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: $X_1, X_2, \ldots, X_n$ = heights of UofT students $\overset{iid}{\sim} N(\mu, 3)$.

  If we want to estimate μ, maybe take $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n$. Seems reasonable!

  - Event of voting for Candidate X in an election $\sim$ Bernoulli(p).

    Maybe want to estimate $\tau(p) = \log\left(\frac{p}{1-p}\right)$ "log-odds of p"

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 $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$. 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!

Not a point estimator: $\frac{1}{2}(X_1 + \mu)$

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 $\mathbb{E}_\theta[\theta] = \theta$ and $\text{Var}_\theta(\theta) = 0$
- 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)

In Module 5, we'll have further demands as $n \to \infty$.

Moments Are (Often) Functions of Parameters

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

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

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

- Example 2.2:

$X \sim N(\mu, \sigma^2) \Rightarrow \mathbb{E}[X] = \mu, \quad \mathbb{E}[X^2] = \mu^2 + \sigma^2$

$X \sim \text{Bin}(k, p) \Rightarrow \mathbb{E}[X] = kp, \quad \mathbb{E}[X^2] = kp(1-p) + k^2p^2$

$X \sim \text{Poisson}(\lambda) \Rightarrow \mathbb{E}[X] = \lambda, \quad \mathbb{E}[X^2] = \lambda + \lambda^2$

$X \sim \text{Exp}(\lambda) \Rightarrow \mathbb{E}[X] = \frac{1}{\lambda}, \quad \mathbb{E}[X^n] = \frac{n!}{\lambda^n}$ (Exercise!)

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 $\hat{\mu}_1(\mathbf{X}) = X_1$, then we'd have $\mathbb{E}_\mu[\hat{\mu}_1(\vec{X})] = \mathbb{E}_\mu[X_1] = \mu$.
- Can we do better? $\hat{\mu}_n(\vec{X}) := \bar{X}_n \Rightarrow \mathbb{E}_\mu[\hat{\mu}_n(\vec{X})] = \mu$ and $\mathrm{Var}_\mu(\hat{\mu}_n(\vec{X})) = \frac{\sigma^2}{n} < \sigma^2 = \mathrm{Var}_\mu(\hat{\mu}_1(\vec{X}))$
- Now suppose we want to estimate both μ and σ^2

Def: the k'th sample moment is $\frac{1}{n}\sum_{i=1}^n X_i^k$

- 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 $m_1(\mathbf{X}) \xrightarrow{d} \mu$ and $m_2(\mathbf{X}) \xrightarrow{d} \mu^2 + \sigma^2 = \mathbb{E}[X^2]$ by LLN!
- Therefore $m_2(\mathbf{X}) - m_1(\mathbf{X})^2 \xrightarrow{d} \sigma^2$ by the continuous mapping theorem (STA257 & Module 5)

So take $\hat{\mu}(\vec{X}) = m_1(\vec{X}) = \bar{X}_n$ and $\hat{\sigma}^2(\vec{X}) = m_2(\vec{X}) - m_1(\vec{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$

The Method of Moments (MOM)

- 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 paremetric 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

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

↳ k^{th} sample moment $= \frac{1}{n}\sum_{i=1}^{n} X_i^k$

Basic principle: LLN + CMT

↑ Although continuity of ψ_j's is not required

Method of Moments: Examples

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

$\lambda = \mathbb{E}[X_i]$

$\Rightarrow$ The MOM estimator is $\hat{\lambda}_{mom}(\vec{X}) = \bar{X}_n$.

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.

$\mathbb{E}_k[X_i] = k\theta \Rightarrow k = \frac{\mathbb{E}_k[X_i]}{\theta} =: \varphi_1(\mathbb{E}[X])$

$\Rightarrow$ Our MOM estimator is $\hat{k}_{mm}(\vec{X}) = \varphi_1(\bar{X}_n) = \frac{1}{\theta} \cdot \bar{X}_n$.

- Could this be a problem? Yes! There's no reason for $\hat{k}(\vec{X})$ to be a natural number, even though $\Theta = \mathbb{N}$.

Poll Time!

$X_1, \ldots, X_n$ iid $\text{Bin}(k, \theta)$, k known, $\theta \in (0,1)$.

$$\mathbb{E}[X] = k\theta \Rightarrow \theta = \frac{\mathbb{E}[X]}{k}$$

$$\Rightarrow \hat{\theta}_{\text{mom}}(\vec{X}) = \frac{1}{k}\bar{X}_n = \frac{1}{nk}\sum_{i=1}^{n} X_i .$$

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.

$$\mathbb{E}[X] = \int_{-1}^{1} x \cdot \left(\frac{1+\alpha x}{2}\right) dx = \frac{1}{2}\left[\frac{x^2}{2} + \frac{\alpha x^3}{3}\right]_{-1}^{1} = \alpha/3.$$

$$\Rightarrow \alpha = 3 \cdot \mathbb{E}[X]$$

$$\Rightarrow \hat{\alpha}_{mom}(\vec{X}) = 3 \cdot \bar{X}_n.$$

Method of Moments: Examples

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

$\psi_1 = \mathbb{E}(X) = \frac{\alpha}{\beta}$ (1)

$\psi_2 = \mathbb{E}(X^2) = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2} = \frac{\alpha + \alpha^2}{\beta^2}$. (2)

(1) $\Rightarrow \alpha = \psi_1 \cdot \beta$

$\Rightarrow \psi_2 = \frac{\psi_1 \beta + \psi_1^2 \beta^2}{\beta^2} = \frac{\psi_1}{\beta} + \psi_1^2$

$\Rightarrow \beta = \frac{\psi_1}{\psi_2 - \psi_1^2}$

$\Rightarrow \alpha = \frac{\psi_1^2}{\psi_2 - \psi_1^2}$

Our MOM estimators are

$\hat{\alpha}_{mom}(\vec{X}) = \frac{m_1(\vec{X})^2}{m_2(\vec{X}) - m_1(\vec{X})^2}$

$\hat{\beta}_{mom}(\vec{X}) = \frac{m_1(\vec{X})}{m_2(\vec{X}) - m_1(\vec{X})^2}$.

