§6.4 充分统计量

- 构造统计量就是对样本进行加工，去粗取精，简化样本，便于统计推断。
- 但在加工过程中是否会丢失样本中关于感兴趣问题的的信息？
- 如果某个统计量包含了样本中关于感兴趣问题的“全部信息”，则这个统计量对将来的统计推断会非常有用，这就是充分统计量的直观含义，它是费希尔（Fisher. R. A.）于1922年正式提出的。

[引例]

设$\xi_1,\xi_2,\cdots,\xi_n$是取自两点分布母体ξ的一个子样,其概率函数为

$$f(x;\theta)=\theta^x(1-\theta)^{1-x}\quad x=0,1\quad 0<\theta<1$$

θ为未知参数

若设统计量$\eta=\sum_{i=1}^{n}\xi_i$,则相应地,观测值$y=\sum_{i=1}^{n}x_i$

由简单随机子样的性质,$\eta\sim B(n,\theta)$,其概率函数为:

$$g(y;\theta)=C_n^y\theta^y(1-\theta)^{n-y}\qquad y=0,1,2,\cdots,\ n.$$

令事件A表示$\{\xi_1=x_1,\xi_2=x_2,\cdots,\xi_n=x_n\}$,$B$表示$\{\eta=y\}$

考虑 $P(A|B)=\dfrac{P(AB)}{P(B)}=\dfrac{P(A)}{P(B)}$

易知,$A\subset B,AB=A$

$$=\frac{P\{\xi_1=x_1\}\cdot P\{\xi_2=x_2\}\cdots P\{\xi_n=x_n\}}{P\{\eta=y\}}$$

$$= \frac{P\{\xi_1 = x_1\} \cdot P\{\xi_2 = x_2\} \cdots P\{\xi_n = x_n\}}{P\{\eta = y\}}$$

$$= \frac{\theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i}}{C_n^{\sum_{i=1}^{n} x_i} \cdot \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i}} = \frac{1}{C_n^{\sum_{i=1}^{n} x_i}}$$

这说明条件概率 $P(A \mid B)$ 不含有未知参数 θ 的信息，因此未知参数 θ 的信息全部包含在 $B = \{\eta = y\}$ 之中，称统计量 η 为 θ 的充分统计量.

定义6.7

设$\xi_1,\xi_2,\cdots,\xi_n$是取自具有概率函数$f(x;\theta),\theta\in\Theta$的母体$\xi$的一个容量为$n$的子样. 设$\eta=u(\xi_1,\xi_2,\cdots,\xi_n)$是一个统计量, 具有概率函数$g(y;\theta)$.

若 $$\frac{f(x_1;\theta)\cdot f(x_2;\theta)\cdots f(x_n;\theta)}{g(y;\theta)}=h(x_1,x_2,\cdots x_n)$$

$$\Leftrightarrow f(x_1;\theta)\cdot f(x_2;\theta)\cdots f(x_n;\theta)=g(y;\theta)\,h(x_1,x_2,\cdots x_n)$$

其中因子h不依赖于参数θ.

则称η为θ的一个充分统计量.(*Fisher*因子分解定理)

例6.16

设母体具有密度函数 $f(x;\theta)=\begin{cases} e^{-(x-\theta)} & \theta<x<\infty,\theta\in\Theta \\ 0 & \text{其他} \end{cases}$，试证：

最小次序统计量 $\xi_{(1)}$ 是 θ 的充分统计量.

(1) $f(x_1;\theta)\cdot f(x_2;\theta)\cdots f(x_n;\theta)=e^{-\sum_{i=1}^{n}x_i+n\theta}\,I(x_1>\theta,\cdots,x_n>\theta)$

其中 I 表示集合的示性函数.

