Chapter 7. Statistical Estimation

7.8: Properties of Estimators III

Slides (Google Drive) Alex Tsun Video (YouTube)

The final property of estimators we will discuss is called sufficiency. Just like we want our estimators to be consistent and efficient, we also want them to be sufficient.

7.8.1 Sufficiency

We first must define what a statistic is.

Definition 7.8.1: Statistic

A **statistic** is any function $T : \mathbb{R}^n \to \mathbb{R}$ of samples $\mathbf{x} = (x_1, \dots, x_n)$. Examples include:

- $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ (the sum)
- $T(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$ (the mean)
- $T(x_1, \dots, x_n) = \max\{x_1, \dots, x_n\}$ (the max/largest value)
- $T(x_1, \dots, x_n) = x_1$ (just take the first sample)
- $T(x_1, \dots, x_n) = 7$ (ignore all samples)

All estimators are statistics because they take in our n data points and produce a single number. We'll see an example which intuitively explains what it means for a statistic to be sufficient.

Suppose we have iid samples $\mathbf{x} = (x_1, \dots, x_n)$ from a known distribution with unknown parameter θ. Imagine we have two people:

- **Statistician A:** Knows the entire sample, gets n quantities: $\mathbf{x} = (x_1, \dots, x_n)$.
- **Statistician B:** Knows $T(x_1, \dots, x_n) = t$, a single number which is a function of the samples. For example, the sum or the maximum of the samples.

Heuristically, $T(x_1, \dots, x_n)$ is a sufficient statistic if Statistician B can do just as good a job as Statistician A, given "less information". For example, if the samples are from the Bernoulli distribution, knowing $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ (the number of heads) is just as good as knowing all the individual outcomes, since a good estimate would be the number of heads over the number of total trials! Hence, we don't actually care the ORDER of the outcomes, just how many heads occurred! The word "sufficient" in English roughly means "enough", and so this terminology was well-chosen.

Now for the formal definition:

Definition 7.8.2: Sufficient Statistic

A statistic $T = T(X_1, \ldots, X_n)$ is a **sufficient statistic** if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ and θ does not depend on θ.

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$$

(if $X_1, \ldots, X_n$ are continuous rather than discrete, replace the probability with a density).

To motivate the definition, we'll go back to the previous example. Again, statistician A has all the samples $x_1, \ldots, x_n$ but statistician B only has the single number $t = T(x_1, \ldots, x_n)$. The idea is, Statistician B only knows $T = t$, but since T is sufficient, doesn't need θ to generate new samples $X_1^{'}, \ldots, X_n^{'}$ from the distribution. This is because $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$ and since she knows $T = t$, she knows the conditional distribution (can generate samples)! Now Statistician B has n iid samples from the distribution, just like Statistician A. So using these samples $X_1', \ldots, X_n'$, statistician B can do just a good a job as statistician A with samples $X_1, \ldots, X_n$ (on average). So no one is at any disadvantage. :)

This definition is hard to check, but it turns out that there is a criterion that helps us determine whether a statistic is sufficient:

Theorem 7.8.1: Neyman-Fisher Factorization Criterion

Let $x_1, \ldots, x_n$ be iid random samples with likelihood $L(x_1, \ldots, x_n \mid \theta)$. A statistic $T = T(x_1, \ldots, x_n)$ is sufficient if and only if there exist non-negative functions g and h such that:

$$L(x_1, \ldots, x_n \mid \theta) = g(x_1, \ldots, x_n) \cdot h(T(x_1, \ldots, x_n), \theta)$$

That is, the likelihood of the data can be split into a product of two terms: the first term g can depend on the entire data, but not θ, and the second term h *can* depend on θ, but **only on the data through the sufficient statistic** T. (In other words, T is the only thing that allows the data $x_1, \ldots, x_n$ and θ to interact!) That is, we don't have access to the n individual quantities $x_1, \ldots, x_n$; just the single number (T, the sufficient statistic).

If you are reading this for the first time, you might not think this is any better...You may be very confused right now, but let's see some examples to clear things up!

But basically, you want to split the likelihood into a product of two terms/functions:

1. For the first term g, you are allowed to know each individual sample if you want, but NOT θ.
2. For the second term h, you can only know the sufficient statistic (single number) $T(x_1, \ldots, x_n)$ and θ. You may not know each individual x_i.

Example(s)

Let $x_1, \ldots, x_n$ be iid random samples from $\text{Unif}(0, \theta)$ (continuous). Show that the MLE $\hat{\theta} = T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ is a sufficient statistic. (The reason this is true is because we don't need to know each individual sample to have a good estimate for θ; we just need to know the largest!)

Solution We saw the likelihood of this continuous uniform in 7.2, which we'll just rewrite:

$$L(x_1,\dots,x_n \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{\{x_i \le \theta\}} = \frac{1}{\theta^n} I_{\{x_1,\dots,x_n \le \theta\}} = \frac{1}{\theta^n} I_{\{\max\{x_1,\dots,x_n\} \le \theta\}} = \frac{1}{\theta^n} I_{\{T(x_1,\dots,x_n) \le \theta\}}$$

Choose

$$g(x_1,\dots,x_n) = 1$$

and

$$h(T(x_1,\dots,x_n),\theta) = \frac{1}{\theta^n} I_{\{T(x_1,\dots,x_n) \le \theta\}}$$

By the Neyman-Fisher Factorization Criterion, $\hat{\theta}_{MLE} = T = \max\{x_1,\dots,x_n\}$ is sufficient. This is a good property of an estimator!

Notice there is no need for a g term (that's why it is $= 1$), because there is no term in the likelihood which just has the data (without θ).

For the h term, notice that we just need to know the max of the samples $T(x_1,\dots,x_n)$ to compute h: we don't actually need to know each individual x_i.

