Mathematical Statistics II

Point Estimation: Principles and Theory

Jesse Wheeler

Introduction

Introduction

- We now introduce principles of parameter estimation, as well as associated theory.
- The three approaches we have considered thus far (MoM, MLE, Bayes) are great starting points, but we would like to study general properties of parameter estimations, and how these approaches align with these properties.
- There is no "correct" order in which these material are presented, so we deviate a bit from our textbooks.
- This work is largely based on Rice (2007, Sections 8.5.2, 8.7–8.8), Casella and Berger (2024, Chapters 6, 7, 10), as well as some supplemental material from Pawitan (2001).

Fisher's Information

Curvature of the Likelihood Function

- Both the MLE and Bayesian estimates are forms of Likelihood based inference.
- For the MLE, we are interested in the maximum of the likelihood function.
- For Bayesian inference, the posterior distribution is the (scaled) product of the likelihood and a prior distribution:

$$\pi(\theta|x) \propto L(\theta) \times \pi(\theta).$$

- In both instances, the shape of the likelihood function $L(\theta)$ is of interest.

Curvature of the Likelihood Function II

- Coming up: We'll introduce the *score* function, and *Fisher's Information*. The latter can be used to obtain uncertainty estimates for *any* parameter estimated using the MLE technique.

Curvature of the Likelihood Function III

Definition: The score function

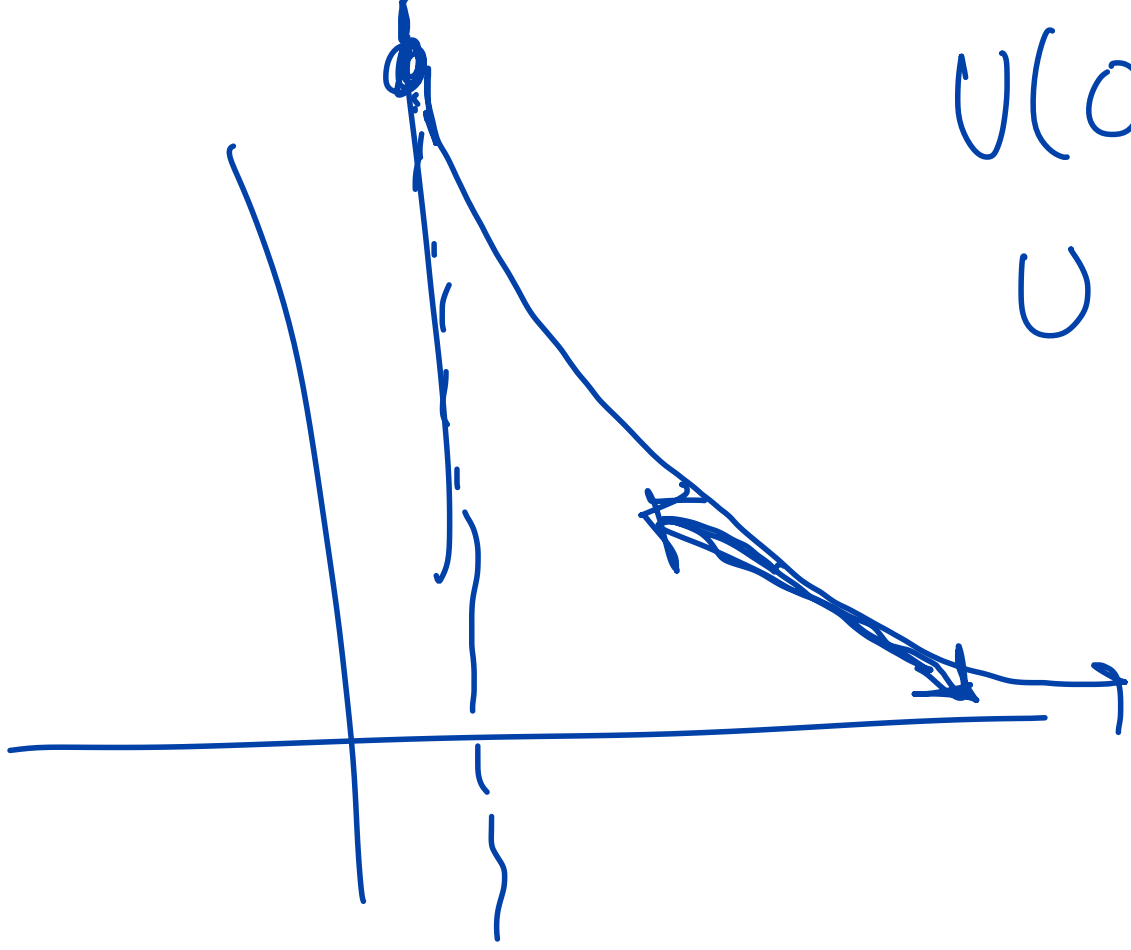
Let $L(\theta) = f(x^*; \theta)$ be the likelihood of some model, conditioned on the observed data (x^* is assumed to be multi-variate). Then we define the score function $S(\theta)$ to be:

$$S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta).$$

- Assuming the MLE $\hat{\theta}$ exists and is not a boundary point, then:

$$S\left(\hat{\theta}\right) = 0.$$

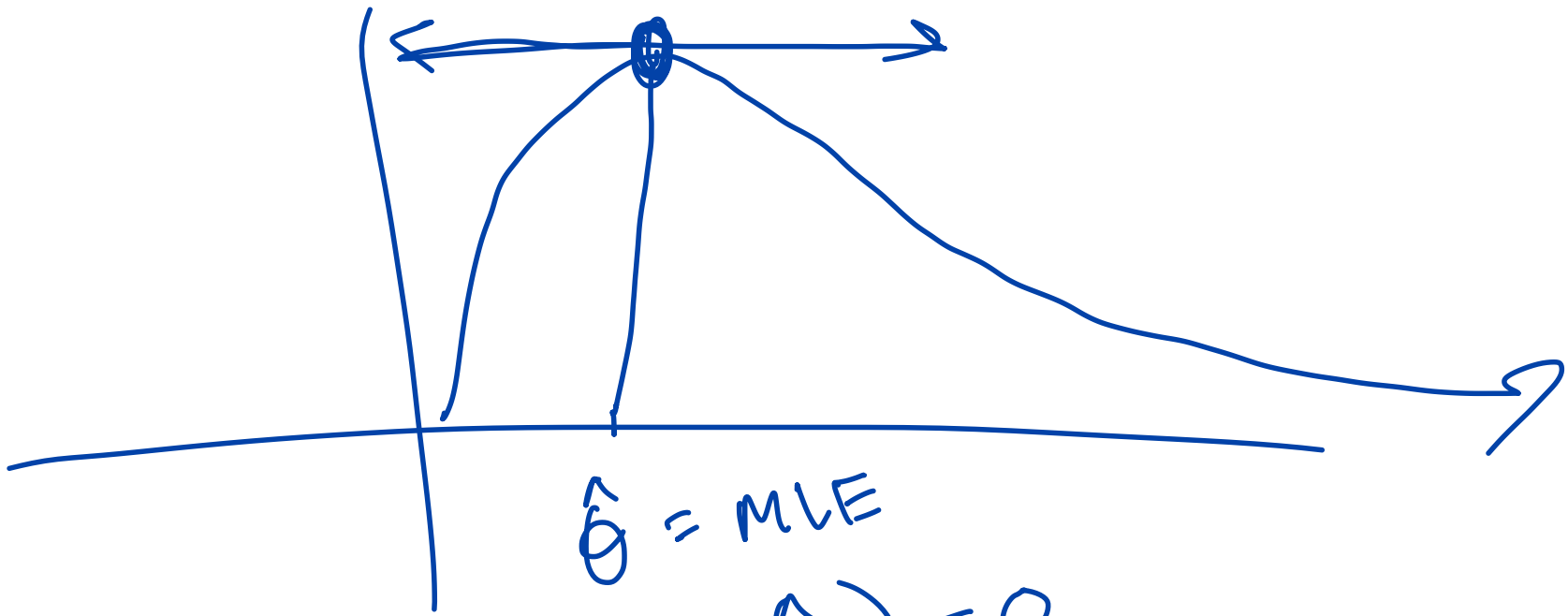
- As written, the score function is dependent on a fixed data set: x^*.



$U(0,\theta)$

$U(0,\theta]$

$\theta = X_{(n)}$


$\hat{\theta}$ = MLE
$S(\hat{\theta}) = 0$

Curvature of the Likelihood Function IV

- We could also expand this by considering it as a random function of any possible data X, which is random.

$$S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta)$$

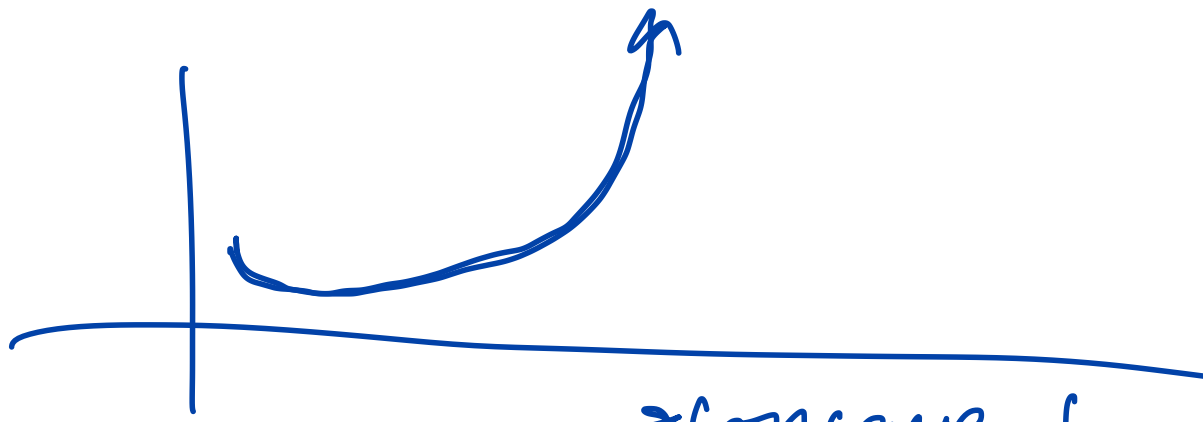
$$S(\theta; X) = \frac{\partial}{\partial \theta} \log L(\theta; X)$$

