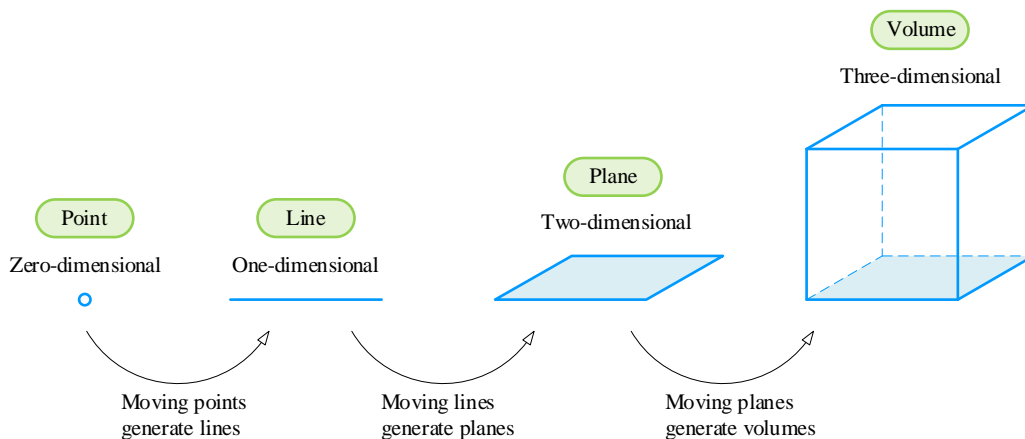


图 1. 点动成线，线动成面，面动成体

规则网格

平面散点图是重要的可视化工具。如图 2 (a) 所示，在平面网格散点基础上用颜色渲染可以可视化 3D 数据。进一步提高颗粒度，我们可以得到更加丰满的平面图像，如图 2 (b) 所示。这一点，我们在本书后文三维散点图中还会看到。