Notice that here the only interaction between the data and parameter θ happens through the sufficient statistic (the max of all the values). □

Example(s)

Let $x_1,\dots,x_n$ be iid random samples from Poi(θ). Show that $T(x_1,\dots,x_n) = \sum_{i=1}^n x_i$ is a sufficient statistic, and hence the MLE $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ is sufficient as well. (The reason this is true is because we don't need to know each individual sample to have a good estimate for θ; we just need to know how many events happened total!)

Solution We take our Poisson likelihood and split it into smaller terms:

$$L(x_1,\dots,x_n \mid \theta) = \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!} = \left(\prod_{i=1}^{n} e^{-\theta}\right)\left(\prod_{i=1}^{n} \theta^{x_i}\right)\left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

$$= \frac{1}{\prod_{i=1}^n x_i!} \cdot e^{-n\theta}\theta^{T(x_1,\dots,x_n)}$$

Choose

$$g(x_1,\dots,x_n) = \frac{1}{\prod_{i=1}^n x_i!}$$

and

$$h(T(x_1,\dots,x_n),\theta) = e^{-n\theta}\ \theta^{T(x_1,\dots,x_n)}$$

By the Neyman-Fisher Factorization Criterion, $T(x_1,\dots,x_n) = \sum_{i=1}^n x_i$ is sufficient. The mean $\hat{\theta}_{MLE} = \dfrac{\sum_{i=1}^n x_i}{n} = \dfrac{T(x_1,\dots,x_n)}{n}$ is as well, since knowing the total number of events and the average number of events is equivalent (since we know n)!

Notice here we had the g term handle some function of only $x_1, \dots, x_n$ but not θ.

For the h term though, we do have θ but don't need the individual samples $x_1, \dots, x_n$ to compute h. Imagine being just given $T(x_1, \dots, x_n)$: now you have enough information to compute h!

Notice that here the only interaction between the data and parameter θ happens through the sufficient statistic (the sum/mean of all the values). We don't actually need to know each individual x_i. □

Example(s)

Let $x_1, \dots, x_n$ be iid random samples from $\text{Ber}(\theta)$. Show that $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ is a sufficient statistic, and hence the MLE $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ is sufficient as well. (The reason this is true is because we don't need to know each individual sample to have a good estimate for θ; we just need to know how many heads happened total!)

Solution The Bernoulli likelihood comes by using the PMF $p_X(k) = \theta^k(1-\theta)^{1-k}$ for $k \in \{0, 1\}$. We get this by observing that $\text{Ber}(\theta) = \text{Bin}(1, \theta)$.

$$L(x_1, \dots, x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \left(\prod_{i=1}^n \theta^{x_i}\right)\left(\prod_{i=1}^n (1-\theta)^{1-x_i}\right)$$

$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = \theta^{T(x_1,\dots,x_n)}(1-\theta)^{n-T(x_1,\dots,x_n)}$$

Choose

$$g(x_1, \dots, x_n) = 1$$

and

$$h(T(x_1, \dots, x_n), \theta) = \theta^{T(x_1,\dots,x_n)}(1-\theta)^{n-T(x_1,\dots,x_n)}$$

By the Neyman-Fisher Factorization Criterion, $T(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ is sufficient. The mean $\hat{\theta}_{MLE} = \dfrac{\sum_{i=1}^n x_i}{n} = \dfrac{T(x_1, \dots, x_n)}{n}$ is as well, since knowing the total number of heads and the sample proportion of heads is equivalent (since we know n)!

Notice that here the only interaction between the data and parameter θ happens through the sufficient statistic (the sum/mean of all the values). We don't actually need to know each individual x_i. □

7.8.2 Properties of Estimators Summary

Here are all the properties of estimators we've talked about from 7.6 to 7.8 (now), in one place!

Definition 7.8.3: Bias

Let $\hat{\theta}$ be an estimator for θ. The **bias** of $\hat{\theta}$ as an estimator for θ is

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

As estimator is **unbiased** if $\text{Bias}(\hat{\theta}, \theta) = 0$ or equivalently, $\mathbb{E}\left[\hat{\theta}\right] = \theta$.

Definition 7.8.4: Mean Squared Error (MSE)

The **mean squared error** of an estimator $\hat{\theta}$ of θ measures the expected squared error from the true value θ, and decomposes into a bias term and variance term. This term results in the phrase "Bias-Variance Tradeoff" - sometimes these are opposing forces and minimizing MSE is a result of choosing the right balance.

$$\text{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}^2(\hat{\theta}, \theta)$$

If $\hat{\theta}$ is an unbiased estimator of θ, then the MSE reduces to just: $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right)$.

Definition 7.8.5: Consistency

An estimator $\hat{\theta}_n$ (depending on n iid samples) of θ is **consistent** if it converges (in probability) to θ. That is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \epsilon\right) = 0$$

Definition 7.8.6: Efficiency

An *unbiased* estimator $\hat{\theta}$ is **efficient** if it achieves the **Cramer-Rao Lower Bound**, meaning it has the lowest variance possible.

$$e(\hat{\theta}, \theta) = \frac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)} = 1 \iff \mathsf{Var}\left(\hat{\theta}\right) = \frac{1}{I(\theta)} = \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln L(\mathbf{x}|\theta)}{\partial \theta^2}\right]}$$

Definition 7.8.7: Sufficiency

An estimator $\hat{\theta} = T(x_1, \ldots, x_n)$ is **sufficient** if it satisfies the **Neyman-Fisher Factorization Criterion**. That is, there exist non-negative functions g and h such that:

$$L(x_1, \ldots, x_n \mid \theta) = g(x_1, \ldots, x_n) \cdot h(\hat{\theta}, \theta)$$