由题意可知，$F(x)=\begin{cases} 0 & x\le\theta \\ 1-e^{-(x-\theta)} & \theta<x<\infty \end{cases}$

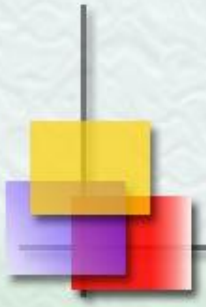
(2) $g(y;\theta)=\begin{cases} n[1-(1-e^{-(y-\theta)})]^{n-1}e^{-(y-\theta)} & \theta<y<\infty \\ 0 & \text{其他} \end{cases}$

$=\begin{cases} ne^{-n(y-\theta)} & \theta<y<\infty \\ 0 & \text{其他} \end{cases}$

$$\frac{f(x_1;\theta)\cdot f(x_2;\theta)\cdots\cdots f(x_n;\theta)}{g(y;\theta)}=\frac{e^{-\sum_{i=1}^{n}x_i+n\theta}}{ne^{-n(y-\theta)}}$$

$$=\frac{e^{-\sum_{i=1}^{n}x_i}\cdot e^{n\theta}}{ne^{-ny}\cdot e^{n\theta}}=\frac{e^{-\sum_{i=1}^{n}x_i}}{ne^{-ny}}=\frac{e^{-\sum_{i=1}^{n}x_i}}{ne^{-n(\min\limits_{1\le i\le n}x_i)}}$$

上述表达式与θ无关,所以$\xi_{(1)}$是一个充分统计量.



例6.17

设母体服从泊松分布，$f(x;\lambda)=\begin{cases}\dfrac{\lambda^x}{x!}e^{-\lambda} & \lambda>0, x=0,1,2,\cdots \\ 0 & \text{其他}\end{cases}$

λ是未知参数.取子样$(\xi_1^{(\omega)},\xi_2^{(\omega)})=(x_1,x_2)=(1,2)$,则试证：$\eta=2\xi_1+\xi_2$不是$\lambda$的充分统计量.$(P_{291})$

解：因为$\eta=2\xi_1+\xi_2$,则$y=2x_1+x_2=4$

$$(1)\ f(x_1;\lambda)\cdot f(x_2;\lambda)=f(1;\lambda)\cdot f(2;\lambda)=\frac{\lambda^1}{1!}e^{-\lambda}\cdot\frac{\lambda^2}{2!}e^{-\lambda}=\frac{\lambda^3}{2}e^{-2\lambda}$$

$(2)\eta=y=4$时的概率函数为：

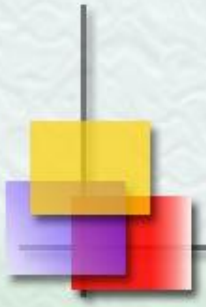
$$g(4;\lambda)=P\{\xi_1=0,\xi_2=4\}+P\{\xi_1=1,\xi_2=2\}+P\{\xi_1=2,\xi_2=0\}$$

$$=e^{-\lambda}\cdot\frac{\lambda^4}{4!}e^{-\lambda}+e^{-\lambda}\lambda\cdot\frac{\lambda^2}{2!}e^{-\lambda}+\frac{\lambda^2}{2!}e^{-\lambda}\cdot e^{-\lambda}$$

$$=e^{-2\lambda}\left(\frac{\lambda^4}{24}+\frac{\lambda^3}{2}+\frac{\lambda^2}{2}\right)$$

$$\frac{f(1;\lambda)f(2;\lambda)}{g(4;\lambda)}=\frac{\frac{\lambda^3}{2}e^{-2\lambda}}{(\frac{\lambda^4}{24}+\frac{\lambda^3}{2}+\frac{\lambda^2}{2})e^{-2\lambda}}=\frac{12\lambda}{12+12\lambda+\lambda^2}$$

所以,$\eta=2\xi_1+\xi_2$不是λ的充分统计量.



