Math/Stat 310: Intro to Mathematical Statistics March 12/14, 2018

Lectures 12&13: MLE & Hypothesis Testing

Instructor: Hyunseung Kang *Scribe: Meenmo Kang*

1 Properties of MLE

In line with the last week, let's discuss about a couple more properties of MLE.

1.1 Invariance of MLE

Theorem
Given any function g: $\mathbb{R} \to \mathbb{R}$, if $\hat{\theta}_{MLE}$ is the MLE of θ from pdf_θ then $g(\hat{\theta}_{MLE})$ is the MLE of $g(\theta)$.

Example
Let us get back to our favorite example again, measuring students' height on campus. Suppose that samples are normally and *i.i.d.* distributed like $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. As we discovered last week, $\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ is for estimation of μ. Now, we want to estimate $sin(\mu)$.

In a naive way, we could probably rewrite the density function $\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$ in terms of $sin(\mu)$ in order to optimize the value. However, simply plugging $sin(\mu)$ into the normal density function is unable since $sin(\mu)$ is an invertible function. Instead, as long as we figure out a MLE, a MLE of $sin(\mu)$ is simply $sin(\hat{\mu}_{MLE})$.

Example
Suppose X_i is income of ith individual, and we collect n *i.i.d.* samples. One of popular economics topics is distribution of income. Typically, the income distribution is not normal. So, in order for income distribution to be more like normal, log transformation is used as a method. As a result, normally and *i.i.d.* log transformed samples are expressed as below.

$$log(X_1), ..., log(X_n) \overset{iid}{\sim} N(\mu, \sigma^2)$$

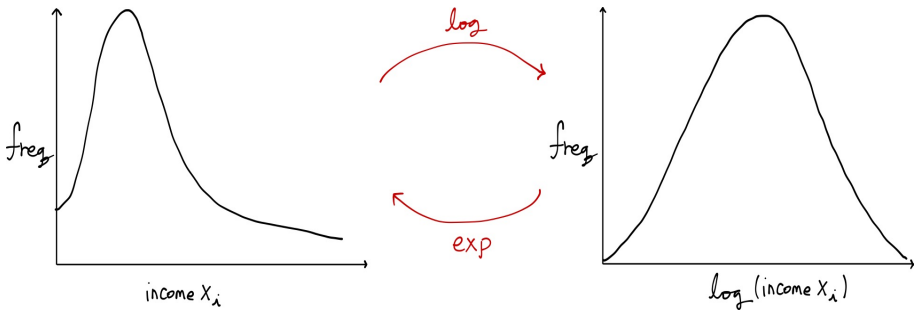
However, what we want to estimate is population mean income in $ scale, not log$ scale. At this point, we can come up with the fact that exponential to the log(A) is just A. Hence, in order for MLE estimator of μ in log scale, $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} \log X_i$ to be

transformed in an original scale, we can take exponential, such as

$$\hat{\mu}_{MLE}(original\ scale) = exp(\frac{1}{n}\sum_{i=1}^{n}\log X_i)$$

.

$$\mu_{orignal\ scale} = exp(\mu_{log\ sclae})$$



2 Point Estimate Optimality

Let's review the two different methods of estimating parameter that we have covered.

- MOM Estimator: $\hat{\theta}_{MOM}$
  Method of moment is to match an expectation value with θs, and replace that expectation with the law of large numbers counterparts like $E(X_1) \approx \frac{1}{n} - \sum_{i=1}^{n} X_i$

- MLE: $\hat{\theta}_{MLE}$
  Maximum likelihood estimation is to find a parameter(s) that maximizes the probability of observing given samples as follows.

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} pdf_\theta X_i \qquad \text{if } X_1, ..., X_n \overset{iid}{\sim} pdf_\theta$$

Then, how do we find optimized a parameter(s)?

2.1 Loss Function: $l(\hat{\theta}, \theta)$

The loss function measures the amount of error by measuring how much $\hat{\theta}$ deviates from the true value θ ($\hat{\theta}$ stands for estimator). Essentially, this is the estimation about our loss per estimate or parameter. Followings are a couple properties of loss function.

- $l(\hat{\theta}, \theta) \geq 0$
- When $l(\hat{\theta}, \theta) = 0$, this implies that $\hat{\theta}$ is a perfect estimator of θ.

Example

- $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$: Squared loss function
  This measures the squared distance between estimators and true value.
- $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$: Absolute loss function
- $l(\hat{\theta}, \theta) = I((\hat{\theta} \neq \theta)$: Indicator loss function
  As an indicator function, value of the function is either 0 or 1. Only if the estimator is equal to true value, it is equal to 1.

2.2 Risk function

Risk function is defined as expected loss. In other words, it is measuring how much we are going to lose or make an error on average.

