STA261 - Module 2

Point Estimation

Rob Zimmerman

University of Toronto

July 12-14, 2022

Extracting Information

- In Module 1, we learned about how a statistic can capture (or not capture) the information provided by our data sample $\mathbf{X} = (X_1, \ldots, X_n) \sim f_\theta$ about the unknown parameter $\theta \in \Theta$
- For the remainder of the course, our focus will be on how to *extract* that information
- In Module 2, we have one goal: to estimate the parameter θ or some function of the parameter $\tau(\theta)$ as best we can (whatever that means)
- Example 2.1:

Point Estimation

- How do we estimate θ from the observed data $\mathbf{x}$?

- Ideally, we want some statistic $T(\mathbf{X})$ such that $T(\mathbf{x})$ will be close to θ

- Definition 2.1: Suppose $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} f_\theta$. A **point estimator** $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is a statistic used to estimate θ.

- How do we find good point estimators?

Poll Time!

Choosing "Good" Point Estimators

- A point estimator $\hat{\theta}(\mathbf{X})$ is a random variable, so it has its own distribution (as does any statistic)

- Definition aside, it would seem that the best point estimator is the constant $\hat{\theta}(\mathbf{X}) = \theta$, but of course this is unattainable

- The constant θ has $\mathbb{E}_\theta\left[\theta\right] = \theta$ and $\text{Var}_\theta\left(\theta\right) = 0$

- It would be nice if the distribution of $\hat{\theta}(\mathbf{X})$ got close to these properties: $\mathbb{E}_\theta\left[\hat{\theta}(\mathbf{X})\right] \approx \theta$ and $\text{Var}_\theta\left(\hat{\theta}(\mathbf{X})\right) \approx 0$

- It would also be good if $\text{Var}_\theta\left(\hat{\theta}(\mathbf{X})\right)$ got lower as the sample size n got bigger (if we're willing to pay good money for more samples, we should demand a higher precision in return)

Moments Are (Often) Functions of Parameters

- Here's one approach to choosing $\hat{\theta}$

- In parametric families, it is often the case that the moments (i.e., $\mathbb{E}_\theta\left[X\right]$, $\mathbb{E}_\theta\left[X^2\right]$, $\mathbb{E}\left[X^3\right]$, and so on) are functions of the parameters

- Example 2.2:

Towards the Method of Moments

- Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ and we want to estimate μ

- We know that $\mathbb{E}\left[X_1\right] = \mu$ and $\mathbb{E}\left[X_1^2\right] - \mathbb{E}\left[X_1\right]^2 = \sigma^2$

- So if we took $\hat{\mu}(\mathbf{X}) = X_1$, then we'd have

- Can we do better?

- Now suppose we want to estimate both μ and σ^2

- If we let $m_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$ and $m_2(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i^2$, then
  $m_1(\mathbf{X}) \xrightarrow{d}$ and $m_2(\mathbf{X}) \xrightarrow{d}$

- Therefore $m_2(\mathbf{X}) - m_1(\mathbf{X})^2 \xrightarrow{d}$

The Method of Moments

- Effectively, we're replacing the true moments with the sample moments

- Definition 2.2: Suppose we have k parameters $\theta_1, \theta_2, \ldots, \theta_k$ to estimate in a parametric model, and each one is some function of the first k moments:

$$\theta_j = \psi_j\left(\mathbb{E}_\theta\left[X\right], \mathbb{E}_\theta\left[X^2\right], \ldots, \mathbb{E}_\theta\left[X^k\right]\right), \quad 1 \leq j \leq k.$$

The **Method of Moments (MOM)** estimator for θ_j is defined by choosing

$$\hat{\theta}_j(\mathbf{X}) = \psi_j\left(m_1(\mathbf{X}), m_2(\mathbf{X}), \ldots, m_k(\mathbf{X})\right), \quad 1 \leq j \leq k.$$

Method of Moments: Examples

- Example 2.3: Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Poisson (λ), where $\lambda > 0$. Find the MOM estimator for λ.

Method of Moments: Examples

- Example 2.4: Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bin}(k, \theta)$, where $k \in \mathbb{N}$ and θ is known. Find the MOM estimator for k.

- Could this be a problem?

Poll Time!

Method of Moments: Examples

- Example 2.5: The angle at which electrons are emitted in muon decay has a distribution with density $f_\alpha(x) = (1 + \alpha x)/2$, where $x \in [-1, 1]$ and $\alpha \in [-\frac{1}{3}, \frac{1}{3}]$. Given a sample $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\alpha$, find the MOM estimator for α.

