

课程名称：数据分析综合训练

一、课程编码：1700144

课程学时：32 学分：2 （其中，课堂学时：32 ，实践学时：0 ）

二、适用学科专业：应用统计专业、数据科学专业

三、先修课程：数学分析，高等代数，概率论，数理统计，Python 或 R 语言

四、教学目标

通过本课程的学习和训练，使学生掌握如何针对实际业务的需求提出和设立有价值的大数据分析项目，综合运用大数据分析的：探索性分析、可视化、大数据建模及优化方法，大数据分析软件等技术研究实际问题，提升研究生的数据分析报告的撰写能力、展示能力和语言表达能力。

五、教学方式

课堂讲授、讨论与分析

六、主要内容及学时分配

主要讲授与讨论以下内容，每一知识点约 2 学时。

1、数据分析的基本原则与实际数据介绍

2、探索性数据分析

2.1 数据可视化

2.2 数据的基本变换

2.3 数据关系的建立

3、试验设计与统计调查

3.1 数据的获取与随机性

3.2 试验设计与追加试验

3.3 观察性数据与伪试验

4、统计模型的建立

4.1 推算和建立关联

4.2 概率分布与量纲分析

4.3 最小二乘

4.4 其他的统计模型

4.5 模型验证与优化

5、当数据不成比例地增长时

5.1 数据的内推与外插

5.2 非同分布数据的统计议题

5.3 稳健统计分析

6、数据分析报告撰写

七、考核与成绩评定

成绩分为 2 部分，出勤与课堂表现（10%），课堂笔记与数据分析小型项目（90%）。其中课堂笔记需要学生以科学小品文的形式呈现，结合课程内容有理有据有实际例子的向普通本科生科普数据分析的宗旨与方法。

八、参考书及学生必读参考资料

参考书籍：

Philipp K. Janert 著，黄权，陆昌辉，邹雪梅，费柳凤译，数据之魅：基于开源工具的数据分析，清华大学出版社，2012

Hadley Wickham et al. R for Data Science, O'Reilly Media, 2023 (Chinese translation: <https://r4ds-zh.readthedocs.io/zh-cn/latest/>)

David Cox 著 周在莹 译, 试验设计, 清华大学出版社, 2024

Joshua D. Angrist and Jörn-Steffen Pischke 著, 郎金焕, 李井奎 译, 基本无害的计量经济学, 格致出版社, 2021

九、大纲撰写人：虞俊