

Lecture 18: Sufficiency

*Instructor: Hyunseung Kang**Scribe: Meenmo Kang*

1 Motivation for Sufficiency

Statisticians have long been concerned about reducing size of data. This is due to limited computing ability in comparison to size of data at all time. As follows, we are going to see if there is an available function that can efficiently store information of collected sample.

In the past, even with a relatively small amount of data, the computing ability was limited. On the other hand, in 2018, we are living with much larger size of data with latest computers. However, as it is called the era of big data, the size of data sets has become massive, so computing ability is still problematic. Therefore, the question how to reduce original data to a simplified function has always been one of the important issues among statisticians.

- Suppose we collect n *iid* samples $X_i \stackrel{iid}{\sim} \text{pdf}_\theta$, to learn about θ .
- Is there a function of the n data points, say $T(X_1, \dots, X_n)$, where T captures **all** the info about θ ?

2 Example: Coin Toss

We want to test whether a coin is fair or not. You flip the coin n times and collect the result X_1, \dots, X_n . Since only two possible outcomes exist, it is binary.

$$X_i = \begin{cases} 1 & \text{if } i\text{th toss is head} \\ 0 & \text{if } i\text{th toss is tail} \end{cases}$$

In case of Bernoulli distribution, instead of the entire vector (X_1, \dots, X_n) , the proportion of the distribution is enough information to represent the whole sample.

$$T(X_1, \dots, X_n) = \frac{\sum_{i=1}^n X_i}{n} = \text{Proportion of heads in } n \text{ tosses}$$

It is noticeable that, through this sufficiency procedure, n data points were summarized as one single number, and there was no loss of information though most of the data was discarded.

3 Formal Definition of Sufficiency

A statistic $T(X_1, \dots, X_n)$ is sufficient for θ if $p(X_1, \dots, X_n | T(X_1, \dots, X_n))$ is not a function of θ .

3.1 Factorization Theorem

A statistic $T(X_1, \dots, X_n)$ is sufficient for θ if and only if

$$pdf_{\theta}(X_1, \dots, X_n) = g[T(X_1, \dots, X_n), \theta] h(X_1, \dots, X_n)$$

3.2 Example(Bernoulli: Coin Toss)

$(X_1, \dots, X_n) \stackrel{iid}{\sim} \text{Bern}(p)$ (p is true proportion of heads)

Goal

Learn about p from n data points (X_1, \dots, X_n) .

Show

That we only need $\frac{\sum_{i=1}^n X_i}{n} = T(X_1, \dots, X_n)$, which is p , to learn about θ (i.e. show T is sufficient)

3.2.1 By Factorization Theorem

$$\begin{aligned} pdf_{\theta}(X_1, \dots, X_n) &= \prod_{i=1}^n [p^{X_i} (1-p)^{1-X_i}] = p^{\sum_{i=1}^n X_i} \cdot (1-p)^{\sum_{i=1}^n (1-X_i)} \\ &= p^{\sum_{i=1}^n X_i} \cdot (1-p)^{n-\sum_{i=1}^n X_i} = p^{nT(X_1, \dots, X_n)} \cdot (1-p)^{n-nT(X_1, \dots, X_n)} \\ g(T(X_1, \dots, X_n), p) &= p^{nT(X_1, \dots, X_n)} \cdot (1-p)^{n(1-T(X_1, \dots, X_n))} \quad h(X_1, \dots, X_n) = 1 \end{aligned}$$

Thus $T(X_1, \dots, X_n) = \frac{\sum_{i=1}^n X_i}{n}$ is sufficient for θ or (p in this case).