Method of Moments: Examples

- Example 2.6: Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Gamma (α, β), where $\alpha, \beta > 0$. Find the MOM estimators for α and β.

Method of Moments: Advantages and Disadvantages

-

The Likelihood Function

- Definition 2.3: Let $\mathbf{X} \sim f_\theta$, where f_θ is a pdf or pmf in a parametric family. Given the observation $\mathbf{X} = \mathbf{x}$, the **likelihood function for $\boldsymbol{\theta}$** is the function $L(\cdot \mid \mathbf{x}) : \Theta \to [0, \infty)$ given by $L(\theta \mid \mathbf{x}) = f_\theta(\mathbf{x})$.

- Interpret this as the "probability" of observing the sample $\mathbf{x}$, given that the sample came from f_θ

- So $L(\theta_1 \mid \mathbf{x}) > L(\theta_2 \mid \mathbf{x})$ says that the chance of observing $\mathbf{X} = \mathbf{x}$ is more likely under f_{θ_1} than under f_{θ_2}

- It could be that the likelihood is very small for all $\theta \in \Theta$, so knowing $L(\theta \mid \mathbf{x})$ for just a single θ is useless

- Instead, we want to know how $L(\theta \mid \mathbf{x})$ compares to the other $L(\theta' \mid \mathbf{x})$'s

The Likelihood Principle

- Much of modern statistics revolves around the likelihood function; it will be with us in some form or another for the rest of our course

- The **likelihood principle** states that if two model and data combinations $L_1(\theta \mid \mathbf{x})$ and $L_2(\theta \mid \mathbf{y})$ are such that $L_1(\theta \mid \mathbf{x}) = c(\mathbf{x}, \mathbf{y}) \cdot L_2(\theta \mid \mathbf{y})$, then the conclusions about θ drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical

- In other words, the likelihood principle says that anything we want to say about θ should be based solely on $L(\cdot \mid \mathbf{x})$, regardless of how $\mathbf{x}$ was actually obtained

- Is this requirement too strong?

- Example 2.7:

Maximizing the Likelihood

- Suppose there were some $\hat{\theta} \in \Theta$ which makes $L(\hat{\theta} \mid \mathbf{x})$ the highest; would it be sensible to use that $\hat{\theta}$ as an estimator?

- If we can maximize $L(\theta \mid \mathbf{x})$ with respect to θ, the resulting maximizer $\hat{\theta}$ will be a function of the sample $\mathbf{x}$

- Example 2.8: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}\,(\theta)$, where $\theta \in (0, 1)$. Maximize the likelihood with respect to θ.

Maximum Likelihood Estimation

- Definition 2.4: Let $\mathbf{X} = (X_1, \ldots, X_n) \sim f_\theta$. Let $L(\theta \mid \mathbf{x})$ be the likelihood function based on observing $\mathbf{X} = \mathbf{x}$. The **maximum likelihood estimate** of θ is given by

$$\hat{\theta}(\mathbf{x}) = \underset{\theta \in \Theta}{\text{argmax}}\, L(\theta \mid \mathbf{x}),$$

  and the **maximum likelihood estimator (MLE)** for θ is the point estimator given by $\hat{\theta}_{\text{MLE}} = \hat{\theta}(\mathbf{X})$.

Maximum Likelihood: Examples

- Nothing says the distribution needs to have a "nice" functional form

- Example 2.9: Suppose $\mathcal{X} = \{1, 2, 3\}$ and $\Theta = \{a, b\}$, and a parametric family is given by the following table:

	$x = 1$	$x = 2$	$x = 3$
$f_a(x)$	0.3	0.4	0.3
$f_b(x)$	0.1	0.7	0.2

  Suppose we observe $X \sim f_\theta$. Find the MLE of θ.

Maximum Likelihood: Examples

- But when the f_θ *does* have a nice form and is continuously differentiable for $\theta \in \Theta$, we can use calculus to find the MLE

- Example 2.10: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (θ), where $\theta \in (0,1)$. Find the MLE of θ.

Maximum Likelihood: Examples

- Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and σ^2 is known

- What happens if we try to find the MLE of μ in the same fashion?

