§2.7 充分统计量

充分统计量

统计量(statistics)是对样本的加工. "好"的统计量应该能够将样本中关于总体分布的未知信息尽可能地集中起来.

§2.7 充分统计量

充分统计量

统计量(statistics)是对样本的加工. "好"的统计量应该能够将样本中关于总体分布的未知信息尽可能地集中起来.

若要研究某参数分布族中的某个未知参数,为此抽取了一组样本,样本中所包含的信息可分成两部分.
其一是关于未知参数的信息;
其二关于样本结构的信息等其它信息.

Example

某厂要了解其半成品的不合格率p, 检验员检查了10件产品, 检查的结果是, 除前两件是不合格品(记为$X_1 = 1$, $X_2 = 1$)外, 其它都是合格品(记为$X_i = 0$, $i = 3, 4, \cdots, 10$).

Example

某厂要了解其半成品的不合格率p, 检验员检查了10件产品, 检查的结果是, 除前两件是不合格品(记为$X_1=1$, $X_2=1$)外, 其它都是合格品(记为$X_i=0$, $i=3,4,\cdots,10$).

这时样本$(X_1,X_2,\cdots,X_{10})$提供了两种信息:

(1) 10 次检验中, 不合格品出现了几次,

(2) 不合格品出现在哪几次试验上.

Example

某厂要了解其半成品的不合格率p, 检验员检查了10件产品, 检查的结果是, 除前两件是不合格品(记为$X_1 = 1$, $X_2 = 1$)外, 其它都是合格品(记为$X_i = 0$, $i = 3, 4, \cdots, 10$).

这时样本$(X_1, X_2, \cdots, X_{10})$提供了两种信息:

(1) 10 次检验中, 不合格品出现了几次,

(2) 不合格品出现在哪几次试验上.

第二种信息(试验编号的信息)对了解不合格率p是没有什么帮助的. 例如, 设在另一次试验中,试验结果为$(0, 0, \cdots, 0, 1, 1)$. 这两次试验结果所含的关于p的信息应该是一样的.

在上例中, 当厂长问及检查结果时, 看看检验员的如下两种回答:

1. 10件中有两件不合格—-所用统计量为$T_1=\sum_{i=1}^{10}X_i$,其值为$t_1=\sum_{i=1}^{10}x_i=2$;

2. 前两件不合格($X_1=1,\ X_2=1$)—-所用统计量为$T_2=X_1+X_2$,其值为$t_2=x_1+x_2$.

在上例中, 当厂长问及检查结果时, 看看检验员的如下两种回答:

1. 10件中有两件不合格—-所用统计量为$T_1=\sum_{i=1}^{10}X_i$,其值为$t_1=\sum_{i=1}^{10}x_i=2$;

2. 前两件不合格$(X_1=1,\ X_2=1)$—-所用统计量为$T_2=X_1+X_2$,其值为$t_2=x_1+x_2$.

显然, 第二种回答是不能令人满意的, 因为统计量T_2不包含样本中有关p的全部信息. 而第一种回答综合了样本中有关p的全部信息.

一个“好”的统计量，应该能够将样本中所包含的关于未知参数的信息全部集中起来. 这样的统计量就是该未知参数的充分统计量.

如何将这样一个直观的想法用严格的数学形式来表示呢?

样本$\boldsymbol{X}$中的信息

$=T(\boldsymbol{X})$中所含样本的信息

$+$ 在知道$T(\boldsymbol{X})$后样本$\boldsymbol{X}$中尚含有的剩余信息.

设有参数分布族$\mathcal{F}=\{F_\theta:\theta\in\Theta\}$,
$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$是从某总体$F_\theta\in\mathcal{F}$中抽取的样本,
$T=T(X_1,X_2,\cdots,X_n)$是一个统计量. 样本$\boldsymbol{X}$有一个样本分布

$$F_\theta(\boldsymbol{x})=\prod_{i=1}^n F_\theta(x_i),$$

统计量T也有一抽样分布$F_\theta^T(t)$.

充分性要求统计量$T(\boldsymbol{X})$包括样本$\boldsymbol{X}$中有关θ的全部信息. 也就是说样本分布$F_\theta(\boldsymbol{x})$所含的关于θ的全部信息都包含在抽样分布$F_\theta^T(t)$中.

充分性要求统计量$T(\boldsymbol{X})$包括样本$\boldsymbol{X}$中有关θ的全部信息. 也就是说样本分布$F_\theta(\boldsymbol{x})$所含的关于θ的全部信息都包含在抽样分布$F_\theta^T(t)$中.

即除了$F_\theta^T(t)$所含的关于θ的信息外, $F_\theta(\boldsymbol{x})$不再含的关于θ的信息.

换言之,即在统计量T的取值给定后, 譬如$T=t$后, 样本的条件分布$F_\theta(\boldsymbol{x}|T=t)$已不再依赖于参数$\theta$.

定义 设有一个分布族$\mathcal{F}=\{F\}$, $(X_1,X_2,\cdots,X_n)$是从某总体$F\in\mathcal{F}$中抽取的一个样本. $T=T(X_1,...,X_n)$为一个（一维或多维的）统计量.
如果当给定$T=t$下, 样本$(X_1,...,X_n)$ 的条件分布与总体分布F无关，则称T为是此分布族的**充分统计量**(*sufficient statistics*).

