Lec3: 指数族和充分完备统计量

张伟平

2011 年 9 月 12 日

1 指数族

在统计理论问题中, 许多统计推断方法的优良性, 对一类范围广泛的统计模型 (亦称为分布族), 有较满意的结果. 这类分布族称为指数族. 常见的分布, 如正态分布、二项分布、Poisson 分布、负二项分布、指数分布和Gamma 分布等都属于这类分布族, 这些表面上看来各不相同的分布, 其实它们都可以统一在一种包罗更广的一类称为指数族的模式中. 当然引进这种分布族的理由, 主要不在于谋求形式上的统一, 而在于这种统一抓住了它们的共性, 因此许多统计理论问题, 对指数族获得较彻底的解决. 本节介绍指数族的定义及简单性质.

一、定义与例子

定义 1. *设$\mathscr{F} = \{f(x,\theta) : \theta \in \Theta\}$ 是定义在样本空间$\mathscr{X}$ 上的样本分布族, 其中Θ 为参数空间. 若其概率函数$f(x,\theta)$ 可表示成如下形式*

$$f(x,\theta) = C(\theta)\exp\Big\{\sum_{i=1}^{k} Q_i(\theta)T_i(x)\Big\}h(x),$$

则称此样本分布族为指数型分布族(简称指数族 (Exponential family). 其中k为自然数, $C(\theta) > 0$ 和$Q_i(\theta)$ $(i = 1,2\cdots,k)$ 都是定义在参数空间Θ 上的(可测) 函数, $h(x) > 0$ 和$T_i(x)$ $(i = 1,2,\cdots,k)$ 都是定义在$\mathscr{X}$ 上的(可测) 函数.

指数族的一个性质是族中的所有分布具有共同的支撑集($G(x)$ 称为概率函数$p(x)$的支撑集, 若$G(x) = \{x : p(x) > 0\}$). 由定义可见指数族支撑集$\{x : f(x,\theta) > 0\} = \{x : h(x) > 0\}$ 与θ 无关. 任一分布族若其支撑集与θ 有关, 则族中分布不再具有共同支撑集, 因而必不是指数族.

例1. *正态分布族$\{N(\mu,\sigma^2):\ -\infty < \mu < \infty,\ \sigma^2 > 0\}$是指数族.*

Proof. 设$\mathbf{X} = (X_1,\cdots,X_n)$为从正态分布$N(\mu,\sigma^2)$中抽取的简单样本, $\mathbf{X}$的联合密度为

$$f(\mathbf{x};\mu,\sigma^2) = \Big(\sqrt{2\pi}\sigma\Big)^{-n}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\Big\}. \tag{1.1}$$

记$\theta = (\mu,\sigma^2)$, 则参数空间为$\Theta = \{\theta = (\mu,\sigma^2):\ -\infty < \mu < +\infty,\ \sigma^2 > 0\}$.将(1.1) 改写为

$$f(\mathbf{x},\theta) = \Big(\sqrt{2\pi}\sigma\Big)^{-n} e^{-\frac{n\mu^2}{2\sigma^2}} exp\Big\{\frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\Big\}$$

$$= C(\theta)\exp\{Q_1(\theta)T_1(\mathbf{x}) + Q_2(\theta)T_2(\mathbf{x})\}h(\mathbf{x}), \tag{1.2}$$

此处$C(\theta) = (\sqrt{2\pi}\sigma)^{-n}e^{-\frac{n\mu^2}{2\sigma^2}}$, $Q_1(\theta) = \mu/\sigma^2$, $Q_2(\theta) = -\frac{1}{2\sigma^2}$, $T_1(\mathbf{x}) = \sum_{i=1}^n x_i$, $T_2(\mathbf{x}) = \sum_{i=1}^n x_i^2$, $h(\mathbf{x}) \equiv 1$. 因此由定义可知正态分布族$N(\mu, \sigma^2)$ 是指数族. □

例2. 二项分布族$\{b(n,\theta):\ 0 < \theta < 1\}$是指数族.

