Evaluating estimators

Jacob M. Montgomery

2018

Evaluating point estimators

Overview

- In this class we will talk about point estimates from four perspectives
  - Frequentist
  - Maximum likelihood
  - Bayesian
  - Nonparametric
- But before we turn to these two, we need to establish some language
  - What is a point estimator
  - What are (some) criteria by which we can evaluate them

Big concepts in statistical estimation and learning

- A **statistical model** is a set of distributions (density functions, regression functions, etc.) that describe some data generating process.

Big concepts in statistical estimation and learning

- A **statistical model** is a set of distributions (density functions, regression functions, etc.) that describe some data generating process.
- A **parametric model** is a set of statistical models that can be parametized by a finite number of parameters.

Big concepts in statistical estimation and learning

- A **statistical model** is a set of distributions (density functions, regression functions, etc.) that describe some data generating process.
- A **parametric model** is a set of statistical models that can be parametized by a finite number of parameters.

Example: A regression model is a parametric model, where the data (Y) is assumed to be generated by the following distribution:

$$Y \sim \mathbf{X}\beta + N(0, \sigma^2)$$

Big concepts in statistical estimation and learning

- A **statistical model** is a set of distributions (density functions, regression functions, etc.) that describe some data generating process.
- A **parametric model** is a set of statistical models that can be parametized by a finite number of parameters.

Example: A regression model is a parametric model, where the data (Y) is assumed to be generated by the following distribution:

$$Y \sim \mathbf{X}\beta + N(0, \sigma^2)$$

$$Y \sim N(\mathbf{X}\beta, \sigma^2)$$

Point estimation

- **Point estimation** refers to providing a single "best guess" of some quantity of interest.

Point estimation

- **Point estimation** refers to providing a single “best guess” of some quantity of interest.
- The quantity of interest could be a parameter in a pdf, an ATE, a regression coefficient, some future event, or an entire pdf.

Point estimation

- **Point estimation** refers to providing a single “best guess” of some quantity of interest.
- The quantity of interest could be a parameter in a pdf, an ATE, a regression coefficient, some future event, or an entire pdf.
- We denote an estimate of some estimand θ by adding a “hat”, $\hat{\theta}$.

Point estimation

- **Point estimation** refers to providing a single "best guess" of some quantity of interest.
- The quantity of interest could be a parameter in a pdf, an ATE, a regression coefficient, some future event, or an entire pdf.
- We denote an estimate of some estimand θ by adding a "hat", $\hat{\theta}$.
- A point estimator is any function $g()$ that maps our data $(X_1, \ldots, X_n)$ into an estimate

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

How to evaluate a point estimator

Many students give little thought as to how estimators are chosen, but the methods you have been taught are by no means obvious.

How to evaluate a point estimator

Many students give little thought as to how estimators are chosen, but the methods you have been taught are by no means obvious. Here are some questions you might think about to motivate this discussion:

- If we are running an experiment, why don't we compare the median outcome in each group rather than the mean?
- Why do we use $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ instead of $s = \sqrt{\frac{(x-\bar{x})^2}{n}}$ to estimate the standard deviation of normally distributed data?

How to evaluate a point estimator

Many students give little thought as to how estimators are chosen, but the methods you have been taught are by no means obvious. Here are some questions you might think about to motivate this discussion:

- If we are running an experiment, why don't we compare the median outcome in each group rather than the mean?
- Why do we use $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ instead of $s = \sqrt{\frac{(x-\bar{x})^2}{n}}$ to estimate the standard deviation of normally distributed data?

In reality, the estimators we use are chosen because they are superior to alternatives in terms of:

- Bias
- Consistency
- Mean squared error
- Finite sample variance
- Efficiency

Bias

We say that $\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta$

Bias

We say that $\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta \to E(\hat{\theta}) - \theta = 0$.

Bias

We say that $\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta \rightarrow E(\hat{\theta}) - \theta = 0$.