如果$\mathcal{F}$是参数型分布族$\{F_\theta:\theta\in\Theta\}$, 当给定$T=t$下, 样本$(X_1,...,X_n)$ 的条件分布与参数θ无关，则亦称T为是参数θ的充分统计量.

Theorem

充分统计量的一一变换仍是充分统计量.

Theorem

充分统计量的一一变换仍是充分统计量.

设$s=\psi(t)$是一一变换, $S=\psi(T)$, 则事件$\{S=s\}$与$\{T=\psi^{-1}(s)\}$等价.一般地,$\{S\in A\}$ 与$\{T\in\psi^{-1}(A)\}$等价.所以

$$P\big(\boldsymbol{X}\in A|S=s\big)=P\big(\boldsymbol{X}\in A|T=\psi^{-1}(s)\big).$$

Theorem

设$\boldsymbol{Y}$是样本$\boldsymbol{X}$的一一变换，则统计量$T = T(\boldsymbol{X})$对$\boldsymbol{X}$的充分性等价于T对$\boldsymbol{Y}$的充分性.

Theorem

设$\boldsymbol{Y}$是样本$\boldsymbol{X}$的一一变换，则统计量$T = T(\boldsymbol{X})$对$\boldsymbol{X}$的充分性等价于T对$\boldsymbol{Y}$的充分性.

设$\boldsymbol{Y} = \widetilde{\psi}(\boldsymbol{X})$是一一变换，则事件$\{\boldsymbol{X} \in A\}$与$\{\boldsymbol{Y} \in \widetilde{\psi}(A)\}$等价. 所以

$$P\big(\boldsymbol{X} \in A | T = t\big) = P\big(\boldsymbol{Y} \in \widetilde{\psi}(A) | T = t\big).$$

Example

设$X_1, X_2, \cdots, X_n$是来自二点分布族$\{B(1,p), 0 < p < 1\}$中的某总体的一个样本, 其中$0 < p < 1$, $n > 2$. 考察统计量

$$T_1 = \sum_{i=1}^{n} X_i, \quad T_2 = X_1 + X_2.$$

这时，样本$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$的联合分布列是

$$\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

其中$x_i=0$或1,$i=1,2,\cdots,n$.而

$$\mathsf{P}(T_1=t)=\binom{n}{t}p^t(1-p)^{n-t},\quad t=0,1,\cdots,n.$$

这时，样本$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$的联合分布列是

$$\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

其中$x_i=0$或1,$i=1,2,\cdots,n$.而

$$\mathsf{P}(T_1=t)=\binom{n}{t}p^t(1-p)^{n-t},\quad t=0,1,\cdots,n.$$

在给定$T_1=t(t=0,1,\cdots,n)$的条件下，若$x_1+\cdots+x_n\neq t$，则

$$\begin{aligned}&\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T_1=t)\\=&\frac{\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n,T_1=t)}{\mathsf{P}(T_1=t)}=0;\end{aligned}$$

若$x_1+\cdots+x_n=t$, 则

$$
\begin{aligned}
&\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T_1=t)\\
=&\frac{\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=t-\sum_{i=1}^{n-1}x_i)}{\mathsf{P}(T_1=t)}\\
=&\frac{p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}}=\binom{n}{t}^{-1},
\end{aligned}
$$

与参数p无关, 故T_1为充分统计量.

对于T_2, 在给定$T_2=t$下, 当$x_1+x_2\neq t$时, 条件概率是0;
当$x_1+x_2=t$时,

$$
\begin{aligned}
&\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T_2=t)\\
=&\frac{\mathsf{P}(X_1=x_1,X_2=t-x_1,\cdots,X_n=x_n)}{\mathsf{P}(T_2=t)}\\
=&\frac{p^{t+\sum_{i=3}^n x_i}(1-p)^{n-t-\sum_{i=3}^n x_i}}{\binom{2}{t}p^t(1-p)^{2-t}}\\
=&\binom{2}{t}^{-1}p^{\sum_{i=3}^n x_i}(1-p)^{n-2-\sum_{i=3}^n x_i}.
\end{aligned}
$$

对于T_2, 在给定$T_2=t$下, 当$x_1+x_2\neq t$时, 条件概率是0;
当$x_1+x_2=t$时,

$$
\begin{aligned}
&\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T_2=t)\\
=&\frac{\mathsf{P}(X_1=x_1,X_2=t-x_1,\cdots,X_n=x_n)}{\mathsf{P}(T_2=t)}\\
=&\frac{p^{t+\sum_{i=3}^n x_i}(1-p)^{n-t-\sum_{i=3}^n x_i}}{\binom{2}{t}p^t(1-p)^{2-t}}\\
=&\binom{2}{t}^{-1}p^{\sum_{i=3}^n x_i}(1-p)^{n-2-\sum_{i=3}^n x_i}.
\end{aligned}
$$

这个条件分布仍与参数p有关,T_2不是充分统计量.

对充分统计量的理解:

设想有两个试验员, 试验员A可观察到样本$\boldsymbol{X}$, 试验员B只能观察到统计量$T(\boldsymbol{X})$. 如果T是充分统计量,他们得到关于θ的信息应该是等价的.

Example

设$X_1, X_2, \cdots, X_n$是来自几何分布

$$\mathsf{P}(X=x)=\theta(1-\theta)^x, \quad x=0,1,2,\cdots$$

的一个样本, 其中$0<\theta<1$, 则$T=\sum_{i=1}^n X_i$是参数θ的充分统计量.

