Lecture 1: 参数与非参数统计: 回顾与简介

张伟平
zwp@ustc.edu.cn
Office: 1006,
Phone: 63600565

作业课件等http://staff.ustc.edu.cn/~zwp/

交流论坛http://fisher.stat.ustc.edu.cn

Course outline

- **主要内容:** U统计量, 一样本方法, 两样本方法, 多样本方法, 成对比较与区组设计, 趋势与关联的检验, 经验分布函数与经验似然, 非参数Bootstrap方法, 非参数回归, 密度估计, 其他光滑方法
- **预修课程:** 概率论, 数理统计, 回归分析, R语言
- **参考书:**

1. 应用非参数统计, 薛留根, 科学出版社.
2. 现代非参数统计, 吴喜之译, 科学出版社.
3. Practical nonparametric statistics, Conover, W.J. 1999, 3rd edition. (实用非参数统计, 人民邮电出版社, 2006)
4. Nonparametric Statistical Inference, 5th Ed. Dickinson-Gibbons J, Chakraborti S, 2010

- **课程评定方法:** 课程评定分为网上和网下评定两部分，最终评分加权得到。

  1. 每人网上提交一个知识点和一道习题解答，网上知识点10分; 网上作业解答提交8分。网上作业和知识点报告最后的分数只有两种情况：满分或者0分。
  2. 课堂表现5分，由到课、提问、小测验等方式评定；网下作业17分。
  3. 作业的截止日期将以周来计算，根据网上的作业任务以及我们上课的进度，请大家在每周上课之前完成相应的习题。
  4. 助教会负责检查每位同学的标准解答，并对没有达到要求的提出修改意见。请同学们认真对待修改意见并作相应的修改。
  5. 课程总评分为期末考试(60%)+课堂评分(课堂5%+知识点10%)+作业(网下作业17%+网上作业8%)

- **网站** http://shjkx.wang

Chapter 0

Introduction and Review

0.1 Parametric and Nonparametric Statistics

- **参数统计**
  - **参数:** 刻画总体分布的(未知)常数
  - **统计量:** 仅依赖于样本的量
  - **参数方法:** 基于总体分布的一些假设下进行估计和推断

例 1. 假设IQ得分$X \sim N(\mu, 10^2)$, 我们观测到的*10*个IQ得分为*121, 98, 95, 94, 102, 106, 112, 120, 108, 109.* 考虑的问题是*:* 平均IQ得分是否显著的大于*100?*

零假设: $H_0 : \mu = 100$
对立假设: $H_1 : \mu > 100.$
检验方法: 在正态假设下, 使用z-test 进行检验.

- **非参数统计**
  - 不假定总体分布的形式
  - 估计和推断过程中对总体分布作较少的假设
  - 多基于大样本性质
  - “nonparametric”一词使用不是很恰当，我们的目的仍然是对参数进行估计或检验. 但不假设分布形式已知，一般仅假设样本是简单随机样本.
  - 更准确的术语: **distribution-free Statistics**

例 **2.** 假设IQ得分X_i 为*i.i.d*, 我们观测到的*10*个IQ得分为*121, 98, 95, 94, 102, 106, 112, 120, 108, 109.* 考虑的问题是: IQ得分的中位数是否显著的大于*100?*

零假设: $H_0 : med(X) = 100$
对立假设: $H_1 : med(X) > 100.$
检验方法: 使用非参数检验方法进行检验.

- **数据的测量尺度**
  - **nominal scale**: 列名尺度, 例如：性别、民族、职业
    数据表现为“类别”,各类之间无等级次序,各类别可以用数字代码表示
  - **ordinal scale**: 顺序尺度, 例如健康状况、质量等级
    数据表现为“类别”, 可对等级、大小等排序, 未测量出类别之间的准确差值
  - **interval scale**: 间隔尺度, 例如年份、摄氏温度
    数据表现为“数值”,可以进行加减运算, “0” 是只是尺度上的一个点，不代表“不存在”
  - **ratio scale**: 比例尺度,例如体重、身高
    数据表现为“数值”,可以进行加减、乘除运算, “0” 表示“没有”或“不存在”

- 为什么学习非参数统计
  - 在很多时候, 对于总体分布没有“先验”知识
  - 若参数统计方法的假设不成立,则统计推断结果可能有误
  - 在小样本场合, 正态逼近表现不佳
- 因此我们需要一类统计方法:
  - 对模型/分布仅作很少假设
  - 对模型/分布假设比较稳健/不敏感

0.2 Review of Probability distribution

0.2.1 Normal distribution

- 非常流行的对称钟形连续分布
- $X \sim N(\mu, \sigma^2)$的概率密度函数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  其中(μ, σ^2)为**参数**.

