

Python数据可视化在统计学课程教学中的应用

何 鑫

西华大学理学院, 四川 成都

收稿日期: 2025年1月13日; 录用日期: 2025年2月14日; 发布日期: 2025年2月24日

摘 要

随着信息技术的发展, Python在统计学课程教学中得到广泛应用。传统的教学方法难以直观地帮助学生理解抽象的统计概念, 而Python提供的可视化工具能够通过图示增强学生对数据统计、概率分布和参数估计等知识点的理解。本文探讨了Python数据可视化在统计学课程教学中的应用, 分析了其在加深学生知识点理解、问题处理和编程技能提升等方面的优势, 并通过具体的教学案例展示了其应用场景。研究表明, Python数据可视化能显著提高学生的学习效果, 为统计学的教学改革有一定的促进作用。

关键词

Python, 数据可视化, 统计学, 教学设计

The Application of Python Data Visualization in Teaching Statistics Courses

Xin He

School of Science, Xihua University, Chengdu Sichuan

Received: Jan. 13th, 2025; accepted: Feb. 14th, 2025; published: Feb. 24th, 2025

Abstract

With the development of information technology, Python has been widely used in the teaching of statistics courses. Traditional teaching methods often fail to intuitively help students understand abstract statistical concepts, while Python's visualization tools can enhance students' understanding of key concepts such as data statistics, probability distributions, and parameter estimation through graphical representation. This paper explores the application of Python data visualization in statistics education, analyzes its advantages in deepening students' understanding of key concepts,

problem analysis, and programming skills, and demonstrates its application scenarios through specific teaching cases. The research indicates that Python data visualization can significantly improve students' learning outcomes and contribute to the reform of statistics education.

Keywords

Python, Data Visualization, Statistics, Teaching Design

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着大数据时代的到来,统计学作为一门基础学科,越来越受到各行各业的重视。然而,传统的统计学教学方法往往侧重于理论推导和数学公式讲解,学生往往难以在实践中理解抽象的统计概念,造成学习效果不理想。近年来,随着信息技术和计算机科学的发展,Python 语言的普及以及其丰富的数据可视化工具,已经成为统计学教育的重要辅助工具[1]-[4]。通过 Python 数据可视化,学生能够更加直观、清晰地理解统计学中的核心概念,如数据分布、假设检验、回归分析等[5]-[7]。本论文旨在探讨 Python 数据可视化在统计学课程教学中的应用,分析其对教学方法和学生学习效果的影响,并通过实际案例探讨如何将其有效融入到课堂教学中。

2. Python 数据可视化在统计学教学中的必要性

2.1. 统计学课程教学中的挑战

统计学的许多概念和方法对大多数学生来说都较为抽象,学生很难通过理论推导来理解这些抽象概念。比如概率分布、假设检验、回归分析等,学生对这些抽象的概念通常会感到枯燥和难以理解[8]。在传统的统计学教学中,多以教师讲解、学生理解和课后练习为主,缺乏互动性和实践性。学生通常很难将学习的统计知识应用到实际问题中。而且,传统的教学方法无法充分调动学生的积极性,导致学习效果不如预期。随着现代信息技术的发展,学生不仅需要掌握理论知识,还得具备一定的数据分析能力[9]。然而,学生无法在课堂中掌握数据分析的经验,且编程能力较弱。

2.2. Python 数据可视化概述

数据可视化是通过图形、图表等形式将数据中的信息展现出来,帮助人们直观地理解数据背后的规律。在统计学教学中,数据可视化可以帮助学生快速理解理论知识,掌握应用场景,获取数据信息,加深学生对统计学方法和结果的理解和认识。Python 和 R 语言都是数据分析领域中常用的编程语言,各自拥有强大的数据处理和可视化能力[10]-[12]。在统计学课程中,这两种语言都能够有效地辅助教学,但它们在可视化方面有一些不同的特点和优势。R 语言专为统计分析和数据可视化设计,拥有如 ggplot2 等强大的库,适合学术研究和复杂的统计图表制作。然而,与 python 相比,R 的语法较为复杂,学习曲线较陡,且应用场景相对比较小。Python 作为通用编程语言,语法简洁、易于上手,适合没有任何编程背景的学生,同时也可以为后续学习人工智能等知识建立程序语言基础。Python 的可视化工具不仅支持基本的统计图表,还能生成交互式图表,与数据处理、机器学习等多种任务无缝集成。Python 的灵活性和广泛性使其在统计和数据科学领域有着更广阔的应用前景,尤其适用于统计教学中的跨学科应用场景。因

