- **问题：** 任意一个统计量$T(\vec{X})$都提供了一种数据简化方式，那么这种简化会导致有关未知总体的信息丢失吗？
- **答案：** 如果一个统计量$T(\vec{X})$可以为相关问题提供与样本$\vec{X}=(X_1,\ldots,X_n)$一样的全部信息，那么我们就可以利用比样本$\vec{X}$更简化的统计量$T(\vec{X})$来进行统计分析。
- **充分性原理：** 如果$T(\vec{X})$是$\vec{\theta}$的一个充分统计量，则$\vec{\theta}$的任意依赖于样本$\vec{X}$的推断都可以经由$T(\vec{X})$值完成。
即，如果$\vec{x}$和$\vec{y}$是满足$T(\vec{x})=T(\vec{y})$的两个不同样本点，则不论观测到的是$\vec{X}=\vec{x}$还是$\vec{X}=\vec{y}$，关于$\vec{\theta}$的推断都完全相同。

- **定义5.1** （【0】定义2.7.1）设样本$\vec{X}$的分布族为$\mathcal{F}=\{f(\vec{x}|\vec{\theta}):\vec{\theta}\in\Theta\}$，这里$\Theta$为参数空间。如果样本$\vec{X}$的条件分布在已知统计量$T(\vec{X})$取值时，与$\vec{\theta}$无关，则称统计量$T$是参数$\vec{\theta}$的充分统计量(Sufficient Statistic).

注1 定义等价于$\mathbb{P}(\vec{X}=\vec{x}|T=t)$不依赖于$\vec{\theta}$;

注2 一个参数或参数向量$\vec{\theta}$的充分统计量不唯一;

注3 $\vec{\theta}$的充分统计量$T(\vec{X})$在一定意义下提炼了样本中有关$\vec{\theta}$的全部信息，即除充分统计量的值以外，样本中其余数据不能再提供关于$\vec{\theta}$的任何更多的信息。

注4 "充分性"概念依赖于给定分布族$\mathcal{F}$。举例而言，考虑参数统计模型$\mathcal{P}_i=\{P_\theta:P_\theta\in\mathcal{F}_i\}, i=1,2,3$, 分布族$\mathcal{F}_1\subset\mathcal{F}_2\subset\mathcal{F}_3$. 若$T$对于$\mathcal{P}_2$是充分统计量。则其对于$\mathcal{P}_1$仍是充分的，但对于$\mathcal{P}_3$不一定是充分的。

举例说明

Example (5.1)

[【0】例2.7.1] 设$X_1,\ldots,X_n$ $i.i.d. \sim Bernoulli(\theta)$, $\theta\in(0,1)$, 证明：统计量$T(\vec{X})=\sum_{i=1}^n X_i$是$\theta$的充分统计量。

Example (5.2)

设$X_1,\ldots,X_n$ $i.i.d. \sim N(\mu,\sigma^2)$，其中$\sigma^2>0$已知，$\mu\in\mathbb{R}$未知.
令$T(\vec{X})=X_1$, 则T不是充分统计量。

注 自行证明统计量$T(\vec{X})=\sum_{i=1}^n X_i$是$\theta$的充分统计量，如果

1. $X_1,\ldots,X_n$ $i.i.d. \sim N(\theta,\sigma^2)$，其中$\sigma^2>0$已知;
2. $X_1,\ldots,X_n$ $i.i.d. \sim Exp(\theta)$，其中θ是速率（rate）参数，即$\mathbb{E}X_i=\frac{1}{\theta}$.

Theorem (5.1, 因子分解定理 The factorization theorem)

[【0】定理2.7.1，【1】TH6.2.6] 设样本$\vec{X} = (X_1, \ldots, X_n)$的联合概率密度/质量函数（p.d.f./p.m.f.）为$f(\vec{x}|\vec{\theta})$。统计量$T(\vec{X})$是$\vec{\theta}$的充分统计量当且仅当存在函数$g(\vec{t}|\vec{\theta})$和$h(\vec{x})$，$s.t. \forall$ 样本点$\vec{x}$和参数向量$\vec{\theta}$，都有

$$f(\vec{x}|\vec{\theta}) = g(\vec{T}(\vec{x})|\vec{\theta})h(\vec{x}). \tag{1}$$

- **证明** 见【0】。

注 此处$h(\vec{x})$不依赖于$\vec{\theta}$。

Corollary (5.1)

[【0】推论2.7.1] 设$T = T(\vec{X})$为θ的充分统计量，若存在一个可测函数φ 和另一个统计量S，$s.t.$ $T = \varphi(S)$，则S也是θ的充分统计量。

举例说明

Example (5.3)

[【0】例2.7.9] 设$\vec{X} = (X_1, \ldots, X_n)$为服从概率密度函数（p.d.f）

$$f(\vec{x}|\vec{\theta}) = C(\vec{\theta})\exp\left\{\sum_{i=1}^{k} Q_i(\vec{\theta})T_i(\vec{x})\right\}h(\vec{x})$$

