“智”汇学习角——概率论与数理统计

数理统计部分

汤纪洋

南开大学人工智能学院

2024 年 6 月 6 日

怎样学好概统？

- 对于一门数学类课程来说，掌握基本的知识点和方法，构建完整的知识体系，是最基础最重要的。
- 在此基础上，需要通过教材练习，加上一定量的课外资料，在练习中不断强化对知识点和方法的理解，对题型和解题技巧进行一定的总结，可以起到很大的提升作用。
- 某种意义上来说，数学类课程（高数、线代、概统）的学习方式具有很大的相似性，可以相互借鉴影响。尤其对于概统而言，非常依赖高数基础，需要注意。

1 数理统计的基本概念

2 矩估计与极大似然估计

3 区间估计与假设检验

样本的相关概念

- 总体、个体、样本
- 简单随机样本及其性质：独立同分布

  简单随机样本的分布函数公式：

$$F_{总}(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} F(x_i)$$

  简单随机样本的概率密度函数公式：

$$f_{总}(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i)$$

统计量

- 概念：样本的不含未知参数的实值连续函数 g
  若样本 $(X_1, X_2, \dots, X_n)$ 取值 $(x_1, x_2, \dots, x_n)$，则
  $g(X_1, X_2, \dots, X_n)$ 为统计量，$g(x_1, x_2, \dots, x_n)$ 为样本值。

统计量

- 概念：样本的不含未知参数的实值连续函数 g
  若样本 $(X_1, X_2, ..., X_n)$ 取值 $(x_1, x_2, ..., x_n)$，则
  $g(X_1, X_2, ..., X_n)$ 为统计量，$g(x_1, x_2, ..., x_n)$ 为样本值。
- 常用统计量：
  (1) 样本均值：

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  (2) 样本方差：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  注意和样本二阶中心矩 S_n^2 的关系：$S^2 = \frac{n}{n-1}S_n^2$
  (3) 样本的 k 阶原点矩 A_k，k 阶中心矩 B_k

抽样分布

- 数理统计常用分布：正态分布、卡方分布、t 分布、F 分布：需要了解概念、定义式和性质 (如数字特征)，对具体分布函数和概率密度函数的表达式要求不高，重点在于这几种分布在正态总体中的应用！

抽样分布

- 数理统计常用分布：正态分布、卡方分布、t 分布、F 分布：需要了解概念、定义式和性质 (如数字特征)，对具体分布函数和概率密度函数的表达式要求不高，重点在于这几种分布在正态总体中的应用！
- 概念 & 定义式：判断分布，根据分布类型求参数
- 性质：和数字特征、切比雪夫不等式、中心极限定理等结合
- (重点)正态总体中的应用：
  若 $X \sim N(\mu, \sigma^2)$，则
  从样本均值出发：
  $$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$
  从样本方差出发：
  $$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

矩估计

- 牢记“三部曲”:
  (1) 取总体矩量 $\mu_1, \mu_2, ...\mu_k$
  (2) 反解待估参数 $\theta_1, \theta_2, ...\theta_k$(用 $\mu_1, \mu_2, ...\mu_k$ 表示)
  (3) 用样本矩 A_k 代替总体矩 μ_k，得到估计值 $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$

矩估计

- 牢记“三部曲”：
  (1) 取总体矩量 $\mu_1, \mu_2, ...\mu_k$
  (2) 反解待估参数 $\theta_1, \theta_2, ...\theta_k$(用 $\mu_1, \mu_2, ...\mu_k$ 表示)
  (3) 用样本矩 A_k 代替总体矩 μ_k，得到估计值 $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$
- 注意：
  (1) 总体矩用期望的一般公式求，样本矩用定义。
  (2) $\mu_1, \mu_2, ..., \mu_k$ 并不一定是一阶、二阶直到 k 阶矩。有的矩量无法获取有效信息，需要跳过。
  (3) 实际做题时，不一定显式地体现“三部曲”，经常列 $\mu_k = \mathbb{E}(X^k) = A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 解出待估参数。