X is Random Variable

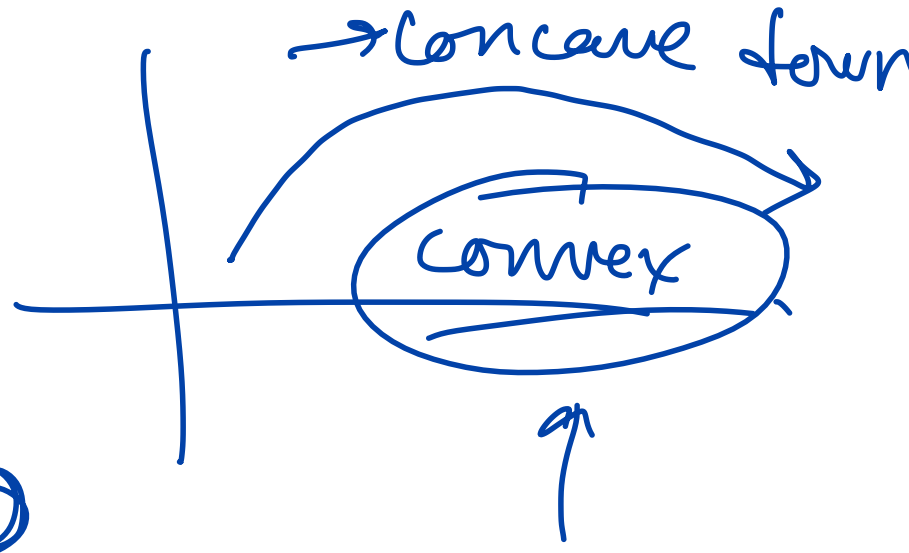
$E[g(X)]$

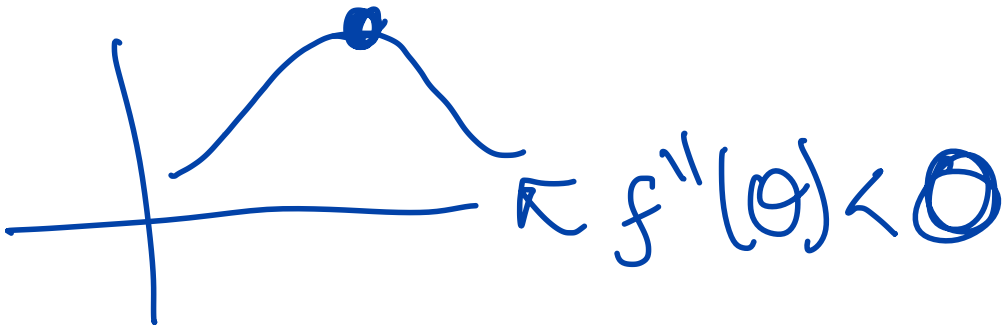
$f(\theta)$.

$f''(\theta) > 0 \iff$ Concave function



$f''(\theta) < 0 \iff$





Curvature of the Likelihood Function V

- Since we are interested in the curvature of the likelihood function, we should also consider the second derivative.
- At the MLE, the second derivative of the log-likelihood will be *negative*, so it's natural to consider the negative second derivative of the log-likelihood:

$$I_n(\theta) = -\frac{\partial^2}{\partial \theta^2} \log L(\theta).$$

- Again, we are treating the likelihood function $L(\theta)$ as a regular (non-random) function, fixed on x^* (this is slightly different than many texts).
- In this case, we call $I(\theta)$ the observed Fisher's Information, at the point θ.

Curvature of the Likelihood Function VI

Comment: joint information

We can define the information for just a single point $X_i = x_i^*$,

$$I_i(\theta) = -\frac{\partial^2}{\partial\theta^2} \log f_{X_i}(x_i^*; \theta),$$

or for the entire collection: $X_{1:n} = x_{1:n}^*$:

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2} \sum_i \log f_{X_i}(x_i^*; \theta) = \sum_i I_i(\theta).$$

- The curvature at the MLE is of particular interest.

Curvature of the Likelihood Function VII

Definition: The Observed Fisher Information

Let $L(\theta)$ be the likelihood function, fixed on the observed data x^*. Unless otherwise specified, the observed Fisher information is understood to be the negative second derivative of $\log L(\theta)$, evaluated at the MLE:

$$I\big(\hat{\theta}\big) = -\frac{\partial^2}{\partial\theta^2} \log L\big(\theta\big)\Big|_{\theta=\hat{\theta}}$$

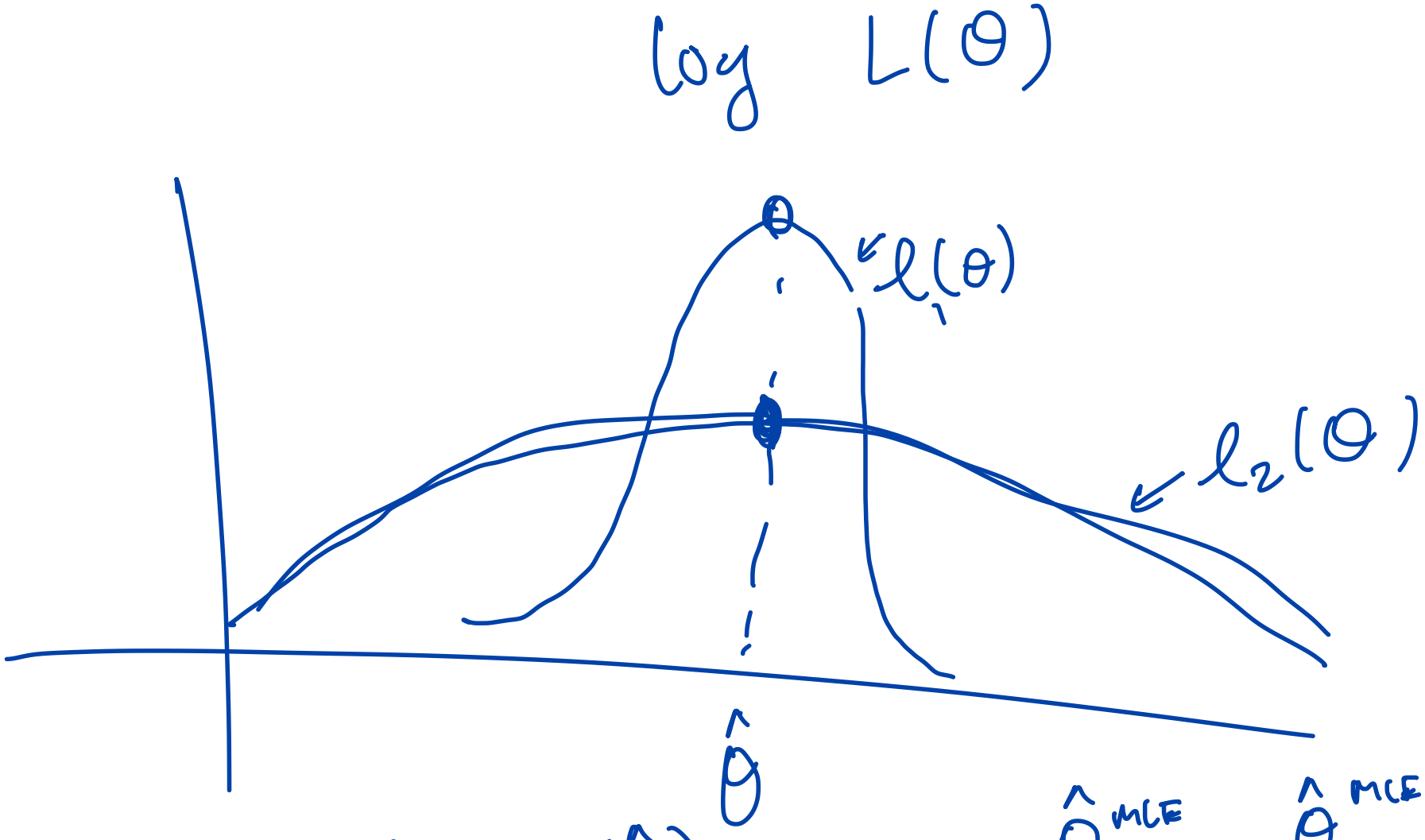
Note that when θ is multi-variate, this is a matrix with entries:

$$\Big[I\big(\hat{\theta}\big)\Big]_{i,j} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta)\Big|_{\theta=\hat{\theta}},$$

also denoted $I\big(\hat{\theta}\big) = -\nabla^2 \log L\big(\hat{\theta}\big)$.

