STA255 Week-7

Shahriar Shams

24/02/2020

Review of Week-6

- A brief introduction to Population, Sample, Parameter and Statistic
- Estimator vs Estimate
- Sampling distribution
- Mean and Variance of $\bar{X}$ under any distribution
- Sampling distribution of sample mean($\bar{X}$) under Normal distribution
- Central Limit Theorem (CLT)
- Law of large number (sample mean approaches true mean)

Learning goals

- Idea of Point Estimation
- Method of Moment Estimation
- Mean Square Error (MSE): measuring accuracy of an estimator
- Bias and Underhandedness
- Maximum Likelihood Estimation

Idea of Point Estimation

- Until now we have learned how to compute different aspects of a population distribution given that the parameter values are known
  - For example, if I say height follows Normal with mean 160 and variance 100, we know how to calculate $P[height < 150]$ or 90th percentile of the height etc.
- But in real life, distributions are often assumptions about the population. And parameter values are often unknown.
- **Estimation** refers to the idea of calculating a "sensible" summary of the sample observations and use it as an educated guess about the parameter.
- Estimation is broadly classified into two categories:
  - *Point estimation:* calculating one single point on the number line and use it as en estimate.
  - *Interval estimation:* calculating a range of values(i.e. an interval) that has a desired chance of containing the true parameter value.

- This week we will cover Point Estimation.
- The choice of an estimator is sometime intuitive.
  - If we are interested in the population mean, intuitively we would use the sample mean.
- There are two formal ways of defining an estimator:
  - Method of Moments Estimation
  - Maximum Likelihood Estimation

Method of Moments Estimation

- Let $X_1, X_2, ..., X_n$ are independently and identically distributed (i.i.d.) random variables.
- Let the k^{th} population moment be

$$\mu_k = E[X^k]$$

- k^{th} sample moment based on sample

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$ ← estimator

estimate

$$\hat{\mu}_K = \frac{1}{n}\sum x_i^k$$

- We use $\hat{\mu}_k$ as an estimator of μ_k
- In other words, we use the sample moments as estimators of the population moments.

Example-1: $X_1, X_2, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. Find the method of moments estimator of λ.

1st pop^n moment, $E[X] = \lambda$

" Sample " , $\frac{1}{n}\sum X_i = \bar{X}$

Method of Moments estimator of λ,

$\bar{X}$

Example-2: $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Find the method of moments estimators of μ and σ^2.

$$E[X] = \mu \leftarrow \bar{X}$$

$$E[X^2] = \sigma^2 + \mu^2 \leftarrow \frac{1}{n}\sum X^2$$

$$\sigma^2 = E[X^2] - (E[X])^2 \leftarrow \boxed{\frac{1}{n}\sum X^2 - (\bar{X})^2}$$

$$= \frac{1}{n}\sum (X_i - \bar{X})^2.$$

Method of moments estimators of $\mu = \bar{X}$ and

$$\sigma^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2$$

- There are two more examples (one for Exponential and one for Gamma distribution) done on page 351 of the text book.

Summary of Method of Moments Estimator

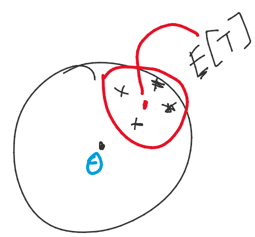
- Express the lower order population moment(s) in terms of the parameter(s).
- Invert the expression(s) to express the parameter(s) in terms of the population moment(s)
- Replace the population moments using the sample moments.

Mean Squared Error (MSE): measuring accuracy of an estimator

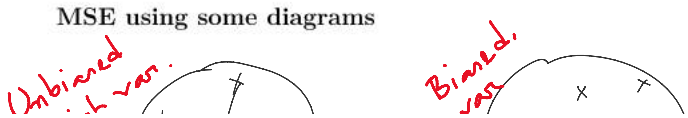
- Let θ be the unknown parameter of interest.
- Suppose, T is an estimator of θ
- The most commonly used measurement of accuracy of an estimator is *Mean Squared Error (MSE)*
- $MSE[T] = E[(T - \theta)^2]$
- The smaller the value of $MSE_\theta(T)$, the more concentrated the sampling distribution of T is about the value θ
- It is possible to express MSE(T) as combination of two components.

$$MSE[T] = var[T] + (E[T] - \theta)^2$$

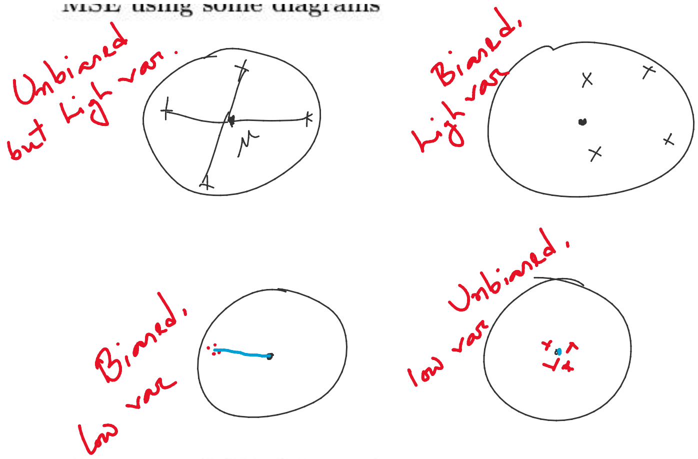
$$MSE[\bar{X}] = V[\bar{X}] = \frac{\sigma^2}{n}$$



