Lecture 10: Sufficiency and the Rao-Blackwell Theorem

MATH 667-01
Statistical Inference
University of Louisville

October 12, 2017

Introduction

- We discuss sufficiency as discussed in Sections 6.1 and 6.2 of Casella and Berger (2002)[1].
- We discuss and prove the Rao-Blackwell Theorem as discussed in Section 7.3.
- The proof of the Rao-Blackwell Theorem uses iterated expectation formulas from Section 4.4.

[1] Casella, G. and Berger, R. (2002). *Statistical Inference, Second Edition*. Duxbury Press, Belmont, CA.

- Now we examine data summarization and data reduction when making inferences about a fixed but unknown parameter θ based on a sample $X_1, \ldots, X_n$.
- When the sample size n is large, simply being given a list of the observed sample values $x_1, \ldots, x_n$ is not very useful.
- Instead, it is useful to provide a statistic $T(X_1, \ldots, X_n)$ and use the observed value $T(x_1, \ldots, x_n)$ to summarize the information about θ in the observed sample.
- Let $\mathcal{X}$ denote the sample space of $X_1, \ldots, X_n$. Then $\mathcal{T} = \{t : t = T(\boldsymbol{x}) \text{ for some } \boldsymbol{x} \in \mathcal{X}\}$ is the image of $\mathcal{X}$ under T.
- So $T(\boldsymbol{x})$ partitions $\mathcal{X}$ into sets $A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$ for $t \in \mathcal{T}$.

- The goal of the *sufficiency principle* is to summarize data while not losing information about θ.
- *Definition L10.1* (Def 6.2.1 on p.272): A statistic $T(\boldsymbol{X})$ is a *sufficient statistic for* θ if the conditional distribution of the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ given the value of $T(\boldsymbol{X})$ does not depend on θ.
- That is, $T(\boldsymbol{X})$ is sufficient for θ if the pdf/pmf $f_{\boldsymbol{X}|T(\boldsymbol{X})=T(\boldsymbol{x})}(\boldsymbol{x}|\theta)$ is the same for all θ.

- *Theorem L10.1* (Thm 6.2.2 on p.274): If $p(\boldsymbol{x}|\theta)$ is the joint pdf/pmf of $\boldsymbol{X}$, and $q(t|\theta)$ is the pdf/pmf of $T(\boldsymbol{X})$, then $T(\boldsymbol{X})$ is a sufficient statistic for θ if, and only if, for every $\boldsymbol{x}$ in the sample space the ratio $p(\boldsymbol{x}|\theta)/q(T(\boldsymbol{x})|\theta)$ is constant as a function of θ.
- *Proof of Theorem L10.1*:

$$
\begin{aligned}
P_\theta(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x})) &= \frac{P_\theta\left(\boldsymbol{X} = \boldsymbol{x} \text{ and } T(\boldsymbol{X}) = T(\boldsymbol{x})\right)}{P_\theta\left(T(\boldsymbol{X}) = T(\boldsymbol{x})\right)} \\
&= \frac{P_\theta\left(\boldsymbol{X} = \boldsymbol{x}\right)}{P_\theta\left(T(\boldsymbol{X}) = T(\boldsymbol{x})\right)} \\
&= \frac{p(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)}.
\end{aligned}
$$

So, $T(X)$ is sufficient if and only if the probability above is constant as a function of θ.

- *Example L10.1*: Let $X_1, \ldots, X_n$ be iid Poisson(λ) random variables. Show that $\sum_{i=1}^{n} X_i$ is sufficient for λ.
- *Answer to Example L10.1*:

$$P\left((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \middle| \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right) =$$

$$\frac{P\left((X_1, \ldots, X_n) = (x_1, \ldots, x_n)\right)}{P\left(\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} x_i\right)} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda} / \left(\prod_{i=1}^{n} x_i!\right)}{(n\lambda)^{\sum_{i=1}^{n} x_i} e^{-n\lambda} / \left(\sum_{i=1}^{n} x_i\right)!}$$

since $\sum_{i=1}^{n} X_i \sim$ Poisson$(n\lambda)$. Simplifying this expression, we obtain $n^{-\sum_{i=1}^{n} x_i} \left(\sum_{i=1}^{n} x_i\right)! / \left(\prod_{i=1}^{n} x_i!\right)$ which does not depend on λ.

- We can use *Theorem L10.1* to verify that a statistic is sufficient for θ, but it is better to have a way of finding sufficient statistics without having a candidate in mind.
- This can be done with the following result known as the Factorization Theorem.
- *Theorem L10.2* (Thm 6.2.6 on p.276): Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf/pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\boldsymbol{x})$ such that, for all sample points $\boldsymbol{x}$ and all parameter points θ, $f(\boldsymbol{x}|\theta) = g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$.