$$I(\hat{\theta}) = \begin{bmatrix} -\frac{\partial^2}{\partial\theta_1 \partial\theta_1} \log L(\hat{\theta}) & -\frac{\partial^2}{\partial\theta_1 \partial\theta_2} \log L(\hat{\theta}) \\ -\frac{\partial^2}{\partial\theta_1 \partial\theta_2} \log L(\hat{\theta}) & -\frac{\partial^2}{\partial\theta_2 \partial\theta_2} \log L(\hat{\theta}) \end{bmatrix}$$

$$\frac{\partial^2}{\partial\theta_1 \partial\theta_2} f(\theta) = \frac{\partial^2}{\partial\theta_2 \partial\theta_1} f(\theta)$$



$$S_1(\hat{\theta}) = S_2(\hat{\theta}) = 0 \quad \Rightarrow \quad \hat{\theta}_1^{MLE} = \hat{\theta}_2^{MLE}$$

$$I^{(1)}(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\ell_1(\hat{\theta}) > -\frac{\partial^2}{\partial\theta^2}\ell_2(\hat{\theta}) = I^{(2)}(\hat{\theta})$$

Curvature of the Likelihood Function VIII

- Note that large values of $I(\hat{\theta})$ in the univariate case (and large eigen-values / determinant in the multivariate case) indicate a "tight peak" around the maximum, suggesting less uncertainty about the parameter θ.

Observed Fisher's information, Binomial Distribution

Let $X \sim \text{Binom}(n, \theta)$. Find the MLE by setting the score function equal to zero, and then use that to compute the observed Fisher's information.

$$L(\theta) = f(x^*; \theta)$$

$$= \binom{n}{x^*} \theta^{x^*} (1-\theta)^{n-x^*}$$

$$C = \log\binom{n}{x^*}$$

$$\ell(\theta) = C + x^* \log(\theta) + (n - x^*) \log(1-\theta)$$

Score function: $S(\theta)$:

$$S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta)$$

$$= \frac{x^*}{\theta} + \frac{(n-x^*)}{1-\theta}(-1) = 0$$

$$\Leftrightarrow$$

$$\frac{x^*}{\theta} = \frac{n-x^*}{1-\theta}$$

$$\hat{\theta} = \frac{x^*}{n}$$

$$I(\theta) = -\frac{\partial^2}{\partial \theta^2} \ell(\theta)$$

$$= -\frac{\partial}{\partial \theta} S(\theta) = \frac{-2}{\partial \theta}\left(\frac{x^*}{\theta} - \frac{n-x^*}{1-\theta}\right)$$

$$= \frac{x^*}{\theta^2} + \frac{n-x^*}{(1-\theta)^2}$$

$$I(\hat{\theta}) = \frac{x^*}{\hat{\theta}^2} + \frac{n - x^*}{(1-\hat{\theta})^2}$$

$$= \frac{n}{\hat{\theta}(1-\hat{\theta})}$$

$$= \frac{n}{\left(\frac{x^*}{n}\right)\left(1 - \frac{x^*}{n}\right)}$$

$X \sim \text{Bin}(n, \theta)$, $\quad \hat{\theta} = \frac{x}{n}$

$$n\hat{\theta} = X \sim \text{Bin}(n, \theta).$$

$$\text{var}(n\hat{\theta}) = n(\theta)(1-\theta)$$

$$n^2 \text{var}(\hat{\theta}) = n(\theta)(1-\theta)$$

$$\text{var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n}$$

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

Curvature of the Likelihood Function IX

Observed Fisher's information, Normal Distribution

Suppose $X_1, \ldots, X_n$ are iid from $N(\theta, \sigma^2)$, with known σ^2. Find the MLE by setting the score function to zero, and then calculate the joint information, $I_n(\theta)$ at the MLE.

Under iid,

$$L(\theta) = \prod_{i=1}^{n} f_{x_i}(x_i^*; \theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2} (x_i^* - \theta)^2 \right\}$$

$$= C \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^* - \theta)^2\right\}$$

$$\ell(\theta) = C_1 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^* - \theta)^2$$

$$S(\theta) = \frac{\partial}{\partial\theta}\ell(\theta) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\theta}\sum_{i=1}^{n}(x_i^* - \theta)^2$$

$$= \frac{1}{\sigma^2}\sum(x_i^* - \theta)$$

$$= \frac{1}{\sigma^2}\left(\sum x_i^* - n\theta\right) = 0$$

$$\Leftrightarrow$$

$$\sum x_i^* = n\hat{\theta}$$

$$\hat{\theta} = \frac{1}{n}\sum x_i^* = \bar{x}_n^*$$

$$\hat{\theta}_{MLE} = \bar{x}_n^* .$$

Observed Information .

$$I(\theta) = \frac{-\partial^2}{\partial\theta^2} \ell(\theta)$$

$$= -\frac{\partial}{\partial\theta} S(\theta)$$

$$= -\frac{\partial}{\partial\theta}\left(\frac{1}{\sigma^2}\sum(x_i - \theta)\right)$$

$\frac{1}{\sigma^2}\sum(x_i - \theta)$ → zero deriv.

$$= \frac{\partial}{\partial\theta}\frac{1}{\sigma^2}\sum\theta = \frac{\partial}{\partial\theta}\frac{n\theta}{\sigma^2}$$

$$= \frac{n}{\sigma^2}$$

$$\boxed{Var(\hat{\theta})} = Var(\bar{X})$$

$$= Var\left(\frac{1}{n}\sum X_i\right)$$

$$= \frac{1}{n^2}\sum Var(X_i)$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

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

$$\mathrm{Var}(\hat{\theta}) \rightarrow \frac{1}{E_\theta[I(\theta)]} \approx \frac{1}{I(\hat{\theta})}$$

Curvature of the Likelihood Function X

- Calculating the observed Fisher's Information at any given point is rather simple: Find the likelihood, take second derivative(s), plug in desired point.
- However, because data are random, we may want to consider the expected Fisher's information.

