Chapter 6
Principle of Data Deduction

Completeness

Outline

Introduction

Definition

Examples

Basu's Theorem

Remarks

Introduction

Completeness addresses the question:

If $E_\nu g(T(X)) = 0$, for some collection of probability distributions $\nu \in \mathcal{F}$, can we have $g(T(X)) \neq 0$ with positive probability for some $\nu \in \mathcal{F}$?

If $\mathcal{F}$ consists solely of the $Bin(2, 1/2)$ distribution and $T(X)$ is the number of successes, then

$$E_{1/2}g(T(X)) = (g(0) + 2g(1) + g(2))\frac{1}{4}$$

Thus, any function that satisfies $g(0) + 2g(1) + g(2) = 0$ has $E_{1/2}g(T(X)) = 0$.

However, If $\mathcal{F}$ consists of the $Bin(2, p)$ distributions, then

$$\begin{aligned} E_p g(T(X)) &= g(0)(1-p)^2 + 2g(1)p(1-p) + g(2)p^2 \\ &= (1-p)^2\left((g(0) + 2g(1)\left(\frac{p}{1-p}\right) + g(2)\left(\frac{p}{1-p}\right)^2\right) \end{aligned}$$

This is 0 for all $p \in (0, 1)$ if and only if $P_p\{g(T(X) = 0\} = 1$.

Definition

Definition. Let $T : \mathcal{X} \to \mathcal{T}$ be a statistic. Then the family of probability densities $\{f_{T(X)}(t|\theta), \theta \in \Theta\}$ is called complete if

$$E_\theta g(T(X)) = 0 \text{ for all } \theta \in \Theta$$

implies that

$$P_\theta\{g(T(X)) = 0\} = 1 \text{ for all } \theta \in \Theta$$

In this case, $T(X)$ is called a complete statistic.

Completeness, as a theoretical concept depends both on the choice of the statistic T and the choice of densities $\{f_{T(X)}(t|\theta), \theta \in \Theta\}$.

Examples

Example. Let $T(X) = X$, then $Pois(\lambda); \lambda > 0$ is complete.

$$E_\lambda g(X) = \sum_{x=0}^{\infty} g(x) \frac{e^{-\lambda}}{x!} \lambda^x$$

is a power series in λ. By the uniqueness of power series expansions, $E_\lambda g(X) = 0$ if and only if the coefficients of λ^x

$$g(x) \frac{e^{-\lambda}}{x!} = 0, \quad x = 0, 1, 2, \ldots.$$

Thus, g(x)=0 for all x, and $P_\lambda\{g(T(X)) = 0\} = 1$.

Examples

Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent $Unif(0, \theta), \theta > 0$ random variables and Let $T(X) = X_{(n)} = \max_{1 \leq i \leq n} X_i$. Then,

$$F_{T(X)}(t) = P_\theta\{T(X) \leq t\} = P_\theta\{X_1 \leq t, \ldots, X_n \leq t\} = P_\theta\{X_1 \leq t\}^n = \left(\frac{t}{\theta}\right)^n = \frac{t^n}{\theta^n}.$$

Thus, the density

$$f_{T(X)}(t) = \frac{nt^{n-1}}{\theta^n}.$$

$$\theta^n E_\theta g(T(X)) = \int_0^\theta g(t) n t^{n-1} \, dt, \quad \frac{d}{d\theta}(\theta^n E_\theta g(T(X))) = g(\theta) n \theta^{n-1}$$

So, if $E_\theta g(T(X)) = 0$ for all $\theta > 0$, $g(\theta) = 0$ for all $\theta > 0$, and $P_\theta\{g(T(X)) = 0\}$=1 for all $\theta > 0$.

Examples

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent random variables from an exponential family, the probability density functions can be expressed in the form

$$\mathbf{f}_X(\mathbf{x}|\eta) = \prod_{j=1}^{n} h(x_i) \cdot \exp\left(\sum_{j=1}^{n} \langle \eta, \mathbf{t}(x_j) \rangle\right) e^{-nA(\eta)}, \quad x \in S.$$

Then, $T(\mathbf{x}) = \sum_{j=1}^{n} \mathbf{t}(x_j)$ is sufficient if the parameter space contains an open subset.

