# STA 522, Spring 2022
# Introduction to Theoretical Statistics II

### Lecture 3

Department of Biostatistics
University at Buffalo

# AGENDA

- ▶ Sufficient Statistics

- ▶ Joint Sufficient Statistics

- ▶ Minimal Sufficient Statistics

# Review: Sufficiency

▶ A statistic $T(\underline{x})$ is a **sufficient statistic** for a parameter $\theta$ if the conditional distribution of the sample $\underline{X}$ given that $T(\underline{x}) = t$ does not depend on $\theta$.

▶ **Checking sufficiency:** $T(\underline{x})$ is a sufficient statistic for $\theta$ if, for every $\underline{x}$ in the sample space, the ratio $\dfrac{p(\underline{x} \mid \theta)}{q(T(\underline{x}) \mid \theta)}$ is constant as a function of $\theta$. Here $p(\underline{x} \mid \theta) =$ joint pdf/pmf of $\underline{X}$ and $q(t \mid \theta) =$ pdf/pmf of $T(\underline{x})$.

▶ **Example:** Let $X_1, X_2, \ldots, X_n$ be iid Bernoulli random variables with parameter $\theta$, $0 < \theta < 1$. Then $T(\underline{x}) = \sum_{i=1}^{n} X_i$ is sufficient for $\theta$.

**Example (sufficient order statistics).** Let $X_1, X_2, \ldots, X_n$ be iid from a distribution with pdf $f(x)$, where we are unable to specify any more information about the pdf (as is the case in nonparametric estimation). Then the order statistics are a sufficient statistic.

To verify this, first let $\theta$ be the vector of all parameters in the density $f$ and define $T(\underline{x}) = \left( X_{(1)}, \ldots, X_{(n)} \right)$ where $X_{(i)}$ are the order statistics.

Then $p(\underline{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} f(x_{(i)} \mid \theta)$ and

$q(T(\underline{x}) \mid \theta) = n! \prod_{i=1}^{n} f(x_{(i)} \mid \theta)$, which means $\dfrac{p(\underline{x} \mid \theta)}{q(T(\underline{x}) \mid \theta)} = \dfrac{1}{n!}$.

## Notes
▶ this is not much of a reduction, but we shouldn't expect more with so little information about the density
▶ for specific specific parametric densities substantial reduction is possible.

# How to find sufficient statistics?

- ▶ The previous theorem allows one to check *if* a statistic is sufficient, but doesn't say *how* to find a sufficient statistic (requires guesswork).

- ▶ The following theorem provides *one way* to find a sufficient statistic. The following form is due to Halmos and Savage (1949), but the original idea can be traced back to Neyman and Fisher (1930-39).

## Theorem 6.2.6 (Factorization Theorem)

Let $f(\underline{x} \mid \theta)$ denote the joint pdf or pmf of a sample $\underline{X}$. A statistic $T(\underline{x})$ is a sufficient statistic for $\theta$ *if and only if* there exist functions $g(t \mid \theta)$ and $h(\underline{x})$ such that, for all sample points $\underline{x}$ and all parameter points $\theta$,

$$f(\underline{x} \mid \theta) = g(T(\underline{x}) \mid \theta) \cdot h(\underline{x}).$$

## Notes

1. Thus the Factorization Theorem says that to find a sufficient statistic we first factor $f$ into $g.h$, where (i) $h$ is free of $\theta$, and (ii) $g$ depends on $\theta$ and on $\underline{x}$ only through some function $T(\underline{x})$. This function is a sufficient statistic for $\theta$.
2. $h(\underline{x})$ can be 1 in some situations.

**Example:** Let $X_1, X_2, \ldots, X_n$ be iid Bernoulli random variables with parameter $\theta$, $0 < \theta < 1$. We know that $T(\underline{X}) = \sum_{i=1}^{n} X_i$ is sufficient for $\theta$.

To obtain this through the Factorization Theorem, note that for $x_i \in \{0, 1\}$, $i = 1, \ldots, n$,

$$f(\underline{x} \mid \theta) = \prod_{i=1}^{n} \left\{ \theta^{x_i} (1 - \theta)^{1 - x_i} \right\}$$

$$= \theta^{\sum_{i=1}^{n} x_i} (1 - \theta)^{n - \sum_{i=1}^{n} x_i} = g(T(\underline{x}) \mid \theta) \, h(\underline{x})$$

where $h(\underline{x}) = 1$, $T(\underline{x}) = \sum_{i=1}^{n} x_i$, and $g(t \mid \theta) = \theta^t (1 - \theta)^{n-t}$.

