CH01 绪论

更新：2026 年 1 月 9 日

目录

1 数理统计的若干基本概念

1.1 总体和样本

定义 1.1 (总体) 一个统计问题所研究的对象的全体称为**总体**，总体中的每一个对象称为**个体**。总体的性质称为**总体特征**，用参数表示。在数理统计学中，总体可以用一个随机变量及其概率分布来描述。

定义 1.2 (样本) 从总体中抽取的一部分个体的集合称为**样本**，样本中的每一个个体称为**样本点**。样本的性质称为**样本特征**，用统计量表示。

1.2 样本空间和样本的两重性

定义 1.3 (样本空间) 设 X 为随机变量，X 的所有可能取值所构成的集合称为**样本空间**，记为 $\mathscr{X}$。

定义 1.4 (样本的两重性) 样本具有两重性：在实施抽样后，它是具体的数；在实施抽样前，它被看作随机变量（随机向量）。

定义 1.5 (简单随机样本) 设有一总体 F，$X_1, X_2, \cdots, X_n$ 为从总体 F 中抽取的容量为 n 的样本，若

(1) $X_1, X_2, \cdots, X_n$ 相互独立
(2) $X_1, X_2, \cdots, X_n$ 相同分布，即同分布于 F

则称 $X_1, X_2, \cdots, X_n$ 为**简单随机样本**，

2 统计量

定义 2.1 (统计量) 由样本算出的量称为**统计量**，即：统计量是样本的函数。

注 统计量只能与样本有关，不能含有任何未知参数。

2.1 常见的统计量

定义 2.2 (样本均值) 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的容量为 n 的样本，则

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

称为**样本均值**，它反映了总体均值的信息。

定义 2.3 (样本方差与样本标准差) 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的容量为 n 的样本，则

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

称为**样本方差**，它反映了总体方差的信息。而 S 称为**样本标准差**，它反映了总体标准差的信息。特别地，$\mathbb{E}(S^2) = \sigma^2 = \mathbb{D}(X)$。

命题 2.1 样本均值和样本方差是两个最常用的统计量，它们具有如下性质：

(1) $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$
(2) 设非零实数 a, b 为常数，作变换 $Y_i = aX_i + b$，则 $\overline{Y} = a\overline{X} + b, \quad S_Y^2 = a^2 S_X^2$
(3) 对于任何常数 c，有

$$\min_c \sum_{i=1}^{n}(X_i - c)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

即样本均值 $\overline{X}$ 是使 $\sum_{i=1}^{n}(X_i - c)^2$ 最小的常数。

定义 2.4 (样本矩) 设 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的容量为 n 的样本，则

$$a_{n,k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \cdots$$

称为**样本 k 阶原点矩**，特别地，$a_{n,1}=\overline{X}$。称

$$m_{n,k}=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k,\quad k=1,2,\cdots$$

为**样本 k 阶中心矩**，特别地，$m_{n,2}=(n-1)S^2/n$。

样本的原点矩和中心矩统称为**样本矩**。

定义 2.5 (二维随机向量的样本矩) 设 $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$ 为来自二维总体 (X,Y) 的容量为 n 的样本，则

$$\begin{aligned}\overline{X}&=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad S_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2,\\ \overline{Y}&=\frac{1}{n}\sum_{i=1}^{n}Y_i,\quad S_X^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2,\\ S_{XY}&=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\end{aligned}$$

分别称为**样本均值、样本方差**和**样本协方差**。

定义 2.6 (次序统计量) 设 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的容量为 n 的样本，记

$$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$$

为 $X_1,X_2,\cdots,X_n$ 的从小到大的排列，则称 $(X_{(1)},X_{(2)},\cdots,X_{(n)})$ 为样本的**次序统计量**。