Proof. 设$X \sim$ 二项分布 $b(n,\theta)$, 其概率函数为

$$\begin{aligned} p(x,\theta) &= P_\theta(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}\Big(\frac{\theta}{1-\theta}\Big)^x(1-\theta)^n, \qquad x = 0,1,2,\cdots,n. \end{aligned} \tag{1.3}$$

此处样本空间$\mathscr{X} = \{0,1,2,\cdots,n\}$, 参数空间$\Theta = \{\theta : 0 < \theta < 1\} = (0,1)$. 将上式改写为

$$\begin{aligned} p(x,\theta) &= (1-\theta)^n \exp\left\{x\log\frac{\theta}{1-\theta}\right\}\cdot\binom{n}{x} \\ &= C(\theta)exp\{Q_1(\theta)T_1(x)\}h(x). \end{aligned} \tag{1.4}$$

此处$C(\theta) = (1-\theta)^n$, $Q_1(\theta) = \log\frac{\theta}{1-\theta}$, $T_1(x) = x$, $h(x) = \binom{n}{x}$, 按定义二项分布族$b(n,\theta)$ 也是指数族. □

例3. 均匀分布族$\{U(0,\theta),\ \theta > 0\}$ 不是指数族.

Proof. 由指数族的定义可知，其支撑集为$\{x:\ p(x,\theta) > 0\} = \{x:\ h(x) > 0\}$, 它与$\theta$ 无关. 而均匀分布族$\{U(0,\theta),\ \theta > 0\}$ 的支撑集为$\{x:\ p(x,\theta) > 0\} = (0,\ \theta)$ 与θ 有关, 因此它不是指数族. □

二、指数族的自然形式及自然参数空间

在指数族的定义$C(\theta)\exp\big\{\sum_{i=1}^k Q_i(\theta)T_i(x)\big\}h(x)$ 中, 若用φ_i 代替$Q_i(\theta)$, 而将$C(\theta)$ 表成φ的函数$C^*(\varphi)$, $\varphi = (\varphi_1,\varphi_2,\cdots,\varphi_k)$, 故其表达式变为$C^*(\varphi)\ \exp\big\{\sum_{i=1}^k \varphi_i T_i(x)\big\}h(x)$. 再改$\varphi_i$为$\theta_i$, $i = 1,2,\cdots,k$, 则上式即为: $C(\theta)\ \exp\big\{\sum_{i=1}^k \theta_i T_i(x)\big\}\ h(x)$, 此式称为指数族的自然形式(或称为标准形式). 故有下列定义

定义 2. 如果指数族有下列形式

$$f(x,\theta) = C(\theta)exp\Big\{\sum_{i=1}^n \theta_i T_i(x)\Big\}h(x), \tag{1.5}$$

则称为指数族的自然形式*(Natural form)*. 此时集合

$$\Theta^* = \Big\{(\theta_1,\theta_2,\cdots,\theta_k) : \int_{\mathscr{X}} \exp\big\{\sum_{i=1}^k \theta_i T_i(x)\big\}h(x)dx < \infty\Big\} \tag{1.6}$$

称为自然参数空间 *(Natural parametric space)*.

例4. 将正态分布族表示为指数族的自然形式, 并求出其自然参数空间.

Proof. 由
$$f(\mathbf{x};\mu,\sigma^2)=\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{n\mu^2}{2\sigma^2}}\exp\Big\{\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\Big\},$$
参数空间$\Theta=\{(\mu,\sigma^2):-\infty<\mu<\infty,\ 0<\sigma^2<\infty\}$.令$\varphi_1=\mu/\sigma^2$, $\varphi_2=-\frac{1}{2\sigma^2}$, 解出$\sigma=\sqrt{-\frac{1}{2\varphi_2}}$, $\mu^2/\sigma^2=\varphi_1^2(-\frac{1}{2\varphi_2})$, 因此有$\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{n\mu^2}{2\sigma^2}}=\Big(\sqrt{\frac{-2\varphi_2}{2\pi}}\Big)^n e^{\frac{n\varphi_1^2}{4\varphi_2}}\triangleq C^*(\varphi)$, $\varphi=(\varphi_1,\varphi_2)$, 故

$$\begin{aligned}f(\mathbf{x},\varphi)&=C^*(\varphi)\exp\big\{\varphi_1\sum_{i=1}^n x_i+\varphi_2\sum_{i=1}^n x_i^2\big\}h(\mathbf{x})\\&=C^*(\varphi)\exp\{\varphi_1T_1(\mathbf{x})+\varphi_2T_2(\mathbf{x})\}h(\mathbf{x}).\end{aligned}$$