证明:我们先求T的分布.设想有一系列独立重复试验,每次试验的成功率为θ, 记T_1为第一次成功时前面试验的总次数(即失败的次数), T_2为第一次成功后到第二次成功之间的失败次数, ..., T_n为第$n-1$次成功后到第n次成功之间的失败次数, 这时$T_1, T_2, \cdots, T_n$独立同分布且与X的分布相同. 因而$T^* = \sum_{i=1}^n T_i$与T同分布. 因为T^*是第n次成功时, 失败试验的总次数, $T^* = t$ 意味着共进行了$t+n$次试验, 其中第$t+n$次成功, 前$t+n-1$次试验中有$n-1$次成功,t次失败. 因此

$$\mathsf{P}(T = t) = \mathsf{P}(T^* = t) = \binom{t+n-1}{n-1}\theta^n(1-\theta)^t, \quad t = 0, 1, 2, \cdots.$$

所以在$T=t$时,样本的条件分布为: 若$x_1+\cdots+x_n\neq t$, 则

$$\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T=t)=0;$$

若$x_1+\cdots+x_n=t$, 则

$$\begin{aligned}
&\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n|T=t)\\
=&\frac{\mathsf{P}(X_1=x_1,X_2=x_2,\cdots,X_n=x_n,T=t)}{\mathsf{P}(T=t)}\\
=&\frac{\mathsf{P}(X_1=x_1)\mathsf{P}(X_2=x_2)\cdots\mathsf{P}(X_n=t-\sum_{i=1}^{n-1}x_i)}{\mathsf{P}(T=t)}\\
=&\frac{\theta(1-\theta)^{x_1}\cdots\theta(1-\theta)^{x_{n-1}}\theta(1-\theta)^{t-\sum_{i=1}^{n-1}x_i}}{\binom{t+n-1}{n-1}\theta^n(1-\theta)^t}\\
=&\binom{t+n-1}{n-1}^{-1}.
\end{aligned}$$

这个条件分布与参数θ无关,所以$T=\sum_{i=1}^{n} X_i$是参数θ的充分统计量.

Example

例 设样本$X_1, ..., X_n$ i.i.d., 服从密度为

$$p(x, \lambda) = \lambda e^{-\lambda x}, \qquad x > 0$$

的指数分布，其中$\lambda > 0$为未知参数，证明：样本均值$\overline{X}$为λ的充分统计量.

证：令$T=\sum_{j=1}^n X_j$. 因为指数分布$E(\lambda)=\Gamma(1,\lambda)$. 由gamma分布的可加性知$T\sim\Gamma(n,\lambda)$, 密度函数为

$$p_T(t;\theta)=[(n-1)!]^{-1}\lambda^n t^{n-1}e^{-\lambda t}, \qquad 0<t.$$

证：令$T=\sum_{j=1}^n X_j$. 因为指数分布$E(\lambda)=\Gamma(1,\lambda)$. 由gamma分布的可加性知$T\sim\Gamma(n,\lambda)$, 密度函数为

$$p_T(t;\theta)=[(n-1)!]^{-1}\lambda^n t^{n-1}e^{-\lambda t}, \qquad 0<t.$$

另一方面, 线性变换

$$\begin{cases} x_1=x_1 \\ \vdots \\ x_{n-1}=x_{n-1} \\ t=\sum_{i=1}^n x_i \end{cases}$$

的Jacobian行列式为

$$\frac{D(x_1,\cdots,x_{n-1},x_n)}{D(x_1,\cdots,x_{n-1},t)}=1.$$

所以$X_1,\cdots,X_{n-1},T$的联合密度函数为

$$\begin{aligned} p(x_1,...,x_{n-1},t;\theta) =&\lambda e^{-\lambda x_1}\cdots\lambda e^{-\lambda x_{n-1}}\lambda e^{-\lambda(t-\sum_{i=1}^{n-1}x_i)} \\ =&\lambda^n e^{-\lambda t}, \end{aligned}$$

$$0<x_i, i=1,2,\cdots,n-1, t\geq\sum_{i=1}^{n-1}x_i.$$

所以$X_1,\cdots,X_{n-1},T$的联合密度函数为

$$\begin{aligned} p(x_1,...,x_{n-1},t;\theta) =&\lambda e^{-\lambda x_1}\cdots\lambda e^{-\lambda x_{n-1}}\lambda e^{-\lambda(t-\sum_{i=1}^{n-1}x_i)} \\ =&\lambda^n e^{-\lambda t}, \end{aligned}$$

$$0<x_i, i=1,2,\cdots,n-1, t\ge\sum_{i=1}^{n-1}x_i.$$

因此当给定$T=t$时，$X_1,X_2,\cdots,X_{n-1}$的条件密度为

$$p_{X_1,\cdots,X_{n-1}|T}(x_1,...,x_{n-1}|t)=(n-1)!t^{-(n-1)},$$

$$0<x_i, i=1,2,\cdots,n-1, t\ge\sum_{i=1}^{n-1}x_i.$$

所以$X_1, \cdots, X_{n-1}, T$的联合密度函数为

$$\begin{aligned} p(x_1, ..., x_{n-1}, t; \theta) =& \lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_{n-1}} \lambda e^{-\lambda(t - \sum_{i=1}^{n-1} x_i)} \\ =& \lambda^n e^{-\lambda t}, \end{aligned}$$

$$0 < x_i, i = 1, 2, \cdots, n-1, t \geq \sum_{i=1}^{n-1} x_i.$$

因此当给定$T = t$时，$X_1, X_2, \cdots, X_{n-1}$的条件密度为

$$p_{X_1, \cdots, X_{n-1}|T}(x_1, ..., x_{n-1}|t) = (n-1)! t^{-(n-1)},$$

$$0 < x_i, i = 1, 2, \cdots, n-1, t \geq \sum_{i=1}^{n-1} x_i.$$

与θ无关.

