# STAT 517:Sufficiency

## Minimal sufficiency and Ancillary Statistics. Sufficiency, Completeness, and Independence

Prof. Michael Levine

March 1, 2016

# Motivation

- Not all sufficient statistics created equal...
- So far, we have mostly had a single sufficient statistic for one parameter or two for two parameters (with some exceptions)
- Is it possible to find the **minimal sufficient statistics** when further reduction in their number is impossible?
- Commonly, for $k$ parameters one can get $k$ minimal sufficient statistics

# Example

- $X_1, \ldots, X_n \sim Unif(\theta - 1, \theta + 1)$ so that

$$f(x;\theta) = \frac{1}{2} I_{(\theta-1,\theta+1)}(x)$$

where $-\infty < \theta < \infty$

- The joint pdf is

$$2^{-n} \left\{ I_{(\theta-1,\theta+1)}(\min x_i) \right\} \left\{ I_{(\theta-1,\theta+1)}(\max x_i) \right\}$$

- It is intuitively clear that $Y_1 = \min x_i$ and $Y_2 = \max x_i$ are joint minimal sufficient statistics

# Occasional relationship between MLE's and minimal sufficient statistics

- Earlier, we noted that the MLE $\hat{\theta}$ is a function of one or more sufficient statistics, when the latter exists
- If $\hat{\theta}$ is itself a sufficient statistic, then it is a function of others...and so it may be a sufficient statistic
- E.g. the MLE $\hat{\theta} = \bar{X}$ of $\theta$ in $N(\theta, \sigma^2)$, $\sigma^2$ is known, is a minimal sufficient statistic for $\theta$
- The MLE $\hat{\theta}$ of $\theta$ in a $P(\theta)$ is a minimal sufficient statistic for $\theta$
- The MLE $\hat{\theta} = Y_{(n)} = \max_{1 \leq i \leq n} X_i$ of $\theta$ in a $Unif(0, \theta)$ is a minimal sufficient statistic for $\theta$
- $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = \frac{n-1}{n} S^2$ of $\theta_1$ and $\theta_2$ in $N(\theta_1, \theta_2)$ are joint minimal sufficient statistics for $\theta_1$ and $\theta_2$

# Formal definition

- A sufficient statistic $T(X_1, \ldots, X_n)$ is called a **minimal sufficient statistic** if it is a function of any other sufficient statistic

# When MLE and minimal sufficient statistics have nothing in common with each other: Example I

- Take again $X_1, \ldots, X_n \sim Unif(\theta - 1, \theta + 1)$
- Clearly, $\theta - 1 < Y_1 < Y_n < \theta + 1$, or

$$Y_n - 1 < \theta < Y_1 + 1$$

- To achieve the maximum possible value of the likelihood function $\left(\frac{1}{2}\right)^n$, choose any $\theta$ between $Y_n - 1$ and $Y_1 + 1$; a common choice as MLE is the average of two endpoints

$$\hat{\theta} = \frac{Y_1 + Y_n}{2}$$

- Note that the resulting $\hat{\theta}$ is not even a sufficient statistic...and, therefore, cannot be a minimal sufficient statistic

# A more general location family setting I

- The above example is a location family $X_i = \theta + W_i$ where $W_i \sim Unif(-1, 1)$
- Take a general location family with $W_i$ having a pdf $f(w)$ and cdf $F(w)$
- We know that the order statistics $Y_1 < Y_2 < \cdots < Y_n$ form a set of sufficient statistics in this case...Can we do better?
- If $f(w)$ is a $N(0, 1)$ pdf, $\bar{X}$ is both the MVUE and MLE of $\theta$; moreover, $\bar{X}$ is a minimal sufficient statistic
- Take $f(w) = e^{-w}$ for $w > 0$ and zero elsewhere; here, $Y_1$ is a sufficient statistic and the MLE - so $Y_1$ is a minimal sufficient statistic

# A more general location family setting II

- On the contrary, for the logistic location family, the MLE of $\theta$ exists and is easy to compute...nevertheless, the order statistics are *minimal sufficient* in this case
- If $f(w)$ is a Laplace pdf with the location parameter $\theta$, the median $Q_2$ is an MLE; however, yet again, the order statistics are *minimal sufficient* in this case
- This latter situation is, in general, more common for location models

