数理统计

数理统计是以概率论为基础，研究社会和自然界中大量随机现象数量变化基本规律的一种方法。其主要内容有参数估计、假设检验、相关分析、试验设计、非参数统计、过程统计等。

- **总体与样本**

> 总体：实验的全部个体集合
>
> 样本：从总体中随机抽取n个个体，**记录个体指标值**$X_1, X_2, X_3 \ldots$，这些指标值称为总体的样本，样本指标之间具有独立性
>
> - **简单随机样本联合分布函数：**
>
> $$F(x_1, x_2, x_3 \ldots) = \prod_{i=1}^{n} F(x_i)$$

- **统计量及其分布**

> 统计量：通过样本反映总体的各种特征，通过样本的**随机变量和已知参数构造的函数为统计量的函数**
>
> 抽样分布：统计量的分布函数
>
> - **经验分布函数**（样本值排序）
>
> $$F_n(x) = \begin{cases} 0 & x < x_{(1)}, \\ \frac{k}{n} & x_{(k)} \le x < x_{(k+1)}, \\ 1, & x \ge x_{(n)}, \end{cases}$$
>
> - **常见统计量**
>
> 1. 样本均值 $\overline{X}$
>
> > - 样本均值方差$D(\overline{x})$ = $\frac{D(X)}{n}$
> >
> > - 总体分布为正态分布N(u,σ^2)，**抽样样本** $\overline{X}$**也服从正态分布N(u,** $\frac{\sigma^2}{n}$**)**
> >
> > > *当总体分布未知时，其E(X)=u，D(X)=*σ^2*，且样本容量较大时，其样本*$\overline{X}$*接近正态分布N(u,* $\frac{\sigma^2}{n}$*)*
>
> 2. 样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *方差减去其数学期望*
> 3. 样本k阶原点矩 $A_K = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
> 4. 样本k阶中心距 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^K$
> 5. 极大/小顺序统计量
> 6. **总体服从正态分布的抽样分布**
>
> > - 卡方分布（其样本相互独立且服从标准正态分布N(0,1)，则$\chi^2 = X_1^2 + X_2^2 + X_3^2 \ldots + X_n^2$的构成的函数为自由度n的$\chi^2$分布，记为 χ^2~ $\chi^2(n)$， **E(χ^2)=n，D(χ^2)=2n**
> >
> > $$\text{卡方分布性质：} P\{\chi^2 > \chi_a^2(n)\} = a \quad 0 < a < 1$$
> > $$\chi_a^2(n)\text{表示自由度为}n\text{的}\chi^2\text{分布的}a\text{分位数}$$
> >
> > - F分布（存在 $X_1 - \chi^2(m)$与$X_2 - \chi^2(n)$，且X_1与X_2相互独立，则存在函数$F = \frac{X_1/m}{X_2/n}$ 为自由度m与n的F分布），记为 F~F(m,n)
> >
> > $$F\text{分布性质：} P\{F > F_a(m,n)\} = a$$
> >
> > - t分布（随机变量X_1与X_2 独立，且 $X_1 - N(0,1)$，$X_2 - \chi^2(n)$，存在函数T=$\frac{X_1}{\sqrt{X_2/n}}$ 为自由度n的t分布，记为T~t(n)
> >
> > 当$n > 1$时，t分布数学期望为0，当$n > 2$时，t分布方差为$n/(n-2)$，当$t > 30$时，t分布可以用正态分布近似$N(0,1)$

- **参数估计**
  - **距法估计**(使用样本代替总体)
    1. 总体的数学期望E(X)等于样本均值$\overline{X}$　E(X) = $\overline{X}$
  - **极大似然估计**（最有可能概率）
    1. 设似然函数 $L(\theta)=\prod_{i=1}^{n} f(x_i,\theta)$或表达 $\prod_{i=1}^{n} P_i(\theta)$，$f(x_i,\theta)$为密度函数
    2. 连边求对数 $ln\ L=\sum_{i=1}^{n} lnf(x_i,\theta)$
    3. 两边求导令结果为0，在求出参数与均值关系
  - **点估计评价标准**（距法估计和极大似然估计）

    无偏估计：系数之和为1，为无偏估计

    有效性：系数方差最小（系数相同方差最小）

    置信区间（类似于标准差）： a：显著性水平，1-a：置信度
    1. **总体标准方差σ 已知**，求置信区间u

       $$u_{=}\frac{\overline{x}-\sigma}{a/\sqrt{n}},\quad 置信区间[\overline{x}-u_{a/2}\frac{\sigma}{\sqrt{n}},\ \overline{x}+u_{a/2}\frac{\sigma}{\sqrt{n}}]$$
    2. **总体标准方差σ 未知**，求u的置信区间，s为样本方差

       $$置信区间：[\overline{x}-t_{a/2}(n-1)\frac{s}{\sqrt{n}},\ \overline{x}+t_{a/2}(n-1)\frac{s}{\sqrt{n}}]$$
    3. 求总体方差σ^2的置信区间

       $$置信区间：[\frac{(n-1)s^2}{X^2u/2(n-1)},\ \frac{(n-1)s^2}{X^2（1-u/2)(n-1)}]$$
- **假设估计**
  - 拒绝域：置信区间外的区域
  - 两类错误
    1. 第一类错误：**在原假设成立情况下**，样本落在拒绝域W中，因而原假设被拒绝，犯第一类错误概率为a （假设总体合格，抽样后存在不合格样本，原假设被拒）
    2. 第二类错误：**在原假设不成立情况下**，样本落在置信区间中，因而原假设被接受，犯第二类错误概率为b（假设总体不合格（否命题），抽样后存在合格样本，原假设接受）