因此当给定$T=t$时，$X_1, X_2, \cdots, X_{n-1}$的条件分布与$\theta$无关.

因此当给定$T=t$时，$X_1, X_2, \cdots, X_{n-1}$的条件分布与$\theta$无关.

故当给定$T=t$时，$X_1, X_2, \cdots, X_n$的条件分布与θ无关.

因此当给定$T=t$时，$X_1, X_2, \cdots, X_{n-1}$的条件分布与$\theta$无关.
故当给定$T=t$时，$X_1, X_2, \cdots, X_n$的条件分布与θ无关.
因此T是参数θ的充分统计量,从而$\overline{X}=T/n$也是参数θ的充分统计量.

因子分解定理

根据充分统计量的定义及其解释，在对总体未知参数进行推断时，应在可能的情况下尽量找出关于未知参数的充分统计量. 虽然可以直接根据定义来验证一个统计量是否充分(这种做法通常比较繁琐), 但是无法提供寻找充分统计量的途径.

用$\boldsymbol{X}$记样本$(X_1, ..., X_n)$.

Theorem

(因子分解定理) 设样本$\boldsymbol{X}$的联合pdf或pmf为$p(\boldsymbol{x};\theta)$, 其中θ为未知参数. 则$T=T(\boldsymbol{X})$为(关于θ的) 充分统计量,当且仅当

$$p(\boldsymbol{x};\theta)=g(T(\boldsymbol{x});\theta)h(\boldsymbol{x}),$$

其中$g(t;\theta)$是定义在统计量$T(\boldsymbol{X})$取值空间$\mathcal{T}$上的函数, $h(\boldsymbol{x})$与θ无关.

证明: 此定理的严格证明需要测度论的知识，超出了本课程的范围. 下面只给出离散场合下的证明. 这时

$$p(\boldsymbol{x};\theta)=\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}).$$

对$t\in\mathcal{T}$, 令集合

$$A(t)=\{\boldsymbol{x}:T(\boldsymbol{x})=t\}.$$

充分性: 设$p(\boldsymbol{x};\theta)$有定理中的因子分解形式. 在给定$T=t$下, 当$\boldsymbol{x}\notin A(t)$时,

$$\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|T=t)=0;$$

充分性: 设$p(\boldsymbol{x};\theta)$有定理中的因子分解形式. 在给定$T=t$下, 当$\boldsymbol{x}\notin A(t)$时,

$$\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|T=t)=0;$$

当$\boldsymbol{x}\in A(t)$时,

$$\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|T=t)=\frac{\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x},T=t)}{\mathsf{P}_\theta(T=t)}$$
$$=\frac{\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x})}{\mathsf{P}_\theta(T=t)}=\frac{p(\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}\in A(t)}p(\boldsymbol{y};\theta)}$$
$$=\frac{g(t;\theta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A(t)}g(t;\theta)h(\boldsymbol{y})}=\frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A(t)}h(\boldsymbol{y})}.$$

此条件分布与参数θ无关. 所以$T(\boldsymbol{X})$是充分统计量.

必要性:设$T(\boldsymbol{X})$是参数θ的充分统计量, 由定义知, $\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|T=t)$与参数$\theta$无关, 它只能是$\boldsymbol{x}$的函数, 记之为$h(\boldsymbol{x})$.

必要性: 设$T(\boldsymbol{X})$是参数θ的充分统计量, 由定义知, $\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|T=t)$与参数$\theta$无关, 它只能是$\boldsymbol{x}$的函数, 记之为$h(\boldsymbol{x})$.

对任意的$\boldsymbol{x}$, 记$t=T(\boldsymbol{x})$. 则

$$
\begin{aligned}
p(\boldsymbol{x};\theta) =&\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}) \\
=&\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}, T=t) \\
=&\mathsf{P}_\theta(\boldsymbol{X}=\boldsymbol{x}|T=t)\mathsf{P}_\theta(T=t) \\
=&h(\boldsymbol{x})g(t;\theta)=h(\boldsymbol{x})g(T(\boldsymbol{x});\theta).
\end{aligned}
$$

显然满足因子分解条件. 这样就证明了定理的必要性.

Example

设$\boldsymbol{X} = (X_1, ..., X_n)$为取自在$(0, \theta)$区间上均匀分布总体的样本，其中$\theta > 0$为未知参数. 样本的联合概率密度函数为

Example

设$\boldsymbol{X} = (X_1, ..., X_n)$为取自在$(0, \theta)$区间上均匀分布总体的样本，其中$\theta > 0$为未知参数. 样本的联合概率密度函数为

$$p(\boldsymbol{x}; \theta) = \theta^{-n} I_{\{0 < x_{(1)} \le x_{(n)} < \theta\}} = \theta^{-n} I_{\{x_{(n)} < \theta\}} \cdot I_{\{0 < x_{(1)} \le x_{(n)}\}},$$

其中I_A为集合A的示性函数, $x_{(1)} = \min\{x_i, i = 1, ..., n\}$, $x_{(n)} = \max\{x_i, i = 1, ..., n\}$. 因此可以定义统计量$T = X_{(n)} = \max\{X_i, i = 1, ..., n\}$.
令$g(t; \theta) = \theta^{-n} I_{\{t < \theta\}}$, $h(\boldsymbol{x}) = I_{\{0 < x_{(1)} \le x_{(n)}\}}$, 则$p(\boldsymbol{x}; \theta)$满足因子分解条件, 因而$X_{(n)}$是充分统计量.

