数据分析统计学基础

1. 统计学概述

1.1 什么是统计学

定义：统计学是一门研究如何**收集、整理、分析、解释**数据的科学，通过数学方法揭示数据背后的规律和模式。

核心价值：

- 从数据中提取有价值的信息
- 支持数据驱动的决策制定
- 建立预测模型和验证假设

1.2 统计学在数据分析中的应用

应用领域	具体应用
描述性分析	数据摘要、可视化、基本统计量
诊断性分析	相关性分析、趋势分析、原因探究
预测性分析	回归分析、时间序列预测、机器学习
规范性分析	优化建议、决策支持、方案评估

2. 数据类型与测量尺度

2.1 数据分类

类型	特点	示例	分析方法
定量数据 (数值型)	可测量，有数值意义	身高、温度、销售额	均值、标准差、回归
定性数据 (类别型)	表示类别或属性	性别、品牌、颜色	频数、比例、卡方检验

2.2 测量尺度

尺度类型	特点	示例	允许的运算
定类尺度	分类，无顺序	性别、颜色	=, ≠
定序尺度	分类，有顺序	满意度评级	=, ≠, >, <
定距尺度	数值，无绝对零点	温度、日期	=, ≠, >, <, +, -
定比尺度	数值，有绝对零点	重量、收入	=, ≠, >, <, +, -, ×, ÷

3. 描述性统计

3.1 集中趋势度量

均值 (Mean)：

- 算术平均值：$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- 适用场景：数据分布均匀，无极端异常值

中位数 (Median)：

- 排序后中间位置的值
- 适用场景：数据有偏或有异常值

众数 (Mode)：

- 出现频率最高的值
- 适用场景：分类数据或寻找典型值

3.2 离散程度度量

方差与标准差：

- 方差：$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- 标准差：$S = \sqrt{S^2}$（与原始数据同单位）

四分位距 (IQR)：

- Q3(75%分位数) - Q1(25%分位数)
- 抗异常值干扰能力强

极差 (Range)：

- 最大值 - 最小值
- 对异常值敏感

3.3 分布形态

偏度 (Skewness)：

- 右偏(正偏)：均值 > 中位数，长尾在右侧
- 左偏(负偏)：均值 < 中位数，长尾在左侧

峰度 (Kurtosis)：

- 高峰度：分布更陡峭，尾部更重
- 低峰度：分布更平缓，尾部更轻

---

4. 概率与概率分布

4.1 基本概念

- **概率**：事件发生的可能性，0≤P(A)≤1
- **条件概率**：P(A|B) = P(A∩B)/P(B)
- **贝叶斯定理**：$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

4.2 常见概率分布

离散分布：

- **二项分布**：n次独立伯努利试验的成功次数
- **泊松分布**：单位时间/空间内随机事件发生次数

连续分布：

- **正态分布**：$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- **t分布**：小样本情况下的均值分布
- **χ^2分布**：方差分布和独立性检验
- **F分布**：方差分析中的分布

4.3 中心极限定理

核心思想：无论总体分布如何，样本均值的抽样分布随样本量增大而趋近正态分布

应用价值：

- justify 使用正态分布进行推断
- 支持大样本下的假设检验

---

5. 推断统计

5.1 抽样与估计

抽样方法：

- 简单随机抽样
- 分层抽样
- 系统抽样
- 整群抽样

点估计与区间估计：

- 点估计：用单个值估计参数（如样本均值）
- 区间估计：给出参数可能范围（置信区间）

5.2 假设检验

基本步骤：

1. 设立假设：H_0(原假设) vs H_1(备择假设)
2. 选择检验统计量
3. 确定显著性水平α(通常0.05)
4. 计算p值或比较临界值
5. 做出统计决策

常见检验方法：

检验类型	适用场景	检验统计量
Z检验	大样本均值检验	$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
t检验	小样本均值检验	$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$
卡方检验	分类变量独立性	$\chi^2 = \sum \frac{(O-E)^2}{E}$
方差分析	多组均值比较	F = 组间方差/组内方差

第一类错误与第二类错误：

- α错误：拒绝真H_0（假阳性）
- β错误：接受假H_0（假阴性）
- 功效(1-β)：正确拒绝H_0的概率

5.3 相关与回归

相关系数：

- Pearson相关系数：线性相关程度，-1到1
- Spearman秩相关：单调相关程度

简单线性回归：

- 模型：$y = \beta_0 + \beta_1 x + \epsilon$
- 最小二乘法估计参数

多元线性回归：

- 模型：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$
- 考虑多个预测变量

模型评估：

- R^2：解释的变异比例
- 调整R^2：考虑变量个数后的R^2
- MSE/RMSE：预测误差大小
