LECTURE 4: SUFFICIENT STATISTICS

Michal Kolesár*

September 9, 2024

We would like to make a decision based on data X. Do we need to keep around all of X, or can we reduce its dimension without loss of information and throw away parts of the data that are not helpful in making our decision? The concept of sufficient statistics allows us to do exactly that.

1. SUFFICIENT STATISTICS

REFERENCE CB, Chapter 6.1–6.2

Let $\mathcal{F} = \{F(\cdot \mid \theta) \colon \theta \in \Theta\}$ be some parametric family. Let X be a random vector with cumulative distribution function (CDF) $F(\cdot \mid \theta)$. X may or may not be a random sample, and it may or may not consist separate observations, $X = (X_1, \ldots, X_n)$. Recall that a *statistic* is any function of the data. Thus, if g is some (measurable) function, then $Y = g(X)$ is a statistic. To stress that its distribution depends on the distribution F of the data, it is often referred to as the *sampling distribution*.

Definition 1. A statistic $S = \phi(X)$ is sufficient for the family $\mathcal{F}$ if the conditional distribution of X given S does not depend on θ.

Remark 2. A sufficient statistic is best understood as a partition of the sample space. When the dimension of S is smaller than the dimension of X, then many x will lead to the same $s = \phi(x)$. A sufficient statistic, just like any statistic, can be thought of as inducing a partition (that is, a collection of mutually disjoint sets) of the possible realizations of X, such that realizations of x that lead to the same s are lumped together in one set. What makes a sufficient statistic special is that the partition contains all information about θ: it is a "sufficient" description of X for all questions one may have about θ.

Example 1. Let X_1, X_2, X_3 be i.i.d. Bernoulli with parameter θ. Let $S = \sum_{i=1}^{n} X_i$. There

*Email: mkolesar@princeton.edu.

are 8 possible realizations of $X = (X_1, X_2, X_3)$. S partitions them into 4 partitions:

$$\begin{aligned}
\{X \mid S = 0\} &= \{(0,0,0)\} \\
\{X \mid S = 1\} &= \{(0,0,1), (0,1,0), (1,0,0)\} \\
\{X \mid S = 2\} &= \{(0,1,1), (1,0,1), (1,1,0)\} \\
\{X \mid S = 3\} &= \{(1,1,1)\}.
\end{aligned}$$

Let's check if the conditional distribution of X given S is independent of θ. When $S = 0$, then $P(X = (0,0,0) \mid S = 0) = 1$, and the other 7 possibilities have zero probability. When $S = 1$, then $P(X = (1,0,0) \mid S = 1) = P(X = (0,1,0) \mid S = 1) = P(X = (0,0,1) \mid S = 1) = 1/3$ and the other 5 possibilities have zero probability. Similarly, $P(X = (1,1,0) \mid S = 2) = P(X = (0,1,1) \mid S = 2) = P(X = (1,0,1) \mid S = 2) = 1/3$, with the other 5 possibilities having zero probability. Finally, $P(X = (1,1,1) \mid S = 3) = 1$, and the other 7 possibilities have zero probability. None of these probabilities depend on θ, so S is sufficient. ⊠

Here is another way of thinking about it. Consider the pair (S, X). Obviously, (S, X) contains the same information about θ as X alone, since S is a function of X. But if we know S, then X itself has no value for us since its conditional distribution given X is independent of θ. Thus, by observing X (in addition to S), we cannot say whether one particular value of parameter θ is more likely than another. Therefore, once we know S, we can discard X completely.

Example 2. Let $X = (X_1, \ldots, X_n)$ be a random sample from $\mathcal{N}(\theta, \sigma^2)$, with σ^2 known. We have already seen that $S = \overline{X}_n \sim N(\theta, \sigma^2/n)$. We claim that S is sufficient.