**Example:** Let $X_1, X_2, \ldots, X_n$ be iid Weibull $(\gamma, \beta)$ with common pdf $f(x \mid \beta) = \dfrac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta}$ for $x > 0$ where $\gamma > 0$ is known and $\beta > 0$ is unknown.

To find a sufficient statistic for $\beta$, write for $\underline{x}$ with $x_i > 0$, all $i$,

$$
\begin{aligned}
f(\underline{x} \mid \beta) &= \prod_{i=1}^{n} \left\{ \frac{\gamma}{\beta} \, x_i^{\gamma-1} \, e^{-x_i^\gamma/\beta} \right\} \\
&= \left( \frac{1}{\beta^n} \, e^{-\frac{1}{\beta} \sum_{i=1}^{n} x_i^\gamma} \right) \left( \prod_{i=1}^{n} \gamma \, x_i^{\gamma-1} \right) = g(T(\underline{x}) \mid \beta) \, h(\underline{x})
\end{aligned}
$$

where $h(\underline{x}) = \displaystyle\prod_{i=1}^{n} \gamma \, x_i^{\gamma-1}$, $T(\underline{x}) = \displaystyle\sum_{i=1}^{n} x_i^\gamma$, and $g(t \mid \beta) = \dfrac{1}{\beta^n} \, e^{-t/\beta}$.

Therefore, from the Factorization Theorem it follows that $T(\underline{X}) = \displaystyle\sum_{i=1}^{n} X_i^\gamma$ is sufficient for $\beta$.

## Note

If the support of $f$ involves $\theta$, then we must appropriately define $h$ and $g$ to ensure that the product is 0 where $f$ is 0.

**Example:** Let $X_1, X_2, \ldots, X_n$ be iid Discrete-Uniform $(1, \ldots, \theta)$, where $\theta$ is a positive integer. To find a sufficient statistic for $\theta$, write

$$
\begin{aligned}
f(\underline{x} \mid \theta) &= \prod_{i=1}^{n} f(x_i \mid \theta) \\
&= \prod_{i=1}^{n} \left\{ \frac{1}{\theta} \; I(x_i \in \{1, \ldots, \theta\}) \right\} \\
&= \frac{1}{\theta^n} \; I\left( \max_{1 \leq i \leq n} x_i \leq \theta \right) \\
&= g(T(\underline{x}) \mid \theta) \; h(\underline{x})
\end{aligned}
$$

where $h(\underline{x}) = 1$, $T(\underline{x}) = \max_{1 \leq i \leq n} x_i$, and $g(t \mid \theta) = \frac{1}{\theta^n} I(t \leq \theta)$.

**Example:** Let $X_1, X_2, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables, where $\sigma^2$ is known. We want a sufficient statistic for $\mu$.

The joint pdf of $\underline{X}$ is given by:

$$
\begin{aligned}
f(\underline{x} \mid \mu) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2\right\}\right) \\
&= \underbrace{\exp\left(-\frac{n}{2\sigma^2}(\overline{x} - \mu)^2\right)}_{=g(T(\underline{x})|\mu)} \underbrace{\left[\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\right]}_{=h(\underline{x})}
\end{aligned}
$$

This means that $T(\underline{X}) = \overline{X}$ is sufficient for $\mu$.

## Proof of Theorem 6.2.6 (Factorization Theorem)

We'll prove this only for discrete cases.

**Only if Part:** Let $T(\underline{x})$ be a sufficient statistic for $\theta$. Choose

$$g(t \mid \theta) = P_\theta \left( T(\underline{X}) = t \right) \text{ and } h(\underline{x}) = P \left( \underline{X} = \underline{x} \mid T(\underline{X}) = T(\underline{x}) \right).$$

Since $T(\underline{X})$ is sufficient, the conditional probability $P \left( \underline{X} = \underline{x} \mid T(\underline{X}) = T(\underline{x}) \right)$ doesn't depend on $\theta$ (definition), and so this choice of $h$ is legitimate.

