第六章 数理统计的基本概念

第一节　样本与统计量

数理统计是一门研究带有随机影响数据的学科，是数学中的一个重要分支. 它主要研究如何有效地收集数据、并利用一定的统计模型对数据进行理论分析和统计推断，从中提取有用的信息，形成统计结论，为决策提供依据. 而分析和推断是以概率论的理论为依据的，因此，数理统计是概率论的一种应用.

数理统计的研究内容可分为抽样理论和统计推断.

抽样理论研究如何更合理、更有效地获取数据．获取数据的过程称为抽样，抽样的方式可分为全面观察、随机抽样观察、安排特定的试验等等．本教材中主要讨论随机抽样观察方式.

统计推断是指如何利用抽样的数据,对所考察的对象（总体）的某种性质作出尽可能准确的推断．在随机抽样观察方式下，体现为“由部分推断总体”．统计推断包括参数估计和假设检验两类基本问题.

一、总体与样本

1. 总体与个体

定义 1 将研究问题中，所有被考察对象的全体称为**总体**，总体中的每一个成员称为**个体**.

例如，考察某品牌电视机的寿命，则所有该品牌电视机的寿命就是总体，而其中每台电视机的寿命就是个体.

由于被考察对象往往是某数量指标 X ，因此总体可以理解为该数量指标 X 的全体. 研究总体就是研究其数量指标 X ． 而在实际问题中，数量指标 X 是一个随机变量.

随机变量 X 的分布称为总体的分布, 总体的特征是由总体的分布刻画的. 为此, 常把总体与总体分布视为等同, 并称总体 X ．

如果从总体中随机抽取一个个体，则此个体也为随机变量，且与总体同分布.

例 1 考察某产品的次品率，令总体

$$X=\begin{cases}1, & \text{产品为次品,}\\ 0, & \text{产品为正品,}\end{cases}$$

因此总体 X 的取值为1和0，总体 X 为有限总体，也是离散型总体，如果记该产品的次品率为 p，则总体 $X\sim B(1,p)$.

例 2 利用仪器测量某物体的高度．假设该物体高度的精确值为 μ，测量值为 X，则总体 X 为无限总体，也是连续型总体．通常认为 $X\sim N(\mu,\sigma^2)$.

2. 样本

定义 2 从总体 X 中，按一定的规则任意抽取的部分个体 $(X_1, X_2, \cdots, X_n)$ 称为来自总体 X 的一个样本，其中 n 称为样本容量或样本大小.

样本必须满足下列两条性质.

(1)代表性：每个 $X_i\ (i = 1, 2, \cdots, n)$ 与总体 X 同分布；

(2)独立性： $X_1, X_2, \cdots, X_n$ 相互独立.

满足上面两条性质的样本称为**简单随机样本**．在数理统计中，所有样本都必须是简单随机样本，通常也简称为样本.

3. 样本的联合分布

如果总体 X 为离散型随机变量，其分布律为

$$P\{X=a_i\}=p(a_i)\text{，}\quad i=1,2,\cdots,$$

则样本 $(X_1,X_2,\cdots,X_n)$ 的（联合）分布律为

$$P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$$

$$\overset{\text{独立性}}{=}\prod_{i=1}^{n}P\{X_i=x_i\}\overset{\text{同分布}}{=}\prod_{i=1}^{n}p(x_i)\text{，}$$

其中 $x_i\ (i=1,2,\cdots,n)$ 为样本值.

如果总体 X 为连续型随机变量，其密度函数为

$$f(x)，\quad -\infty < x < +\infty，$$

则样本 $(X_1, X_2, \cdots, X_n)$ 的（联合）密度函数为

$$f^*(x_1, x_2, \cdots, x_n) \overset{\text{独立性}}{=} f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

$$\overset{\text{同分布}}{=} f(x_1) f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i)，$$

其中 x_i $(i = 1, 2, \cdots, n)$ 为样本值.

例 3 设(X_1, X_2, X_3, X_4)为来自总体 $X \sim \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$的一个样本，求$P\{\min\limits_{1\le i\le 4} X_i \le 1\}$.

解 $P\{\min\limits_{1\le i\le 4} X_i \le 1\} = 1 - P\{\min\limits_{1\le i\le 4} X_i > 1\}$

$= 1 - P\{X_1 > 1, X_2 > 1, X_3 > 1, X_4 > 1\}$

$\overset{\text{独立性}}{=} 1 - P\{X_1 > 1\}P\{X_2 > 1\}P\{X_3 > 1\}P\{X_4 > 1\}$

$\overset{\text{同分布}}{=} 1 - (P\{X > 1\})^4 = 1 - (P\{X = 2\})^4 = 1 - (\frac{1}{3})^4 = \frac{80}{81}$.

