



Topic 4: Sufficiency, and MVUEs

Ethan P. Marzban University of California, Santa Barbara PSTAT 120B



Outline

1. Sufficiency

2. MVUEs

Sufficiency



Leadup

- Perhaps you've noticed that certain quantities arise repeatedly in the context of estimating certain parameters.



Leadup

- Perhaps you've noticed that certain quantities arise repeatedly in the context of estimating certain parameters.
- For example, when estimating a *population* mean μ (using either the method of moments or maximum likelihood estimation), the *sample* mean \bar{Y}_n appears often.



Leadup

- Perhaps you've noticed that certain quantities arise repeatedly in the context of estimating certain parameters.
- For example, when estimating a *population* mean μ (using either the method of moments or maximum likelihood estimation), the *sample* mean \bar{Y}_n appears often.
- When estimating the population variance of a zero-mean distribution, the quantity $\sum_{i=1}^n Y_i^2$ arises frequently.



Leadup

- Perhaps you've noticed that certain quantities arise repeatedly in the context of estimating certain parameters.
- For example, when estimating a *population* mean μ (using either the method of moments or maximum likelihood estimation), the *sample* mean \bar{Y}_n appears often.
- When estimating the population variance of a zero-mean distribution, the quantity $\sum_{i=1}^n Y_i^2$ arises frequently.
- As such, let's take a brief **break** from estimation and return back to the general notion of a **statistic**.



Statistics

Definition (Statistic)

Given a random sample $\vec{Y} = \{Y_i\}_{i=1}^n$, a **statistic** T is simply a function of \vec{Y} :

$$T := T(\vec{Y}) = T(Y_1, \dots, Y_n)$$



Statistics

Definition (Statistic)

Given a random sample $\vec{Y} = \{Y_i\}_{i=1}^n$, a **statistic** T is simply a function of \vec{Y} :

$$T := T(\vec{Y}) = T(Y_1, \dots, Y_n)$$

- Example: sample mean $\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$



Statistics

Definition (Statistic)

Given a random sample $\vec{Y} = \{Y_i\}_{i=1}^n$, a **statistic** T is simply a function of \vec{Y} :

$$T := T(\vec{Y}) = T(Y_1, \dots, Y_n)$$

- Example: sample mean $\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$
- Example: sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$



Statistics

Definition (Statistic)

Given a random sample $\vec{Y} = \{Y_i\}_{i=1}^n$, a **statistic** T is simply a function of \vec{Y} :

$$T := T(\vec{Y}) = T(Y_1, \dots, Y_n)$$

- Example: sample mean $\bar{Y}_n := \frac{1}{n} \sum_{i=1}^n Y_i$
- Example: sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$
- Example: sample maximum $Y_{(n)}$



Statistics as Data Reduction

- A statistic, inherently, is a form of *data reduction*.



Statistics as Data Reduction

- A statistic, inherently, is a form of *data reduction*.
- That is, we take a sample \vec{Y} consisting of n elements (i.e. observations) and *reduce* it to a single quantity (like the mean, variance, maximum, etc.).



Statistics as Data Reduction

- A statistic, inherently, is a form of *data reduction*.
- That is, we take a sample \vec{Y} consisting of n elements (i.e. observations) and *reduce* it to a single quantity (like the mean, variance, maximum, etc.).
 - Again, this is just a more heuristic way of saying that a statistic is a *function* of our sample!



Statistics as Data Reduction

- A statistic, inherently, is a form of *data reduction*.
- That is, we take a sample \vec{Y} consisting of n elements (i.e. observations) and *reduce* it to a single quantity (like the mean, variance, maximum, etc.).
 - Again, this is just a more heuristic way of saying that a statistic is a *function* of our sample!
- For this reason, statistics are sometimes referred to as **summary statistics**, as they *summarize* our sample in some way (e.g. summarize where the “center” of our sample is, summarize how “spread out” our sample is, etc.)



Leadup

- Intuitively (as was mentioned at the beginning of this lecture), the sample mean seems like a pretty good proxy for the population mean.



Leadup

- Intuitively (as was mentioned at the beginning of this lecture), the sample mean seems like a pretty good proxy for the population mean.
- Conversely, the sample variance might not give us a lot of information about the population mean (unless we have a very specific distribution).



Leadup

- Intuitively (as was mentioned at the beginning of this lecture), the sample mean seems like a pretty good proxy for the population mean.
- Conversely, the sample variance might not give us a lot of information about the population mean (unless we have a very specific distribution).
- So, our intuition is telling us that the sample mean is doing a better job of summarizing information about μ (the population mean) than the sample variance.



