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：实验的全部个体集合
>
> 参数空间（θ）：随机变量X的概率范围
>
> 样本（n）：总体的任何子集合
>
> 简单随机抽样：随机（个体抽中概率相等），独立性（样本变量相互独立），同分布（总体与样本分布性相同）
>
> $$\text{总体分布函数与样本分布函数之间关系}$$
> $$F(X_1,X_2,\dots X_n)=\prod_{i=1}^{n}F(X_i)$$
>
> **统计量：样本变量X构成的函数集合**（不含任何未知参数）
>
> 
>
>
> **常见的统计量**
>
> 1. 样本均值（数学期望）
> 2. 样本标准差 S，样本方差S^2
>
> $$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{x})^2=D(\overline{x})=\frac{D(X)}{n}$$
>
> 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$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^K$
>
> **统计量与抽样分布**：
>
> - 样本函数$T=T(x_1,x_2,\dots x_n)$中不含有任何未知参数，则**T为统计量，统计量的分布函数为抽样分布**
>
> 样本均值与抽样分布
>
> 1. 总体为正态分布$X-N(u,\sigma^2)$，其样本均值$\overline{x}$也为正态分布$X-N(u,\frac{\sigma^2}{n})$
> 2. 总体分布不明确，当样本容量n较大时，其样本分布接近与正态分布（中心极限定理）

- ## 常见的抽样分布函数

  > - 卡方分布：随机变量$X_1,X_2\dots X_n$相互独立，且满足N(0,1)标准分布，则$\chi^2=X_1^2+X_2^2+X_3^2+\dots+X_n^2$是自由度为n的卡方分布，**EX =n，DX=2n**
  >   1. X，Y相互独立，且为卡方分布，X，Y构成的卡方分布$(X+Y)-\chi^2(m+n)$
  >   2. 上α分位数：随机变量$\chi_\alpha^2(n)$右侧的概率为 a

$$P\{\chi^2 > \chi^2_\alpha(n)\} = a$$

$$n：自由度，\alpha：该点右侧的概率，\chi^2_\alpha(n)：上位a分位数$$

- t分布：X~N(0,1)，Y~$\chi^2(n)$，且X，Y独立，则$\frac{X}{\sqrt{Y/n}} - t(n)$
  1. t分布关于y轴对称，存在 $t_{1-a}(n) = -t_a(n)$
  2. 当n>2时，t分布方差为n/(n-2)
  3. 当n>1时，t分布数学期望为0
  4. 当t>30时，t分布可以用正态分布近似N(0,1)

- F分布：$X_1 - \chi^2(m)$，$X_2 - \chi^2(n)$，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)
  1. $F - F(m,n)$，　$\frac{1}{F} = F(n,m)$
  2. $F_{1-a}(n_1, n_2) = 1/F_a(n_2, n_1)$

正态总体下的抽样分布

1. 样本均值$\overline{x}$也为正态分布$X - N(u, \frac{\sigma^2}{n})$
2. $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \chi^2(n-1)$
3. $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - u)^2 - \chi^2(n)$
4. $\frac{\overline{x}-u}{S}\sqrt{n} - t(n-1)$

- **参数估计**
  - 点估计
  - 区间估计
  - **矩估计** (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** （选择概率最大的事件作为总体）
    > 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）