第六章 数理统计的基本概念

北京化工大学数学系

苏贵福

与概率论一样, 数理统计也是研究大量随机现象的统计规律的一门数学学科, 它以概率论 为理论基础, 根据试验或观察到的数据, 对研究对象的客观规律性做出合理的估计和科学的推断.

数理统计的内容包括: 如何收集并整理数据资料; 如何对所得到的数据资料进行分析, 从而对所研究对象的性质或特征做出推理.

- 在概率论中, 我们总是假设所研究的随机变量的分布都是已知的, 在该前提下去研究随机变量的性质和规律.

- 在数理统计中, 我们研究的随机变量的分布是未知的, 或者 是不完全确定, 人们是通过对研究的随机变量进行重复独立的考察, 得到许多观测值. 然后对这些数据进行分析, 从而对研究的随机变量的分布做出各种推断.

6.1数理统计中的基本概念

一. 随机样本

定义1 在数理统计中, 我们往往研究有关对象的某一项数量指标. 为此考虑与这一数量指标 相联系的随机试验, 对这一数量指标进行试验或观察. 我们将试验的全部可能的观察值称为总体. 每一个可能的观察值称为个体. 总体中包含个体的数目称为总体的容量.

- 在考察某大学一年级男生的身高这一试验中，若一年级男生共3000人，每个男生的身高是一个可能观察值，所形成的总体中共含有3000个可能观察值(有限总体).
- 考察全国正在使用的某种型号灯泡的寿命所形成的总体，由于可能观察值 的个数很多，可以认为是无限总体.

- 总体中的每一个个体是随机试验的一个观察值, 因此它是某一随机变量X的值. 这样一个总体对应于一个随机变量. 故对总体的研究就是对一个随机变量的研究.

- 在实际中, 总体的分布一般是未知的, 或只知道它具有某种形式而其中包含着未知参数. 在数理统计中, 人们都是通过从总体中抽取一部分个体, 根据获得的数据来对总体的分布加以 推断. 被抽取的部分个体叫做总体的一个样本.

- 从总体中抽取样本时，不仅要求每一个个体被抽到的机会均等，同时还要求每次抽取是独立的，即每次抽样的结果不影响其它各次抽样的结果，这种抽样方法称为随机抽样.

- 从总体X中抽取一个个体，就是对随机变量X进行一次试验. 抽取n个个体就是对随机变量X进行n次试验，分别记作$X_1, X_2, \cdots, X_n$，则样本就是n维随机变量$(X_1, X_2, \cdots, X_n)$.

- 在一次抽样以后，n维随机变量$(X_1, X_2, \cdots, X_n)$就有了一组确定的值$(x_1, x_2, \cdots, x_n)$，将其称为样本的一个观测值.

定义2 设X是具有分布函数$F(x)$的随机变量，若$X_1, X_2, \cdots, X_n$ 是具有同一分布函数$F(x)$的相互独立的随机变量，则称$X_1, X_2, \cdots, X_n$为从分布函数$F(x)$或总体X得到的容量 为n的随机样本，简称样本. 它们的观察值$(x_1, x_2, \cdots, x_n)$称为样本值.

- $X_1, X_2, \cdots, X_n$相互独立.
- $X_1, X_2, \cdots, X_n$与总体具有相同的分布.

定理1 若$(X_1, X_2, \cdots, X_n)$为X的一个样本，则$(X_1, X_2, \cdots, X_n)$的联合分布函数为

$$F^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i).$$

若X具有概率密度$f(x)$，则$(X_1, X_2, \cdots, X_n)$的联合密度函数为

$$f^*(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i).$$

二. 统计量

定义3 设$(X_1,X_2,\cdots,X_n)$是来自总体的一个样本, $g(X_1,X_2,\cdots,X_n)$是关于$X_1,X_2,\cdots,X_n$的一个连续函数且$g(X_1,X_2,\cdots,X_n)$中不含 任何未知参数, 则称$g(X_1,X_2,\cdots,X_n)$是样本$(X_1,X_2,\cdots,X_n)$的一个<u>统计量</u>. 若$(x_1,x_2,\cdots,x_n)$是相应于样本$(X_1,X_2,\cdots,X_n)$的一个样本值, 则称$g(x_1,x_2,\cdots,x_n)$为$g(X_1,X_2,\cdots,X_n)$的<u>观察值</u>.