Method of Moments: Advantages and Disadvantages

GOOD: Very simple and easy to calculate; they exist wherever the moments do

BAD: They may not be in the correct parameter space ☹

BAD: If moments don't exist (ie, Cauchy), it won't work

BAD: May not be sufficient statistics (if we care)

BAD: May not have lowest variance possible among all point estimators of θ (as well see...)

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 θ** is the function $L(\cdot \mid \mathbf{x}) : \Theta \to [0, \infty)$ given by $L(\theta \mid \mathbf{x}) = f_\theta(\mathbf{x})$. ($L(\theta|\vec{X})$ is a random function of θ)

  If $\theta_1, \theta_2 \in \Theta$ then $L(\theta_1|\vec{x}) = f_{\theta_1}(\vec{x})$ and $L(\theta_2|\vec{x}) = f_{\theta_2}(\vec{x})$.

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

  NOT $P(\theta = \theta | \vec{X} = \vec{x})$!!! $L(\theta|\vec{x}) \notin [0,1]$ in general!

- 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}

  So $L(\cdot|\vec{x})$ ranks the elements of Θ given the observed data

- 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

ie, if $\frac{L_1(\theta|\vec{x})}{L_2(\theta|\vec{y})}$ is free of θ

- 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:

Experiment 1: toss a coin w/ $\mathbb{P}(H) = \theta$ 10 times; let $X = \#$ of $H \sim Bin(10, \theta)$
Say we observe $X = 4$. $L_1(\theta|4) = \binom{10}{4}\theta^4(1-\theta)^6$
Experiment 2: toss the same coin until we observe 4 H; let $Y = \#$ of T until that happens.
Then $Y \sim NegBin(4, \theta)$. Say we observe $Y = 6$. Then $L_2(\theta|6) = \binom{9}{6}\theta^4(1-\theta)^6$

Then $L_1(\theta|x=4) \propto L_2(\theta|y=6)$. The Likelihood Principle says we should be indifferent to Exp 1 or Exp 2.

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

$$L(\theta|\vec{x}) = f_\theta(\vec{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.$$

We'll soon see that the maximum occurs at $\hat{\theta} = \bar{x}_n$.

So with this idea, a point estimator could be $\hat{\theta}(\vec{X}) = \overline{X}_n$.

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

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

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

Equivalently, $\hat{\theta}(\vec{x})$ is s.t. $L(\hat{\theta}(\vec{x}) \mid \vec{x}) \geq L(\theta \mid \vec{x})$ $\forall \theta \in \Theta$.

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.

$X=1 \Rightarrow f_a(1) > f_b(1) \Rightarrow \hat{\theta}(1) = a$

$X=2 \Rightarrow f_b(2) > f_a(2) \Rightarrow \hat{\theta}(2) = b$

$X=3 \Rightarrow f_a(3) > f_b(3) \Rightarrow \hat{\theta}(3) = a$

Therefore $\hat{\theta}(x) = \begin{cases} a, & x \in \{1,3\} \\ b, & x = 2 \end{cases} = a \cdot \mathbb{1}_{x \in \{1,3\}} + b \cdot \mathbb{1}_{x=2}$

So $\hat{\theta}_{MLE}(X) = a \cdot \mathbb{1}_{X \in \{1,3\}} + b \cdot \mathbb{1}_{X=2}$.

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.

$L(\theta|\vec{x}) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$

$\frac{dL}{d\theta} = (\sum x_i)\theta^{\sum x_i - 1}(1-\theta)^{n-\sum x_i} - (n-\sum x_i)\theta^{\sum x_i}(1-\theta)^{n-\sum x_i - 1} \overset{\text{set}}{=} 0$

$\Rightarrow (\sum x_i)\theta^{-1} - (n-\sum x_i)(1-\theta)^{-1} = 0$ (divide by $\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \neq 0$)

$\Rightarrow \frac{\sum x_i}{n-\sum x_i} = \frac{\theta}{1-\theta} \Rightarrow \hat{\theta} = \frac{1}{n}\sum x_i = \bar{x}_n.$

Is this a local max? Need to find $\frac{d^2L}{d\theta^2}$, plug in $\hat{\theta} = \bar{x}_n$, and check that $\frac{d^2L}{d\theta^2} < 0$ there.

(you can verify). So $\hat{\theta}_{MLE}(\vec{x}) = \bar{X}_n$.

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?

$$L(\mu|\vec{x}) = \prod_{i=1}^{n} f_\mu(x_i)$$

$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{1}{2\sigma^2}\left\{\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right\}\right)$$

$$\frac{dL}{d\mu} = \underbrace{(2\pi\sigma^2)^{-n/2}}_{\neq 0} \cdot \underbrace{\left(\frac{\sum x_i - n\mu}{2\sigma^2}\right)}_{\text{must be } 0} \cdot \underbrace{\exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right)}_{\neq 0} \overset{\text{set}}{=} 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum x_i = \bar{x}_n.$$

But differentiating $\frac{dL}{d\mu}$ w.r.t. μ would be a nightmare!

Is there an easier way?!

...yes.

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

because it's a monotonically increasing function of $L(\theta|\vec{x})$

- ...but usually way easier

...because it's easier to differentiate a sum than a product (❤️ linearity!)