Therefore

$$\begin{aligned}
f(\underline{x} \mid \theta) &= P_\theta(\underline{X} = \underline{x}) \\
&= P_\theta \left( \underline{X} = \underline{x} \text{ and } T(\underline{X}) = T(\underline{x}) \right) \\
&= P_\theta \left( T(\underline{X}) = T(\underline{x}) \right) P \left( \underline{X} = \underline{x} \mid T(\underline{X}) = T(\underline{x}) \right) \\
&= g(T(\underline{x}) \mid \theta) \, h(\underline{x})
\end{aligned}$$

i.e., factorization holds.

**If part:** Suppose factorization holds.

Let $q(t \mid \theta)$ be the pmf of $T(\underline{X})$. Define $A_{T(\underline{x})} = \{\underline{y} : T(\underline{y}) = T(\underline{x})\}$.

Then

$$
\begin{aligned}
\frac{f(\underline{x} \mid \theta)}{q(T(\underline{X}) \mid \theta)} &= \frac{g(T(\underline{x}) \mid \theta)\ h(\underline{x})}{q(T(\underline{X}) \mid \theta)} \\
&= \frac{g(T(\underline{x}) \mid \theta)\ h(\underline{x})}{\sum_{\underline{y} \in A_T} g(T(\underline{y}) \mid \theta)\ h(\underline{y})} \\
&= \frac{g(T(\underline{x}) \mid \theta)\ h(\underline{x})}{g(T(\underline{x}) \mid \theta)\ \sum_{\underline{y} \in A_T} h(\underline{y})} \\
&= \frac{h(\underline{x})}{\sum_{\underline{y} \in A_T} h(\underline{y})}
\end{aligned}
$$

is free of $\theta$.

Therefore, $T(\underline{X})$ is sufficient for $\theta$.

# Joint Sufficiency

**Definition:** Let $X_1, X_2, \ldots, X_n$ be a random sample from the density $f(x \mid \theta)$. The statistics $T_1(\underline{X}), \ldots, T_r(\underline{X})$ are **jointly sufficient for** $\theta$ if and only if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T_1(\underline{X}) = t_1, \ldots, T_r(\underline{X}) = t_r$ does not depend on $\theta$.

## Notes

▶ A set of jointly sufficient statistics may also be referred to as a **vector-valued sufficient statistic**.

▶ The sample itself, $X_1, X_2, \ldots, X_n$, is always jointly sufficient since the conditional distribution of the sample given the sample does not depend on $\theta$. Also, as seen in a previous example, the order statistics are jointly sufficient as well.

▶ The Factorization Theorem can still be used to find jointly sufficient statistics.

**Example:** Let $X_1, X_2, \ldots, X_n$ be iid Uniform $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. We want a sufficient statistic for $\theta$.

$$f(\underline{x} \mid \theta) = \prod_{i=1}^{n} \left\{ I\left(\theta - \frac{1}{2} < x_i < \theta + \frac{1}{2}\right) \right\}$$
$$= I\left(\theta - \frac{1}{2} < x_{(1)} < x_{(n)} < \theta + \frac{1}{2}\right)$$
$$= g(T_1(\underline{x}), T_2(\underline{x}) \mid \theta) \, h(\underline{x})$$

where $T_1(\underline{x}) = x_{(1)}$, $T_2(\underline{x}) = x_{(n)}$, and $h(\underline{x}) = 1$. This shows that $(T_1, T_2) = (X_{(1)}, X_{(n)})$ are jointly sufficient for $\theta$.

### Notes

- It is likely that a set of jointly sufficient statistics $(T_1(\underline{X}), \ldots, T_r(\underline{X}))$ is needed when the parameter is also a vector, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$.
- Usually the sufficient statistic and the parameter vectors are of equal lengths ($r = s$), but different combinations of lengths are possible.

**Example (Contd.):** Suppose that $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ with both $\mu$ and $\sigma^2$ unknown. Then

$$
\begin{aligned}
f(\underline{x} \mid \mu, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2\right\}\right) \\
&= \underbrace{\left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\overline{x} - \mu)^2\right\}\right)}_{=g(T_1(\underline{x}), T_2(\underline{x})|\mu,\sigma^2)} \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n}_{=h(\underline{x})}
\end{aligned}
$$

where $T_1(\underline{x}) = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $T_2(\underline{x}) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. This shows that $(\overline{X}, S^2)$ are jointly sufficient for $(\mu, \sigma^2)$.