再改φ_i 为θ_i $(i=1,2)$, 上式变为

$$f(\mathbf{x},\theta)=C^*(\theta)\exp\{\theta_1T_1(\mathbf{x})+\theta_2T_2(\mathbf{x})\}h(\mathbf{x}).\tag{1.7}$$

此为其自然形式. 其自然参数空间为

$$\Theta^*=\{(\theta_1,\theta_2):-\infty<\theta_1<+\infty,-\infty<\theta_2<0\}.$$

□

三、指数族的性质

定理 1. *在指数族的自然形式下, 自然参数空间为凸集.*

证明的方法如下：设任给$\theta^{(1)}=(\theta_1^1,\cdots,\theta_k^1)$, $\theta^{(0)}=(\theta_1^0,\cdots,\theta_k^0)$ 皆属于自然参数空间Θ^*, 设$0<\alpha<1$, 令$\theta=\alpha\theta^{(1)}+(1-\alpha)\theta^{(0)}$ (即$\theta_i=\alpha\theta_i^1+(1-\alpha)\theta_i^0$, $i=1,2,\cdots,k$),若能证明$\theta\in\Theta^*$, 则按凸集的定义, 定理得证.

定理 2. *设指数族的自然形式中, 自然参数空间有内点, $g(x)$ 是任一有界可积函数, 则对*

$$G(\theta)=\int_{\mathscr{X}}g(x)\exp\Big\{\sum_{j=1}^k\theta_jT_j(x)\Big\}h(x)dx,$$

有
$$\frac{\partial^m G(\theta)}{\partial\theta_1^{m_1}\cdots\partial\theta_k^{m_k}}=\int_{\mathscr{X}}\frac{\partial^m}{\partial\theta_1^{m_1}\cdots\partial\theta_k^{m_k}}\Big[g(x)\exp\Big\{\sum_{j=1}^k\theta_jT_j(x)\Big\}h(x)\Big]dx,$$

其中$\sum_{j=1}^k m_j=m$, 即对$G(\theta)$ 关于θ 的任意阶偏导数可在积分下求得.

此定理的更一般的形式及其证明查看参考文献[1] P_{21}定理1.2.1.

2 充分统计量

我们知道, 统计量是对样本的简化, 希望达到: (i) 简化的程度高; (ii) 信息的损失少. 一个统计量能集中样本中信息的多少, 与统计量的具体形式有关, 也依赖于问题的统计模型. 最好的情

况是统计量把样本中的全部信息都集中起来, 也就是说信息无损失, 我们称这样的统计量为充分统计量.

关于样本$\mathbf{X}=(X_1,X_2,\cdots,X_n)$的信息可以设想成如下的公式:

$$\begin{aligned}\{\text{样本}\mathbf{X}\text{中包含参数的信息}\} &= \{\text{统计量}T(\mathbf{X})\text{中所含参数的信息}\}\\ &+\{\text{在知道}T(\mathbf{X})\text{后样本}\mathbf{X}\text{尚含有关于参数的剩余信息}\}\end{aligned}$$

故$T(\mathbf{X})$为充分统计量的要求归结为: 要求后一项信息为0. 用统计的语言来描述, 即要求$P_\theta(\mathbf{X}|T=t)$与θ无关. 因此我们得到如下的定义:

定义 1. *设样本$\mathbf{X}$的分布族$\{F_\theta(\mathbf{x}),\ \theta\in\Theta\}$, θ为分布的参数. 设$T=T(\mathbf{X})$为一统计量, 若在已知T的条件下, 样本$\mathbf{X}$的条件分布与θ无关, 则称$T(\mathbf{X})$为充分统计量 (Sufficient statistic).*

实际应用时条件分布用条件概率(离散情形) 或条件密度(连续情形) 来代替.

例1. *设$\mathbf{X}=(X_1,X_2,\cdots,X_n)$为从$0-1$分布中抽取的简单样本, 则$T(\mathbf{X})=\sum_{i=1}^n X_i$为充分统计量.*

Proof. 按定义我们只要证明下列条件概率与θ无关.

$$\begin{aligned}&P(X_1=x_1,\cdots,X_n=x_n|T(\mathbf{x})=t_0)\\ &=\begin{cases}\frac{P(X_1=x_1,\cdots,X_n=x_n,T=t_0)}{P(T(x)=t_0)}=1/\binom{n}{t_0}, & \text{当}\ \sum\limits_{i=1}^n x_i=t_0\\ 0, & \text{当}\ \sum\limits_{i=1}^n x_i\neq t_0.\end{cases}\end{aligned}$$