此, Python 在统计学课程教学中具有易学性、灵活性和多功能性, 是比 R 语言更适合的工具。接下来简单介绍 Python 常用的几个数据可视化工具库。

Matplotlib: Matplotlib 是 Python 中最常用的绘图库之一, 能够绘制各种二维图形, 如折线图、柱状图、直方图、散点图、小提琴图、饼图等, 适用于基本的数据可视化需求。

Seaborn: Seaborn 是基于 Matplotlib 的高级可视化库, 提供了更加美观和易用的图表绘制功能, 尤其适合统计学分析中常见的热力图、箱线图、分布图等。

Plotly: Plotly 是一个交互式图表库, 能够生成动态和交互式的图表, 适用于展示大规模数据和复杂数据关系。

Pandas: Pandas 不仅是 Python 中最常用的数据处理库, 也提供了一些简单的可视化功能, 能够帮助用户直接通过 DataFrame 对象绘制常见的图表。

2.3. Python 数据可视化在统计学教学中的优势和作用

直观展示抽象的统计概念: 统计学的许多概念(如概率分布、回归分析、假设检验等)通常较为抽象, 学生难以通过纯粹的数学公式或理论推导理解。Python 的数据可视化工具能够将这些概念通过图形化的方式呈现, 使学生能够直观地看到不同参数下的数据分布和统计分析的结果。例如, 通过绘制卡方分布图, 学生可以清晰地看到不同自由度下概率密度图像的分布特征以及与正态分布的惯性, 帮助他们理解卡方分布的性质及其在实际问题中的应用。

增强数据分析和编程能力: 在教学中, 学生不仅要理解统计学的理论, 还需要掌握数据分析的技能。Python 提供了强大的数据处理和分析库(如 Numpy, Pandas), 结合可视化工具, 学生可以在实践中学习如何进行数据预处理、数据分析、统计建模和数据可视化。通过编程, 学生能够动手解决实际问题, 从而提升他们的数据分析能力和编程技巧。这对于培养学生的学习兴趣, 统计思维和编程技能具有重要意义。

增加学习的互动性与参与感: 传统的统计学教学往往侧重于教师的课堂讲解和学生听课做笔记, 缺乏互动性。Python 的数据可视化能够促进交互式教学, 尤其是通过生成交互式图表, 学生可以通过操作图表实时调整参数、探索参数和数据间的关系。这种交互性极大提高了学生的参与感, 激发他们的学习兴趣, 也有助于深入理解数据分析和统计方法。

分解复杂的统计分析过程: 统计分析过程中会涉及很多的数学计算和理论推导, 这导致很多学生无法很好的理解这些统计方法和计算过程背后的原理。Python 能够将这些复杂的统计过程可视化, 例如在进行一元和多元线性回归分析时, 学生不仅可以看到不同数据下线性回归模型的拟合曲线, 还能通过交互式图表查看回归结果的变化, 从而更加好地理解线性回归分析的原理以及应用场景。

促进理论与实践的结合: Python 数据可视化能够帮助学生将理论知识与实际数据结合起来, 增强理论的实际应用感。通过实际操作, 学生可以更好地掌握统计学方法, 理解如何将这些方法应用于现实问题中。由于 Python 是一门通用编程语言, 它不仅可以应用于统计学课程, 还可以与其他学科的内容相结合, 促进跨学科的学习。例如, Python 可以与机器学习、人工智能、经济学等其他领域相结合, 帮助学生理解如何在不同的学科背景下应用统计学方法进行分析。这种跨学科的能力对于学生未来的学术研究或职业发展都有重要意义。