# Minimal sufficiency and completeness

- In general, if the minimal sufficient statistic exists (and it almost always does), any complete sufficient statistic is also a minimal sufficient statistic
- The converse is not true, however; from the uniform example, note that

$$\mathbb{E}\left[\frac{Y_n - Y_1}{2} - \frac{n-1}{n+1}\right] = 0$$

for all $\theta$

# Ancillary statistics

- A quick example - for $X_1, \ldots, X_n \sim N(\theta, 1)$ the distribution of $S^2$ does not depend on $\theta$
- Alternatively, take $X_1, X_2 \sim \Gamma(\alpha, \theta)$ where $\alpha > 0$ is known and recall that $Z = \frac{X_1}{X_1 + X_2}$ has a beta distribution that does not depend on $\theta$...Thus, $Z$ is an **ancillary** statistic for this sample size 2 w.r.t $\theta$
- In general, select a location family $X_i = \theta + W_i$, $i = 1, \ldots, n$, where $-\infty < \theta < \infty$ is a parameter and $W_1, \ldots, W_n \sim f(w)$ that doesn't depend on $\theta$
- The common pdf of $X_i$ is $f(x - \theta)$...any **location-invariant** statistic $Z = u(X_1, \ldots, X_n)$ s.t. $Z = u(W_1 + \theta, \ldots, W_n + \theta) = u(W_1, \ldots, W_n)$ for all $\theta$ is an ancillary statistic
- Sample variance is one such statistic...Sample range $R = \max X_i - \min X_i$ is another...Finally, the absolute mean deviation from the sample median

$$\frac{1}{n} \sum_{i=1}^{n}$$

# Basu's theorem

- Let $X_1, \ldots, X_n \sim f(x;\theta)$ where $\theta \in \Omega$ and $\Omega$ is an interval. Let $Y_1$ be a complete and sufficient statistic for $\theta$
- $Z = u(X_1, \ldots, X_n)$ is another statistic
- Distribution of $Y_1$ doesn't depend on $\theta \rightarrow Z$ is independent of $Y_1$

# Main Ideas of Basu's Theorem

- If $Y_1$ is a sufficient statistic and is independent of $Z \rightarrow$ distribution of $Z$ doesn't depend on $\theta$
- If $Y_1$ is a sufficient and a *complete* statistic, distribution of $Z$ doesn't depend on $\theta \rightarrow Y_1$ and $Z$ are independent
- The second case is very easily satisfied for regular exponential families

# Example

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- $\bar{X}$ is a sufficient statistic for $\mu$ while $S^2$ is ancillary for $\mu$ (location-invariant)
- Thus, $\bar{X}$ and $S^2$ are independent

# Example

- $X_1, \ldots, X_n \sim e^{-(x-\theta)}$ for $\theta < x < \infty$
- $Y_1 = \min_{1 \leq i \leq n} X_i$ is a complete sufficient statistic for $\theta$
- Any location invariant statistic, e.g. $S^2$ or the sample range are independent of $Y_1$

# Example

- $X_1, X_2 \sim f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$
- $Y_1 = X_1 + X_2$ is a complete sufficient statistic for $\theta$
- $Y_1$ is independent of the scale invariant statistic $\frac{X_1}{X_1+X_2}$ that is beta distributed

# Example

- $X_1, \ldots, X_n \sim N(\theta_1, \theta_2)$
- $\bar{X}$ and $S^2$ are joint sufficient statistics for $\theta_1$ and $\theta_2$
- Thus, the location and scale invariant statistic

$$Z = \frac{\sum_{i=1}^{n}(X_{i+1} - X_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

is independent of both $\bar{X}$ and $S^2$

# When a sufficient statistic is not complete

- An incomplete sufficient statistic may still contain some information about the parameter
- $X_1, \ldots, X_n \sim \frac{1}{2} I_{(\theta-1,\theta+1)}(x)$
- Let $Y_1 = \min X_i$ and $Y_n = \max X_i$; $T_1 = \frac{Y_1+Y_n}{2}$ is an MLE of $\theta$
- $T_2 = Y_n - Y_1$ is an ancillary statistic...however

$$Var(T_1|t_2) = \frac{(2-t_2)^2}{12}$$

and so the distribution of $T_2$ has some information about $\theta$