3. Sufficiency

3.1. Sufficient statistics

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

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

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 $\theta$. □

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

$$f_X(x \mid T = t) = \frac{P_\theta(X = x, T(X) = t)}{P_\theta(T = t)} = \frac{g(T(x), \theta)h(x)}{g(t, \theta) h(x)} = \frac{g(t, \theta)}{g(t, \theta) \sum_{x' : T(x') = t} h(x')} h(x),$$

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

$$P_\theta(X = x) = P_\theta(X = x, T(X) = t(x)) = P_\theta(X = x \mid T = t) P_\theta(T = t).$$

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. □
Example 3.1 continued
For Bernoulli trials, $f_X(x \mid \theta) = \theta^{x} (1 - \theta)^{n-x}$.
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

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

Then $T(X) = \max_i X_i$ is sufficient for $\theta$. □

Minimal sufficient statistics

3.2. Minimal sufficient statistics

Minimal sufficient statistics can be found using the following theorem.

**Theorem 3.4**
Suppose $T = T(X)$ is a statistic such that $f_X(x \mid \theta)/f_X(y; \theta)$ 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 $T(x) = T(x_t)$, i.e. $T(x) = T(x_{T(x)})$. Hence, by hypothesis, the ratio $f_{X}(x_{T(x)}; \theta) / f_{X}(x; \theta)$ does not depend on $\theta$, so let this be $h(x)$. Let $g(t, \theta) = f_X(x_t \mid \theta)$. Then

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

and so $T = T(X)$ is sufficient for $\theta$ by the Factorisation Criterion.

Next we aim to show that $T(X)$ is a function of every other sufficient statistic.
Suppose that $S(X)$ 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

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

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

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

because $S(x) = S(y)$. This means that the ratio $f_X(x \mid \theta) / f_X(y \mid \theta)$ 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. □
Example 3.3
Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$.
Then
\[
\frac{f_X(x | \mu, \sigma^2)}{f_X(y | \mu, \sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \right\}}{(2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right\}}
= \exp \left\{ -\frac{1}{2\sigma^2} \left( \sum_i x_i^2 - \sum_i y_i^2 \right) + \frac{\mu}{\sigma^2} \left( \sum_i x_i - \sum_i y_i \right) \right\}.
\]
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 $T'(X) = (\bar{X}, (\sum_i (X_i - \bar{X})^2)$ is also sufficient for $(\mu, \sigma^2)$, where $\bar{X} = \sum_i X_i/n$.
We write $S_{XX}$ for $\sum (X_i - \bar{X})^2$.

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 $\hat{\theta}$ be an estimator for $\theta$ with $E(\hat{\theta}^2) < \infty$ for all $\theta$. Let $\hat{\theta} = E[\hat{\theta} | T]$. Then for all $\theta$,
\[
E[(\hat{\theta} - \theta)^2] \leq E[(\hat{\theta} - \theta)^2].
\]
The inequality is strict unless $\hat{\theta}$ is a function of $T$.

Proof By the conditional expectation formula we have $E\hat{\theta} = E[E(\hat{\theta} | T)] = E\hat{\theta}$, so $\hat{\theta}$ and $\theta$ have the same bias. By the conditional variance formula,
\[
\text{var}(\hat{\theta}) = E[\text{var}(\hat{\theta} | T)] + E[\text{var}(\hat{\theta} | T)] = E[\text{var}(\hat{\theta} | T)] + \text{var}(\hat{\theta}).
\]
Hence $\text{var}(\hat{\theta}) \geq \text{var}(\hat{\theta})$, and so $\text{mse}(\hat{\theta}) \geq \text{mse}(\hat{\theta})$, with equality only if $\text{var}(\hat{\theta} | T) = 0$. □
Example 3.4
Suppose \( X_1, \ldots, X_n \) are iid Poisson(\( \lambda \)), and let \( \theta = e^{-\lambda} \) (\( = P(X_1 = 0) \)).

Then \( p_X(x | \lambda) = (e^{-n\lambda}\lambda^x) / \prod x_i! \), 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 \( \sum X_i \sim \text{Poisson}(n\lambda) \).

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

Then
\[
E[\hat{\theta} | T = t] = \frac{P(X_1 = 0) \prod \sum_{i=1}^n X_i = t}{\prod \sum_{i=1}^n X_i = t} \left( \frac{n - 1}{n} \right)^t \text{ (check).}
\]

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

[Common sense check: \( \hat{\theta} = (1 - \frac{1}{n}) nX \approx e^{-X} = e^{-\lambda} \)]

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
\[
E[\tilde{\theta} | T = t] = 2E[X_1 | \max X_i = t] = 2\left( E[X_1 | \max X_i = t, X_1 = \max X_i] P(X_1 = \max X_i) + E[X_1 | \max X_i = t, X_1 \neq \max X_i] P(X_1 \neq \max X_i) \right) = 2 \left( \frac{1}{n} + \frac{t}{n} \right) = \frac{n + 1}{n} t,
\]

so that \( \hat{\theta} = \frac{n + 1}{n} \max X_i \). \( \Box \)

In Lecture 4 we show directly that this is unbiased.

N.B. Why is \( E[X_1 | \max X_i = t, X_1 \neq \max X_i] = t/2? \)

Because
\[
f_{X_1}(x_1 | X_1 < t) = \frac{f_{X_1}(x_1, X_1 < t)}{f_{X_1}(X_1 < t)} = \frac{\frac{1}{\theta} 1_{[0 \leq X_1 < t]}}{t/\theta} = \frac{1}{t} \frac{1}{\theta} 1_{[0 \leq X_1 < t]}, \text{ and so } X_1 | X_1 < t \sim U[0, t].
\]