Probability & Statistics Review

21 CST H3Art

Chapter 0: Introduction to probability & statistics

- 课程介绍

Chapter 1: Overview and Descriptive Statistics

- 绘图
  - 茎叶图
    - 注意标注Stem和Leaf
  - 点图
    - 横着的数轴，重复的数值点向上堆积
  - 直方图
    - 类别宽度（数据范围）× 矩形高度（数据密度）= 该类别的相对频率
    - 典型样式
      - 对称单峰
      - 双峰
      - 正向偏斜
      - 负向偏斜
  - 盒图
    - 四分点——中值、下四分位、上四分位
    - 四价差fs——上四分位 - 下四分位
    - 离群值：距离最近的第四个值超过 1.5 fs 的任何观察值都是离群值
    - 极端：如果距最近的第四个值超过 3 fs，则异常值为极端值
    - 轻度：如果离群值在最接近的四分之一的 (1.5 fs, 3 fs] 范围内，则异常值是轻度的
- 样本均值（mean）
- 样本中位数（median）
- 裁剪均值（trimmed mean）
  - 对最大和最小值各裁切给定百分比的值
- 样本方差（sample variance）
  - 除以n-1
  - $s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1}$
  - $s^2 = \dfrac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$
- 总体方差（population variance）
  - 除以n

Chapter 2: Probability

- 实验与事件
- 集合论（交、并、补）

Chapter 3: Discrete RV

- **主要分布类型：**
  - **二项分布**
    - 在n次试验中成功数量x的分布（有放回抽样）
    - pmf: $b(x;n,p)=\binom{n}{x}p^x(1-p)^{n-x}$
    - cdf: $B(x;n,p)=\sum_{i=0}^{x}b(i;n,p)$（可查表 Table A.1）
    - 期望：$E(X)=np$
    - 方差：$V(X)=np(1-p)=npq, q=1-p$
  - **几何分布**
    - 在试验中取得1次成功时所需的试验数量x的分布（有放回抽样）
    - pmf: $g(x;p)=(1-p)^{x-1}p$
    - cdf: $G(x;p)=1-(1-p)^x$
    - 期望：$E(X)=\frac{1}{p}$
    - 方差：$V(X)=\frac{1-p}{p^2}$
  - **超几何分布**
    - 一次性抽出n个样本，已知总体数和总成功数，求抽出的n个样本中成功数量x的分布（不放回抽样）
    - pmf: $h(x;n,M,N)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$
    - cdf: $H(x;n,M,N)=\sum_{i=0}^{x}h(i;n,M,N)$
    - 期望：$E(X)=np, p=\frac{M}{N}$
    - 方差：$V(X)=\left(\frac{N-n}{N-1}\right)np(1-p), p=\frac{M}{N}$
  - **负二项分布**
    - 在试验中取得r次成功时所需的失败次数x的分布（有放回抽样）
    - pmf: $nb(x;r,p)=\binom{x+r-1}{r-1}p^r(1-p)^x$
    - cdf: $nB(x;r,p)=\sum_{i=0}^{x}nb(i;r,p)$
    - 期望：$E(X)=\frac{r(1-p)}{p}$
    - 方差：$V(X)=\frac{r(1-p)}{p^2}$
  - **泊松分布**
    - 单位时间内随机事件发生的次数x的概率分布
    - pmf: $p(x,\lambda)=\frac{e^{-\lambda}\lambda^x}{x!}, x=0,1,2,3,\ldots$
    - cdf: $P(x,\lambda)=\sum_{i=0}^{x}p(i,\lambda)$（可查表 Table A.2）
    - 期望：$E(X)=\lambda$
    - 方差：$V(X)=\lambda$

Chapter 4: Continuous RV

- **主要分布类型：**

- **正态分布/高斯分布**
  - pdf: $f(x;\mu,\sigma)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\quad(-\infty<x<\infty)$
  - 标准化pdf: $f(z;0,1)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}, Z\sim N(0,1)$
  - 标准化cdf: $\Phi(z)=\int_{-\infty}^{z}f(t)\mathrm{d}t=\int_{-\infty}^{z}\dfrac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}\mathrm{d}t$（可查表 Table A.3）
  - 期望：$E(X)=\mu$
  - 方差：$V(X)=\sigma^2$
- **伽马分布**
  - 伽马函数: $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x$
  - pdf:

$$f(x;\alpha,\beta)=\begin{cases}\frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}} & x\geq 0\\ 0 & \text{otherwise}\end{cases}$$

