Lecture 3. Sufficiency

Sufficient statistics

The concept of sufficiency addresses the question

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

Example 3.1

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

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

This depends on the data only through $T(\mathbf{x}) = \sum x_i$, the total number of ones.

Note that $T(\mathbf{X}) \sim \mathrm{Bin}(n, \theta)$.

If $T(\mathbf{x}) = t$, then

$$f_{\mathbf{X}|T=t}(\mathbf{x} \mid T = t) = \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T = t)}{\mathbb{P}_\theta(T = t)} = \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x})}{\mathbb{P}_\theta(T = t)} = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1},$$

ie the conditional distribution of $\mathbf{X}$ given $T = t$ does not depend on θ.

Thus if we know T, then additional knowledge of $\mathbf{x}$ (knowing the exact sequence of 0's and 1's) does not give extra information about θ. $\square$

Definition 3.1

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

Note that T and/or θ may be vectors. In practice, the following theorem is used to find sufficient statistics.

Theorem 3.2

(The Factorisation criterion) T is sufficient for θ iff $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$ for suitable functions g and h.

Proof (Discrete case only)

Suppose $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$.

If $T(\mathbf{x}) = t$ then

$$\begin{aligned} f_{\mathbf{X}|T=t}(\mathbf{x} \mid T = t) &= \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{\mathbb{P}_\theta(T = t)} = \frac{g(T(\mathbf{x}), \theta)h(\mathbf{x})}{\sum_{\{\mathbf{x}' : T(\mathbf{x}') = t\}} g(t, \theta)h(\mathbf{x}')} \\ &= \frac{g(t, \theta)h(\mathbf{x})}{g(t, \theta)\sum_{\{\mathbf{x}' : T(\mathbf{x}') = t\}} h(\mathbf{x}')} = \frac{h(\mathbf{x})}{\sum_{\{\mathbf{x}' : T(\mathbf{x}') = t\}} h(\mathbf{x}')}, \end{aligned}$$

which does not depend on θ, so T is sufficient.

Now suppose that T is sufficient so that the conditional distribution of $\mathbf{X} \mid T = t$ does not depend on θ. Then

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

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

Example 3.1 continued

For Bernoulli trials, $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$.

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

Example 3.2

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

Write $1_{[A]}$ for the indicator function of A.

We have

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} 1_{[0 \le x_i \le \theta]} = \frac{1}{\theta^n} 1_{[\max_i x_i \le \theta]} 1_{[\min_i x_i \ge 0]}.$$

Then $T(\mathbf{X}) = \max_i X_i$ is sufficient for θ. □

Minimal sufficient statistics

Sufficient statistics are not unique. If T is sufficient for θ, then so is any (1-1) function of T.

$\mathbf{X}$ itself is always sufficient for θ; take $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, $g(\mathbf{t},\theta) = f_{\mathbf{X}}(\mathbf{t} \mid \theta)$ and $h(\mathbf{x}) = 1$. But this is not much use.

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

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

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

Definition 3.3

A sufficient statistic $T(\mathbf{X})$ is *minimal sufficient* if it is a function of every other sufficient statistic:
i.e. if $T'(\mathbf{X})$ is also sufficient, then $T'(\mathbf{X}) = T'(\mathbf{Y}) \to T(\mathbf{X}) = T(\mathbf{Y})$
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 $T = T(\mathbf{X})$ is a statistic such that $f_{\mathbf{X}}(\mathbf{x};\theta)/f_{\mathbf{X}}(\mathbf{y};\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then T is minimal sufficient for θ.

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 $\mathbf{x} \sim \mathbf{y}$ when $T(\mathbf{x}) = T(\mathbf{y})$. (Check that this is indeed an equivalence relation.) Let $\mathcal{U} = \{T(\mathbf{x}) : \mathbf{x} \in \mathcal{X}^n\}$, and for each u in $\mathcal{U}$, choose a representative $\mathbf{x}_u$ from the equivalence class $\{\mathbf{x} : T(\mathbf{x}) = u\}$. Let $\mathbf{x}$ be in $\mathcal{X}^n$ and suppose that $T(\mathbf{x}) = t$. Then $\mathbf{x}$ is in the equivalence class $\{\mathbf{x}' : T(\mathbf{x}') = t\}$, which has representative $\mathbf{x}_t$, and this representative may also be written $\mathbf{x}_{T(\mathbf{x})}$. We have $\mathbf{x} \sim \mathbf{x}_t$, so that $T(\mathbf{x}) = T(\mathbf{x}_t)$, ie $T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$. Hence, by hypothesis, the ratio $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_{T(\mathbf{x})};\theta)}$ does not depend on θ, so let this be $h(\mathbf{x})$. Let $g(t,\theta) = f_{\mathbf{X}}(\mathbf{x}_t,\theta)$. Then

$$f_{\mathbf{X}}(\mathbf{x};\theta) = f_{\mathbf{X}}(\mathbf{x}_{T(\mathbf{x})};\theta)\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_{T(\mathbf{x})};\theta)} = g(T(\mathbf{x}),\theta)h(\mathbf{x}),$$

and so $T = T(\mathbf{X})$ is sufficient for θ 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 θ, 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

$$f_{\mathbf{X}}(\mathbf{x};\theta) = g_S(S(\mathbf{x}),\theta)h_S(\mathbf{x}).$$

