《数理统计》

第六章　数理统计的基本概念

数理统计的分类

描述统计学——

对随机现象进行观测、试验，以取得有代表性的观测值

推断统计学——

对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性

推断统计学

- 数参估计 (第七章)
- 假设检验 (第八章)
- 方差分析 (第九章)
- 回归分析 (第九章)

共同获得

2003年诺贝尔经济学奖

20 世纪 80 年代两位获奖者发明了新的统计方法来处理许多经济时间数列中两个关键属性:

随时间变化的 $\left\{\begin{array}{l}\text{易 变 性} \\ \text{非稳定性}\end{array}\right.$

研究方向主要是

利率、汇率和期权的金融计量分析

提出谱分析回归等创新性统计方法

的研究涉及

统计和经济计量学　特别是

时间序列分析、预测、金融、人口统计学、方法论等领域.

§ 6.1 基本概念

- 总体和样本

总体 —— 研究对象全体元素组成的集合

所研究的对象的某个(或某些)数量指标的全体,它是一个随机变量(或多维随机变量). 记为X.

X 的分布函数和数字特征称为总体的分布函数和数字特征.

个体 —— 组成总体的每一个元素

即总体的每个数量指标，可看作随机变量 X 的某个取值. 用 X_i 表示.

样本 —— 从总体中抽取的部分个体.

用 $(X_1, X_2, \cdots, X_n)$ 表示，n 为样本容量.

称 $(x_1, x_2, \cdots, x_n)$ 为总体 X 的一个容量为 n 的样本观测值，或称样本的一个实现.

样本空间 —— 样本所有可能取值的集合.

简单随机样本

若总体 X 的样本 $(X_1, X_2, \cdots, X_n)$ 满足:

(1) $X_1, X_2, \cdots, X_n$ 与 X 有相同的分布

(2) $X_1, X_2, \cdots, X_n$ 相互独立

则称 $(X_1, X_2, \cdots, X_n)$ 为简单随机样本.

一般, 对有限总体, 放回抽样所得到的样本为简单随机样本, 但使用不方便, 常用不放回抽样代替. 而代替的条件是

$$N / n \geq 10.$$

总体中个体总数 | 样本容量

设总体 X 的分布函数为$F(x)$, 则样本 $(X_1, X_2, \cdots, X_n)$ 的联合分布函数为

$$\boxed{F_{总}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} F(x_i)}$$

若总体X 的密 d. f. 为 $f(x)$, 则样本的联合 d. f. 为

$$\boxed{f_{总}(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i)}$$

例如　设某批产品共有N个，其中的次品数为M，其次品率为

$$p = M / N$$

若p是未知的，则可用抽样方法来估计它.

从这批产品中任取一个产品，用随机变量X来描述它是否是次品：

$$X = \begin{cases} 1, & \text{所取的产品是次品} \\ 0, & \text{所取的产品不是次品} \end{cases}$$

X服从参数为p的0−1分布，可用如下表示方法：　$f(x,p) = p^x(1-p)^{1-x}, \quad x = 0,1$

设有放回地抽取一个容量为 n 的样本

$$(X_1, X_2, \cdots, X_n)$$

其样本值为 $(x_1, x_2, \cdots, x_n)$

样本空间为

$$\{(x_1, x_2, \cdots, x_n) \mid x_i = 0,1,\ i = 1,2,\cdots,n\}$$

$(X_1, X_2, \cdots, X_n)$ 的联合分布为

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

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}$$

若抽样是无放回的，则前次抽取的结果会影响后面抽取的结果. 例如

$$P(X_2=1 \mid X_1=1)=\frac{M-1}{N-1}=\frac{p-\frac{1}{N}}{1-\frac{1}{N}}\xrightarrow{N\to\infty}p$$

$$P(X_2=1 \mid X_1=0)=\frac{M}{N-1}=\frac{p}{1-\frac{1}{N}}\xrightarrow{N\to\infty}p$$

所以，当样本容量 n 与总体中个体数目 N 相比很小时，可将无放回抽样近似地看作放回抽样.

统计量

定义

设$(X_1, X_2, \cdots, X_n)$ 是取自总体X的一个样本，$g(r_1, r_2, \cdots, r_n)$ 为一实值连续函数，且不含有未知参数，则称随机变量$g(X_1, X_2, \cdots, X_n)$为**统计量**.

若$(x_1, x_2, \cdots, x_n)$是一个样本值，称 $g(x_1, x_2, \cdots, x_n)$ 为统计量$g(X_1, X_2, \cdots, X_n)$的一个样本值

例 $X \sim N(\mu,\sigma^2)$，μ,σ^2 是未知参数,

$(X_1,X_2,\cdots,X_n)$ 是一样本, 则

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

是统计量，其中 $X_i \sim N(\mu,\sigma^2)$

但 $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ 不是统计量.