MSE using some diagrams



MSE using some diagrams



Bias and Unbiasedness

- There are several characteristics of a good estimator.
- Unbiasedness is one of them.
- Sufficiency, consistency and efficiency are other three well known characteristics that we will not study in this course.
- Bias relates to the sampling distribution of the estimator.
- In week-6 we learned that estimator is a random variable.
- If T is an estimator for θ, we want the expected value of T to be θ.
- Bias of an estimator is defined as

$$Bias[T] = E[T] - \theta$$

- An estimator, T is called **unbiased** if $Bias[T] = 0 \implies E[T] = \theta$

Unbiasedness of sample mean ($\bar{X}$)

- In week_6, we showed $E[\bar{X}] = \mu$ irrespective of the distribution. Here μ is the population mean and $\bar{X}$ is the sample mean.
- Since $E[\bar{X}] = \mu$, we can say sample mean is an unbiased estimator of the population mean (for any distribution).

Unbiasedness of sample variance

- For sample variance often we see two estimators in practice,
  -
  $$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$$ • Unbiased
  - and
  $$\hat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$$ • biased
  - the denominator is different. One is divided by $n-1$ the other is divided by n.

- One of these are unbiased and the other is biased. $E[S^2] = \sigma^2$ but $E[\hat{\sigma}^2] \neq \sigma^2$

- Standard Error is the standard deviation of an estimator.

$SE[\bar{x}] = \frac{\sigma}{\sqrt{n}}$

Maximum Likelihood Estimation

- In estimation, our goal is to guess the value of the parameter (θ) based on the sample observations $(x_1, x_2, ..., x_n)$

- Let us start with an example that will give the intuition behind maximum likelihood estimation.
- Assume we have two distributions: $P1$ and $P2$, both discrete uniforms.
- Under $P1$, $X \sim Unif\{1, 2, ..., 10^3\}$
- Under $P2$, $X \sim Unif\{1, 2, ..., 10^6\}$

- We observe one sample value of X.
- Say we observe x=5000
  - Which distribution did it come from? (this is a easy one)

- Say we observe another sample value, x=10
- Which distribution did it come from?
  - Which distribution is more *likely* to have produced this number?

Definition of likelihood function

- Suppose $X_1, X_2, ..., X_n$ has a joint density or mass function $f(x_1, x_2, ..., x_n|\theta)$ θ
- We observe sample, $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$
- Given the sample, the likelihood function of θ, denoted by $L(\theta|x_1, x_2, ..., x_n)$, is defined as

$$L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta)$$

- $L(\theta|x_1, x_2, ..., x_n)$, or often written as $L(\theta)$, is a function of θ

- The text book does not use the notation $L(\theta|x_1, x_2, ..., X_n)$ or $L(\theta)$.
- We will use the notation $L(\theta)$ to denote likelihood function.
- In short, likelihood function is simply the joint pdf or pmf of your observed data.

Likelihood function where observations are independent

- If the observations are independent we can write the joint pmf or pdf as the product of individual pdfs or pmfs.
- So if $X_1, X_2, ..., X_n$ are independent we can write,

$$L(\theta) = f(x_1, x_2, ..., x_n|\theta) = f_\theta(x_1) * f_\theta(x_2) * ... * f_\theta(x_n)$$

Example:

- Suppose we have a coin.
- We don't know the value of probability of getting a head (θ) for this particular coin.
- Let's define a random variable X,

$$X = 1, \text{ if it's a Head}$$
$$X = 0, \text{ if it's a Tail}$$

$P[X=1]=\theta$

$P[X=0]=1-\theta$

- So We can say $X \sim Bern(\theta)$
- Say we toss the coin 5 times and observe HTHHT.
- So we have

$$X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0$$

* The likelihood function for this data,

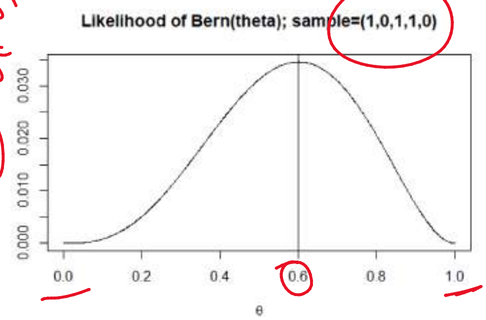
$$L(\theta) = \theta * (1-\theta) * \theta * \theta * (1-\theta) = \theta^3(1-\theta)^2$$

$\theta=\frac{1}{2}$, $L(\theta)=0.03125$

$\theta=\frac{1}{4}$ $L(\theta)=0.0087$

- Clearly this is a function of θ. Let's plot this function.

