Lecture 3. Sufficiency
Sufficient statistics

The concept of sufficiency addresses the question

"Is there a statistic $T(X)$ that in some sense contains all the information about $\theta$ that is in the sample?"

**Example 3.1**

$X_1, \ldots, X_n$ iid Bernoulli($\theta$), so that $P(X_i = 1) = 1 - P(X_i = 0) = \theta$ for some $0 < \theta < 1$.

So $f_X(x \mid \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum x_i} (1 - \theta)^{n-\sum x_i}$.

This depends on the data only through $T(x) = \sum x_i$, the total number of ones. Note that $T(X) \sim \text{Bin}(n, \theta)$.

If $T(x) = t$, then

$$f_{X \mid T=t}(x \mid T=t) = \frac{P_{\theta}(X=x, T=t)}{P_{\theta}(T=t)} = \frac{P_{\theta}(X=x)}{P_{\theta}(T=t)} = \frac{\theta^{\sum x_i} (1 - \theta)^{n-\sum x_i}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} = \left(\frac{n}{t}\right)^{-1},$$

ie the conditional distribution of $X$ given $T=t$ does not depend on $\theta$.

Thus if we know $T$, then additional knowledge of $x$ (knowing the exact sequence of 0's and 1's) does not give extra information about $\theta$. □
Definition 3.1

A statistic $T$ is **sufficient** for $\theta$ if the conditional distribution of $X$ given $T$ does not depend on $\theta$.

Note that $T$ and/or $\theta$ may be vectors. In practice, the following theorem is used to find sufficient statistics.
Theorem 3.2

(The Factorisation criterion) \( T \) is sufficient for \( \theta \) iff \( f_X(x \mid \theta) = g(T(x), \theta)h(x) \) for suitable functions \( g \) and \( h \).

**Proof** (Discrete case only)

Suppose \( f_X(x \mid \theta) = g(T(x), \theta)h(x) \).

If \( T(x) = t \) then

\[
f_X(x \mid T = t) = \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T = t)} = \frac{g(T(x), \theta)h(x)}{\sum_{x' : T(x') = t} g(t, \theta)h(x')}
\]

\[
= \frac{g(t, \theta)h(x)}{g(t, \theta) \sum_{x' : T(x') = t} h(x')} = \frac{h(x)}{\sum_{x' : T(x') = t} h(x')},
\]

which does not depend on \( \theta \), so \( T \) is sufficient.

Now suppose that \( T \) is sufficient so that the conditional distribution of \( X \mid T = t \) does not depend on \( \theta \). Then

\[
\mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(X = x, T(X) = t(x)) = \mathbb{P}_\theta(X = x \mid T = t)\mathbb{P}_\theta(T = t).
\]

The first factor does not depend on \( \theta \) by assumption; call it \( h(x) \). Let the second factor be \( g(t, \theta) \), and so we have the required factorisation. \( \square \)
Example 3.1 continued

For Bernoulli trials, \( f_X(x | \theta) = \theta \sum x_i (1 - \theta)^{n - \sum x_i} \).

Take \( g(t, \theta) = \theta^t (1 - \theta)^{n - t} \) and \( h(x) = 1 \) to see that \( T(X) = \sum X_i \) is sufficient for \( \theta \). □

Example 3.2

Let \( X_1, \ldots, X_n \) be iid \( U[0, \theta] \).

Write \( 1_A(x) \) for the indicator function, = 1 if \( x \in A \), = 0 otherwise.

We have

\[
fx(x | \theta) = \prod_{i=1}^{n} \frac{1}{\theta} 1_{[0,\theta]}(x_i) = \frac{1}{\theta^n} 1_{\{\max_i x_i \leq \theta\}}(\max_i x_i) 1_{\{0 \leq \min_i x_i\}}(\min_i x_i).
\]

Then \( T(X) = \max_i X_i \) is sufficient for \( \theta \). □
Sufficient statistics are not unique. If $T$ is sufficient for $\theta$, then so is any (1-1) function of $T$.

$X$ itself is always sufficient for $\theta$; take $T(X) = X$, $g(t, \theta) = f_X(t | \theta)$ and $h(x) = 1$. But this is not much use.

The sample space $\mathcal{X}^n$ is partitioned by $T$ into sets $\{x \in \mathcal{X}^n : T(x) = t\}$.

If $T$ is sufficient, then this data reduction does not lose any information on $\theta$.

We seek a sufficient statistic that achieves the maximum-possible reduction.

**Definition 3.3**

A sufficient statistic $T(X)$ is *minimal sufficient* if it is a function of every other sufficient statistic:

i.e. if $T'(X)$ is also sufficient, then $T'(X) = T'(Y) \rightarrow T(X) = T(Y)$

i.e. the partition for $T$ is coarser than that for $T'$. 
Minimal sufficient statistics can be found using the following theorem.