Leadup

- Intuitively (as was mentioned at the beginning of this lecture), the sample mean seems like a pretty good proxy for the population mean.
- Conversely, the sample variance might not give us a lot of information about the population mean (unless we have a very specific distribution).
- So, our intuition is telling us that the sample mean is doing a better job of summarizing information about μ (the population mean) than the sample variance.
- Can we make this more explicit?



Leadup

- Well, the answer is “yes” and we’ve actually taken some pretty good steps to making our intuition more explicit, by way of estimation!



Leadup

- Well, the answer is “yes” and we’ve actually taken some pretty good steps to making our intuition more explicit, by way of estimation!
- Said differently, used as an estimator for μ , \bar{Y}_n possess *many* more desirable properties than, say, S_n^2 .



Leadup

- Well, the answer is “yes” and we’ve actually taken some pretty good steps to making our intuition more explicit, by way of estimation!
- Said differently, used as an estimator for μ , \bar{Y}_n possess *many* more desirable properties than, say, S_n^2 .
 - For example, \bar{Y}_n is an unbiased estimator for μ whereas S_n^2 is, in general, not.



Leadup

- Well, the answer is “yes” and we’ve actually taken some pretty good steps to making our intuition more explicit, by way of estimation!
- Said differently, used as an estimator for μ , \bar{Y}_n possess *many* more desirable properties than, say, S_n^2 .
 - For example, \bar{Y}_n is an unbiased estimator for μ whereas S_n^2 is, in general, not.
 - Similarly, \bar{Y}_n is a consistent estimator for μ whereas S_n^2 is, in general, not.



Leadup

- Well, the answer is “yes” and we’ve actually taken some pretty good steps to making our intuition more explicit, by way of estimation!
- Said differently, used as an estimator for μ , \bar{Y}_n possess *many* more desirable properties than, say, S_n^2 .
 - For example, \bar{Y}_n is an unbiased estimator for μ whereas S_n^2 is, in general, not.
 - Similarly, \bar{Y}_n is a consistent estimator for μ whereas S_n^2 is, in general, not.
- But let’s see if there’s perhaps a *different* way to quantify our intuitions.



Example

- This is all very abstract - let's make things more concrete.



Example

- This is all very abstract - let's make things more concrete.
- Specifically, suppose $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$.



Example

- This is all very abstract - let's make things more concrete.
- Specifically, suppose $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$.
 - In other words, you can imagine Y_i to be the outcome of tossing a coin once and observing whether it landed on heads or tails, where θ represents the probability the coin will land "heads" on any particular toss.



Example

- This is all very abstract - let's make things more concrete.
- Specifically, suppose $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$.
 - In other words, you can imagine Y_i to be the outcome of tossing a coin once and observing whether it landed on heads or tails, where θ represents the probability the coin will land "heads" on any particular toss.
- One statistic we could consider is $U := \sum_{i=1}^n Y_i$.



Example

- This is all very abstract - let's make things more concrete.
- Specifically, suppose $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$.
 - In other words, you can imagine Y_i to be the outcome of tossing a coin once and observing whether it landed on heads or tails, where θ represents the probability the coin will land "heads" on any particular toss.
- One statistic we could consider is $U := \sum_{i=1}^n Y_i$.
 - In words, U denotes the number of heads in the n coin tosses.



Example

- This is all very abstract - let's make things more concrete.
- Specifically, suppose $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$.
 - In other words, you can imagine Y_i to be the outcome of tossing a coin once and observing whether it landed on heads or tails, where θ represents the probability the coin will land "heads" on any particular toss.
- One statistic we could consider is $U := \sum_{i=1}^n Y_i$.
 - In words, U denotes the number of heads in the n coin tosses.
- Does U capture the maximal amount of information about θ ? That is, can we gain any further information about θ by looking at other statistics?



Example

- Here is one way to answer this question: let's look at the distribution of $(Y_1, \dots, Y_n \mid U)$.



Example

- Here is one way to answer this question: let's look at the distribution of $(Y_1, \dots, Y_n \mid U)$.
- Before we do, let's convince ourselves that examining this distribution is a good idea.



Example

- Here is one way to answer this question: let's look at the distribution of $(Y_1, \dots, Y_n \mid U)$.
- Before we do, let's convince ourselves that examining this distribution is a good idea.
- If the distribution of $(Y_1, \dots, Y_n \mid U)$ does not depend on θ , then, in essence, U will have captured all of the necessary information about θ .



