第一次作业

1907402030 熊雄

2021 年 9 月 14 日

题目 1. 设 $y_1, \dots, y_n$ 是一组样本，其中 μ 和 σ 都是未知的。构建下面模型

$$y_i = \mu + \varepsilon_i, \quad i = 1, 2, \dots, n$$

我们可以采用

- 最小二乘法估计: 最小化 $\sum_{i=1}^{n}(y_i - \mu)^2$
- 最小绝对值估计: 最小化 $\sum_{i=1}^{n}|y_i - \mu|$

回答下面问题:

1. 证明 μ 的最小二乘估计是样本均值;
2. 证明 μ 的最小绝对值估计是样本中位数;
3. 列出样本均值的一个优点和一个缺点;
4. 列出样本中位数的一个优点和一个缺点;
5. 你会选择 μ 的两个估计量的哪一个? 说出理由.

解答.

1. **证明:** 设 $f(t) = \sum_{i=1}^{n}(y_i - t)^2$, 对 t 求导得到:

$$f'(t) = -2\sum_{i=1}^{n}(y_i - t)$$

令 $f'(t) = 0$, 可以得到

$$t_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

又因为 $f''(t_0) = 2n > 0$, 故 $t_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ 为 $f(\mu)$ 的极小值. 而我们知道 $\frac{1}{n}\sum_{i=1}^{n} y_i$ 就是样本均值 μ. 故 μ 的最小二乘估计是样本均值，证毕.

2. **证明：** 设样本中位数为 y_{mid}. 由于 $y_1, \ldots, y_n$ 是一组样本的观测值，则将其任意排序后总能使其递增排列，故我们不妨假设其为递增数列，即

$$y_1 \leq y_2 \leq \cdots \leq y_n$$

   (a) 当 n 为奇数时，记 $U_i = [y_i, y_{n+1-i}]$,$i = 1, 2, \ldots, \frac{n+1}{2}$. 我们知道, $|y_1 - a| + |y_n - a|$ 在 $a \in U_1$ 时取得最小值; $|y_2 - a| + |y_{n-1} - a|$ 在 $a \in U_2$ 时取得最小值...... 以此类推, 最后只剩下 $\left|y_{\frac{n+1}{2}} - a\right|$ 在 $a \in U_{\frac{n+1}{2}}$ 时取得最小值, 即 $a = y_{\frac{n+1}{2}} = y_{mid}$. 将所有的 U_i 取交集:

$$\cap_{i=1}^{\frac{n+1}{2}} U_i = U_{\frac{n+1}{2}} = y_{mid}$$

   即 $a = y_{mid}$ 时使得所求离差绝对值的和最小, μ 的最小绝对值估计是样本中位数.

   (b) 当 n 为偶数时，记 $V_i = [y_i, y_{n+1-i}]$,$i = 1, 2, \ldots, \frac{n}{2}$. 我们知道, $|y_1 - a| + |y_n - a|$ 在 $a \in V_1$ 时取得最小值; $|y_2 - a| + |y_{n-1} - a|$ 在 $a \in V_2$ 时取得最小值...... 以此类推, 最后只剩下 $\left|y_{\frac{n}{2}} - a\right| + \left|y_{\frac{n}{2}+1} - a\right|$ 在 $a \in V_{\frac{n}{2}}$ 时取得最小值. 将所有的 V_i 取交集:

$$\cap_{i=1}^{\frac{n}{2}} V_i = V_{\frac{n}{2}} = y_{mid}$$

   即 $a = y_{mid}$ 时使得所求离差绝对值的和最小, μ 的最小绝对值估计是样本中位数.

   综上所述, μ 的最小绝对值估计是样本中位数, 证毕.

3. **答：**

   (a) **样本均值一个优点：** 样本均值为总体均值的无偏估计、有效估计和相合估计，是总体均值的最优的估计量;

(b) **样本均值一个缺点：** 样本均值容易受极端值影响。当一组观测值为明显的偏态分布时，用样本均值来估计总体均值的效果比较差。

4. **答：**

(a) **样本中位数一个优点：** 样本中位数不受极端值影响；

(b) **样本中位数一个缺点：** 当样本容量较大时，计算样本中位数比较复杂，比如若采用快速排序算法的时间复杂度为 $O(n\log(n))$。

5. **答：** 我会选择样本均值. 由第 3 问的样本均值的优点可知，样本均值是总体均值的最优的估计量. 我们研究样本的观测值的主要目的就是来估计总体, 而样本中位数不能很好地反映总体的特征，因此我会选择样本均值.