If the data are iid, then $\ell(\theta|\vec{x}) = \log(L(\theta|\vec{x}))$

$$= \log\left(\prod_{i=1}^{n} f_\theta(x_i)\right)$$

$$= \sum_{i=1}^{n} \log(f_\theta(x_i))$$

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

$L(\mu|\vec{x}) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left(\frac{-\sum(x_i-\mu)^2}{2\sigma^2}\right)$

$\Rightarrow \ell(\mu|\vec{x}) = c - \frac{\sum(x_i-\mu)^2}{2\sigma^2}$ where $c \in \mathbb{R}$ is free of μ

$\Rightarrow S(\mu|\vec{x}) = \frac{\partial}{\partial\mu}\left(\frac{-\sum x_i^2 + 2\mu\sum x_i - n\mu^2}{2\sigma^2}\right)$

$= \frac{\sum x_i - n\mu}{\sigma^2} \overset{\text{set}}{=} 0$

$\Rightarrow \hat{\mu} = \bar{x}_n.$

Second derivative test: $\frac{\partial}{\partial\mu} S(\mu|\vec{x}) = \frac{-n}{\sigma^2}$

$\Rightarrow \frac{\partial}{\partial\mu} S(\mu|\vec{x})\Big|_{\mu=\bar{x}_n} = \frac{-n}{\sigma^2} < 0.$

Therefore, $\hat{\mu}(\vec{X}) = \bar{X}_n$ is the MLE for 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.

$$L(\lambda|\vec{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \cdot \exp(-\lambda \textstyle\sum x_i) = \lambda^n \cdot \exp(-\lambda n \bar{x}).$$

$$\Rightarrow \ell(\lambda|\vec{x}) = n \cdot \log(\lambda) - \lambda n \bar{x}$$

$$\Rightarrow S(\lambda|\vec{x}) = \frac{n}{\lambda} - n\bar{x} \overset{set}{=} 0$$

$$\Rightarrow \hat{\lambda} = 1/\bar{x}$$

Second derivative test: $\frac{\partial}{\partial \lambda} S(\lambda|\vec{x}) = \frac{-n}{\lambda^2}$

$$\Rightarrow \frac{\partial}{\partial \lambda} S(\lambda|\vec{x})\Big|_{\lambda = 1/\bar{x}} = \frac{-n}{(1/\bar{x})^2} < 0.$$

So the MLE of λ is $\hat{\lambda}(\vec{X}) = 1/\bar{X}_n$.

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.

$$L(\alpha|\vec{x}) = \prod_{i=1}^{n} \frac{2^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-2x_i} = \frac{2^{n\alpha}}{\Gamma(\alpha)^n} \cdot \left(\prod_i x_i\right)^{\alpha-1} \cdot e^{-2\sum x_i}$$

$$\Rightarrow \ell(\alpha|\vec{x}) = n\alpha \cdot \log(2) - n \cdot \log(\Gamma(\alpha)) + (\alpha-1) \cdot \log\left(\prod_{i=1}^{n} x_i\right) + c \quad \text{where } c \in \mathbb{R} \text{ is free of } \alpha$$

$$\Rightarrow S(\alpha|\vec{x}) = n \cdot \log(2) - ??? + \log\left(\prod_i x_i\right)$$

No closed form! We can't differentiate this!

Maximum Likelihood: More Examples

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

(Completed after lecture)

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

$$L(\mu, \sigma^2 | \vec{x}) = (2\pi\sigma^2)^{-n/2} \cdot \exp\left(-\frac{\sum (x_i - \mu)^2}{2\sigma^2}\right)$$

$$\Rightarrow \ell(\mu, \sigma^2 | \vec{x}) = c - \frac{n}{2}\log(\sigma^2) - \frac{\sum (x_i - \mu)^2}{2\sigma^2}$$

$$\Rightarrow S(\mu, \sigma^2 | \vec{x}) = \nabla S = \left(\frac{\partial S}{\partial \mu}, \frac{\partial S}{\partial \sigma^2}\right) = \left(\frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu), \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \mu)^2\right) \overset{\text{set}}{=} \vec{0} = (0,0)$$

$$\overset{\text{solve tediously}}{\Longrightarrow} (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{x}, \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2\right)$$

Second derivative test:

$$\frac{\partial^2 S}{\partial \mu^2} = \frac{-n}{\sigma^2} < 0$$

$$\frac{\partial^2 S}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (x_i - \mu)^2$$

$$\frac{\partial^2 S}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu)$$

The determinant of the Hessian is

$$\begin{vmatrix} \frac{\partial^2 S}{\partial \mu^2} & \frac{\partial^2 S}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 S}{\partial \mu \partial \sigma^2} & \frac{\partial^2 S}{\partial (\sigma^2)^2} \end{vmatrix} = \begin{vmatrix} \frac{-n}{\sigma^2} & -\frac{1}{\sigma^4}\sum (x_i - \mu) \\ -\frac{1}{\sigma^4}\sum (x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum (x_i - \mu)^2 \end{vmatrix}_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2}$$

$$\cdots = \frac{1}{\hat{\sigma}^6} \cdot \frac{n^2}{2} > 0. \quad \text{So } (\hat{\mu}, \hat{\sigma}^2) \text{ is the MLE.}$$

Note: we could've also done this by maximizing $\ell(\mu, \sigma^2 | \vec{x})$ for fixed σ^2 to find $\hat{\mu}$, and then maximizing $\ell(\hat{\mu}, \sigma^2 | \vec{x})$ in σ^2. This works since μ and σ^2 aren't functions of each other.

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.

$$L(\theta|\vec{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \theta^{-n} \cdot \mathbb{1}_{0 \le x_{(1)} \wedge x_{(n)} \le \theta} = \mathbb{1}_{0 \le x_{(1)}} \cdot \theta^{-n} \cdot \mathbb{1}_{x_{(n)} \le \theta}$$

If $\theta = x_{(n)}$, then $L(x_{(n)}|\vec{x}) = \mathbb{1}_{0 \le x_{(1)}} \cdot x_{(n)}^{-n}$

If $\theta > x_{(n)}$, then $L(\theta|\vec{x}) = \mathbb{1}_{0 \le x_{(1)}} \cdot \theta^{-n} \le \mathbb{1}_{0 \le x_{(1)}} \cdot x_{(n)}^{-n} = L(x_{(n)}|\vec{x})$

If $\theta < x_{(n)}$, then $L(\theta|\vec{x}) = \mathbb{1}_{0 \le x_{(1)}} \cdot \theta^{-n} \cdot 0 = 0 \le L(x_{(n)}|\vec{x})$

Hence $\hat{\theta}_{MLE}(\vec{x}) = x_{(n)}$. But we couldn't use calculus to find it, because $L(\theta|\vec{x})$ is not differentiable at $\theta = x_{(n)}$.