散点大小

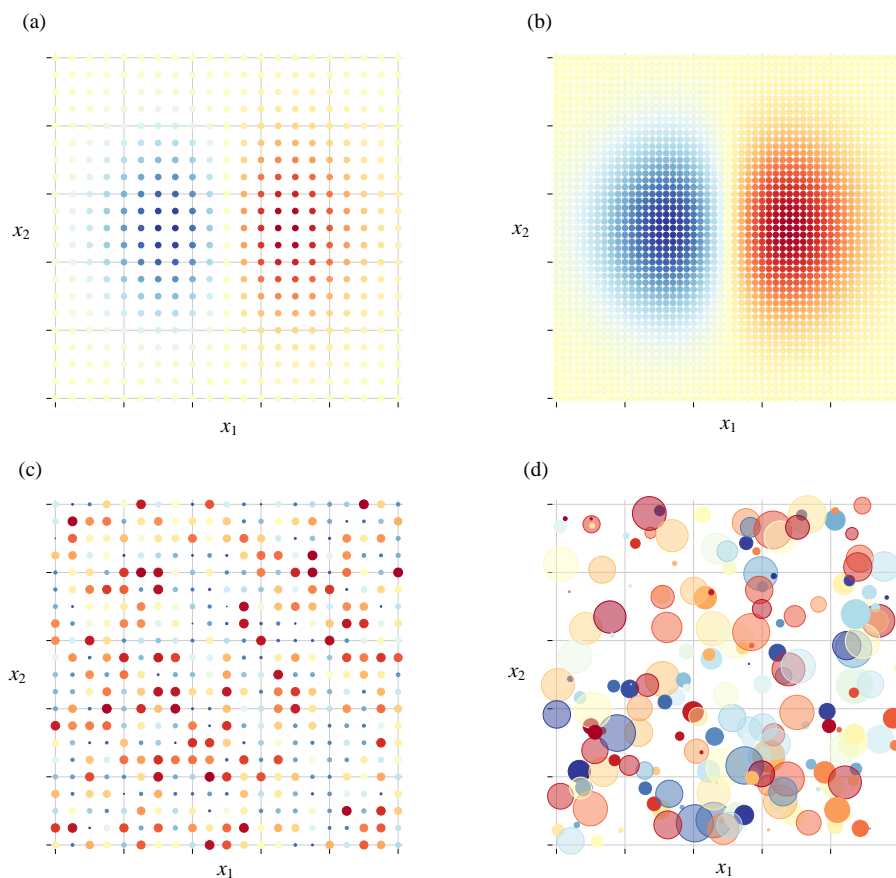
如图 2 (c) 所示，除了颜色，我们还可以用散点大小展示数据特征。

随机散点

除了规则网格散点，我们更常用平面散点可视化随机散点，比如图 2 (d)。因此，平面散点常用来可视化样本数据。



Jupyter 笔记 BK_2_Ch10_1.ipynb 绘制图 2。

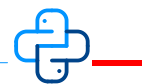
图 2. 使用 `matplotlib.pyplot.scatter()` 绘制平面散点图

10.2 样本数据

可视化样本数据

图 3 所示为平面散点可视化鸢尾花样本数据。这幅图中，我们可以用颜色、大小、标记符号可视化更多特征。

本书后文将专门介绍用 `seaborn.scatterplot()` 绘制散点图。



Jupyter 笔记 BK_2_Ch10_2.ipynb 绘制图 3。

10.3 使用面具

图 4、图 5 所示为用面具 (mask)，也叫蒙皮，区分满足不同条件的散点。

68-95-99.7 法则

图 4 中，大家会看到一组服从高斯分布的散点。以 0 ± 2 为界， $[-2, 2]$ 区间之内的散点用原点 ● 展示； $[-2, 2]$ 区间之外的散点用叉 × 代表。这体现的实际上是 68-95-99.7 法则。

68-95-99.7 法则是一种统计学中的规则，也被称为“三个标准差法则”或“标准差法则”。该法则用于描述服从高斯分布样本数据分布情况。根据 68-95-99.7 法则，对于一个符合正态分布的数据集，大约：68% 的数据值会落在均值的一个标准差范围内；95% 的数据值会落在均值的两个标准差范围内；99.7% 的数据值会落在均值的三个标准差范围内。

注意，图 4 中样本数据的均值为 0，标准差为 1。 $[-2, 2]$ 区间之内约有 95% 样本数据。

换句话说，大约 68% 的数据会分布在均值左右一个标准差的范围内，约 95% 的数据会分布在均值左右两个标准差的范围内，而约 99.7% 的数据会分布在均值左右三个标准差的范围内。这个法则在统计学和数据分析中被广泛应用，用于估计数据的分布情况和识别异常值。它提供了一种简单而有用的方法来理解和描述正态分布的特性。

鸢尾花书《统计至简》第 9 章将专门讲解一元高斯分布。



Jupyter 笔记 BK_2_Ch10_3.ipynb 绘制图 4。请大家想办法区分 68-95-99.7 对应的不同区间。

蒙特卡罗模拟估算圆周率

蒙特卡罗模拟是一种使用随机抽样的方法来估算数值的技术，可以用于估算圆周率。下面是使用蒙特卡罗模拟来估算圆周率的一般步骤。

- ▶ 假设有一个边长为 2 的正方形，其中包含一个半径为 1 的圆。
- ▶ 在正方形内部随机生成一组点，可以通过在正方形内均匀抽样得到。每个点都有一个 x 和 y 坐标，均在 $[-1, 1]$ 的范围内。
- ▶ 对于每个生成的点，计算其到原点的距离。
- ▶ 如果距离小于等于 1，表示该点在圆内或圆上，否则在圆外。
- ▶ 统计在圆内的点的数量和正方形内生成的总点数。
- ▶ 估算圆周率的值可以通过以下公式计算： $\pi \approx (4 \times \text{圆内点的数量}) / (\text{正方形内生成的总点数})$ 。