的指数族中抽取的简单样本，则$\vec{T}(\vec{X}) = (T_1(\vec{X}), \cdots, T_k(\vec{X}))$ 为充分统计量。

练习 自行利用例5.3和推论5.1，证明：

1. 对于正态总体$N(\mu, \sigma^2)$，则$\vec{T}(\vec{X}) = (\overline{X}, S^2)$是$\vec{\theta} = (\mu, \sigma^2)$的充分统计量；
2. 对于$Bernoulli(p)$，则$T(\vec{X}) = \sum_{i=1}^{n} X_i$是p的充分统计量。

- 上述练习参考【0】例2.7.5，2.7.6，先证是否是指数族。

Example (5.4)

[【0】例2.7.7] 设$X_1, \ldots, X_n$ $i.i.d. \sim U(0, \theta)$, $\theta > 0$, 证明：统计量$T(\vec{X}) = X_{(n)}$为θ的充分统计量。

作业1 设$X_1, \ldots, X_n$ $i.i.d. \sim Exp(\theta)$，证明：$T(\vec{X}) = X_{(1)}$不是充分统计量。

作业2 设$\vec{X} = (X_1, \ldots, X_n)$是从具有如下概率密度函数（p.d.f）的总体中抽取的简单样本

$$f(x|a,b) = c(a,b)\phi(x)1_{(a,b)}(x), \quad -\infty < a < b < \infty,$$

其中$0 < c(a,b) = [\int_a^b \phi(x)dx]^{-1} < \infty$. 证明: $T(\vec{X}) = (X_{(1)}, X_{(n)})$是$(a,b)$的充分统计量。

作业3 习题2：Ex.42, 43, 46.

- 一个统计模型可以有许多充分统计量。举例而言，
  1. 样本$\vec{X}$（整个数据集合）就是一个充分统计量；
  2. 回顾推论5.1，设$X_1, \ldots, X_n$ $i.i.d. \sim Bernoulli(p)$，令$\varphi(x,y) = x + y$, $x, y \in \mathbb{R}^2$, $S = (\sum_{i=1}^m X_i, \sum_{i=m+1}^n X_i)$，则$S$也是充分的. 但$T = \varphi(S)$显然比$S$更有用，因为在不丢失信息的前提下，$T$比$S$在数据的简化程度上更高。

问题 是否存在一个充分统计量实现最大程度的数据简化？

- **定义5.2**(【0】定义2.7.2，【1定义6.2.11】) 对于一个统计模型$\mathcal{P} = \{P_\theta : P_\theta \in \mathcal{F}\}$，设$T$是$\mathcal{P}$的充分统计量，则$T$ 称为**极小充分统计量**（Minimal Sufficient Statistic）当且仅当对于$\mathcal{P}$的任意其它充分统计量S，均存在可测函数φ，$s.t. T = \varphi(S)$, $a.s.\mathcal{P}$.

注1 如果一个结论对于除了事件A，并且$P_\theta(A) = 0, \forall P_\theta \in \mathcal{P}$之外，皆成立，那么我们就称这个结论$a.s.\mathcal{P}$ 成立。

注2 存在性：对于有分布族的参数统计模型，极小充分统计量总是存在的（Bahadur, 1957）.

*判别和寻找极小充分统计量

注3 唯一性：对于一个分布族，如果同时存在两个不同的极小充分统计量T和S，则由定义5.2即可知，存在一个$1-1$映射φ, $s.t.$ $T=\varphi(S)$, $a.s.\mathcal{P}$. 因此在"相互存在1-1映射即可被认为是同一（类）统计量"这个意义上，极小充分统计量是唯一。

- 直接利用定义5.2判别和寻找极小充分统计量不可行。

Theorem (5.2)

[【1】TH 6.2.13] 设$f(\vec{x}|\vec{\theta})$是样本$\vec{X}$的概率密度/质量函数（p.d.f./p.m.f.）. 如果存在函数$T(\vec{x})$, $s.t.\forall$两个样本点$\vec{x}$和$\vec{y}$，比值$\frac{f(\vec{x}|\vec{\theta})}{f(\vec{y}|\vec{\theta})}$是$\vec{\theta}$的常函数，当且仅当$T(\vec{x})=T(\vec{y})$，则$T(\vec{x})$是$\vec{\theta}$的极小充分统计量。

注 这里$\vec{\theta}$的常函数可与$\vec{x},\vec{y}$有关。

- 定义**5.3** 如果一个统计量$V(\vec{X})$的分布不依赖于$\vec{\theta}$，则称$V(\vec{X})$ 是辅助统计量（Ancillary Statistic）.
- **注意**
  1. 一个平凡的辅助统计量就是常数统计量$V(\vec{X}) \equiv c, c \in \mathbb{R}$;
  2. 对于一个统计量$S(\vec{X})$，如果存在函数V, $s.t. V(S(\vec{X}))$是不平凡的辅助统计量，则由$S(\vec{X})$ 生成的σ-代数中就包含一个非平凡的由$V(S(\vec{X}))$生成的σ-代数，这个$\sigma(V \circ S(\vec{X}))$显然不包含$\vec{\theta}$的任何信息，也即"$S(\vec{X})$可进一步简化"。
  3. 如果一个充分统计量T，它的任意不平凡函数表达式$\varphi(T)$ 均不是辅助统计量，则我们认为在简化数据方面，T是最成功的统计量之一，这样的T我们称之为充分完全统计量。