定义 2.7 (样本变异系数) 设 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的容量为 n 的样本，S 为样本标准差，$\overline{X}$ 为样本均值，则称

$$\hat{\nu}=\frac{S_n}{\overline{X}}$$

为样本的**变异系数**。

定义 2.8 (样本偏度) 设 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的容量为 n 的样本，S 为样本标准差，$\overline{X}$ 为样本均值，则称

$$\hat{\beta}_1=\frac{m_{n,3}}{m_{n,2}^{3/2}}=\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)^{3/2}}$$

为样本的**偏度**。它反映了总体偏度的信息，总体偏度的定义为 $\beta_1=\mu_3/\mu_2^{3/2}$。此处，$\mu_i,i=2,3$ 分别为总体的 i 中心距。

定义 2.9 (样本峰度) 设 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的容量为 n 的样本，S 为样本标准差，$\overline{X}$ 为样本均值，则称

$$\hat{\beta}_2=\frac{m_{n,4}}{m_{n,2}^2}-3=\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)^2}-3$$

为样本的**峰度**。它反映了总体峰度的信息，总体峰度的定义为 $\beta_2=\mu_4/\mu_2^2-3$。此处，$\mu_i,i=2,4$ 分别为总体的 i 中心距。

2.2 经验分布函数

定义 2.10 (经验分布函数) 设 $X_1, X_2, \cdots, X_n$ 为来自总体 $F(x)$ 的容量为 n 的独立同分布样本，将其按大小排序为 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$，对任意的实数 $x \in \mathbb{R}$ 称下列函数：

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

为经验分布函数。

易见经验分布函数是单调、非降、左连续函数，具有分布函数的基本性质。若记示性函数

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

则 $F_n(x)$ 可表示为

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x)}(X_i)$$

注 经验分布函数 $F_n(x)$ 是样本的函数，因此也是统计量。它可能取值为 $\{0, 1/n, 2/n, \cdots, 1\}$ 中的某一值，因此 $F_n(x)$ 是离散型随机变量。

若记 $Y_i = I_{(-\infty, x)}(X_i)$ 则 $P(Y_i = 1) = F(x), P(Y_i = 0) = 1 - F(x)$，且 $Y_1, Y_2, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} b(1, F(x))$ Bernoulli 分布。故

$$nF_n(x) = \sum_{i=1}^{n} Y_i \sim b(n, F(x))$$

因此对 $k = 0, 1, \cdots, n$ 有

$$P(F_n(x) = \frac{k}{n}) = P\left(\sum_{i=1}^{n} Y_i = k\right) = \binom{n}{k}[F(x)]^k[1 - F(x)]^{n-k}$$

从而

$$\mathbb{E}(F_n(x)) = F(x), \quad \mathbb{D}(F_n(x)) = \frac{F(x)[1 - F(x)]}{n}$$

命题 2.2 利用二项分布的性质，可知对任一固定的 $x \in \mathbb{R}$，$F_n(x)$ 具有如下大样本性质：

(1) 由中心极限定理，当 $n \to \infty$ 时

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F(x)[1 - F(x)]}} \xrightarrow{\mathscr{L}} N(0, 1)$$

其中 $\xrightarrow{\mathscr{L}}$ 表示依分布收敛。

(2) 由大数定律，当 $n \to \infty$ 时

$$F_n(x) \xrightarrow{P} F(x)$$

其中 $\xrightarrow{P}$ 表示依概率收敛。

(3) 由 Borel 强大数定律，当 $n \to \infty$ 时

$$F_n(x) \xrightarrow{a.s.} F(x)$$

其中 $\xrightarrow{a.s.}$ 表示几乎必然收敛，即

$$P\left(\lim_{n\to\infty} F_n(x) = F(x)\right) = 1$$

定理 2.3 (Glivenko-Cantelli 定理) 设 r.v. $X_1, X_2, \cdots, X_n$ 为来自总体 $F(x)$ 的容量为 n 的简单随机样本，$F_n(x)$ 为经验分布函数，记 $D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ 则

$$D_n \xrightarrow{a.s.} 0$$

即

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

其中 sup 表示上确界。