随着生成的点数增多，根据蒙特卡罗模拟的原理，估算得到的圆周率值会逐渐接近真实值 π 。因此，增加生成的点数可以提高估算的准确性。

需要注意的是，蒙特卡罗模拟是一种概率估算方法，结果的准确性取决于随机性和抽样点的数量。在实际应用中，通常需要生成大量的点才能得到比较准确的估算结果。

鸢尾花书《统计至简》第 15 章将专门讲解蒙特卡罗模拟。



Jupyter 笔记 BK_2_Ch10_3.ipynb 绘制图 5。

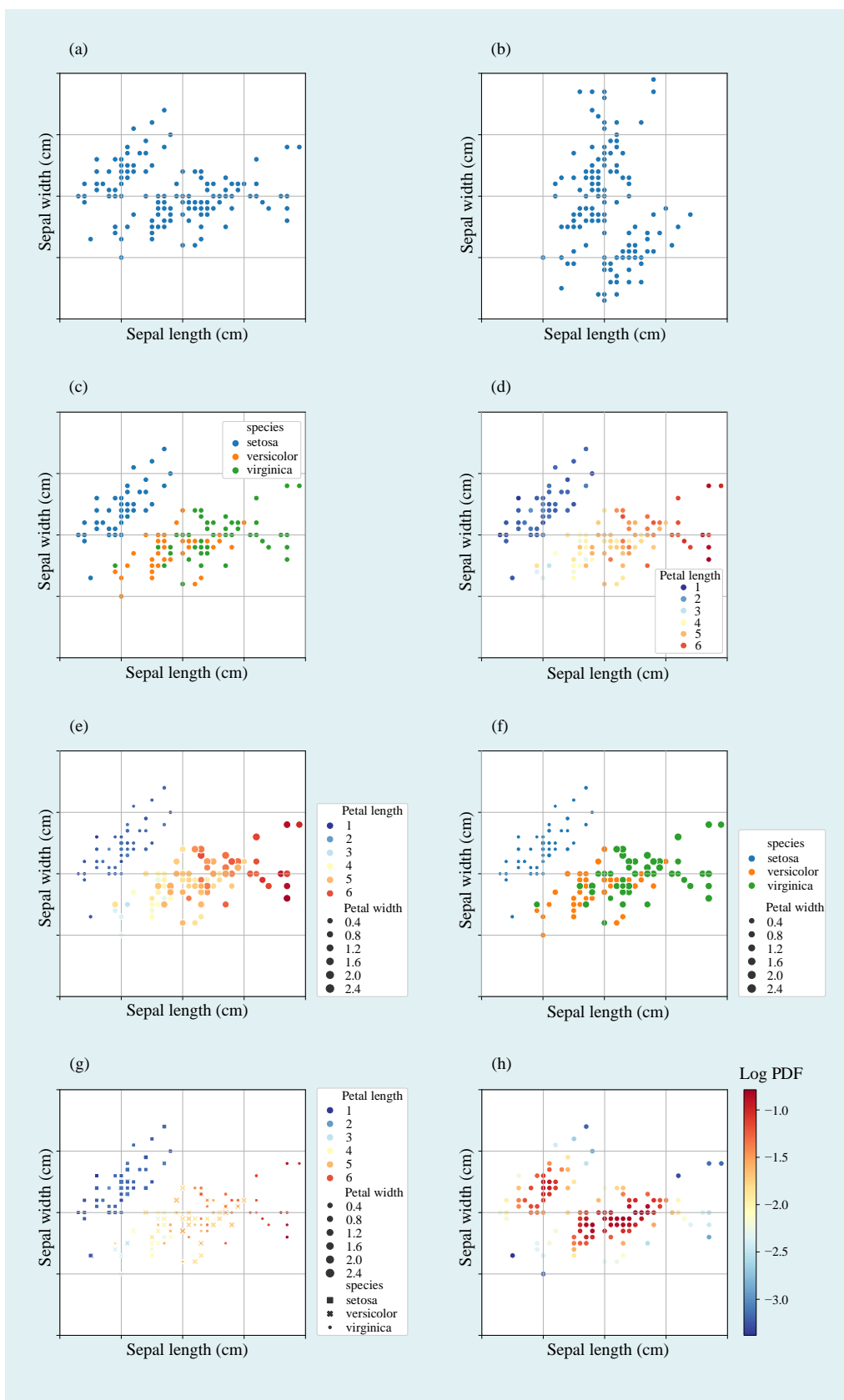


图 3. 用平面散点图可视化鸢尾花数据

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

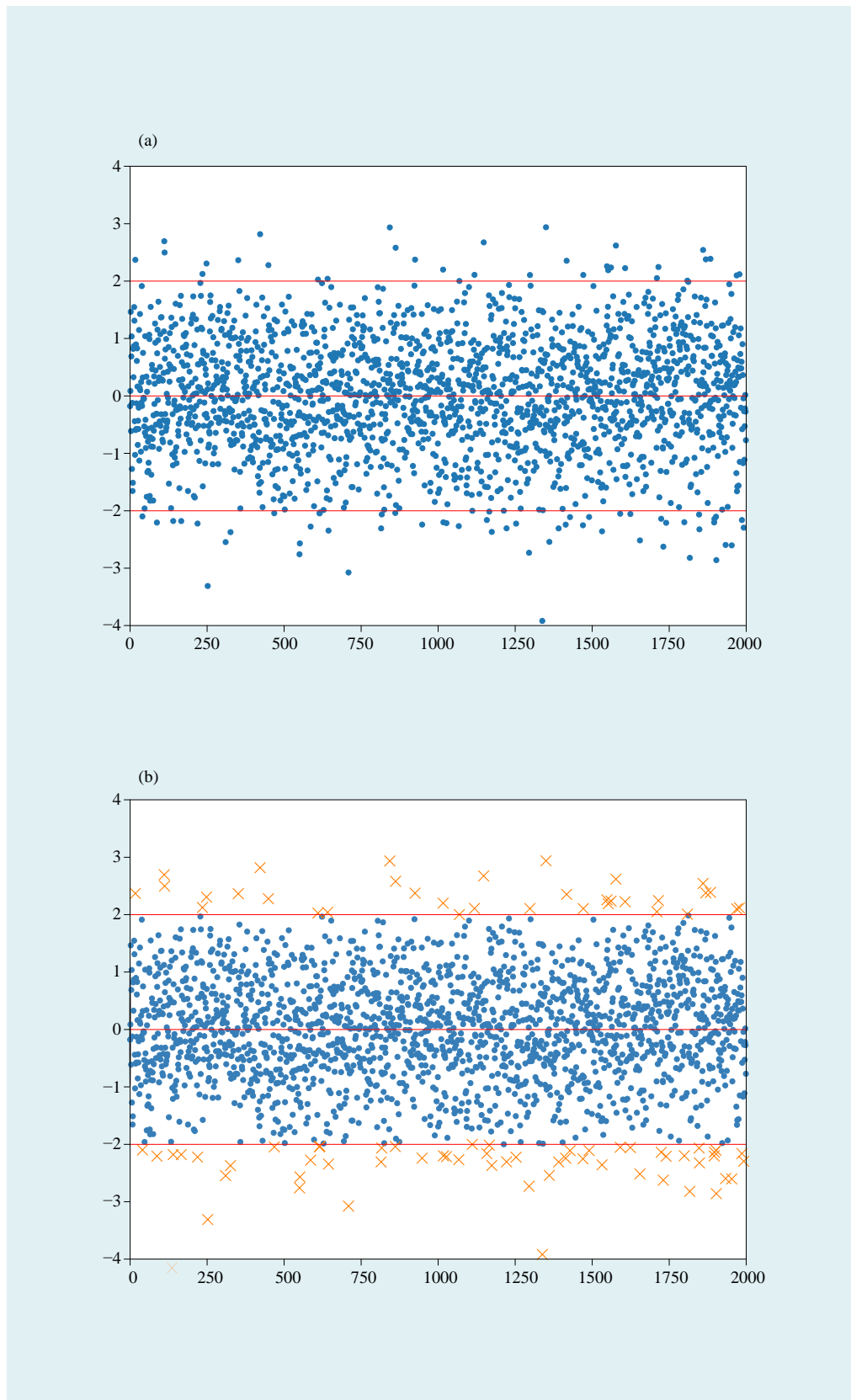


