STAT 4352 - Mathematical Statistics Notes

JaimeGoB

February 27, 2021

1 Chapter 10.2 - Unbiased Estimators

Definition: 10.2 Unbiased Estimator
A statistic $\hat{\theta}$ is an ***Unbiased Estimator*** of parameter θ if and only if:

$$\mathrm{E}[\hat{\theta}] = \theta$$

That is $\hat{\theta}$ on average its value equals θ.

Definition: Bias
Bias of $\hat{\theta}$: $\mathrm{b}_n(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta$
When:

$\mathrm{b}_n(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta = 0$ **(Unbiased Estimator)**
$\mathrm{b}_n(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta \neq 0$ **(Biased Estimator)**

Definition: Asymptotically Unbiased Estimator
Based on a random sample n, from a given distribution. We say $\hat{\theta}$ is a ***Asymptotically Unbiased Estimator*** if and only if:

$$\lim_{n\to\infty} b_n(\hat{\theta}) = 0$$

Properties of Unbiased Estimators

- $\overline{x}$ is always unbiased for all distributions.
- NOT UNIQUE. (there can be multiple unbiased estimators).
  If you can have multiple unbiased estimators which one is best?
  Next desireable properties are sufficiency and low variance.
- Does not have invariance property.
  $\overline{x}$ is unbiased for $\mu \not\Rightarrow \overline{x}^2$ is unbiased for μ^2

2 Chapter 10.3 - Efficiency

How to measure accuracy of estimators?

1) Mean Absolute Error (MAE)
$\text{MAE}_\theta = \text{E}[\ |\hat{\theta} - \theta|\ ]$

2) Mean Absolute Deviation (MAD)
$\text{MAD}_\theta = \text{median}[\ |\hat{\theta} - \theta|\ ]$

3) Mean Squared Error (MSE)
$\text{MSE}_\theta = \text{E}(\hat{\theta} - \theta)^2 = \text{Var}_\theta(\hat{\theta}) + \text{Bias}^2_\theta(\hat{\theta})$

For an unbiased estimator (Bias = 0)
$\text{MSE}_\theta = \text{Var}_\theta(\hat{\theta})$

Definition: Relative Efficiency
Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of θ.

$$\text{If} \quad \frac{Var_\theta(\hat{\theta}_1)}{Var_\theta(\hat{\theta}_2)} < 1$$

We can say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.
You would want to pick the estimator $\hat{\theta}$ that is more efficient (lowest variance).

Efficiency Example 1: If $\dfrac{Var_\theta(\hat{\theta}_1)}{Var_\theta(\hat{\theta}_2)} = 0.50 \Longrightarrow Var_\theta(\hat{\theta}_1) = 0.5Var_\theta(\hat{\theta}_2)$

$\hat{\theta}_1$ is 50% **MORE** efficient than $\hat{\theta}_2$

Efficiency Example 2: If $\dfrac{Var_\theta(\hat{\theta}_1)}{Var_\theta(\hat{\theta}_2)} = 1.50 \Longrightarrow Var_\theta(\hat{\theta}_1) = 1.5Var_\theta(\hat{\theta}_2)$

$\hat{\theta}_1$ is 50% **LESS** efficient than $\hat{\theta}_2$
OR
$\hat{\theta}_2$ is 50% **MORE** efficient than $\hat{\theta}_1$

Definition: Asymptotic Relative Efficiency

Based on a random sample n, from a given distribution. We define the comparison of estimators ($\hat{\theta}_1$,$\hat{\theta}_2$) is ***Asymptotically Relative Efficiency*** when:

$$ARE = \lim_{n \to \infty} \frac{Var_\theta(\hat{\theta}_1)}{Var_\theta(\hat{\theta}_2)} < 1$$

The efficiency(gain) is reduced as sample size $n \to \infty$. For huge sample sizes both unbiased estimators are equally good. For small n one estimator is better than other.

Definition: Uniformly Minimum Variance Unbiased Estimator

An unbiased estimator $\hat{\theta}$ is ***Uniformly Minimum Variance Unbiased Estimator (UMVUE)*** for θ if it has the smallest variance in the class of all unbiased estimators for θ.

Theorem 10.2: Cramer-Rao Inequality

It is possible to obtain a lower bound on the variance of all ***unbiased estimators*** θ.

- $\hat{\theta}$ - unbiased estimator of parameter θ, based on a random sample of n observations.
- f(x, θ) is the probability distribution of random variable X.
- n is a random sample size

The ***Lower Bound of Variance of an Unbiased Estimator*** is the defined by the Cramer-Rao inequality:

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)} \qquad \text{where } I(\theta) = nE\left[\left(\frac{\partial}{\partial\theta} lnf(X,\theta)\right)^2\right]$$

$I(\theta)$ is the Fisher Information in a random sample of size n and $\dfrac{\partial}{\partial\theta} lnf(X,\theta)$ is known as score function. It is the smallest possible value variance can have.