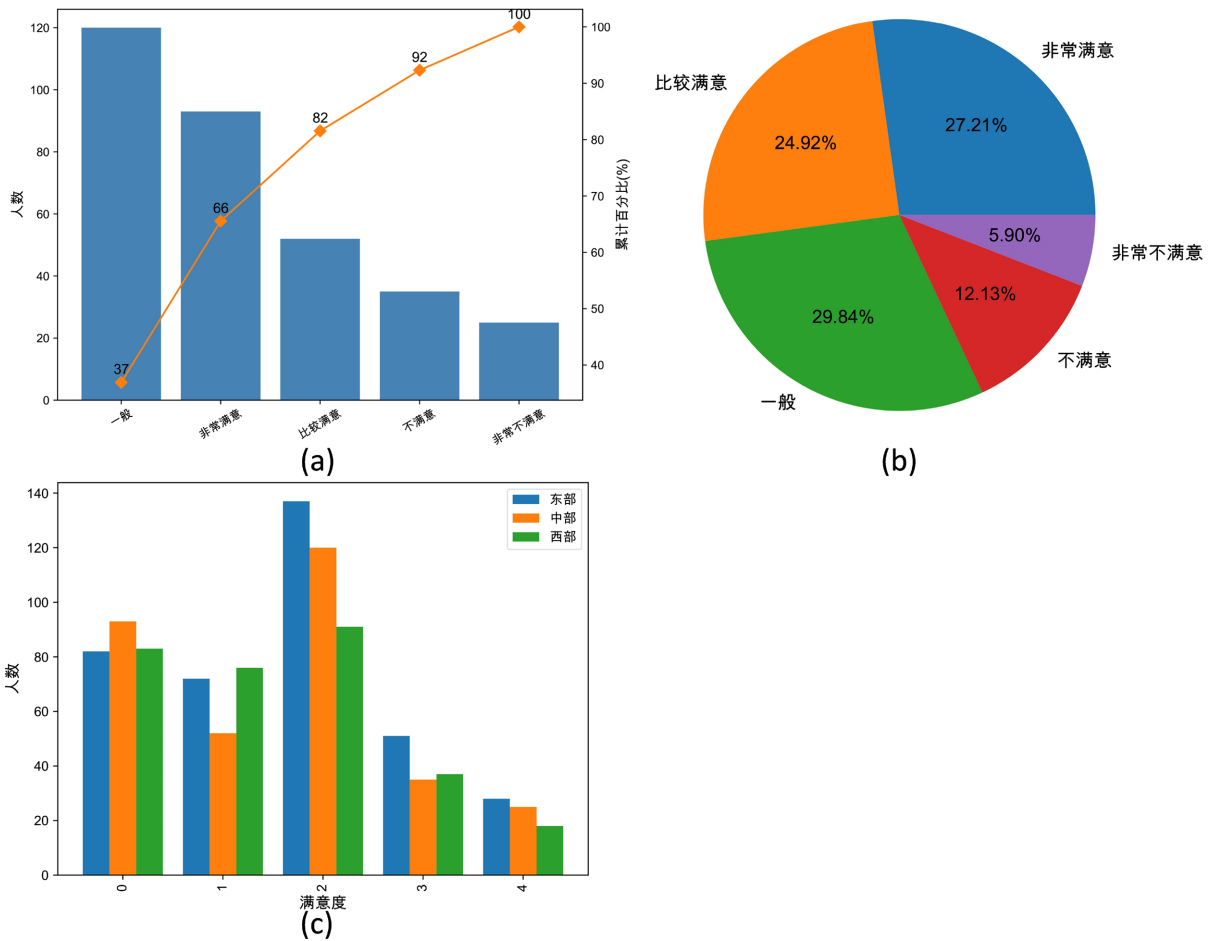
3. Python 数据可视化在统计学教学中的应用案例

本章我们将通过三个简单的教学案例说明 Python 数据可视化在统计学课堂教学中的应用, 说明其对学生理解知识的帮助。

3.1. 案例 1：统计数据的多种可视化方式

问题：为了解不同地区的消费者对网上购物的满意度，随机抽取东部、中部和西部 1000 名消费者进行调查，得到的存储在 data 文件夹的 table1.csv 文件中。通过绘制下述图形对数据进行分析：

- 根据中部地区的满意度数据，绘制帕累托图；
- 根据西部地区的满意度数据，绘制饼图；
- 根据东部地区、中部地区和西部地区的满意度数据，绘制并列条形图。



(a) 帕累托图；(b) 饼图；(c) 并列条形图。

Figure 1. Visualization result of Case 1

图 1. 案例 1 的可视化结果

解答：首先引入库并读取数据

```
import matplotlib.pyplot as plt
import pandas as pd
df = pd.read_csv('data/table1.csv')
```

