

STA261: Probability and Statistics II

Shahriar Shams

Week 4 (Sufficiency & Consistency of an estimator, Score & Fisher Information)



Winter 2020

Recap of Week 2-3

- Learned two formal ways of defining an estimator
 - Method of Moments Estimator.
 - Maximum Likelihood Estimator (MLE).
- Sampling distribution of an estimator
 - Sampling distribution of (\bar{X}) (*under Normal and non-Normal dist*)
 - Sampling distribution of (S^2) (*only under Normal dist*)
- Properties of an estimator
 - Unbiasedness

Learning goals for this week

- Sufficient statistic (Rice page 305, E&R page 302)
- Consistent estimator (Rice page 266, ER page 325)
- ~~Efficient estimator~~ (will cover next week)
- Score and Fisher Information (Rice page-276, E&R page 365)

These are selected topics from [Evans and Rosenthal: chapter 6](#) and [John A. Rice: Chap 8](#)

Section 1

Sufficient Statistic

An example to explain the intuition

- Suppose I give these following two sets of information to the two sections of this course.
- It involves estimating the parameter (θ) of a Bernoulli distribution.

Info-1 for day section

I have tossed a fair coin 5 times and the outcomes are (1,0,1,1,0).
can you calculate the MLE of θ ?

Info-2 for evening section

I have tossed a fair coin 5 times and got 3 Heads.
can you calculate the MLE of θ ?

An example to explain the intuition (cont...)

Note:

- For info-1, $L(\theta) = \theta^3(1 - \theta)^2$
- For info-2, $L(\theta) = \binom{5}{3}\theta^3(1 - \theta)^2$
- Both sections will give me the same answer $[\hat{\theta} = 0.6]$
- In info-2, I calculated a summary of the sample observations.
- This summary (Let's call it $T(x_1, x_2, \dots, x_n)$) contains the same info about θ as it is contained in the entire sample (x_1, x_2, \dots, x_n) .
- So we say $T(x_1, x_2, \dots, x_n)$ is sufficient for θ .
- This is considered as a data reduction.
- Sufficient statistic is parameter specific.

Definition of sufficient statistic (Rice-P305)

A statistic $T(X_1, X_2, \dots, X_n)$ is said to be **sufficient** for θ if the conditional distribution of X_1, X_2, \dots, X_n , given $T = t$, does not depend on θ .

In other words: once we have the value of the sufficient statistic, the actual sample observations don't add any more information about the parameter.

For example, for the evening section where I have told the total number of heads already, giving them the actual sequence (1,0,1,1,0) won't add anything new.

Example using $Poisson(\lambda)$

Suppose $X_1, X_2, X_3 \stackrel{iid}{\sim} Poisson(\lambda)$. Verify that $T = \sum_{i=1}^3 X_i$ is a sufficient statistic for λ .

Note: a similar example is given in the Rice text book (page 306) for Bernoulli distribution.

Factorization theorem - An easier way of finding sufficient statistic

$T(X_1, X_2, \dots, X_n)$ is said to be **sufficient** for θ if the joint probability function factors in the form

$$f(x_1, x_2, \dots, x_n | \theta) = g[T(x_1, x_2, \dots, x_n), \theta] * h(x_1, x_2, \dots, x_n)$$

where,

- $h(x_1, x_2, \dots, x_n)$ is a function of sample observations only
- $g[T(x_1, x_2, \dots, x_n), \theta]$ involves θ and the sufficient statistic T

Note: By now we know $f(x_1, x_2, \dots, x_n | \theta)$ is just the likelihood function. Proof of this theorem is available on page 307 of Rice book (not needed for the course)

Factorization theorem applied on $Poisson(\lambda)$

$$\begin{aligned} L(\lambda) &= e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} * \frac{1}{\prod_{i=1}^n x_i!} \\ &= g\left[\sum_{i=1}^n x_i, \lambda\right] * h(x_1, x_2, \dots, x_n) \end{aligned}$$

Therefore, according to the factorization theorem, $T = \sum_{i=1}^n x_i$ is a sufficient statistic for λ

Note: when we maximize likelihood, we maximize $g[T(x_1, x_2, \dots, x_n), \theta]$. Hence, MLE is a function of sufficient statistic $T(x_1, x_2, \dots, x_n)$.

Section 2

Consistent Estimator

Definition of consistent estimator (E&R-P325)

- Let T_n be an estimator of parameter θ
- T_n is said to be consistent(in probability) if $T_n \xrightarrow{P} \theta$
- In words, T_n converges to θ in probability.

Note:

- If $T_n \xrightarrow{a.s.} \theta$ then T_n is called consistent(almost surely).
- In this course we will only talk about consistent(in probability)

Proving consistency using LLN

- LLN tells us, $\bar{X} = \frac{1}{n} \sum X_i \xrightarrow{P} E[X_i]$ for any distribution.
- Immediately that tells us:
 - If $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ then \bar{X} is a consistent estimator of μ
 - If $X_i \stackrel{iid}{\sim} Poisson(\lambda)$ then \bar{X} is a consistent estimator of λ
 - And we can say this for few other known distributions (do it yourself)
- How can we prove consistency when the estimator is not simply \bar{X} ?
 - We can still use LLN but with the help of a well known Lemma and the continuous mapping theorem.

