第七章 数理统计的基本概念

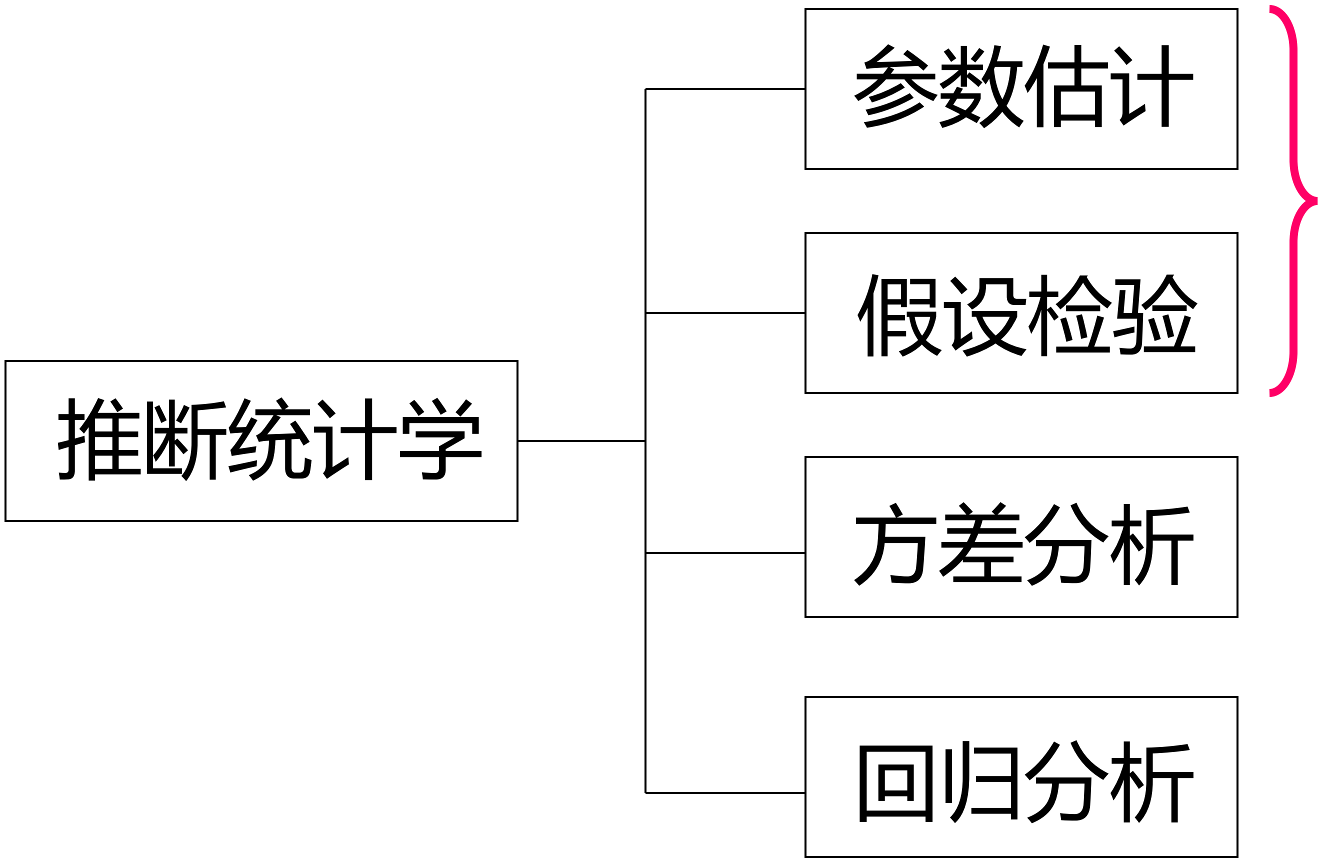
数理统计的分类

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
参数估计
假设检验
方差分析
回归分析

§ 7.1 基本概念

总体和样本

总体 —— 研究对象的全体组成的集合

一般地，研究对象全体的某个(或某些)数量指标,是一个随机变量(或多维随机变量).

若为一个随机变量,可记为X. 例如, 某钢铁厂生产的钢锭的强度.

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(单元)

总体中每个元素的数量指标,可以看作随机变量 X 的某个取值.用x_i表示.

抽样 —— 从总体中抽取个体，做随机试验并记录其结果

样本 ——从总体中抽取的部分个体.