Note that the joint distribution of $(\overline{X}_n, S)$ is multivariate normal, but with a degenerate covariance matrix, so that the joint density is not defined on $\mathbb{R}^{n+1}$, and we can't check sufficiency by checking that $f_{S,X}(s,x)/f_S(s)$ doesn't depend on θ. Instead, we use the result from Lecture note 1 that

$$X \mid S = s \sim \mathcal{N}(E[X] + \operatorname{cov}(X,S)\operatorname{var}(S)^{-1}(s - E[S])\ ,\ V),$$

where $V = \operatorname{var}(X) - \operatorname{cov}(X,S)\operatorname{var}(S)^{-1}\operatorname{cov}(S,X)$. Now,

$$\operatorname{cov}(X_i, S) = n^{-1}E[(X_i - \theta)\sum_j (X_j - \theta)] = \sigma^2/n.$$

Letting ι_n denote an n-vector of ones, we therefore have

$$X \mid S = s \sim \mathcal{N}(\iota_n s\ ,\ \sigma^2(I_n - \iota_n\iota_n'/n)),$$

which doesn't depend on θ. ⊠

2. FACTORIZATION THEOREM

The Factorization Theorem gives a general approach to finding a sufficient statistic:

Theorem 3 (Factorization Theorem, Halmos and Savage 1949). Let $f(x \mid \theta)$ be the probability density function (PDF) or the probability mass function of a random vector X. Then $S = \phi(X)$ is a sufficient statistic if and only if there exist functions $g(s, \theta)$ and $h(x)$ such that $f(x \mid \theta) = g(\phi(x), \theta)h(x)$.

Proof. We prove this for the discrete case. For rigorous proof in the continuous case, see Lehmann and Romano (2005, Corollary 2.6.1).

Consider first necessity. If S is sufficient, then the conditional probability $P_\theta(X = x \mid S = s) = P_\theta(X = x \mid S = \phi(x))$ must be independent of θ, denote it by $h(x)$. But then

$$f(x \mid \theta) = P_\theta(X = x, S = \phi(x)) = h(x)P_\theta(S = \phi(x)),$$

which gives the required factorization.

To prove sufficiency, let $A(s) = \{x \colon \phi(x) = s\}$ denote the partition induced by $S = s$. Then the marginal pmf of s is given by $f_S(s \mid \theta) = P_\theta(S = s) = \sum_{y \colon y \in A(s)} f(y \mid \theta) = g(s, \theta) \sum_{y \colon y \in A(s)} h(y)$, with the conditional probability therefore given by

$$P_\theta(X = x \mid S = s) = \frac{h(x)}{\sum_{y \colon y \in A(s)} h(y)},$$

which doesn't depend on θ. □

Example 3 (Casella and Berger 2002, Example 6.2.9). Let $X_1, \ldots, X_n$ be an i.i.d. sample from $\mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown, so that $\theta = (\mu, \sigma^2)$. Then

$$\begin{aligned} f(x \mid \theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + \mu^2\right)\right). \end{aligned}$$

Thus, the vector $(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is a sufficient statistic (here $h(x) = 1$ and g is the whole thing). In addition, as we have seen before,

$$\begin{aligned} f(x \mid \theta) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2\right)\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left((n-1)s_n^2 + n(\bar{x}_n - \mu)^2\right)\right). \end{aligned}$$

Thus, $(\bar{X}_n, S_n^2)$ is another sufficient statistic. ⊠

For future use, we record the following basic facts about the sufficient statistic $(\bar{X}_n, S_n^2)$:

Theorem 4 (Casella and Berger 2002, Theorem 5.3.1). If $X_1, \ldots, X_n$ are i.i.d. random variables with $N(\mu, \sigma^2)$ distribution, then

(i) $\overline{X}_n$ *and* S_n^2 *are independent;*

(ii) $\overline{X}_n \sim N(\mu, \sigma^2/n)$*; and*

(iii) $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$.

We've shown before that regardless of the distribution of X_i, $(\overline{X}_n, S_n)$ is unbiased for (μ, σ^2). Saying more about their distribution is in general complicated. The significance of Theorem 4 is that if the distribution is normal, then we actually know the whole distribution.

Proof. Let $Y_i = X_i - \overline{X}_n$. Then the random variables $\overline{X}_n, Y_1, \ldots, Y_n$ are jointly normal (why?). We already know that $E[\overline{X}_n] = \mu$ and that $\text{var}(\overline{X}_n) = \sigma^2/n$, which implies part (ii). To show part (i), observe that for any $j = 1, \ldots, n$,

$$\begin{aligned}\text{cov}(\overline{X}_n, Y_j) &= \text{cov}(\overline{X}_n, X_j - \overline{X}_n) \\ &= \text{cov}(\overline{X}_n, X_j) - \text{var}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \text{cov}(X_i, X_j) - \sigma^2/n \\ &= \sigma^2/n - \sigma^2/n = 0.\end{aligned}$$

Since uncorrelated jointly normal random variables are independent, we conclude that $\overline{X}_n$ is independent of $Y_1, Y_2, \ldots, Y_n$. Moreover, since $S_n^2 = \sum_{i=1}^n Y_i^2/(n-1)$, part (i) holds since functions of independent random variables are independent as well.