Regression Through the Origin

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

all known!

$$L(\beta|\vec{y}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \cdot \exp\left(\frac{-(y_i - \beta x_i)^2}{2\sigma^2}\right)$$
$$= (2\pi\sigma^2)^{-n/2} \cdot \exp\left(\frac{-\sum (y_i - \beta x_i)^2}{2\sigma^2}\right)$$

$$\Rightarrow \ell(\beta|\vec{y}) = C - \frac{\sum (y_i - \beta x_i)^2}{2\sigma^2} \quad \text{where } C \in \mathbb{R} \text{ is free of } \beta$$

$$\Rightarrow S(\beta|\vec{y}) = \frac{\sum x_i (y_i - \beta x_i)}{\sigma^2} \overset{\text{set}}{=} 0$$

$$\Rightarrow \sum x_i (y_i - \beta x_i) = 0$$

$$\Rightarrow \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Second derivative test:

$$\frac{\partial}{\partial \beta} S(\beta|\vec{y}) = -\frac{\sum x_i^2}{\sigma^2} < 0 \quad \forall \beta \in \mathbb{R}.$$

Hence $\hat{\beta}_{MLE}(\vec{Y}) = \dfrac{\sum x_i Y_i}{\sum x_i^2}$.

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

Let $\tau = \theta^2$.

Then $L(\tau(\vec{x})) = \sqrt{\tau}^{\sum x_i}(1-\sqrt{\tau})^{n-\sum x_i}$

$\Rightarrow \ell(\tau(\vec{x})) = n\bar{x}\cdot\log(\sqrt{\tau}) + (n-n\bar{x})\cdot\log(1-\sqrt{\tau})$

$\Rightarrow S(\tau(\vec{x})) = \frac{n\bar{x}}{2\tau} + \frac{n-n\bar{x}}{2(\tau-\sqrt{\tau})} \overset{\text{set}}{=} 0$

$\Rightarrow \sqrt{\hat{\tau}} = \bar{x}$

$\Rightarrow \hat{\tau} = (\bar{x})^2$

$\Rightarrow \hat{\tau}(\vec{x}) = (\bar{X}_n)^2 = (\theta_{MLE}(\vec{x}))^2$

Exercise: second derivative test.

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 $\tau(\hat{\theta}(\mathbf{X}))$.

ie, $\tau(\theta)_{MLE}(\vec{X}) = \tau(\underset{MLE}{\theta}(\vec{X}))$ "plug-in estimator"

Proof. Let $\psi = \tau(\theta)$ so that $\theta = \tau^{-1}(\psi)$, and let $\hat{\psi} := \tau^{-1}(\hat{\theta})$.

Let the likelihood under θ be $L(\theta|\vec{x})$, and the likelihood under ψ be $L^*(\psi|\vec{x})$.

Then for any $\psi = \tau(\theta) \in \tau(\Theta)$,

$$
\begin{aligned}
L^*(\hat{\psi}|\vec{x}) &= f_{\tau^{-1}(\hat{\psi})}(\vec{x}) \\
&= L(\tau^{-1}(\hat{\psi})|\vec{x}) \\
&= L(\hat{\theta}|\vec{x}) \\
&\geq L(\theta|\vec{x}) \\
&= L(\tau^{-1}(\psi)|\vec{x}) \\
&= f_{\tau^{-1}(\psi)}(\vec{x}) \\
&= L^*(\psi|\vec{x}).
\end{aligned}
$$

Hence $\hat{\psi}$ maximizes L^*. □

Discussion prompt: what if τ is not one-to-one?

Reparameterization

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

From Ex 2.11, $\hat{p}_{MLE}(\vec{X}) = \overline{X}_n$.

Since $\log\left(\frac{x}{1-x}\right)$ is one-to-one on $(0,1)$, the MLE of $\tau(p)$ is $\log\left(\frac{\overline{X}_n}{1-\overline{X}_n}\right)$

by the invariance property.

Poll Time!

Invariance principle! $\hat{\lambda}_{MLE} = \bar{X}_n$

$\Rightarrow \widehat{(-\log(\lambda))}_{MLE} = -\log(\bar{X}_n)$.

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):

GOOD: $\hat{\theta}_{MLE}$ is always in $\mathcal{H}$

GOOD: Widely applicable; requires very few assumptions (don't need moments, etc)

GOOD: (Relatively) easy to easy to implement (usually general-purpose optimization software does it)

"BAD": $\hat{\theta}_{MLE}$ may not have the "right" expectation (eg, $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ but $\mathbb{E}[\sigma^2_{MLE}] \neq \sigma^2$ in the $N(\mu, \sigma^2)$ model)

- 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

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

- 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 $\tau(\theta)$.

$\Rightarrow \mathbb{E}_\theta[T(\vec{X})] = \tau(\theta)$

- Example 2.18:

$X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Then $T(\vec{X}) = \bar{X}_n$ is unbiased for μ.

$X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$, $p \in (0,1)$. Then $T(\vec{X}) = \bar{X}_n$ is unbiased for p,

because $\text{Bias}_p(T(\vec{X})) = \mathbb{E}_p[T(\vec{X})] - p = \mathbb{E}_p\left[\frac{1}{n}\sum_{i=1}^n X_i\right] - p$

$= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_p[X_i] - p$

$= \frac{1}{n} \cdot np - p = p - p = 0.$

- Example 2.19:

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

$\text{Bias}_{\sigma^2}\left(\hat{\sigma}^2_{MLE}(\vec{X})\right) = \text{Bias}_{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = \frac{\sigma^2}{n} \neq 0.$ So MLEs can be biased!

(Assignment 0: $\mathbb{E}[S^2] = \text{Var}(X)$)

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 $\tau(\theta) = \frac{1}{\theta}$.

