Stat210A: Theoretical Statistics **Lecture Date: September 4, 2014**

Lecture 3: Sufficient Statistics

Lecturer: Michael I. Jordan *Scribe: Dave Moore*

1 Prelude and Intuition

Informally, a statistic T is sufficient for a parameter θ if, knowing $T(X)$, we can throw away the data X without compromising our ability to infer θ. In this sense, sufficient statistics are like a form of lossless compression for datasets. Since modern computers are fast and have lots of memory, this sort of compression is less necessary than it once was, but often still useful.

A simple example of a sufficient statistic is the sample mean for estimating the mean of a Gaussian: once we compute the sample mean, nothing else in the data gives us any additional information about the underlying parameter.

For i.i.d. samples, the *order statistics* are sufficient: given data $X = (X_1, X_2, \ldots, X_n)$, the order statistics $\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right)$ are defined such that the first order statistic $X_{(1)}$ is equal to the smallest of all the X_i's, the second order statistic $X_{(2)}$ is the second-smallest, and so on. Since order statistics throw away the original ordering of the data, they are lossy in a sense, but for i.i.d. samples the original order is unimportant and so the order statistics are sufficient. Depending on the distribution being sampled from, this might be the best we can do, or we might be able to find another set of statistics that compress the data better. The existence of a *finite-dimensional* set of sufficient statistics implies (subject to some regularity conditions) that we are in the exponential family.

2 Definitions

Definition 1 (Frequentist sufficiency[1])**.** Let X be an rv with distribution from a family $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$. Then $T(X)$ is a sufficient statistic for $\mathcal{P}$ if, for every t and θ, the conditional distribution $P_\theta(X|T = t)$ does not depend on θ.

If we're willing to allow a distribution on θ, we can also give a Bayesian definition:

Definition 2 (Bayesian sufficiency)**.** A statistic T is sufficient if X and θ are conditionally independent given $T(X) = t$, i.e., if

$$P(x, \theta|t) = P(x|t)P(\theta|t).$$

Remark 3. This definition implies the frequentist definition. To see this, divide both sides by $P(\theta|t)$, yielding

$$P(x|t, \theta) = P(x|t).$$

The left side is just $P_\theta(x|T = t)$, written with Bayesian notation (conditioning on θ instead of indexing), and the right side shows that this quantity does not depend on θ, satisfying the frequentist definition.

[1] Definition 3.2 from keener's book.

Example 4. (Exponential families) Earlier we defined exponential family distributions in terms of a quantity $T(x)$, which we called a sufficient statistic. Now we'll show that this quantity actually meets the definition of a sufficient statistic.

Suppose $P_\theta(X)$ is in the exponential family, i.e., it has a density

$$p_\theta(x) = h(x) \exp\left(\eta(\theta)^T T(x) - B(\theta)\right).$$

We want to show that $P_\theta(X|T=t)$ does not depend on θ. To do this, we write the conditional density in terms of the joint density, then cancel terms:

$$p_\theta(X|T=t) = \frac{p_\theta(X, T=t)}{p(T=t)} \tag{1}$$

$$= \frac{h(x) \exp\left(\eta(\theta)^T t - B(\theta)\right)}{\int_{T(x)=t} h(x) \exp\left(\eta(\theta)^T T(x) - B(\theta)\right) \mu(dx)} \tag{2}$$

$$= \frac{h(x) \exp\left(\eta(\theta)^T t - B(\theta)\right)}{\exp\left(\eta(\theta)^T t - B(\theta)\right) \int_{T(x)=t} h(x) \mu(dx)} \tag{3}$$

$$= \frac{h(x)}{\int_{T(x)=t} h(x) \mu(dx)} \tag{4}$$

and finally note that the quantity at the end does not depend on θ, satisfying the definition of sufficiency.