常用的统计量是由样本的如下几类数字特征构造的:

- 样本均值 $$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$
- 样本方差 $$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$
- 样本标准差 $$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$$
- 样本k阶原点矩 $$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k\ (k=1,2,\cdots)$$
- 样本k阶中心矩 $$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k\ (k=1,2,\cdots)$$

定理2 若总体X的期望$E(X)$存在，方差$D(X)$存在，则有

$$\lim_{n} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i^k - E(X^k)\right| < \epsilon\right\} = 1.$$

证明 由$(X_1, X_2, \cdots, X_n)$的独立同分布性可知

$$E(X_1) = E(X_2) = \cdots = E(X_n) = E(X).$$

由于$X_1^k, X_2^k, \cdots, X_n^k$也具有独立性，且与$X^K$同分布，因此

$$E(X_1^k) = E(X_2^k) = \cdots = E(X_n^k) = E(X^k).$$

于是由独立同分布大数定律知

$$\lim_{n} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i^k - E(X^k)\right| < \epsilon\right\} = 1.$$

定理3 设总体X的均值和方差分别为μ和σ^2. 若$(X_1, X_2, \cdots, X_n)$ 是X的一个样本, 则$E(\overline{X}) = \mu$, $D(\overline{X}) = \frac{\sigma^2}{n}$, $E(S^2) = \sigma^2$.

解 根据定义有

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n}\sigma^2$$

$$\begin{aligned}
E(S^2) =& E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\} = E\left\{\frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \mu) - (\overline{X} - \mu)]^2\right\} \\
=& E\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\right]\right\} \\
=& \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\overline{X} - \mu)^2\right] = \frac{1}{n-1}\left(n\sigma^2 - n\cdot\frac{\sigma^2}{n}\right) = \sigma^2.
\end{aligned}$$

三. 理论分布与经验分布

定义4 设$X_1, X_2, \cdots, X_n$是总体$F(x)$的一个样本, 用$S(x)$, $-\infty < x < \infty$, 表示$X_1, X_2, \cdots, X_n$中不大于x的随机变量的个数. 我们称函数

$$F_n(x) = \frac{1}{n} S(x), \quad -\infty < x < \infty$$

为总体的经验分布函数.

设总体F具有一个样本值1,2,3, 则经验分布函数$F_3(x)$为

$$F_3(x) = \begin{cases} 0, & x < 1 \\ \frac{1}{3}, & 1 \le x < 2 \\ \frac{2}{3}, & 2 \le x < 3 \\ 1, & x \ge 3 \end{cases}$$

一般地, 设$x_1, x_2, \cdots, x_n$是总体F的一个容量为n的样本值. 先将$x_1, x_2, \cdots, x_n$按从小到大的次序排列, 并重新编号为

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

则经验分布函数为

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{k}{n}, & x_{(1)} \leq x < x_{(k+1)}, \ (k = 1, 2, \cdots, n-1) \\ 1, & x \geq x_{(n)} \end{cases}$$

♠ 总体分布函数的求法

定理4 (格列汶科定理) 当$n \to \infty$时, 经验分布函数$F_n(x)$以概率1关于x均匀收敛于理论分布函数$F(x)$. 即

$$P\left\{\lim_{n\to\infty}\sup_{-\infty<x<\infty}|F_n(x)-F(x)|=0\right\}=1.$$

格列汶科定理表明: 当n很大时我们可用$F_n(x)$近似代替$F(x)$, 而且这种近似程度相当高. 这也是我们借助样本推断总体的理论依据.

♠ 总体密度函数的求法

设总体X的密度函数$f(x)$未知，$x_1, x_2, \cdots, x_n$为总体X的样本值. 我们可以根据这组样本值来近似求出总体X的密度函数. 这就是直方图法，具体步骤如下:

1. 整理资料

把样本值$x_1, x_2, \cdots, x_n$重新编号排序为$x_1^* \le x_2^* \le \cdots \le x_n^*$. 在包含$[x_1^*, x_n^*]$的区间$[a, b]$插入一些等分点: $a < t_1 < t_2 < \cdots < t_m < b$. 分点的数目一般在6到17间为宜，但要求每个区间$(t_i, t_{i+1}]$内须有样本观察值x_i落入其中.

2. 计算频率

统计出样本观察值落入各小区间$(t_i, t_{i+1}]$中的个数，即频数v_i. 然后计算样本值落入各区间内的频率$f_i = \frac{v_i}{n}$.

当n充分大时，f_i可近似地表示为随机变量X落入区间$(t_i, t_{i+1}]$内的概率，即

$$f_i \approx P\{t_i < X \le t_{i+1}\} = p_i = \int_{t_i}^{t_{i+1}} f(x)dx.$$

3. 作频率分布图 (详见教材P101)

4. 作频率直方图

在xOy平面上，在x轴上以各小区间$(t_i, t_{i+1}]$为底，以频率与组距之比$y_i = \frac{f_i}{t_{i+1}-t_i}$为第$i$个长方形的高，画一排竖着的长方形，即频率直方图.

5. 作分布密度曲线

过每一个小长方形的"上边"作一条光滑曲线, 这条曲线可以作为X的密度函数$y = f(x)$的近似曲线. 不难发现, 如果样本容量越大, 分组越细, 则这样画出的密度曲线就越准确.