上述条件概率与θ无关,因此$T(\mathbf{X})=\sum_{i=1}^n X_i$为充分统计量. □

例2. *设$\mathbf{X}=(X_1,X_2,\cdots,X_n)$为从正态总体$N(\theta,1)$中抽取的简单样本, 则$T(\mathbf{X})=\frac{1}{n}\sum_{i=1}^n X_i=\bar{X}$ 为充分统计量.*

Proof. 再做正交变换

$$\begin{cases} y_1=\frac{1}{\sqrt{n}}\sum\limits_{i=1}^n x_i,\\ y_j=\sum\limits_{k=1}^n a_{jk}x_k, \quad j=2,\cdots,n.\end{cases}$$

由正态总体下样本均值和样本方差的分布导出过程可知$\sum_{i=1}^n {Y_i}^2=\sum_{i=1}^n {X_i}^2$,且$Y_1,Y_2,\cdots,Y_n$是相互独立的, $Y_1\sim N(\sqrt{n}\theta,1)$, $Y_i\sim N(0,1)$, $i=2,\cdots,n$. 因此$(Y_1,\cdots Y_n)$的联合密度为

$$f(y_1,\cdots y_n)=(2\pi)^{-\frac{n}{2}}e^{-\frac{1}{2}\sum\limits_{i=2}^n y_i^2-\frac{1}{2}(y_1-\sqrt{n}\theta)^2}.$$

再由Y_1的密度函数为

$$f_{Y_1}(y_1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y_1-\sqrt{n}\theta)^2}$$

知在给定Y_1时, $(Y_1,\cdots,Y_n)$的条件密度是

$$f(y_1,\cdots,y_n|y_1)=\frac{f(y_1,\cdots,y_n)}{f_{Y_1}(y_1)}=(2\pi)^{-\frac{n-1}{2}}e^{-\frac{1}{2}\sum\limits_{i=2}^n y_i^2} \tag{2.1}$$

与θ无关. □

这里利用了下列事实: 曲面$\{(Y_1\cdots Y_n):Y_1=\sqrt{n}t=y_1\}$是由曲面$\{(X_1,\cdots,X_n):T(\mathbf{X})=t\}$经正交旋转而来,曲面保持不变. 因此在曲面$\{(X_1,\cdots,X_n):T(\mathbf{X})=t\}$上的条件概率与在曲面$\{Y_1,\cdots,Y_n):Y_1=y_1\}$ 上的条件概率相同. 故有

$$f(x_1,\cdots,x_n|T=t)=f(y_1,\cdots,y_n|Y_1=y_1)=(2\pi)^{-\frac{n-1}{2}}e^{-\frac{1}{2}\sum\limits_{i=2}^{n}y_i^2}$$

与θ无关, 所以$T(\mathbf{X})=\bar{X}$是充分统计量.

二、 充分性的判别准则—因子分解定理

因子分解定理是由R.A. Fisher 在二十世纪二十年代提出来,它的最一般形式和严格数学证明, 是Halmos 和Savage在1949年作出来的.

定理 3. *(因子分解定理) 设样本$\mathbf{X}=(X_1,\cdots,X_n)$的概率函数$f(\mathbf{x},\theta)=f(x_1,\cdots,x_n;\theta)$ 依赖于参数θ, $\mathbf{T}=T(\mathbf{X})=(T_1(\mathbf{X}),\cdots,T_k(\mathbf{X})$是一个统计量, 则$\mathbf{T}$为充分统计量的充要条件是$f(\mathbf{x},\theta)$可以分解为*

$$f(\mathbf{x},\theta)=g(T(\mathbf{x}),\theta)h(\mathbf{x}) \tag{2.2}$$

的形状.注意此处函数$h(\mathbf{x})=h(x_1,\cdots,x_n)$不依赖于$\theta$.

这里概率函数是指: 若$\mathbf{X}$为连续型, 则$f(\mathbf{x},\theta)$是其密度函数; 若$\mathbf{X}$是离散型, 则$f(\mathbf{x},\theta)=P_\theta(X_1=x_1,\cdots,X_n=x_n)$, 即样本$\mathbf{X}$的概率分布.

