数理统计

点估计

- 最大似然估计
  - 基本原理：运用实验推测数据
  - 计算实验结果的概率，概率越大假设的参数越可能是真的
- 点估计的评价
  - 相合性
    - 当拥有足够多的样本时，被估计的参数可以精确到指定的精度，即依概率收敛
      - 大样本
  - 无偏性
    - 对于小样本精确度的评价
      - 小样本
  - 有效性
    - 无偏估计的方差：评估波动
  - 均方误差
    - 方差+偏差的平方

贝叶斯估计

- 信息种类
  - 总体信息：总体分布提供的信息
  - 样本信息：抽取样本时观测到的信息
  - 先验信息：抽样前得到的信息，一般来自于经验
- 贝叶斯学派观点
  - 任一未知量都可以看作随机变量，可以用一个概率分布去描述，这个分布称为先验分布
- 贝叶斯估计步骤
  - 根据参数的先验信息得到先验分布
  - 从贝叶斯学派的观点看，样本的产生要分为两个步骤
    - 从先验分布中产生一个样本，设想得到
    - $p(X|\theta)$
      得到一组样本，此时样本的条件概率函数综合了总体信息及样本信息
  - $h(X,\theta)=p(\theta|X)\pi(\theta)$
    由于第一个样本是设想来的，为把先验信息综合进去，因此要用先验分布进行综合
    - 此时，把总体信息，样本信息，先验信息都综合考虑了
  - 子主题 4

区间估计

- 理解
  - 一个区间能够覆盖住参数的概率(置信水平)
- 轴度量法
  - 常用来构造置信区间的方法
- μ
  单个正态总体的置信区间
  - 方差已知
  - 方差未知
  - 大样本
- $\mu_1-\mu_2$
  两个正态总体的置信区间
  - 方差已知
  - 方差未知
  - 方差之比已知
  - 大样本
- σ_1^2/σ_2^2
  两个正态总体的置信区间
  - F分布

假设检验

- 假设检验步骤
  - 建立假设
  - 选择检验统计量，给出拒绝域形式
  - 选择显著性水平
    - 第一类错误：拒真错误/假阳性
      发生第一类错误的概率被称为显著水平
    - 第二类错误：存伪错误/假阴性
    - 势函数：
      样本观测值落在拒绝域内的概率
  - 给出拒绝域
    - p值：当零假设成立时，观测到样本数据出现零假设事件的概率
    - 置信区间
- 正态总体均值的假设检验
  - 单个正态总体

σ	检验方式	统计量	H_0	拒绝域
已知	u检验	$u=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$\mu\le\mu_0$	$W=\{u\ge u_{1-\alpha}\}$
			$\mu\ge\mu_0$	$W=\{u\le u_{-\alpha}\}$
			$\mu=\mu_0$	$W=\{\lvert u\rvert\ge u_{1-\alpha/2}\}$
未知	t检验	$t=\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}$	$\mu\le\mu_0$	$W=\{t\ge t_{1-\alpha}(n-1)\}$
			$\mu\ge\mu_0$	$W=\{t\le t_{-\alpha}(n-1)\}$
			$\mu=\mu_0$	$W=\{\lvert t\rvert\ge t_{1-\alpha/2}(n-1)\}$

  - 两个正态总体

检验法	条件	原假设 H_0	备择假设 H_1	检验统计量	拒绝域
u 检验	σ_1,σ_2 已知	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$u=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$	$\{u\geqslant u_{1-\alpha}\}$ $\{u\leqslant u_\alpha\}$ $\{\lvert u\rvert\geqslant u_{1-\alpha/2}\}$
t 检验	σ_1,σ_2 未知 $\sigma_1=\sigma_2$	$\mu_1\leqslant\mu_2$ $\mu_2\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$t=\frac{(\bar{x}-\bar{y})}{s_w\sqrt{\frac{1}{m}+\frac{1}{n}}}$	$\{t\geqslant t_{1-\alpha}(m+n-2)\}$ $\{t\leqslant t_\alpha(m+n-2)\}$ $\{\lvert t\rvert\geqslant t_{1-\alpha/2}(m+n-2)\}$
大样本检验	σ_1,σ_2 未知 m,n 充分大	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$u=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$	$\{u\geqslant u_{1-\alpha}\}$ $\{u\leqslant u_\alpha\}$ $\{\lvert u\rvert\geqslant u_{1-\alpha/2}\}$
近似 t 检验	σ_1,σ_2 未知 m,n 不很大	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$t=\frac{(\bar{x}-\bar{y})}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$	$\{t\geqslant t_{1-\alpha}(l-1)\}$ $\{t\leqslant t_\alpha(l-1)\}$ $\{\lvert t\rvert\geqslant t_{1-\alpha/2}(l-1)\}$