极大似然估计

- 牢记“两步走”：
  (1) 求似然函数 (样本取观测值的概率)：
  连续用概率密度：

$$L(\theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, ..., \theta_k)$$

  离散用分布律：

$$L(\theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} P(X_i = x_i; \theta_1, \theta_2, ..., \theta_k)$$

  (2) 令其最 (极) 大
  一般使用对数似然，需注意不可导情形

矩估计与极大似然估计

(“概率论与数理统计”结课考试)

已知随机变量 X 的密度函数为

$$f(x)=\begin{cases}(\theta+1)(x-5)^{\theta},5<x<6 \\ 0,\text{其他}\end{cases}\quad(\theta>0)$$

其中 θ 为未知参数，$X_1,X_2,...X_n$ 为来自总体的样本，$x_1,x_2,...x_n$ 为对应的样本值，求 θ 的矩估计量与极大似然估计量。

(“概率论与数理统计”结课考试)

设 $X_1,X_2,...X_n$ 是来自总体 X 的一个样本，总体 X 服从参数为 λ 的泊松分布。此样本的一组观察值为 1,2,...,100，试求未知参数 λ 的极大似然估计值。

1. 数理统计的基本概念
2. 矩估计与极大似然估计
3. 区间估计与假设检验

区间估计与假设检验

- 注意两者的对应

  (1) 取枢轴量——取统计量

  (2) 考察分布 (常见统计分布)

  (3) 根据分位数确定范围 (置信区间——拒绝域)

  (4) 注意两者的区别 (已知未知，主观性)
- 利用分位数时注意：有没有 $\frac{1}{2}$(单双侧)，自由度是 $n/n-1$(有没有未知)
- 常见结论 (置信区间、拒绝域) 分类掌握，务必牢记!

假设检验

(“概率论与数理统计”结课考试)

某台机器加工某种零件，规定零件长度为 100cm ，标准差不超过 2cm ，每天定时检查机器运行情况，某日抽取 10 个零件，测得平均长度 $\overline{x} = 101$cm，样本标准差 $s = 2$cm，设加工的零件长度服从正态分布，问该日机器工作是否正常？($\alpha = 0.05$)

假设检验

(“概率论与数理统计”结课考试)

某药厂生产一种新的止痛药，厂方希望验证服用新药片后至开始起作用的时间间隔较原有止痛片至少缩短半，因此厂方提出需检验假设
$H_0 : \mu_1 = 2\mu_2$; $H_1 : \mu_1 \neq 2\mu_2$。
此处 μ_1,μ_2 分别是服用原有止痛片和服用新止痛片后至起作用的时间间隔的总体的均值。设两总体均为正态分布，且方差分别为已知值 σ_1^2,σ_2^2。现分别在两总体中取一样本 $X_1, X_2, \dots X_n$ 和 $Y_1, Y_2, \dots Y_m$，设两个样本独立。试给出上述假设检验的拒绝域，取显著性水平为 α。

假设检验

(“概率论与数理统计”结课考试)

市级地标建筑国际俱乐部为了要大修而重新测量。天大建筑学院的 6 名同学对该大厦的高度进行测量结果如下 (单位: 米)

87.4 87.0 86.9 86.8 87.5 87.0

据记载该大厦的高度为 87.4。设大厦的高度服从正态分布，问在检验水平 $\alpha = 0.1$ 下

(1) 你认为该大厦的高度是否要修改? (要写出计算过程)

(2) 若测量的方差不得超过 0.04，那么你是否认为这次测量的方差偏大?(要写出计算过程)

$t_{0.005}(5) = 4.0322$, $t_{0.005}(6) = 3.7074$, $t_{0.01}(5) = 3.3649$, $t_{0.01}(6) = 3.1427$

$\chi^2_{0.005}(5) = 16.750$, $\chi^2_{0.005}(6) = 18.548$, $\chi^2_{0.01}(5) = 15.086$, $\chi^2_{0.01}(6) = 16.812$

Thanks!