Example

- Here is one way to answer this question: let's look at the distribution of $(Y_1, \dots, Y_n \mid U)$.
- Before we do, let's convince ourselves that examining this distribution is a good idea.
- If the distribution of $(Y_1, \dots, Y_n \mid U)$ does not depend on θ , then, in essence, U will have captured all of the necessary information about θ .
 - Remember that the distribution of $(X \mid Y)$ can be interpreted as our beliefs on X after knowing Y .



Example

- Here is one way to answer this question: let's look at the distribution of $(Y_1, \dots, Y_n \mid U)$.
- Before we do, let's convince ourselves that examining this distribution is a good idea.
- If the distribution of $(Y_1, \dots, Y_n \mid U)$ does not depend on θ , then, in essence, U will have captured all of the necessary information about θ .
 - Remember that the distribution of $(X \mid Y)$ can be interpreted as our beliefs on X after knowing Y .
 - Saying that the distribution of $(Y_1, \dots, Y_n \mid U)$ doesn't depend on θ means, after knowing U , our beliefs on (Y_1, \dots, Y_n) no longer depend on θ .



Example

- Alright, let's go!



Example

- Alright, let's go!
- Specifically, we examine $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n \mid U = u)$.



Example

- Alright, let's go!
- Specifically, we examine $\mathbb{P}(Y_1 = y_1, \dots, Y_1 = y_n \mid U = u)$.
- We're conditioning on an event with nonzero probability, meaning we can invoke the definition of conditional probability to write

$$\mathbb{P}(Y_1 = y_1, \dots, Y_1 = y_n \mid U = u) = \frac{\mathbb{P}(Y_1 = y_1, \dots, Y_1 = y_n, U = u)}{\mathbb{P}(U = u)}$$



Example

- Alright, let's go!
- Specifically, we examine $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n \mid U = u)$.
- We're conditioning on an event with nonzero probability, meaning we can invoke the definition of conditional probability to write

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n \mid U = u) = \frac{\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u)}{\mathbb{P}(U = u)}$$

- Since $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$, we know that $U := (\sum_{i=1}^n Y_i) \sim \text{Bin}(n, p)$, meaning

$$\mathbb{P}(U = u) = \binom{n}{u} p^u (1-p)^{n-u}$$



Example

- What about the numerator, $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u)$?



Example

- What about the numerator, $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u)$?
- Well, if $\sum_{i=1}^n y_i \neq u$, the probability is zero.



Example

- What about the numerator, $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u)$?
- Well, if $\sum_{i=1}^n y_i \neq u$, the probability is zero.
 - Here's how we can think through this: say $n = 3$, and $y_1 = 1$, $y_2 = 0$, $y_3 = 3$. (That is, the first coin landed heads, the second landed tails, and the third landed tails).



Example

- What about the numerator, $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u)$?
- Well, if $\sum_{i=1}^n y_i \neq u$, the probability is zero.
 - Here's how we can think through this: say $n = 3$, and $y_1 = 1$, $y_2 = 0$, $y_3 = 3$. (That is, the first coin landed heads, the second landed tails, and the third landed tails).
 - What's the probability of the first coin landing heads, the second landing tails, the third landing tails, and observing a total number of heads that is not equal to 1 (i.e. $1 + 0 + 0$)?



Example

- What about the numerator, $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u)$?
- Well, if $\sum_{i=1}^n y_i \neq u$, the probability is zero.
 - Here's how we can think through this: say $n = 3$, and $y_1 = 1$, $y_2 = 0$, $y_3 = 3$. (That is, the first coin landed heads, the second landed tails, and the third landed tails).
 - What's the probability of the first coin landing heads, the second landing tails, the third landing tails, and observing a total number of heads that is not equal to 1 (i.e. $1 + 0 + 0$)?
 - The answer is zero!



Example

- If $\sum_{i=1}^n y_i = u$, the event we're taking the probability of is

$$\{Y_1 = y_1, \dots, Y_n = y_n, U = u\}$$

which is just the probability of an independent sequences of zeros and ones with a total of u ones and $(n - u)$ zeroes.



Example

- If $\sum_{i=1}^n y_i = u$, the event we're taking the probability of is

$$\{Y_1 = y_1, \dots, Y_n = y_n, U = u\}$$

which is just the probability of an independent sequences of zeros and ones with a total of u ones and $(n - u)$ zeroes.

- That is,

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u) = \theta^u (1 - \theta)^{n-u}$$



Example

- If $\sum_{i=1}^n y_i = u$, the event we're taking the probability of is

$$\{Y_1 = y_1, \dots, Y_n = y_n, U = u\}$$

which is just the probability of an independent sequences of zeros and ones with a total of u ones and $(n - u)$ zeroes.

