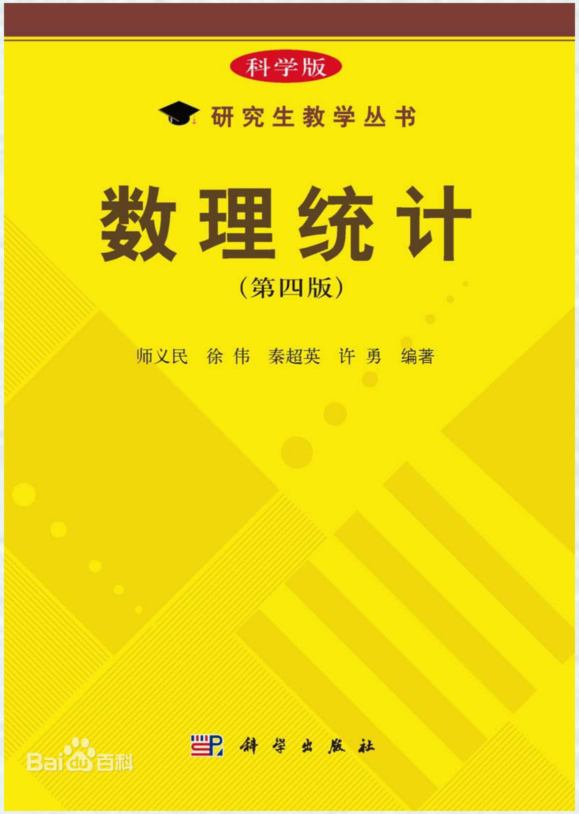
QQ群号：717167659



第一章　统计量与抽样分布

数理统计

数理统计可以分为两大类：

一类是如何科学地安排试验以获取有效的随机数据。

该类为**描述统计学**如：试验设计、抽样方法。

另一类是研究如何分析所获得的随机数据，对所研究的问题进行科学的、合理的估计和推断，尽可能地为采取一定的决策提供依据，作出精确而可靠的结论.

该类为**推断统计学**，如：参数估计、假设检验等。

例如　某厂生产一型号的合金材料，用随机的方法选取100个样品进行强度测试，于是面临下列几个问题：

1、估计这批合金材料的强度均值是多少?（**参数的点估计问题**）

2、强度均值在什么范围内？（**参数的区间估计问题**）

3、若规定强度均值不小于某个定值为合格，那么这批材料是否合格？（**参数的假设检验问题**）

4、这批合金的强度是否服从正态分布？（**分布检验问题**）

5、若这批材料是由两种不同工艺生产的，那么不同的工艺对合金强度有否影响？若有影响，哪一种工艺生产的强度较好？ (**方差分析问题**)

6、若这批合金由几种原料用不同的比例合成，那么如何表达这批合金的强度与原料比例之间的关系？

(**回归分析问题**)

我们依次讨论参数的点估计、区间估计、假设检验、方差分析、回归分析

第1.1节　基本概念

一、总体和样本

二、统计量和样本矩

三、经验分布函数

一、总体与样本

1. 总体

一个统计问题总有它明确的研究对象.

研究对象的全体元素组成的集合称为总体(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

所有国产轿车百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.

从另一方面看

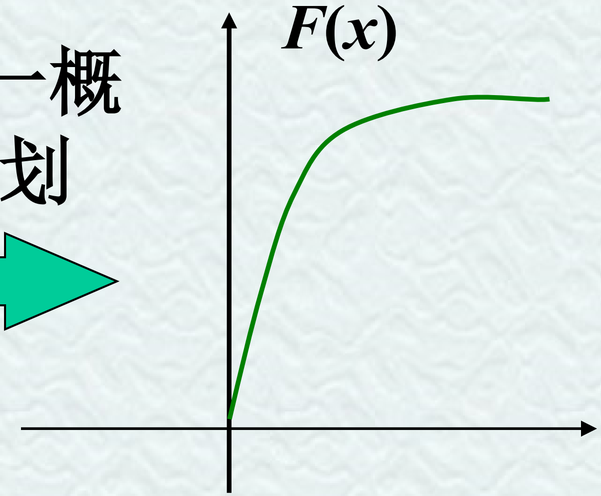
统计的任务，是根据从总体中抽取的样本，去推断总体的性质.

