数理统计

冯伟

数学与系统科学学院

wfeng_323@buaa.edu.cn

第一章 数理统计学的基本概念

- 总体与个体
- 抽样、简单随机抽样
- 样本、简单随机样本与样本空间
- 分布族、参数空间
- 统计量与样本矩

总体与个体

- 在数理统计中，把研究对象的全体称为总体，把组成总体的每一个单元称为个体
- 在实际中，总体通常是某个随机变量取值的全体，其中每一个个体都是一个实数
- 以后我们把总体和数量指标X可能取值的全体组成的集合等同起来。
- 随机变量X的分布就是总体的分布

抽样与简单随机抽样

从一总体X中随机抽取n个个体$x_1,x_2,\ldots,x_n$,

- 其中每个x_i是一次抽样观察结果，我们称$x_1,x_2,\ldots,x_n$为总体X的 一组样本（观察）值。
- 这里的x_i具有二重性：1.对每一次抽样结果，它是完全确定的一组数；2.由于抽样的随机性，每一个x_i都可以看作某一个随机变量$X_i(i=1,2,\ldots,n)$所取的观察值。
- 我们称$(X_1,X_2,\ldots,X_n)$是容量为n的样本。

抽样与简单随机抽样

定义： 设（$X_1, X_2, \ldots, X_n$）为来自总体X的容量为n的样本，如果随机变量$X_1, X_2, \ldots, X_n$相互独立且与总体有相同的分布，则称这样的样本为总体X的**简单随机样本**，简称样本。这样获得简单随机样本的方法称为简单随机抽样。

- **抽样方式：** 随机抽样,分层抽样,等距抽样,整群抽样,多阶段抽样
- 以后如不特别声明，所提到的样本都是简单随机样本。

- 综上所述，所谓总体就是一个随机变量X，所谓样本（指简单随机样本）就是n个相互独立且与总体X有相同的分布的随机变量$X_1,X_2,\ldots,X_n$,并称$X_1,X_2,\ldots,X_n$为来自于总体X的样本. 显然，若总体具有分布函数$F(x)$,则（$X_1,X_2,\ldots,X_n$）的联合分布函数为

$$\prod_{i=1}^{n} F(x_i)$$

抽样与简单随机抽样

- 以后对（$X_1, X_2, \ldots, X_n$)作两种理解：

✓在理论推导中把其作为随机向量

✓在用理论推导所得出的结论进行具体推断时，作为实数向量，代入具体的观察值进行计算。

样本空间

定义：样本$(X_1,X_2,\ldots,X_n)$所有可能取值的全体称为样本空间，或称为子样空间。

- ✓ 样本空间为n维欧氏空间或它的一个子集。
- ✓ 一个样本观察值$(x_1,x_2,\ldots,x_n)$是样本空间中的一个点。

分布族与参数空间

- 在概率论中，总假定所用随机变量的分布函数已知，而在数理统计中，认为其是未知的，但总假定其是某一个分布族的成员。
- 一般可凭经验，直方图或经验分布函数来对总体给出假定。

分布族与参数空间

- ·如果对总体了解甚少，那么总体所在的分布族可设为$\{F(x):F(x)$为分布函数，其它条件$\}$

- 如果知道总体的分布形式，只是不知道具体参数，那么总体所在的分布族可设为$\{F_\theta(x):\theta\in\Theta\}$，这里$\theta$为总体的分布函数中的未知参数(可以是向量)，未知参数的全部可容许值组成的集合称为参数空间，记为Θ

分布族与参数空间

定义：若一个分布族中只含有有限个未知参数，或参数空间为欧氏空间的一部分，则称此分布族为参数分布族。凡不是参数分布族的分布族称为非参数分布族。

- 由参数分布族出发所得到的统计方法称为参数统计方法；由非参数分布族出发所得到的统计方法称为非参数统计方法。这两类分布族在研究方法上有很大差异。

统计量与样本矩

- 我们对某一个问题归纳出所在的分布族，并从总体中抽出了一个样本后，就要进行统计推断，即判断这个样本是来自总体分布族中哪一个基本的分布.
- 虽然样本含有总体的信息，但仍比较分散。为了使统计推断成为可能，首先必须把分散在样本中的信息集中起来，用样本的某种函数表示，这种函数称为统计量(Statistic) 。