- That is,

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u) = \theta^u (1 - \theta)^{n-u}$$

- So, in all,

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u) = \begin{cases} \theta^u (1 - \theta)^{n-u} & \text{if } \sum_{i=1}^n y_i = u \\ 0 & \text{otherwise} \end{cases}$$



Example

- Therefore, dividing by $\mathbb{P}(U = u) = \binom{n}{u} \theta^u (1 - \theta)^{n-u}$, we have

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u) = \begin{cases} \frac{1}{\binom{n}{u}} & \text{if } \sum_{i=1}^n y_i = u \\ 0 & \text{otherwise} \end{cases}$$



Example

- Therefore, dividing by $\mathbb{P}(U = u) = \binom{n}{u} \theta^u (1 - \theta)^{n-u}$, we have

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u) = \begin{cases} \frac{1}{\binom{n}{u}} & \text{if } \sum_{i=1}^n y_i = u \\ 0 & \text{otherwise} \end{cases}$$

- So, does this distribution depend on θ ?



Example

- Therefore, dividing by $\mathbb{P}(U = u) = \binom{n}{u} \theta^u (1 - \theta)^{n-u}$, we have

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, U = u) = \begin{cases} \frac{1}{\binom{n}{u}} & \text{if } \sum_{i=1}^n y_i = u \\ 0 & \text{otherwise} \end{cases}$$

- So, does this distribution depend on θ ?
- Nope! So, after conditioning on $U := \sum_{i=1}^n Y_i$, we have removed all dependency on θ - said differently, U has captured all of the necessary information about θ .



Sufficiency

Definition (Sufficiency)

Let Y_1, \dots, Y_n denote a random sample from a distribution with parameter θ . A statistic $U := g(Y_1, \dots, Y_n)$ is said to be **sufficient** for θ if the conditional distribution $(Y_1, \dots, Y_n \mid U)$ does not depend on θ .



Sufficiency

- Now, we almost never use the definition of sufficiency.



Sufficiency

- Now, we almost never use the definition of sufficiency.
- Firstly, it only allows us to check whether a given statistic is sufficient - not how to actually *find* a sufficient statistic.



Sufficiency

- Now, we almost never use the definition of sufficiency.
- Firstly, it only allows us to check whether a given statistic is sufficient - not how to actually *find* a sufficient statistic.
- Furthermore, it requires us to find conditional distributions which are, in general, not particularly easy to find.



Sufficiency

- Now, we almost never use the definition of sufficiency.
- Firstly, it only allows us to check whether a given statistic is sufficient - not how to actually *find* a sufficient statistic.
- Furthermore, it requires us to find conditional distributions which are, in general, not particularly easy to find.
- As such, in practice, we rely more heavily on the following theorem:



Factorization Theorem

Theorem (Factorization Theorem)

Let U be a statistic based on the random sample $\vec{Y} = (Y_1, \dots, Y_n)$. Then U is a sufficient statistic for the estimation of a parameter θ if and only if the likelihood $\mathcal{L}_{\vec{Y}}(\theta)$ factors as

$$\mathcal{L}_{\vec{Y}}(\theta) = g(U, \theta) \times h(\vec{Y})$$

where $g(U, \theta)$ is a function of only U and θ (and possibly fundamental constants) and $h(\vec{Y})$ does *not* depend on θ .



Example

Example

Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$, where $\theta \in (0, 1)$ is an unknown parameter. Show that $U := \sum_{i=1}^n Y_i$ is a sufficient statistic for θ .



Example

Example

Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$, where $\theta \in (0, 1)$ is an unknown parameter. Show that $U := \sum_{i=1}^n Y_i$ is a sufficient statistic for θ .

- We've actually already shown this using the definition of sufficiency (at the start of today's lecture) - let's show this again, this time using the Factorization Theorem.



Solution

$$\mathcal{L}_{\bar{\mathbf{Y}}}(\theta) = \prod_{i=1}^n p(Y_i; \theta) = \prod_{i=1}^n [\theta^{Y_i} (1 - \theta)^{1-Y_i}]$$



Solution

$$\begin{aligned}\mathcal{L}_{\bar{\mathbf{Y}}}(\theta) &= \prod_{i=1}^n p(Y_i; \theta) = \prod_{i=1}^n [\theta^{Y_i} (1 - \theta)^{1-Y_i}] \\ &= \theta^{\sum_{i=1}^n Y_i} \cdot (1 - \theta)^{n - \sum_{i=1}^n Y_i}\end{aligned}$$