a. 帕累托图

处理数据

```
df = df.sort_values(by='中部', ascending=False)
```

```

p = 100*df['中部'].cumsum()/df['中部'].sum()
df['累计百分比']=p
绘图
fig, ax = plt.subplots(figsize = (10,7))
ax.bar(df['满意度'], df["中部"], color="steelblue")
ax.set_ylabel('人数',size=12); ax.set_xlabel('满意度',size=12)
plt.xticks(range(5),df['满意度'],rotation=30)
ax2 = ax.twinx()
ax2.plot(df['满意度'], p, color="C1", marker="D", ms=7)
ax2.set_ylabel('累计百分比(%)',size=12)
for a,b in zip(df['满意度'],df['累计百分比']):
    plt.text(a,b+1, '%.0f%%b',ha='center',va='bottom',color='black',fontsize=12)
plt.title('中部地区满意度的帕累托',size=15);plt.show()
b. 饼图
p1=plt.pie(df['西部'],labels=df['满意度'],autopct='%1.2f%%')
plt.title('西部地区满意度饼图',size=13)
c. 并列条形图
df.plot(kind='bar',stacked=False,width=0.8)
plt.xlabel('满意度',size=12); plt.ylabel('人数',size=12)
plt.xticks(range(8),df['支出项目'],rotation=30)
plt.title('(a) 并列条形图',fontsize=13,color='black')
plt.tight_layout(); plt.show()

```

图1展示了python的不同数据的三种可视化方式，可以看出这种数据可视化可以直观的展示数据的信息和趋势。

3.2. 案例 2：参数估计的模拟可视化

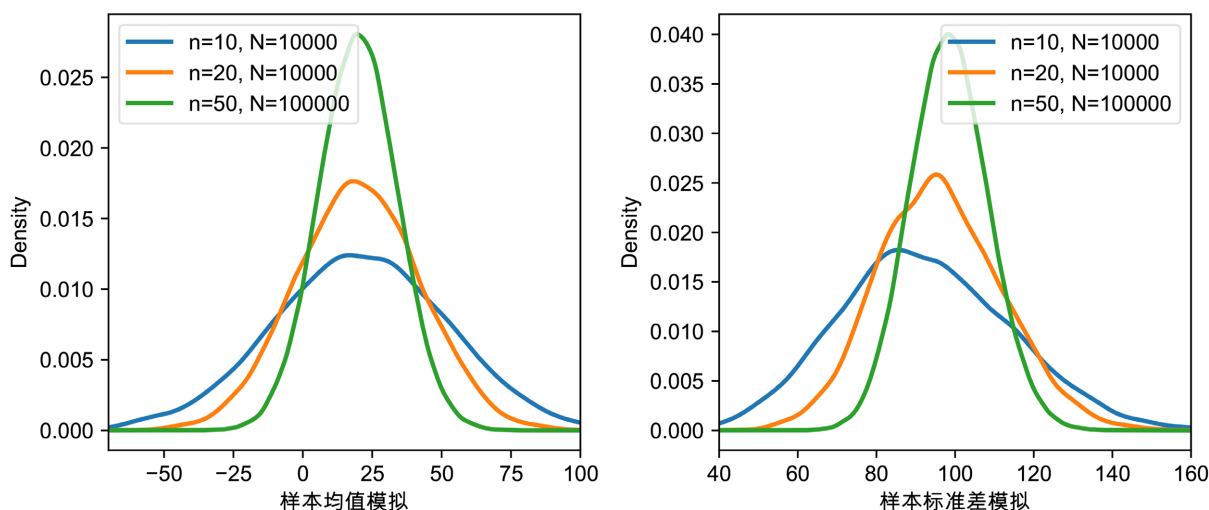


Figure 2. Parameter estimation simulation of sample mean and standard deviation

图2. 样本均值和标准差的参数估计模拟

问题：对比不同抽样的次数和样本数的模拟估计参数与真实值比较。利用编写 python 函数三次模拟一个均值为 20，标准差为 100 的正态分布。三次模拟的抽样次数和单次抽样样本个数分别为 $[N = 10,000, n = 10]$ ， $[N = 10,000, n = 20]$ ， $[N = 100,000, n = 100]$ 。通过两张图分别绘制三次模拟的样本均值和样本标准差的核密度图对比，说明抽样样本次数和个数对抽样结果的影响。

解答：第一步：编写参数模拟函数 simul_fun，函数的输入包含一下 4 个参数：n (每次抽样的样本数)，N (抽样的次数)，loc (抽样的正态分布的均值，默认值为 0) 和 scale (抽样的正态分布的标准差，默认值为 1)。函数返回值为两个列表：基于均值为 loc，标准差为 scale 的正态分布进行 N 次抽样的均值列表和标准差列表。