- Remember that $\hat{\theta}$ is itself a function of the data, $\hat{\theta} = g(X_1, \ldots, X_n)$
- Thus, the expectation here is taken over $\mathbf{X}$ not θ.

Bias

We say that $\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta \rightarrow E(\hat{\theta}) - \theta = 0$.

- Remember that $\hat{\theta}$ is itself a function of the data, $\hat{\theta} = g(X_1, \ldots, X_n)$
- Thus, the expectation here is taken over $\mathbf{X}$ not θ.

Example: If we used $\hat{\theta} = \bar{x}$ to estimate the mean of normally iid varaibles $(X_1, \ldots, X_n)$, we would calculate

$$E(\hat{\theta}) = \int \bar{x} f(x) dx$$

Example 1:

Let $(X_1, \ldots, X_n)$ be iid distributed data from a normal population with mean μ and variance σ^2. Prove that $\bar{X}$ is an unbiased estimator of μ.

Example 1:

Let $(X_1, \ldots, X_n)$ be iid distributed data from a normal population with mean μ and variance σ^2. Prove that $\bar{X}$ is an unbiased estimator of μ.

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i)$$

Example 1:

Let $(X_1, \dots, X_n)$ be iid distributed data from a normal population with mean μ and variance σ^2. Prove that $\bar{X}$ is an unbiased estimator of μ.

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \int_{-\infty}^{\infty} \frac{1}{n} \sum X_i f(x) dx$$

Example 1:

Let $(X_1, \ldots, X_n)$ be iid distributed data from a normal population with mean μ and variance σ^2. Prove that $\bar{X}$ is an unbiased estimator of μ.

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \int_{-\infty}^{\infty} \frac{1}{n} \sum X_i f(x) dx$$

$$= \frac{1}{n} \sum \int_{-\infty}^{\infty} X_i f(x) dx$$

Example 1:

Let $(X_1, \ldots, X_n)$ be iid distributed data from a normal population with mean μ and variance σ^2. Prove that $\bar{X}$ is an unbiased estimator of μ.

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \int_{-\infty}^{\infty} \frac{1}{n}\sum X_i f(x)dx$$

$$= \frac{1}{n}\sum \int_{-\infty}^{\infty} X_i f(x)dx = n\frac{1}{n}\int_{-\infty}^{\infty} X_1 f(x)dx$$

Example 1:

Let $(X_1, \dots, X_n)$ be iid distributed data from a normal population with mean μ and variance σ^2. Prove that $\bar{X}$ is an unbiased estimator of μ.

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \int_{-\infty}^{\infty} \frac{1}{n}\sum X_i f(x)dx$$

$$= \frac{1}{n}\sum \int_{-\infty}^{\infty} X_i f(x)dx = n\frac{1}{n}\int_{-\infty}^{\infty} X_1 f(x)dx$$

$$= \frac{n}{n}\mu = \mu$$

Example 2:

Let $(X_1, \ldots, X_n)$ *be iid distributed data from a uniform population with distribution* $f(x) = \frac{1}{\theta}$*. A reasonable approach to estimating* θ *is to use the maximum observed value* $\hat{\theta} = max(x)$*. But is it biased?*

Example 2:

Let $(X_1, \ldots, X_n)$ be iid distributed data from a uniform population with distribution $f(x) = \frac{1}{\theta}$. A reasonable approach to estimating θ is to use the maximum observed value $\hat{\theta} = max(x)$. But is it biased?

$$P(X_{max}) = P(X_i < x, \forall i)$$

Example 2:

Let $(X_1, \ldots, X_n)$ be iid distributed data from a uniform population with distribution $f(x) = \frac{1}{\theta}$. A reasonable approach to estimating θ is to use the maximum observed value $\hat{\theta} = max(x)$. But is it biased?

$$P(X_{max}) = P(X_i < x, \forall i) = \prod_i P(X_i < x) =$$

Example 2:

Let $(X_1, \ldots, X_n)$ be iid distributed data from a uniform population with distribution $f(x) = \frac{1}{\theta}$. A reasonable approach to estimating θ is to use the maximum observed value $\hat{\theta} = max(x)$. But is it biased?

$$P(X_{max}) = P(X_i < x, \forall i) = \prod_i P(X_i < x) = \begin{cases} 1 & \text{if } x > \theta \\ (\frac{x}{\theta})^n & \text{if } 0 \leq x \leq \theta \\ 0 & \text{if } x < 0 \end{cases}$$

1. So that is the CDF. Find the pdf
2. Find the expected value of $\hat{\theta}$. Set up the bounds of integration correctly.
3. Is it biased?