假设抽取n个个体，则每一个体的数量指标分别为一个随机变量，记为：$X_1, X_2,\ldots ,X_n$，则$(X_1, X_2,\ldots ,X_n)$为样本变量，简称样本.

n 称为样本容量.

样本值 ——
从总体中抽取的部分个体.其数量指标的观察值$(x_1, x_2, \dots, x_n)$，称为总体 X 的一个容量为n的样本值.

或称为样本$(X_1, X_2, \dots, X_n)$的一个实现.

样本空间 —— 样本$(X_1, X_2, \dots, X_n)$所有可能取值的集合.

简单随机样本

设$(X_1, X_2, \ldots ,X_n)$是来自总体 X 的一个样本,它满足:

(1) 同分布：$X_1, X_2, \ldots ,X_n$都与X 有相同的分布

(2) 独立性: $X_1, X_2, \ldots ,X_n$相互独立

则称$(X_1, X_2, \ldots ,X_n)$为简单随机样本.

一般地,对有限总体,采用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.当总体中个体的数目N 与样本容量 n 之比$N / n \geq 10$时,可将不放回抽样近似地看作放回抽样.

设总体 X 的分布函数为$F(x)$, $(X_1, X_2, \ldots, X_n)$为总体 X 的简单随机样本,

则$(X_1, X_2, \ldots, X_n)$的联合分布函数为

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

若总体X的概率密度函数为$f(x)$,

则$(X_1, X_2, \ldots, X_n)$的联合概率密度函数为

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

统计量

定义

设$(X_1, X_2, \ldots, X_n)$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$g(X_1, X_2, \ldots, X_n)$为**统计量**.

若$(x_1, x_2, \ldots, x_n)$是一个样本值,

则称 $g(x_1, x_2, \cdots, x_n)$

为统计量$g(X_1, X_2, \ldots, X_n)$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 $\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$ 不是统计量.

若 μ ,σ 已知,则为统计量

常用的统计量

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

称统计量

(1) $$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$ 为**样本均值**

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

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

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

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

例如

$$A_1 = \overline{X}$$

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

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

注 样本方差 S^2 与样本二阶中心矩 S_n^2 不同:

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

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

1) **关系式** $S^2 = \dfrac{n}{n-1} S_n^2$

2) 设 $E(X)=\mu$, $D(X)=\sigma^2$ 则

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

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

推导如下：

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

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

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

则

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

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

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

$$EX_i^2$$

$$=E^2(X_i)+DX_i$$

$$=\sigma^2+\mu^2$$

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

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

(5) **顺序统计量与极差**

设$(X_1, X_2,\ldots ,X_n)$为样本,

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

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

定义随机变量 $X_{(k)} = x_k^*, k = 1, 2, \cdots, n$

则称统计量$(X_{(1)}, X_{(2)},\ldots ,X_{(n)})$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**

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$(x_1, x_2,\dots ,x_n)$
$=(210，243，185， 240，215，$
$228，196， 235，200，199)$

则 $$\bar{x} = \frac{1}{10}(230 + 243 + 185 + 240 + 215$$
$$+228 + 196 + 235 + 200 + 199)$$
$$= 217.19$$

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

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

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

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

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

故 $P(50.8 < \overline{X} < 53.8) = F_{\overline{X}}(53.8) - F_{\overline{X}}(50.8)$

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

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

$$= 0.8239$$

例3 设总体X的概率密度函数为

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

$(X_1, X_2,\dots ,X_{50})$为总体的样本,求

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

(3) 利用中心极限定理，估计$P(|\overline{X}|>0.02)$

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

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

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

(2) $$E(S^2) = D(X)$$

$$= E(X^2) - E^2(X)$$

$$= \frac{1}{2} - 0$$

$$= \frac{1}{2}$$

(3) $\overline{X} \sim N(0, \frac{1}{100})$ (近似)， 由中心极限定理

$$
\begin{aligned}
P(\left|\overline{X}\right| > 0.02) &= 1 - P(\left|\overline{X}\right| \le 0.02) \\
&= 1 - P(-0.02 \le \overline{X} \le 0.02) \\
&= 1 - \left[P(\overline{X} \le 0.02) - P(\overline{X} \le -0.02) \right] \\
&= 1 - \left\{ P(\overline{X} \le 0.02) - \left[1 - P(\overline{X} \le 0.02) \right] \right\} \\
&= 2\left(1 - P(\overline{X} \le 0.02) \right) \\
&= 2\left(1 - \Phi\left(\frac{0.02 - 0}{0.1} \right) \right) \\
&= 0.8414
\end{aligned}
$$