第五章 数理统计的基本概念与抽样分布

第一节 基本概念

- 一、总体与个体
- 二、随机样本的定义
- 三、统计量
- 四、内容小结

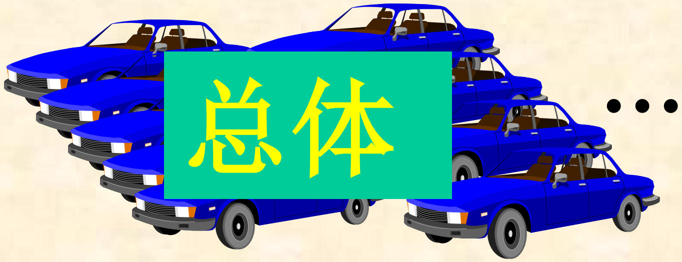
一、总体与个体

一个统计问题总有它明确的研究对象.

研究对象的全体称为总体(母体),

总体中每个成员称为个体.

研究某批灯泡的质量



考察国产 轿车的质量

然而在统计研究中，人们关心总体仅仅是关心其每个个体的一项(或几项)数量指标和该数量指标在总体中的分布情况. 这时，每个个体具有的数量指标的全体就是总体.

该批灯泡寿命的全体就是总体

所有国产轿车每公里耗油量的全体就是总体

总体可以用一个随机变量来表示

考察某大学一年级学生的年龄

某大学一年级全体学生的年龄构成问题的总体

设该大学一年级学生的年龄分布如下表

年龄	18	19	20	21	22
比例	0.5	0.3	0.1	0.07	0.03

若从该大学一年级学生中任意抽查一个学生的年龄，所得结果为一随机变量，记作X.

X的概率分布是：

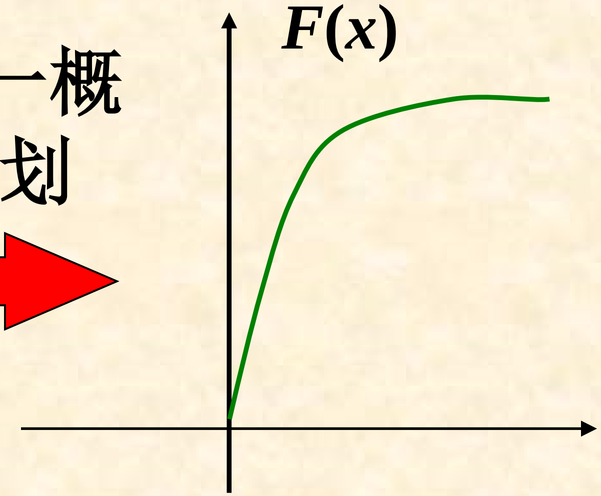
X	18	19	20	21	22
p	0.5	0.3	0.1	0.07	0.03

可见，X的概率分布反映了总体中各个值的分布情况. 很自然地，我们就用随机变量X来表示所考察的总体.

也就是说，总体可以用一个随机变量及其分布来描述.

又如:研究某批灯泡的寿命时，关心的数量指标就是寿命，那么，此总体就可以用随机变量X表示，或用其分布函数$F(x)$表示.

寿命X可用一概率分布来刻划



鉴于此，常用随机变量的记号或用其分布函数表示总体. 如说总体X或总体$F(x)$.

有限总体和无限总体

实例　某工厂10月份生产的灯泡寿命所组成的总体中, 个体的总数就是10月份生产的灯泡数, 这是一个有限总体; 而该工厂生产的所有灯泡寿命所组成的总体是一个无限总体, 它包括以往生产和今后生产的灯泡寿命.

当有限总体包含的个体的总数很大时, 可近似地将它看成是无限总体.

二、随机样本的定义

1. 样本的定义

从总体X中，随机地抽取n个个体：

$$X_1, X_2, \cdots, X_n$$

称为总体X的一个样本，记为

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

样本中所包含个体的总数n称为样本容量.

注　样本 $(X_1, X_2, \cdots, X_n)$ 是一个n 维随机变量

2. 样本值

每一次抽取 $X_1, X_2, \cdots, X_n$ 所得到的n个确定的具体数值，记为

$$(x_1, x_2, \cdots, x_n)$$

称为样本 $(x_1, x_2, \cdots, x_n)$ 的一个样本值(观察值).

3. 简单随机样本

若来自总体 X的样本 $(X_1, X_2, \cdots, X_n)$ 具有下列两个特征:

(1) 代表性： $X_1, X_2, \ldots, X_n$中每一个与所考察的总体有相同的分布.

(2) 独立性： $X_1, X_2, \ldots, X_n$是相互独立的随机变量.

则称 $(X_1, X_2, \cdots, X_n)$ 是来自总体X 容量为n 的**简单随机样本**.