Slutsky's Lemma and Continuous mapping theorem

- **Slutsky's Lemma:**

- We have two different sequences X_n and Y_n
- $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y \implies X_n + Y_n \xrightarrow{P} X + Y$
- $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y \implies X_n Y_n \xrightarrow{P} XY$

- **Continuous mapping theorem:**

- Let $X_n \xrightarrow{P} X$ and $g(\cdot)$ be a continuous function
- then $g(X_n) \xrightarrow{P} g(X)$

Proving S^2 is a consistent estimator of σ^2

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_i (X_i - \bar{X})^2 = \left(\frac{n}{n-1} \right) \left(\frac{1}{n} \sum_i (X_i - \bar{X})^2 \right) \\ &= \left(\frac{n}{n-1} \right) \left(\frac{1}{n} \left[\sum_i X_i^2 - n\bar{X}^2 \right] \right) \\ &= \left(\frac{n}{n-1} \right) \left(\frac{1}{n} \sum_i X_i^2 - \bar{X}^2 \right) \\ &= \left(\frac{n}{n-1} \right) \left(\frac{1}{n} \sum_i X_i^2 - (\bar{X})^2 \right) \\ &\implies S^2 \xrightarrow{p} (1) (E[X^2] - (E[X])^2) = \sigma^2 \end{aligned}$$

Note: Using continuous mapping theorem, we can say, S is a consistent estimator of σ

MSE consistent (often used in practice)

- An estimator T_n is called **MSE consistent** if

$$MSE(T_n) \rightarrow 0 \text{ as } n \rightarrow \infty$$

- **Example:** for $N(\mu, \sigma^2)$
 - $MSE(\bar{X}) = \sigma^2/n \rightarrow 0$ as $n \rightarrow \infty$
 - Therefore \bar{X} is a MSE consistent estimator of μ
- In naive words, after you have calculated the MSE of an estimator, just check if it goes to zero for large n ...

MLE is consistent

Before making this claim, Let us introduce a new notation and revisit few of the old ones

- θ_0 : The TRUE value of the parameter which produced the data. (which is a unknown constant)
- Suppose $(X_1, X_2, \dots, X_n) \stackrel{iid}{\sim} f(x|\theta_0)$
- $\hat{\theta}$ is MLE

Claim: $\hat{\theta}$ converges to θ_0 in probability ($\hat{\theta} \xrightarrow{P} \theta_0$)

MLE is consistent (illustration of the proof)

- Let us consider the situation where we haven't observed the samples yet.
- log-likelihood, $l(\theta) = \sum_{i=1}^n \log f(X_i|\theta)$
- In naive words, the log-likelihood function varies from one set of sample to the other.
- Dividing both side by the sample size n

$$\frac{1}{n}l(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(X_i|\theta)$$

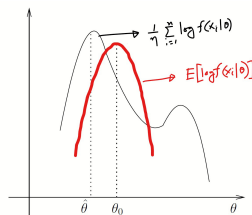
- The right hand side of the above equation is a sample mean!
- Applying LLN,

$$\frac{1}{n} \sum_{i=1}^n \log f(X_i|\theta) \xrightarrow{P} E[\log f(X_i|\theta)]$$

MLE is consistent (illustration of the proof cont...)

- The main idea:

since $\frac{1}{n}l(\theta)$ gets closer to $E[\log f(X_i|\theta)]$, the θ that maximizes $\frac{1}{n}l(\theta)$ should be close to the θ that maximizes $E[\log f(X_i|\theta)]$



$$E[\log f(X_i|\theta)] = \int_x \log f(x|\theta) f(x|\theta_0) dx$$

- Show that $E[\log f(X_i|\theta)]$ is maximized at θ_0 (Rice page 276)

source of the graph: <https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture3.pdf>

Section 3

Score and Fisher Information

- **Score function, $S(\theta)$:**

- it's the derivative of the log-likelihood

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta} = l'(\theta)$$

- When we say “score function” we mean it's a function of θ

- **Score equation:**

- $S(\theta) = 0$
- the solution of this equation is the the MLE
- we can say $S(\theta)|_{\theta=\hat{\theta}} = 0$