- *Sketch of proof of Theorem L10.2 for the discrete case*:
- Suppose $T(\boldsymbol{X})$ is a sufficient statistic. Then

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= P_\theta(\boldsymbol{X}=\boldsymbol{x}) \\ &= P_\theta\left(\boldsymbol{X}=\boldsymbol{x} \text{ and } T(\boldsymbol{X})=T(\boldsymbol{x})\right) \\ &= P_\theta\left(T(\boldsymbol{X})=T(\boldsymbol{x})\right) P_\theta\left(\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=T(\boldsymbol{x})\right) \\ &= g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x}). \end{aligned}$$

- Suppose that $f(\boldsymbol{x}|\theta)=g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})$. Then

$$\begin{aligned} \frac{f(\boldsymbol{x}|\theta)}{q(T(\boldsymbol{x})|\theta)} &= \frac{g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{q(T(\boldsymbol{x})|\theta)} \\ &= \frac{g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A_{T(\boldsymbol{x})}} g(T(\boldsymbol{y})|\theta)h(\boldsymbol{y})} \\ &= \frac{g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})}{g(T(\boldsymbol{x})|\theta)\sum_{\boldsymbol{y}\in A_{T(\boldsymbol{x})}} h(\boldsymbol{y})} = \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A_{T(\boldsymbol{x})}} h(\boldsymbol{y})} \end{aligned}$$

does not depend on θ.

- *Example L10.2*: Let $X_1, \dots, X_n$ be iid random variables from a Normal$(\mu, 1)$ distribution. Find a sufficient estimator for μ.
- *Answer to Example L10.2*: Let $\boldsymbol{x} = (x_1, \dots, x_n)$. The joint pdf of $X_1, \dots, X_n$ is

$$\begin{aligned} f(\boldsymbol{x}|\mu) &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \exp\left(n\bar{x}\mu - \frac{n}{2}\mu^2\right) \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) e^{-\frac{n}{2}(\bar{x}-\mu)^2} \\ &= h(\boldsymbol{x}) g(\bar{x}|\mu) \end{aligned}$$

where $h(\boldsymbol{x}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$ does not depend on μ and $g(t|\mu) = e^{-\frac{n}{2}(t-\mu)^2}$. Thus, $\bar{X}$ is sufficient for μ.

- *Example L10.3*: Let $X_1, \ldots, X_n$ be iid random variables from a Uniform$\{1, \ldots, \theta\}$ distribution. Show that $X_{(n)}$ is sufficient for θ.
- *Answer to Example L10.3*: Let $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\mathcal{N}_\theta = \{1, 2, \ldots, \theta\}$, and $\mathcal{N}$ is the set of positive integers. The joint pmf of $X_1, \ldots, X_n$ is

$$
\begin{aligned}
f(\boldsymbol{x}|\theta) &= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{\mathcal{N}_\theta}(x_i) \\
&= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{\mathcal{N}}(x_i) I_{\mathcal{N}_\theta}\left(x_{(n)}\right) \\
&= \frac{1}{\theta^n} I_{\mathcal{N}_\theta}\left(x_{(n)}\right) \prod_{i=1}^{n} I_{\mathcal{N}}(x_i) \\
&= g(x_{(n)}|\theta) h(\boldsymbol{x})
\end{aligned}
$$

where $g(t|\theta) = \frac{1}{\theta^n} I_{\mathcal{N}_\theta}(t)$ and $h(\boldsymbol{x}) = \prod_{i=1}^{n} I_{\mathcal{N}}(x_i)$ does not depend on θ. Thus, $X_{(n)}$ is sufficient for θ.

- Sometimes, the information about the parameter cannot be summarized with a single number. The sufficient statistic might be a vector and the parameter itself might be vector-valued.
- *Theorem L10.3* (Thm 6.2.10 on p.279): Let $X_1, \ldots, X_n$ be iid observations from a pdf or pmf, $f(x|\boldsymbol{\theta})$ that belongs to an exponential family given by

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, $d \leq k$. Then $T(\boldsymbol{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$ is a sufficient statistic for $\boldsymbol{\theta}$.

- *Example L10.4*: Suppose that $X_1, \ldots, X_n$ is a random sample from a Normal distribution with unknown mean μ and unknown variance σ^2. Find a sufficient statistic for (μ, σ^2).
- *Answer to Example L10.4*: Recall from Example L6.5 that the normal family of densities with mean μ and variance σ^2 can be expressed as

$$f(x|\boldsymbol{\eta}) = h(x)c(\boldsymbol{\eta})e^{\eta_1 t_1(x) + \eta_2 t_2(x)}$$

where $h(x) = \frac{1}{\sqrt{2\pi}}$, $c^*(\boldsymbol{\eta}) = \sqrt{\eta_1}\exp\left(-\frac{\eta_2^2}{2\eta_1}\right)$, $t_1(x) = -\frac{x^2}{2}$, and $t_2(x) = x$ with $\eta_1 = 1/\sigma^2$ and $\eta_2 = \mu/\sigma^2$.