图 4. 使用面具可视化可能的离群值

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

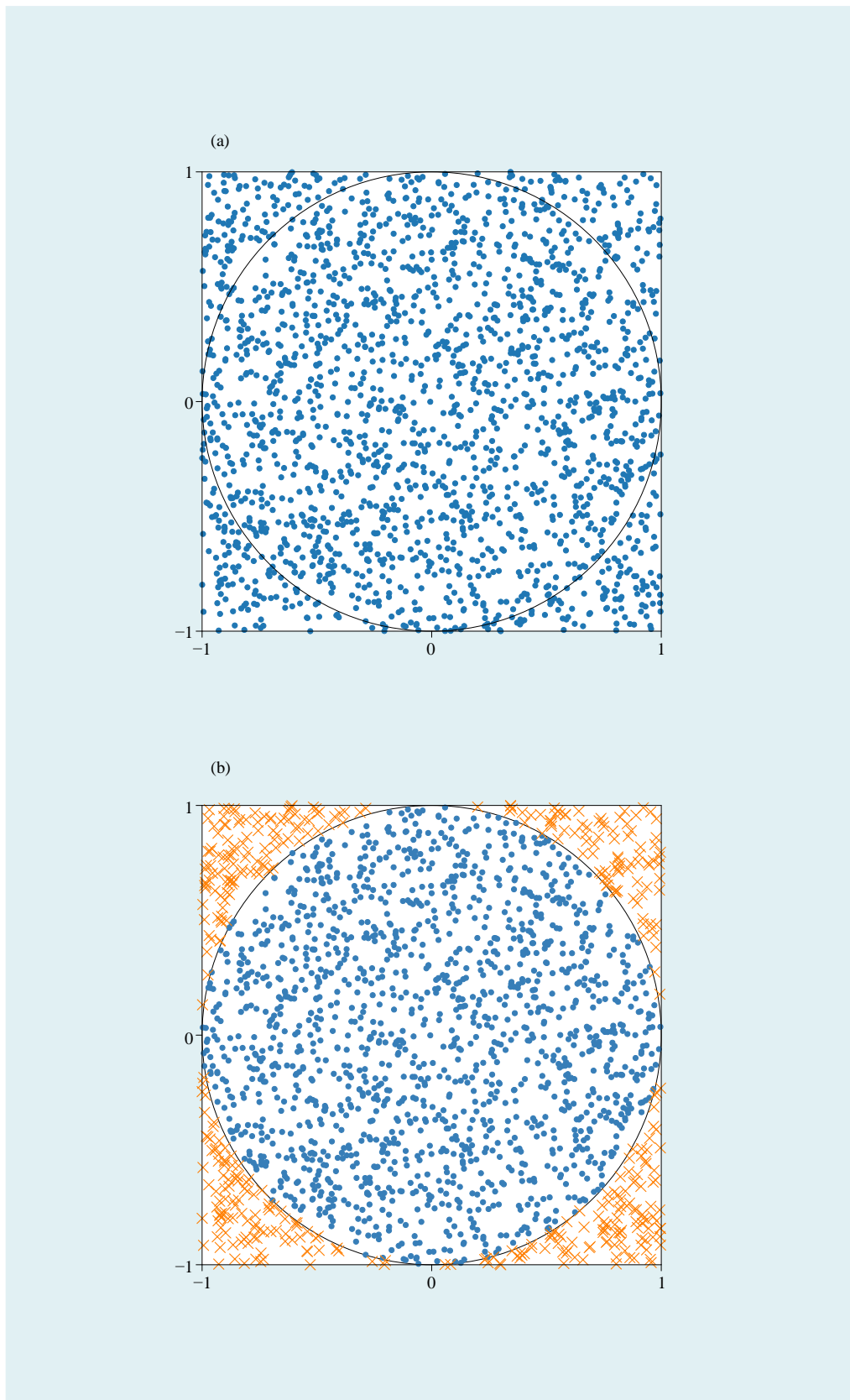


图 5. 蒙特卡洛模拟估算圆周率

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