二、统计量与样本矩

1. 统计量的概念

定义 3 设 $(X_1, X_2, \cdots, X_n)$ 为来自总体 X 的一个样本，$g(x_1, x_2, \cdots, x_n)$ 为一个 n 元函数，且 $g(x_1, x_2, \cdots, x_n)$ 不依赖总体 X 中的任何未知参数，就称随机变量 $g(X_1, X_2, \cdots, X_n)$ 为一个**统计量**.

如果 $(x_1, x_2, \cdots, x_n)$ 为样本观察值，也称 $g(x_1, x_2, \cdots, x_n)$ 为**统计量的观察值**.

例 4 设总体 $X \sim N(\mu,\sigma^2)$，其中 μ 已知，σ^2 未知，

则

$$g_1(X_1, X_2, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \ ,$$

$$g_2(X_1, X_2, \cdots, X_n) = \max_{1 \le i \le n} X_i - \mu \ ,$$

都是统计量，但由于

$$g_3(X_1, X_2, \cdots, X_n) = \frac{1}{\sigma^2}\sum_{i=1}^{n} (X_i - \mu)^2$$

中含有未知参数 σ^2，故该样本的函数不是统计量.

统计量是样本的函数．而在实际问题中，统计量表现为处理统计问题的方法．不同的统计量表示不同的处理方法，“好”的统计量表示“好”的处理方法（参见第七章中估计量的评价标准）．

例如，在比较某两个平行班的数学测验成绩中，通常采用取平均值的方法进行比较，这时，构造的统计量为$\frac{1}{n}\sum_{i=1}^{n} X_i$；

如果是数学竞赛，可选用统计量$\max\limits_{1\le i\le n} X_i$进行比较等等．

又如，像相声小品比赛，青年歌手大奖赛等均采用去掉最高分和最低分后，以剩下的分数之和作为选手最后得分的方法衡量选手的比赛水平，此时采用的统计量为

$$\sum_{i=1}^{n} X_i - \max_{1\le i\le n} X_i - \min_{1\le i\le n} X_i \text{ .}$$

2. 常见统计量

(1) 样本矩

定义 4 设$(X_1, X_2, \cdots, X_n)$为来自总体 X 的一个样本，定义

① **样本均值** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$；

② **样本方差** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$；

③ **样本标准差** $S = \sqrt{S^2}$；

④ **样本 k 阶原点矩** $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$， $k = 1, 2, \cdots$；

⑤ **样本 k 阶中心矩** $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$， $k = 1, 2, \cdots$.

例 5 设总体 X 的数学期望 $EX=\mu$，方差 $DX=\sigma^2$，

$(X_1, X_2, \cdots, X_n)$ $(n>1)$ 为来自总体 X 的一个样本，则

$E\overline{X}=\mu$， $D\overline{X}=\dfrac{\sigma^2}{n}$， $E(S^2)=\sigma^2$.（务必记住结论！）

证 本章第四章第二节例 5 已证 $E\overline{X}=\mu$， $D\overline{X}=\dfrac{\sigma^2}{n}$.

$$E(X_i^2)=DX_i+(EX_i)^2=\sigma^2+\mu^2,\quad i=1,2,\cdots,n,$$

$$E(\overline{X}^2)=D\overline{X}+(E\overline{X})^2=\frac{\sigma^2}{n}+\mu^2.$$

所以

$$E(S^2)=\frac{1}{n-1}[\sum_{i=1}^{n}E(X_i^2)-nE(\overline{X}^2)]$$

$$=\frac{1}{n-1}[\sum_{i=1}^{n}(\sigma^2+\mu^2)-n(\frac{\sigma^2}{n}+\mu^2)]=\sigma^2.$$

(2) 顺序统计量

定义 5 设$(X_1,X_2,\cdots,X_n)$为来自总体X的一个样本，将其排序为$X_1^*\le X_2^*\le\cdots\le X_n^*$，称$X_1^*,X_2^*,\cdots,X_n^*$为**顺序统计量**.

特别地

$$X_1^*=\min\{X_1,X_2,\cdots,X_n\},\quad X_n^*=\max\{X_1,X_2,\cdots,X_n\}.$$

结论： 设总体X的分布函数和密度函数分别为

$f(x)$, $F(x)$， 则X_n^*和X_1^*的分布函数和密度函数分别为

$$F_{X_n^*}(x)=[F(x)]^n,\quad f_{X_n^*}(x)=n[F(x)]^{n-1}f(x);$$

$$F_{X_1^*}(x)=1-[1-F(x)]^n,\quad f_{X_1^*}(x)=n[1-F(x)]^{n-1}f(x).$$