  - 让$\beta=1$，获得标准化伽马分布，pdf:

$$f(\frac{x}{\beta};\alpha)=\begin{cases}\frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & x\geq 0\\ 0 & \text{otherwise}\end{cases}$$

  - 标准化cdf:

$$F(\frac{x}{\beta};\alpha)=\begin{cases}\int_0^x\frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\mathrm{d}y & x\geq 0\\ 0 & x<0\end{cases}$$

    （可查表 Table A.4）
  - 期望：$E(X)=\mu=\alpha\beta$
  - 方差：$V(X)=\sigma^2=\alpha\beta^2$
- **指数分布**
  - 令$\alpha=1,\beta=\dfrac{1}{\lambda}(\lambda>0)$，伽马分布变成了指数分布
  - pdf:

$$f(x;\lambda)=\begin{cases}\lambda e^{-\lambda x} & x\geq 0\\ 0 & \text{otherwise}\end{cases}$$

  - cdf:

$$F(x;\lambda)=\begin{cases}0 & x<0\\ 1-e^{-\lambda x} & x\geq 0\end{cases}$$

  - 期望：$E(X)=\dfrac{1}{\lambda}$
  - 方差：$V(X)=\dfrac{1}{\lambda^2}$
- **韦布尔分布（仅了解）**
- **卡方分布（仅了解）**
- **柯西分布（仅了解）**
- **对数分布（仅了解）**
- **贝塔分布（仅了解）**

Chapter 5: Joint Probability Distributions and Random Sample

- 条件分布
  - $f_{Y|X}(y|x)=\dfrac{f(x,y)}{f_X(x)}$ or $f_{Y|X}(y|x)=\dfrac{p(x,y)}{p_X(x)}$
- 期望
- 协方差

    - $\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_X)]$
    - $\text{Cov}(X,Y) = E(XY) - \mu_X \cdot \mu_Y = E(XY) - E(X)E(Y)$
- 相关性/标准化协方差
    - $\text{Corr}(X,Y) = \rho_{X,Y} = \rho = \dfrac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$
        - $\text{Corr}(aX+b, cY+d) = \text{Corr}(X,Y)$
        - 对于任何二元随机变量的X和Y，$-1 \le \text{Corr}(X,Y) \le 1.$
- 中心极限定理（CLT）
    - $\bar{X}$ 是样本均值
        - $E(\bar{X}) = \mu_{\bar{X}} = \mu$
        - $V(\bar{X}) = \sigma^2_{\bar{X}} = \sigma^2/n$
        - $\sigma_{\bar{X}} = \sigma/\sqrt{n}$
    - 统计量T_0 是 $T_0 = X_1 + X_2 + \cdots + X_n$
        - $E(T_0) = \mu_{T_0} = n\mu$
        - $V(T_0) = \sigma^2_{T_0} = n\sigma^2$
    - n越大时候近似效果越好，通常当$n > 30$的时候，可以使用中心极限定理(CLT)
- 分布的线性组合
    - $E(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_i$
    - 如果$X_1, X_2, \ldots, X_n$是独立的，$V(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 V(X_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$
    - 对任意$X_1, X_2, \ldots, X_n$, $V(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j)$
    - $E(X_1 - X_2) = E(X_1) - E(X_2)$
    - 如果X_1和X_2是独立的，$V(X_1 - X_2) = V(X_1) + V(X_2)$

Chapter 6: Point Estimation

- 点估计
    - 原则
        - 无偏估计
        - 方差最小
        - 最小方差无偏估计量（MVUE）
    - 方法
        - 矩估计
        - 极大似然估计

Chapter 7: Statistical Intervals Based on A Single Sample

- 置信区间
- t分布（查表 Table A.5）