**Theorem 3.4**

Suppose $T = T(X)$ is a statistic such that $f_X(x; \theta)/f_X(y; \theta)$ is constant as a function of $\theta$ if and only if $T(x) = T(y)$. Then $T$ is minimal sufficient for $\theta$.

**Sketch of proof**: Non-examinable

First, we aim to use the Factorisation Criterion to show sufficiency. Define an equivalence relation $\sim$ on $\mathcal{X}^n$ by setting $x \sim y$ when $T(x) = T(y)$. (Check that this is indeed an equivalence relation.) Let $\mathcal{U} = \{T(x) : x \in \mathcal{X}^n\}$, and for each $u$ in $\mathcal{U}$, choose a representative $x_u$ from the equivalence class $\{x : T(x) = u\}$. Let $x$ be in $\mathcal{X}^n$ and suppose that $T(x) = t$. Then $x$ is in the equivalence class $\{x' : T(x') = t\}$, which has representative $x_t$, and this representative may also be written $x_{T(x)}$. We have $x \sim x_t$, so that $T(x) = T(x_t)$, ie $T(x) = T(x_{T(x)})$. Hence, by hypothesis, the ratio $\frac{f_X(x; \theta)}{f_X(x_{T(x)}; \theta)}$ does not depend on $\theta$, so let this be $h(x)$. Let $g(t, \theta) = f_X(x_t, \theta)$. Then

$$f_X(x; \theta) = f_X(x_{T(x)}; \theta) \frac{f_X(x; \theta)}{f_X(x_{T(x)}; \theta)} = g(T(x), \theta)h(x),$$

and so $T = T(X)$ is sufficient for $\theta$ by the Factorisation Criterion.
Next we aim to show that $T(\mathbf{X})$ is a function of every other sufficient statistic. Suppose that $S(\mathbf{X})$ is also sufficient for $\theta$, so that, by the Factorisation Criterion, there exist functions $g_S$ and $h_S$ (we call them $g_S$ and $h_S$ to show that they belong to $S$ and to distinguish them from $g$ and $h$ above) such that

$$f_X(x; \theta) = g_S(S(x), \theta) h_S(x).$$

Suppose that $S(x) = S(y)$. Then

$$\frac{f_X(x; \theta)}{f_X(y; \theta)} = \frac{g_S(S(x), \theta) h_S(x)}{g_S(S(y), \theta) h_S(y)} = \frac{h_S(x)}{h_S(y)},$$

because $S(x) = S(y)$. This means that the ratio $\frac{f_X(x; \theta)}{f_X(y; \theta)}$ does not depend on $\theta$, and this implies that $T(x) = T(y)$ by hypothesis. So we have shown that $S(x) = S(y)$ implies that $T(x) = T(y)$, i.e. $T$ is a function of $S$. Hence $T$ is minimal sufficient. $\square$
Example 3.3

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$.

Then

$$
\frac{f_X(x \mid \mu, \sigma^2)}{f_X(y \mid \mu, \sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \right\}}{(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right\}}
$$

$$
= \exp \left\{ -\frac{1}{2\sigma^2} \left( \sum_i x_i^2 - \sum_i y_i^2 \right) + \frac{\mu}{\sigma^2} \left( \sum_i x_i - \sum_i y_i \right) \right\}.
$$

This is constant as a function of $(\mu, \sigma^2)$ iff $\sum_i x_i^2 = \sum_i y_i^2$ and $\sum_i x_i = \sum_i y_i$. So $T(X) = (\sum_i X_i^2, \sum_i X_i)$ is minimal sufficient for $(\mu, \sigma^2)$. □

1-1 functions of minimal sufficient statistics are also minimal sufficient.

So $T'(X) = (\bar{X}, \sum(X_i - \bar{X})^2)$ is also sufficient for $(\mu, \sigma^2)$, where $\bar{X} = \sum_i X_i/n$. We write $S_{XX}$ for $\sum(X_i - \bar{X})^2$. 

Lecture 3. Sufficiency
Notes

- Example 3.3 has a vector $T$ sufficient for a vector $\theta$. Dimensions do not have to the same: e.g. for $N(\mu, \mu^2)$, $T(X) = (\sum_i X_i^2, \sum_i X_i)$ is minimal sufficient for $\mu$ [check]

- If the range of $X$ depends on $\theta$, then "$f_X(x; \theta) / f_X(y; \theta)$ is constant in $\theta$" means "$f_X(x; \theta) = c(x, y) f_X(y; \theta)$"
The Rao–Blackwell Theorem

The Rao–Blackwell theorem gives a way to improve estimators in the mse sense.

**Theorem 3.5**

(The Rao–Blackwell theorem) Let $T$ be a sufficient statistic for $\theta$ and let $\tilde{\theta}$ be an estimator for $\theta$ with $\mathbb{E}(\tilde{\theta}^2) < \infty$ for all $\theta$. Let $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T]$. Then for all $\theta$,

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\tilde{\theta} - \theta)^2].$$

The inequality is strict unless $\tilde{\theta}$ is a function of $T$.