统计量与样本矩

- 定义：设$(X_1,X_2,\ldots,X_n)$为总体X的一个样本，若样本的实值连续（可扩大为可测）函数

$$T=T(X_1,X_2,\ldots,X_n)$$

不依赖于可能含于总体中的未知参数，则称T为此分布族的一个统计量(Statistic)。

- 往往从直观或某些一般性原则考虑提出统计量，再考虑它是否在某种意义下较好地集中了样本中与所讨论问题有关的信息量。

例如，$X\sim N(\mu,\sigma^2)$，其中μ已知，σ^2未知。(X_1,X_2)是从X中抽取的一个样本，问X_1+X_2，$(X_1-\mu)/\sigma$，$\frac{1}{2}\sum_{i=1}^{2}(X_i-\mu)^2$ 哪个是统计量？

样本矩（Sample Moment）

设（$X_1,X_2,\ldots,X_n$)是来自于总体X的一个样本

- 样本均值（Sample Mean）

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

- 样本方差（Sample Variance）

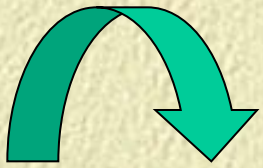
$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

- 样本的k阶原点矩

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

- 样本的k阶中心矩

$$B_k(M_k')=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k$$



再设（ $Y_1, Y_2, \ldots, Y_n$ ）是来自总体Y的样本。

两个样本之间的协方差：

$$S_{12} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

两个样本之间的相关系数：

$$\rho_{12} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(\sum_{i=1}^{n}(X_i - \bar{X})^2)(\sum_{i=1}^{n}(Y_i - \bar{Y})^2)}}$$

记 $E(X)=\mu$, $D(X)=\sigma^2$, $E(X^k)=a_k$

定理1 若X的二阶矩存在，则有

$$E(\bar{X})=\mu, D(\bar{X})=\frac{\sigma^2}{n} \quad E(S^2)=\sigma^2$$

定理2 若X的$2k$阶矩存在，则有

$$E(A_k)=a_k, D(A_k)=\frac{a_{2k}-a_k^{\ 2}}{n}$$

经验分布函数

定义 设$(X_1,X_2,\ldots,X_n)$为总体X的一个样本，$(x_1,x_2,\ldots,x_n)$是样本的一观察值，把其从小到大重新排列得到 $x_{(1)},x_{(2)},\cdots,x_{(n)}$，定义函数如下

$$F_n^*(x)=\begin{cases}0, & x\le x_{(1)}\\ \dfrac{k}{n}, & x_{(k)}<x\le x_{(k+1)},(k=1,2,\cdots,n-1)\\ 1, & x>x_{(n)}\end{cases}$$

称其为总体X的经验分布函数。

•此经验分布函数是一个分布函数；

•对于x的 每一个固定的值，它又是样本（$X_1,X_2,\ldots,X_n$)的函数，因而它是一个统计量.

定理 （格列文科定理） 设总体的分布函数为$F(x)$,经验分布函数为$F_n^*(x)$,则对任何实数x有

$$P\left(\lim_{n\longrightarrow\infty}\sup_{-\infty<x<\infty}\left|F_n^*(x)-F(x)\right|=0\right)=1$$

证明参看M.费史著，王福保译《概率论与数理统计》345页。

- 从上面定理知道，经验分布函数$F_n^*(x)$依概率1收敛于（理论）分布函数$F(x)$。
- 可以利用经验分布函数构造出非参数统计推断中许多常用的统计量。

习题 某厂从一批指示灯中抽出了10个，测其寿命，得数据如下(单位千时）：

95.5,15.8,13.1,26.5,31.7,33.8,8.7,15.0,48.8,49.3,求它的经验分布函数。

第一章 抽样分布 数理统计初步

统计量的分布称为抽样分布，求出统计量的分布函数是数理统计的基本问题之一。

- 精确分布与小样本问题
- 极限分布与大样本问题

Γ分布族

若连续型 r.v X具有概率密度

$$f(x)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}, & x>0\\ 0, & x\le 0\end{cases}$$

则称X服从参数为α,β的Γ分布,记作$X\sim\Gamma(\alpha,\beta)$

其中α,β均为正常数,分别称为形状参数和尺度参数。$\Gamma(\alpha)=\int_0^{+\infty}t^{\alpha-1}e^{-t}dt$是含参变量的广义积分。

这里$\Gamma(\alpha)=\int_0^{+\infty}t^{\alpha-1}e^{-t}dt$是含参变量的广义积分。

Γ函数具有以下性质:

(1) $\Gamma(1)=1,\Gamma(\frac{1}{2})=\sqrt{\pi}$

(2) $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)\qquad \alpha>0$

(3)对自然数n, $\Gamma(n+1)=n!$

- Γ分布在水文统计、最大风速或最大风压的概率计算中经常要用到.