The proof of part (iii) is left as a homework exercise. □

Digression. The statistic $T = (\overline{X}_n - \mu)/(S_n/\sqrt{n})$ is called the *t-statistic*. Using Theorem 4,

$$t = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \frac{1}{\sqrt{S_n^2/\sigma^2}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} \sim t(n-1)$$

since the $\mathcal{N}(0,1)$ and $\chi^2(n-1)$ random variables are independent. Thus, we proved that if $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, then the t-statistic has t-distribution with $n-1$ degrees of freedom. ⊠

Example 4. Let $X_1, \ldots, X_n$ be a random sample from $U[\theta, 1+\theta]$. Then $f(x \mid \theta) = 1$ if $\theta \le \min_i X_i \le \max_i X_i \le 1 + \theta$ and 0 otherwise. In other words,

$$f(x \mid \theta) = I(\theta \le X_{(1)}) I(1 + \theta \ge X_{(n)}).$$

So $(X_{(1)}, X_{(n)})$ is sufficient. ⊠

Some final observations:

- A sufficient statistic is not unique: one can always add another statistic to a given sufficient statistic to obtain a higher dimensional sufficient statistic, and $\phi(X) = X$ is also always a (trivial) sufficient statistic.
- Any one-to-one function of a sufficient statistic is also sufficient (since they induce the same partition). We have seen an example of this in Example 3.

3. MINIMAL SUFFICIENT STATISTICS (OPTIONAL)

This is optional material that resolves the non-uniqueness problem with sufficient statistics: if there are many sufficient statistics, is there a way to determine which sufficient statistic is "best" at data reduction?

Definition 5. A sufficient statistic $\phi^*(X)$ is a *minimal sufficient statistic* if for any sufficient statistic $\phi(X)$ there exists some function r such that $\phi^*(X) = r(\phi(X))$.

The minimal sufficient statistic $\phi^*(x)$ induces the coarsest possible partition of the sample space, because we can calculate it from any other sufficient statistic. In this sense, the minimal sufficient statistic gives us the greatest data reduction without a loss of information about parameters.[1] The following theorem (see also Theorem 6.2.13 in Casella and Berger 2002), gives a characterization of minimal sufficient statistics:

Theorem 6 (Lehmann and Scheffé 1950, Theorem 6.3). Let $f(x \mid \theta)$ be the PDF of X and $\phi(X)$ be a function such that, for any x, y, $f(x \mid \theta)/f(y \mid \theta)$ is constant as a function of $\theta \iff \phi(x) = \phi(y)$. Then $\phi(X)$ is minimal sufficient.

Proof. Let $\mathcal{X}$ denote the support of X and assume for simplicity that $f(x \mid \theta) > 0$ for $x \in \mathcal{X}$.

First, we show that the statistic $S = \phi(X)$ is in fact sufficient. Let $A(s) = \{x \in \mathcal{X} : \phi(x) = s\}$ denote the partition induced by S. Fix some element $x_s \in A(s)$ on each partition, and define $h : \mathcal{X} \to \mathbb{R}$ by $h(x) = f(x \mid \theta)/f(x_{\phi(x)} \mid \theta)$. Since $\phi(x) = \phi(x_{\phi(x)})$, $h(x)$ doesn't depend on θ by assertion of the theorem. Let $g(s, \theta) = f(x_s \mid \theta)$. Then

$$f(x \mid \theta) = h(x)g(s, \theta),$$

and by the factorization theorem, S is sufficient.

Let us now show that S is minimal sufficient. Take some sufficient statistic $\phi'(X)$. By the factorization theorem, there exist functions g' and h' such that $f(x \mid \theta) = g'(\phi'(x), \theta)h'(x)$. If $\phi'(x) = \phi'(y)$, then

$$\frac{f(x \mid \theta)}{f(y \mid \theta)} = \frac{g'(\phi'(x), \theta)h'(x)}{g'(\phi'(y), \theta)h'(y)} = \frac{h'(x)}{h'(y)}.$$

Since the ratio is independent of θ, it must be that $\phi(x) = \phi(y)$ as well by assertion of the theorem. Since $\phi'(x) = \phi'(y) \implies \phi(x) = \phi(y)$, it is possible to define a function r such that $\phi(X) = r(\phi'(X))$. □