推论 1. *设$\mathbf{T}=T(\mathbf{X})$为θ的充分统计量, $S=\varphi(\mathbf{T})$是单值可逆函数, 则$S=\varphi(\mathbf{T})$也是θ的充分统计量.*

例3. *设$\mathbf{X}=(X_1,\cdots,X_n)$为从正态总体$N(a,\sigma^2)$中抽取的简单样本, 令$\theta=(a,\sigma^2)$, 则$T(\mathbf{X})=(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2)$为充分统计量.*

Proof. 样本$\mathbf{X}$的联合密度为

$$\begin{aligned}
f(\mathbf{x},\theta)&=\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-a)^2\Big\}\\
&=\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n exp\Big\{-\frac{1}{2\sigma^2}\Big(\sum_{i=1}^n x_i^2-2a\sum_{i=1}^n x_i+na^2\Big)\Big\}\\
&=g(T(\mathbf{x}),\theta)\cdot h(\mathbf{x}).
\end{aligned}$$

此处$h(\mathbf{x})\equiv 1$, 故由因子分解定理可知$T(\mathbf{X})=(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2)$为充分统计量. □

由于$(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2)$与$(\bar{X},S^2)$为一一对应的变换, 由推论可知$(\bar{X},S^2)$也是充分统计量.

例4. *设$\mathbf{X}=(X_1,\cdots,X_n)$为从总体$b(1,\theta)$中抽取的简单样本, 则$T(\mathbf{X})=\sum_{i=1}^n X_i$是充分统计量.*

Proof. 样本$\mathbf{X}$的联合分布是

$$\begin{aligned}
f(\mathbf{x},\theta)&=P_\theta(X_1=x_1,\cdots,X_n=x_n)\\
&=\theta^{\sum\limits_{i=1}^n x_i}(1-\theta)^{n-\sum\limits_{i=1}^n x_i}=g(T(\mathbf{x}),\theta)h(\mathbf{x}).
\end{aligned}$$

此处$h(\mathbf{x})\equiv 1$, 故由因子分解定理可知$T(\mathbf{X})=\sum_{i=1}^n X_i$为充分统计量. □

三、极小充分统计量*

一个分布族$\mathscr{F}$的充分统计量往往不止一个, 那么在使用中应该如何挑选呢? 我们知道, 统计量是由样本加工而来的, 如本章引言所述, 对样本的加工显然可以提出两条要求: (1) 在加工中, 样本所含参数θ的信息损失越少越好. 若加工中此种信息毫无损失, 那就是充分性的要求. (2) 加工中, 所得统计量愈简化越好, 简化的程度可以用统计量的维数来衡量, 也可以用函数关系来表示. 例如对一个二维统计量$T_1(\mathbf{X}) = (\sum_{i=1}^m X_i, \sum_{i=m+1}^n X_i)$,再进一步加工得到一维统计量$T_2 = \sum_{i=1}^n X_i$. 直观上容易看出, T_2比T_1简化. 而且可以看出, T_2是T_1的函数. 一般来说, 若T与s是两个统计量, 且T是S 的函数, 即$T = q(S)$, 那么由函数的定义可知, T比S简化.

定义 2. *设T是分布族$\mathscr{F}$的充分统计量, 若对$\mathscr{F}$的任一充分统计量$S(\mathbf{X})$, 存在一个函数$q_S(\cdot)$使得$T(\mathbf{X}) = q_S(S(\mathbf{X}))$, 则称$T(\mathbf{X})$是此分布族的极小充分统计量.*

3 完全统计量*

定义 1. *设$\mathscr{F} = \{F_\theta(x), \theta \in \Theta\}$为一分布族, Θ是参数空间.设$T = T(X)$为一统计量, 若对任一实函数$\varphi(\cdot)$,由*

$$E_\theta\varphi(T(X)) = 0, \text{ 一切 } \theta \in \Theta, \tag{3.1}$$

总可推出

$$P_\theta(\varphi(T(X)) = 0) = 1, \text{ 一切 } \theta \in \Theta, \tag{3.2}$$

则称$T(X)$是一完全统计量 (Complete Statistic).

由定义可见, 若$T(X)$是完全统计量, 则它的任一实函数$g(T)$也是完全统计量.

