

Lecture 10: Sufficiency and the Rao-Blackwell Theorem

MATH 667-01
Statistical Inference
University of Louisville

October 12, 2017

Last modified: 10/18/2017

- 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.

¹Casella, 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, \dots, X_n .
- When the sample size n is large, simply being given a list of the observed sample values x_1, \dots, x_n is not very useful.
- Instead, it is useful to provide a statistic $T(X_1, \dots, X_n)$ and use the observed value $T(x_1, \dots, x_n)$ to summarize the information about θ in the observed sample.
- Let \mathcal{X} denote the sample space of X_1, \dots, X_n . Then $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ is the image of \mathcal{X} under T .
- So $T(\mathbf{x})$ partitions \mathcal{X} into sets $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$ 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 $T(\mathbf{X})$ is a *sufficient statistic for θ* if the conditional distribution of the sample $\mathbf{X} = (X_1, \dots, X_n)$ given the value of $T(\mathbf{X})$ does not depend on θ .
- That is, $T(\mathbf{X})$ is sufficient for θ if the pdf/pmf $f_{\mathbf{X}|T(\mathbf{X})=T(\mathbf{x})}(\mathbf{x}|\theta)$ is the same for all θ .

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

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

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

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

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

since $\sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$. Simplifying this expression, we obtain $n^{-\sum_{i=1}^n x_i} (\sum_{i=1}^n x_i)! / (\prod_{i=1}^n x_i!)$ 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 $f(\mathbf{x}|\theta)$ denote the joint pdf/pmf of a sample \mathbf{X} . A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points \mathbf{x} and all parameter points θ , $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$.

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

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

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

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

does not depend on θ .

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

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

where $h(\mathbf{x}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2\right)$ 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, \dots, X_n be iid random variables from a $\text{Uniform}\{1, \dots, \theta\}$ distribution. Show that $X_{(n)}$ is sufficient for θ .
- *Answer to Example L10.3:* Let $\mathbf{x} = (x_1, \dots, x_n)$, $\mathcal{N}_\theta = \{1, 2, \dots, \theta\}$, and \mathcal{N} is the set of positive integers. The joint pmf of X_1, \dots, X_n is

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

where $g(t|\theta) = \frac{1}{\theta^n} I_{\mathcal{N}_\theta}(t)$ and $h(\mathbf{x}) = \prod_{i=1}^n I_{\mathcal{N}}(x_i)$ 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, \dots, 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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)$, $d \leq k$. Then

$T(\mathbf{X}) = \left(\sum_{j=1}^n t_1(X_j), \dots, \sum_{j=1}^n t_k(X_j) \right)$ is a sufficient statistic for $\boldsymbol{\theta}$.

- *Example L10.4:* Suppose that X_1, \dots, 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 $T(\mathbf{X})$ is a sufficient statistic for θ , and suppose r is a one-to-one function (with inverse r^{-1}) such that $T^*(\mathbf{x}) = r(T(\mathbf{x}))$ for all \mathbf{x} .
- By the Factorization Theorem (*Theorem L10.2*), there exist g and h such that

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

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

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

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

- *Example L10.5:* Suppose that X_1, \dots, 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 $T(\mathbf{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$.
- *Theorem L10.4* (Thm 6.2.13 on p.281): Let $f(\mathbf{x}|\theta)$ be a pmf/pdf of a sample \mathbf{X} . Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points \mathbf{x} and \mathbf{y} , the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ .

- *Example L10.6:* Let X_1, \dots, X_n be iid $\text{Normal}(\mu, \sigma^2)$, 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 \mathbf{x} and \mathbf{y} , respectively. It can be shown that

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

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) .

- 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 $\phi(T) = E[W|T]$. Then
 - (1) $E_{\theta}[\phi(T)] = \tau(\theta)$ and
 - (2) $\text{Var}_{\theta}[\phi(T)] \leq \text{Var}_{\theta}[W]$ 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.

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

$$E[X] = E[E[X|Y]],$$

provided that the expectations exist.

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

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

provided that the expectations exist.

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

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

and

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

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

- *Example L10.7:* Let X_1 and X_2 be independent identically distributed (iid) $\text{Poisson}(\theta)$ random variables.
 - (a) Find a sufficient statistic for θ .
 - (b) Show that $W = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{otherwise} \end{cases}$ is an unbiased estimator of $\tau(\theta) = e^{-\theta}$.
 - (c) Compute $E[W | X_1 + X_2 = y]$.
 - (d) For the estimator W 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 $h(\mathbf{x}) = \frac{1}{x_1! x_2!}$. So, $X_1 + X_2$ is sufficient for θ .

- (b) $E[\mathbf{W}] = P(\mathbf{W} = 1) = P(X_1 = 0) = \frac{\theta^0 e^{-\theta}}{0!} = e^{-\theta}$
- (c) Since $X_1 + X_2 \sim \text{Poisson}(2\theta)$, we have

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

- *Answer to Example L10.7 continued:*

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

- (d) Since $\textcolor{red}{W}$ 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

$$\phi(X_1 + X_2) = \mathbb{E}[\textcolor{red}{W} | X_1 + X_2] = \left(\frac{1}{2}\right)^{X_1 + X_2}$$

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