西北工業大學
NORTHWESTERN POLYTECHNICAL UNIVERSITY

概率论与数理统计

徐 爽

西北工业大学理学院概率统计教研室

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、问题的提出
- 二、总体与个体
- 三、随机样本的定义
- 四、统计量

一、问题的提出

概率论　假设：**研究对象的分布已知**

$$X \sim F(x), p(x), P(x = x_k) \Rightarrow EX, DX, P(A)$$

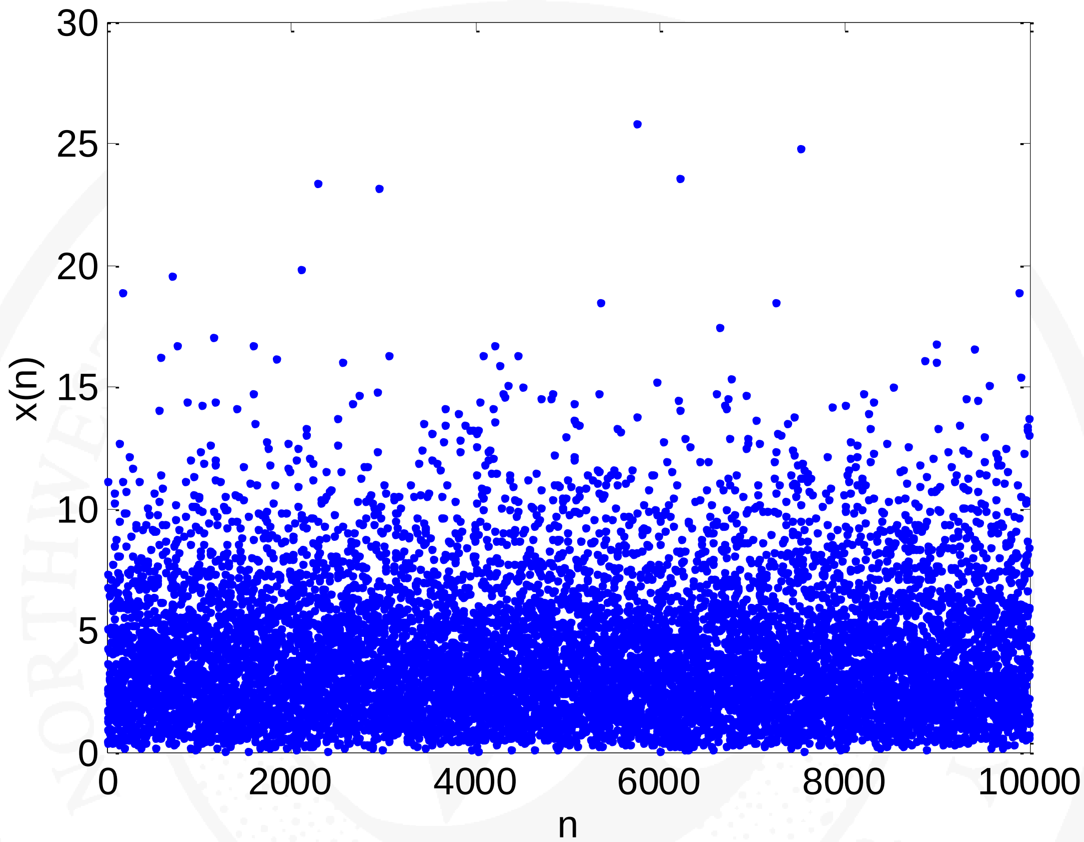
数理统计　实际：**研究对象的分布未知**

需要用已有的**部分信息**去**推断整体**情况。

$$(x_1, x_2, \cdots, x_n) \Rightarrow EX, DX, F(x)$$

什么是统计？

就是瞎子摸象



$$(x_1, x_2, \cdots, x_{10000}) \to EX, DX, F(x, \lambda)?$$

一叶落而知秋天下，一粒米见一世界

二、总体与个体

总体: 在数理统计中，把研究对象的全体称为总体（或母体）.