Thus, $\left(-\frac{1}{2}\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i\right)$ is sufficient for (μ, σ^2).

- Any one-to-one function of a sufficient statistic is also a sufficient statistic, as shown below.
- Suppose $T(\boldsymbol{X})$ is a sufficient statistic for $\boldsymbol{\theta}$, and suppose r is a one-to-one function (with inverse r^{-1}) such that $T^*(\boldsymbol{x}) = r(T(\boldsymbol{x}))$ for all $\boldsymbol{x}$.
- By the Factorization Theorem(*Theorem L10.2*), there exist g and h such that

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = g(T(\boldsymbol{x})|\boldsymbol{\theta})h(\boldsymbol{x}) = g(r^{-1}(T^*(\boldsymbol{x}))|\boldsymbol{\theta}))h(\boldsymbol{x}).$$

Letting $g^*(t|\boldsymbol{\theta}) = g(r^{-1}(t)|\boldsymbol{\theta})$, we have

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = g^*(T^*(\boldsymbol{x})|\boldsymbol{\theta})h(\boldsymbol{x})$$

so that $T^*(\boldsymbol{X})$ is sufficient for $\boldsymbol{\theta}$.

- *Example L10.5*: Suppose that $X_1, \ldots, X_n$ is a random sample from a Normal distribution with unknown mean μ and unknown variance σ^2. Show that $(\bar{X}, S^2)$ is sufficient for (μ, σ^2).
- *Answer to Example L10.5*: It was shown in *Example L10.4* that $\left(-\frac{1}{2}\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i\right)$ is sufficient for (μ, σ^2). Let

$$r(t_1, t_2) = \left(\frac{t_2}{n}, \frac{-2nt_1 - t_2^2}{n(n-1)}\right).$$

Since r is one-to-one, $r\left(-\frac{1}{2}\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i\right) = (\bar{X}, S^2)$ is sufficient for (μ, σ^2).

- *Definition L10.2* (Def 6.2.11 on p.280): A sufficient statistic $T(\boldsymbol{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\boldsymbol{X})$, $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$.
- *Theorem L10.4* (Thm 6.2.13 on p.281): Let $f(\boldsymbol{x}|\theta)$ be a pmf/pdf of a sample $\boldsymbol{X}$. Suppose there exists a function $T(\boldsymbol{x})$ such that, for every two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ is constant as a function of θ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

- *Example L10.6*: Let $X_1, \dots, X_n$ be iid Normal(μ, σ^2), with μ and σ^2 unknown. Show that $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).
- *Answer to Example L10.6*: From *Example L10.5*, this statistic is sufficient. Let $(\bar{x}, s_x^2)$ and $(\bar{y}, s_y^2)$ denote the sample means and sample variances corresponding to the observed samples $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. It can be shown that

$$
\begin{aligned}
\frac{f(\boldsymbol{x}|\mu,\sigma^2)}{f(\boldsymbol{y}|\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-\left[n(\bar{x}-\mu)^2+(n-1)s_x^2\right]/(2\sigma^2)\right\}}{(2\pi\sigma^2)^{-n/2}\exp\left\{-\left[n(\bar{y}-\mu)^2+(n-1)s_y^2\right]/(2\sigma^2)\right\}} \\
&= \exp\left\{-\left[n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)(s_x^2-s_y^2)\right]/(2\sigma^2)\right\},
\end{aligned}
$$

which is constant if and only if $\bar{x} = \bar{y}$ and $s_x^2 = s_y^2$.

Thus, $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).

Rao-Blackwell Theorem

- Sufficient statistics are related to unbiased estimators through a well-known result known as the Rao-Blackwell Theorem.
- *Theorem L10.5* (Thm 7.3.17 on p.342): Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = \mathsf{E}(W|T)$. Then
  (1) $\mathsf{E}_\theta \phi(T) = \tau(\theta)$ and
  (2) $\mathsf{Var}_\theta\ \phi(T) \leq \mathsf{Var}_\theta\ W$ for all θ;

  that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.
- Consequently, conditioning any unbiased estimator on a sufficient statistic will uniformly "improve" the estimator, so the Rao-Blackwell Theorem shows that we only need to consider statistics which are functions of sufficient statistics when searching for a UMVUE.