注1. *统计量$T(X)$的完全性不仅取决于T的形状, 还取决于样本X的分布族. 完全性(亦称完备性) 这个名称, 是来源于正交函数理论中的一个类似概念. 为简单计, 设统计量$T(X)$有密度函数$g_\theta(t)$,则(3.1)式可写为*

$$\int \varphi(t)g_\theta(t)dt = 0, \text{ 一切 } \theta \in \Theta. \tag{3.3}$$

积分$\int \varphi(t)g_\theta(t)dt = 0$形式上可看成"$\varphi$与$g_\theta$正交". 于是, 条件(3.1)可说成是"$\varphi$与函数系$\{g_\theta, \theta \in \Theta\}$正交". 在正交函数论, 若$M$表示一正交函数系, 且不存在与$M$正交的非零函数, 则称$M$为完全正交系. 由(3.3) 看出, 我们这里的完全性正好与此相当. 不过我们不称密度函数系$\{g_\theta, \theta \in \Theta\}$完全, 而称统计量$T$完全. 由于$\{g_\theta, \theta \in \Theta\}$是由统计量$T$决定的, 这种称呼不影响实质.

例1. *设$\mathbf{X} = (X_1, \cdots, X_n)$为从总体$b(1, \theta)$中抽取的简单样本, 则$T(\mathbf{X}) = \sum_{i=1}^n X_i$是完全统计量.*

Proof. 显然, $T(\mathbf{X}) \sim b(n, \theta)$, 故有

$$P(T(\mathbf{X}) = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad k = 0, 1, 2, \cdots, n.$$

设$\varphi(t)$为任一实函数, 满足$E_\theta\varphi(T) = 0$, 一切$0 < \theta < 1$,此即

$$\sum_{k=0}^n \varphi(k)\binom{n}{k}\theta^k(1-\theta)^{n-k} = 0 \Longleftrightarrow$$

$$\sum_{k=0}^{n}\varphi(k)\binom{n}{k}\Big(\frac{\theta}{1-\theta}\Big)^k=0,\ 0<\theta<1.$$

令$\theta/(1-\theta)=\delta$, 则上式等价于

$$\sum_{k=0}^{\infty}\Big[\varphi(k)\binom{n}{k}\Big]\delta^k=0,\ 0<\delta<\infty.$$

上式左边是δ的多项式, 故必有

$$\varphi(k)\binom{n}{k}=0,\ k=0,1,2,\cdots,n.$$

即$\varphi(k)=0,\ k=0,1,2,\cdots,n$. 这就证明了$T(\mathbf{X})=\sum_{i=1}^n X_i$是完全统计量. □

例2. 设$\mathbf{X}=(X_1,X_2,\cdots,X_n)$为从正态总体$N(\theta,1)$中抽取的简单样本, 则$T(\mathbf{X})=\bar{X}$为完全统计量.

Proof. 显然$T(\mathbf{X})=\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i\sim N(\theta,1/n)$, 设$\varphi(t)$为$t$的任一实函数, 满足$E_\theta\varphi(T)=0$, 对一切$-\infty<\theta<\infty$. 此即

$$\sqrt{\frac{n}{2\pi}}\int_{-\infty}^{\infty}\varphi(t)e^{-\frac{n(t-\theta)^2}{2}}dy=\sqrt{\frac{n}{2\pi}}\int_{-\infty}^{\infty}\varphi(t)e^{-\frac{nt^2}{2}}\cdot e^{-\frac{n\theta^2}{2}}\cdot e^{nt\theta}dt=0.$$

所以

$$\int_{-\infty}^{\infty}\varphi(t)e^{-\frac{nt^2}{2}}\cdot e^{nt\theta}dt=0,\ -\infty<\theta<\infty$$

令$z=n\theta$,则

$$G(z)=\int_{-\infty}^{\infty}\varphi(t)e^{-\frac{nt^2}{2}}e^{tz}dt.$$

将z视为复数, $G(z)$为全平面上的解析函数, 且$G(z)$当z取实数时为0, 由解析函数的唯一性定理, $G(z)$在整个复平面上为0, 特别取$z=i\mu$,则

$$G(\mu)=\int_{-\infty}^{\infty}\varphi(t)e^{-\frac{nt^2}{2}}\cdot e^{-i\mu t}dt=0.$$

由$Fourier$变换的逆变换公式, 可知

$$\varphi(t)e^{-nt^2/2}=0.$$

故有$\varphi(t)=0,\ |t|<\infty$, 因此$T(\mathbf{X})=\bar{X}$为完全统计量. □

二、指数族中统计量的完全性

定理 4. 设样本$\mathbf{X}=(X_1,X_2,\cdots,X_n)$的概率函数

$$f(\mathbf{x},\theta)=C(\theta)exp\Big\{\sum_{i=1}^{k}Q_i(\theta)T_i(\mathbf{x})\Big\}h(\mathbf{x}),\quad \theta\in\Theta$$

为指数族. 令$T(\mathbf{X})=(T_1(\mathbf{X}),\cdots,T_k(\mathbf{X}))$, 若自然参数空间$\Theta^*$作为$R_k$的子集有内点, 则$T(X)$是完全统计量.

