STAT 713 sp 2023 Lec 01 slides

Data reduction part 1: Sufficiency

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Setup and notation

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ represent the set of rvs from an experiment.
- Use $\mathbf{x} = (x_1, \ldots, x_n)$ to represent a specific set of values for the rvs in $\mathbf{X}$.
- Let $f(\mathbf{x}; \theta)$ or $p(\mathbf{x}; \theta)$ denote the joint pdf or pmf of the rvs in $\mathbf{X}$, resp.
- The distribution of $\mathbf{X}$ depends on a parameter (or some parameters) $\theta \in \Theta$.
- A function $T(\mathbf{X})$ of the rvs in $\mathbf{X}$ is called a *statistic*.

Goal: Learn about θ from a realization of $\mathbf{X}$ via the value of a statistic $T(\mathbf{X})$.

Key concepts of data reduction in the rough:

1. A *sufficient statistic* carries all the information about θ from $\mathbf{X}$.
2. A *minimal sufficient statistic* carries the above and no more than this.
3. An *ancillary statistic* carries no information about θ.
4. A *complete statistic* retains info about θ under any non-deg. transformation.

We think of computing a statistic $T(\mathbf{X})$ as "reducing" or summarizing the data.

Example: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Exponential}(\lambda)$, $\lambda > 0$, and consider the statistics

$$T_1(\mathbf{X}) = X_{(1)}, \qquad T_2(\mathbf{X}) = \bar{X}_n, \qquad T_3(\mathbf{X}) = S_n^2,$$

$$T_4(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right), \quad T_5(\mathbf{X}) = X_{(1)}/X_{(n)}, \quad T_6(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)}).$$

Sufficiency, *minimality*, *ancillarity*, and *completeness* address (resp) the questions

1. Which reductions of the data do not discard any information about λ?
2. Which ones keep all relevant information about λ, but discard all else?
3. Which ones discard all information about λ?
4. Which ones retain info about λ under any non-deg. transformation?

Sufficient statistic

A statistic $T(\mathbf{X})$ is *sufficient for* θ if the joint pdf/pmf of $\mathbf{X}$, conditional on the value of $T(\mathbf{X})$, does not depend on θ.

Affect of θ on the distribution of $\mathbf{X}$ is expressed fully in the value $T(\mathbf{X})$.

Two samples $\mathbf{X}_1$, $\mathbf{X}_2$ carry same info about θ if $T(\mathbf{X}_1) = T(\mathbf{X}_2)$.

Example: For the Bernoulli(p) distribution, it seems like the random samples

$$\mathbf{X}_1 = (0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1)$$
$$\mathbf{X}_2 = (1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0)$$

should lead to the same inferences about $p \in (0, 1)$. Why?

More notation

- Denote by $\mathcal{X}$ the support of $\mathbf{X}$.
- Let $\mathcal{T} = \{t : T(\mathbf{x}) = t \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the support of $T(\mathbf{X})$.
- Use subscript θ in $\mathbb{E}_\theta$ and P_θ to indicate dependence on θ.

For the discrete case, if $T(\mathbf{X})$ is a sufficient statistic, then

$$P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t)$$

is free of θ for every $t \in \mathcal{T}$ and $\mathbf{x} \in \mathcal{X}$.

Theorem (Checking for sufficiency, cf. Thm 6.2.2 in CB)

For $\mathbf{X}$ with joint pdf/pmf $f(\mathbf{x};\theta)$, $T(\mathbf{X})$ is a sufficient statistic for θ if

$$\frac{f(\mathbf{x};\theta)}{f_T(T(\mathbf{x});\theta)}$$

is free of θ for all $\mathbf{x} \in \mathcal{X}$, where f_T is the pdf/pmf of $T = T(\mathbf{X})$.

Exercise: Prove for the discrete case.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Poisson}(\lambda)$, where $\lambda > 0$.

1. Check whether $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for λ...
2. Interpret.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Uniform}(0, \theta)$, where $\theta > 0$.

Check whether $T(\mathbf{X}) = X_{(n)}$ is a sufficient statistic for θ.