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
def simul_fun(n,N,loc=0,scale=1):
    x, v = [], []
    for i in range(N):
        d = np.random.normal(loc=loc,scale=scale,size=n)
        x.append(np.mean(d))
        v.append(np.std(d))
    return x, v
```

第二步：编写绘图函数 plot_dis，输入值 df: 给定的 DataFrame 数据；ax: 绘图的 axes 对象；xlabel: 图的 x 轴标签名。将 df 数据的直方图及其核密度函数绘制在 axes 对象中并设置 title。

```
def plot_dis(df, ax, xlabel):
    df.plot(bins=20, kind='hist', density=True, ax=ax, legend=False)
    df.plot(kind='density', linewidth=2, ax=ax, legend=False)
    ax.set_xlabel(xlabel); ax.set_title(xlabel + '模拟分布')
```

第三步：模拟实现

```
loc1, scale1 = 20, 100
n1, N1 = 10, 10000; n2, N2 = 20, 10000; n3, N3 = 50, 100000
x1, v1 = simul_fun(n1,N1,loc1,scale1)
x2, v2 = simul_fun(n2,N2,loc1,scale1)
x3, v3 = simul_fun(n3,N3,loc1,scale1)
plt.subplots(1, 2, figsize=(8, 3.5)); ax1 = plt.subplot(121)
pd.DataFrame(x1).plot(kind='density',linewidth=2,ax=ax1)
pd.DataFrame(x2).plot(kind='density',linewidth=2,ax=ax1)
pd.DataFrame(x3).plot(kind='density',linewidth=2,ax=ax1)
ax1.legend(['n='+str(n1)+'',N='+str(N1)',n='+str(n2)+'',N='+str(N2)',n='+str(n3)+'',N='+str(N3)])
ax1.set_xlabel("样本均值模拟"); ax1.set_xlim(-70, 100)
ax2 = plt.subplot(122)
pd.DataFrame(v1).plot(kind='density',linewidth=2,ax=ax2)
pd.DataFrame(v2).plot(kind='density',linewidth=2,ax=ax2)
```



```
pd.DataFrame(v3).plot(kind='density',linewidth=2,ax=ax2)
ax2.legend(['n='+str(n1)+'',N='+str(N1)',n='+str(n2)+'', N='+str(N2)',n='+str(n3)+'',N='+str(N3)'])
ax2.set_xlabel("样本标准差模拟"); ax2.set_xlim(40, 160)
plt.tight_layout(); plt.show()
```

图 2 展示了样本均值和标准差的参数估计模拟结果，通过图示可以看出，抽样的样本数和抽样的次数越多，参数估计结果越趋于准确值，同时估计参数的方差也越小，这可以加深同学们对参数估计理论的认识和理解。

3.3. 案例 3：卡方分布的概率密度函数可视化

问题：编写函数实现自由度为 df 的卡方分布的概率密度函数

$$f(x, df) = \frac{x^{\frac{df}{2}-1} e^{-\frac{x}{2}}}{\Gamma\left(\frac{df}{2}\right) 2^{df/2}}$$

其中 $\Gamma(x)$ 表示 Gamma 函数，其计算可调用 `scipy. special` 库中的 `gamma` 函数之间计算。结合 `matplotlib` 库和编写的概率密度函数对自由度 df 分别为 3, 8, 15, 30 的概率分布曲线进行可视化。对比不同自由度下，卡方分布的概率密度曲线的变化趋势。

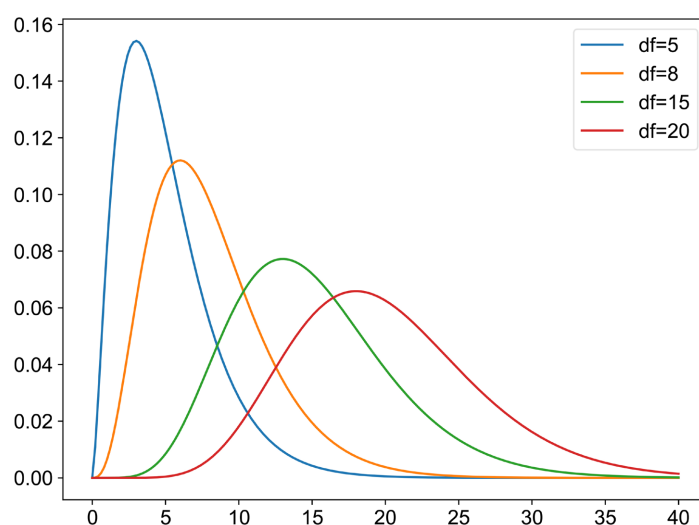


Figure 3. Probability density plot of the chi-square distribution with different degrees of freedom
图 3. 不同自由度下卡方分布的概率密度图

解答：第一步：编写函数计算卡方分布的概率密度函数，输入值 x 为需要计算概率密度的点， df 为卡方分布的自由度。输出值为在 x 点处自由度为 df 的卡方分布的概率密度。