- $N(0,1)$为标准正态分布, 相应的分布函数记为Φ, 概率密度函数为ϕ
- 标准化: $X \sim N(\mu, \sigma^2)$, 则

$$\frac{X-\mu}{\sigma} \sim N(0,1)$$

- $N(\mu, \sigma^2)$的分布函数(CDF)

$$F_X(x) = P(X \le x) = P((X-\mu)/\sigma \le (x-\mu)/\sigma) = \Phi((x-\mu)/\sigma)$$

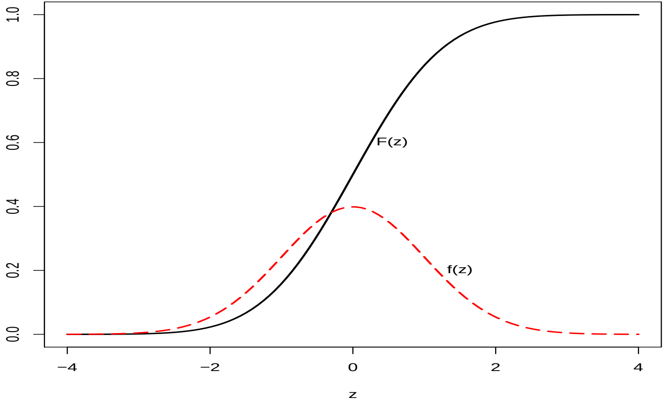

Figure 1: PDF and CDF of standard normal distribution.

0.2.2 Binomial distribution

- 由n(个数已知)个Bernoulli试验构成
- 每个Bernoulli试验只有“成功”(S) 和“失败”(F) 两种可能试验结果
- 每次试验的成功的概率均为$P(S) = p$
- 试验相互独立
- 二项分布随机变量X={n次试验中成功的次数}, 其分布律为

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n$$

  记为$X \sim Bin(n, p)$. p为**参数**.

- 均值$EX = np$, 方差$Var(X) = np(1-p)$.

0.3 Quantile

p分位数 假设总体分布函数为$F(x)$, 则$F(x)$的p分位数ξ_p是满足下述条件的一个数:

$$F(\xi_p) \geq p, F(\xi_p - 0) \leq p, \ 0 < p < 1$$

这样定义的p分位数可能不唯一(易证: 若p分位数不唯一, 则它充满一个有界区间) 为此, 定义p分位数为

$$\xi_p = \inf\{x : F(x) \geq p\}, p \in (0, 1)$$

上α分位数 称数x_α为连续分布$F(x)$的上α分位数, 如果

$$F(x_\alpha) = 1 - \alpha$$

- 显然对连续分布$F(x)$有$\xi_{1-\alpha} = x_\alpha$.
- 常使用该分布的名称来记上分位数, 如标准正态的上α分位数为z_α, t_n的上α分位数为$t_\alpha(n)$等

0.3.1 Descriptive Statistics

假设$X_1, \ldots, X_n\ i.i.d \sim f(x)$, 相应的CDF为$F(x)$. 常用的统计量包括

- 均值$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- 方差、标准差

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

- 次序统计量. 记其次序统计量为$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$, 则
  - 中位数

$$m_{1/2} = \begin{cases} X_{(\frac{n+1}{2})}, & n\text{为奇数} \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}, & n\text{为偶数} \end{cases}$$

  - p分位数

$$m_{n,p} = X_{\lfloor np \rfloor}$$

  - $X_{(n)}$有密度函数

$$f_n(t) = n[F(t)]^{n-1} f(t)$$

- $X_{(1)}$有密度函数

$$f_1(t) = n[1 - F(t)]^{n-1} f(t)$$

- $(X_{(i)}, X_{(j)}), (i < j)$有联合概率密度

$$f_{ij}(s,t) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(s)]^{i-1} \cdot$$
$$[F(t) - F(s)]^{j-i-1}[1 - F(t)]^{n-j} f(s) f(t) I(s < t)$$

- $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$有联合密度

$$f_{1\ldots n}(t_1, \ldots, t_n) = n! f(t_1) \cdots f(t_n) I(t_1 \leq \ldots \leq t_n)$$

- 经验分布函数

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x).$$

0.4 Review of Statistical Inference

0.4.1 Estimation

例 **3.** 若$X_1,\ldots,X_n$为来自正态总体$N(\mu,\sigma^2)$的一组简单随机样本，试求总体均值$a(F)=\mu$，总体方差$a(F)=\sigma^2$ 以及总体分布的上q 分位数$a(F)=\mu+\sigma z_q$ 的估计.