Example 5. Consider the setup from Example 4. The ratio $f(x \mid \theta)/f(y \mid \theta)$ is independent of $\theta \iff x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)} \iff \phi(x) = \phi(y)$. Therefore, $\phi(X) = (X_{(1)}, X_{(n)})$ is minimal sufficient. ⊠

Example 6. Let $X_1, \ldots, X_n$ be a random sample from the Cauchy distribution with parameter θ, i.e. the distribution with the PDF $f(x \mid \theta) = (\pi(x - \theta))^{-2}$. Then $f(x_1, \ldots, x_n \mid \theta) = (\pi^n \prod_{i=1}^n (x_i - \theta))^{-2}$. By Theorem 6, $\phi(X) = (X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient.⊠

1. In pathological cases, a minimal sufficient statistic may not exist. A sufficient condition for its existence is that the family $\{F_\theta\}_{\theta \in \Theta}$ is dominated, and $X \in \mathbb{R}^k$. See Lehmann and Casella (1998, p. 37)

Finally, note that a minimal sufficient statistic is not unique: any one-to-one function of it is also minimal sufficient. For example, $(\overline{X}_n, S_n)$, $(\overline{X}_n, S_n^2)$, and $(\overline{X}_n, \sum_{i=1}^n X_i^2)$ are all minimal sufficient in Example 3.

4. SUFFICIENCY AND ADMISSIBILITY

We have said that reducing the data to a sufficient statistic does not sacrifice any information about θ. We now justify this statement in two ways:

1. We show that for any decision rule, we can find a randomized decision rule that is based only on the sufficient statistic and that has the same risk function.
2. We show that any estimator that is not a function of the sufficient statistic can be improved upon.

4.1. *Sufficient statistics are sufficient*

Let $\delta(X)$ be a decision rule, and let S be a sufficient statistic. Consider the following (randomized) decision rule $\tilde{\delta}(S)$: draw $\tilde{X}$ from the distribution of $X \mid S$, and take action $\delta(\tilde{X})$. Since X and $\tilde{X}$ have the same distribution, $\delta(X)$ and $\tilde{\delta}(S) = \delta(\tilde{X}(S))$ have the same distribution. Since they have the same distribution, they must have the same risk.

4.2. *Rao-Blackwell Theorem*

Theorem 7 (Rao-Blackwell, Lehmann and Casella 1998, Theorem 1.7.8, p. 47). Suppose that S is a sufficient statistic, the loss function $L(\hat{\theta}, \theta)$ is convex in $\hat{\theta}$, and that the risk of $\hat{\theta}$ is finite. Let $\hat{\theta}_{RB}(s) = E[\hat{\theta}(X) \mid S = s]$. Then for all θ, $R(\hat{\theta}_{RB}, \theta) \leq R(\hat{\theta}, \theta)$.

Proof. Fix θ. By Jensen's inequality, $L(\hat{\theta}_{RB}, \theta) \leq E[L(\hat{\theta}, \theta) \mid S]$. Taking expectation on both sides yields the result. □

1. Note that sufficiency ensures that $\hat{\theta}_{RB}(S)$ is an estimator (it doesn't depend on θ).
2. If the loss is strictly convex, then the inequality is strict unless $\hat{\theta}_{RB}(S) = \hat{\theta}(X)$ with probability one (by strictly convex version of Jensen's inequality). Therefore, estimators that are not a function of sufficient statistics are in general inadmissible. In particular, there exists a non-randomized estimator which is uniformly at least as good as any randomized estimator. So for estimation (when the loss is convex), randomization is not needed (Lehmann and Casella 1998, Corollary 1.7.9).
3. Sometimes (e.g. Casella and Berger 2002, Theorem 7.3.17, or HMC, Theorem 7.3.1), the Rao-Blackwell theorem is stated as a statement about variance of unbiased

estimators. This follows from the version of the theorem stated above. In particular, by the law of iterated expectations, if $\hat{\theta}(X)$ is unbiased, then so is $\hat{\theta}_{RB}(X)$. Since the squared error loss is convex, this implies that for all θ, $\text{var}(\hat{\theta}_{RB}(X)) \leq \text{var}(\hat{\theta})$.