The Log-Likelihood

- Definition 2.5: Given data $\mathbf{x}$ and a parametric model with likelihood function $L(\theta \mid \mathbf{x})$, the **log-likelihood function** is defined as by

$$\ell(\theta \mid \mathbf{x}) = \log \left(L(\theta \mid \mathbf{x}) \right).$$

- Maximizing the log-likelihood is equivalent to maximizing the likelihood

- ...but usually way easier

The Score Function

- Definition 2.6: Given data $\mathbf{x}$ and a parametric model with log-likelihood function $\ell(\theta \mid \mathbf{x})$, the **score function** is defined as

$$S(\theta \mid \mathbf{x}) = \frac{\partial}{\partial \theta} \ell(\theta \mid \mathbf{x}),$$

  when it exists.

- When $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a vector, this is interpreted as the gradient

$$S(\boldsymbol{\theta} \mid \mathbf{x}) = \nabla \ell(\boldsymbol{\theta} \mid \mathbf{x}) = \left(\frac{\partial}{\partial \theta_1} \ell(\boldsymbol{\theta} \mid \mathbf{x}), \ldots, \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta} \mid \mathbf{x}) \right)$$

- If the likelihood function is nice enough, then any extremum $\hat{\theta}$ will satisfy the *score equation* $S(\hat{\theta} \mid \mathbf{x}) = 0$

- So finding the MLE amounts to finding $\hat{\theta}$ such that $S(\hat{\theta} \mid \mathbf{x}) = 0$ and then checking that $\hat{\theta}$ is a global maximum

Maximum Likelihood: More Examples

- Example 2.11: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in \mathbb{R}$ and σ^2 known. Find the MLE of μ.

Maximum Likelihood: More Examples

- Example 2.12: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Exp}\,(\lambda)$ with $\lambda > 0$. Find the MLE of λ.

Maximum Likelihood: More Examples

- Even if the likelihood is smooth and well-behaved, this method doesn't always work

- Example 2.13: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \Gamma(\alpha, 2)$ with $\alpha > 0$. Try to find the MLE of α.

Maximum Likelihood: More Examples

- What about when θ is multidimensional? We need to bring out our multivariate calculus

- Example 2.14: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Find the MLE of $\theta = (\mu, \sigma^2)$.

Maximum Likelihood: More Examples

- The likelihood may not be differentiable, but that doesn't mean it can't be maximized
- Example 2.15: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Unif}\,(0, \theta)$ with $\theta > 0$. Find the MLE of θ.

Regression Through the Origin

- Example 2.16: Let $Y_1, Y_2, \ldots, Y_n$ be independent where $Y_i \sim \mathcal{N}\left(\beta x_i, \sigma^2\right)$ with $\beta \in \mathbb{R}$, $x_i \in \mathbb{R}$, and $\sigma^2 > 0$. Find the MLE of β.

- This is a particular case of **linear regression**; see Assignment 2 for more

Reparameterization

- Instead of θ itself, what if we want to find the MLE of some one-to-one function of the parameter $\tau(\theta)$?

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bernoulli (θ), where $\theta \in (0, 1)$. Find the MLE of θ^2.

Reparameterization

- That wasn't a coincidence

- Theorem 2.1 (**Invariance Property**): If $\hat{\theta}(\mathbf{X})$ is an MLE of $\theta \in \Theta$ and $\tau(\cdot)$ is one-to-one on Θ, then the MLE of $\tau(\theta)$ is given by $\tau(\hat{\theta}(\mathbf{X}))$.

Proof.

Reparameterization

- Example 2.17: Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim}$ Bernoulli (p) where $p \in (0,1)$. Find the MLE of $\tau(p) = \log\left(\frac{p}{1-p}\right)$.

Poll Time!

Maximum Likelihood Estimation

- Maximum likelihood is *by far* the most common method that statisticians use to find point estimates[1]

- Maximum likelihood estimators tend to have quite good properties (especially for large sample sizes):

- When in doubt, it's usually a good idea to use maximum likelihood if you can

[1] Assuming those statisticians aren't Bayesians – more on that in Module 6

Evaluating Estimators

- Back to the idea of what makes a point estimator "good"
- From now on, we focus on point estimators of $\tau(\theta)$, rather than θ
- It turns out there's a much more convenient way to assess the quality of a point estimator estimator than our earlier thoughts
- Consider the *error* (or *absolute deviation*) of an estimator $|T(\mathbf{X}) - \tau(\theta)|$, which is of course a random variable
- It's too much to ask for this to *always* be small; some random sample $\mathbf{X}_j$ may be an "outlier", so that $T(\mathbf{X}_j)$ is far from $\tau(\theta)$
- But we can ask for it to be small on average