Example Mean Square Error (MSE)
MSE is a special type of risk where $l(\hat{\theta}, \theta)$ is squared error loss. MSE = risk with the squared error loss function, defining $E[(\hat{\theta} - \theta)^2]$

Theorem If loss function is squared error, then

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

$$\text{where} \;\; \text{Bias}(\hat{\theta^2}) = \text{E}[\hat{\theta}] - \theta \quad \& \quad \text{Var}(\hat{\theta}) = \text{E}[\hat{\theta} - \theta)^2]$$

Proof
$\text{MSE} = \text{E}[\hat{\theta} - \theta)^2] = E[\hat{\theta} - \text{E}(\hat{\theta}) + \text{E}(\hat{\theta}) - \theta)] =$
$E[(\hat{\theta} - \text{E}[\hat{\theta}])^2 + (\text{E}(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - \text{E}(\hat{\theta}))(\text{E}(\hat{\theta}) - \theta)]$

As we discovered above,

$$\text{E}[\hat{\theta}] - \theta = \text{Bias}(\hat{\theta^2}) \& \text{E}[\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta})$$

As for the last term,

$$\text{E}[2(\hat{\theta} - \text{E}(\hat{\theta}))(\text{E}(\hat{\theta}) - \theta)] = 2(\text{E}(\hat{\theta}) - \theta)\text{E}[(\hat{\theta} - \text{E}(\hat{\theta}))]$$

since $2(\hat{\theta} - \mathrm{E}(\hat{\theta}))$ is constant

.

$$\Rightarrow 2(\mathrm{E}(\hat{\theta}) - \theta)[\mathrm{E}(\hat{\theta}) - \mathrm{E}(\mathrm{E}(\hat{\theta}))] = 2(\mathrm{E}(\hat{\theta}) - \theta)[\mathrm{E}(\hat{\theta}) - \mathrm{E}(\hat{\theta})]) = 0$$

Thus, MSE $= E[(\hat{\theta} - \mathrm{E}[\hat{\theta}])^2 + (\mathrm{E}(\hat{\theta}) - \theta)^2$, implying that under the squared of loss, when estimating θ, we should either minimize $Var(\hat{\theta})$ or Bias$(\hat{\theta}^2)$ due to the trade-off.

2.3 Optimality

To find all possible estimators $\hat{\theta}$ that minimize MSE, MLE is used as below, because $\hat{\theta}$ is a function, so derivative cannot be applied.

$$\hat{\theta}_{opt} = \underset{\hat{\theta}}{\operatorname{argmin}}\, E[(\hat{\theta} - \theta)^2]$$

In a narrower sense, we aim to find an estimator among all **unbiased** estimators, **called optimal unbiased estimator**.

$$\hat{\theta}_{opt\ unbiased} = \underset{opt\ unbiased}{\operatorname{argmin}}\ E[(\hat{\theta}_{opt\ unbiased} - \theta)^2] \Leftrightarrow \underset{opt\ unbiased}{\operatorname{argmin}}\ Var[\hat{\theta}_{opt\ unbiased}]$$

2.4 Relative Efficiency

To find optimal unbiased estimators, we need to compare variances between all possible types of estimators and all biases to figure out what minimizes variances. Given two estimators $\hat{\theta}_1, \hat{\theta}_2$, relative efficiency is defined as follows.

$$\text{Rel.Eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{var(\hat{\theta}_1)}{var(\hat{\theta}_2)}, \quad \text{Asymp Rel.Eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\text{Asym Var}(\hat{\theta}_1)}{\text{Asym Var}(\hat{\theta}_2)}$$

Example $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\hat{\theta}_{MLE} = \bar{X} \qquad E[\bar{X}] = \mu \Rightarrow \text{unbiased}$$
$$\hat{\theta}_{crazy} = X_1 \qquad E[X_1] = \mu \Rightarrow \text{unbiased}$$
$$\text{Rel.Eff}(\hat{\theta}_{MLE}, \hat{\theta}_{crazy}) = \frac{\text{Var}(\hat{\theta}_{MLE})}{\text{Var}(\hat{\theta}_{crazy})} = \frac{\frac{\sigma^2}{n}}{\sigma^2} = \frac{1}{n} < 1$$
$$\Rightarrow \hat{\theta}_{MLE} \text{ is more optimal}$$

Generally,
If Rel.Eff$(\hat{\theta}_1, \hat{\theta}_2) < 1$, then $\hat{\theta}_1$ is more precise because $var(\hat{\theta}_2) > var(\hat{\theta}_1)$.
If Rel.Eff$(\hat{\theta}_1, \hat{\theta}_2) > 1$, then $\hat{\theta}_2$ is more precise because $var(\hat{\theta}_2) < var(\hat{\theta}_1)$.

2.5 Cramer-Rao Lower Bound

Suppose we have population pdf_θ. If $\hat{\theta}$ is an unbiased estimator, then

$$var(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

.