1. So that is the CDF. Find the pdf
2. Find the expected value of $\hat{\theta}$. Set up the bounds of integration correctly.
3. Is it biased? $\frac{n}{n+1}\theta$

1. So that is the CDF. Find the pdf
2. Find the expected value of $\hat{\theta}$. Set up the bounds of integration correctly.
3. Is it biased? $\frac{n}{n+1}\theta$
4. What would be an unbiased estimator?

Consistency

- It turns out that many useful estimators are biased given a limited amount of data.

Consistency

- It turns out that many useful estimators are biased given a limited amount of data.
- However, one important criteria is that our estimator should generally converge to the right answer as we add more and more data.

Consistency

- It turns out that many useful estimators are biased given a limited amount of data.
- However, one important criteria is that our estimator should generally converge to the right answer as we add more and more data.
- A point estimator $\hat{\theta}$ of a parameter θ is consistent if $\hat{\theta}$ converges in probability to θ.

Consistency

- It turns out that many useful estimators are biased given a limited amount of data.
- However, one important criteria is that our estimator should generally converge to the right answer as we add more and more data.
- A point estimator $\hat{\theta}$ of a parameter θ is consistent if $\hat{\theta}$ converges in probability to θ.
- This means as $n \to \infty$, $P(|\hat{\theta} - \theta| > \epsilon) \to 0$ for every value of $\epsilon > 0$.
- This is an asymptotic (as opposed to finite sample) property of an estimator

Let $(X_1, \ldots, X_n)$ be iid distributed data from a uniform population with distribution $f(x) = \frac{1}{\theta}$. A reasonable approach to estimating θ is to use the maximum observed value $\hat{\theta} = max(x)$. But is it consistent?

1. We showed above that $E(\hat{\theta}) = \frac{n}{n+1}\theta$
2. $\lim_{n\to\infty} \frac{n}{n+1}\theta =$

Let $(X_1, \ldots, X_n)$ be iid distributed data from a uniform population with distribution $f(x) = \frac{1}{\theta}$. A reasonable approach to estimating θ is to use the maximum observed value $\hat{\theta} = max(x)$. But is it consistent?

1. We showed above that $E(\hat{\theta}) = \frac{n}{n+1}\theta$
2. $\lim_{n\to\infty} \frac{n}{n+1}\theta = \theta$

Let $(X_1, \ldots, X_n)$ be iid distributed data from a uniform population with distribution $f(x) = \frac{1}{\theta}$. A reasonable approach to estimating θ is to use the maximum observed value $\hat{\theta} = max(x)$. But is it consistent?

1. We showed above that $E(\hat{\theta}) = \frac{n}{n+1}\theta$
2. $\lim_{n\to\infty} \frac{n}{n+1}\theta = \theta$
3. Thus $\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) \to 0 \ \forall \epsilon > 0$

Mean squared error

The **mean squared error** for an estimator is defined as

$$E(\hat{\theta} - \theta)^2.$$

Mean squared error

The **mean squared error** for an estimator is defined as

$$E(\hat{\theta} - \theta)^2.$$

A nice feature of MSE is that

$$MSE = (Bias(\hat{\theta}))^2 + Var(\hat{\theta})$$

Proof (remember that the expectations and variance are in terms of X)

Let $\bar{\theta} = E(\hat{\theta})$

$$E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2$$

$$= \left(E(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta)\right)^2$$

Proof (remember that the expectations and variance are in terms of X)