Suppose $T(X)$ is unbiased for $\tau(\theta) = 1/\theta$.

Then $\frac{1}{\theta} = \mathbb{E}_\theta[T(X)] = T(0) \cdot P_\theta(X=0) + T(1) \cdot P_\theta(X=1)$

$= T(0) \cdot (1-\theta) + T(1) \cdot \theta \quad \forall \theta \in (0,1)$.

But LHS is unbounded as $\theta \to 0$, but RHS $\to T(0) \in \mathbb{R}$. So this can't happen!

The Bias-Variance Tradeoff

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

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

Proof. $MSE_\theta(T(\vec{X})) = \mathbb{E}_\theta[(T(\vec{X}) - \tau(\theta))^2]$

$= Var_\theta(T(\vec{X}) - \tau(\theta)) + \mathbb{E}_\theta[T(\vec{X}) - \tau(\theta)]^2$

$= Var_\theta(T(\vec{X})) + Bias_\theta(T(\vec{X}))^2$. $\square$

$Var(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$

Take $Y = T(\vec{X}) - \tau(\theta)$.

So among all estimators with a fixed MSE, you must choose between more accuracy ($\downarrow$ bias) and less precision ($\uparrow$ variance), or vice versa..

Poll Time!

$$MSE_\theta[T(\vec{X})] = Bias_\theta(T(\vec{X}))^2 + Var_\theta(T(\vec{X}))$$

If $T(\vec{X})$ is unbiased, then $Bias_\theta(T(\vec{X})) = 0 \Rightarrow MSE = Var.$

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. In particular, if $X = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)} \cdot Y + a$ for some $a \in \mathbb{R}$.

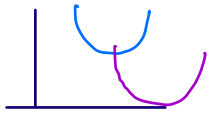
Proof.

Proof: Let $\mu_X = \mathbb{E}[X]$, $\mu_Y = \mathbb{E}[Y]$. If $\mathrm{Var}(Y) = 0$, then $Y = \mu_Y$. Then

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)\underbrace{(Y-\mu_Y)}_{0}] = 0 \leq 0 = \sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}.$$

So suppose $\mathrm{Var}(Y) > 0$. Let $Z = X - \mu_X$ and $W = Y - \mu_Y$. Let $t \in \mathbb{R}$. Then

$$\begin{aligned}\mathbb{E}[(Z-tW)^2] &= \mathbb{E}[Z^2] - 2t\cdot\mathbb{E}[ZW] + t^2\cdot\mathbb{E}[W^2] \\ &= \mathrm{Var}(X) - 2t\cdot\mathrm{Cov}(X,Y) + t^2\cdot\mathrm{Var}(Y).\end{aligned}$$



This is quadratic in t and non-negative, so it has at most one real root.

Therefore, $4\cdot\mathrm{Cov}(X,Y)^2 - 4\cdot\mathrm{Var}(X)\cdot\mathrm{Var}(Y) \leq 0$

$$\Rightarrow \mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X)\cdot\mathrm{Var}(Y)$$

$$\Rightarrow |\mathrm{Cov}(X,Y)| \leq \sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}$$

Quadratic formula: $0 = ax^2+bx+c$

$$\Rightarrow x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

Equality holds iff $\mathbb{E}[(Z-tW)^2] = 0$ for some t

iff $Z = tW$

iff $X - \mu_X = t\cdot(Y-\mu_Y)$

iff $X = tY - t\mu_Y + \mu_X$ (i.e., iff X and Y are linearly related).

Moreover, check that $t = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}$. □

UMVUEs Are Unique

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

Proof. Let W and W' be two UMVUEs of $\tau(\theta)$. Let $W^* = \frac{1}{2}(W + W')$.

Clearly W^* is unbiased for $\tau(\theta)$, and moreover,

$$\begin{aligned}
\mathrm{Var}_\theta(W^*) &= \tfrac{1}{4}\mathrm{Var}_\theta(W) + \tfrac{1}{4}\mathrm{Var}_\theta(W') + \tfrac{1}{2}\cdot\mathrm{Cov}_\theta(W, W') \\
&\le \tfrac{1}{4}\mathrm{Var}_\theta(W) + \tfrac{1}{4}\mathrm{Var}_\theta(W') + \tfrac{1}{2}\sqrt{\mathrm{Var}(W)\cdot\mathrm{Var}(W')} \quad \text{by Cauchy-Schwarz} \\
&= \mathrm{Var}_\theta(W) \quad \text{since all variances are equal (by assumption).}
\end{aligned}$$

But W^* can't beat a UMVUE, so equality must hold (i.e., $\mathrm{Cov}(W, W') = \mathrm{Var}_\theta(W)$ ✱)

$\Rightarrow W' = a(\theta)\cdot W + b(\theta)$. What are $a(\theta)$ and $b(\theta)$?

✱ implies $\mathrm{Var}_\theta(W) = \mathrm{Cov}_\theta(W, W')$

$= \mathrm{Cov}_\theta(W, a(\theta)\cdot W + b(\theta))$

$= \mathrm{Cov}_\theta(W, a(\theta)\cdot W)$

$= a(\theta)\cdot\mathrm{Cov}(W, W)$

$= a(\theta)\cdot\mathrm{Var}(W)$

$\Rightarrow a(\theta) = 1$

Finally, $\tau(\theta) = \mathbb{E}_\theta[W']$

$= \mathbb{E}_\theta[1\cdot W + b(\theta)]$

$= \tau(\theta) + b(\theta)$

$\Rightarrow b(\theta) = 0$.

Hence $W = W'$. □

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, $\text{Var}_\theta\left(W_T(\mathbf{X})\right) \leq \text{Var}_\theta\left(W(\mathbf{X})\right)$.

i.e., conditioning on sufficient statistics never hurts!

Proof. Unbiasedness: $\mathbb{E}_\theta[W_T(\vec{X})] = \mathbb{E}_\theta\left[\mathbb{E}_\theta(W(\vec{X})|T(\vec{X}))\right] \overset{\text{"tower rule"}}{=} \mathbb{E}_\theta[W(\vec{X})] = \tau(\theta)$ since W is unbiased.