**Proof** By the conditional expectation formula we have $\mathbb{E}\hat{\theta} = \mathbb{E}[\mathbb{E}(\tilde{\theta} \mid T)] = \mathbb{E}\tilde{\theta}$, so $\hat{\theta}$ and $\tilde{\theta}$ have the same bias. By the conditional variance formula,

$$\text{var}(\tilde{\theta}) = \mathbb{E}[\text{var}(\tilde{\theta} \mid T)] + \text{var}\left[\mathbb{E}(\tilde{\theta} \mid T)\right] = \mathbb{E}[\text{var}(\tilde{\theta} \mid T)] + \text{var}(\hat{\theta}).$$

Hence $\text{var}(\tilde{\theta}) \geq \text{var}(\hat{\theta})$, and so $\text{mse}(\tilde{\theta}) \geq \text{mse}(\hat{\theta})$, with equality only if $\text{var}(\tilde{\theta} \mid T) = 0$. □
Notes

(i) Since $T$ is sufficient for $\theta$, the conditional distribution of $X$ given $T = t$ does not depend on $\theta$. Hence $\hat{\theta} = \mathbb{E}[\tilde{\theta}(X) \mid T]$ does not depend on $\theta$, and so is a bona fide estimator.

(ii) The theorem says that given any estimator, we can find one that is a function of a sufficient statistic that is at least as good in terms of mean squared error of estimation.

(iii) If $\tilde{\theta}$ is unbiased, then so is $\hat{\theta}$.

(iv) If $\tilde{\theta}$ is already a function of $T$, then $\hat{\theta} = \tilde{\theta}$. 
Example 3.4

Suppose $X_1, \ldots, X_n$ are iid Poisson($\lambda$), and let $\theta = e^{-\lambda}$ ($= P(X_1 = 0)$).

Then $p_X(x | \lambda) = (e^{-n\lambda} \lambda \sum x_i) / \prod x_i!$, so that $p_X(x | \theta) = (\theta^n (- \log \theta) \sum x_i) / \prod x_i!$.

We see that $T = \sum X_i$ is sufficient for $\theta$, and $\sum X_i \sim$ Poisson($n\lambda$).

An easy estimator of $\theta$ is $\tilde{\theta} = 1_{[X_1 = 0]}$ (unbiased) [i.e. if do not observe any events in first observation period, assume the event is impossible!]

Then

$$\mathbb{E}[\tilde{\theta} | T = t] = P(X_1 = 0 | \sum_{1}^{n} X_i = t)$$

$$= \frac{P(X_1 = 0)P(\sum_{2}^{n} X_i = t)}{P(\sum_{1}^{n} X_i = t)} \left(\frac{n - 1}{n}\right)^t \text{ (check).}$$

So $\hat{\theta} = (1 - \frac{1}{n}) \sum X_i$. □

[Common sense check: $\hat{\theta} = (1 - \frac{1}{n})^n \bar{X} \approx e^{-\bar{X}} = e^{-\hat{\lambda}}$]
Example 3.5
Let $X_1, \ldots, X_n$ be iid $U[0, \theta]$, and suppose that we want to estimate $\theta$. From Example 3.2, $T = \max X_i$ is sufficient for $\theta$. Let $\tilde{\theta} = 2X_1$, an unbiased estimator for $\theta$ [check].

Then

$$E[\tilde{\theta} \mid T = t] = 2 \mathbb{E}[X_1 \mid \max X_i = t]$$

$$= 2(\mathbb{E}[X_1 \mid \max X_i = t, X_1 = \max X_i] \mathbb{P}(X_1 = \max X_i)$$

$$+ \mathbb{E}[X_1 \mid \max X_i = t, X_1 \neq \max X_i] \mathbb{P}(X_1 \neq \max X_i))$$

$$= 2(t \times \frac{1}{n} + \frac{t}{2} \frac{n - 1}{n}) = \frac{n + 1}{n} t,$$

so that $\hat{\theta} = \frac{n + 1}{n} \max X_i$. □

In Lecture 4 we show directly that this is unbiased.

N.B. Why is $\mathbb{E}[X_1 \mid \max X_i = t, X_1 \neq \max X_i] = t/2$?

Because

$$f_{X_1}(x_1 \mid X_1 < t) = \frac{f_{X_1}(x_1, X_1 < t)}{\mathbb{P}(X_1 < t)} = \frac{f_{X_1}(x_1)1_{[0 \leq x_1 < t]}}{t/\theta} = \frac{1/\theta \times 1_{[0 \leq x_1 < t]}}{t/\theta} = \frac{1}{t} 1_{[0 \leq x_1 < t]},$$

and so $X_1 \mid X_1 < t \sim U[0, t]$. 