Let $\bar{\theta} = E(\hat{\theta})$

$$E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2$$

$$= \Big(E(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta)\Big)^2$$

- Finish the proof
- Remember that

$$E(\hat{\theta} - \bar{\theta}) = \bar{\theta} - \bar{\theta} = 0,$$

- and $\bar{\theta} = E(\hat{\theta})$

MSE and consistency

Theorem: If bias$(\hat{\theta}) \to 0$ *and Var*$(\hat{\theta}) \to 0$ *then,* $\hat{\theta}$ *is consistent.*

MSE and consistency

Theorem: If bias$(\hat{\theta}) \to 0$ *and Var*$(\hat{\theta}) \to 0$ *then,* $\hat{\theta}$ *is consistent.*

1. If bias and variance are zero, then MSE is zero.

MSE and consistency

Theorem: If bias$(\hat{\theta}) \to 0$ *and Var*$(\hat{\theta}) \to 0$ *then,* $\hat{\theta}$ *is consistent.*

1. If bias and variance are zero, then MSE is zero.
2. It follows that $\hat{\theta}$ converges in L2 (see definition)

MSE and consistency

Theorem: If bias$(\hat{\theta}) \to 0$ *and Var*$(\hat{\theta}) \to 0$ *then,* $\hat{\theta}$ *is consistent.*

1. If bias and variance are zero, then MSE is zero.
2. It follows that $\hat{\theta}$ converges in L2 (see definition)
3. It follows that $\hat{\theta}$ converges in probability

Example

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Note that we are not assuming anything else about the distribution. Show that $\bar{X}$ is a consistent estimator of μ.

Example

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Note that we are not assuming anything else about the distribution. Show that $\bar{X}$ is a consistent estimator of μ.

You are going to help me do this proof:

1. Find the bias of $\bar{X}$.
2. Find the variance of $\bar{X}$.

Example

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Note that we are not assuming anything else about the distribution. Show that $\bar{X}$ is a consistent estimator of μ.

You are going to help me do this proof:

1. Find the bias of $\bar{X}$.
2. Find the variance of $\bar{X}$.
3. Show that $MSE \to 0$ as $n \to \infty$

Example:

Let $X_1, \ldots, X_n \sim Bern(p)$ *and let* $\hat{p} = \frac{\sum X_i}{n}$. *Show that* $\hat{p}$ *is a consistent estimator for* p

Example:

Let $X_1, \ldots, X_n \sim Bern(p)$ *and let* $\hat{p} = \frac{\sum X_i}{n}$. *Show that* $\hat{p}$ *is a consistent estimator for* p *(ignoring the fact that we have already proved this more generically).*

1. Use our results from above to easily find $E(\hat{p})$ and $Var(\hat{p})$.
2. Show what happens as $n \to \infty$.
3. Profit.

Risk and MSE

- Another way to think of MSE is in terms of risk.

Risk and MSE

- Another way to think of MSE is in terms of risk.
- We can define risk as the squared error loss of an estimator

$$\mathit{Loss}(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$$

Risk and MSE

- Another way to think of MSE is in terms of risk.
- We can define risk as the squared error loss of an estimator

$$Loss(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$$

- The risk of some estimator is then:

$$R(\theta, \hat{\theta}) = E\left(Loss(\hat{\theta}, \theta)\right)$$

- $Risk = Bias^2 + Variance$
- Often there is a bias-variance tradeoff

Risk and MSE

- Another way to think of MSE is in terms of risk.
- We can define risk as the squared error loss of an estimator

$$Loss(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$$

- The risk of some estimator is then:

$$R(\theta, \hat{\theta}) = E\left(Loss(\hat{\theta}, \theta)\right)$$

- $Risk = Bias^2 + Variance$
- Often there is a bias-variance tradeoff (especially for nonparametric statistics)
  - Overfitting the data can lead to estimators with small variances that are high in bias
  - Underfitting can lead to less biased estimates that are high in bias

Fininite variance

- MSE seems like a reasonable criteria by which to evaluate an estimator, but it is not sufficiently precise.