The requirements for an open subset allow us to take advantage of the uniqueness of power series for the analytic function $E_\eta g(T(X))$.

Thus, the sufficient statistics from the normal, binomial, gamma, beta, Poisson, ... are also complete.

Basu's Theorem

Theorem. Any boundedly complete minimal sufficient statistic is independent of any ancillary statistic.

Let $T(X)$ be a boundedly complete minimal sufficient statistic with distribution $\mu_\theta^T(B) = P_\theta\{T(X) \in B\}$ and let $V(X)$ be ancillary $\mu_\theta^V(B) = P_\theta\{V(X) \in B\}$

To guarantee independence we will show that the condition probability of V given T does not depend on t. In other words, for every t,

$$\mu_\theta^V(B|T(X) = t) = \mu_\theta^V(B).$$

Basu's Theorem

Use, first the law of total probability, the sufficiency of $T(X)$ and then ancillarity of $V(X)$ to obtain

$$\begin{aligned}
\mu_\theta^V(B) &= \int_{\mathcal{T}} \mu_\theta^V(B|T(X)=t)\mu_\theta^T(dt) \\
&= \int_{\mathcal{T}} \mu^V(B|T(X)=t)\mu_\theta^T(dt) \\
\mu^V(B) &= \int_{\mathcal{T}} \mu^V(B|T(X)=t)\mu_\theta^T(dt) \\
0 &= \int_{\mathcal{T}} (\mu^V(B|T(X)=t) - \mu^V(B))\mu_\theta^T(dt)
\end{aligned}$$

Basu's Theorem

$$
\begin{aligned}
0 &= \int_{\mathcal{T}} (\mu^V(B|T(X)=t) - \mu^V(B))\mu_\theta^T(dt) \\
&= \int_{\mathcal{T}} g(t)\mu_\theta^T(dt) = E_\theta g(T(X))
\end{aligned}
$$

where

$$g(t) = \mu^V(B|T(X)=t) - \mu^V(B).$$

Because $T(X)$ is boundedly complete, $g(t)$ is equal to 0 for every t.

Examples

Example. For $X_1, \ldots, X_n \sim Exp(\beta)$, we have that

- $Exp(\beta)$ is a single parameter exponential family, with complete minimal sufficient statistic
$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$
- $Exp(\beta)$ is a scale family, thus
$$V(\mathbf{X}) = \frac{X_1}{T(\mathbf{X})}$$
is ancillary.

By Basu's Theorem, $T(\mathbf{X})$ and $V(\mathbf{X})$ are independent. Thus, for any β,

$$\frac{1}{\beta} E_\beta[X_1] = E_\beta[T(\mathbf{X})V(\mathbf{X})] = E_\beta[T(\mathbf{X})]E_\beta[V(\mathbf{X})] = \frac{n}{\beta} E_\beta[V(\mathbf{X})]$$

and $E_\beta[V(\mathbf{X})] = 1/n$.

Examples

Example. For $X_1, \ldots, X_n \sim Unif(0, \theta), \theta > 0$ random variables. Then, $T(X) = X_{(n)} = \max_{1 \leq i \leq n} X_i$ is complete.

Define

$$U_i = \frac{X_i}{\theta},$$

then $U_1, \ldots, U_n \sim Unif(0, 1)$. For i and j between 1 and n, the ratio

$$\frac{X_{(i)}}{X_{(j)}} = \frac{\theta U_{(i)}}{\theta U_{(j)}} = \frac{U_{(i)}}{U_{(j)}}$$

is ancillary. Moreover $U_{(i)} \sim Beta(i, n+1-i)$

Examples

By Basu's Theorem, $T(\mathbf{X}) = X_{(n)}$ and $V(\mathbf{X}) = X_{(i)}/X_{(n)}$ are independent. Thus, for any θ,

$$E_\theta[X_{(i)}] = E_\theta\left[X_{(n)}\frac{X_{(i)}}{X_{(n)}}\right] = E_\theta X_{(n)} E_\theta\left[\frac{X_{(i)}}{X_{(n)}}\right]$$

Thus,

$$E_\theta\left[\frac{X_{(i)}}{X_{(n)}}\right] = \frac{E_\theta X_{(i)}}{E_\theta X_{(n)}} = \frac{E[\theta U_{(i)}]}{E[\theta U_{(n)}]} = \frac{E[U_{(i)}]}{E[U_{(n)}]} = \frac{i/(n+1)}{n/(n+1)} = \frac{i}{n}$$

Examples

Example. For $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$, σ_0^2 known.