Curvature of the Likelihood Function XI

Theorem: Expected Score

Now consider treating the score function $S(\theta)$ as a random function, dependent on random input data X. That is:

$$S(\theta) = \frac{\partial}{\partial \theta} \log f(X; \theta).$$

Then, under mild regularity conditions, the expected score is zero:

$$E\big[S(\theta)\big] = 0.$$

- NOTE: This is the most common way to define the score (and later, information). This holds for either a single observation, or the entire collection of data.

Proof:

$$E[S(\theta)] = E\left[\frac{\partial}{\partial\theta}\log L(\theta)\right]$$

$$= E\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]$$

$$= \int\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right) f(x;\theta)\,dx$$

$$= \int\left(\frac{1}{f(x;\theta)}\cdot\frac{\partial}{\partial\theta} f(x;\theta)\right)\cdot f(x;\theta)\,dx$$

Assume:

We can swap. DLT.

$$= \int \frac{\partial}{\partial \theta} f(x;\theta)\, dx$$

$$= \frac{\partial}{\partial \theta} \int f(x;\theta)\, dx$$

$$= \frac{\partial}{\partial \theta} (1)$$

$$= 0$$

Curvature of the Likelihood Function XII

Definition: Fisher's Information

Continuing with the treatment of the score function as a random function of X, we will define the Expected Fisher's Information to be the following:

$$\mathcal{I}(\theta) = E\Big[\Big(\frac{\partial}{\partial\theta}\log f(X;\theta)\Big)^2\Big] = E\Big[\big(S(\theta)\big)^2\Big].$$

This is often referred to as just Fisher's Information.

$$\mathrm{Var}(S(\theta)) = E[(S(\theta))^2] - [E[S(\theta)]]^2$$

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2}\log(x^*;\theta) \quad \text{(obs)}$$

$$E[S(\theta)] = 0$$

if we can swap $\frac{\partial}{\partial\theta}\int dx$.

Curvature of the Likelihood Function XIII

Proposition: Variance of the Score

If the partial derivative and integral operations can be exchanged, then

$$\mathcal{I}(\theta) \overset{\checkmark}{=} \operatorname{Var}\big(S(\theta)\big) = -E\Big[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\Big],$$

where the score function is treated as random.

$$I(\theta) = E\left[\left(S(\theta)\right)^2\right]$$

$$= \operatorname{Var}(S(\theta)) = E\left[\left(S(\theta)\right)^2\right] - \left(E\left[S(\theta)\right]\right)^2$$

$$E\left[S(\theta)\right] = 0$$

(normal (μ, σ^2) model)

$$S(\theta) = \frac{1}{\sigma^2} \sum (x_i^* - \theta)$$

$$\underline{I(\theta)} = E\left[\left(\frac{1}{\sigma^2} \sum (X - \theta) \right)^2 \right]$$

$$= \frac{1}{\sigma^4} E\left[\left(n\bar{x} - n\theta \right)^2 \right]$$

$$= c\, E\left[\bar{x}^2 \right] + d(\theta)$$

$$\mathcal{I}(\theta) = E\left[(S(\theta))^2\right] \leftarrow \text{hard}$$

$$= -E\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] \downarrow$$

$$= -E\left[\frac{n}{\sigma^2}\right] = \frac{-n}{\sigma^2}$$

Curvature of the Likelihood Function XIV

- This proposition is useful because:
  - If the log-likelihood is differentiable, the first definition $\mathcal{I}(\theta) = E[S(\theta)^2]$ always exists.
  - Typically, we can express this as $\mathcal{I}(\theta) = \mathrm{Var}(S(\theta))$, which is a nice interpretable quantity.
  - The final expression $\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(X;\theta)\right]$ is usually the easiest to calculate.

Curvature of the Likelihood Function XV

- For the multivariate case, this is equivalent to writing:

$$\begin{aligned}\left[\mathcal{I}(\theta)\right]_{i,j} &:= E_\theta\left[\frac{\partial \log f(X;\theta)}{\partial \theta_i}\frac{\partial \log f(X;\theta)}{\partial \theta_j}\right] \\ &= \mathrm{Cov}\left(\frac{\partial \log f(X;\theta)}{\partial \theta_i}, \frac{\partial \log f(X;\theta)}{\partial \theta_j}\right) \\ &= -E\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j}\log f(X;\theta)\right]\end{aligned}$$

$$\mathcal{I}(\theta) = -E\begin{bmatrix} \frac{\partial^2}{\partial\theta_1\partial\theta_1}\log f(x;\theta) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\log f(x;\theta) \\ \frac{\partial^2}{\partial\theta_1\partial\theta_2}\log f(x;\theta) & \frac{\partial^2}{\partial\theta_2^2}\log f(x;\theta) \end{bmatrix}$$