Solution

$$\begin{aligned}\mathcal{L}_{\vec{Y}}(\theta) &= \prod_{i=1}^n p(Y_i; \theta) = \prod_{i=1}^n [\theta^{Y_i} (1 - \theta)^{1-Y_i}] \\ &= \theta^{\sum_{i=1}^n Y_i} \cdot (1 - \theta)^{n - \sum_{i=1}^n Y_i} \\ &= \underbrace{\left[\theta^{\sum_{i=1}^n Y_i} \cdot (1 - \theta)^{n - \sum_{i=1}^n Y_i} \right]}_{:=g(\sum_{i=1}^n Y_i, \theta)} \times \underbrace{[1]}_{:=h(\vec{Y})}\end{aligned}$$



Solution

$$\begin{aligned}\mathcal{L}_{\vec{Y}}(\theta) &= \prod_{i=1}^n p(Y_i; \theta) = \prod_{i=1}^n [\theta^{Y_i} (1 - \theta)^{1-Y_i}] \\ &= \theta^{\sum_{i=1}^n Y_i} \cdot (1 - \theta)^{n - \sum_{i=1}^n Y_i} \\ &= \underbrace{\left[\theta^{\sum_{i=1}^n Y_i} \cdot (1 - \theta)^{n - \sum_{i=1}^n Y_i} \right]}_{:=g(\sum_{i=1}^n Y_i, \theta)} \times \underbrace{[1]}_{:=h(\vec{Y})}\end{aligned}$$

where $g(U, \theta) = \theta^U \cdot (1 - \theta)^{n-U}$ and $h(\vec{Y}) = 1$.



Solution

$$\begin{aligned}\mathcal{L}_{\vec{Y}}(\theta) &= \prod_{i=1}^n p(Y_i; \theta) = \prod_{i=1}^n [\theta^{Y_i} (1 - \theta)^{1 - Y_i}] \\ &= \theta^{\sum_{i=1}^n Y_i} \cdot (1 - \theta)^{n - \sum_{i=1}^n Y_i} \\ &= \underbrace{\left[\theta^{\sum_{i=1}^n Y_i} \cdot (1 - \theta)^{n - \sum_{i=1}^n Y_i} \right]}_{:=g(\sum_{i=1}^n Y_i, \theta)} \times \underbrace{[1]}_{:=h(\vec{Y})}\end{aligned}$$

where $g(U, \theta) = \theta^U \cdot (1 - \theta)^{n - U}$ and $h(\vec{Y}) = 1$. Therefore, by the Factorization Theorem, $U := \sum_{i=1}^n Y_i$ is a sufficient statistic for θ .



Example

Example

Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\theta)$, where $\theta > 0$ is an unknown parameter. Propose a sufficient statistic for θ , and show that it is sufficient.



Example

Example

Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\theta)$, where $\theta > 0$ is an unknown parameter. Propose a sufficient statistic for θ , and show that it is sufficient.

- We'll do this one on the board.



Questions (to be answered together)

- **Question:** are sufficient statistics unique?



Questions (to be answered together)

- **Question:** are sufficient statistics unique?
- **Question:** do sufficient statistics always exist?



Questions (to be answered together)

- **Question:** are sufficient statistics unique?
- **Question:** do sufficient statistics always exist?
- Let's discuss!

MVUEs



Leadup

- Alright, let's dip our toes back into the realm of estimation.



Leadup

- Alright, let's dip our toes back into the realm of estimation.
- Recall that, a few lectures ago, I tried to convince everyone that one notion of an “ideal” estimator should be unbiased and with as little variance as possible.



Leadup

- Alright, let's dip our toes back into the realm of estimation.
- Recall that, a few lectures ago, I tried to convince everyone that one notion of an “ideal” estimator should be unbiased and with as little variance as possible.
- Let's run with this idea a bit!



Leadup

- Alright, let's dip our toes back into the realm of estimation.
- Recall that, a few lectures ago, I tried to convince everyone that one notion of an “ideal” estimator should be unbiased and with as little variance as possible.
- Let's run with this idea a bit!
- Indeed, we have the notion of a **Minimum Variance Unbiased Estimator** (MVUE) as a sort of “gold-standard” estimator.



Leadup

- Alright, let's dip our toes back into the realm of estimation.
- Recall that, a few lectures ago, I tried to convince everyone that one notion of an “ideal” estimator should be unbiased and with as little variance as possible.
- Let's run with this idea a bit!
- Indeed, we have the notion of a **Minimum Variance Unbiased Estimator** (MVUE) as a sort of “gold-standard” estimator.
- As the name suggests, an MVUE is an estimator that is unbiased and possesses the smallest possible variance.



Leadup

- “Smallest possible variance.-” is it possible to get an unbiased estimator with zero variance?