Example

次序统计量是充分统计量.

Example

次序统计量是充分统计量.

设$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$为样本$(X_1, X_2, \cdots, X_n)$的次序统计量, 总体有密度函数(或分布列函数)$p_\theta(x)$. 则样本的联合密度函数(或分布列函数)为

$$p(\boldsymbol{x};\theta) = p_\theta(x_1)p_\theta(x_2)\cdots p_\theta(x_n) = p_\theta(x_{(1)})p_\theta(x_{(2)})\cdots p_\theta(x_{(n)}).$$

Example

设样本$X_1, ..., X_n$取自正态总体$\sim N(\mu, \sigma^2)$, 其中μ和σ^2均未知. 样本的联合密度函数为

$$\begin{aligned} p(\boldsymbol{x}; \mu, \sigma^2) =& (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ =& (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right\}. \end{aligned}$$

Example

设样本$X_1, ..., X_n$取自正态总体$\sim N(\mu, \sigma^2)$, 其中μ和σ^2均未知. 样本的联合密度函数为

$$\begin{aligned} p(\boldsymbol{x}; \mu, \sigma^2) =&(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\} \\ =&(2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)\right\}. \end{aligned}$$

取$h(\boldsymbol{x}) = 1$, 由因子分解定理知,
$(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$是$(\mu, \sigma^2)$的充分统计量.

Example

设样本$X_1, ..., X_n$取自正态总体$\sim N(\mu, \sigma^2)$, 其中μ和σ^2均未知. 则由前一页知$\widetilde{T} = (T_1, T_2)$为$(\mu, \sigma^2)$的充分统计量, 其中$T_1 = \sum_i X_i^2$, $T_2 = \sum_i X_i$. 又$(\overline{X}, S^2)$为$\widetilde{T}$的一一变换, 所以$(\overline{X}, S^2)$为(μ, σ^2)的充分统计量.

利用因子分解定理可以对指数型分布族找到充分统计量.

假定样本的pdf或pmf所在的分布族是指数型分布族, 即

$$p(\boldsymbol{x};\theta)=c^*(\theta)\exp\left\{\sum_{j=1}^{k}Q_j(\theta)T_j^*(\boldsymbol{x})\right\}h^*(\boldsymbol{x}),$$

则$\widetilde{T}=(T_1^*,...,T_k^*)$为充分统计量.

极小充分统计量

Example

设$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$是取自$B(1,p)$的一个样本，$0<p<1$. 样本的联合分布列(pmf)为

$$
\begin{aligned}
&P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)\\
&=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\cdot I\{x_i=0,1,i=1,2,\cdots,n\}.
\end{aligned}
$$

取$h(\boldsymbol{x}) = I\{x_i = 0, 1, i = 1, 2, \cdots, n\}$,则由因子分解定理,知

$$
\begin{aligned}
T_1 =& (X_1, X_2, \cdots, X_n) \\
T_2 =& (X_1 + X_2, X_3, \cdots, X_n) \\
& \cdots \\
T_k =& (X_1 + X_2 + \cdots + X_k, X_{k+1}, \cdots, X_n) \\
& \cdots \\
T_n =& X_1 + X_2 + \cdots + X_n
\end{aligned}
$$

都是p的充分统计量.

定义 设S是分布族$\mathcal{F}$的充分统计量，假如对$\mathcal{F}$的任意一个充分统计量T，存在一个函数$f(\cdot)$，使得

$$S = f(T),$$

则称S是此分布族$\mathcal{F}$的极小充分统计量.

Theorem

设总体是指数型分布族，则样本$\boldsymbol{X}$的联合pdf或pmf可以写为

$$p(\boldsymbol{x};\theta)=c^*(\theta)\exp\left\{\sum_{j=1}^{k}Q_j(\theta)T_j^*(\boldsymbol{x})\right\}h^*(\boldsymbol{x}),$$

若$\widetilde{Q}(\theta)=(Q_1(\theta),\cdots,Q_k(\theta))$的值域有非空的内部.
则$\widetilde{T}=(T_1^*,...,T_k^*)$为极小充分统计量.

§2.8 完全统计量

分布族完备性(或称完全性)

§2.8 完全统计量

分布族完备性(或称完全性)

Definition

定义 设$\boldsymbol{X}$的分布所在的分布族为$\mathcal{F}=\{F(\boldsymbol{x};\theta):\theta\in\Theta\}$. 假如对任一个实函数$\varphi(\boldsymbol{x})$, 由

$$\mathsf{E}_\theta\varphi(\boldsymbol{X})=0,\quad \forall\theta\in\Theta$$

可推出

$$\mathsf{P}_\theta\{\varphi(\boldsymbol{X})=0\}=1,\quad \forall\theta\in\Theta.$$

则称此分布族$\mathcal{F}$是完备的(或称为完全的).