2.5.1 Implication 1

If you find an unbiased estimate $\hat{\theta}$ whose var$(\hat{\theta})$ is the Cramer-Rao lower bound, you found the most precise estimator with smallest variance possible.

2.5.2 Implication 2

$$Var(\hat{\theta}_{opt}) = \frac{1}{nI(\theta)}$$

Theorem
The asymptotic MLE $\hat{\theta}$ achieves optimally
Proof
Asymptotic Var$(\hat{\theta}_{MLE}) \approx \frac{1}{nI(\theta)}$. Thus, we achieved the optimal criterion.

2.6 The Greatness of MLE

- Intuitive formulation: Find $\hat{\theta}$ that maximizes of observing data.
- Easy asymptotic variance formula: $Var(\hat{\theta}_{opt}) = \frac{1}{nI(\theta)}$
- $Var(\hat{\theta}_{opt}) = \frac{1}{nI(\theta)} \approx N(\theta, \frac{1}{nI(\theta})$ is useful
  because $\hat{\theta}_{MLE} \pm z_{\frac{1-\alpha}{2}} \sqrt{\frac{1}{nI(\theta)}}$ is $1-\alpha$ confidence interval for θ
- $\hat{\theta} \to \theta$ as $n \to \infty$ (consistency)
- Invariance if $\hat{\theta}_{MLE}$ is MLE for θ, then g$(\hat{\theta}_{MLE})$ is MLE for g(θ) for any function g.

- $\hat{\theta}_{MLE}$ is the asymptotic optimal estimate for θ.
  (e.g. $\bar{X}$ for $N(\mu, \sigma^2)$ is optimal way of estimate μ by achieving Cramer-Rao lower bound).

3 Hypothesis Testing

From the beginning of the semester, we have discovered MOM and MLE which were ways to construct point estimator $\hat{\theta}$ at a particular point for θ. From now on, we will figure out a slightly different topic: Hypothesis Testing. Hypothesis testing tests hypotheses about θ. Assuming we already know parameter θ, we are going to determine which hypothesis is correct.

Let's get back to our favorite example, measuring students' height on campus. Suppose scientist A claims that the population average height of students is 72 inches following the national trend. On the other hand, scientist B claims that the population average height of students is 74 inches because a half of the students are from Wisconsin which is well known for dairy products. In this case, hypothesis testing is to determine whose hypothesis is more plausible. Accordingly, the hypothesis test can be set up as below.

$$\begin{cases} H_0 : \mu = 72 & \text{Null Hypothesis} \\ H_1 : \mu = 74 & \text{Alternative Hypothesis} \end{cases}$$

Simple Hypothesis
A simple hypothesis is where the distribution of data is fully specified. For example,

$$pdf_\theta \sim N(\mu = 72, \sigma^2 = 3^2)$$

implies that $H_0 \& H_1$ are simple hypotheses because the distribution is given as normal with uniquely specified mean and variance.

Composite Hypothesis
Any hypothesis that is not a simple hypothesis is called a composite hypothesis. Suppose we know the distribution and mean about a hypothesis, but do not know what its variance is. Thus it is considered as a composite hypothesis.

Procedure of Hypothesis Testing

- Set up the hypothesis, collecting data n *i.i.d.* samples $X_1, ..., X_n \overset{iid}{\sim} pdf_\theta$.
  e.g.) n=3, Height $\sim N(\mu, 3^2)$, $\quad X_1 = 74, X_2 = 73, X_3 = 75$

- Construct a **decision rule** or **test statistic** to choose between H_0 and H_1 from collected sample. In order to identify more plausible hypothesis, we will see intuitively if the sample mean is closer to 72 or 74.
- If $\bar{X} > 73$, then reject H_0 in favor of H_1. Otherwise, if $\bar{X} \leq 73$, then accept H_0.
- As a result, since the sample mean turned out to be 74, we should reject H_0 for H_1.

Error Types
It might be wrong, though we made a decision based on the MLE above. Let me introduce two major errors.

- Type 1 Error: Reject H_0 even though H_0 is true
  P(Type I) = P(Reject $H_0|H_0$ is true) = P($\bar{X} > 73|H_0$ is True)
  $=P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > \frac{73-\mu}{\frac{\sigma}{\sqrt{n}}}|H_0 \text{ is true}\right) = P\left(\frac{\bar{X}-72}{\frac{3}{\sqrt{3}}} > \frac{73-72}{\frac{3}{\sqrt{3}}}|H_0 \text{ is true}\right) = 28\%$
- Type 2 Error: Accept H_0 even though H_1 is true
  P(Type II) = P(Accept $H_0|H_1$ is true) = P($\bar{X} \leq 73)|H_1$ is true)
  $= P\left(\frac{\bar{X}-74}{\frac{3}{\sqrt{3}}} \leq \frac{73-74}{\frac{3}{\sqrt{3}}}|H_1 \text{ is true}\right) = 28\%$