Review: Iterated expectation formulas

- *Theorem L10.6* (Thm 4.4.3 on p.164): If X and Y are any two random variables, then

  $$\mathsf{E}[X] = \mathsf{E}[\mathsf{E}[X|Y]],$$

  provided that the expectations exist.
- *Theorem L10.7* (Thm 4.4.7 on p.167): For any two random variables X and Y,

  $$\mathsf{Var}[X] = \mathsf{E}[\mathsf{Var}[X|Y]] + \mathsf{Var}[\mathsf{E}[X|Y]]$$

  provided that the expectations exist.

- *Proof of Theorem L10.5*: Since T is sufficient, $W|T$ does not depend on θ and thus $\phi(T) = \mathsf{E}[W|T]$ is only a function of the sample and thus an estimator. Using the iterated formulas, we have

$$\mathsf{E}[\phi(T)] = \mathsf{E}[\mathsf{E}[W|T]] = \mathsf{E}[W] = \tau(\theta)$$

and

$$\begin{aligned} \mathsf{Var}[W] &= \mathsf{E}[\mathsf{Var}[W|T]] + \mathsf{Var}[\mathsf{E}[W|T]] \\ &= \mathsf{E}[\mathsf{Var}[W|T]] + \mathsf{Var}[\phi(T)] \\ &\geq \mathsf{Var}[\phi(T)] \end{aligned}$$

since $\mathsf{Var}[W|T] \geq 0$, and thus, $\mathsf{E}[\mathsf{Var}[W|T]] \geq 0$.

- *Example L10.7*: Let X_1 and X_2 be independent identically distributed (iid) Poisson(θ) random variables.
  (a) Find a sufficient statistic for θ.
  (b) Show that $T(X_1) = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{otherwise} \end{cases}$
  is an unbiased estimator of $\tau(\theta) = e^{-\theta}$.
  (c) Compute $\mathsf{E}[T(X_1) | X_1 + X_2 = y]$.
  (d) For the estimator $T(X_1)$ in part (b), find a uniformly better unbiased estimator of $e^{-\theta}$.

- *Answer to Example L10.7*: (a) The joint pmf of X_1 and X_2 is

$$\begin{aligned} f(x_1, x_2|\theta) &= f(x_1|\theta)f(x_2|\theta) = \frac{\theta^{x_1}e^{-\theta}}{x_1!}\frac{\theta^{x_2}e^{-\theta}}{x_2!} \\ &= \frac{\theta^{x_1+x_2}e^{-2\theta}}{x_1!x_2!} = g(x_1 + x_2|\theta)h(x_1, x_2) \end{aligned}$$

  where $g(t|\theta) = \theta^t e^{-2\theta}$ and $h(\boldsymbol{x}) = \frac{1}{x_1!x_2!}$. So, $X_1 + X_2$ is sufficient for θ.
- (b) $\mathsf{E}[T(X_1)] = P(T(X_1) = 1) = P(X_1 = 0) = \frac{\theta^0 e^{\theta}}{0!} = e^{-\theta}$
- (c) Since $X_1 + X_2 \sim \text{Poisson}(2\theta)$, we have

$$\begin{aligned} \mathsf{E}[T(X_1)|X_1 + X_2 = y] &= P(T(X_1) = 1|X_1 + X_2 = y) \\ &= P(X_1 = 0|X_1 + X_2 = y) \\ &= \frac{P(X_1 = 0 \text{ and } X_1 + X_2 = y)}{P(X_1 + X_2 = y)} \\ &= \frac{P(X_1 = 0 \text{ and } X_2 = y)}{P(X_1 + X_2 = y)} \end{aligned}$$

- *Answer to Example L10.7 continued*:

$$
\begin{aligned}
\mathsf{E}[T(X_1)|X_1+X_2=y] &= \frac{P(X_1=0 \text{ and } X_2=y)}{P(X_1+X_2=y)} \\
&= \frac{P(X_1=0)P(X_2=y)}{P(X_1+X_2=y)} \\
&= \frac{e^{-\theta}(\theta^y e^{-\theta}/y!)}{(2\theta)^y e^{-2\theta}/y!} \\
&= \frac{\theta^y}{(2\theta)^y} = \left(\frac{1}{2}\right)^y .
\end{aligned}
$$

- (d) Since $T(X_1)$ is an unbiased estimator of $e^{-\theta}$ and X_1+X_2 is sufficient for θ (and consequently $e^{-\theta}$), the Rao-Blackwell Theorem implies that

$$\phi(X_1+X_2) = \mathsf{E}[T(X_1)|X_1+X_2] = \left(\frac{1}{2}\right)^{X_1+X_2}$$

is a uniformly better unbiased estimator of $e^{-\theta}$.