个体: 总体中每个研究对象称为个体.

在实际中，我们并不关心总体的各个方面，而往往关心它的**某项或几项数量指标**.

例如, 在考察我校某届本科生**学习质量**时，该届本科生的**全体成绩**称为**总体**，每一个本科生的成绩称为**个体**.

当我们说到**总体**，就是指数量指标（具有确定概率分布的随机变量）**可能取值**的全体。每一个可能的取值为个体。

总体 ←----→ 随机变量

定义5.1 一个随机变量或者其相应的分布函数 $F(x)$称为一个总体.

通常，我们用随机变量 $X, Y, Z \cdots$ 等表示**总体**.

三、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.

例 1 为了了解数学专业本科毕业生的月薪情况，调查了某地区100名2013届数学专业的本科生的月薪情况，试问

（1）什么是总体？

（2）什么是样本？

（3）样本容量是多少？

解 （1）总体是该地区2013届数学本科毕业生的月薪；

（2）样本是被调查的100名2013届数学本科毕业生的月薪；

（3）样本容量是100.

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).

■ 数理统计的**基本任务**是：根据从总体中抽取的样本，利用样本的信息推断总体的性质。

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: $X_1,X_2,\cdots,X_n$ 中每一个体与总体X有相同的分布.

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

则称 $(X_1,X_2,\cdots,X_n)$ 为n维简单随机样本.

获得简单随机样本的抽样方法称为简单随机抽样.

样本的严格数学定义：

定义5.2 设随机变量X的分布函数为$F(x)$，若$X_1, X_2, \cdots, X_n$是具有同一分布函数$F(x)$,且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

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

(2)若总体X的分布密度为$p(x)$,则样本

$$(X_1, X_2, \cdots, X_n)\text{的分布密度为}\prod_{i=1}^{n} p(x_i)$$

(3)若总体X的分布律为$P(X=x_i)=p(x_i)(i=1,\cdots,n)$

则样本$(X_1,X_2,\cdots,X_n)$的分布律为$\prod_{i=1}^{n}p(x_i)$.

样本的分布：

***n*维独立同分布随机变量的联合分布**

四、统计量 $(x_1, x_2, \cdots, x_{10000}) \to F(x, \lambda)$

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

定义5.3 设$(X_1 \cdots X_2, \cdots, X_n)$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$(X_1, X_2, \cdots, X_n)$的观察值

则称$f(x_1, x_2, \cdots, x_n)$是统计量 $f(X_1, X_2, \cdots, X_n)$的观察值

注 $1°$统计量$Y = f(X_1, X_2, \cdots, X_n)$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;

$3°$ 统计量是样本的函数，它是一个**随机变量**，统计量的分布称为**抽样分布**.

2. 几个常用统计量

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

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

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

它反映了总体均值的信息

3.9735， 3.9820， 3.9735…

可用于推断：$E(X)$.

2) 样本方差

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

它反映了总体方差的信息

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

其观察值

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

7.8544，7.7280, 7.8440

可用于推断：$D(X)$.

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**：

$$S_n^2=\frac{n-1}{n}S_n^{*2}.\quad \le S_n^{*2}$$

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$EX=\mu$,方差$DX=\sigma^2$,
$(X_1,X_2,\cdots,X_n)$为来自总体X的样本，则有

$$(1)\quad E(\bar{X})=E(X)=\mu$$

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

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

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

证 (1) $E(\overline{X})=\mu$

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

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

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

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

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

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

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

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

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

性质5.2 若总体X的k阶矩$E(X^k)=a_k$存在，

则当$n\to\infty$时，$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \xrightarrow{p} a_k, k=1,2,\cdots$

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

再根据第四章辛钦大数定理，即

若$r.v$ $X_1\cdots X_n$独立同分布，且$EX_k=\mu$,则

$\forall\varepsilon>0$ 有$\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{P} EX_i$

由上述定理可得

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{p} a_k\ , \quad k = 1, 2, \cdots;$$

$$\Rightarrow \overline{X} \xrightarrow{p} EX;$$

由第四章关于依概率收敛的序列的性质知

$$g(A_1, A_2, \cdots, A_k) \xrightarrow{P} g(a_1, a_2, \cdots, a_k),$$

其中g是连续函数.

$$\Rightarrow S_n^2 = A_2 - A_1^2 \xrightarrow{P} DX = a_2 - a_1^2;$$

注 性质5.2是下一章矩估计法的理论根据.

