**《大数据分析中的统计方法与应用》教学大纲**

**（**Statistical Methods and Applications in Big Data Analysis**）**

1. **编写说明**

本课程为统计学专业硕士研究生专业选修课，2学分，共32课时。在本专业研究生教学计划中占有重要地位。

**（一）本课程的教学目的和要求**

本课程以统计理论为基础，深入阐释数理统计分析方法的理论体系，并以R软件为工具，结合实际的大数据案例对各分析方法开展实践教学。旨在通过本课程的学习，系统地掌握经典统计推断的原理和方法，并了解相关理论的最新的研究进展，以丰富学生的统计理论知识体系；熟练地运用R软件对大数据案例进行统计分析，以培养学生的科研兴趣以及理论学习与实践结合的能力。

大数据分析的学习起点是大数据的统计分析，而其学习特点则应兼具案例化和工具化。本课程以数理统计分析方法为理论基础，以R软件为分析工具，以大数据案例为线索，以解决问题为导向讲解统计方法及R语言的实现。突出大数据应用特色，兼顾统计方法的经典性和普适性、理论讲解的通俗性和严谨性、R语言代码的实操性和示范性。

学习本课程之前，要求已经选修过《高等数学》、《统计学原理》、《概率论》、《数理统计学》、《多元统计学》等课程。

**（二）大纲的教学体系**

本课程教学体系主要包括理论与实验两部分内容。理论授课内容以推断统计经典方法为核心，同时兼顾当前国内外最新研究动态，并实时进行调整。实验操作采用R语言，围绕教材案例数据，并适当引入相关的大数据，开展数据处理和案例分析。同时尝试让同学们自行查找大数据，并围绕相应的统计分析方法，开展大数据分析实践。

1. **教学大纲内容**

**第1章 R语言与统计分析概述**

**重点内容：** R的基本概念；基于R的统计分析基本框架

**难点内容：** R程序的运行；R的帮助系统

**教学内容：**

1.1 R中的基本概念

1.2 R的下载安装

1.3 R程序的运行

1.4 R帮助系统

1.5 基于R的统计分析基本框架

**第2章 R的数据组织**

**重点内容：**R数据的各种组织方式；R数据组织的读入

**难点内容：**R共享包的查找及调用；基于具体案例的R数据结构分析

**教学内容：**

2.1 R的数据对象

2.2 R的数据组织基本方式

2.3 R数据组织的保存、读入和R的共享数据包

2.4 基于具体案例的数据结构和R组织分析

**第3章 R的数据整理和编程基础**

**重点内容：**R的文件的读入和写出；R语言分支语句和循环语句；自定义R的函数

**难点内容：**R的分支与循环语句；R函数的编写

**教学内容：**

3.1 数据的初步整理

3.2 数据资料评估

3.3 数据加工

3.4 数据管理中的R编程基础

3.5 基于具体案例的R数据整理分析和基本编程实践

**第4章 R的基本分析和统计图形**

**重点内容：**R的图形设备和文件；R的图形组成和参数；分类型单变量的频数分布表和基本统计图形；数值型变量的基本描述统计量和统计图形

**难点内容：** R基本统计分析和绘图实践

**教学内容：**

4.1 R的绘图基础

4.2 分类型单变量的基本分析

4.3 数值型单变量的基本分析

4.4 基于具体案例的R基本统计分析和绘图

**第5章 R变量的相关性分析和统计图形**

**重点内容：**分类型变量的列联表及卡方统计量、等级相关系数、统计绘图函数（马赛克图）；数值型变量相关性的简单相关系数、统计绘图函数（散点图）

**难点内容：**分类型变量和数值型变量相关性的统计绘图实践

**教学内容：**

5.1 分类型变量的相关性描述及统计绘图

5.2 数值型变量的相关性描述及统计绘图

5.3 基于具体案例的相关性分析和统计绘图实践

**第6章 R的统计检验**

**重点内容：**单个总体的均值检验过程；两个总体的均值的*t*检验过程；两个配对样本的*t*检验过程；多因素方差分析的过程

**难点内容：**基于R的统计检验实践

**教学内容：**

6.1 单个总体的均值推断

6.2 两个总体的均值比较

6.3 方差分析

6.4 基于具体案例的R统计检验实践

**第7章 R的回归分析**

**重点内容：**线性回归的原理与R实现；Logistic回归的原理与R实现

**难点内容：**R回归分析的具体实践

**教学内容：**

7.1 回归方程的建立

7.2 回归方程的应用

7.3 回归方程的检验

7.4基于具体案例的R回归分析实践

**第8章 R的聚类分析**

**重点内容：**K-Means 聚类的原理和R实现；分层聚类的原理与R实现

**难点内容：**R聚类分析的具体实践

**教学内容：**

8.1 K-Means 聚类

8.2 分层聚类

8.3 基于具体案例的R聚类实践

**第9章 R的判别分析**

**重点内容：**距离判别的原理和R实现；Fisher判别的原理与R实现

**难点内容：**R判别分析的具体实践

**教学内容：**

9.1 距离判别

9.2 Fisher判别

9.3 基于具体案例的R判别实践

**第10章 R的因子分析**

**重点内容：**因子分析的原理和R实现

**难点内容：**R因子分析的具体实践

**教学内容：**

10.1 因子分析

10.2 基于具体案例的R因子分析

1. **考核方式及成绩评定标准**

**考核方式**：期末大作业。

**成绩评定标准：**平时课堂表现(30%) + 期末大作业(70%)

1. **教材及主要参考书**

[1]薛薇著.大数据分析中的统计方法及应用[M].电子工业出版社,2018.

[2]薛毅,陈立萍编著.R语言在统计中的应用[M].人民邮电出版社,2017.

[3]吕书龙,梁飞豹.应用统计分析与R语言实战[M].北京大学出版社,2017.

[4]Paul Teetor著,李洪成,朱文佳,沈毅诚译.R语言经典实例[M].机械工业出版社,2013.

[5]薛毅,陈立萍编著.统计建模与R软件[M].清华大学出版社,2007.

**执笔人：曾慧**