Leadup

- “Smallest possible variance.-” is it possible to get an unbiased estimator with zero variance?
- It turns out (and the reasoning behind *why* is outside the scope of this course) the answer is, in general, “no.”



Leadup

- “Smallest possible variance.-” is it possible to get an unbiased estimator with zero variance?
- It turns out (and the reasoning behind *why* is outside the scope of this course) the answer is, in general, “no.”
- Indeed, there exists a lower bound for the variance of *any* unbiased estimator, called the **Cramér-Rao Lower Bound** (CRLB).



Cramér-Rao Lower Bound

Theorem (Cramér-Rao Lower Bound)

Consider an i.i.d. sample Y_1, \dots, Y_n from a distribution with unknown parameter θ . Under appropriate “regularity conditions”, every unbiased estimator $\hat{\theta}$ obeys the inequality

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

where

$$\mathcal{I}_n(\theta) = \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{\mathbf{Y}}}(\theta) \right]$$



Some Terminology

- The Cramér-Rao Lower Bound refers to the lower bound on the variance, $[\mathcal{I}_n(\theta)]^{-1}$.



Some Terminology

- The Cramér-Rao Lower Bound refers to the lower bound on the variance, $[\mathcal{I}_n(\theta)]^{-1}$.
- The term $\mathcal{I}_n(\theta)$ is referred to as the **Fisher Information** of the sample \vec{Y} . Note that the fisher information is the expectation of the negative second-derivative of the log-likelihood of the sample.



Some Terminology

- The Cramér-Rao Lower Bound refers to the lower bound on the variance, $[\mathcal{I}_n(\theta)]^{-1}$.
- The term $\mathcal{I}_n(\theta)$ is referred to as the **Fisher Information** of the sample \vec{Y} . Note that the fisher information is the expectation of the negative second-derivative of the log-likelihood of the sample.
- Note that the CRLB is not a strict inequality, meaning that certain estimators actually achieve the lower bound. An estimator that achieves the CRLB (i.e. an estimator satisfying $\text{Var}(\hat{\theta}) = [\mathcal{I}_n(\theta)]^{-1}$) is said to be a **efficient** estimator.



A Note

- The Cramér-Rao Lower Bound only applies to *unbiased* estimators. It is possible to construct *biased* estimators that have variance smaller than the CRLB (a very popular example of such an estimator, used throughout a wide array of different disciplines, is the so-called “James-Stein estimator”)



Example

Example

Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\theta)$, where $\theta > 0$ is an unknown parameter.

- (a) Find the lowest attainable variance by an unbiased estimator for θ .
- (b) Is the estimator $\hat{\theta}_n := \bar{Y}_n$ an efficient estimator for θ ?



Solutions

- Part (a) is essentially just asking us to compute the CRLB.



Solutions

- Part (a) is essentially just asking us to compute the CRLB.
- From previous work, we have that the log-likelihood of the sample is given by

$$\ell_{\mathbf{Y}}(\theta) = -n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^n Y_i + \sum_{i=1}^n \ln \mathbb{1}_{\{Y_i \geq 0\}}$$



Solutions

- Part (a) is essentially just asking us to compute the CRLB.
- From previous work, we have that the log-likelihood of the sample is given by

$$\ell_{\mathbf{Y}}(\theta) = -n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^n Y_i + \sum_{i=1}^n \ln \mathbb{1}_{\{Y_i \geq 0\}}$$

- We now take the first and second derivatives:



Solutions



Solutions

$$\frac{\partial}{\partial \theta} \ell_{\vec{y}}(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n Y_i$$



Solutions

$$\frac{\partial}{\partial \theta} \ell_{\vec{y}}(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n Y_i$$

$$\frac{\partial^2}{\partial \theta^2} \ell_{\vec{y}}(\theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n Y_i$$



Solutions

$$\frac{\partial}{\partial \theta} \ell_{\vec{y}}(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n Y_i$$

$$\frac{\partial^2}{\partial \theta^2} \ell_{\vec{y}}(\theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n Y_i$$

$$-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{y}}(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n Y_i$$



Solutions

$$\frac{\partial}{\partial \theta} \ell_{\vec{y}}(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n Y_i$$

$$\frac{\partial^2}{\partial \theta^2} \ell_{\vec{y}}(\theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n Y_i$$

$$-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{y}}(\theta) = -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n Y_i$$

- The Fisher Information is just the expectation of this last quantity:



Solutions



Solutions

$$\mathcal{I}_n(\theta) = \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{\mathbf{Y}}}(\theta) \right]$$



Solutions

$$\mathcal{I}_n(\theta) = \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{\mathbf{Y}}}(\theta) \right]$$