The careful reader will be uncomfortable with the joint density $p_\theta(X, T=t)$, since T is actually a deterministic function of X, so this density is only supported on a measure-zero subset of the joint space. A fully rigorous proof would require more careful treatment of conditional probabilities; for an example, see the proof of the factorization theorem 3.6 in section 6.4 keener's book.

2.1 Intuition: "fake data"

One way to motivate the frequentist definition (definition 1) is to show that knowledge of the sufficient statistic T allows us to construct a "fake dataset" $\widetilde{X}$ having the same distribution as the true dataset X.

Let X and Y be independent r.v.s with density

$$f_\theta(x) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Also consider an additional r.v., $U \sim \text{Uniform}(0,1)$, independent of X and Y. Now let

$$T = X + Y,$$
$$\widetilde{X} = UT,$$
$$\widetilde{Y} = (1-U)T.$$

Here $\widetilde{X}$ and $\widetilde{Y}$ will consitute our "fake dataset"; note that $\widetilde{X} + \widetilde{Y} = T$.

To derive the joint density on $\widetilde{X}$ and $\widetilde{Y}$, we first need the density of T, which we can get by differentiating the cumulative distribution function (cdf). To find the cdf, $P(T \leq t)$, we use a trick: first write down the *conditional* cdf $P(T \leq t|Y=y)$, then take the expectation over Y using the tower rule to get an unconditional

cdf. Let's begin with the conditional cdf:

$$\begin{aligned}
P(T \le t|Y = y) &= P(X + Y \le t|Y = y) \\
&= E\left[1_{X+Y\le t}|Y = y\right] \text{(rewriting the probability as expectation of an indicator)}, \\
&= \int 1_{(x+y\le t)} P(dx) \\
&= F_X(t - y).
\end{aligned}$$

Now we'll use this to derive the unconditional cdf:

$$\begin{aligned}
F_T(t) &= P(T \le t) \\
&= E\left[P(T \le t|Y)\right] \\
&= E\left[F_X(t - Y)\right] \\
&\text{and we substitute } F_X(t - Y) = 1 - e^{-\theta(t-Y)}, \text{ obtained by integrating the density, to yield} \\
&= E\left[1 - e^{-\theta(t-Y)}\right] \\
&= \int_0^t \left(1 - e^{-\theta(t-y)}\right) \theta e^{-\theta y} dy \\
&= 1 - e^{-\theta t} - t\theta e^{-\theta t}.
\end{aligned}$$

Finally, we differentiate this cdf to obtain the density for T,

$$p_T(t) = F_T'(t) = t\theta^2 e^{-\theta t}.$$

Since T and U are independent, their joint density is just

$$p_\theta(t, u) = \begin{cases} t\theta^2 e^{-\theta t} & t \ge 0, 0 \le u \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

Using this, we can finally write the joint measure on $\widetilde{X}$ and $\widetilde{Y}$:

$$P\left(\begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} \in B\right) = \iint 1_B \begin{pmatrix} tu \\ t(1-u) \end{pmatrix} p_\theta(t, u)\, dudt.$$

Substituting $x = ut$, $du = dx/t$, and using Fubini's theorem to swap the integrals, we get

$$P\left(\begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} \in B\right) = \iint 1_B \begin{pmatrix} x \\ t - x \end{pmatrix} t^{-1} p_\theta\left(t, \frac{x}{t}\right) dtdx$$

Substituting $y = t - x$,

$$P\left(\begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} \in B\right) = \iint 1_B \begin{pmatrix} x \\ y \end{pmatrix} (x + y)^{-1} p_\theta\left(x + y, \frac{x}{x + y}\right) dydx$$

from which we conclude that $\widetilde{X}$ and $\widetilde{Y}$ have joint density

$$\frac{p_\theta\left(x + y, \frac{x}{x+y}\right)}{x + y} = \begin{cases} \theta^2 e^{-\theta(x+y)} & x \ge 0, y \ge 0 \\ 0 & \text{otherwise} \end{cases}.$$

Note this is the same joint density as for X and Y. Since our fake data have the same distribution as the real data, they must give us the same information about the parameter θ. But we constructed the fake data using only T and U, and U is clearly irrelevant. This suggests that all relevant information about θ is contained in T.