UMVUE exists when:

If $Var(\hat{\theta}) = \dfrac{1}{I(\theta)}$ It has smallest possible value for variance.

$\Longrightarrow \hat{\theta}$ ***is UMVU of*** $\hat{\theta}$

UMVUE does not exists when:

If $Var(\hat{\theta}) \neq \dfrac{1}{I(\theta)}$ ***You can't say*** $\hat{\theta}$ ***is UMVUE as lower bound is not achievable.***

3 Chapter 10.4 - Consistency

Definition: Consistency

If $\hat{\theta}$ is an estimator of θ based on a random sample of size n, we say that $\hat{\theta}$ is ***consistent (closed)*** for θ, if $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 1 \qquad \theta - targetparameter, \hat{\theta} - estimator$$

ϵ - estimator (small distance ex: 0.0001)

Consistency is an Asymptotic Property:

Error in estimation using $\hat{\theta}$ is small

$\hat{\theta}$ converges in probability to θ

When $n \to \infty$ we can be practically certain that the error made with a consistent estimator will be less than any small preassigned positive constant ϵ.

Theorem 10.3

If $\hat{\theta}$ is an unbiased estimator of the parameter θ and Var($\hat{\theta}$) $\to 0$ Bias($\hat{\theta}$) $\to 0$ as $n \to \infty$ then $\hat{\theta}$ is a consistent estimator of θ.

4 Chapter 10.5 - Sufficiency

Definition: Sufficient Principle

- Reduce data without loosing information about θ.
- Captures all information about a sample relevant to estimation of θ, that is, if all the knowledge about θ that can be gained from the individual sample values and their order can just as well be gained from the value of $\hat{\theta}$ alone.

Definition: Sufficient Estimator
The statistic $\hat{\theta}$ is a sufficient estimator of parameter θ of a given distribution ***iff*** for each value of $\hat{\theta}$ ***the conditional probability distribution or density*** of a random sample $x_1, x_2, ... x_n$ given $\hat{\theta} = \theta$ ***is independent of*** θ.

Sufficient property from conditional probability distribution or density when $\hat{\theta} = \theta$:

$$f(x_1, x_2, ... x_n \, ; \hat{\theta}) = \frac{f(x_1, x_2, ... x_n, \hat{\theta})}{g(\hat{\theta})}$$

Note: Ratio should not contain θ in order to be sufficient estimator of θ

Theorem 10.4: Factorization Theorem
The statistic $\hat{\theta}$ is a sufficient estimator of the parameter θ ***iff*** the joint probability distribution or density of the random sample can be factored so that:

$$f(x_1, x_2, ... x_n \, ; \hat{\theta}) = g(\hat{\theta}, \theta) * h(x_1, x_2, ... x_n)$$

where $g(\hat{\theta}, \theta)$ depends on θ and $\hat{\theta}$ and $h(x_1, x_2, ... x_n)$ does not depend on θ.

Using factorization you want to identify:

- g function $\Longrightarrow$ function θ
- h function $\Longrightarrow$ function without θ
  (h(x) = 1 if not present)

Properties of Sufficiency:

- Complete data is always sufficient.
- Any 1-1 function of a sufficient statistic is also sufficient.
- Good estimators should be functions of sufficient statistic.
  (a good estimator is sufficient)

5 Chapter 10.8 - Method of Maximum Likelihood

Notation
$X = (X_1, X_2, ... X_n) \overset{i.i.d.}{\sim} f_\theta(x)$
Data before observed - r.v.'s with same distribution.
θ may be a vector, $\theta \in \Theta$
Θ is parameter space. Ex: parameter space of $\mathcal{N}(\mu, \sigma^2)$ is $-\infty < \mu < \infty$, $-\infty < \sigma^2 < \infty$,
$x = (x_1, x_2, ... x_n)$ Data that has been observed. (Set of sample elements)

Definition: Likelihood Function
Joint pdf/pmf of X considered as a function of θ keeping the data X fixed.

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = f(x_1, x_2, ... x_n ; \hat{\theta}) = f(x_1, \hat{\theta}) f(x_2, \hat{\theta}) ... f(x_n, \hat{\theta})$$

vary θ to find value that maximizes product, this value for θ is known as MLE.

Definition: Maximum Likelihood Estimator
The Maximum Likelihood Estimator (MLE) of θ is the value θ that maximizes the Likelihood function $\mathcal{L}(\theta)$.

- Complete data is always sufficient.
- Value of θ that maximizes $\mathcal{L}(\theta)$ also maximizes $\log \mathcal{L}(\theta) / \ln \mathcal{L}(\theta)$.
- First Derivative: $\dfrac{\partial \mathcal{L}(\theta)}{\partial \theta} \implies$ Critical Points (Max/Min)
- Second Derivative: $\dfrac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2} < 0 \implies$ Maximum exists.

