Basics of Probs and Stats

6. 参数估计

梁雪吟

地球科学与资源学院

June 16, 2024

Outline

基本概念

- 估计量 (Estimator)：是一个随机变量
  - 因为我们无法观测完全总体的信息，所以通过随机样本包含的信息来构造估计量，从而达到用样本估计总体的目的
  - 估计量是样本的函数，若是知道了样本的观测值，估计量的估计值就可以确定
- 估计值 (Estimate)：是一个常数，是估计量的实现值

矩估计 MoM

- 矩估计的基础：大数定律，样本矩依概率收敛于总体矩

总体 K 阶矩

连续型：$\mu_l = E(X^l) = \int_{-\infty}^{\infty} x^l f(x; \theta_1, \theta_2, \ldots, \theta_n)\, dx$

离散型：$\mu_l = E(X^l) = \sum_{x \in \Omega_x} x^l f(x; \theta_1, \theta_2, \ldots, \theta_n)\, dx$

$$(l = 1, 2, \ldots, k)$$

样本 K 阶矩

$$A_l = \frac{1}{n} \sum_{i=1}^{n} X_i^l$$

$$(l = 1, 2, \ldots, k)$$

- $\mu_1 = E(X)$ 即总体期望，$\mu_2 = E(X^2) = D(X) + [E(X)]^2$

练习

最大似然估计 MLE

最大似然函数

离散型： $L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$

连续型： $L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$

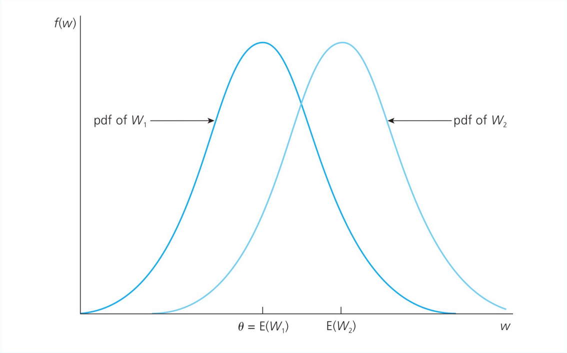
- 最大似然估计： $L(x_1, x_2, ..., x_n; \hat{\theta}) = maxL(x_1, x_2, ..., x_n; \theta)$
- 最大似然估计量： $\hat{\theta}(X_1, X_2, ..., X_n)$
- 最大似然估计值： $\hat{\theta}(x_1, x_2, ..., x_n)$
- 求解步骤：
  1. 写出似然函数（或方程组）
  2. 等式两边取对数
  3. 求出关于估计量的一阶导数，并令其等于 0
  4. 解得估计量

练习

无偏性 Unbiasedness

- 无偏性的含义：$E(\hat{\theta})=\theta$ 估计量在“平均意义上”非常接近待估计的参数
  - 多次抽样获得多个样本数据集，每个数据集算出一个估计值，这些估计值们的期望与总体的参数相等

FIGURE C.1 An unbiased estimator, W_1, and an estimator with positive bias, W_2.



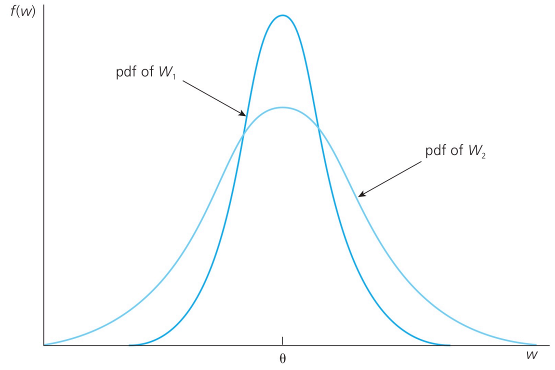
几个重要的无偏估计量

- 样本均值是总体均值的无偏估计: $E(\bar{X}) = \mu$
- 样本 k 阶矩是总体 k 阶矩的无偏估计
- 样本方差是总体方差的无偏估计: $E(S^2) = \sigma^2$

有效性 Efficiency

- 在无偏估计的前提下，方差更小更有效

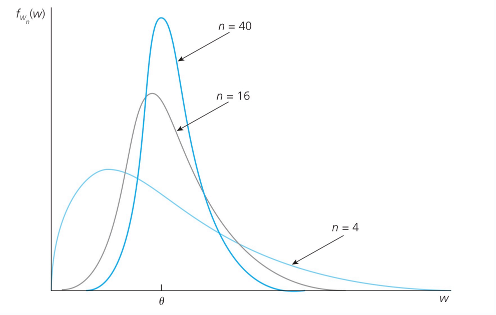
FIGURE C.2 The sampling distributions of two unbiased estimators of θ.