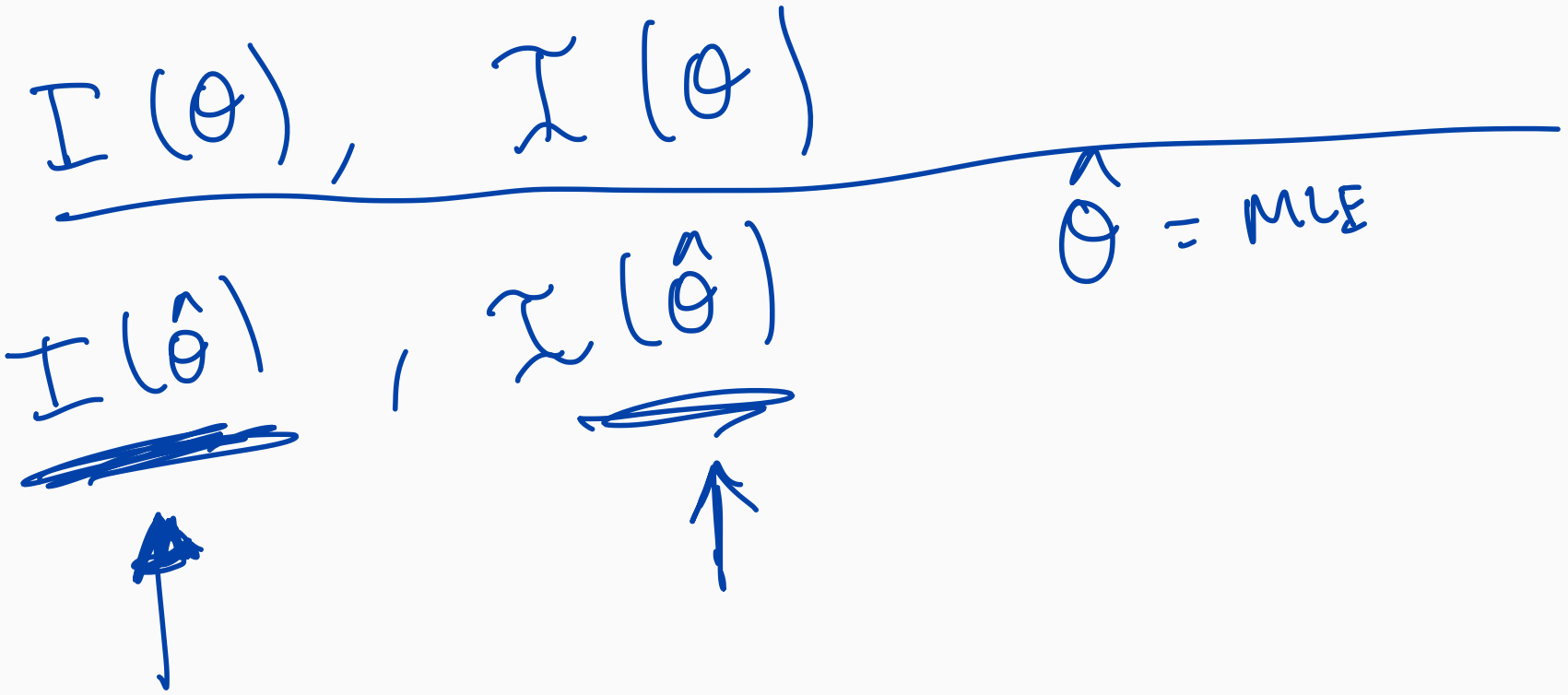
$$= \begin{bmatrix} -E\left[\frac{\partial^2}{\partial\theta_1 \partial\theta_2} \log f(x;\theta)\right] & \cdots \\ \cdots & \cdots \end{bmatrix}$$

Curvature of the Likelihood Function XVI

- Note that the curvature $I(\theta)$, defined as the second derivative of $L(\theta)$ (fixed function based on the observed data), exists for all points θ.
- Many theorem's instead rely on $\mathcal{I}(\theta)$, which is the expected Fisher's information. This later quantity "integrates out" the effect of the randomness in the data.
- In practice, the expected information is often hard to find, while the observed information is easy to calculate. For instance, in cases where the expected information is available, we might be able to get exact confidence intervals, which are preferred.
- In practice, we often replace $\mathcal{I}(\hat{\theta})$ in theory with $I(\hat{\theta})$.

Curvature of the Likelihood Function XVII

- However, we're primarily only interested in this function at the MLE, *or* when we regard $S(\theta)$ as a random function of arbitrary X (not fixed on x^*).



Curvature of the Likelihood Function XVIII

Comment on numeric optimization

If numeric optimization is used to find the MLE, many approaches estimate the Hessian of the loss-function. If the negative log-likelihood is used, setting `optim(..., hessian = TRUE)` can be used to compute the Observed Fisher's Information (don't forget to adjust for variable transformations!)

$$I(\hat{\theta}) = -\nabla^2 \log f(x^x; \hat{\theta})$$

Asymptotic properties of the MLE and Bayesian estimators

$$X_1^*, X_2^*, \ldots, X_n^* \sim f_{\theta_0}(x)$$

$$\uparrow$$

$$\theta_0$$

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$$

$$E[\theta \mid X^*] \xrightarrow{p} \theta_0.$$

Normality of MLE

- Under weak regularity conditions, both the MLE and Bayesian posterior converge to a Normal distribution, centered at the true value.
- We'll begin with a sketch proof of the distribution of the MLE.

Theorem: Asymptotic distribution of the MLE

Let $x_1^*, x_2^*, \ldots, x_n^*$ come from a family of distributions with density $f(x;\theta)$. If the true value of θ is θ_0, i.e., $X_i \overset{iid}{\sim} f(x;\theta_0)$, and we denote $\hat{\theta}$ as the MLE of θ, then

$$\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} Z$$

where $Z \sim N(0,1)$.

$$\tilde{I}_1(\theta) := E\left[\left(\frac{\partial}{\partial\theta} \log f(X_i;\theta) \right)^2 \right]$$

$$= -E\left[\frac{\partial^2}{\partial\theta^2} \log f(X_i;\theta) \right]$$

$$I(\theta) := E\left[\left(\frac{\partial}{\partial\theta} \log f_{X_{1:N}}(X_1,\ldots,X_n;\theta) \right)^2 \right]$$

$$\updownarrow$$

$$I_n(\theta)$$

Proof: $X_1, X_2, \ldots, X_n$.

$$L(\theta) := \prod_{i=1}^{n} f(x_i; \theta)$$

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta) = \underline{\text{MLE}}.$$

$$\ell'(\hat{\theta}) = 0$$

Taylor series expansion of $\ell'(\hat{\theta})$ around θ_0.