Log-Likelihood Function $\log\mathcal{L}(\theta)/\ln\mathcal{L}(\theta)$
Because log/ln is a monotone function, if we ***maximize log-Likelihood it is the same as maximizing Likelihood.*** The reason why because $\log \mathcal{L}(\theta) / \ln \mathcal{L}(\theta)$ is used because taking the derivative is much easier. When referring to log we mean $\log_e = \ln$.

- $\log(ab) = \log(a) + \log(b)$
  Ex: $f(x) = \prod_{i=1}^{n} g(x_i) \implies \ln f(x) = \sum_{i=1}^{n} \ln g(x_i)$
  where $x = (x_1, x_2, ... x_n)$ Set of sample elements

This property applies to both log/ln. ***The inner product can be expressed as a sum of individual elements.*** This comes super handy when taking derivatives.

Properties of a Maximum Likelihood Estimator

- $\hat{\theta}_{MLE}$ is always a function of sufficient statistics whenever they exist.
- Optimal when n is large.
- May not be good when the distribution assumptions are wrong.
- $\hat{\theta}_{MLE} \in \Theta$ (MLE is included in parameter space)
- $\hat{\theta}_{MLE}$ has invariance property:

$$\hat{\theta} \text{ is MLE for } \theta$$
$$\Longleftrightarrow$$
$$\hat{\theta}^2 \text{ is MLE for } \theta^2$$
$$\Longleftrightarrow$$
$$\ldots\ldots\ldots$$

6 10.9 Bayesian Inference

Bayes Rule
Conditional probability can be rewritten with a Bayes Rules.

$$\text{P(A | B)} = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Difference Between Classical and Bayesian Approach
Classical Approach:

- θ is a unknown parameter and *fixed.*
- Does not have a probability distribution f(θ)

Bayesian Approach:

- θ is a random variable and *not fixed.*
- Has a probability distribution f(θ)

Prior Distribution for θ:

- Denoted f(θ) or π(θ)
- Specified before seeing data.
- Reflects personal degree of belief about what are the possible values of θ and how likely they are.
- Can be discrete or continuous.
- Can be vague: *All values equally likely*

Let data X = (X_1, X_2,...,X_n) be a random sample from a population.
The distribution of random variable X , will have a pdf/pmf: f($\mathbf{x}|\theta$) "distribution depends on θ" will have a **joint density of likelihood of sample** is:

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

Posterior Distribution of θ

We define posterior distribution of θ as the conditional distribution of θ given the sample results.

$$f(\theta|\ x_1,\ x_2,...,x_n) = \frac{f(x_1, x_2, ..., x_n; \theta)}{f(x_1, x_2, ..., x_n)} = \frac{f(x_1, x_2, ..., x_n|\theta)f(\theta)}{\int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n|\theta)f(\theta)\,d\theta}$$

where,

Joint Distribution of Sample and θ		**Joint Density of Likelihood of Sample**	**Prior**
$f(x_1, x_2, ..., x_n; \theta)$	$=$	$f(x_1,\ x_2,...,x_n|\theta)$	$f(\theta)$

Marginal Distribution of Sample (independent of θ)
Continuous Case:

$$f(x_1, x_2, ..., x_n) = \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n|\theta)f(\theta)\,d\theta$$

In summary,

*Posterior Distribution = Likelihood * Prior / Normalizing Constant*

Posterior Distribution of a parameter can be used to:

- make estimates
- make probability statements about the parameter.

Def: Conjugate Family
When prior and posterior distribution belong to the same distribution family.

NOTES:

- Because denominator of Posterior Distribution is a normalizing constant and independent of θ:

$$\frac{1}{f(x_1, x_2, ..., x_n)}$$

it is known as *proportional factor* and it will get absorbed in the $\propto$ sign.

We can summarize the Posterior Distribution to:

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

$$\textit{Posterior Distribution} \propto \quad \textit{Likelihood * Prior}$$

- The Posterior Distribution depends only sufficient statistics.

  Let T(x) be a sufficient static for θ using *factorization theorem* :

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$$

  We can conclude:

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta)$$

  factorization theorem

$$f(\theta|\mathbf{x}) \propto h(\mathbf{x})g(T(\mathbf{x})|\theta)f(\theta)$$

  h($\boldsymbol{x}$) does not depend on θ so constant gets absorbed in $\propto$

$$f(\theta|\mathbf{x}) \propto g(T(\mathbf{x})|\theta)f(\theta)$$

  $\implies$ f($\theta|\mathbf{x}$) = $f(\theta|T(\mathbf{x}))$

  Beliefs of θ having observed full data $\mathbf{x}$ are same as if we had observed only the sufficient statistic T($\mathbf{x}$).

,

Properties of Bayes Estimator

- Depends on Sufficient Statistic.
- Under certain circumstances, as n$\rightarrow \infty$ it is equivalent to MLE.
- Effect of prior diminishes as data dominates (n$\rightarrow \infty$) .
- Optimal for large n.
- Posterior Distribution can be estimated for non-conjugate priors using Markov Chains or Monte Carlo.

,

7 Chapter number - Chapter Name

Theorem number: Theorem Name