- $N(\mu, \sigma_0^2)$ s a single parameter exponential family, with complete minimal sufficient statistic
$$T(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$
- $N(\mu, \sigma_0^2)$ is a location family, thus
$$V(\mathbf{X}) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
is ancillary.

By Basu's Theorem, $T(\mathbf{X})$ and $V(\mathbf{X})$ are independent.

$$E_\mu T = E_\mu \left[\frac{\bar{X} - \mu}{S/\sqrt{n}}\right]$$

Examples

Student's T distribution with $n-1$ degrees of freedom is based on $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$,

$$E_\mu T = E_\mu\left[\frac{\bar{X}-\mu}{S/\sqrt{n}}\right] = E_\mu[\sqrt{n}(\bar{X}-\mu)] \cdot E\left[\frac{1}{S}\right] = 0$$

$$\mathrm{Var}_\mu(T) = E_\mu[T^2] = E_\mu\left[\left(\frac{\bar{X}-\mu}{S/\sqrt{n}}\right)^2\right] = E_\mu[n(\bar{X}-\mu)^2] \cdot E\left[\frac{1}{S^2}\right] = 1 \cdot E\left[\frac{1}{S^2}\right]$$

Now $(n-1)S^2 \sim \chi^2_{n-1}$. Its reciprocal is called an inverse χ^2_{n-1}. Its mean is $1/(n-3)$.

$$\mathrm{Var}_\mu(T) = E\left[\frac{n-1}{(n-1)S^2}\right] = \frac{n-1}{n-3}$$

Thus, a T distribution with ν degrees of freedom has mean

$$\frac{\nu}{\nu-2}.$$

Remarks

Remark. If $T(X)$ is a complete statistic, $c(T(X))$ is also complete.

- Go back to the definition of completeness.and replace g with $g \circ c$

Remark. If $T(X)$ is a complete statistic with respect to a family of distributions $\mathcal{F}$, then $T(X)$ is also complete to any family of distributions $\mathcal{F}^* \supset \mathcal{F}$.

- Adding more distrbiutions makes it harder for $P_F\{g(T(X)) = 0\} < 1$.

Remark. If $V(X)$ is a nondegenerate ancillary statistic, then it is *not* complete.

- Take $g(t) = t - EV(X)$, then $E_\theta g(V(X)) = 0$ and $P_\theta\{g(V(X)) = 0\} < 1$.
- Intuitively, if expection does not depend on the parameter θ then the size of the parameter space cannot force $P_\theta\{g(V(X)) = 0\} = 1$.

Remarks

Theorem. (Bahadur) If $T(X)$ is complete sufficient, then $T(X)$ is minimal sufficient.

- Thus, we can limit our search for complete sufficient statistics to those that are minimal sufficient.
- If a minimal sufficient statistic is not complete, then there are no complete and sufficient statistics for the family.
  If $\tilde{T}(X)$ is also minimal sufficient, then $T(X) = c(\tilde{T}(X))$ for some one to one function c. If $T(X)$ is not complete then there exist a function g and parameter value θ so that $E_\theta g(T(X)) = 0$, but $P_\theta\{g(T(X)) = 0\} < 1$. However

$$E_\theta(g \circ c)(\tilde{T}(X)) = 0 \quad P_\theta\{(g \circ c)(\tilde{T}(X)) = 0\} < 1$$

  and $\tilde{T}(X)$ is not complete.
- Completeness formalizes our ideal notion of optimal data reduction, whereas minimal sufficiency is our achievable notion of data reduction.