**Example (Sufficient Statistic for Exponential Family):** Let $X_1, X_2, \ldots, X_n$ be iid observations from a pdf/pmf $f(x \mid \boldsymbol{\theta})$ that belongs to an exponential family given by

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left( \sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x) \right),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ and $d \leq k$.

Then

$$T(\underline{X}) = \left( \sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j) \right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

**Proof:** Homework (Exercise 6.4).

# Sufficient Statistics are NOT unique

## Theorem (Invariance of Sufficient Statistic)

Suppose $T(\underline{X})$ is sufficient for a parameter $\theta$, and let $u$ be a one-to-one function. Then $T^*(\underline{X}) = u(T(\underline{X}))$ is also sufficient for $\theta$.

**Remarks:**

- Based on the Invariance Principle, we see that one-to-one functions of sufficient statistics are also sufficient.

- Note that the sample itself, i.e., $T(\underline{X}) = \underline{X}$, is a sufficient statistic. So are the order statistics $T'(\underline{X}) = (X_{(1)}, \ldots, X_{(n)})$.

- Question: Is one sufficient statistic better than another?

- Recall that the purpose of sufficient statistics is to achieve data reduction without loss of information about the parameter $\theta$.

- A statistic that achieves the most data reduction while still retaining all the information about $\theta$ might be preferable.

# Minimal Sufficient Statistic

**Definition:** A sufficient statistic $T(\underline{X})$ is called a **minimal sufficient statistic** if, for any other sufficient statistic $T'(\underline{X})$, $T(\underline{x})$ is a function of $T'(\underline{x})$, i.e., whenever $T'(\underline{x}) = T'(\underline{y})$, then $T(\underline{x}) = T(\underline{y})$.

Among all sufficient statistics, a minimal sufficient statistic achieves the greatest possible data reduction.

**Example:** Suppose that $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, with $\sigma$ known.

We have seen that $T(\underline{X}) = \overline{X}$ is sufficient for $\mu$.

We can also show that the statistic $T'(\underline{X}) = (\overline{X}, S^2)$ is sufficient for $\mu$.

Since both $T(\underline{X})$ and $T'(\underline{X})$ are sufficient for $\mu$, they each contain the same information about $\mu$. But clearly $T(\underline{X})$ achieves greater data reduction than $T'(\underline{X})$.

If $\sigma$ were unknown, however, things would be different.

# Results on Minimal Sufficiency

## Theorem 6.2.13 (Checking for Minimal Sufficiency)

Let $f(\underline{x} \mid \theta)$ be the pdf/pmf of a sample $\underline{X}$. Suppose there exists a function $T(\underline{X})$ such that, for every $\underline{x}$ and $\underline{y}$, the ratio

$$\frac{f(\underline{x} \mid \theta)}{f(\underline{y} \mid \theta)}$$

is constant as a function of $\theta$ if and only if $T(\underline{x}) = T(\underline{y})$. Then $T(\underline{X})$ is a minimal sufficient statistic for $\theta$.

## Invariance Theorem (Establishes non uniqueness)

Suppose $T(\mathbf{X})$ is a minimal sufficient statistic for a parameter $\theta$, and let $u$ be a one-to-one function. Then $T^*(\mathbf{X}) = u(T(\mathbf{X}))$ is also a minimal sufficient statistic for $\theta$.

**Example:** $\left( \sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right)$ is a minimal sufficient statistic for $(\mu, \sigma^2)$.

**Example (Contd.):** Let $X_1, X_2, \ldots, X_n$ be iid $N\left(\mu, \sigma^2\right)$, with both $\mu$ and $\sigma^2$ unknown. Given a sample $\underline{X}$ we saw that $(\overline{X}, S^2)$ is sufficient for $(\mu, \sigma^2)$.