(2) 次序统计量

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，$X_{(1)},X_{(2)},\cdots,X_{(n)}$也是随机变量，但它们一般不相互独立.

$(X_1, X_2, X_3, X_4, X_5)$ $X_k \sim N(2,3)$

第1次抽样	-1.6225	4.1517	6.8907	3.4667	5.1041
第2次抽样	2.9756	-0.2648	6.1109	-3.1345	1.6933
第3次抽样	-1.4412	-1.2066	-0.4285	-6.8329	6.3151

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

-1.6225	3.4667	4.1517	5.1041	6.8907
-3.1345	-0.2648	1.6933	2.9756	6.1109
-6.8329	-1.4412	-1.2066	-0.4285	6.3151

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

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

证明思路：极值分布

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

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

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

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

$$= F^n(x)$$

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

(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.

注 1° $\mu_n(x)$ 为样本中不超过x的样本的最大个数，即在n次重复独立试验中，事件 $A=\{X\le x\}$ 发生的次数. $P(A)=F(x)$

$(\because\ x_{(1)}\le x_{(2)}\le\cdots\le x_{(k)}\le x$,有$\mu_n(x)$个样品的取值$\le x)$

$$2^\circ\ F_n(x)=\frac{\mu_n(x)}{n}$$ 为事件$\{X\le x\}$发生的频率.

事实上，令 $\mu_n(x)=\sum_{i=1}^{n}I_i$，其中

$$I_i=\begin{cases}1, & \{X_i\le x\}\text{发生}\\ 0, & \{X_i\le x\}\text{不发生}\end{cases}\sim B(1,F(x)),$$ 则$\mu_n(x)\sim B(n,F(x))$

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：$0 \le F_n(x) \le 1, F_n(-\infty)=0, F_n(+\infty)=1,$ 单调非降右连续，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$nF_n(x)=\sum^{n} I_i \sim B(n, F(x))$，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

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

例10 设从总体 X 中取得一个容量为5的样本，样本观测值为 -2，-1，2.5，3.1，3.7，试求此样本经验分布分布函数 $F(x)$.

解 由经验分布函数的定义知

$$F(x)=\begin{cases} 0 & x<-2 \\ 1/5 & -2\le x<-1 \\ 2/5 & -1\le x<2.5 \\ 3/5 & 2.5\le x<3.1 \\ 4/5 & 3.1\le x<3.7 \\ 1 & 3.7\le x \end{cases}$$

内容小结

基本概念： 总体 X (随机变量)

个体 $X_1, X_2, \cdots, X_n$ （随机变量）

（简单随机）样本 $(X_1, X_2, \cdots, X_n)$ （n维随机向量）

样本值 $(x_1, x_2, \cdots, x_n)$ （常量）

总体、样本、样本值的相互关系：

X 总体(理论分布)

估计

$(X_1, X_2, \cdots, X_n)$ 样本 $\longrightarrow$ 样本值 $(x_1, x_2, \cdots, x_n)$

统计量: $f(X_1, X_2, \cdots, X_n)$ 随机变量

观测值: $f(x_1, x_2, \cdots, x_n)$ 常量

1、样本矩

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

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

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

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

样本矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{p} \mu_k = E(X^k)$

2、次序统计量:

(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

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

3、经验分布函数:

$$F_n(x)=\frac{\mu_n(x)}{n}\xrightarrow{p,a.e.}F(x)$$

$$nF_n(x)=\mu_n(x)=\sum_{i=1}^{n}I_i\sim B(n,F(x))$$

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

5-1 基本概念

Thank You !