Lecture 33: Sufficiency and Rao-Blackwell

Ciaran Evans

Sufficient statistics

Question: Given an unbiased estimator, can I improve its variance?

- Answering this requires us to introduce a new concept: sufficient statistics

Definition (sufficient statistic): Let $X_1, \ldots, X_n$ be a sample from a distribution $f(x|\theta)$. Let $T = T(X_1, \ldots, X_n)$ be a statistic. We say that T is a sufficient statistic for θ if the conditional distribution of $X_1, \ldots, X_n \mid T$ does not depend on θ

Example

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Poisson(\lambda)$

Let $T = \sum_{i=1}^{n} X_i$ $\qquad T \sim Poisson(n\lambda)$

$$f(x_1,\ldots,x_n \mid T) = \frac{f(x_1,\ldots,x_n, T)}{f(T)} = \frac{f(x_1,\ldots,x_n)}{f(T)}$$

$$f(x_1,\ldots,x_n,T) : P(X_1=x_1, X_2=x_2, \ldots, X_n=x_n, T=\textstyle\sum_i x_i)$$
$$= P(X_1=x_1,\ldots,X_n=x_n)$$

$$\Rightarrow \frac{f(x_1,\ldots,x_n)}{f(T)} = \frac{\prod_{i=1}^{n} e^{-\lambda}\lambda^{x_i}/x_i!}{e^{-n\lambda}(n\lambda)^T/T!} = \frac{\cancel{e^{-n\lambda}}\,\cancel{\lambda^{\sum_i x_i}}/\prod_i x_i!}{\cancel{e^{-n\lambda}}\,\cancel{\lambda^{T}}\,n^T/T!}$$

$$= \frac{T!}{n^T \prod_i x_i!}$$

← does not depend on λ

$\Rightarrow T = \sum_i X_i$ is a sufficient statistic

Example

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$

$T = \sum_i X_i \quad \sim \text{Binomial}(n, p)$

$$f(x_1, \ldots, x_n \mid T) = \frac{f(x_1, \ldots, x_n)}{f(T)}$$

$$= \frac{\prod_i p^{x_i}(1-p)^{1-x_i}}{\binom{n}{T} p^T (1-p)^{n-T}}$$

$$= \frac{p^{\sum_i x_i}(1-p)^{n-\sum x_i}}{\binom{n}{T} p^T (1-p)^{n-T}} = \frac{1}{\binom{n}{T}}$$

does not depend on p

$\Rightarrow$ T is sufficient for p

Rao-Blackwell

Let θ be a parameter of interest, and let $\tau(\theta)$ be some function of θ.

Let $\hat{\tau}$ be an unbiased estimator of $\tau(\theta)$, and T be a sufficient statistic for θ.

Let $\tau^* = \mathbb{E}[\hat{\tau} \mid T]$. Then:

1. $\mathbb{E}[\tau^*] = \tau$ (unbiased)
2. $Var(\tau^*) \leq Var(\hat{\tau})$

Example

Suppose $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$

sufficient statistic: $T = \sum_i X_i$

consider $\hat{\lambda} = X_1$ $\quad \mathbb{E}[X_1] = \lambda \quad Var(X_1) = \lambda$

$\lambda^* = \mathbb{E}[\hat{\lambda} \mid T]$

want: $\mathbb{E}[X_1 \mid T = t]$ i.e. $\mathbb{E}[X_1 \mid \sum_i X_i = t]$

$\mathbb{E}[T \mid T = t] = t$

$\Rightarrow \mathbb{E}[\sum_i X_i \mid T = t] = t$

$\Rightarrow \sum_i \mathbb{E}[X_i \mid T = t] = t$

$\Rightarrow n\,\mathbb{E}[X_1 \mid T = t] = t \Rightarrow \mathbb{E}[X_1 \mid T = t] = \frac{t}{n}$

$\Rightarrow \lambda^* = \mathbb{E}[\hat{\lambda} \mid T] = \frac{T}{n} = \frac{1}{n}\sum_i X_i$

(Rao-Blackwellized estimator) $\quad Var(\lambda^*) = \frac{\lambda}{n} < Var(X_1)$

Factorization theorem

Let $X_1, \dots, X_n$ be a sample with joint probability function $f(x_1, \dots, x_n \mid \theta)$.

A statistic $T = T(X_1, \dots, X_n)$ is sufficient for θ if and only if there exist functions $g(t \mid \theta)$ and $h(x_1, \dots, x_n)$ such that for any possible $x_1, \dots, x_n$ and any possible θ,

$$f(x_1, \dots, x_n \mid \theta) = \underbrace{g(T(x_1, \dots, x_n) \mid \theta)}_{\text{joint distribution only depends on } \theta \text{ through the sufficient statistic}} h(x_1, \dots, x_n)$$

joint distribution only depends on θ through the sufficient statistic

Ex: $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$

$$f(x_1, \dots, x_n \mid p) = p^{\sum_i x_i} (1-p)^{n - \sum x_i}$$

$T = \sum_i X_i \qquad g(T \mid p) = p^T (1-p)^{n-T} \qquad h(x) = 1$

$$\Rightarrow f(x_1, \dots, x_n \mid p) = g\left(\sum_i x_i \mid p\right) h(x_1, \dots, x_n)$$

$X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ $\qquad$ μ, σ^2 are unknown

$$f(x_1, \dots, x_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\}$$

$$\sum_i (X_i - \mu)^2 = \sum_i \left(X_i^2 - 2\mu X_i + \mu^2 \right)$$

$$= \sum_i X_i^2 - 2\mu \sum_i X_i + n\mu^2$$

$$\Rightarrow \quad f(x_1, \dots, x_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_i X_i^2 - 2\mu \sum_i X_i + n\mu^2 \right) \right\}$$

$$= g(T_1, T_2 \mid \mu, \sigma^2)$$

$$T_1 = \sum_i X_i$$

$$T_2 = \sum_i X_i^2$$

$T = (T_1, T_2) \in \mathbb{R}^2$ (sufficient statistic)

equivalently: $\left(\bar{X}, \ \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 \right)$