To show that $(\overline{X}, S^2)$ is minimal sufficient, consider two sample points $\underline{x}$ and $\underline{y}$ with sample mean and variances $(\overline{x}, s_x^2)$ and $(\overline{x}, s_x^2)$ respectively. Then

$$
\begin{aligned}
\frac{f(\underline{x} \mid \mu, \sigma^2)}{f(\underline{y} \mid \mu, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left(-[n(\overline{x} - \mu)^2 + (n-1)s_x^2]/(2\sigma^2)\right)}{(2\pi\sigma^2)^{-n/2} \exp\left(-[n(\overline{y} - \mu)^2 + (n-1)s_y^2]/(2\sigma^2)\right)} \\
&= \exp\left([-n(\overline{x}^2 - \overline{y}^2) + 2n\mu(\overline{x} - \overline{y}) - (n-1)(s_x^2 - s_y^2)]/(2\sigma^2)\right)
\end{aligned}
$$

is constant as a function of $(\mu, \sigma^2)$ if and only if $\overline{x} = \overline{y}$ and $s_x^2 = s_y^2$.

This shows, from the previous theorem, that $(\overline{X}, S^2)$ is minimal sufficient.

**Example (Contd.):** Let $X_1, X_2, \ldots, X_n$ be iid Uniform $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. We saw that $\left(X_{(1)}, X_{(n)}\right)$ are jointly sufficient for $\theta$. To see if they are minimal sufficient, consider two sample points $\underline{x}$ and $\underline{y}$, and obtain $(x_{(1)}, x_{(n)})$ and $(y_{(1)}, y_{(n)})$. Then

$$
\begin{aligned}
\frac{f(\underline{x} \mid \theta)}{f(\underline{y} \mid \theta)} &= \frac{I\left(\theta - \frac{1}{2} < x_{(1)} < x_{(n)} < \theta + \frac{1}{2}\right)}{I\left(\theta - \frac{1}{2} < y_{(1)} < y_{(n)} < \theta + \frac{1}{2}\right)} \\
&= \frac{I\left(x_{(n)} - \frac{1}{2} < \theta < x_{(1)} + \frac{1}{2}\right)}{I\left(y_{(n)} - \frac{1}{2} < \theta < y_{(1)} + \frac{1}{2}\right)}
\end{aligned}
$$

The numerator and the denominator of this ratio are both one if and only if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$.

Therefore, $\left(X_{(1)}, X_{(n)}\right)$ is minimal sufficient for $\theta$.

**Example:** Let $X_1, X_2, \ldots, X_n$ be iid beta $(\alpha, \beta)$. Show that
$\left( \sum\limits_{i=1}^{n} \log X_i, \sum\limits_{i=1}^{n} \log (1 - X_i) \right)$ is a minimal sufficient statistic for $(\alpha, \beta)$.

First note on the outset that

$$
\begin{aligned}
f(\underline{x} \mid \alpha, \beta) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \prod_{i=1}^{n} \left\{ x_i^{\alpha-1}(1-x_i)^{\beta-1} \right\} \\
&= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \; \exp\left[ (\alpha - 1)\sum_{i=1}^{n} \log x_i + (\beta - 1)\sum_{i=1}^{n} \log(1 - x_i) \right]
\end{aligned}
$$

Now show minimal sufficiency:

$$\frac{f(\underline{x} \mid \alpha, \beta)}{f(\underline{y} \mid \alpha, \beta)} = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \exp\left[(\alpha-1)\sum_{i=1}^{n}\log x_i + (\beta-1)\sum_{i=1}^{n}\log(1-x_i)\right]}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \exp\left[(\alpha-1)\sum_{i=1}^{n}\log y_i + (\beta-1)\sum_{i=1}^{n}\log(1-y_i)\right]}$$

$$= \exp\left[(\alpha-1)\left(\sum_{i=1}^{n}\log x_i - \sum_{i=1}^{n}\log y_i\right)\right.$$

$$\left. + (\beta-1)\left(\sum_{i=1}^{n}\log(1-x_i) - \sum_{i=1}^{n}\log(1-y_i)\right)\right].$$

This is constant in $(\alpha, \beta)$ if and only if $\sum_{i=1}^{n}\log x_i = \sum_{i=1}^{n}\log y_i$ and $\sum_{i=1}^{n}\log(1-x_i) = \sum_{i=1}^{n}\log(1-y_i)$.