$$0 = \ell'(\hat{\theta}) = \ell'(\theta_0) + (\hat{\theta} - \theta_0)\ell''(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2 \ell'''(\theta^*)$$

$$\sqrt{n\, I_1(\theta_0)}\,(\hat{\theta} - \theta_0) \longrightarrow N \ldots$$

$$-(\hat{\theta}-\theta_0)\left[\ell''(\theta_0)+\frac{1}{2}(\hat{\theta}-\theta_0)\ell'''[\theta^*]\right]=\ell'(\theta_0)$$

$$\sqrt{n}\,(\hat{\theta}-\theta_0)=\frac{\sqrt{n}\,\ell'(\theta_0)}{-\ell''(\theta_0)-\frac{1}{2}(\hat{\theta}-\theta_0)\ell'''(\theta^*)}$$

$$\sqrt{n}\,(\hat{\theta}-\theta_0)=\frac{\frac{1}{\sqrt{n}}\ell'(\theta_0)}{-\frac{1}{n}\ell''(\theta_0)-\frac{1}{2n}(\hat{\theta}-\theta_0)\ell'''(\theta^*)}$$

$$\sqrt{n}\,(\hat{\theta}-\theta_0)=\frac{W_n}{Y_n}$$

W_n converges in dist. W,

Y_n converges in prob Y,

$$\frac{W_n}{Y_n} \xrightarrow{d} \frac{W}{Y}$$

Numerator (W_n):

$$\frac{1}{\sqrt{n}} l'(\theta_0) = \frac{1}{\sqrt{n}} \sum \frac{\partial}{\partial \theta} \log(\theta_0)$$

$$= \frac{1}{\sqrt{n}} \sum \frac{\partial}{\partial \theta} \log(\theta_0, X_i)$$

CLT: $\frac{1}{\sqrt{n}} \sum \frac{\partial}{\partial \theta} \log(\theta_0 ; X_i)$

Converges in Dist'n to:

$$E\left[\frac{\partial}{\partial \theta} \log(\theta_0 ; X_i)\right]$$

$$= E[S_1(\theta)] = 0$$

$$\text{Var}\left(\frac{1}{\sqrt{n}} \sum \frac{\partial}{\partial \theta} \log f(\theta_0 ; X_i)\right) \to E\left[S_1(\theta_0)^2\right]$$

$$= \mathcal{I}_1(\theta_0)$$

$$W_n \xrightarrow{d} N(0, \mathcal{I}_1(\theta_0)).$$

Y_n (denominator) :

$$\underbrace{-\frac{1}{n}\ell''(\theta_0)} - \underbrace{\frac{1}{2n}(\hat{\theta}-\theta_0)\ell'''(\theta^*)}$$

$\xrightarrow{P}$ 0

$$-\frac{1}{n}\ell''(\theta_0) = -\frac{1}{n}\sum \frac{\partial^2}{\partial\theta_0^2}\log f(X_i;\theta_0)$$

$$\xrightarrow[P]{WLLN} -E\left[\frac{\partial^2}{\partial\theta_0}\log f(X_i;\theta_0)\right]$$

$$= -I_1(\theta_0)$$

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) = \frac{W_n}{Y_n} \to \frac{W}{Y},$$

$$\underline{W \sim N\left(0, \mathcal{I}_1(\theta_0)\right)}$$

$$Y = \mathcal{I}_1(\theta_0),$$

$$\frac{W}{Y} \sim \frac{1}{\mathcal{I}_1(\theta_0)} N\left(0, \mathcal{I}_1(\theta_0)\right)$$

$$\sim N\left(0, \frac{1}{\mathcal{I}_1(\theta_0)}\right)$$

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}_1^{-1}(\theta_0)\right)$$

$$\sqrt{n\mathcal{I}_1(\theta_0)}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(0,1)$$

□.

Normality of MLE II

- How do we interpret / use the result above?
- For large iid samples: the MLE $\hat{\theta}$ is approximately normally distributed, centered at the true value θ_0, with shrinking variance:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{n\mathcal{I}(\theta_0)}.$$

- Note that there is sometimes some confusion regarding whether or not $\mathcal{I}$ is the information from *a single* observation, or *all observations*.
- Because the data are IID, it's really just a difference in scaling constants.

X_i are iid:

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\ell_i(\theta; x_i)\right]$$

$$= -E\left[\sum \frac{\partial^2}{\partial\theta^2} \ell_i(\theta; X_i)\right]$$

$$= -\sum E\left[\frac{\partial^2}{\partial\theta^2} \ell_i(\theta; X_i)\right]$$

$$= \sum \mathcal{I}_1(\theta)$$

$$\mathcal{I}(\theta) = n\,\mathcal{I}_1(\theta)$$

Normality of MLE III

- In our definition and proof, we let $\mathcal{I}(\theta_0)$ be the information from a single observation:

$$\mathcal{I}(\theta_0) = E\Big[\Big(\frac{\partial}{\partial\theta}\log f(X_i;\theta)\Big)^2\Big].$$

- We could have instead considered the likelihood / information using all of the data points:

$$\mathcal{I}_n(\theta_0) = E\Big[\Big(\frac{\partial}{\partial\theta}\log f_{X_1:X_n}(X_1,\ldots,X_n;\theta)\Big)^2\Big].$$

Normality of MLE IV

- In the later case, the theorem would have stated:

$$\sqrt{\mathcal{I}_n}(\hat{\theta} - \theta_0) \xrightarrow{d} Z \sim N(0, 1),$$

  or that

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{\mathcal{I}_n(\theta_0)}$$

- In the IID case, these are the same, as

$$\begin{aligned}\mathcal{I}_n(\theta) &= -E\Big[\frac{\partial^2}{\partial\theta^2}\sum_i \log f(X_i;\theta)\Big] \\ &= -\sum_i E\Big[\frac{\partial^2}{\partial\theta^2}\log f(X_i;\theta)\Big] = n\mathcal{I}(\theta).\end{aligned}$$

Normality of MLE V

- Our derivation treated θ as univariate. The more common multivariate case has a similar result.

Multivariate parameters

If $\theta \in \mathbb{R}^d$ is multivariate, than using the matrix definition of $\mathcal{I}(\theta)$, then:

$$\sqrt{n}\big(\hat{\theta} - \theta_0\big) \xrightarrow{d} Z_d \sim N_d\big(0_d, \mathcal{I}(\theta_0)^{-1}\big),$$

where $\hat{\theta}$ is the MLE, θ_0 is the true parameter value, and $N_d(\mu, \Sigma)$ is a d-dimensional multivariate normal distribution with mean vector μ and covariance matrix Σ.