获得简单随机样本的抽样方法称为简单随机抽样.

总体和样本的数学严格定义：

定义5.1 一个随机变量 X 或其相应的分布函数 $F(x)$称为一个总体.

定义5.2 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$F(x)$、相互独立的随机变量，则称$X_1, X_2, \cdots, X_n$为来自总体X的容量为n的简单随机样本，简称样本.

4. 样本的分布

定理5.1 设$(X_1, X_2, \cdots, X_n)$为来自总体X的样本.

(1)若总体X的分布函数为$F(x)$，则样本

$(X_1, X_2, \cdots, X_n)$的分布函数为$\prod_{i=1}^{n} F(x_i)$.

(2)若总体X的分布密度为$p(x)$,则样本$(X_1,X_2,\cdots,X_n)$的分布密度为$\prod_{i=1}^{n} p(x_i)$.

(3)若总体X的分布率为$P\{X=x_i^*\}=p(x_i^*)(i=1,2,\cdots)$,则样本$(X_1,X_2,\cdots,X_n)$的分布率为$\prod_{i=1}^{n} p(x_i)$.

例1 设总体 X 服从参数为 $\lambda\ (\lambda>0)$ 的指数分布, $(X_1, X_2, \cdots, X_n)$ 是来自总体的样本, 求样本 $(X_1, X_2, \cdots, X_n)$ 的概率密度.

解 总体 X 的概率密度为 $p(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & x\le 0\end{cases}$

因为 $X_1, X_2, \cdots, X_n$ 相互独立, 且与 X 有相同的分布,

所以 $(X_1, X_2, \cdots, X_n)$ 的概率密度为

$$p(x_1, x_2, \cdots, x_n)=\prod_{i=1}^{n} p(x_i) = \begin{cases}\lambda^n e^{-\lambda\sum\limits_{i=1}^{n} x_i}, & x_i>0 \quad (i=1,2,\cdots,n)\\ 0, & \text{其它}\end{cases}$$

目录 上页 下页 返回 结束

例2 设总体 X 服从两点分布 $B(1, p)$, 其中 $0 < p < 1$, $(X_1, X_2, \cdots, X_n)$ 是来自总体的样本, 求样本 $(X_1, X_2, \cdots, X_n)$ 的分布律.

解 总体 X 的分布律为

$$P\{X = i\} = p^i (1-p)^{1-i} \qquad (i = 0, 1)$$

因为 $X_1, X_2, \cdots, X_n$ 相互独立,

且与 X 有相同的分布,

所以 $(X_1, X_2, \cdots, X_n)$ 的分布律为

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

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

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

其中 $x_1, x_2, \cdots, x_n$ 在集合$\{0,1\}$中取值.

三、统计量

由样本推断总体情况,需要对样本值进行“加工”,这就需要构造一些样本的函数,它把样本中所含的信息集中起来.

1. 统计量的定义5.3

设$(X_1, X_2, \cdots, X_n)$是来自总体 X 的一个样本, $f(X_1, X_2, \cdots, X_n)$是$X_1, X_2, \cdots, X_n$的函数, 若 f 中不含任何关于总体X的未知参数, 则称 $f(X_1, X_2, \cdots, X_n)$ 是一个统计量.

设$(x_1,x_2,\cdots,x_n)$是样本

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

的样本值, 则称$f(x_1,x_2,\cdots,x_n)$是

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

的观察值.

用于估计分布中参数的统计量，称为**估计量.**

注 1° 统计量 $f(X_1,X_2,\cdots,X_n)$是**随机变量**;

2°统计量用于统计推断，故**不应含**任何关于总体X的**未知参数**.

例1 设 X_1, X_2, X_3是来自总体$N(\mu,\sigma^2)$的一个样本, 其中μ为已知, σ^2 为未知, 判断下列各式哪些是统计量, 哪些不是?

$$T_1 = X_1, \qquad T_4 = \max(X_1, X_2, X_3),$$

$$T_2 = X_1 + X_2 e^{X_3}, \qquad T_5 = X_1 + X_2 - 2\mu,$$

$$T_3 = \frac{1}{3}(X_1 + X_2 + X_3),$$

是

$$T_6 = \frac{1}{\sigma^2}(X_1{}^2 + X_2{}^2 + X_3{}^2).$$

不是

2. 几个常用统计量的定义

(1) 样本矩

设 $X_1, X_2, \cdots, X_n$ 是来自总体的一个样本,
$x_1, x_2, \cdots, x_n$ 是这一样本的观察值.

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

其观察值 $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i.$

它反映了总体均值的信息

可用于推断：$E(X)$.

2) 样本方差

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

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

它反映了总体方差的信息

可用于推断：$D(X)$.

