## 附件3

## 《探索性数据分析》教学大纲

**一、基本信息**

|  |  |  |  |
| --- | --- | --- | --- |
| **课程编号**： | 2042001 | **课程性质**： | 专业必修课 |
| **课程名称**： | 探索性数据分析 | **英文名称**： | Exploratory Data Analysis |
| **学时/学分**： | 32学时/2.0学分 | **开课时间**： | 第2学期 |
| **适用对象**： | 应用统计硕士 | | |
| **大纲执笔人**： | 陈骥、王伟伟、方冠奇、汪懿君 | **当前版本**： |  |

**二、课程概述**

探索性数据分析主要围绕统计数据预处理问题，通过系统讲述数据整理、数据质量诊断和数据填补、模拟等关键问题，为数据建模和统计分析开展探索性应用。通过该课程的讲解，培养学生统计思维能力，提高研究生在数据的综合分析能力和实际应用能力，并能熟练运用JMP、SPSS软件分析和处理实际问题。

**三、教学目标**

主要实现以下目标：

1.掌握数据预处理的基本方法，为统计建模提供完整的数据基础。

2.提升应用统计硕士的统计数据敏感性，培养形成面向数据处理的研究范式。

3.能自主开展统计数据建模前的数据探索和相关软件工具的应用。

**四、课程内容**

**第1讲 探索性数据分析概述**

**教学内容：**

1.数据分析与统计数据

2.数据的基本分类

3.探索性数据分析的内容

4.探索性数据分析的应用介绍（思政模块：我国改革开放以来的经济增长驱动因素）

**第2讲 缺失数据分析**

**重点内容：**缺失数据形成的机制以及相应的处理方法。

教学内容：

1.缺失数据的产生机制

2.缺失数据的常用插补方法

3.缺失数据的高级插补方法

4.基于SPSS软件的缺失值分析（思政模块：宏观经济数据的插补方法）

**第3讲 异常数据的处理与分析**

**重点内容：**异常数据形成的机制以及相应的处理方法。

**教学内容：**

1.异常值的基本问题

2.回归模型中的异常值诊断统计量（思政模块：新冠肺炎疫情冲击中的旅游流量数据异常分析）

3.正态分布下的异常值诊断

4.极值分布下的异常值诊断

**第4讲 随机数的产生与检验**

重点内容：利用计算机产生各种不同分布的随机变量的随机样本。

教学内容：

1. 概论
2. 均匀随机数的产生
3. 均匀随机数的检验
4. 非均匀随机数的产生
5. 搜索材料：自然科学、工程相关领域的随机现象，并做好统计解读

**第5讲 数据模拟方法**

重点内容：利用R语言产生非常规分布的随机变量的数据模拟。

教学内容：

1. 逆变换方法
2. 合成法
3. 筛选抽样
4. 随机向量的抽样方法（思政模块：居民幸福感调查的模拟方法）

**第6讲 Bootstrap和Jacknife方法**

重点内容：介绍在总体分布未知的情况下，如何利用样本抽样。

教学内容：

1. Bootstrap方法
2. Jacknife方法
3. 思政模块：双碳战略背景下的冲击数据模拟

**第7讲 数据探索与检验**

重点内容：对数据进行初步探索与检验

教学内容：

1. 茎叶图
2. 箱线图
3. 残差分析
4. 思政模块：经济建设成就的图表描述

**第8讲 模型的评估**

重点内容：对候选模型的泛化误差进行评估

教学内容：

1. 经验误差与过拟合
2. 评估方法
3. 性能度量
4. 比较检验
5. 偏差与方差

6. 思政模块：空气治理成效的评估模型选择

**第9讲 线性模型**

重点内容：介绍几种经典的线性模型

教学内容：

1. 基本形式
2. 线性回归
3. 线性判别分析
4. 应用案例（思政模块：科技自立自强的影响因素分析）

**第10讲 分类与聚类**

重点内容：介绍机器学习中常用的分类与聚类方法

教学内容：

1. 决策树
2. 支撑向量机
3. 聚类
4. 应用案例（思政模块：纳税主体的社会信用类型判别）

**第11讲 课程评估（课余安排）**

重点内容：考察学生对知识点的掌握情况

教学内容：以小组为单位书面汇报。

要求：选题立足中国经济改革实践、科技强国、中国式现代化等背景下的，数据产业、数据应用等主题。

**五、教学安排**

该课程每周3学时，11周，32学时均为课堂授课教学时间。由于五一假期放假一次，实际授课为30次课时，建议教学进度如下：

|  |  |
| --- | --- |
| **章节** | **学时数** |
| 第1讲 探索性数据分析概述 | 3.0 |
| 第2讲 缺失数据分析 | 3.0 |
| 第3讲 异常数据的处理与分析 | 3.0 |
| 第4讲 随机数的产生与检验 | 3.0 |
| 第5讲 数据模拟 | 3.0 |
| 第6讲 Bootstrap和Jacknife方法 | 3.0 |
| 第7讲 数据探索与检验 | 3.0 |
| 第8讲 模型的评估与选择 | 3.0 |
| 第9讲 线性模型 | 3.0 |
| 第10讲 分类与聚类 | 3.0 |
| 第11讲 课程评估 | 课余安排，书面形式 |

**六、授课方式与手段**

教师课堂讲授，学生小组讨论完成课程论文。

**七、考核方式及成绩评定**

**考核方式**：期末考试闭卷，平时作业、出勤、课堂情况，实验作业。

**成绩评定标准**：总成绩（百分制）=40%\*平时成绩 +60%\*期末成绩

**八、教材及主要参考书**

|  |
| --- |
| **指定教材：** |
| Frederick Mosteller，John W. Tukey，David C. Hoaglin，探索性数据分析，中国统计出版社1998年 |
| **参考书目：** |
| Robert Carver.探索性数据分析:基于JMP软件，上海财经大学出版社，2013年 |
| 周志华. 机器学习.清华大学出版社，2016年 |