To tie this back to the definition of sufficiency, consider

$$P_\theta\left(\begin{pmatrix} X \\ Y \end{pmatrix} \in B | T = t\right) = P_\theta\left(\begin{pmatrix} \widetilde{X} \\ \widetilde{Y} \end{pmatrix} \in B | T = t\right) \tag{5}$$

$$= P\left(\begin{pmatrix} Ut \\ (1-U)t \end{pmatrix} \in B\right). \tag{6}$$

Since the last line has no dependence on θ, T is a sufficient statistic.

Theorem 5. *Suppose that $X \sim \mathcal{P} = \{P_\theta : \theta \in \Omega\}$ and that T is sufficient for the family $\mathcal{P}$. Then for any estimatior $\delta(X)$ of $g(\theta)$, there exists a (possibly randomized) estimator based on T that has the same risk function as $\delta(X)$.*

Proof. The estimator $\delta(\widetilde{X})$, with $\widetilde{X}$ constructed as "fake data" from T, has the same risk as $\delta(X)$ because X and $\widetilde{X}$ have the same distribution. □

3 Factorization

We can give a different definition of sufficiency that's often more useful in practice. Let $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ be a family of distributions *dominated* by some measure μ. That is, for all θ, $P_\theta \ll \mu$, so by Radon-Nikodym we can assume the existence of densities p_θ.

Theorem 6. (Factorization) *A statistic T is sufficient if there exists $g_\theta \geq 0$ and $h \geq 0$ such that*

$$p_\theta(x) = g_\theta\left(T(x)\right) h(x)$$

almost everywhere. (we allow exceptions on set of measure zero wrt μ, since such discrepancies won't affect the values of integrals over these densities).

Note that g_θ, the factor involving θ, depends only on the statistic T, not the full dataset x.

Proof. See Theorem 3.6 in keener's book. □

Example 7. For distributions in the exponential family,

$$p_\theta(x) = h(x) \exp\left(\eta(\theta)^T T(x) - B(\theta)\right),$$

the factorization is straightforward: let

$$g_\theta(x) = \exp\left(\eta(\theta)^T T(x) - B(\theta)\right)$$

and

$$h(x) = h(x).$$

Example 8. Let $X_i \sim \text{Uniform}(\theta, \theta+1)$ be iid random variables, with density $f_\theta(x_i) = 1_{(\theta,\theta+1)}(x_i)$. The joint density is given by

$$p_\theta(x) = \prod_{i=1}^n 1_{(\theta,\theta+1)}(x_i).$$

Using some cleverness we can rewrite this as

$$p_\theta(x) = 1_{(\theta,\infty)}\left(\min_i x_i\right) 1_{(-\infty,\theta+1)}\left(\max_i x_i\right);$$

it's easy to check that this is equivalent. By the factorization theorem, using the trivial $h(x) = 1$, this implies that

$$T = \left(\min_i x_i, \max_i x_i\right)$$

is a sufficient statistic for θ.

Example 9. Let X_i be independent random variables with cdf

$$P(X_i \le x) = x^{t_i\theta},$$

for $x \in (0,1)$ and a sequence of known constants t_i. Differentiating the cdf gives a density

$$f_\theta(x_i) = t_i\theta x^{t_i\theta}/x,$$

which implies a joint density

$$p_\theta(x) = \theta^n \left(\prod_{i=1}^n \frac{t_i}{x}\right)\left(\prod_{i=1}^n x_i^{t_i}\right)^\theta.$$

Now the factorization theorem allows us to read off the sufficient statistic

$$T(x) = \prod_{i=1}^n x_i^{t_i}.$$

Since any one-to-one transformation of a sufficient statistic is also sufficient, we could equivalently choose

$$T(x) = \sum_{i=1}^n t_i \log x_i,$$

which might be nicer to work with.