```
from scipy.special import gamma
def chi2_fun(x, df):
    up = x**(df/2-1) * np.exp(-x/2)
    down = gamma(df/2) * 2**(df/2)
    return up/down
```

第二步：`matplotlib` 可视化概率密度图，确定绘图区间 $[0, 40]$ ，对区间等差划分为 200 个点，分别计

算每个点在不同自由度下卡方分布的概率密度值，利用 matplotlib.pyplot 进行可视化。

```
import matplotlib.pyplot as plt
n = 200; x = np.linspace(0,40,n)
plt.figure(figsize=(8,6))
plt.plot(x, chi2_fun(x,5)); plt.plot(x, chi2_fun(x,8)); plt.plot(x, chi2_fun(x,15)); plt.plot(x, chi2_fun(x,20))
plt.title("自由度为 df 的卡方分布的概率密度图")
plt.legend(["df=5","df=8","df=15","df=20"]); plt.show()
```

通过图 3 可以看出卡方分布的自由度越大，其概率密度函数的图像越趋于对称的，辅助同学们理解卡方分布中自由度的重要性。

4. 结束语

通过对教学实践的分析，Python 数据可视化在统计学教学中可以有效提高学生的学习兴趣 and 参与度。学生在上机实验中能够更加主动地使用 Python 进行数据分析，并展示分析结果。教学效果的评估表明，数据可视化不仅提升了学生的统计学知识理解，还增强了其编程技能。同时为了将 Python 数据可视化有效地融入统计学教学中，教师也需要更加合理的设计课程内容和教学方法。可以通过任务驱动学习，让学生在实际问题中应用数据可视化技术，同时结合 Python 编程技巧，提升学生的数据处理与分析能力。同时老师也应该提供丰富的教学资源，如教学示范代码、数据集、课后练习等，以帮助学生更好地掌握相关技能。综上 Python 的数据可视化工具为学生提供了更为直观和互动的学习体验，使其能够更好地理解统计学的概念和方法。未来，随着教学技术的发展，Python 数据可视化有望在更多领域的教学中发挥重要作用。

基金项目

西华大学校内人才引进项目(RX2400002043)。

参考文献

- [1] 丁毅涛,任水利. 基于 Python 语言的数理统计课程应用改革探索——以应用统计学专业为例[J]. 科技风, 2021(28): 25-27.
- [2] 何小年,段凤华. Python 在统计学课程实训中的应用探索[J]. 软件, 2021, 42(9): 32-34.
- [3] 郭鹏. Python 在应用统计学课程教学中的应用[J]. 现代信息科技, 2024, 8(14): 183-189.
- [4] 杜江,戴君,曹瑞元. 人工智能与数字技术背景下统计学专业实践教学体系优化及设计[J]. 高教学刊, 2024, 10(9): 115-118.
- [5] 李毕祥,李兰芳,朱琳. 标准化统计分析方法在 Python 教学中的应用与课程设计[J]. 中国标准化, 2024(10): 235-238.
- [6] 何小年,段凤华. 基于 Python 的线性回归案例分析[J]. 微型电脑应用, 2022, 38(11): 35-37.
- [7] 袁欧,何山. 基于 python 的 Z 检验法和 T 检验法研究[J]. 大众标准化, 2022(15): 174-176.
- [8] 茆诗松,吕晓玲. 数理统计学[M]. 第 2 版. 北京: 中国人民大学出版社, 2016.
- [9] 陈华. Python 在数据分析中的应用[J]. 上海轻工业, 2024(3): 108-110.
- [10] 贾俊平,吴翌琳. 统计学——基于 Python[M]. 北京: 中国人民大学出版社, 2024.
- [11] 王家武,姜庆. Python 在医学文献检索与统计学课程教学实践中的应用[J]. 现代医药卫生, 2023, 39(9): 1590-1595.
- [12] 胡国治,曾婕. 基于 R 软件的统计学专业课程教学研究与实践[J]. 许昌学院学报, 2023, 42(2): 21-26.