# STAT 517:Sufficiency

## Exponential families of distributions. Estimation of functions of parameters

Prof. Michael Levine

April 21, 2015

# The definition of joint sufficiency

- Let $X_1, \ldots, X_n \sim f(x; \boldsymbol{\theta})$ where $\theta \in \Omega \subset \mathbb{R}^p$ while $\mathcal{S}$ is the support of $X$
- Let $\mathbf{Y} = (Y_1, \ldots, Y_m)'$ be a vector of statistics where each $Y_i = u_i(X_1, \ldots, X_n)$, $i = 1, \ldots, m$
- The pdf of $\mathbf{Y}$ is $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ - a function of $\mathbf{y} \in \mathbb{R}^m$
- $\mathbf{Y}$ is **jointly sufficient** for $\boldsymbol{\theta}$ iff

$$\frac{\prod_{i=1}^n f(x_i; \theta)}{f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} = H(x_1, \ldots, x_n)$$

  for all $x_i \in \mathcal{S}$ where the function $H$ does not depend on $\boldsymbol{\theta}$
- Most of the time, $m = p$

# The multivariate factorization theorem

- The vector $\mathbf{Y}$ is jointly sufficient for the parameter $\boldsymbol{\theta} \in \Omega$ iff we can find two non-negative functions $k_1$ and $k_2$ s.t.

$$\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = k_1(\mathbf{y}; \boldsymbol{\theta}) k_2(x_1, \ldots, x_n)$$

for all $x_i \in \mathcal{S}$ where $k_2(x_1, \ldots, x_n)$ does not depend on $\boldsymbol{\theta}$

# Example

- Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with both parameters unknown; $p = 2$
- Easily verified that

$$\prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}} e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}}$$

- The joint sufficient statistic is $\mathbf{Y} = (\bar{X}, \sum_{i=1}^{n}(X_i - \bar{X})^2)$

# Completeness

- The pdf of the $k$-variate random variable $\mathbf{V} = (V_1, \ldots, V_k)'$ is

$$f(v_1, \ldots, v_k; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega \subset \mathbb{R}^p)$$

- Let $u(v_1, \ldots, v_k)$ be the function of any or all $v_1, \ldots, v_k)$ (but not a function of $\boldsymbol{\theta}$)
- If

$$\mathbb{E}\left[u(V_1, \ldots, V_k)\right] = 0$$

for all $\boldsymbol{\theta} \in \Omega$ implies that $u(v_1, \ldots, v_k) = 0$, we say that the family of pdfs $f(v_1, \ldots, v_k; \boldsymbol{\theta})$ is a complete family

# Some remarks

- In the multivariate case, we typically look for estimators of functions of $\boldsymbol{\theta}$
- For example, if the distribution is $N(\theta_1, \theta_2)$, we have $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ and the objects of interest are $\delta_1 = g_1(\boldsymbol{\theta}) = \theta_1$ and $\delta_2 = g_2(\boldsymbol{\theta}) = \sqrt{\theta_2}$
- The Rao-Blackwell and Lehmann-Scheffe theorems can be easily extended in the natural way. Let $\delta = g(\boldsymbol{\theta})$ be the parameter of interest and $\mathbf{Y}$ is a complete sufficient statistic
- If $\mathbb{E}T = \delta$ and $T = T(\mathbf{Y})$, $T$ is the unique MVUE of $\delta$

# A regular exponential family

- Let $X \sim f(x;\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Omega \subset \mathbb{R}^p$, $\mathcal{S}$ is a support of $X$
- Continuous $X$ - $\mathcal{S} = (a, b)$ with potentially infinite $a, b$; discrete $X$ - $\mathcal{S} = \{a_1, a_2, \ldots, \}$
- Represent

$$f(x;\boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{m} p_j(\boldsymbol{\theta})K_j(x) + H(x) + q(\boldsymbol{\theta}_1, \ldots, \theta_m)\right]$$

  for all $x \in \mathcal{S}$ and zero elsewhere
- Here $m = p$ - the number of sufficient statistics equals the number of parameters
- A vector $p_1(\boldsymbol{\theta}), \ldots, p_m(\boldsymbol{\theta})$ is called a **natural parameter**