- **定义5.4** 设$\mathcal{F}=\{f(\vec{x}|\vec{\theta}),\vec{\theta}\in\Theta\}$是某一分布族，$\Theta$是参数空间，统计量$T(\vec{X})$称为完全统计量（Complete Statistic），当且仅当对于任意满足条件$\mathbb{E}_{\vec{\theta}}\varphi(T(\vec{X}))\equiv 0,\forall\vec{\theta}\in\Theta$的函数$\varphi$，都有$\varphi\equiv 0, a.s.\mathcal{P}$.
- **注意**
  1. 在上述条件中，如果对于任意有界（或a.s.有界）函数φ，都有$\varphi\equiv 0, a.s.\mathcal{P}$，则称$T(\vec{X})$为有界完全统计量（Boundedly Complete Statistic）。一个完全统计量总是有界完全的。
  2. 这里$\varphi(t)\equiv 0$中的t值只需仅限$T(\vec{X})$的所有可能取值;
  3. 如果T是完全的，则对任意可测函数ψ，$\psi(T)$仍是完全的。
  4. 对于任意统计量T，如果存在非平凡可测函数V，$s.t.V(T)$是辅助统计量，则T必不是完全统计量。

举例说明

Example (5.5)

[【0】例2.8.1] 设$X_1,\dots,X_n$ $i.i.d. \sim Bernoulli(p)$, $0<p<1$未知。证明：$T(\vec{X})=\sum_{i=1}^n X_i$是完全统计量。

Example (5.6)

[【0】例2.8.2] 设$X_1,\dots,X_n$ $i.i.d. \sim U(0,\theta)$, $\theta>0$未知。证明：$T(\vec{X})=X_{(n)}$是完全统计量。

练习 设$X_1,\dots,X_n$ $i.i.d. \sim N(\mu,\sigma^2)$, $\mu\in\mathbb{R}$未知, $\sigma^2>0$已知。证明：$T(\vec{X})=\sum_{i=1}^n X_i$ 是完全统计量。（参考【0】例2.8.3）.

特例：指数族

Theorem (5.3)

[【0】定理2.8.1] 设简单样本$\vec{X}=(X_1,\ldots,X_n)$的总体是具有如下自然参数形式的概率密度函数的指数族

$$f(x|\vec{\theta})=C(\vec{\theta})\exp\left\{\sum_{i=1}^{k}\theta_i T_i(x)\right\}h(x),$$

其中$\vec{\theta}=(\theta_1,\cdots,\theta_k)\in\Theta$. 如果参数空间$\Theta$作为$\mathbb{R}^k$的子集有内点，则统计量$\vec{T}(\vec{X})=(\sum_{j=1}^{n}T_1(X_j),\cdots,\sum_{j=1}^{n}T_k(X_j))$是完全统计量。

- 定理条件“自然参数形式”可去除，见【1】定理6.2.25。

Example (5.7)

设$X_1,\ldots,X_n$ $i.i.d.\sim N(\mu,\sigma^2)$, $\mu\in\mathbb{R},\sigma^2>0$未知。证明:
$T(\vec{X})=(\overline{X},S^2)$是完全统计量。

Theorem (5.4, Basu's Theorem)

[【0】定理2.8.2] 设随机样本$\overrightarrow{X}$取自服从统计模型$\mathcal{P}=\{P_\theta,\theta\in\Theta\}$的总体，$V$和$T$是它的两个统计量。如果$V$是辅助统计量，而$T$是（有界）完全充分统计量，则$V$和$T$关于$\mathcal{P}$相互独立。

- **证明** 参考【1】定理6.2.24证明。
- **应用**

Example (5.8)

[【0】例2.8.7] 设$X_1,\ldots,X_n$ $i.i.d.\sim N(\mu,\sigma^2)$, $\sigma^2>0$已知, $\mu\in\mathbb{R}$未知。证明：$T(\overrightarrow{X})=\overline{X}$与极差$R=X_{(n)}-X_{(1)}$独立。

Example (5.9)

[【1】例6.2.26] 设$X_1, \ldots, X_n$ $i.i.d. \sim Exp(\lambda)$, $\lambda > 0$未知。
令$g(\vec{X}) = \frac{X_n}{X_1+\cdots+X_n}$，计算$\mathbb{E}[g(\vec{X})]$。

Example (5.10)

[【1】例6.2.27] 设$X_1, \ldots, X_n$ $i.i.d. \sim N(\mu, \sigma^2)$, $\sigma^2 > 0, \mu \in \mathbb{R}$均未知。
证明：$\overline{X}$与S^2独立。

作业 习题2：Ex. 5(用Basu定理证明), 48, 49, 51, 54.