STAT 713 sp 2023 Lec 03 slides

Data reduction part 3: Completeness

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 $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Exponential}(\lambda)$ and consider the statistics

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

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

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?

Complete statistic

A statistic $T(\mathbf{X})$ is called *complete* if for every function g such that $\mathbb{E}_\theta g(T(\mathbf{X})) = 0$ for all θ, we have $P_\theta(g(T(\mathbf{X})) = 0) = 1$ for all θ.

- If we can't construct an unbiased estimator of 0 from $T(\mathbf{X})$, it is complete.
- If we can find nonzero g st $\mathbb{E}_\theta g(T(\mathbf{X})) = 0$ for all θ, $T(\mathbf{X})$ is not complete.

This property becomes important later on, when we search for "best" estimators.

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

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

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

Determine which properties each statistic has among *sufficiency*, *minimality*, *ancillarity*, and *completeness*.

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\theta, \theta^2)$, $\theta \in \mathbb{R}$.

1. Find a sufficient statistic for θ.
2. Check whether it is a complete statistic.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Bernoulli}(p)$, $0 < p < 1$.

1. Check whether X_1 is a complete statistic. Is it a sufficient statistic?
2. Find a minimal sufficient statistic for p.
3. Check whether it is a complete statistic.

Exercise: Let $X_1, \dots, X_n \stackrel{\text{ind}}{\sim} f(x; \mu) = e^{-(x-\mu)}\mathbf{1}(x > \mu)$, where $\mu \in \mathbb{R}$.

1. Find a minimal sufficient statistic for μ.
2. Check whether it is a complete statistic.

Theorem (Functions of complete statistics)

Any known function of a complete statistic is complete.

Exercise: Prove the result.

In most situations it is hard to check completeness using the definition. Try this:

Theorem (Completeness via exponential family, cf. Thm 6.2.25 in CB)

Suppose $X_1, \dots, X_n \overset{\text{ind}}{\sim} f(x;\theta)$, *where*

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

and the range of $(w_1(\theta), \dots, w_k(\theta))$ *over* $\theta \in \Theta$ *contains an open set in* $\mathbb{R}^k$. *Then*

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

is a complete minimal sufficient statistic for θ.

- Use to find complete statistics in full exponential families.
- Completeness not guaranteed for curved exponential families.

Exercise: Check for completeness in family Normal(θ, θ^2), $\theta > 0$.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Gamma}(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$.
Identify, if possible, a complete sufficient statistic.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} (2x/\theta)e^{-x^2/\theta} \cdot \mathbf{1}(x > 0)$, where $\theta > 0$.

Identify, if possible, a complete sufficient statistic.

Theorem

Let $T(\mathbf{X})$ be a complete sufficient statistic for a parameter θ. Then

1. *(Lehmann, cf. Thm 6.2.28 in CB): $T(\mathbf{X})$ is minimal sufficient if at least one minimal sufficient statistic exists.*
2. *(Basu, cf. Thm 6.2.24 in CB): $T(\mathbf{X})$ is indep. of every ancillary statistic.*

Lehmann's says only minimal sufficient statistics can be complete and sufficient.

Exercise: Go through proofs of the above results.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma_0^2)$, where $\mu \in \mathbb{R}$ and σ_0^2 is known.

Use Basu's theorem to argue that $\bar{X}_n$ and S_n^2 are independent.

Exercise: Let $X_1, \ldots, X_n \stackrel{\text{ind}}{\sim} \text{Gamma}(\alpha_0, \beta)$, where $\beta > 0$ and α_0 is known.

1. Find a complete sufficient statistic for β.
2. Find a statistic that is ancillary to β.
3. Are these two statistics independent?

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

1. Check sufficiency and completeness of $T(\mathbf{X}) = \prod_{i=1}^n X_i(1 - X_i)$.
2. Check whether $T'(\mathbf{X}) = (\prod_{i=1}^n X_i, \prod_{i=1}^n (1 - X_i))$ is complete.

Theorem (Cannot construct an ancillary from a complete statistic)

If $U(\mathbf{X})$ is a known function of $T(\mathbf{X})$ with a non-degenerate distribution and $U(\mathbf{X})$ is ancillary, then $T(\mathbf{X})$ is not complete.

A *complete statistic* retains info about θ under any non-deg. transformation.

Exercise: Prove the above.

Hopefully a useful summary:

To show that $T(\mathbf{X})$ is complete, you can:

1. Use the exponential family result (which also gives sufficiency).
2. Use the definition: Let g be any function such that $\mathbb{E}_\theta g(T(\mathbf{X})) = 0$ for all θ and show that this implies $g(T(\mathbf{X})) = 0$ with probability 1.
   (Try going this route if the support depends on the parameter).

To show that $T(\mathbf{X})$ is not complete, you can:

1. Show that a function of $T(\mathbf{X})$ is ancillary.
2. Use the definition: Let g be a function such that $\mathbb{E}_\theta g(T(\mathbf{X})) = 0$ for all θ and show that we can still have $g(T(\mathbf{X})) \neq 0$ with positive probability.