Fininite variance

- MSE seems like a reasonable criteria by which to evaluate an estimator, but it is not sufficiently precise.
- For example, using the estamator $\hat{\theta} = 17$ will have a very low MSE when $\theta = 17$

Fininite variance

- MSE seems like a reasonable criteria by which to evaluate an estimator, but it is not sufficiently precise.
- For example, using the estamator $\hat{\theta} = 17$ will have a very low MSE when $\theta = 17$, but will be a poor estimator for all other cases.

Fininite variance

- MSE seems like a reasonable criteria by which to evaluate an estimator, but it is not sufficiently precise.
- For example, using the estamator $\hat{\theta} = 17$ will have a very low MSE when $\theta = 17$, but will be a poor estimator for all other cases.
- Statisticians have therefore settled on focusing on *unbiased* estimators with the least MSE.

Fininite variance

- MSE seems like a reasonable criteria by which to evaluate an estimator, but it is not sufficiently precise.
- For example, using the estamator $\hat{\theta} = 17$ will have a very low MSE when $\theta = 17$, but will be a poor estimator for all other cases.
- Statisticians have therefore settled on focusing on *unbiased* estimators with the least MSE.
- This is the "best"" unbiased estimator.

Fininite variance

- MSE seems like a reasonable criteria by which to evaluate an estimator, but it is not sufficiently precise.
- For example, using the estamator $\hat{\theta} = 17$ will have a very low MSE when $\theta = 17$, but will be a poor estimator for all other cases.
- Statisticians have therefore settled on focusing on *unbiased* estimators with the least MSE.
- This is the "best"" unbiased estimator.
- Implicitly, this criteria favors estimators with low variance.

Fininite variance

- MSE seems like a reasonable criteria by which to evaluate an estimator, but it is not sufficiently precise.
- For example, using the estamator $\hat{\theta} = 17$ will have a very low MSE when $\theta = 17$, but will be a poor estimator for all other cases.
- Statisticians have therefore settled on focusing on *unbiased* estimators with the least MSE.
- This is the “best”" unbiased estimator.
- Implicitly, this criteria favors estimators with low variance. Why?

Which estimator is best?

- It turns out to be difficult to establish the MSE for all *possible* unbiased estimators.

Which estimator is best?

- It turns out to be difficult to establish the MSE for all *possible* unbiased estimators.
- Instead we will follow this strategy:
  - Establish the lower bound for the variance that *any* estimator can have.

Which estimator is best?

- It turns out to be difficult to establish the MSE for all *possible* unbiased estimators.
- Instead we will follow this strategy:
  - Establish the lower bound for the variance that *any* estimator can have.
  - Show that some candidate estimator $\hat{\theta}$ has a variance that equals this lower bound.

Which estimator is best?

- It turns out to be difficult to establish the MSE for all *possible* unbiased estimators.
- Instead we will follow this strategy:
  - Establish the lower bound for the variance that *any* estimator can have.
  - Show that some candidate estimator $\hat{\theta}$ has a variance that equals this lower bound.
- Profit.

Cramer-Rao Inequality/Information inequality

Let $X_1, \ldots, X_n$ be a be a sample with pdf $f(\mathbf{x}|\theta)$ and let $\hat{\theta}$ be any unbiased estimator such that $Var(\hat{\theta}) < \infty$.

Cramer-Rao Inequality/Information inequality

Let $X_1, \ldots, X_n$ be a be a sample with pdf $f(\mathbf{x}|\theta)$ and let $\hat{\theta}$ be any unbiased estimator such that $Var(\hat{\theta}) < \infty$.

Define the fisher information as

$$I(\theta) = E\left[\left(\frac{\partial \mathcal{L}(\theta|\mathbf{x})}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \mathcal{L}(\theta|\mathbf{x})}{\partial \theta^2}\right]$$

Cramer-Rao Inequality/Information inequality

Let $X_1, \ldots, X_n$ be a be a sample with pdf $f(\mathbf{x}|\theta)$ and let $\hat{\theta}$ be any unbiased estimator such that $Var(\hat{\theta}) < \infty$.