The Rao-Blackwell Theorem is constructive in the sense that it says how a given estimator that's not based on a sufficient statistic can be improved upon. This improvement of an estimator is sometimes called Rao-Blackwellization, and it is often hard to do. Mainly, the theorem is useful because it tells us that we can restrict attention to functions of a sufficient statistic when looking for a good estimator.[2]

Example 3 (continued). If σ^2 is known, the sample mean $\overline{X}_n$ is a sufficient statistic, while the sample median $X_{\lceil n/2 \rceil}$ is not. So the estimator $E[X_{\lceil n/2 \rceil} \mid \overline{X}_n]$ should be better. This is hard to evaluate directly, though we do know from the Rao-Blackwell theorem that the sample median will not be admissible in general. ⊠

Example 7. Let $X_1, \ldots, X_n$ be a random sample from Binomial(k, p). Suppose our parameter of interest is the probability of one success, i.e. $\theta = P(X_j = 1) = kp(1-p)^{k-1}$. One possible estimator is $\hat{\theta} = \sum_{i=1}^n \mathbb{1}\{X_i = 1\}/n$. This estimator is unbiased. Since p is the only unknown parameter, the statistic $S = \sum_{i=1}^n X_i$ is sufficient.

Using the Rao-Blackwell theorem, we can improve $\hat{\theta}$ by considering its conditional expectation given S. We have

$$\begin{aligned}
\hat{\theta}_{RB}(s) &= E\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i = 1\} \;\Big|\; \sum_{j=1}^n X_j = s\right] \\
&= \frac{1}{n}\sum_{i=1}^n P\left(X_i = 1 \;\Big|\; \sum_{j=1}^n X_j = s\right) = P\left(X_1 = 1 \;\Big|\; \sum_{j=1}^n X_j = s\right) \\
&= \frac{P(X_1 = 1, \sum_{j=1}^n X_j = s)}{P(\sum_{j=1}^n X_j = s)} = \frac{P(X_1 = 1, \sum_{j=2}^n X_j = s-1)}{P(\sum_{j=1}^n X_j = s)} \\
&\overset{(1)}{=} \frac{P(X_1 = 1)P(\sum_{j=2}^n X_j = s-1)}{P(\sum_{j=1}^n X_j = s)} \\
&\overset{(2)}{=} \frac{k \cdot p(1-p)^{k-1} \cdot \binom{k(n-1)}{s-1} p^{s-1}(1-p)^{k(n-1)-(s-1)}}{\binom{nk}{s} p^s (1-p)^{kn-s}} \\
&= \frac{k \cdot \binom{k(n-1)}{s-1}}{\binom{nk}{s}} = \frac{k(k(n-1))!(kn-s)!s}{(kn)!(kn-k+1-s)!}
\end{aligned}$$

where (1) uses the fact that X_1 is independent of $(X_2, \ldots, X_n)$, and (2) uses $\sum_{i=1}^n X_i \sim$ Binomial(kn, p), and $\sum_{i=2}^n X_i \sim$ Binomial$(k(n-1), p)$. So our new estimator is

$$\hat{\theta}_{RB}(X) = \frac{k(k(n-1))!(kn - n\overline{X}_n)! \cdot n\overline{X}_n}{(kn)!(kn-k+1-n\overline{X}_n)!}.$$ ⊠

2. An interesting bit of history: Blackwell was briefly at IAS, but the President of Princeton organized a protest against extending his appointment. A famous quote: "Basically, I'm not interested in doing research and I never have been", Blackwell said. "I'm interested in understanding, which is quite a different thing. And often to understand something you have to work it out yourself because no one else has done it."

REFERENCES

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* 2nd ed. Pacific Grove, CA: Duxbury/Thomson Learning.

Halmos, Paul R., and Leonard J. Savage. 1949. "Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics." *Annals of Mathematical Statistics* 20, no. 2 (June): 225–241. https://doi.org/10.1214/aoms/1177730032.

Lehmann, Erich L., and George Casella. 1998. *Theory of Point Estimation.* 2nd ed. New York, NY: Springer. https://doi.org/doi.org/10.1007/b98854.

Lehmann, Erich Leo, and Joseph P. Romano. 2005. *Testing Statistical Hypotheses.* 3rd ed. New York, NY: Springer. https://doi.org/10.1007/0-387-27605-X.

Lehmann, Erich Leo, and Henry Scheffé. 1950. "Completeness, Similar Regions, and Unbiased Estimation–Part I." *Sankhyā* 10 (4): 305–340. https://doi.org/10.1007/978-1-4614-1412-4.