若分布族中的分布$F(\boldsymbol{x};\theta)$ 有密度$p(\boldsymbol{x};\theta)$, 完全性即为

$$\int \varphi(\boldsymbol{x})p(\boldsymbol{x};\theta)d\boldsymbol{x} = 0, \quad \forall\theta \in \Theta$$

$$\Downarrow$$

$$\varphi(\boldsymbol{x}) = 0 \ \ a.e.$$

若分布族中的分布$F(\boldsymbol{x};\theta)$ 有密度$p(\boldsymbol{x};\theta)$, 完全性即为

$$\int \varphi(\boldsymbol{x})p(\boldsymbol{x};\theta)d\boldsymbol{x} = 0, \quad \forall\theta \in \Theta$$

$$\Downarrow$$

$$\varphi(\boldsymbol{x}) = 0 \;\; a.e.$$

即 "若φ与函数系$\{p(\boldsymbol{x},\theta);\theta \in \Theta\}$正交, 则$\varphi$必为0", 也就是说函数系$\{p(\boldsymbol{x},\theta);\theta \in \Theta\}$是完全的.

Example

二项分布族$\{B(n,p): 0 < p < 1\}$是完备的.

Example

二项分布族$\{B(n,p): 0<p<1\}$是完备的.

证：假如$\varphi(x)$满足

$$\mathsf{E}_p\varphi(X)=\sum_{x=0}^{n}\varphi(x)\binom{n}{x}p^x(1-p)^{n-x}=0,\quad 0<p<1.$$

Example

二项分布族$\{B(n,p): 0<p<1\}$是完备的.

证：假如$\varphi(x)$满足

$$\mathsf{E}_p\varphi(X)=\sum_{x=0}^{n}\varphi(x)\binom{n}{x}p^x(1-p)^{n-x}=0,\quad 0<p<1.$$

令$\theta=p/(1-p)$,则上式可改写为

$$\sum_{x=0}^{n}\binom{n}{x}\varphi(x)\theta^x=0,\quad \theta>0.$$

上式为θ的多项式，而$\binom{n}{x}\varphi(x)$是多项式的系数，所以$\varphi(x)=0\ (x=0,1,\cdots,n)$.故$\mathsf{P}\{\varphi(X)=0\}=1$.所以二项分布族是完备的.

Example

正态分布族$\{N(0,\sigma^2):\sigma>0\}$是不完备的.

Example

正态分布族$\{N(0,\sigma^2):\sigma>0\}$是不完备的.

$$\varphi(x)=x.$$

Example

Gamma分布族$\{gamma(\alpha, \lambda) : \lambda > 0\}$是完备的(其中$\alpha > 0$已知).

Example

Gamma分布族$\{gamma(\alpha,\lambda):\lambda>0\}$是完备的(其中$\alpha>0$已知).

证明:设实函数$\varphi(t)$ 满足$\mathsf{E}_\lambda\varphi(\Gamma)=0$, 即

$$\int_0^\infty \varphi(t)t^{\alpha-1}e^{-\lambda t}dt=0, \forall\lambda>0.$$

Example

Gamma分布族$\{gamma(\alpha,\lambda):\lambda>0\}$是完备的(其中$\alpha>0$已知).

证明:设实函数$\varphi(t)$ 满足$\mathsf{E}_\lambda\varphi(\Gamma)=0$, 即

$$\int_0^\infty \varphi(t)t^{\alpha-1}e^{-\lambda t}dt=0, \forall\lambda>0.$$

上式左边是$\varphi(t)t^{\alpha-1}$的Laplace变换, 由Laplace变换的唯一性知:

$$\varphi(t)t^{\alpha-1}=0 \quad a.e. \text{ 故 } \varphi(t)=0 \quad a.e..$$

所以

$$\mathsf{P}_\lambda\{\varphi(\Gamma)=0\}=1, \quad \lambda>0.$$

Example

正态分布族$\{N(\mu,\sigma^2): -\infty < \mu < \infty\}$是完备的,其中$\sigma$已知.

Example

正态分布族$\{N(\mu,\sigma^2):-\infty<\mu<\infty\}$是完备的,其中$\sigma$已知.

证明:设实函数$\varphi(x)$ 满足$\mathsf{E}_\mu\varphi(X)=0$, 其中$X\sim N(\mu,\sigma^2)$, 即

$$\int_{-\infty}^{\infty}\varphi(x)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx=0\ \ \forall\mu,$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\mu^2}{2\sigma^2}}\int_{-\infty}^{\infty}\varphi(x)e^{-\frac{x^2}{2\sigma^2}}e^{\frac{\mu}{\sigma^2}x}dx=0\ \ \forall\mu.$$

也即

$$\int_{-\infty}^{\infty}\left[\varphi(x)e^{-\frac{x^2}{2\sigma^2}}\right]e^{-\lambda x}dx=0\quad\forall\lambda.$$

上式左边是$\varphi(x)e^{-\frac{x^2}{2\sigma^2}}$的Laplace变换，从而

$$\varphi(x)e^{-\frac{x^2}{2\sigma^2}}=0\quad a.e.$$

即$\varphi(x)=0$ a.e. 所以

$$\mathsf{P}\{\varphi(X)=0\}=1.$$

统计量的完全性

Definition

设有参数分布族$\mathcal{F} = \{F(x;\theta) : \theta \in \Theta\}$. T 是一个统计量. 假如对任一实函数$\varphi(t)$, 由

$$\mathsf{E}_\theta \varphi(T) = 0, \quad \forall \theta \in \Theta$$

可推出

$$\mathsf{P}_\theta\{\varphi(T) = 0\} = 1, \forall \theta \in \Theta,$$

也就是说, 由T诱导出的分布族$\mathcal{F}^T = \{F_T(t;\theta) : \theta \in \Theta\}$是完备的, 则称$T$是完备统计量.