例3. 设$\mathbf{X} = (X_1, \cdots, X_n)$是从均匀分布$U(\theta - 1/2, \theta + 1/2)$中抽取的简单样本, 则$T(\mathbf{X}) = (X_{(1)}, X_{(n)})$是充分统计量, 但不是完全统计量.

Proof. $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$的充分性在例2.7.9中已证. 下面来证明它不是完全的.

要证明一个统计量$T(\mathbf{X})$不是完全的, 只要找到一个实函数$\varphi(t)$使得$E_\theta\varphi(T) = 0$, 但"$\varphi(T) = 0$, $a.e.\ P_\theta$"是不成立的即可.

令$Z = X_{(n)} - X_{(1)}$, $Y_i = X_i - (\theta - 1/2)$, $i = 1, 2, \cdots, n$,则$Y_1, \cdots, Y_n$ i.i.d. $\sim U(0,1)$,与θ无关. 而此时$Z = X_{(n)} - X_{(1)} = Y_{(n)} - Y_{(1)}$的分布也与$\theta$无关. 找常数$a < b$使得

$$P(Z < a) = P(Z > b) > 0.$$

定义

$$\varphi(t) = \begin{cases} 1, & Z < a, \\ -1, & Z > b, \\ 0, & \text{其它}. \end{cases}$$

则易见$\varphi(t)$满足:$E_\theta\varphi(T) = 0$, 但$\varphi(t) \not\equiv 0$ (即$P(\varphi(t) \neq 0) > 0$). 按定义$T(\mathbf{X}) = (X_{(1)}, X_{(n)})$不是完全统计量. □

三、有界完全统计量及其性质

定义 2. 若对任何满足

$$E_\theta\varphi(T(X)) = 0, \text{ 对一切 } \theta \in \Theta$$

的有界(或$a.e.$有界)的函数$\varphi(\cdot)$ 都有

$$P_\theta\varphi(T(X) = 0) = 1, \text{ 对一切 } \theta \in \Theta,$$

则称$T(X)$为有界完全统计量.

由定义可见: 一个"完全统计量"必为"有界完全统计量", 反之不必对.

定理 5. *(*Basu定理 设$\mathscr{F} = \{F_\theta(x), \theta \in \Theta\}$为一分布族, Θ是参数空间.样本$\mathbf{X} = (X_1, \cdots, X_n)$是从分布族$\mathscr{F}$中抽取的简单样本, 设$T(\mathbf{X})$是一有界完全统计量,且是充分统计量. 若$r.v.\ V(\mathbf{X})$的分布与$\theta$无关, 则对任何$\theta \in \Theta$, $V(\mathbf{X})$与$T(\mathbf{X})$独立.

例4. 设$\mathbf{X} = (X_1, \cdots, X_n)$是从$N(\theta, 1)$中抽取的简单样本, $R(\mathbf{X}) = X_{(n)} - X_{(1)}$称为极差, 则$T(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$与$R(\mathbf{X})$ 独立.

Proof. 由于正态分布$N(\theta, 1)$为指数族, 自然参数空间$\Theta^* = \{\theta : -\infty < \theta < \infty\}$作为$R_1$的子集有内点. 故$T(\mathbf{X})$为充分完全统计量.

令$Y_i = X_i - \theta$,则$Y_i \sim N(0,1)$, $i = 1, 2, \cdots, n$. 因此$Y_1, \cdots, Y_n$ i.i.d. $\sim N(0,1)$,与θ无关.从而$Y_{(n)} - Y_{(1)}$的分布也与θ无关.故

$$R(\mathbf{X}) = X_{(n)} - X_{(1)} = Y_{(n)} - Y_{(1)}$$

之分布与θ 无关, 由推论2.8.1可知$T(\mathbf{X})$与$R(\mathbf{X})$独立. □