- 正态总体方差的检验

检验法	H_0	H_1	检验统计量	拒绝域
χ^2 检验	$\sigma^2\leqslant\sigma_0^2$ $\sigma^2\geqslant\sigma_0^2$ $\sigma^2=\sigma_0^2$	$\sigma^2>\sigma_0^2$ $\sigma^2<\sigma_0^2$ $\sigma^2\neq\sigma_0^2$	$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2\geqslant\chi_{1-\alpha}^2(n-1)$ $\chi^2\leqslant\chi_\alpha^2(n-1)$ $\chi^2\leqslant\chi_{\alpha/2}^2(n-1)$ 或 $\chi^2\leqslant\chi_{1-\alpha}^2(n-1)$
F 检验	$\sigma^2\leqslant\sigma_0^2$ $\sigma^2\geqslant\sigma_0^2$ $\sigma^2=\sigma_0^2$	$\sigma^2>\sigma_0^2$ $\sigma^2<\sigma_0^2$ $\sigma^2\neq\sigma_0^2$	$F=\frac{s_x^2}{s_y^2}$	$F\geqslant F_{1-\alpha}(m-1,n-1)$ $F\leqslant F_\alpha(m-1,n-1)$ $F\leqslant F_{\alpha/2}(m-1,n-1)$或 $F\geqslant F_{1-\alpha/2}(m-1,n-1)$

- 分布拟合的检验
  - 分布总体只取有限个值
    - pi完全已知
    - pi不完全已知
  - 列联表检验
    - 考察各属性之间有无关联，即判断二属性是否独立
- 正态性检验
  - qq-plot
  - Shapiro-Wilk检验(W检验)
    建立次序统计量的基础上，将n个独立观测值按非降次序排序
    - 原假设成立，即总体分布为正态分布时，W的值应该接近于1
- 其他
  - 指数分布参数的假设检验
    - 卡方检验
  - 比例P检验
  - 大样本检验

概率统计

- 乘法公式
  全概率公式
  贝叶斯公式
- 常见的概率分布函数
  - 泊松分布
    - 用来描述某个固定时间内一个事件发生的次数
  - 指数分布
    - 用于描述事件发生的时间间隔
  - 正态分布
  - 卡方分布
    - 独立随机变量平方和的分布
    - 一个随机变量服从正态分布时，其期望除以方差的比值的分布
  - F分布
    - 两个独立卡方分布比值的分布
  - t分布
    - 正态分布总体中随机抽取的样本均值与总体均值之间偏差的分布
- 分布的特征数
  - k阶矩
    - 用来描述分布的形态及特征
      - 一阶矩：描述分布的中心位置
      - 二阶矩：描述分布的离散程度
      - 三阶矩：描述分布的偏斜程度
  - 变异系数
    - 标准差/期望，是以数学期望为单位去度量随机变量取值的波动程度
      - 比较两个随机变量的波动程度
  - 偏度系数
    - 描绘数据分布的偏态和倾斜程度
    - 分布的对称性
  - 峰度系数
    - 描绘数据分布的形状-峰态
    - 分布的峰峭性
- 概率的收敛性
  - 强收敛：依概率收敛
  - 弱收敛：按分布收敛
  - 强收敛VS弱收敛：强收敛要求点点收敛(一致收敛)，弱收敛在间断点可以为0(不考虑间断点)
- 大数定律与中心极限定理
  - 大数定律
    - 随着样本量的增大，事件发生的频率与其概率的偏差越来越小，即频率趋近于概率
  - 中心极限定理
    - 在独立重复实验中，当样本量足够大时，样本平均值会趋向于正态分布
      - 意义：通过中心极限定理，可以将数据随机变化的影响逐渐抵消，得到一个稳定的、可预测的分布