# Regularity

- The support does not depend on $\boldsymbol{\theta}$
- $\Omega$ contains a non-empty, $m$-dimensional open rectangle
- $p_j(\boldsymbol{\theta})$, $j = 1, \ldots, m$ are non-trivial, functionally independent, continuous functions of $\boldsymbol{\theta}$
  - Continuous $X$ - $K_j^{'}(x)$ are continuous, none of them is a linear function of the others, $H(x)$ is a continuous function
  - Discrete $X$ - $K_j(x)$ are nontrivial functions, none of them is a linear function of the others

# Sufficiency and completeness in exponential families

- Clearly,

$$\prod_{i=1}^{m} f(x_i;\boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{m} p_j(\boldsymbol{\theta}) \sum_{i=1}^{n} K_j(x_i) + nq(\boldsymbol{\theta})\right] \exp\left[\sum_{i=1}^{n} H(x_i)\right]$$

- By factorization theorem, $Y_1 = \sum_{i=1}^{n} K_1(x_i)$, $\ldots, Y_m = \sum_{i=1}^{n} K_m(x_i)$, are joint sufficient statistics for the $m$-dimensional parameter vector $\boldsymbol{\theta}$

# Sufficiency and completeness in exponential families

- Can be shown that the joint pdf of $\mathbf{Y}$ is of the form

$$R(\mathbf{y}) \exp\left[\sum_{j=1}^{m} p_j(\boldsymbol{\theta}) y_j + nq(\boldsymbol{\theta})\right]$$

  where the function $R(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$
- Using some results from analysis, one can also show that this sufficient statistic is *always* complete for a regular exponential family when $n > m$ ; thus, $Y_1, \ldots, Y_m$ are **joint complete sufficient statistics** for $\boldsymbol{\theta}$

# Example

- Take $X_1, \ldots, X_n \sim N(\theta_1, \theta_2)$ with $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$
- The pdf is

$$f(x; \theta_1, \theta_2) = \exp\left\{ -\frac{1}{2\theta_2} x^2 + \frac{\theta_1}{\theta_2} x - \frac{\theta_1^2}{2\theta_2} - \log\sqrt{2\pi\theta_2} \right\}$$

- Clearly, $Y_1 = \sum_{i=1}^n X_i^2$ and $Y_2 = \sum_{i=1}^n X_i$ are joint complete sufficient statistics for $\theta_1$ and $\theta_2$
- Their one-to-one transformations $Z_1 = \frac{Y_2}{n}$ and $Z_2 = \frac{Y_1 - Y_2^2/n}{n-1} = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are also joint complete sufficient statistics for $\theta_1$ and $\theta_2$
- Thus, since $\mathbb{E}Z_1 = \theta_1$ and $\mathbb{E}Z_2 = \theta_2$, $Z_1$ and $Z_2$ are MVUE's of $\theta_1$ and $\theta_2$

# Example II

- Let $X_1, \ldots, X_n \sim N(\mu, a^2\mu^2)$ with a known $a > 0$
- This model may describe a series of measurements on an object using a measuring device(e.g. a series of weights) whose accuracy (standard deviation) is relative to the true (known) measure of the object rather than being constant
- Write

$$f(x;\mu) = \exp\left\{x\frac{1}{a^2\mu} - x^2\frac{1}{2a^2\mu^2} - \frac{1}{2}\left(\frac{1}{a^2} + \log(2\pi a^2\mu^2\right)\right\}$$

- This is a *two-parameter* exponential family where $K_1(x) = x$ and $K_2(x) = x^2$; thus, we have $m = 2$ while $p = 1$

# Example II

- The natural parameters are $p_1(\mu) = \frac{1}{a^2\mu}$ and $p_2(\mu) = -\frac{1}{2a^2\mu^2}$
- Note - the set $\left\{ \left( \frac{1}{a^2\mu}, -\frac{1}{2a^2\mu^2} \right), \mu \in \mathbb{R} \right\}$ consists only of a single parabola and does not contain a two-dimensional rectangle

# Sufficiency of order statistics

- Take $X_1, \ldots, X_n \sim F(x)$ and $Y_1 < Y_2 < \cdots Y_n$ be the order statistics
- Recall that, given $Y_1 = y_1, \ldots, Y_n = y_n$, the conditional distribution of $X_1, \ldots, X_n$ is discrete with prob. $\frac{1}{n!}$ on each of $n1$ possible permutations
- Thus, conditional distribution does not depend on $F(x)$ and, w.r.t., $F(x)$, the set of order statistics is always a sufficient statistic
- It can also be shown that a set of order statistics is complete as well