Mean-Squared Error

- In other words, it's reasonable to ask for $\mathbb{E}_\theta\left[|T(\mathbf{X}) - \tau(\theta)|\right]$ to be small
- That's fine, but it turns out that for mathematical reasons, it's much more convenient to ask for the *squared error* $(T(\mathbf{X}) - \tau(\theta))^2$ to be small on average
- Definition 2.7: Let $T(\mathbf{X})$ be an estimator for $\tau(\theta)$. The **mean-squared error (MSE)** is defined as

$$\mathsf{MSE}_\theta\left(T(\mathbf{X})\right) = \mathbb{E}_\theta\left[(T(\mathbf{X}) - \tau(\theta))^2\right].$$

- So why not look for the $T(\mathbf{X})$ that minimizes the MSE for all $\theta \in \Theta$?
- Because unfortunately, such a $T(\mathbf{X})$ almost never exists
- Let's try to restrict the class of estimators under consideration to one where minimizers of the MSE are easier to find

Bias

- Definition 2.8: The **bias** of a point estimator $T(\mathbf{X})$ is defined as

$$\text{Bias}_\theta\left(T(\mathbf{X})\right) = \mathbb{E}_\theta\left[T(\mathbf{X})\right] - \tau(\theta).$$

If $\text{Bias}_\theta\left(T(\mathbf{X})\right) = 0$, then $T(\mathbf{X})$ is said to be an **unbiased estimator** of $\tau(\theta)$.

- Example 2.18:

- Example 2.19:

Unbiased Estimators Don't Always Exist

- Example 2.20: Let $X \sim \text{Bernoulli}(\theta)$, where $\theta \in (0, 1)$. There exists no unbiased estimator of $\tau(\theta) = \frac{1}{\theta}$.

The Bias-Variance Tradeoff

- Theorem 2.2 (**Bias-Variance Tradeoff**): If a point estimator $T(\mathbf{X})$ has a finite second moment, then

$$\mathsf{MSE}_\theta\left(T(\mathbf{X})\right) = \mathsf{Bias}_\theta\left(T(\mathbf{X})\right)^2 + \mathsf{Var}_\theta\left(T(\mathbf{X})\right).$$

Proof.

Poll Time!

Best Unbiased Estimation

- So let's restrict our attention to the class of unbiased estimators, and *then* choose the one (or ones?) with the lowest MSE

- Equivalently, choose the unbiased estimator (or estimators?) with the lowest variance

- Definition 2.9: An unbiased estimator $T^*(\mathbf{X})$ of $\tau(\theta)$ is a **best unbiased estimator** of $\tau(\theta)$ if

$$\text{Var}_\theta\left(T^*(\mathbf{X})\right) \leq \text{Var}_\theta\left(T(\mathbf{X})\right) \quad \text{for all } \theta \in \Theta$$

  where $T(\mathbf{X})$ is any other unbiased estimator of $\tau(\theta)$. A best unbiased estimator is also called a **uniform minimum variance unbiased estimator (UMVUE)** of $\tau(\theta)$.

Questions That We Will Answer

- How do we know whether or not an estimator $T(\mathbf{X})$ is a UMVUE for $\tau(\theta)$?
- How do we find a UMVUE for $\tau(\theta)$?
- Are UMVUEs unique?

An Ubiquitous Inequality in Mathematics

- Theorem 2.3 (**Cauchy-Schwarz Inequality**): Let X and Y be random variables, each having finite, nonzero variance. Then

$$|\text{Cov}\,(X, Y)\,| \leq \sqrt{\text{Var}\,(X)\,\text{Var}\,(Y)}.$$

Furthermore, if $\text{Var}\,(Y) > 0$, then equality is attained if and only if X and Y are linearly related.

Proof.

UMVUEs Are Unique

- Theorem 2.4: If a UMVUE exists for $\tau(\theta)$, then it is unique.

Proof.

The Rao-Blackwell Theorem

- It turns out that sufficiency can help us in our search for the UMVUE in powerful ways

- Theorem 2.5 (**Rao-Blackwell**): Let $W(\mathbf{X})$ be unbiased for $\tau(\theta)$, and let $T(\mathbf{X})$ be sufficient for θ. Define $W_T(\mathbf{X}) = \mathbb{E}_\theta \left[W(\mathbf{X}) \mid T(\mathbf{X})\right]$. Then $W_T(\mathbf{X})$ is also an unbiased point estimator of $\tau(\theta)$, and moreoever, $\mathrm{Var}_\theta \left(W_T(\mathbf{X})\right) \leq \mathrm{Var}_\theta \left(W(\mathbf{X})\right)$.