•不少常见的重要分布是Γ分布的特殊情形.

当$\alpha=1$时,Γ分布即是参数为β的指数分布;

当$\alpha=n/2,\beta=1/2$时,Γ分布则是统计学中十分重要的$\chi^2(n)$分布,其概率密度为

$$f(y)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}, & y>0\\ 0, & y\le 0\end{cases}$$

性质1 设$X\sim\Gamma(\alpha,\beta)$，则$E(X)=\alpha/\beta$，$D(X)=\alpha/\beta^2$.

$$\Gamma(\alpha)=\int_0^{+\infty}t^{\alpha-1}e^{-t}dt$$

性质2 设$X_i\sim\Gamma(\alpha_i,\beta)$，$i=1,2,\ldots,n$，且$X_i$相互独立，则$X_1+X_2+\ldots+X_n\sim\Gamma(\alpha_1+\alpha_2+\ldots+\alpha_n,\beta)$

定义 称$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$为β函数。

引理

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

β分布族

定义 若连续型 r.v X具有概率密度

$$f(x,\alpha,\beta)=\begin{cases}\dfrac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, & 0<x<1\\ 0, & \text{其它}\end{cases}$$

则称X服从参数为α,β的β分布,记作 $X\sim Be(\alpha,\beta)$

其中α,β均为正常数。

性质3 设$X\sim Be(\alpha,\beta)$，则$E(X)=\alpha/(\alpha+\beta)$,

$$D(X)=\alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1).$$

二维正态随机变量的性质

设$(X,Y)\sim N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$

则(1) $X\sim N(\mu_1,\sigma_1^2)$, $Y\sim N(\mu_2,\sigma_2^2)$

$EX=\mu_1, DX=\sigma_1^2, EY=\mu_2, DY=\sigma_2^2$

(2) $\mathrm{cov}(X,Y)=\rho\sigma_1\sigma_2$

(3) $Z=k_1X+k_2Y+b$服从正态分布,

$EZ=k_1\mu_1+k_2\mu_2+b$

$DZ=k_1^2\sigma_1^2+k_2^2\sigma_2^2+2k_1k_2\rho\sigma_1\sigma_2$

(4) X与Y独立$\Leftrightarrow\ \rho=0\Leftrightarrow X$与$Y$不相关

多元正态分布族

定义 如果p维随机向量 (随机变量)

$$X=(X_1,X_2,\cdots,X_p)^T$$

(联合)概率密度函数为

$$f(x_1,x_2,\cdots,x_p)=\frac{1}{(2\pi)^{\frac{p}{2}}|V|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(X-\mu)^TV^{-1}(X-\mu)\right\}$$

则称随机向量X为p维正态随机向量，其中μ称为均值向量，V为协方差矩阵(协差阵)，且$V>0$.

多元正态分布的性质：

（1） p维正态分布由其均值向量和协方差阵唯一确定。

（2） 对于任一 p 维向量μ及p阶非负定矩阵V，必存在p维正态随机向量$X \sim N_p(\mu, V)$。

（3） 设$X \sim N_p(\mu, V)$，A是$m \times p$常数矩阵，b是m维向量，若令$Y = AX + b$，则

$$Y \sim N_m(A\mu + b, AVA^T).$$

（4）X为 p 维正态随机向量的充要条件为对任一p维向量c，$c^T X$是一维正态随机变量。

（5）设$X=(X_1^T, X_2^T)^T$为多维正态随机向量，则X_1与X_2互不相关的充要条件是 X_1与 X_2相互独立。

注：若$Cov(X,Y)=0$，则称X与Y互不相关。

（6）若 $X \sim N_p(\mu, V)$，且 $|V| \neq 0$，则

$$\eta \triangleq (X-\mu)^T V^{-1}(X-\mu) \sim \chi^2(p).$$

（7） n个独立的p元正态随机向量的和仍服从p元正态分布,即若$x_{(1)}, x_{(2)}, \ldots x_{(n)}$相互独立,且$x_{(\alpha)} \sim N_p(\mu_\alpha, V_\alpha)(\alpha = 1,2,\cdots,n)$,则

$$y = \sum_{\alpha=1}^{n} x_{(\alpha)} \sim N_p\left(\sum_{\alpha=1}^{n} \mu_\alpha, \sum_{\alpha=1}^{n} V_\alpha\right)$$

（8）设$X \sim N_p(\mu, V), V > 0.$ 则存在$p \times p$矩阵$B(BB^T = V)$使得

$$X = BY + \mu$$

其中$Y \sim N_p(0, I_p)$。