Define the fisher information as

$$I(\theta) = E\left[\left(\frac{\partial \mathcal{L}(\theta|\mathbf{x})}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \mathcal{L}(\theta|\mathbf{x})}{\partial \theta^2}\right]$$

Then

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

Cramer-Rao Inequality/Information inequality

Let $X_1, \ldots, X_n$ be a be a sample with pdf $f(\mathbf{x}|\theta)$ and let $\hat{\theta}$ be any unbiased estimator such that $Var(\hat{\theta}) < \infty$.

Define the fisher information as

$$I(\theta) = E\left[\left(\frac{\partial \mathcal{L}(\theta|\mathbf{x})}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \mathcal{L}(\theta|\mathbf{x})}{\partial \theta^2}\right]$$

Then

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

(Proofs of this to come later)

Fun properties of best unbiased estimators

- If $\hat{\theta}$ is a best unbiased estimator of θ, then $\hat{\theta}$ is unique (Casella Berger Theorem 7.3.19)
- The theorem does not apply in cases where the range of the pdf depends on the parameter (the scale uniform distribution discussed above).
- The equality

$$I(\theta) = E\left[\left(\frac{\partial ln(L(\theta|\mathbf{x}))}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 ln(L(\theta|\mathbf{x}))}{\partial\theta^2}\right]$$

  does not hold for all distributions, but does for the exponential family.

Example: Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let $\bar{X}$ be the sample mean. Show that $\bar{X}$ is the best unbiased estimator of λ.

Example: Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let $\bar{X}$ be the sample mean. Show that $\bar{X}$ is the best unbiased estimator of λ.

1. From the results above, we know that $E(\bar{X}) = \lambda$ since the expected value of the distribution is λ. Thus it is unbiased.
2. Recall from above that $V(\bar{X}) = \frac{\lambda}{n}$.
3. Recall that the Poisson distribution is in the exponential family.

Example: Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let $\bar{X}$ be the sample mean. Show that $\bar{X}$ is the best unbiased estimator of λ.

1. From the results above, we know that $E(\bar{X}) = \lambda$ since the expected value of the distribution is λ. Thus it is unbiased.
2. Recall from above that $V(\bar{X}) = \frac{\lambda}{n}$.
3. Recall that the Poisson distribution is in the exponential family.
4. Find the log likelihood.
5. Take the second derivative and multiply by -1.
6. Show that $1/I(\theta) = \frac{\lambda}{n}$

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. The statistic S^2 is an unbiased estimator of σ^2 (you will show in the problem set).

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. The statistic S^2 is an unbiased estimator of σ^2 (you will show in the problem set). Further,

$$E(S^2 - \sigma^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$

(assertion).

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. The statistic S^2 is an unbiased estimator of σ^2 (you will show in the problem set). Further,

$$E(S^2 - \sigma^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$

(assertion). Show that S^2 does not attain the information bound.

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. The statistic S^2 is an unbiased estimator of σ^2 (you will show in the problem set). Further,

$$E(S^2 - \sigma^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$

(assertion). Show that S^2 does not attain the information bound.

1. Find the log-likelihood.
2. Take the second derivative in terms of σ^2.

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. The statistic S^2 is an unbiased estimator of σ^2 (you will show in the problem set). Further,

$$E(S^2 - \sigma^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$

(assertion). Show that S^2 does not attain the information bound.

1. Find the log-likelihood.
2. Take the second derivative in terms of σ^2. $n(\frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6})$
3. Take the expected value and multiply by -1.

Example: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. The statistic S^2 is an unbiased estimator of σ^2 (you will show in the problem set). Further,

$$E(S^2 - \sigma^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$

(assertion). Show that S^2 does not attain the information bound.