由于我们关心的是总体中的个体的某项指标(如人的身高、体重，灯泡的寿命，汽车的耗油量…)，所谓总体的性质，无非就是这些指标值的集体的性质.

而概率分布正是刻画这种集体性质的适当工具. 因此在理论上可以把总体与概率分布等同起来.

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$F(x)$.

2. 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

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

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.

3. 样本值

每次抽取$X_1, X_2, \cdots, X_n$ 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).

4. 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 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 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.

5. 样本的分布

定理1.1 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$f(x)$,则样本$(X_1, X_2, \cdots, X_n)$的分布密度为$\prod_{i=1}^{n} f(x_i)$.

(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 $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases}$

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

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

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

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

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

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

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

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

二、统计量与样本矩

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

定义1.1 设$(X_1, X_2, \cdots, X_n)^T$是来自总体$X$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$(x_1,x_2,\cdots,x_n)^T$是样本$(X_1,X_2,\cdots,X_n)^T$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3 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. 常用统计量—样本矩

(1) 样本矩

定义1.2 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, 则称

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

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)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

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

S_n^2 与 S_n^{*2} 的关系： $S_n^2=\frac{n-1}{n}S_n^{*2}.$

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．

4)样本标准差

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

其观察值

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

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

3、样本矩的性质

定理1.2 设总体X具有$2k$阶矩，则来自总体X的样本k阶原点距A_k的数学期望和方差分别为：

$$E(A_k)=\alpha_k$$

$$D(A_k)=\frac{\alpha_{2k}-\alpha_k^2}{n}$$

其中α_k＝$E(X^k)(k=1,2,\cdots,)$表示总体的k阶原点距.

证 $E(A_k)=E(\frac{1}{n}\sum_{i=1}^{n}X_i^k)$

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

$$D(A_k)=D(\frac{1}{n}\sum_{i=1}^{n}X_i^k)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i^k)=\frac{1}{n^2}\sum_{i=1}^{n}D(X^k)$$

$$=\frac{1}{n}((E(X^{2k})-E^2X^k)=\frac{\alpha_{2k}-\alpha_k^2}{n}$$

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

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

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

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

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

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

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

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

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

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

样本矩的极限性质

若总体的k阶原点矩存在, 则$A_k \xrightarrow{p} \alpha_k$,如果函数$g(x_1, x_2, \cdots, x_k)$连续,则$g(A_1, A_2, \cdots, A_k) \xrightarrow{p} g(\alpha_1, \alpha_2, \cdots, \alpha_k)$.

利用辛钦大数定理可以证明此性质. 同时也可以得到

$$\bar{X} \xrightarrow{P} E(X)$$

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = A_2 - A_1^2 \xrightarrow{P} \alpha_2 - \alpha_1^2 = \sigma^2$$

三、经验分布函数

1、次序统计量

设$(X_1,X_2,\cdots,X_n)^T$是从总体X中抽取的一个样本，$(x_1,x_2,\cdots,x_n)^T$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$(X_1,X_2,\cdots,X_n)^T$取值为$(x_1,x_2,\cdots x_n)^T$时，定义$X_{(k)}$取值为$x_{(k)}(k=1,2,\cdots n)$，由此得到

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

称为样本 $(X_1,X_2,\cdots,X_n)^T$ 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.

2、次序统计量的分布

设总体X的分布密度为$f(x)$(或分布函数为$F(x)$), $(X_{(1)},X_{(2)},\cdots,X_{(n)})$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

$$f_{X_{(n)}}(x)=n[F(x)]^{n-1}f(x)$$

(2) 最小次序统计量$X_{(1)}$的分布密度为

$$f_{X_{(1)}}(x)=n[1-F(x)]^{n-1}f(x)$$

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

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

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

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

$$= F^n(x)$$

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

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

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

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

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

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

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

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

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

例4 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$X_{(n)}$的分布密度.

解 总体X的分布密度为

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

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

由上述性质得$X_{(1)}$的分布密度为

$$f_{X_{(1)}}(x)=n[1-F(x)]^{n-1}f(x)$$

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

而$X_{(n)}$的分布密度为

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

3、 经验分布函数

定义1.3 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

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

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

$$v_n(x) \sim B(n, F(x))$$

即

$$P\{v_n(x)=k\}=C_n^k(P\{X\le x\})^k(1-P\{X\le x\})^{n-k}$$

4、 经验分布函数的性质

(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(共四条),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

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

Thank You!