Solutions

$$\begin{aligned}\mathcal{I}_n(\theta) &= \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{Y}}(\theta) \right] \\ &= \mathbb{E} \left[-\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n Y_i \right]\end{aligned}$$



Solutions

$$\begin{aligned}\mathcal{I}_n(\theta) &= \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{Y}}(\theta) \right] \\ &= \mathbb{E} \left[-\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n Y_i \right] \\ &= -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n \mathbb{E}[Y_i]\end{aligned}$$



Solutions

$$\begin{aligned}\mathcal{I}_n(\theta) &= \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{Y}}(\theta) \right] \\ &= \mathbb{E} \left[-\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n Y_i \right] \\ &= -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n \mathbb{E}[Y_i] = -\frac{n}{\theta^2} + \frac{2n}{\theta^2} = \frac{n}{\theta^2}\end{aligned}$$



Solutions

$$\begin{aligned}\mathcal{I}_n(\theta) &= \mathbb{E} \left[-\frac{\partial^2}{\partial \theta^2} \ell_{\vec{\mathbf{Y}}}(\theta) \right] \\ &= \mathbb{E} \left[-\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n Y_i \right] \\ &= -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^n \mathbb{E}[Y_i] = -\frac{n}{\theta^2} + \frac{2n}{\theta^2} = \frac{n}{\theta^2}\end{aligned}$$

- The CRLB is just the reciprocal of this last quantity: θ^2/n .



Solutions

- So, in other words, *any* unbiased estimator for θ (in the context of the exponential distribution) will have variance greater than or equal to θ^2/n .



Solutions

- So, in other words, *any* unbiased estimator for θ (in the context of the exponential distribution) will have variance greater than or equal to θ^2/n .
- To answer part (b), first note that $\hat{\theta}_n := \bar{Y}_n$ is an unbiased estimator for θ . Hence, we simply need to check whether or not its variance attains the CRLB:

$$\text{Var}(\hat{\theta}_n) = \text{Var}(\bar{Y}_n) = \frac{\text{Var}(Y_1)}{n} = \frac{\theta^2}{n}$$



Solutions

- So, in other words, *any* unbiased estimator for θ (in the context of the exponential distribution) will have variance greater than or equal to θ^2/n .
- To answer part (b), first note that $\hat{\theta}_n := \bar{Y}_n$ is an unbiased estimator for θ . Hence, we simply need to check whether or not its variance attains the CRLB:

$$\text{Var}(\hat{\theta}_n) = \text{Var}(\bar{Y}_n) = \frac{\text{Var}(Y_1)}{n} = \frac{\theta^2}{n}$$

- Since this is exactly equal to the CRLB, we conclude that \bar{Y}_n is a **efficient** estimator for θ .



Solutions

- Finally, let's try and tie the notion of efficiency back to our initial discussions on MVUEs.



Solutions

- Finally, let's try and tie the notion of efficiency back to our initial discussions on MVUEs.
- First note: perhaps counterintuitively, it's possible that the MVUE in a given situation *won't* be efficient. We won't worry too much about why that is, for the purposes of this class.



Solutions

- Finally, let's try and tie the notion of efficiency back to our initial discussions on MVUEs.
- First note: perhaps counterintuitively, it's possible that the MVUE in a given situation *won't* be efficient. We won't worry too much about why that is, for the purposes of this class.
- I would, however, like to stress that we would like to construct an unbiased estimator that has as low variance as possible.



Solutions

- Finally, let's try and tie the notion of efficiency back to our initial discussions on MVUEs.
- First note: perhaps counterintuitively, it's possible that the MVUE in a given situation *won't* be efficient. We won't worry too much about why that is, for the purposes of this class.
- I would, however, like to stress that we would like to construct an unbiased estimator that has as low variance as possible.
- So, given an estimator $\hat{\theta}_1$ for a parameter θ , is it possible to “improve” (i.e. obtain a new estimator $\hat{\theta}_2$ with a lower variance than $\hat{\theta}_1$?)



Solutions

- Finally, let's try and tie the notion of efficiency back to our initial discussions on MVUEs.
- First note: perhaps counterintuitively, it's possible that the MVUE in a given situation *won't* be efficient. We won't worry too much about why that is, for the purposes of this class.
- I would, however, like to stress that we would like to construct an unbiased estimator that has as low variance as possible.
- So, given an estimator $\hat{\theta}_1$ for a parameter θ , is it possible to “improve” (i.e. obtain a new estimator $\hat{\theta}_2$ with a lower variance than $\hat{\theta}_1$?) Yes!