Moreover, $\text{Var}_\theta(W(\vec{X})) = \mathbb{E}_\theta\left[\underbrace{\text{Var}_\theta(W(\vec{X})|T(\vec{X}))}_{\geq 0}\right] + \text{Var}_\theta\left(\mathbb{E}[W(\vec{X})|T(\vec{X})]\right)$

$\geq \text{Var}_\theta\left(\mathbb{E}_\theta[W(\vec{X})|T(\vec{X})]\right)$

$= \text{Var}_\theta(W_T(\vec{X}))$. □

What about sufficiency? If $T(\vec{X})$ weren't sufficient, then $\mathbb{E}_\theta[W(\vec{X})|T(\vec{X})]$ wouldn't be free of θ

$\Rightarrow$ not a point estimator!

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: $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$, $\lambda > 0$.

  We have at least two unbiased estimators for λ: $\bar{X}_n$ and S^2.

  But $\bar{X}_n$ is sufficient for λ by Theorem 1.2, so $\mathbb{E}[S^2 | \bar{X}_n]$ is better than S^2 itself,

  but by Theorem 2.9, it can't be better than $\mathbb{E}[\bar{X}_n | \bar{X}_n] = \bar{X}_n$.

Rao-Blackwell: Examples

$\sum_{i=1}^n X_i \sim \text{Bin}(nk, \theta)$

$\Rightarrow \sum_{i=2}^n X_i \sim \text{Bin}((n-1)k, \theta)$ (Exercise)

- Example 2.22: Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bin}(k, \theta)$, 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.

Unbiasedness: $\mathbb{E}_\theta[W(\vec{X})] = \mathbb{E}_\theta(\mathbb{1}_{X_1=1}] = P_\theta(X_1=1) = k\theta(1-\theta)^{k-1} = \tau(\theta)$.

Recall that $T(\vec{X}) = \sum_{i=1}^n X_i$ is sufficient for θ.

Let $W_T(\vec{x}) = \mathbb{E}[W(\vec{X}) \mid T(\vec{X}))$. How do we use this? Suppose $T(\vec{x}) = t$.

Then $\mathbb{E}[W(\vec{X}) \mid T(\vec{X}) = t]$

$= P(X_1=1 \mid \sum_{i=1}^n X_i = t)$

$= \dfrac{P(X_1=1 \wedge \sum X_i = t)}{P(\sum X_i = t)}$

$= \dfrac{P(X_1=1 \wedge \sum_{i=2}^n X_i = t-1)}{P(\sum X_i = t)}$

$\overset{\text{indep}}{=} \dfrac{P(X_1=1) \cdot P(\sum_{i=2}^n X_i = t-1)}{P(\sum X_i = t)}$

$= \dfrac{k\theta(1-\theta)^{k-1} \cdot \binom{k(n-1)}{t-1} \theta^{t-1}(1-\theta)^{(n-1)k-(t-1)}}{\binom{kn}{t}\theta^t(1-\theta)^{kn-t}}$

$= k\binom{k(n-1)}{t-1} \Big/ \binom{kn}{t}$

Therefore, $W_T(\vec{x}) = \dfrac{k\binom{k(n-1)}{\sum x_i - 1}}{\binom{kn}{\sum x_i}}$.

Suppose $V_T(\vec{x}) = \mathbb{E}[V(\vec{X}) \mid T(\vec{X})]$ is also an MVUE, where $V(\vec{X})$ is unbiased for $\tau(\theta)$, which we can do by Rao-Blackwell.

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 $W_T(\mathbf{X}) = \mathbb{E}[W(\mathbf{X}) \mid T(\mathbf{X})]$ is the unique UMVUE.

Proof. Suppose that $V(\vec{X})$ is a UMVUE of $\tau(\theta)$. Then $V_T(\vec{X}) := \mathbb{E}[V(\vec{X}) \mid T(\vec{X})]$ is also unbiased for $\tau(\theta)$, and $\mathrm{Var}(V_T(\vec{X})) \le \mathrm{Var}(V(\vec{X}))$ by Rao-Blackwell, so it too is a UMVUE. By Theorem 2.4, $V(\vec{X}) = V_T(\vec{X})$.

Then $0 = \mathbb{E}[V_T(\vec{X})] - \mathbb{E}[W_T(\vec{X})]$

$= \mathbb{E}_\theta\left[\mathbb{E}[V(\vec{X}) \mid T(\vec{X})]\right] - \mathbb{E}\left[\mathbb{E}[W(\vec{X}) \mid T(\vec{X})]\right]$

$= \mathbb{E}_\theta\left[\underbrace{\mathbb{E}_\theta[V(\vec{X}) - W(\vec{X}) \mid T(\vec{X})]}_{=: h(T(\vec{X}))}\right]$

$= \mathbb{E}_\theta[h(T(\vec{X}))] \quad \forall \theta \in \Theta.$

$\Rightarrow$ By completeness, $P_\theta(h(T(\vec{X})) = 0) = 1 \ \forall \theta \in \Theta$.

$\Rightarrow W_T(\vec{X}) = V_T(\vec{X})$

$= V(\vec{X})$.

So the UMVUE is $\mathbb{E}[W(\vec{X}) \mid T(\vec{X})]$. $\square$

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

We know that $(\bar{X}_n, S^2)$ is complete (eg, Ex 1.29, Theorem 1.28, A1 Q18).

Also $\bar{X}_n$ is unbiased for μ, and S^2 is unbiased for σ^2.

$\Rightarrow (\bar{X}_n, S^2)$ is unbiased for (μ, σ^2).

By Lehmann-Scheffé, $T(\vec{X}) = (\bar{X}_n, S^2)$ is the UMVUE for (μ, σ^2).

That's not the MLE of (μ, σ^2)!

Lehmann-Scheffé: Examples

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

We know that $\bar{X}_n$ is unbiased for λ, and it's also a complete sufficient statistic.

