非参数统计分析

Lollins

2023 年 9 月 21 日

前言

记点非参数统计分析的笔记。

Lollins

2023 年 9 月 21 日

目录

第一章　绪论

1.1　序

1.1.1　非参数统计概念及学习意义

1、意义

2、概念

- **参数统计方法：** 数据样本被视为从分布族的某个参数族抽取出来的总体的代表，未知的仅仅是总体分布具体数值，这样推断问题就转化为分布族的若干未知参数的估计问题，用样本来对这些参数进行估计或进行假设检验，从而得知背后的分布，这类推断方法称为参数统计方法。

- **非参数统计方法：** 不假定总体分布的具体形式，尽量从数据（或样本）本身获得所需要的信息，通过估计而获得分布的结构，并逐步建立对事物的数学描述和统计模型的方法。

1.1.2　非参数统计的历史及发展

1.2　引言

1.2.1　参数统计方法与非参数统计方法的区别

- **参数统计方法**：假定总体的分布形式，既利用样本的数据信息，又利用产生数据总体的信息，是一个有效的数据分析方法，针对性强，但可能出现大的错误。

- **非参数统计方法**：不假定总体的分布形式，更接近大多数实际情况，故不会出现大的错误。

1.2.2　非参数统计方法的特点

(1) 有广泛的适用性（广）

(2) 样本方法是非参数统计的基本方法（样本）

(3) 计算简单（简）

(4) 良好的稳定性（稳）

第二章　描述性统计

定义 2.0.1 (描述性统计). 是在对产生数据的总体的分布不作任何假设的情况下，整理数据、显示数据、分析数据，将数据中有用的信息提取出来的统计方法。本章介绍常用的描述性统计方法：**表格法、图形法和数值方法**。

2.1　图表法

表格法、图形法描述统计数据主要是频数（率）分布表和直方图。

2.2　数值方法

数值方法主要是用数值来表示数据的中心位置和离散程度等的方法。

2.2.1　表示中心位置的数值

我们要求数据的中心位置满足这样一个**条件**：它到各个数据点的距离的和比较小。表示中心位置的数值有平均数、中位数、众数、切尾平均数。

1、平均数

如果用平方值距离法，则点 a 到各数据点 $x_1, x_2, ..., x_n$ 的距离的和可以用 $\sum_{i=1}^{n}(x_i - a)^2$ 来衡量。平均数 $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ 满足条件：

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min_a \sum_{i=1}^{n}(x_i - a)^2 \tag{2.1}$$

上式表示**平均数这一点到各个数据点的平方值距离和最短**。所以在**平方值距离方法下**, 数据中心位置的代表是**平均数**。

2、中位数

如果用绝对值距离法，则点 a 到各数据点 $x_1, x_2, ..., x_n$ 的距离的和可以用 $\sum_{i=1}^{n}|x_i - a|$ 来衡量，中位数 me 满足条件：

$$\sum_{i=1}^{n}|x_i - \max| = \min_a \sum_{i=1}^{n}|x_i - a| \tag{2.2}$$

上式表示**中位数这一点到各个数据点的绝对值距离和最短**。所以在**绝对值距离方法下**, 数据中心位置的代表是**中位数**。

注：

- 中位数是非线性规划选址问题的解；
- 中位数不受极大（小）的影响，有时能较好地表示数据的中心位置。

3、众数

众数：一组数据中出现频数最高的数据。

注：

- 众数也能描述数据的中心位置。特别是定性数据；
- 一组数据有偏时，若数据右偏 (Positively Skewed)，通常有 $\bar{x} < me < mo$，若数据左偏 (Negatively Skewed)，通常有 $mo < me < \bar{x}$，见图2.1。

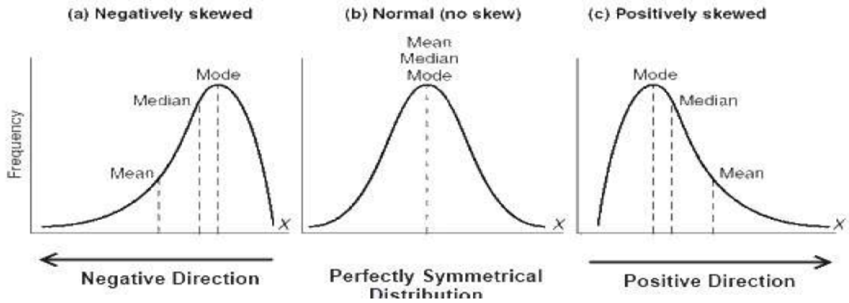

图 2.1

4、切尾平均数

设 $X_{(1)},...,X_{(n)}$是来自总体 X 的简单随机样本$X_1,...,X_n$ 的次序统计值，称

$$T_{nk}=\frac{1}{n-2k}(x_{(k+1)}+\ldots+x_{(n-k)}) \tag{2.3}$$

为原样本的的切尾均值。

2.2.2 表示离散程度的数值

样本方差、标准差、全距（范围）、四分位数间距。

2.2.3 标准误

$$se=\frac{s}{\sqrt{n}},s\text{为样本方差} \tag{2.4}$$

2.2.4 偏度

偏度反映单峰分布对称性，常用 β_s 表示总体偏度，

$$\beta_s=E[(\frac{x-\mu}{\sigma})^3]=\frac{\mu_3}{\sigma^3},\text{其中}\mu_3=E(x-\mu)^3 \tag{2.5}$$

注：对称分布的偏度 $\beta_s=0$；反之不成立，即 $\beta_s=0$，不一定是对称分布。

样本偏度用 b_s 表示，

$$b_s = \frac{m_3}{m_2^{\frac{3}{2}}}, \text{其中} m_j = \frac{1}{n}\sum_i (x_i - \overline{x})^j \tag{2.6}$$

注：$b_s > 0$ 时，倾向于认为数据分布右偏；$b_s < 0$ 时，倾向于认为数据分布左偏；$b_s \approx 0$ 时，倾向认为数据分布是对称的。

2.2.5　峰度

峰度反映分布峰的尖峭程度，常用 β_k 表示总体峰度。

$$\beta_k = E[(\frac{x-\mu}{\sigma})^4] = \frac{\mu_4}{\sigma^4} \tag{2.7}$$

注：若 $X \sim N(\mu, \sigma^2)$，则 $\beta_k = 3$。当 $\beta_k > 3$ 时，该分布具有过度的峰度，当 $\beta_k < 3$ 时，该分布具有不足的峰度，

样本峰度用 b_k 表示，

$$b_k = \frac{m_4}{(m_2)^2} \tag{2.8}$$

参考文献

[1] 孙山泽. 非参数统计讲义. 北京大学出版社

[2] 陈希孺. 非参数统计. 中国科学技术大学出版社

[3] 李裕奇. 非参数统计方法. 西南交通大学出版社