相合性 Consistency

- 渐进，比无偏估计更放松
- 当样本的容量 N 越来越大的时候，$\hat{\theta}$ 的取值（即估计值）将越来越接近总体参数 θ
  - 极端情况：当 N 无穷大，样本包含了总体的所有信息的时候，此时，$\hat{\theta}$ 的取值（即估计值）几乎等于总体参数 θ

FIGURE C.3 The sampling distributions of a consistent estimator for three sample sizes.



- 由图我们可以看出相合性要求样本容量较大

练习

置信区间的解释

- 为什么需要置信区间？
  - 在实际应用中，我们很难通过单个样本数据集计算出的估计值是否偏离总体的参数，也不知道偏离程度具体有多大。此外，点估计的无偏性只是代表大样本情况下，样本的均值（平均意义上）与总体参数相等。
  - 我们希望有一个“较窄的、大概率能能**包含到**总体参数的区间”，即置信区间，此区间覆盖到参数 θ 的概率叫做置信度。
- 置信区间的意思是：有 95% 的概率能**包含到**总体参数。（不能说参数落在置信区间，因为总体参数是固定的常数，置信区间的上下限才是需要我们计算的随机变量）

单个正态总体参数的区间估计

- 单个正态总体均值的区间估计
  - 方差 σ 已知，求均值 μ 的区间估计，取 $U=\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1)$
    - 置信区间为：$(\bar{X}-u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{X}+u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$
  - 方差 σ 未知，求均值 μ 的区间估计，取 $T=\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}\sim t(n-1)$
    - 置信区间为：$(\bar{X}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\bar{X}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}})$
- 单个正态总体方差和标准差的区间估计
  - 均值 μ 已知，求方差 σ 的区间估计，取 $\frac{X_i-\mu}{\sigma}\sim N(0,1)$
    - 方差 σ^2 置信区间为：
      $$\left(\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{a}{2}}(n)},\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{a}{2}}(n)}\right)=\left(\frac{(n-1)S^2}{\chi^2_{\frac{a}{2}}(n)},\frac{n-1)S^2}{\chi^2_{1-\frac{a}{2}}(n)}\right)$$
    - 标准差 σ 置信区间为：$\left(\frac{\sqrt{n-1}S}{\sqrt{\chi^2_{\frac{a}{2}}(n-1)}},\frac{\sqrt{n-1}S}{\sqrt{\chi^2_{1-\frac{a}{2}}(n-1)}}\right)$

两个正态总体参数的区间估计

- 两个正态总体均值的区间估计
  - σ_1^2 和 σ_2^2 已知，求 $\mu_1-\mu_2$ 的区间估计，取：
    $\bar{X}-\bar{Y}\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$ 或
    $\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1)$
    - 置信区间为：$\left(\bar{X}-\bar{Y}\pm z_{a/2}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}\right)$
    - 当 n>50 时，样本方差满足无偏估计，置信区间为：
      $\left(\bar{X}-\bar{Y}\pm z_{a/2}\sqrt{S_1^2/n_1+S_2^2/n_2}\right)$
  - 当 $\sigma_1^2=\sigma_2^2=\sigma^2$，但$\sigma^2$ 为未知 时，
    - $\mu_1-\mu_2$ 的置信度为 $1-\alpha$ 的置信区间为：
      $\left(\bar{X}-\bar{Y}\pm S_w t_{a/2}\left(n_1+n_2-2\right)\sqrt{1/n_1+1/n_2}\right)$
      其中 $S_w^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, $S_w=\sqrt{S_w^2}$

两个正态总体参数的区间估计

- 两个正态总体方差的区间估计
  - μ_1，μ_2 未知，求 $\frac{\sigma_1^2}{\sigma_2^2}$ 的区间估计，取：

    $$\frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

    - $\frac{\sigma_1^2}{\sigma_2^2}$ 置信区间为：

      $$\left(\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{a/2}(n_1-1,n_2-1)}, \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right)$$

    - $\frac{\sigma_1}{\sigma_2}$ 置信区间为：

      $$\left(\sqrt{\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{a/2}(n_1-1,n_2-1)}}, \sqrt{\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}}\right)$$

非正态总体的区间估计（了解）

- 思路：因为我们已经知道了正态总体的区间估计，我们又学过中心极限定理，样本量足够大的时候样本均值满足中心极限定理——依分布收敛到正态分布。所以中心极限定理（标准化参数）是此类型的桥梁。
- 假设总体 X 的均值 μ 与方差 σ^2 存在, $X_1, X_2, \cdots, X_n$ 为来自总体 X 的样本, $\bar{X}$ 与 S^2 为样本均值和样本方差, 取 $T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$
  - 均值 μ 的置信度为 $1-\alpha$ 的近似置信区间为:

$$\left(\bar{X} - \frac{S}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{n}} z_{a/2}\right)$$

Thank You

June 16, 2024