解:

利用矩估计方法. 由总体分布$X\sim N(\mu,\sigma^2)$有

$$EX=\mu,\qquad Var(X)=\sigma^2,$$

从而由矩估计方法知μ和σ^2的估计分别为

$$\hat{\mu}=\bar{X},\qquad \hat{\sigma}^2=S^2.$$

进而上q分位数的估计为$\bar{X}+S^2 z_q$.

利用最大似然估计方法. 由于对数似然函数为

$$l(\mu,\sigma^2)=\sum_{i=1}^{n} log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x_i-\mu)^2/(2\sigma^2)}\right)$$

$$\propto \frac{n}{2}log\sigma^2 - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}.$$

由于$l(\mu, \sigma^2)$为参数的可导函数, 令$\frac{\partial l(\mu,\sigma^2)}{\partial \mu} = 0$, $\frac{\partial l(\mu,\sigma^2)}{\partial \sigma^2} = 0$. 易知得到的解是对数似然函数的最大值点, 从而得到μ, σ^2的最大似然估计

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n}S^2.$$

根据最大似然估计的定义, 总体分布的上q分位数的最大似然估计为$\hat{\mu} + \hat{\sigma}^2 z_q$.

- **矩估计方法**(Method of moments): 需要已知关于总体分布的(一些)矩方程在满足适当正则条件下矩估计量$\tilde{\theta}_n$满足
  - 相合性: $\tilde{\theta}_n \to \theta$
  - 渐近正态性: $\sqrt{n}(\tilde{\theta}_n - \theta) \rightsquigarrow N(0, b^2)$
- **最大似然方法**(Maximum likelihood): 需要已知总体分布, 在满足适当正则条件下最大似然估计量$\hat{\theta}_n$满足
  - 相合性: $\hat{\theta}_n \to \theta$

- 渐近正态性: $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow N(0, I^{-1}(\theta))$, 一般地, 对可微函数$g$有

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \rightsquigarrow N(0, \nabla_\theta^T I^{-1}(\theta)\nabla_\theta) \qquad (\textbf{delta method})$$

其中$I(\theta)$为Fisher信息阵, $\nabla_\theta = \nabla g(\theta)$.

- 估计量优良性评定
  - 无偏性
  - 有效性
  - 最小方差无偏估计
    - **Cramér-Rao** 不等式:

$$Var_\theta(\hat{g}) \geq [g'(\theta)]^2/(nI(\theta))$$

    - **Lehmann-Scheffé** 定理: 设T为充分完全统计量, $\hat{g}(T)$为$g(\theta)$的无偏估计, 则$\hat{g}$为g 的最小方差无偏估计.
    - **若g的有效估计存在, 则MLE $\hat{g}$ 必为有效估计.**

0.4.2 Confidence set

令$\mathcal{F}$为分布函数F的一个类, 而θ为感兴趣的量，例如均值、密度或者回归函数等. 令C_n为θ的可能取值的集合, 它依赖于样本$X_1, \ldots, X_n$. 则

- 如果
$$\inf_{F \in \mathcal{F}} P(\theta \in C_n) \geq 1 - \alpha$$
对所有的n都成立, 那么C_n称为θ的一个**(有穷样本)$1-\alpha$置信集**.

- 如果
$$\lim_{n \to \infty} \inf \inf_{F \in \mathcal{F}} P(\theta \in C_n) \geq 1 - \alpha$$
那么C_n称为θ的一个**一致渐近$1-\alpha$置信集**.

- 如果
$$\lim_{n \to \infty} \inf P(\theta \in C_n) \geq 1 - \alpha$$
对所有的$F \in \mathcal{F}$都成立, 那么C_n称为θ的一个**逐点渐近$1-\alpha$置信集**.