3.3 Example (Normal)

$(X_1, \dots, X_n) \stackrel{iid}{\sim} N(\mu, \sigma^2) \Rightarrow$ What are the sufficient stats for μ and σ^2 ?

$$\begin{aligned}
 T(X_1, \dots, X_n) &= \left(\frac{\sum_{i=1}^n X_i}{n}, \frac{\sum_{i=1}^n X_i^2}{n} \right) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^n \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^n \exp\left(-\frac{X_i^2 - 2\mu X_i + \mu^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^n \exp\left(-\frac{X_i^2}{2\sigma^2} - \frac{2\mu X_i}{2\sigma^2} + \frac{\mu^2}{2\sigma^2}\right) \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^n \frac{X_i^2}{2\sigma^2} - \sum_{i=1}^n \frac{2\mu X_i}{2\sigma^2} + \sum_{i=1}^n \frac{\mu^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n X_i^2 - 2\mu \sum_{i=1}^n X_i + n\mu^2\right)\right) \\
 &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(T(X_1, \dots, X_n) \cdot \left[\frac{n\mu}{\sigma^2}, \frac{-n}{2\sigma^2}\right]\right) \cdot \exp\left(\frac{-n\mu^2}{2\sigma^2}\right) \Rightarrow g(T(X_1, \dots, X_n))
 \end{aligned}$$

Since there is no terms with respect to X_i , we assume $h(X_1, \dots, X_n) = 1$. Therefore, this is sufficient for (μ, σ^2)

Theorem

If $T(X_1, \dots, X_n)$ is sufficient for θ , then your MLE estimation of θ **must be** functions of $T(X_1, \dots, X_n)$.

Proof

$$\begin{aligned}
 \hat{\theta}_{MLE} &= \operatorname{argmax}_{\theta} \operatorname{pdf}_{\theta}(X_1, \dots, X_n) \\
 &= \operatorname{argmax}_{\theta} g(T(X_1, \dots, X_n), \theta) h(X_1, \dots, X_n) = \operatorname{argmax}_{\theta} \log[g(T(X_1, \dots, X_n), \theta) h(X_1, \dots, X_n)] \\
 &= \operatorname{argmax}_{\theta} (\log[g(T(X_1, \dots, X_n), \theta)] + \log[h(X_1, \dots, X_n)]) = \operatorname{argmax}_{\theta} \log[g(T(X_1, \dots, X_n), \theta)]
 \end{aligned}$$

Hence, since g is only dependent on $T(X_1, \dots, X_n)$ and θ , every MLE must be a function of $T(X_1, \dots, X_n)$.

4 Exponential Family

A $\operatorname{pdf}_{\theta}$ is a part of a family of functions known as exponential family if the pdf is of the form

$$\operatorname{pdf}_{\theta} X_i = \exp[c(\theta) \cdot T(x_i) + d(\theta) + f(x_i)] \text{ where } c, T, d, \text{ and } f \text{ are functions}$$

Note that if $f(x|\theta)$ of n iid samples. $T(x_i)$ is considered to be sufficient as below.

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^n pdf(X_i, \dots, X_n) = \prod_{i=1}^n \exp(c(\theta)T(x_i) + d(\theta) + f(x_i)) \\ &= \exp(c(\theta) \sum_{i=1}^n T(x_i)) + n \cdot d(\theta) \cdot \exp(\sum_{i=1}^n f(x_i)) \end{aligned}$$

4.1 Example: Bernoulli Distribution (Coin Toss)

Suppose that we flip a coin n times again.

n coin toss example (Show its in the exponential family)

$$\begin{aligned} pdf(X_i) &= p^{X_i}(1-p)^{1-X_i} \\ &= \exp(\log(p^{X_i}(1-p)^{1-X_i})) = \exp(X_i \log(p) + (1-X_i) \log(1-p)) \\ &= \exp(X_i \cdot \log(\frac{p}{1-p}) + \log(1-p)) \end{aligned}$$

$$\begin{aligned} \prod_{i=1}^n pdf(X_i) &= \prod_{i=1}^n \exp(X_i \cdot \log(\frac{p}{1-p}) + \log(1-p)) \\ &= \exp(\log(\frac{p}{1-p}) \cdot \sum_{i=1}^n X_i + n \cdot \log(1-p)) \end{aligned}$$

$$\Leftrightarrow c(p) = \log(\frac{p}{1-p}) \quad T(X_i) = X_i \quad d(p) = \log(1-p) \quad f(X_i) = 1$$

Therefore, a coin toss distribution is a part of exponential family.