By Lehmann-Scheffé, $\bar{X}_n$ is the UMVUE of λ.

$=$

$\mathbb{E}[\bar{X}_n | \bar{X}_n]$

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. Apply Cauchy-Schwarz to $X = T(\vec{X})$ and $Y = U(\vec{X})$ and square everything. □

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

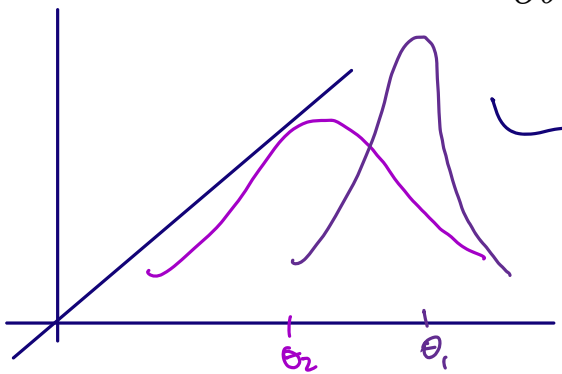
$$I_n(\theta) = \mathsf{Var}_\theta \left(S(\theta \mid \mathbf{X}) \right).$$

- 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

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

$= \frac{\partial}{\partial \theta} \ell(\theta \mid \vec{x})$

$= -\frac{\partial^2}{\partial \theta^2} \ell(\theta \mid \vec{x})$



The Fisher Information: Examples

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

$$L(\lambda|\vec{x}) = \prod_i \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$\Rightarrow \ell(\lambda|\vec{x}) = \Sigma x_i \cdot \log(\lambda) - n\lambda + c, \text{ where } c \in \mathbb{R} \text{ is free of } \lambda$$

$$\Rightarrow S(\lambda|\vec{x}) = \frac{\Sigma x_i}{\lambda} - n$$

$$I_n(\lambda) = \text{Var}_\lambda(S(\lambda|\vec{X}))$$
$$= \text{Var}_\lambda\left(\frac{\Sigma X_i}{\lambda} - n\right)$$
$$= \frac{1}{\lambda^2} \cdot \text{Var}_\lambda(\Sigma X_i)$$
$$= \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}.$$

For $J_n(\vec{x})$, recall $\hat{\lambda}_{MLE}(\vec{x}) = \bar{X}_n$.

Then $-\frac{\partial}{\partial\lambda} S(\lambda|\vec{x}) = \frac{\Sigma x_i}{\lambda^2}$, so

$$J_n(\vec{x}) = \left.\frac{\Sigma X_i}{\lambda^2}\right|_{\lambda = \bar{X}_n}$$
$$= \frac{n\bar{X}_n}{(\bar{X}_n)^2} = \frac{n}{\bar{X}_n} = \frac{n^2}{\Sigma X_i}.$$

The Fisher Information: Examples

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

From Ex 2.12, $S(\mu|\vec{x}) = \frac{\sum x_i - n\mu}{\sigma^2}$.

$$I_n(\mu) = \text{Var}_\mu\left(\frac{\sum X_i - n\mu}{\sigma^2}\right)$$
$$= \frac{1}{\sigma^4}\text{Var}_\mu(\sum X_i)$$
$$= \frac{n}{\sigma^2}.$$

Recall $\hat{\mu}_{mle}(\vec{x}) = \bar{X}_n$. Then

$$J_n(\vec{x}) = -\frac{\partial}{\partial \mu} S(\mu|\vec{x})\Big|_{\mu=\bar{x}_n}$$
$$= \frac{n}{\sigma^2}\Big|_{\mu=\bar{x}_n}$$
$$= \frac{n}{\sigma^2}.$$

They're usually not the same!

The Cramér-Rao Lower Bound

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

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

  Then

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

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

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

Proof.

The Cramér-Rao Lower Bound

Proof: In the Covariance inequality, let $U(\vec{X}) = S(\theta|\vec{X}) = \frac{\partial}{\partial\theta}\ell(\theta|\vec{X})$.

Then $\mathrm{Cov}_\theta(T(\vec{X}), S(\theta|\vec{X}))$

$= \underbrace{\mathbb{E}_\theta[T(\vec{X})\cdot S(\theta|\vec{X})]}_{①} - \mathbb{E}_\theta[T(\vec{X})]\cdot\underbrace{\mathbb{E}_\theta[S(\theta|\vec{X})]}_{②} = \frac{d}{d\theta}\mathbb{E}_\theta[T(\vec{X})] - 0 = \frac{d}{d\theta}\mathbb{E}_\theta[T(\vec{X})]$. ③

$$① = \int_{\mathcal{X}^n} T(\vec{x})\cdot S(\theta|\vec{x})\cdot f_\theta(\vec{x})\,d\vec{x}$$
$$= \int T(\vec{x})\left(\frac{\partial}{\partial\theta}\ell(\theta|\vec{x})\right) f_\theta(\vec{x})\,d\vec{x}$$
$$= \int T(\vec{x})\left(\frac{1}{f_\theta(\vec{x})}\cdot\frac{\partial}{\partial\theta}f_\theta(\vec{x})\right) f_\theta(\vec{x})\,d\vec{x}$$
$$= \int T(\vec{x})\cdot\frac{\partial}{\partial\theta}f_\theta(\vec{x})\,d\vec{x}$$
$$= \int \frac{\partial}{\partial\theta}\left(T(\vec{x})\cdot f_\theta(\vec{x})\right)d\vec{x}$$
$$\overset{\text{ass.}}{=} \frac{d}{d\theta}\int T(\vec{x})\cdot f_\theta(\vec{x})\,d\vec{x}$$
$$= \frac{d}{d\theta}\mathbb{E}_\theta[T(\vec{X})]$$

$$② = \int_{\mathcal{X}^n}\left(\frac{\partial}{\partial\theta}\log(f_\theta(\vec{x}))\right) f_\theta(\vec{x})\,d\vec{x}$$
$$= \int_{\mathcal{X}^n}\frac{1}{f_\theta(\vec{x})}\cdot\left(\frac{\partial}{\partial\theta}f_\theta(\vec{x})\right) f_\theta(\vec{x})\,d\vec{x}$$
$$= \int_{\mathcal{X}^n}\frac{\partial}{\partial\theta}f_\theta(\vec{x})\,d\vec{x}$$
$$\overset{\text{ass.}}{=} \frac{d}{d\theta}\int_{\mathcal{X}^n} f_\theta(\vec{x})\,d\vec{x}$$
$$= \frac{d}{d\theta}1$$
$$= 0$$