Proof.

Interpreting Rao-Blackwellization

- The process of replacing an estimator with its conditional expectation (with respect to a sufficient statistic) is called **Rao-Blackwellization**

- Theorem 2.5 says that we can always improve on (or at least make no worse) any unbiased estimator $W(\mathbf{X})$ with a second moment by Rao-Blackwellizing it

- Example 2.21:

Rao-Blackwell: Examples

- Example 2.22: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Bin (k, θ), where $\theta \in (0, 1)$ and k is known. Let $\tau(\theta) = k\theta(1-\theta)^{k-1}$. Show that $W(\mathbf{X}) = \mathbb{1}_{X_1=1}$ is unbiased for $\tau(\theta)$, and then Rao-Blackwellize it.

The Lehmann-Scheffé Theorem

- Theorem 2.6 (**Lehmann-Scheffé Theorem**): Let $W(\mathbf{X})$ be unbiased for $\tau(\theta)$ and let $T(\mathbf{X})$ be a complete sufficient statistic, for all $\theta \in \Theta$. Then $W_T(\mathbf{X}) = \mathbb{E}\left[W(\mathbf{X}) \mid T(\mathbf{X})\right]$ is the unique UMVUE.

Proof.

More On Lehmann-Scheffé

- This is a bit startling
- If we take some unbiased estimator and condition it on a complete sufficient statistic, then the resulting estimator is *the* UMVUE
- As such, if we find an unbiased estimator $T(\mathbf{X})$ of $\tau(\theta)$ which is also a complete sufficient statistic, then we're done
- However, Lehmann-Scheffé assumes that a complete sufficient statistic exists (which isn't always the case, as we know from Module 1), so it doesn't subsume Theorem 2.4
- In fact, there do exist models where UMVUEs exist but complete sufficient statistics don't

Lehmann-Scheffé: Examples

- Example 2.23: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Find the UMVUE of (μ, σ^2).

Lehmann-Scheffé: Examples

- Example 2.24: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim} \text{Poisson}\,(\lambda)$, where $\lambda > 0$. Find the UMVUE of λ.

Poll Time!

What About the Likelihood?

- Rao-Blackwellization and Lehmann-Scheffé tell us how to get the unique UMVUE (if it exists) via complete sufficient statistics
- The likelihood wasn't involved
- It turns out there exists a very helpful tool that helps us with finding the UMVUE (if it exists) by exploiting the likelihood
- It doesn't always work...
- But when it does, it works like a charm
- But we need several auxiliary results to produce it

The Covariance Inequality

- Theorem 2.7 (**Covariance Inequality**): Let $T(\mathbf{X})$ and $U(\mathbf{X})$ be two statistics such that $0 < \mathbb{E}_\theta\left[T(\mathbf{X})^2\right], \mathbb{E}_\theta\left[U(\mathbf{X})^2\right] < \infty$ for all $\theta \in \Theta$. Then

$$\text{Var}_\theta\left(T(\mathbf{X})\right) \geq \frac{\text{Cov}_\theta\left(T(\mathbf{X}), U(\mathbf{X})\right)^2}{\text{Var}_\theta\left(U(\mathbf{X})\right)} \qquad \text{for all } \theta \in \Theta.$$

  Equality holds if and only if

$$T(\mathbf{X}) = \mathbb{E}_\theta\left[T(\mathbf{X})\right] + \frac{\text{Cov}_\theta\left(T(\mathbf{X}), U(\mathbf{X})\right)}{\text{Var}_\theta\left(U(\mathbf{X})\right)}\left(U(\mathbf{X}) - \mathbb{E}_\theta\left[U(\mathbf{X})\right]\right)$$

  almost surely.

Proof.

The Fisher Information

- Definition 2.10: Let $\mathbf{X} = (X_1, \ldots, X_n) \sim f_\theta$, and let $S(\theta \mid \mathbf{x})$ be the score function for the parametric model. The **(expected) Fisher information** is the function $I_n : \Theta \to [0, \infty)$ defined by

$$I_n(\theta) = \mathsf{Var}_\theta \left(S(\theta \mid \mathbf{X}) \right).$$

- Definition 2.11: Let $\mathbf{X} = (X_1, \ldots, X_n) \sim f_\theta$, and let $S(\theta \mid \mathbf{x})$ be the score function for the parametric model. The **observed Fisher information** is the function $J_n : \mathcal{X}^n \to [0, \infty)$ defined by

$$J_n(\mathbf{X}) = -\frac{\partial}{\partial \theta} S(\theta \mid \mathbf{X}_n)\Big|_{\theta = \hat{\theta}_{\mathsf{MLE}}}.$$

The Fisher Information: Examples

- Example 2.25: Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim}$ Poisson$\,(\lambda)$, where $\lambda > 0$. Calculate the observed and expected Fisher information for λ.

