STAT 517: Sufficiency
Sufficient statistics

Prof. Michael Levine

January 6, 2015
The idea of data summary

▶ Suppose you have $X_1, \ldots, X_n \sim f(x; \theta)$

▶ What if you were only given $\bar{X}$ and $S^2$ - how much do you know about the data?

▶ What about an arbitrary statistic $Y = u(X_1, \ldots, X_n)$?
The idea

- We are looking for \( Y_1 = u_1(X_1, \ldots, X_n) \) such that, given

\[
(X_1, \ldots, X_n) \in \{(x_1, \ldots, x_n) : u_1(x_1, \ldots, x_n) = y_1\}
\]

the conditional probability of \( X_1, \ldots, X_n \) does not depend on \( \theta \)

- Thus, when \( Y_1 = y_1 \) is fixed, the distribution of any other \( Y_2 = u_2(X_1, \ldots, X_n) \) does not depend on \( \theta \)

- It is impossible to use \( Y_2 \) given \( Y_1 = y_1 \), to make any inference about \( \theta \)... \( Y_1 \) exhausts the information about \( \theta \) that is contained in the sample
Example

- Let $X_1, \ldots, X_n$ be Bernoulli with the parameter $0 < \theta < 1$
- The pmf is $f(x; \theta) = \theta^x(1 - \theta)^{1-x}$ when $x = 0, 1$ and zero otherwise
- $Y_1 = \sum_{i=1}^{n} X_i$ is Bin($n$, $\theta$)
- Its pdf is $f_{Y_1}(y_1; \theta) = \binom{n}{y_1} \theta^{y_1}(1 - \theta)^{n-y_1}$ if $y_1 = 0, 1, \ldots, n$ and zero otherwise
Clearly, $P(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1) = P(A|B)$ is zero if $y_1 \neq \sum_{i=1}^{n} x_i$

Otherwise, $A \subset B$ and $P(A|B) = P(A)/P(B)$

Thus, $P(A|B)$ becomes

$$
\frac{\theta^{x_1}(1 - \theta)^{1-x_1} \cdots \theta^{x_n}(1 - \theta)^{1-x_n}}{inom{n}{y_1} \theta^{y_1}(1 - \theta)^{n-y_1}}
= \frac{1}{\left(\sum_{i=1}^{n} x_i\right)}
$$

Thus, the ratio above (conditional probability) does not depend on $\theta$
More formal definition

- \( X_1, \ldots, X_n \sim f(x; \theta) \) where \( f(x; \theta) \) is either pdf or pmf
- \( Y_1 = u_1(X_1, \ldots, X_n) \) has the distribution \( f_{Y_1}(y_1; \theta) \)
- \( Y_1 \) is a **sufficient statistic** for \( \theta \) iff

\[
\frac{\prod_{i=1}^n f(x_i; \theta)}{f_{Y_1}[u_1(x_1, \ldots, x_n); \theta]} = H(x_1, \ldots, x_n)
\]

does not depend on \( \theta \)

- In the pmf case, this clearly implies that the conditional distribution of \( X_1 = x_1, \ldots, X_n = x_n \) given \( Y_1 = y_1 \) does not depend on \( \theta \)

- In the continuous case, we still use the definition regardless
Remark

- Note that the definition of a sufficient statistic does not require that $X_1, \ldots, X_n$ be independent.

- In general, $Y_1 = u_1(X_1, \ldots, X_n)$ is a sufficient statistic iff

\[
\frac{f(x_1, \ldots, x_n; \theta)}{f_{Y_1}[u_1(x_1, \ldots, x_n); \theta]} = H(x_1, \ldots, x_n)
\]

does not depend upon $\theta$.

- Here, $f(x_1, \ldots, x_n; \theta)$ is simply the joint pdf or pmf of $X_1, \ldots, X_n$. 

Example

- $Y_1 < Y_2 < \ldots < Y_n$ the order statistics of a sample size $n$ from the shifted exponential distribution

\[ f(x; \theta) = e^{-(x-\theta)} I_{(\theta, \infty)}(x) \]

- The pdf of $Y_1 = \min_i X_i$ is

\[ f_{Y_1}(y_1; \theta) = ne^{-n(y_1-\theta)} I_{(\theta, \infty)}(y_1) \]

- The ratio is

\[
\frac{\prod_{i=1}^{n} e^{-(x_i-\theta)} I_{\theta, \infty}(x_i)}{ne^{-n(\min x_i-\theta)} I_{(\theta, \infty)}(\min x_i)}
= \frac{e^{-x_1-x_2-\cdots-x_n}}{ne^{-n \min x_i}}
\]

- Clearly, $Y_1$ is a sufficient statistic for $\theta$
Factorization Theorem - due to J. Neyman

- Let $X_1, \ldots, X_n \sim f(x; \theta)$
- $Y_1 = u_1(X_1, \ldots, X_n)$ is a sufficient statistic for $\theta$ iff
  
  $$f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta) = k_1[u_1(x_1, x_2, \ldots, x_n); \theta]k_2(x_1, \ldots, x_n)$$

  where $k_2(x_1, \ldots, x_n)$ does not depend on $\theta$
Example

- Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, $-\infty < \theta < \infty$, with the known $\sigma^2 > 0$
- Verify that $\sum_{i=1}^{n}(x_i - \theta)^2 = n(\bar{x} - \theta)^2 + \sum_{i=1}^{n}(x_i - \bar{x})^2$
- Factorize the joint pdf to find out that
  \[ f(x_1, \ldots, x_n; \theta) = e^{-n(\bar{x}-\theta)^2/2\sigma^2} \times k_2(x_1, \ldots, x_n) \]
  where $k_2(x_1, \ldots, x_n)$ does not depend on $\theta$
- Thus, $\bar{X}$ is a sufficient statistic for $\theta$
- Note that we could have used the definition of a sufficient statistic directly here since we know that $\bar{X} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$
Now, the definition is not useful at all...let $X_1, \ldots, X_n \sim f(x; \theta)$ with $f(x; \theta) = \theta x^{\theta-1}$ $0 < x < 1$, and zero otherwise, with $\theta > 0$.

The distribution above is a beta distribution with one parameter fixed...

The joint pdf is

$$f(x_1, \ldots, x_n; \theta) = \left[ \theta^n \left( \prod_{i=1}^n x_i \right)^{\theta} \right] \left( \frac{1}{\prod_{i=1}^n x_i} \right)$$

By factorization theorem, $\prod_{i=1}^n X_i$ is a sufficient statistic for $\theta$.