Also, by def, $\mathrm{Var}_\theta(S(\theta|\vec{X})) = I_n(\theta)$. Plug ③ into the covariance inequality and we're done! □

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: If $X_1, \ldots, X_n$ iid $N(\mu, 1)$, don't try $T(\vec{X}) = X_1/X_n$. Won't work!
- 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: $\text{Unif}(0, \theta) \Rightarrow$ Support $(\theta) = (0, \theta)$ depends on θ

  $\Rightarrow \frac{d}{d\theta} E_\theta[T(\vec{x})] \neq \int_0^\theta \left(\frac{\partial}{\partial \theta} T(\vec{x}) \cdot \frac{1}{\theta}\right) d\vec{x}$ in general.

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. $I_n(\theta) = \mathrm{Var}_\theta(S(\theta|\vec{X}))$ by def

$= \mathbb{E}[S(\theta|\vec{X})^2] - \mathbb{E}[S(\theta|\vec{X})]^2$

$= 0$ from the proof of CRLB

$= \mathbb{E}[S(\theta|\vec{X})^2]$.

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

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

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. RHS $= -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\log(f_\theta(\theta|x))\right)\right]$

$= -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\left(\frac{1}{f_\theta(x)}\cdot\frac{\partial}{\partial\theta} f_\theta(x)\right)\right]$

$= -\mathbb{E}_\theta[\ldots - \ldots]$

Exercise: you finish off!

Use the assumption somewhere..

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 $\tau(\theta)$**.

- What's the connection between UMVUEs and efficient estimators?

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

We need to calculate the CRLB for estimators of μ, and also $\text{Var}_\mu(T(\vec{X}))$, and show that they're equal.

We know that $\text{Var}_\mu(T(\vec{X})) = \text{Var}_\mu(\bar{X}_n) = \sigma^2/n$.

What about the CRLB? Numerator: $\left(\frac{d}{d\mu} E_\mu[T(\vec{X})]\right)^2 = \left(\frac{d}{d\mu}\mu\right)^2 = 1$.

Denominator: $I_n(\mu) = n/\sigma^2$ from Ex 2.26.

So the CRLB is $\frac{1}{n/\sigma^2} = \sigma^2/n = \text{Var}_\mu(\bar{X}_n)$.

So $T(\vec{X}) = \bar{X}_n$ is efficient 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. From the covariance inequality, equality holds in the CRLB iff

$$T(\vec{X}) = E_\theta[T(\vec{X})] + \frac{\mathrm{Cov}_\theta(T(\vec{X}), S(\theta|\vec{X}))}{\mathrm{Var}_\theta(S(\theta|\vec{X}))}\left(S(\theta|\vec{X}) - E_\theta[S(\theta|\vec{X})]\right)$$

$$= \tau(\theta) + \frac{[\tau'(\theta)]^2}{I_n(\theta)} \cdot S(\theta|\vec{X})$$

$$\text{iff } S(\theta|\vec{X}) = \underbrace{\frac{I_n(\theta)}{[\tau'(\theta)]^2}}_{=: a(\theta)}(T(\vec{X}) - \tau(\theta)). \qquad \square$$

Efficiency: Examples

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

If there did exist one, say $T(\vec{x})$, then there would also be some $a(\sigma^2)$ s.t.

$$S(\sigma^2|\vec{x}) = a(\sigma^2)\left(T(\vec{x}) - \sigma^2\right).$$

But $S(\sigma^2|\vec{x}) = \frac{n}{2\sigma^4}\left(\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} - \sigma^2\right)$. ← (Exercise)

By Theorem 2.12, the only candidate for $T(\vec{x})$ is $T(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$.

But μ is unknown! So $T(\vec{x})$ is not a point estimator.

So no efficient estimator of σ^2 exists.

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

$= e^{-\lambda}$.

$S(\lambda|\vec{x}) = \frac{\sum x_i}{\lambda} - n = \frac{\sum x_i}{\lambda} - n + e^{-\lambda} - e^{-\lambda}$. Clearly no efficient estimator of $e^{-\lambda}$ exists, using Theorem 2.12.

However, consider $W(\vec{X}) = \mathbb{1}_{X_1=0}$, which is unbiased for $\tau(\lambda)$. Also, $T(\vec{X}) = \bar{X}_n$ is a CSS for λ.

By Lehmann-Scheffé, $\mathbb{E}[W(\vec{X})|T(\vec{X})] = \mathbb{P}(X_1 = 1 | \bar{X}_n)$ is the UMVUE of $\tau(\lambda)$.

How do we use it? $n\bar{X}_n = \sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$ Exercise: use mgfs! Check that $\vec{X}|\sum X_i = t$ has pmf $\binom{t}{x_1, \ldots x_n}\left(\frac{1}{n}\right)^{x_1}\cdots\left(\frac{1}{n}\right)^{x_n} \sim \text{Multinomial}\left(t; \frac{1}{n}, \ldots, \frac{1}{n}\right)$.

$\Rightarrow X_1 | \sum X_i = t \sim \text{Bin}\left(t, \frac{1}{n}\right)$.

Therefore, $\mathbb{E}[\mathbb{1}_{X_1=0}|\bar{X}_n] = \mathbb{E}[\mathbb{1}_{X_1=0}|\sum X_i] = \mathbb{P}(X_1 = 0|\sum X_i) = \left(1 - \frac{1}{n}\right)^{\sum X_i}$.