其观察值

$$s_n^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

3)样本标准差

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

其观察值

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\,.$$

4)修正样本方差

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

其观察值

$$s_n^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right).$$

样本方差与修正样本方差的关系：

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{n-1}{n}S_n^{*2}.$$

注　1°当n较大时，S_n^{*2}与S_n^2差别微小；

2°当n较小时，S_n^{*2}比S_n^2有更好的统计性质.

5) 样本 k 阶(原点)矩

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k,\ k=1,2,\cdots;$$ 特例：$A_1=\overline{X}$

其观察值 $$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,\ k=1,2,\cdots.$$

6)样本 k 阶中心矩

特例：$B_2=S_n^2$

$$B_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^k,\ k=2,3,\cdots;$$

其观察值 $$b_k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k,\ k=2,3,\cdots.$$

样本矩具有下列性质：

性质5.1 设总体X的期望$E(X)=\mu$,方差$D(X)=\sigma^2$, $(X_1,X_2,\cdots,X_n)$为来自总体X的样本, 则有

(1) $E(\overline{X})=\mu$;

(2) $D(\overline{X})=\dfrac{1}{n}\sigma^2$;

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

(4) $E(S_n^{*2})=\sigma^2$.

证 (1) $E(\bar{X}) = \mu$

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

(2) $D(\bar{X}) = \frac{1}{n}\sigma^2$

$$D(\bar{X}) = D(\frac{1}{n}\sum_{i=1}^{n} X_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n}\sigma^2.$$

$$(3)\ E(S_n^2)=\frac{n-1}{n}\sigma^2$$

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

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

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

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

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

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

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

性质5.2 若总体 X 的 k 阶矩 $E(X^k) = \mu_k$ 存在，

则当 $n \to \infty$ 时，$A_k \xrightarrow{P} \mu_k, k = 1, 2, \cdots.$

证 因为 $X_1, X_2, \cdots, X_n$ 独立且与 X 同分布，

所以 $X_1^k, X_2^k, \cdots, X_n^k$ 独立且与 X^k 同分布，

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

再根据第四章辛钦定理知，

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k, \quad (n \to \infty), \quad k = 1, 2, \cdots;$$

由第四章关于依概率收敛的序列的性质知

$$\boxed{g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(\mu_1, \mu_2, \cdots, \mu_k),}$$

其中g是连续函数.

以上结论是下一章所要介绍的矩估计法的理论根据.

(2) 次序统计量

定义 设$(X_1, X_2, \cdots, X_n)$是从总体X中抽取的一个样本,$(x_1, x_2, \cdots, x_n)$是其一个观测值,将观测值按由小到大的次序重新排列为

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

当$(X_1, X_2, \cdots, X_n)$取值为$(x_1, x_2, \cdots x_n)$时,定义$X_{(k)}$取值为$x_{(k)}(k = 1,2,\cdots n)$, 由此得到

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

称为样本$(X_1, X_2, \cdots, X_n)$的次序统计量.

对应的$(x_{(1)}, x_{(2)}, \cdots x_{(n)})$称为其观测值.

$X_{(k)}$： 样本$(X_1, X_2, \cdots, X_n)$的第k个次序统计量.

特别地， $X_{(1)} = \min\limits_{1 \le i \le n} X_i$ 称为最小次序统计量.

$X_{(n)} = \max\limits_{1 \le i \le n} X_i$ 称为最大次序统计量.

注 由于每个$X_{(k)}$都是样本$(X_1, X_2, \cdots, X_n)$的函数,所以,$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$也都是随机变量,并且它们一般不相互独立.

定理5.2 设总体X的分布密度为$p(x)$(或分布函数为$F(x)$), $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$为总体$X$的样本$(X_1, X_2, \cdots, X_n)$的次序统计量. 则有

(1) 最大次序统计量$X_{(n)}$的分布密度为

$$p_{X_{(n)}}(x) = n[F(x)]^{n-1} p(x)$$

(2) 最小次序统计量$X_{(1)}$的分布密度为

$$p_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} p(x)$$

证 (1) $F_{X_{(n)}}(x) = P\{X_{(n)} \le x\}$

$$= P\{\max_{1\le i\le n} X_i \le x\}$$

$$= P\{X_1 \le x, X_2 \le x, \cdots, X_n \le x\}$$

$$= P\{X_1 \le x\} \cdot P\{X_2 \le x\} \cdot \cdots \cdot P\{X_n \le x\}$$

$$= F^n(x)$$

$$\therefore \quad p_{X_{(n)}}(x) = \frac{\mathrm{d}F_{X_{(n)}}(x)}{\mathrm{d}x} = nF^{n-1}(x) \cdot p(x)$$