Suppose that $S(\mathbf{x}) = S(\mathbf{y})$. Then

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

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

Example 3.3

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

Then

$$\begin{aligned}\frac{f_{\mathbf{X}}(\mathbf{x} \mid \mu, \sigma^2)}{f_{\mathbf{X}}(\mathbf{y} \mid \mu, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right\}}{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_i x_i - \sum_i y_i\right)\right\}.\end{aligned}$$

This is constant as a function of (μ, σ^2) iff $\sum_i x_i^2 = \sum_i y_i^2$ and $\sum_i x_i = \sum_i y_i$.
So $T(\mathbf{X}) = \left(\sum_i X_i^2, \sum_i X_i\right)$ is minimal sufficient for (μ, σ^2). $\square$

1-1 functions of minimal sufficient statistics are also minimal sufficient.
So $\mathbf{T}'(\mathbf{X}) == (\bar{X}, \sum(X_i - \bar{X})^2)$ is also sufficient for (μ, σ^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 θ. Dimensions do not have to the same: e.g. for $N(\mu, \mu^2)$, $T(\mathbf{X}) = \left(\sum_i X_i^2, \sum_i X_i\right)$ is minimal sufficient for μ [check]
- If the range of X depends on θ, then "$f_{\mathbf{X}}(\mathbf{x}; \theta)/f_{\mathbf{X}}(\mathbf{y}; \theta)$ is constant in θ" means "$f_{\mathbf{X}}(\mathbf{x}; \theta) = c(\mathbf{x}, \mathbf{y})\, f_{\mathbf{X}}(\mathbf{y}; \theta)$"

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 θ and let $\tilde{\theta}$ be an estimator for θ with $\mathbb{E}(\tilde{\theta}^2) < \infty$ for all θ. Let $\hat{\theta} = \mathbb{E}\left[\tilde{\theta} \mid T\right]$. Then for all θ,

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

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

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

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

Hence $\mathrm{var}(\tilde{\theta}) \ge \mathrm{var}(\hat{\theta})$, and so $\mathrm{mse}(\tilde{\theta}) \ge \mathrm{mse}(\hat{\theta})$, with equality only if $\mathrm{var}(\tilde{\theta} \mid T) = 0$. $\square$

Notes

(i) Since T is sufficient for θ, the conditional distribution of $\mathbf{X}$ given $T = t$ does not depend on θ. Hence $\hat{\theta} = \mathbb{E}\left[\tilde{\theta}(\mathbf{X}) \mid T\right]$ does not depend on θ, 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(λ), and let $\theta = e^{-\lambda}$ ($= \mathbb{P}(X_1 = 0)$).

Then $p_{\mathbf{X}}(\mathbf{x} \mid \lambda) = \left(e^{-n\lambda}\lambda^{\sum x_i}\right) / \prod x_i!$, so that $p_{\mathbf{X}}(\mathbf{x} \mid \theta) = \left(\theta^n(-\log\theta)^{\sum x_i}\right) / \prod x_i!$.

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

An easy estimator of θ 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

$$\begin{aligned}\mathbb{E}\big[\tilde{\theta} \mid T = t\big] &= \mathbb{P}\Big(X_1 = 0 \mid \sum_1^n X_i = t\Big) \\ &= \frac{\mathbb{P}(X_1=0)\mathbb{P}\big(\sum_2^n X_i = t\big)}{\mathbb{P}\big(\sum_1^n X_i = t\big)} \left(\frac{n-1}{n}\right)^t \text{ (check).}\end{aligned}$$

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

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

Example 3.5

Let $X_1, \ldots, X_n$ be iid $U[0, \theta]$, and suppose that we want to estimate θ. From Example 3.2, $T = \max X_i$ is sufficient for θ. Let $\tilde{\theta} = 2X_1$, an unbiased estimator for θ [check].

Then

$$\begin{aligned}\mathbb{E}\big[\tilde{\theta} \mid T = t\big] &= 2\mathbb{E}\big[X_1 \mid \max X_i = t\big] \\ &= 2\Big(\mathbb{E}\big[X_1 \mid \max X_i = t, X_1 = \max X_i\big]\mathbb{P}(X_1 = \max X_i) \\ &\quad + \mathbb{E}\big[X_1 \mid \max X_i = t, X_1 \neq \max X_i\big]\mathbb{P}(X_1 \neq \max X_i)\Big) \\ &= 2\Big(t \times \frac{1}{n} + \frac{t}{2}\frac{n-1}{n}\Big) = \frac{n+1}{n}t,\end{aligned}$$

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

In Lecture 4 we show directly that this is unbiased.

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

Because

$f_{X_1}(x_1 \mid X_1 < t) = \frac{f_{X_1}(x_1, X_1 < t)}{\mathbb{P}(X_1 < t)} = \frac{f_{X_1}(x_1)1_{[0 \le X_1 < t]}}{t/\theta} = \frac{1/\theta \times 1_{[0 \le X_1 < t]}}{t/\theta} = \frac{1}{t}1_{[0 \le X_1 < t]}$, and so $X_1 \mid X_1 < t \sim U[0, t]$.