Example

$\{N(0,\sigma^2):\sigma>0\}$不完备, 但统计量$T=X_1^2+\cdots+X_n^2\sim gamma(\frac{n}{2},\frac{1}{2\sigma^2})$是完备统计量.

Example

设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$ 的样本. 证明$(\overline{X}, S^2)$ 是完备统计量.

Example

设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$ 的样本. 证明$(\overline{X}, S^2)$ 是完备统计量.

证明: 设有$\varphi(x, t)$使得$E_{\mu,\sigma^2}\varphi(\overline{X}, S^2) = 0$.

Example

设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$ 的样本. 证明$(\overline{X}, S^2)$ 是完备统计量.

证明: 设有$\varphi(x,t)$使得$E_{\mu,\sigma^2}\varphi(\overline{X}, S^2) = 0$. 由于$\overline{X} \sim N(\mu, \sigma^2/n)$, $(n-1)S^2/\sigma^2 \sim \chi^2(n-1) = \Gamma(\frac{n-1}{2}, \frac{1}{2})$, 所以$S^2 \sim \Gamma(\frac{n-1}{2}, \frac{n-1}{2\sigma^2})$.且$\overline{X}$ 和S^2相互独立.

Example

设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$ 的样本. 证明$(\overline{X}, S^2)$ 是完备统计量.

证明: 设有$\varphi(x,t)$使得$E_{\mu,\sigma^2}\varphi(\overline{X}, S^2) = 0$. 由于$\overline{X} \sim N(\mu, \sigma^2/n)$, $(n-1)S^2/\sigma^2 \sim \chi^2(n-1) = \Gamma(\frac{n-1}{2}, \frac{1}{2})$, 所以$S^2 \sim \Gamma(\frac{n-1}{2}, \frac{n-1}{2\sigma^2})$.且$\overline{X}$ 和S^2相互独立. 从而

$$\int_{-\infty}^{\infty}\left[\int_0^{\infty}\varphi(x,t)f_{\sigma^2}(t)dt\right]\frac{1}{\sqrt{2\pi\sigma^2/n}}e^{-\frac{(x-\mu)^2}{2\sigma^2/n}}dx = 0, \quad \forall \sigma > 0, \mu,$$

其中$f_{\sigma^2}(t)$为$\Gamma(\frac{n-1}{2}, \frac{n-1}{2\sigma^2})$的密度函数.

Example

设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$ 的样本. 证明$(\overline{X}, S^2)$ 是完备统计量.

证明: 设有$\varphi(x,t)$使得$E_{\mu,\sigma^2}\varphi(\overline{X}, S^2) = 0$. 由于$\overline{X} \sim N(\mu, \sigma^2/n)$, $(n-1)S^2/\sigma^2 \sim \chi^2(n-1) = \Gamma(\frac{n-1}{2}, \frac{1}{2})$, 所以$S^2 \sim \Gamma(\frac{n-1}{2}, \frac{n-1}{2\sigma^2})$.且$\overline{X}$ 和S^2相互独立. 从而

$$\int_{-\infty}^{\infty}\left[\int_0^{\infty}\varphi(x,t)f_{\sigma^2}(t)dt\right]\frac{1}{\sqrt{2\pi\sigma^2/n}}e^{-\frac{(x-\mu)^2}{2\sigma^2/n}}dx = 0, \quad \forall\sigma > 0, \mu,$$

其中$f_{\sigma^2}(t)$为$\Gamma(\frac{n-1}{2}, \frac{n-1}{2\sigma^2})$的密度函数. 先固定$\sigma^2$, 上式对$\forall\mu$成立, 由正态分布族$\{N(\mu, \sigma^2/n), -\infty < \mu < \infty\}$的完备性知

$$\int_0^{\infty} \varphi(x,t) f_{\sigma^2}(t)dt = 0, a.e. x, \quad \forall \sigma > 0.$$

$$\int_0^{\infty} \varphi(x,t) f_{\sigma^2}(t)dt = 0, a.e.x, \quad \forall \sigma > 0.$$

再由gamma分布族的完备性知道

$$\varphi(x,t) = 0, a.e.(x,t).$$

所以$(\overline{X}, S^2)$是完备统计量.

Example

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

证明： $\boldsymbol{X}$的密度函数为

$$p(\boldsymbol{x};\theta)=\prod_{i=1}^{n} 1\cdot I\{\theta-1/2<x_i<\theta+1/2\}$$
$$=I\{\theta-1/2<x_{(1)}\le x_{(n)}<\theta+1/2\}.$$

证明: $\boldsymbol{X}$ 的密度函数为

$$\begin{aligned} p(\boldsymbol{x};\theta) =& \prod_{i=1}^{n} 1 \cdot I\{\theta - 1/2 < x_i < \theta + 1/2\} \\ =& I\{\theta - 1/2 < x_{(1)} \le x_{(n)} < \theta + 1/2\}. \end{aligned}$$

由因子分解定理, $T = (X_{(1)}, X_{(n)})$ 为充分统计量.