1. Find the log-likelihood.
2. Take the second derivative in terms of σ^2. $n(\frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6})$
3. Take the expected value and multiply by -1. $(\frac{n}{2\ \sigma^4})$.
4. Show that it is greater than $(\frac{n}{2\sigma^4})^{-1}$.

Rao-Blackwell Theorem

- Let $\hat{\theta}$ be any unbiased estimator of θ, and let $T(\mathbf{X})$ be a sufficient statistic for θ.
- Define $\phi(\theta) = E(\hat{\theta}|T(\mathbf{X}))$.
- Then $E(\phi) = \theta$ and $Var(\phi) \leq Var(\hat{\theta})$ for all θ; that is, ϕ is a uniformly better unbiased estimator of θ.

Example: Let $X_1, \ldots, X_n$ be iid $Pois(\lambda)$. We choose the estimator X_1. Show how conditioning on the sufficient statistic $\sum_i X_i = T$ "Rao-Blackwellizes" the estimator.

Example: Let $X_1, \ldots, X_n$ be iid $Pois(\lambda)$. We choose the estimator X_1. Show how conditioning on the sufficient statistic $\sum_i X_i = T$ "Rao-Blackwellizes" the estimator.

1. We are trying to find $E(X_1 | \sum X_i = T)$

Example: Let $X_1, \ldots, X_n$ be iid $Pois(\lambda)$. We choose the estimator X_1. Show how conditioning on the sufficient statistic $\sum_i X_i = T$ "Rao-Blackwellizes" the estimator.

1. We are trying to find $E(X_1 | \sum X_i = T)$
2. Note that $E(X | \sum X_i)$ must be the same for all observations due to iid assumption. That is $E(X | \sum X_i) = \frac{T}{n} \; \forall X$

Example: Let $X_1, \ldots, X_n$ be iid $Pois(\lambda)$. We choose the estimator X_1. Show how conditioning on the sufficient statistic $\sum_i X_i = T$ "Rao-Blackwellizes" the estimator.

1. We are trying to find $E(X_1 | \sum X_i = T)$
2. Note that $E(X | \sum X_i)$ must be the same for all observations due to iid assumption. That is $E(X | \sum X_i) = \frac{T}{n} \ \forall X$
3. $\dfrac{T}{n} = \hat{\lambda} = \bar{X}$

Efficiency

Frustratingly enough, statisticians can mean one of two things when they talk about efficiency:

- An estimator can be considered efficient if it obtains the information lower bound (or can be considered relatively more efficient if it is closer to the bound).

Efficiency

Frustratingly enough, statisticians can mean one of two things when they talk about efficiency:

- An estimator can be considered efficient if it obtains the information lower bound (or can be considered relatively more efficient if it is closer to the bound).
- An estimator can be considered asymptotically efficient if the asymptotic variance of the estimator achieves the information lower bound.

Finite sample efficiency

- An unbiased estimator $\hat{\theta}$ is efficient for a parameter θ if

$$\frac{\frac{1}{I(\theta)}}{Var(\hat{\theta})} = 1$$

Finite sample efficiency

- An unbiased estimator $\hat{\theta}$ is efficient for a parameter θ if

$$\frac{\frac{1}{I(\theta)}}{Var(\hat{\theta})} = 1$$

- An unbiased estimator can be considered relatively more efficient if this ratio is closer to one (relative to some competing estimator).

Asyptotic efficiency

- An unbiased estimator $\hat{\theta}$ is asymptotically efficient for a parameter θ if $\sqrt{n}[\hat{\theta} - \theta] \to N(0, \nu)$ in distribution and

$$\nu = \frac{1}{I(\theta)}$$

Asyptotic efficiency

- An unbiased estimator $\hat{\theta}$ is asymptotically efficient for a parameter θ if $\sqrt{n}[\hat{\theta} - \theta] \to N(0, \nu)$ in distribution and

$$\nu = \frac{1}{I(\theta)}$$

- We will return to this topic when we discuss maximum likelihood estimators next class.