(2) $F_{X_{(1)}}(x) = P\{X_{(1)} \le x\}$

$= P\{\min\limits_{1\le i\le n} X_i \le x\} = 1 - P\{\min\limits_{1\le i\le n} X_i > x\}$

$= 1 - P\{X_1 > x, X_2 > x, \cdots, X_n > x\}$

$= 1 - [1 - F(x)]^n$

$\therefore \quad p_{X_{(1)}}(x) = \dfrac{\mathrm{d}F_{X_{(1)}}(x)}{\mathrm{d}x}$

$= -n[1 - F(x)]^{n-1} \cdot [-F'(x)] = n[1 - F(x)]^{n-1} \cdot p(x).$

例2 设总体X服从区间$[0,\theta]$上的均匀分布，$(X_1, X_2, \cdots, X_n)$为总体X的样本，试求$X_{(1)}$和$X_{(n)}$的分布.

解 总体X的分布密度为 $p(x)=\begin{cases}\dfrac{1}{\theta}, & 0\le x\le\theta\\ 0, & \text{其他}\end{cases}$

X的分布函数为 $F(x)=\begin{cases}0, & x<0\\ \dfrac{x}{\theta}, & 0\le x\le\theta\\ 1, & x>\theta\end{cases}$

由定理5.2得$X_{(1)}$的分布密度为

$$p_{X_{(1)}}(x)=n[1-F(x)]^{n-1}p(x)$$

$$=\begin{cases}\dfrac{n}{\theta}(1-\dfrac{x}{\theta})^{n-1}, & 0\le x\le\theta\\ 0, & \text{其他}\end{cases}$$

而$X_{(n)}$的分布密度为

$$p_{X_{(n)}}(x)=nF^{n-1}(x)p(x)=\begin{cases}\dfrac{n}{\theta^n}x^{n-1}, & 0\le x\le\theta\\ 0, & \text{其他}\end{cases}$$

(3) 经验分布函数

定义5.5 设 $X_1, X_2, \cdots, X_n$是总体X的一个样本, $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$为总体$X$的样本$(X_1, X_2, \cdots, X_n)$的次序统计量. $(x_{(1)}, x_{(2)}, \cdots x_{(n)})$为其观测值, 设$x$是任一实数, 称函数

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ \dfrac{k}{n}, & x_{(k)} \le x < x_{(k+1)}, \\ 1, & x \ge x_{(n)}. \end{cases}$$

为总体X的**经验分布函数**，即对于任何实数x,，经验分布函数$F_n(x)$为样本值中不超过x的个数再除以n，亦即

$$F_n(x) = \frac{\mu_n(x)}{n}$$

其中$\mu_n(x)\,(-\infty < x < +\infty)$表示 $x_1, x_2, \cdots, x_n$ 中不超过于 x 的个数.

性质

(1)对于给定的一组样本值$(x_1, x_2, \cdots, x_n)$，$F_n(x)$满足分布函数的特征,是一个分布函数.

(2)由于$F_n(x)$是样本的函数,故$F_n(x)$是随机变量.可以证明$nF_n(x) \sim B(n, F(x))$,所以

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

(3)$F_n(x)$依概率收敛于$F(x)$. 即

$$\lim_{n\to\infty} P\{|F_n(x) - F(x)| < \varepsilon\} = 1 \qquad (\forall \varepsilon > 0)$$

格里汶科定理5.3

格里汶科

对于任一实数x,当$n\to\infty$时,$F_n(x)$以概率1一致收敛于分布函数$F(x)$,即

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

对于任一实数x当n充分大时,经验分布函数的任一个观察值$F_n(x)$与总体分布函数$F(x)$只有微小的差别,从而在实际上可当作$F(x)$来使用.

例3 设总体X具有一个样本值1, 2, 3,

则经验分布函数 $F_3(x)$的观察值为

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

例4 设总体X具有一个样本值1, 1, 2,

则经验分布函数$F_3(x)$的观察值为

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

四、内容小结

基本概念：个体　总体$\begin{cases}\text{有限总体}\\ \text{无限总体}\end{cases}$　随机样本

说明1　一个总体对应一个随机变量X, 我们将不区分总体和相应的随机变量, 统称为总体X.

说明2　在实际中遇到的总体往往是有限总体, 它对应一个离散型随机变量; 当总体中包含的个体的个数很大时, 在理论上可认为它是一个无限总体.

总体, 样本, 样本值的关系:

总体(理论分布)

?

样本 样本值

统计是从手中已有的资料--样本值,去推断总体的情况--总体的分布$F(x)$的性质.

样本是联系二者的桥梁.

总体分布决定了样本取值的概率规律,也就是样本取到样本值的规律,因而可以由样本值去推断总体.

两个最重要的统计量:

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

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