Rao-Blackwell Theorem

Theorem (Rao-Blackwell Theorem)

Let $\hat{\theta}_1$ be an unbiased estimator for θ with finite variance. If U is a sufficient statistic for θ , define $\hat{\theta}_2 := \mathbb{E}[\hat{\theta}_1 | U]$. Then, for all θ ,

$$\mathbb{E}[\hat{\theta}_2] = \theta \quad \text{and} \quad \text{Var}(\hat{\theta}_2) \leq \text{Var}(\hat{\theta}_1)$$



Rao-Blackwell Theorem

Theorem (Rao-Blackwell Theorem)

Let $\hat{\theta}_1$ be an unbiased estimator for θ with finite variance. If U is a sufficient statistic for θ , define $\hat{\theta}_2 := \mathbb{E}[\hat{\theta}_1 | U]$. Then, for all θ ,

$$\mathbb{E}[\hat{\theta}_2] = \theta \quad \text{and} \quad \text{Var}(\hat{\theta}_2) \leq \text{Var}(\hat{\theta}_1)$$

- So, given an initial unbiased estimator $\hat{\theta}_1$ and a sufficient statistic U , we can “improve” (or, at least, never do worse) by conditioning our unbiased estimator on our sufficient statistic.



Rao-Blackwell Theorem

- Now, in practice, using the Rao-Blackwell theorem can be a bit tricky, mainly due to the intractability of some of the conditional expectations it requires us to compute.



Rao-Blackwell Theorem

- Now, in practice, using the Rao-Blackwell theorem can be a bit tricky, mainly due to the intractability of some of the conditional expectations it requires us to compute.
 - I walk you through one particular example in problem 4 of your HW05



Rao-Blackwell Theorem

- Now, in practice, using the Rao-Blackwell theorem can be a bit tricky, mainly due to the intractability of some of the conditional expectations it requires us to compute.
 - I walk you through one particular example in problem 4 of your HW05
- However, the Rao-Blackwell Theorem can be used to tell us that the following procedure generally gives us an MVUE:



Rao-Blackwell Theorem

- Now, in practice, using the Rao-Blackwell theorem can be a bit tricky, mainly due to the intractability of some of the conditional expectations it requires us to compute.
 - I walk you through one particular example in problem 4 of your HW05
- However, the Rao-Blackwell Theorem can be used to tell us that the following procedure generally gives us an MVUE:
- Say we have a sufficient statistic U that best summarizes our data. Additionally, say we have an estimator $\hat{\theta} := h(U)$ that is unbiased for θ . Then, typically, $\hat{\theta}$ will be an MVUE.



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)
 - So, it’s really a function of a *minimal* sufficient statistic that will give us the MVUE in a given situation.



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)
 - So, it’s really a function of a *minimal* sufficient statistic that will give us the MVUE in a given situation.
 - But, again, for the purposes of this class, we won’t concern ourselves with this too much.



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)
 - So, it’s really a function of a *minimal* sufficient statistic that will give us the MVUE in a given situation.
 - But, again, for the purposes of this class, we won’t concern ourselves with this too much.
- Indeed, in general, constructing MVUEs can be a pain! But, it’s useful to at least know about their existence, and how sufficiency and the Rao-Blackwell theorem tie into constructing them.



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)
 - So, it’s really a function of a *minimal* sufficient statistic that will give us the MVUE in a given situation.



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)
 - So, it’s really a function of a *minimal* sufficient statistic that will give us the MVUE in a given situation.
 - But, again, for the purposes of this class, we won’t concern ourselves with this too much.



Rao-Blackwell Theorem

- Of course, there are some details missing. For one, it turns out that even among sufficient statistics, some are “better” at capturing the information about a parameter than others. (These are called **minimal sufficient statistics**, which we won’t cover in this course.)
 - So, it’s really a function of a *minimal* sufficient statistic that will give us the MVUE in a given situation.
 - But, again, for the purposes of this class, we won’t concern ourselves with this too much.
- Indeed, in general, constructing MVUEs can be a pain! But, it’s useful to at least know about their existence, and how sufficiency and the Rao-Blackwell theorem tie into constructing them.



Example

Example

Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$, where $\theta > 0$ is an unknown parameter.

- (a) Show that $Y_{(n)}$ is a sufficient statistic for θ . (It turns out that this is a *minimal* sufficient statistic for θ , but you do not need to show that.)
- (b) Find an MVUE for θ .



Example

Example

Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$, where $\theta > 0$ is an unknown parameter.

- (a) Show that $Y_{(n)}$ is a sufficient statistic for θ . (It turns out that this is a *minimal* sufficient statistic for θ , but you do not need to show that.)
- (b) Find an MVUE for θ .

- Try this on your own, and feel free to ask me about it during Office Hours!