若 μ , σ 已知, 则为统计量

常用的统计量

设 $(X_1, X_2, \cdots, X_n)$ 是来自总体 X 的容量为 n 的样本, 称统计量

(1) $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ 为样本均值

(2) $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ 为样本方差

$S = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$ 为样本标准差

(3) $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ 为样本的k 阶原点矩

(4) $B_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k$ 为样本的k 阶中心矩

例如

$$A_1 = \overline{X}$$

$$B_2 = \frac{n-1}{n}S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \triangleq S_n^2$$

(5) 顺序统计量与极差

设$(X_1, X_2, \cdots, X_n)$为样本,

$(x_1, x_2, \cdots, x_n)$ 为样本值, 且$x_1^* \le x_2^* \le \cdots \le x_n^*$

当$(X_1, X_2, \cdots, X_n)$取值为$(x_1, x_2, \cdots, x_n)$ 时,

定义 r. v. $\quad X_{(k)} = x_k^*, k = 1,2, \cdots, n$

则称统计量$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$为**顺序统计量**.

其中,$X_{(1)} = \min\limits_{1 \le k \le n} \{X_k\}, X_{(n)} = \max\limits_{1 \le k \le n} \{X_k\}$

称 $D_n = X_{(n)} - X_{(1)}$ 为**极差**

注 样本方差 $\boxed{S^2}$ 与样本二阶中心矩 $\boxed{S_n^2}$ 的不同

1） 关系式 $S^2 = \dfrac{n}{n-1} S_n^2$

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

$$= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n}\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2$$

$$= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = n(A_2 - \overline{X}^2)$$

故 $B_2 = A_2 - \overline{X}^2$ $\quad S^2 = \dfrac{n}{n-1}(A_2 - \overline{X}^2) = \dfrac{n}{n-1} S_n^2$

2） $E(S_n^2)=\dfrac{n-1}{n}\sigma^2$　　$E(S^2)=\sigma^2$

推导 设 $E(X)=\mu$, $D(X)=\sigma^2$　　则

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

$$E(S_n^2)=EA_2-E\left(\overline{X}^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-\left[D\left(\overline{X}\right)+E^2\left(\overline{X}\right)\right]$$

$$=\sigma^2+\mu^2-\left(\frac{1}{n}\sigma^2+\mu^2\right)=\frac{n-1}{n}\sigma^2$$

$$E(S^2)=E\left[\frac{n}{n-1}S_n^2\right]=\frac{n}{n-1}ES_n^2=\sigma^2$$

例1 从一批机器零件毛坯中随机地抽取10件，测得其重量为(单位：公斤)：

210, 243, 185, 240, 215,

228, 196, 235, 200, 199

求这组样本值的均值、方差、二阶原点矩与二阶中心矩.

解 令 $(x_1, x_2, \cdots, x_{10})$

$$=(210, 243, 185, 240, 215, 228, 196, 235, 200, 199)$$

则 $\bar{x}=\frac{1}{10}(230+243+185+240+215$

$+228+196+235+200+199)$

$=217.19$

$$s^2=\frac{1}{9}\sum_{i=1}^{10}(x_i-\bar{x})^2=433.43$$

$$A_2=\frac{1}{10}\sum_{i=1}^{10}x_i^2=47522.5$$

$$B_2=\frac{9}{10}s^2=\frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2=390.0$$

例2 在总体 $N(52, 6.3^2)$中，随机抽取一个容量为36的样本，求样本均值 $\overline{x}$ 落在50.8到53.8之间的概率.

解 $\overline{X} \sim N(52,\ 6.3^2/36)$

故 $P(50.8 < \overline{X} < 53.8)$

$$=\Phi\left(\frac{53.8-52}{6.3/6}\right)-\Phi\left(\frac{50.8-52}{6.3/6}\right)$$

$$=\Phi(1.7143)-\Phi(-1.1429)$$

$$=0.8239$$

例3 设总体X的概率密度函数为

$$f(x)=\begin{cases}|x| & |x|<1\\ 0 & |x|\ge 1\end{cases}$$

$(X_1, X_2, \cdots, X_{50})$ 为总体的样本, 求

(1) $\overline{X}$ 的数学期望与方差 (2) $E(S^2)$

(3) $P(|\overline{X}|>0.02)$

解 (1) $E(\overline{X})=E(X)=\int_{-1}^{1} x|x|\mathrm{d}x=0$

$$D(\overline{X})=\frac{1}{50}D(X)=\frac{1}{50}E(X^2)$$

$$=\frac{1}{50}2\int_{0}^{1} x^2|x|\mathrm{d}x=\frac{1}{100}$$

(2) $E(S^2) = D(X) = E(X^2) = 1/2.$

(3) $\overline{X} \overset{\text{近似}}{\sim} N(0,0.01)$ 由中心极限定理

$$P(|\overline{X}| > 0.02) = 1 - P(|\overline{X}| \le 0.02)$$

$$= 2\left(1 - \Phi\left(\frac{0.02-0}{0.1}\right)\right) = 2(1 - \Phi(0.2))$$

$$= 0.8414$$

作业 P. 202 习题六

2　　5

6　　7