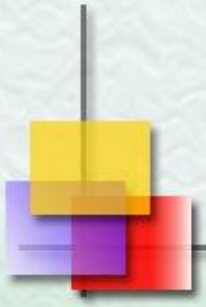
定理6.2 Neyman因子分解定理

*Th*6.2 :设$\xi_1,\xi_2,\cdots,\xi_n$为取自具有概率函数$f(x;\theta),\theta\in\Theta$的母体$\xi$的一个子样,则统计量$\eta=u(\xi_1,\xi_2,\cdots,\xi_n)$是一个充分统计量

$\Leftrightarrow$ $\exists$非负函数K_1和K_2,使得

$$f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta)=K_1[u(x_1,x_2,\cdots,x_n);\theta]K_2(x_1,x_2,\cdots,x_n)$$

且当$y=u(x_1,x_2,\cdots,x_n)$取定值时,函数K_2不依赖于θ.



补例：设$\xi_1,\xi_2,\cdots,\xi_n$是独立同分布的随机变量，都服从泊松分布，

则证明：$T_n=\sum_{i=1}^{n}\xi_i$是关于λ的充分统计量.

解：$f(x_1;\lambda)f(x_2;\lambda)\cdots f(x_n;\lambda)=\dfrac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}$

取$K_1=\lambda^{\sum_{i=1}^{n}x_i}e^{-n\lambda}$，$K_2=(\prod_{i=1}^{n}x_i!)^{-1}$与$\lambda$无关.

由因子分解定理，T_n是λ的充分统计量.

例6.19

设$\xi_1,\xi_2,\cdots\xi_n$是取自$[0,\theta]$上的均匀分布母体的一个子样，θ为未知参数，试证：$\xi_{(n)}$是θ的一个充分统计量.

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

$0\le x_i\le\theta,\ i=1,2,\cdots,n$

所以，$f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta)=\begin{cases}\dfrac{1}{\theta^n} & \max\limits_{1\le i\le n}x_i\le\theta,\min\limits_{1\le i\le n}x_i\ge 0\\ 0 & \text{其他}\end{cases}$

$$=\frac{1}{\theta^n}I(\max_{1\le i\le n}x_i\le\theta)I(\min_{1\le i\le n}x_i\ge 0)=\frac{1}{\theta^n}I(x_{(n)}\le\theta)I(x_{(1)}\ge 0)$$

其中I表示集合的示性函数.

取$K_1=\dfrac{1}{\theta^n}I(x_{(n)}\le\theta),K_2=I(x_{(1)}\ge 0)$，则由因子分解定理立得.

例6.18 设$\xi_1,\xi_2,\cdots,\xi_n$是来自正态总体$N(\mu,1)$的样本

则证明：$T_n=\frac{1}{n}\sum_{i=1}^{n}\xi_i=\bar{\xi}$是关于$\mu$的充分统计量.

解：$$f(x_1;\mu)f(x_2;\mu)\cdots f(x_n;\mu)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2}}$$

$$=\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2}}=\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2}{2}}$$

$$=e^{-\frac{n(\bar{x}-\mu)^2}{2}}\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2}}=K_1(\bar{x};\mu)K_2(x_1,x_2,\cdots,x_n)$$

由因子分解定理，T_n是λ的充分统计量.

无偏，有效，充分统计量，指数型分布

$$\sum_{i=1}^{n}\frac{\partial \ln f(\xi_i;\theta)}{\partial\theta}=K(\theta)(\eta-g(\theta))\text{以概率1成立},C-R\text{不等式等号成立}$$

$$\int\sum_{i=1}^{n}\frac{\partial \log f(\xi_i;\theta)}{\partial\theta}d\theta=A(\theta)\eta+B(\theta)+C(\xi_1,\cdots,\xi_n)$$

即 $$\int\frac{\partial}{\partial\theta}\left[\ln\prod_{i=1}^{n}f(\xi_i;\theta)\right]d\theta=A(\theta)\eta+B(\theta)+C(\xi_1,\cdots,\xi_n)$$

或 $$\ln\prod_{i=1}^{n}f(x_i;\theta)=A(\theta)y+B(\theta)+C(x_1,\cdots,x_n),y=u(x_1,\cdots,x_n)$$

$$L(\theta)=\prod_{i=1}^{n}f(x_i;\theta)=\exp\left\{A(\theta)y+B(\theta)+C(x_1,\cdots,x_n)\right\}$$