The Fisher Information: Examples

- Example 2.26: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$, where $\mu \in \mathbb{R}$ and σ^2 is known. Calculate the observed and expected Fisher information for μ.

The Cramér-Rao Lower Bound

- Theorem 2.8 (**Cramér-Rao Lower Bound**): Let $\mathbf{X} = (X_1, \dots, X_n) \sim f_\theta$, and let $T(\mathbf{X})$ be any estimator such that

$$\mathrm{Var}_\theta\left(T(\mathbf{X})\right) < \infty \quad \text{and} \quad \frac{d}{d\theta}\mathbb{E}_\theta\left[T(\mathbf{X})\right] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[T(\mathbf{x}) f_\theta(\mathbf{x})]\, \mathrm{d}\mathbf{x}.$$

  Then

$$\mathrm{Var}_\theta\left(T(\mathbf{X})\right) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta\left[T(\mathbf{X})\right]\right)^2}{I_n(\theta)}.$$

  In particular, if $T(\mathbf{X})$ is unbiased for $\tau(\theta)$ and $\tau(\cdot)$ is differentiable on Θ, then

$$\mathrm{Var}_\theta\left(T(\mathbf{X})\right) \geq \frac{\left(\tau'(\theta)\right)^2}{I_n(\theta)}.$$

Proof.

The Cramér-Rao Lower Bound

The Cramér-Rao Lower Bound Conditions

- Unfortunately, the conditions of the Cramér-Rao Lower Bound don't always hold
- The first says that our estimator must actually have a variance to minimize, which seems reasonable
- Example 2.27:
- The second says that we need to be able to push a derivative inside an integral, which is more subtle
- When would this condition fail to hold?
- Example 2.28:

Easing the Computation

- Theorem 2.9: Under the conditions of Theorem 2.8,

$$I_n(\theta) = \mathbb{E}_\theta\left[S(\theta \mid \mathbf{X})^2\right].$$

Proof.

- Theorem 2.10: If $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ and conditions of Theorem 2.8 hold,

$$I_n(\theta) = n\mathbb{E}_\theta\left[S(\theta \mid X)^2\right].$$

More Easing

- Theorem 2.11 (**Second Bartlett Identity**): If $X \sim f_\theta$ and f_θ satisfies

$$\frac{d}{d\theta}\mathbb{E}_\theta\left[S(\theta \mid X)\right] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[S(\theta \mid x) f_\theta(x)]\, \mathrm{d}x,$$

(which is true when f_θ is in an exponential family) then

$$\mathbb{E}_\theta\left[S(\theta \mid X)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}S(\theta \mid X)\right].$$

Proof.

Efficiency

- Definition 2.12: An estimator $T(\mathbf{X})$ of $\tau(\theta)$ that attains the Cramér-Rao Lower Bound is called an **efficient estimator of $\boldsymbol{\tau(\theta)}$**.
- What's the connection between UMVUEs and efficient estimators?
- If an efficient estimator exists, then it must be the UMVUE
- But an efficient estimator doesn't always exist, as we'll soon see

Efficiency: Examples

- Example 2.29: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that $T(\mathbf{X}) = \bar{X}_n$ is an efficient estimator for μ.

A Criterion for Efficiency

- Is there a better way to find efficient estimators than simply making an educated guess?

- Theorem 2.12: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$ satisfy the conditions of Theorem 2.8. An unbiased estimator $T(\mathbf{X})$ of $\tau(\theta)$ is efficient if and only if there exists some function $a : \Theta \to \mathbb{R}$ such that

$$S(\theta \mid \mathbf{x}) = a(\theta)[T(\mathbf{x}) - \tau(\theta)].$$

Proof.

Efficiency: Examples

- Example 2.30: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Show that there exists no efficient estimator of σ^2.

Efficiency: Examples

- If an unbiased point estimator is efficient, then it's the UMVUE – but the converse is not true in general

- Example 2.31: Let $X_1, X_2, \dots, X_n \overset{iid}{\sim}$ Poisson (λ), where $\lambda > 0$. Show that an efficient estimator of $\tau(\lambda) = \mathbb{P}_\lambda(X = 0)$ does not exist, and find its UMVUE.