Theorem (Easier check for suff. by factorization, cf. Thm 6.2.6 in CB)

For $\mathbf{X}$ with joint pdf/pmf $f(\mathbf{x};\theta)$,

1. *$T = T(\mathbf{X})$ is a suff. stat. for θ iff there exist non-neg fns $g(t,\theta)$ and $h(\mathbf{x})$ st*

$$f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta) \cdot h(\mathbf{x})$$

   for all $\theta \in \Theta$ and all $\mathbf{x} \in \mathcal{X}$.

2. *As a consequence, the pdf/pmf $f_T(t;\theta)$ of T is given by*

$$f_T(t;\theta) = g(t;\theta)\tilde{h}(t)$$

   for some function $\tilde{h}(t) \geq 0$ not depending on θ.

Can be used to *find* a sufficient statistic.

Exercise: Prove the result for the discrete case.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Beta}(1, \theta)$, where $\theta > 0$.

Check whether $T(\mathbf{X}) = \prod_{i=1}^{n}(1 - X_i)$ is a sufficient statistic for θ.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} f_X(x; \mu) = \pi^{-1}[1 + (x - \mu)^2]^{-1}$, where $\mu \in \mathbb{R}$.

Check whether $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for μ.

Exercise: Let $Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} f_Y(y; \alpha) = \alpha y^{-(\alpha+1)} \mathbf{1}(y > 1)$.

Find a sufficient statistic for α.

Corollary (1:1 function of a sufficient statistic)

If $T(\mathbf{X})$ is a sufficient statistic for θ and $a(t)$ is a 1:1 function not depending on θ, then $a(T(\mathbf{X}))$ is a sufficient statistic for θ.

Exercise: Prove result.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Bernoulli}(p)$, where $p \in (0, 1)$. Check sufficiency of

1. $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$
2. $T'(\mathbf{X}) = \bar{X}_n$

A statistic may consist of multiple functions of the data:

$$T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X})), \quad k \geq 1.$$

A parameter θ may consist of several values $\theta = (\theta_1, \ldots, \theta_d)$, $d \geq 1$.

We can still establish sufficiency using the factorization theorem.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$. Check sufficiency of

1. $T(\mathbf{X}) = (\bar{X}_n, S_n^2)$
2. $T'(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$

Theorem (Cannot make a suff. statistic from a non-sufficient statistic)

Let $T(\mathbf{X})$ and $U(\mathbf{X})$ be statistics such that $T(\mathbf{X}) = a(U(\mathbf{X}))$ for a function a not depending on θ. If $T(\mathbf{X})$ is sufficient for θ, $U(\mathbf{X})$ is also sufficient for θ.

If you start with a statistic that is not sufficient, no function of it will be sufficient.

If you reduce the data too much, you cannot recover any info you have discarded.

Note that the entire sample $T(\mathbf{X}) = \mathbf{X}$ is always sufficient!

Exercise: Prove the result.

Exercise: Let $Y_1, \ldots, Y_n$ be ind. rvs and $x_1, \ldots, x_n$ real numbers such that

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{where} \quad \log(\lambda_i) = \beta_0 + \beta_1 x_i, \quad \text{for } i = 1, \ldots, n.$$

Show that we lose no information about $(\beta_0, \beta_1) \in \mathbb{R}^2$ by reducing $Y_1, \ldots, Y_n$ to

$$T(\mathbf{Y}) = \left(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} x_i Y_i \right).$$

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} f_X(x; \theta)$, $\theta \in \Theta$.

Show that $T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$ is sufficient for θ.

Theorem (Find suff. stat. in exp. family, cf. Thm 6.2.10 in CB)

If $X_1, \ldots, X_n \overset{\text{ind}}{\sim} f(x;\theta)$, *where* $f(x;\theta)$ *belongs to an exponential family such that*

$$f(x;\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right), \quad x \in \mathbb{R}, \quad \theta \in \Theta,$$

then a sufficient statistic for θ *is given by*

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right).$$

Exercise: Prove the result.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Beta}(\alpha, \beta)$, where $\alpha > 0$, $\beta > 0$.

Find a sufficient statistic for (α, β) using the exponential family result.