充分统计量与指数族分布

何元珍

2021-03-04

充分统计量（Sufficient Statistics）

- **统计量** $T = T(X)$ 是样本的函数，它是对样本 X 的“加工”或压缩
- 通俗理解：

$$\{X \text{ 所包含的关于参数 } \theta \text{ 的信息}\} = \{T(X) \text{ 所包含的关于参数 } \theta \text{ 的信息}\} + \{\text{已知 } T(X) \text{ 后，} X \text{ 还包含的关于参数 } \theta \text{ 的信息}\}$$

- 若后一项为零，则称 $T(X)$ 是**充分统计量**
  - 已知 $T(X)$ 后，此时 X 的条件分布不再与 θ 有关

充分统计量（Sufficient Statistics）

充分统计量 定义：给定 $X \sim \left(\mathbb{X}, \mathcal{B}X, P^X\theta\right)$，$\theta \in \Theta$，$T(X)$ 称为**充分统计量**，若条件概率 $P_\theta^X(A \mid T(X)=t) \stackrel{d}{=} P_\theta^X(A \mid t)$ 与 θ 无关，即条件分布 $F(x \mid t, \theta)$ 或条件密度 $f(x \mid t; \theta)$ 与 θ 无关

因子分解定理（Fisher–Neyman Factorization theorem）：设总体概率函数为 $f(x; \theta)$，$x_1, \ldots, x_n$ 为样本，那么 $T = T\left(x_1, \ldots, x_n\right)$ 为充分统计量的**充要条件**为：存在函数 $g(t, \theta)$ 与 $h\left(x_1, \ldots, x_n\right)$，使得对任意的 θ 和任意一组的观测值 $\left(x_1, \ldots, x_n\right)$，都有：

$$f\left(x_1, \cdots, x_n; \theta\right) = g\left(T\left(x_1, \cdots, x_n\right), \theta\right) h\left(x_1, \cdots, x_n\right)$$

充分统计量（Sufficient Statistics）

示例 1

设 $x_1, \cdots, x_n$ 为来自 $p(x;\theta)=\theta x^{\theta-1},\ 0<x<1,\ \theta>0$ 的样本，求出它的一个充分统计量。

首先，求出其联合概率密度：

$$p\left(x_1, \cdots, x_n, \theta\right)=\theta^n\left(\prod_{i=1}^n x_i\right)^{\theta-1}$$

取：

- $g(t,\theta)=\theta^n t^{\theta-1},\ h\left(x_1, x_2, \cdots, x_n\right)=1$

可得**充分统计量** $T=\prod_{i=1}^n x_i$

充分统计量（Sufficient Statistics）

示例 2

设 $x_1, \cdots, x_n$ 独立同分布，并且 $x_i \sim p\left(x_i; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma x_1} \exp\left\{-\frac{1}{2\sigma^2}\left(\ln x_i - \mu\right)^2\right\}$ 的充分统计量。

其联合密度函数：

$$
\begin{aligned}
p\left(x_1, \cdots, x_n;\ \mu, \sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma x_i} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\ln x_i - \mu\right)^2\right\} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} x_i} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \ln^2 x_i + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} \ln x_i - \frac{n\mu^2}{2\sigma^2}\right\}
\end{aligned}
$$

- 只要让 $h(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} x_i}$，其他项都放到函数 $g(T, \theta)$ 中即可

充分统计量（Sufficient Statistics）

示例 2

可见以下统计量即为**充分**的：

$$T = \left(\sum_{i=1}^{n} \ln^2 x_i, \sum_{i=1}^{n} \ln x_i \right)$$

> 一般情况下，n 个统计量解决 n 个未知参数。

指数族分布

指数族分布的定义

如果概率密度函数 $f(x;\theta)$ 可以写成

$$f(x;\theta) = c(\theta)\exp\left\{\sum_{j=1}^{k} c_j(\theta)T_j(x)\right\} h(x)$$

- 其中 $c(\theta)$，$c_j(\theta)$ 不含 x，（$c(\theta)$，$c_j(\theta)$ 是不同的项，后者不是前者的分量），$T_j(x)$，$h(x)$ 不含 θ 。分布的支撑 $\{x \mid f(x;\theta) > 0\}$ 不依赖于参数 θ（所以均匀分布不是指数族分布）

指数族分布

指数族分布的标准形式

- 可以令 $w_j = c_j(\theta),\ j = 1, \ldots, k$ 解出 $\theta = \theta\left(w_1, \ldots, w_k\right)$，从而 $c(\theta) = c(\theta(w)) = c^*(w)$
- 即 $c(\theta)$，$c_j(\theta)$ 均用 $\left(w_1, \ldots, w_k\right)$ 向量表示，代回密度函数得到指数族的标准形式：

$$
f(x;\theta) = c^*(w) \exp\left\{\sum_{j=1}^{k} w_j T_j(x)\right\} h(x)
$$

> 大部分常用分布都是指数族分布。

指数族分布

示例1：二项分布

$$P_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}(1-\theta)^n \exp\left\{x \ln \frac{\theta}{1-\theta}\right\}$$

指数族分布

示例2：正态分布

$$
\begin{aligned}
P_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} &= \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}}\right] e^{-\frac{x^2}{2\sigma^2}+\frac{\mu}{\sigma^2}x} \\
&= c(\mu,\sigma)\exp\left\{c_1(\mu,\sigma)x + c_2(\mu,\sigma)x^2\right\}
\end{aligned}
$$

其中 $T_1(x) = x$，$T_2(x) = x^2$，$h(x) = 1$。

> 指数族分布有一些很好的特性，例如必定存在共轭先验分布等。

参考资料

- 数理统计|笔记整理（3）——充分统计量
- 高等数理统计—第二章 充分统计量、完备性、样本信息
- Sufficient statistic - Wikipedia
- 03 统计学基础--指数族和充分完备统计量
- 指数族分布