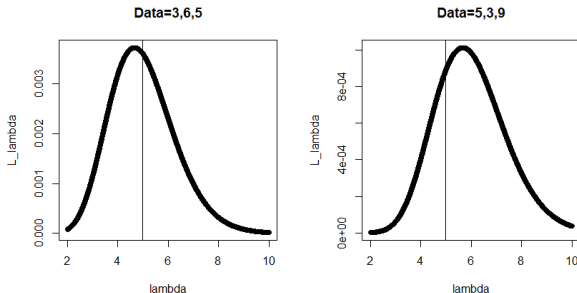
- **Score as a random variable:**

- For the random variable X_i , $S(\theta|X_i) = \frac{\partial}{\partial \theta} \log f(X_i|\theta)$
- For *iid* (X_1, X_2, \dots, X_n) ,

$$S(\theta|X_1, \dots, X_n) = \frac{\partial}{\partial \theta} \sum_i \log f(X_i|\theta) = \sum_i \frac{\partial}{\partial \theta} \log f(X_i|\theta) = \sum_i S(\theta|X_i)$$

A plot showing the randomness of $S(\theta)$

- Both of these likelihood plots are for $Poisson(\lambda)$ distribution.
- In both cases we have 3 observations ($n=3$) generated from a Poisson with $\lambda = 5$ (true value, $\lambda_0 = 5$)
- The likelihood function looks different for different data!
- The slopes at $\lambda = 5$ differs \implies Score evaluated at $\lambda = 5$ is a random variable



Randomness of Score using Poisson distribution

- Let $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} Pois(\lambda)$ with true value of λ being λ_0
- We can show, $S(\lambda|X_1, \dots, X_n) = -n + \frac{\sum_i X_i}{\lambda}$
- The score contains $\sum_i X_i$ which will change value from one set of sample to the other.
- It's the sample obs. (X_1, X_2, \dots, X_n) in the score func. which makes it a random variable

One important property of $S(\theta)$

Under some assumptions,

- $E[S(\theta|X)]|_{\theta=\theta_0} = 0$ (proof...)
- This expectation is taken over X .

Example: Let $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \text{Pois}(\lambda)$ with true value of λ being λ_0

$$\begin{aligned} E[S(\lambda|X_1, \dots, X_n)] &= E\left[-n + \frac{\sum_i X_i}{\lambda}\right] \\ &= -n + \frac{1}{\lambda} E\left[\sum_i X_i\right] \\ &= -n + \frac{1}{\lambda} n E[X_i] \\ &= -n + \frac{1}{\lambda} n \lambda_0 \end{aligned}$$

at $\lambda = \lambda_0$, $E[S(\lambda|X_1, \dots, X_n)] = 0$

Fisher Information, $I(\theta_0)$

Definition:

$$I(\theta_0) = \text{var}[S(\theta|X)|_{\theta=\theta_0}] = E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)|_{\theta=\theta_0}\right]^2$$

- It's the amount of “information” that each observable random variable X contains about θ

For $Pois(\lambda_0)$

$$I(\lambda_0) = \text{var}[S(\lambda|X)|_{\lambda=\lambda_0}] = \text{var}\left[-1 + \frac{X}{\lambda_0}\right] = \frac{1}{\lambda_0^2} \text{var}[X] = \frac{1}{\lambda_0}$$

Fisher Information from a sample of size n

Fisher Information of a set of sample of size n ,

$$\begin{aligned} & \text{var}[S(\theta|X_1, X_2, \dots, X_n)|_{\theta=\theta_0}] \\ &= \text{var}\left[\sum_i S(\theta|X_i)|_{\theta=\theta_0}\right] \\ &= \sum_i \text{var}[S(\theta|X_i)|_{\theta=\theta_0}] \\ &= nI(\theta_0) \end{aligned}$$

An easier way of calculating $nI(\theta_0)$

- It can be shown that

$$I(\theta_0) = E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\Big|_{\theta=\theta_0}\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\Big|_{\theta=\theta_0}\right]$$

- Often in practice, we use the second derivative of the log-likelihood to calculate $nI(\theta_0)$

Summarizing the steps:

- Write down the log-likelihood in terms of random variable X
- Differentiate twice with respect to θ and then put $\theta = \theta_0$
- Take Expectation over X and finally multiply by (-1)

For $X_1, X_2, \dots, X_n \stackrel{iid}{\sim} \text{Pois}(\lambda_0)$

$$nI(\lambda_0) = -E\left[\frac{\partial^2}{\partial\lambda^2}(-n\lambda + \sum_i X_i \log\lambda)\Big|_{\lambda=\lambda_0}\right] = -E\left[-\frac{\sum_i X_i}{\lambda_0^2}\right] = \frac{n}{\lambda_0}$$

Assignment (Non-credit)

Finding sufficient statistic and Fisher information for all the known distributions

Evans and Rosenthal

6.5.1-6.5.3

John A. Rice

Exercise 8: 16(d), 17(e), 18(d), 21(c), 47(d), 52(d), 68(we did parts of this in the lecture), 69-73

R home work

1. Write a function that generates 30 random samples from $Poisson(\lambda = 5)$ dist and calculates the score function for $\lambda = 5$
2. Run this function 100K times and calculate the mean and variance of the output (mean should be ≈ 0 and $\text{var} \approx 30/5 = 6$)