$L(\theta) = \theta^3(1-\theta)^2$



$\frac{d}{d\theta}\left[\theta^3(1-\theta)^2\right] = 0$

$\hat{\theta} = 0.6$

- Keeping the intuition presented to you using the two discrete uniform distributions on page 5 of this note,
  - What θ value will you pick as your best guess?
- We will pick the value for which the likelihood function is maximized.
  - This is the reason why it's called maximum likelihood estimation.
- This value is called the maximum likelihood estimate and denoted by $\hat{\theta}$

Calculation of Maximum Likelihood Estimate (MLE)

- From Calculus, we know that finding maximum of a function is done through differentiation.
- $L(\theta)$ is often a product of complicated pdf functions that makes it harder to differentiate.
- For this reason, instead of maximizing $L(\theta)$ we maximize $logL(\theta)$, which is called the log-likelihood function.
- $logL(\theta)$ converts the products into sum. $\log AB = \log A + \log B$

For our example,

- $logL(\theta) = 3log\theta + 2log(1-\theta)$
- To find maxima, we calculate $\frac{d}{d\theta}logL(\theta)$ and set it equal to zero which gives us an equation.
- Solving that equation for θ gives us the maxima

- In previous part, we found θ where the slope of the log-likelihood is zero.
- So the solution that we got can very well be a point where the function is minimized.
- So technically we have to check the second derivative and see if it's negative or not.

- But the text book avoided getting into this. And we will skip this. But bare in mind the calculation that we will do is not complete.

Example:

$X_1, X_2, ..., X_n$ are interdependently drawn from $Exp(\beta)$ with pdf

$$f(x) = \frac{1}{\beta}e^{-\frac{1}{\beta}x}; \quad x > 0; \, \beta > 0$$

Find the maximum likelihood estimate of β.

$$L(\beta) = f(x_1, x_2 \cdots x_n | \beta)$$

$$= f(x_1) f(x_2) \cdots f(x_n)$$

$$= \frac{1}{\beta} e^{-\frac{1}{\beta}x_1} * \frac{1}{\beta} e^{-\frac{1}{\beta}x_2} * \cdots * \frac{1}{\beta} e^{-\frac{1}{\beta}x_n}$$

$$= \left(\frac{1}{\beta}\right)^n \cdot e^{-\frac{1}{\beta}\Sigma x_i}$$

$$\log L(\beta) = n \log \frac{1}{\beta} - \frac{1}{\beta}\Sigma x_i$$

$$= -n \log \beta - \frac{1}{\beta}\Sigma x_i$$

$$\frac{d \log L(\beta)}{d\beta} = -\frac{n}{\beta} + \frac{1}{\beta^2}\Sigma x_i = 0$$

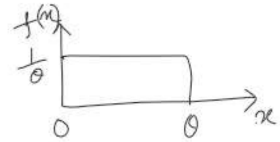
$$\Rightarrow \frac{\Sigma x_i}{\beta^2} = \frac{n}{\beta}$$

$$\Rightarrow \hat{\beta} = \frac{\Sigma x_i}{n} = \bar{x}$$

estimator $\hat{\beta} = \bar{X}$

Condition where differentiation does not work.

- Anytime we have the parameter in the range of the random variable, the differentiation technique does not work. eg. $x \geqslant \theta$ or $x \leq \theta$
- For example, say out observations are from a $Unif[0, \theta]$.
- The likelihood function is not continuous every where.



$$X_1, X_2, X_3 \sim Unif[0, \theta]$$

$$f(x) = \frac{1}{\theta} I(0 \leq x \leq \theta)$$

$$L(\theta) = f(x_1) f(x_2) f(x_3)$$
$$= \frac{1}{\theta} \cdot \frac{1}{\theta} \cdot \frac{1}{\theta} I(0 \leq x_1, x_2, x_3 \leq \theta)$$
$$= \frac{1}{\theta^3} I(0 \leq x_1, x_2, x_3 \leq \theta)$$

Ex: Suppose we have observed 3, 4 and 2 from this dist$^{\underline{n}}$. $\Rightarrow 0 \leq 3, 4, 2 \leq \theta$

$\Rightarrow \theta \not< 4$.

$L(\theta)$ will be maximized when θ is minimized

$$\min(\theta) = \max(X_1, X_2 \cdots X_n)$$

Max likelihood estimator of θ.

$$\hat{\theta} = \max(X_1, X_2, \ldots X_n)$$

Some properties of MLE

- Using the idea of estimate vs estimator
  - MLE is a number that can vary from one set of sample to the other.
  - Hence we can treat it as a random variable
  - At that point it's called maximum likelihood estimator.
- MLE (the estimator) may or may not be unbiased.
- But for large n, it is possible to show that

$$E[\hat{\theta}] \approx \theta$$

- So we say MLE is asymptotically unbiased.

- Out of all the available unbiased estimators, MLE can be shown to have lowest variance.

Chapter 7.3 and 7.4 not needed

Homework

Chapter 7.1

1, 5, 9, 11, 12, 15, 19

Chapter 7.2

21, 23, 24, 27, 28, 30