为证T不完备,只要找一个非0的函数$\varphi(T)$, 使得$\mathsf{E}_\theta\left[\varphi(T)\right] = 0$.

为证T不完备,只要找一个非0的函数$\varphi(T)$, 使得$\mathsf{E}_\theta\left[\varphi(T)\right]=0.$

考虑$X_{(n)}-X_{(1)}$. 令$Y_i=X_i-(\theta-1/2)$. 则$Y_1,\cdots,Y_n$为独立的$U(0,1)$变量, 分布不依赖于参数θ,

为证T不完备,只要找一个非0的函数$\varphi(T)$, 使得$\mathsf{E}_\theta\left[\varphi(T)\right]=0$.

考虑$X_{(n)}-X_{(1)}$. 令$Y_i=X_i-(\theta-1/2)$. 则$Y_1,\cdots,Y_n$为独立的$U(0,1)$变量, 分布不依赖于参数θ, 且

$$X_{(n)}-X_{(1)}=Y_{(n)}-Y_{(1)}.$$

所以

$$\mathsf{E}_\theta\left[X_{(n)}-X_{(1)}\right]=\mathsf{E}[Y_{(n)}-Y_{(1)}]=c_n=\frac{n-1}{n+1}.$$

为证T不完备,只要找一个非0的函数$\varphi(T)$, 使得$\mathsf{E}_\theta\left[\varphi(T)\right] = 0$.

考虑$X_{(n)} - X_{(1)}$. 令$Y_i = X_i - (\theta - 1/2)$. 则$Y_1, \cdots, Y_n$为独立的$U(0,1)$变量, 分布不依赖于参数$\theta$, 且

$$X_{(n)} - X_{(1)} = Y_{(n)} - Y_{(1)}.$$

所以

$$\mathsf{E}_\theta\left[X_{(n)} - X_{(1)}\right] = \mathsf{E}[Y_{(n)} - Y_{(1)}] = c_n = \frac{n-1}{n+1}.$$

取$\varphi(T) = X_{(n)} - X_{(1)} - c_n$. 则

$$\mathsf{E}_\theta\left[\varphi(T)\right] = 0, \quad \forall \theta.$$

指数型分布族中统计量的完全性

Theorem

设总体来自指数型分布族,从总体中抽取的样本的联合pdf或pmf为

$$p(\widetilde{x};\theta)=c^*(\theta)\exp\left\{\sum_{j=1}^{k}Q_j(\theta)T_j^*(\widetilde{x})\right\}h^*(\widetilde{x}),$$

若$\mathbf{Q}=(Q_1(\theta),\cdots,Q_k(\theta))$的值域有非空的内部.

则$T=(T_1^*,...,T_k^*)$为充分完备统计量.

Example

设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$ 的样本. 证明$(\overline{X}, S^2)$ 是完备统计量.

Example

设$X_1, X_2, \cdots, X_n$是来自正态总体$N(\mu, \sigma^2)$ 的样本. 证明$(\overline{X}, S^2)$ 是完备统计量.

解: $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$的联合密度函数为

$$\begin{aligned} p(\boldsymbol{x}; \mu, \sigma) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\} \\ &\quad \cdot \exp\left\{\left[\sum_i x_i\right]\left[\frac{\mu}{\sigma^2}\right] + \left[\sum_i x_i^2\right]\left[-\frac{1}{2\sigma^2}\right]\right\} \end{aligned}$$

$(\mu/\sigma^2, -1/(2\sigma^2))$的值域有非空的内部.

故$(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$是充分完备统计量.

从而它一一变换$(\overline{X}, S^2)$也是充分完备统计量.

完备统计量的一个性质

Theorem

(Basu定理) 设X的分布所在的分布族为$\mathcal{F}=\{p(x;\theta):\theta\in\Theta\}$, $\boldsymbol{X}=(X_1,\cdots,X_n)$是从分布族$\mathcal{F}$中抽取的简单样本. 设$T=T(\boldsymbol{X})$是充分完备统计量. 若随机变量$V=V(\boldsymbol{X})$的分布与参数$\theta$无关, 则对任何$\theta\in\Theta$, V与T独立.

Example

设$(X_1,\cdots,X_n)$是抽取自正态总体$N(\mu,\sigma^2)$的简单样本，则$(\overline{X},S^2)$为充分完备统计量. 偏度系数

$$\begin{aligned}\frac{m_3}{m_2^{3/2}} &= \frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^3}{\left(\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\right)^{3/2}}\\ &= \frac{\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})^3}{\left(\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})^2\right)^{3/2}},\end{aligned}$$

$$\left(\text{其中} Y_i=\frac{X_i-\mu}{\sigma}, i=1,\cdots,n, \text{i.i.d.} \sim N(0,1)\right)$$

Example

设$(X_1,\cdots,X_n)$是抽取自正态总体$N(\mu,\sigma^2)$的简单样本，则$(\overline{X},S^2)$为充分完备统计量. 偏度系数

$$\frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^3}{\left(\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^2\right)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})^3}{\left(\frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})^2\right)^{3/2}},$$

$$\left(\text{其中} Y_i = \frac{X_i-\mu}{\sigma}, i=1,\cdots,n, \text{i.i.d.} \sim N(0,1)\right)$$

的分布与参数无关, 从而它与$(\overline{X},S^2)$独立.