单参数指数族分布 $\{f(x;\theta),\theta\in\Theta\}$

$$f(x;\theta)=\exp\{c(\theta)T(x)+d(\theta)+S(x)\}I(a<x<b)$$

例6.20 $N(0,\sigma^2)$ **例6.21** $b(m,p)$, m已知

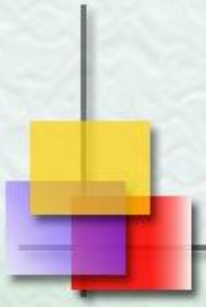
定理6.3 设$r.v.\xi$具有单参数指数族分布$f(x;\theta),\theta\in\Theta$, $\xi_1,\xi_2,\cdots,\xi_n$为来自总体ξ的一个样本,则统计量$\sum_{i=1}^{n}T(\xi_i)$是参数θ的充分统计量.

$$\prod_{i=1}^{n}f(x_i;\theta)=\exp\left\{c(\theta)\sum_{i=1}^{n}T(x_i)+nd(\theta)+\sum_{i=1}^{n}S(x_i)\right\}$$
$$I(a<x_{(1)}<x_{(n)}<b)$$

（补充）均方误差 (MSE) Mean Squared Error

相合性和渐近正态性是在大样本场合下评价估计好坏的两个重要标准,在样本量不是很大时,人们更加倾向于使用一此基于小样本的评价标准,此时,对无偏估计使用方差,对有偏估计使用均方误差.

$$MSE(\hat{\theta}) = E\left(\hat{\theta} - \theta\right)^2$$



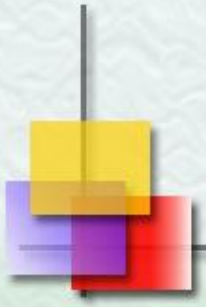
（补充）均方误差 (MSE) Mean Squared Error

$$MSE(\hat{\theta}) = E\left(\hat{\theta}-\theta\right)^2$$

$$= E\left(\hat{\theta}-E\hat{\theta}+E\hat{\theta}-\theta\right)^2$$

$$= E\left(\hat{\theta}-E\hat{\theta}\right)^2 + E\left(E\hat{\theta}-\theta\right)^2 + 2E\left[\left(\hat{\theta}-E\hat{\theta}\right)\left(E\hat{\theta}-\theta\right)\right]$$

$$= Var(\hat{\theta}) + \left(E\hat{\theta}-\theta\right)^2 + 0 \quad = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$



例如：正态总体下方差的估计量

定理5.4

$$S_{n-1}^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \qquad MSE(S_{n-1}^2)=E\left(S_{n-1}^2-\sigma^2\right)^2=\frac{2}{n-1}\sigma^4$$

$$S_n^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \qquad MSE(S_n^2)=E\left(S_n^2-\sigma^2\right)^2=\frac{2n-1}{n^2}\sigma^4$$

$$S_{n+1}^2=\frac{1}{n+1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \qquad MSE(S_{n+1}^2)=E\left(S_{n+1}^2-\sigma^2\right)^2=\frac{2}{n+1}\sigma^4$$

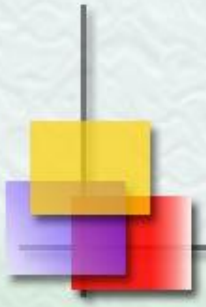
$$MSE(S_{n+1}^2)<MSE(S_n^2)<MSE(S_{n-1}^2)$$

一致最小方差无偏估计(UMVUE) P301

Uniformly Minimum-Variance Unbiased Estimator

定义6.8　设$\hat{\theta}_1$是θ的一个无偏估计量，若对θ的任何无偏估计量$\hat{\theta}$,都有$D\hat{\theta}_1 \le D\hat{\theta}$, 的则称$\hat{\theta}_1$是$\theta$的**UMVUE.**

注：有效估计一定是**UMVUE**，反之不真.



补充内容见PDF文件

作业： p.310 6.30（6.29，选做）， 6.42（6.40），6.43（6.41）