从而当θ为一维未知参数，$\mathcal{F}=\{F_\theta(x): F_\theta\text{连续}, \theta\in\Theta\}$且$C_n=[A_n, B_n]$时候,

- 置信区间：

$$\inf_{F\in\mathcal{F}} P(\theta\in C_n)\geq 1-\alpha \Longleftrightarrow$$
$$P(A_n\leq\theta\leq B_n)=1-\alpha, \forall\theta\in\Theta, \forall n$$

- 一致渐近置信区间：

$$\lim_{n\to\infty}\inf_{F\in\mathcal{F}} P(\theta\in C_n)\geq 1-\alpha \Longleftrightarrow$$
$$\lim_{n\to\infty}\inf_{\theta\in\Theta} P(A_n\leq\theta\leq B_n)=1-\alpha$$

- 逐点渐近区间：

$$\lim_{n\to\infty} P(\theta\in C_n)\geq 1-\alpha\text{对所有的}F\in\mathcal{F}\text{都成立} \Longleftrightarrow$$
$$\lim_{n\to\infty} P(A_n\leq\theta\leq B_n)=1-\alpha, \forall\theta\in\Theta$$

例 4. 假设$X_1,\ldots,X_n i.i.d \sim Bin(1,p)$，求$p$的$1-\alpha$置信区间。

记$X=\sum_{i=1}^n X_i$，则$X\sim Bin(n,p)$，在观测到$X=x$时p的$1-\alpha$置信区间为

(**Clopper-Pearson Interval**)

(精确区间)
$$\left[(1+\frac{n-x+1}{xF_{1-\alpha/2}(2x,2(n-x+1))})^{-1},\ (1+\frac{n-x}{(x+1)F_{\alpha/2}(2(x+1),2(n-x))})^{-1}\right]$$

(**Wald Interval**)

(渐近区间)
$$\left[\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

0.4.3 Testing

- 原假设与对立假设$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta_1$
- 拒绝域：使得原假设被拒绝的样本区域
- 检验φ的功效函数

$$\beta_\varphi(\theta) = P_\theta(\text{使用检验}\varphi\text{拒绝}H_0)$$

- p值$=P$(在H_0下，得到检验统计量T观测值T_{obs}那么大或者更极端)
    - 双边假设$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0$

$$p\text{ 值} = P_{\theta_0}(|T| \geq |T_{obs}|)$$

    - 左边假设$H_0 : \theta \geq \theta_0 \leftrightarrow H_1 : \theta < \theta_0$

$$p\text{ 值} = P_{\theta_0}(T < T_{obs})$$

    - 右边假设$H_0 : \theta \leq \theta_0 \leftrightarrow H_1 : \theta > \theta_0$

$$p\text{ 值} = P_{\theta_0}(T > T_{obs})$$

检验的水平 称检验法则φ有水平α，如果

$$\beta_\varphi(\theta) \leq \alpha, \theta \in \Theta_0$$

UMP 检验 称水平α检验φ为假设$H_0: \theta \in \Theta_0 \leftrightarrow H_1: \theta \in \Theta_1$的一致最优检验(UMPT), 如果对该假设的任意水平α检验φ_1都有

$$\beta_\varphi(\theta) \geq \beta_{\varphi_1}(\theta), \theta \in \Theta_1$$

寻找UMPT:

- **Neyman-Pearson lemma**
- MLR (Monotone Likelihood ratio) 族

基于似然的常用检验方法

对假设$H_0 : \theta \in \Theta_0 \leftrightarrow H_1 : \theta \in \Theta - \Theta_0$, 假设基于i.i.d样本$X_1, \ldots, X_n$的似然函数为$L_n(\theta)$, $\hat{\theta}_n$为最大似然估计. 则

- **The likelihood ratio test**:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta} L_n(\theta)}{\sup_{\theta \in \Theta_0} L_n(\theta)}$$

可证在H_0成立条件下，当$n \to \infty$时有

$$2log\lambda(x) \rightsquigarrow \chi^2_{dim\Theta - dim\Theta_0}$$

对假设$H_0 : \theta = \theta_0$,

- **The score test statistic**:

$$\frac{1}{\sqrt{n}} S(\theta_0)^T I^{-1}(\theta_0) \frac{1}{\sqrt{n}} S(\tilde{\theta}_0)$$

其中$S(\theta) = \partial log L_n(\theta)/\partial\theta$.

- **The Wald test statistic**:

$$n(\hat{\theta}_n - \theta_0)^T \left\{ I^{-1}(\hat{\theta}_n) \right\}^{-1} (\hat{\theta}_n - \theta_0)$$

三种检验的比较

- Wald检验使用的不是零假设下的标准误差
- Score检验使用的是零假设下的标准误差
- 似然比检验综合利用了零假设和观测似然
- 对小规模和中等规模样本量, 似然比检验一般好一些