Implications for estimating parameter uncertainty

- This result gives us an immediate way to approximate the uncertainty associated with the MLE, for rather arbitrary models.
- If $\hat{\theta}$ is the MLE, then:

$$\text{asy. Var}(\hat{\theta}) = \frac{1}{n\mathcal{I}_1(\theta_0)}$$

$$\text{Var}(\hat{\theta}) \approx \frac{1}{n\mathcal{I}(\hat{\theta})} = \frac{1}{\mathcal{I}_n(\hat{\theta})}.$$

- There are two different approximations here:
  - Asymptotic normality.
  - $\hat{\theta} \xrightarrow{p} \theta_0$ (Slutksy's Theorem).
- It is a good approximation for even moderate n, but exact calculations are preferred when possible.

expectation

Implications for estimating parameter uncertainty II

- Finally, recall that $\mathcal{I}(\hat{\theta})$ is an expectation.
- When an exact calculation is not readily available, calculating the expectation $\mathcal{I}(\hat{\theta})$ usually isn't either!
- In practice, we often replace the Fisher's Information with the Observed Fisher's information:

$$\mathrm{Var}(\hat{\theta}) \approx \frac{1}{I_n(\hat{\theta})},$$

where I_n is the information provided by all n data points.

- The details justifying this final approximation are a bit tricky. See Keener (2010, Chapter 9) for more details.
- In short, for many distributions (regular exponential families), $\mathcal{I}(\hat{\theta}) = I(\hat{\theta})$. In other cases, suitable conditions imply that $I(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\hat{\theta})$.

Wald - Confidence intervals -

$$\hat{\theta}_{MLE} \pm z_{\alpha/2} \cdot \frac{1}{I(\hat{\theta})}$$

95% CI: ~~(approximate CI)~~:

$$\hat{\theta}_{MLE} \pm 1.96 \cdot \left[I(\hat{\theta}) \right]^{-1}$$

Asymptotic results for Bayesian posteriors

- Similar results hold for Bayesian estimates.

Bernstein–von Mises Theorem

We are not ready to state the full Theorem, so we'll just provide a heuristic approach: under suitable conditions, the posterior distribution behaves asymptotically like a normal distribution with mean $\hat{\theta}_{MLE}$, and variance $1/\mathcal{I}(\theta_0)$.

Sketch:

$$\pi(\theta|x) \propto L(\theta)\pi(\theta)$$
$$\approx L(\theta)$$
$$= \exp(\log L(\theta))$$

$$\simeq \exp\left(\ell(\hat{\theta}) + (\theta - \hat{\theta})\,\ell'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}) \right)$$

$$\propto \exp\left(\frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}) \right)$$

$$= \exp\left(-\frac{1}{2\left(-\ell''(\hat{\theta})\right)^{-1}} (\theta - \hat{\theta})^2 \right)$$

$\Rightarrow$ pdf of

$$N\left(\hat{\theta}, \frac{1}{I(\hat{\theta})}\right)$$

MLE and Bayesian Posteriors, with enough data, behave approximately normal, with mean centered at truth θ_0, variance $(I(\theta_0))^{-1}$.

Bias, MSE, and Consistency

Bias

- We have already talked a little about bias last semester.
- It's an important concept / principle for point-estimates.

Definition: Bias of an estimator

The bias of a point estimator $\hat{\theta}$ of a parameter θ is the difference between the expected value of $\hat{\theta}$ and θ; that is,

$$\text{Bias}_\theta\left[\hat{\theta}\right] = E_\theta\left[\hat{\theta}\right] - \theta$$

- If $\text{Bias}_\theta\left[\hat{\theta}\right] = 0$, then $\hat{\theta}$ is an unbiased estimator of θ, and $E_\theta\left[\hat{\theta}\right] = 0$.

Bias II

- unbiasedness is a nice property, but alone does not mean we have a good estimator.
- Another thing that needs to be considered is the variance of an estimator.
- We encountered this with Monte-Carlo methods: many approaches converge to the true value, but lower variance estimates are generally preferred among unbiased methods.

Bias III

Unbiased estimators for normal distribution

Let $X_1, X_2, \ldots, X_n$ be iid from $N(\mu, \sigma^2)$. Consider two estimates of the parameter $\theta = \mu$:

$$\hat{\theta}_1 = \bar{X}_n, \qquad \hat{\theta}_2 = X_1.$$

Show that both are unbiased estimators, but that $\hat{\theta}_1$ has lower variance.

MSE

- The mean square error (MSE) describes the expected squared difference between an estimator and the true parameter value.
- Often, attempts as lowering the *variance* results in increasing the *variance*.
- It can be used to balance the bias-variance tradeoff.

Definition + Theorem: MSE

Let $\hat{\theta}$ be an estimator of θ. The MSE of $\hat{\theta}$ is defined as:

$$\mathrm{MSE}_\theta\big(\hat{\theta}\big) := E\big[(\hat{\theta} - \theta)^2\big].$$

Furthermore, $\mathrm{MSE}_\theta\big(\hat{\theta}\big) = \mathrm{Var}_\theta\big(\hat{\theta}\big) + \big(\mathrm{Bias}_\theta(\hat{\theta})\big)^2$.

MSE II

Example: Variance of MLE

Let X_i be iid from $N(\mu, \sigma^2)$, and consider estimating $\theta = (\mu, \sigma^2)$ in two ways:

$$\hat{\theta}_{MLE} = \left(\bar{X}_n, \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{X} - \sigma^2\right)^2\right),$$

$$\hat{\theta}_1 = \left(\bar{X}_n, S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{X} - \sigma^2\right)^2\right).$$

What is the MSE of $\hat{\mu}_{MLE}$? What about the bias and MSE of S^2 and $\hat{\sigma}^2$?

MSE III

- Based on the examples above, should $\hat{\sigma^2}$ be used instead of S^2? The argument above proves that, on average, $\hat{\sigma}^2_{MLE}$ will be closer to σ^2 than S^2.
- However: statisticians generally agree that S^2 is a better estimate. Why?
- Though it has lower MSE, on average, $\hat{\sigma}^2_{MLE}$ will be *smaller* than σ^2 (biased low).
- There's also some agreement that MSE is *not* a reasonable criterion for *scale* parameters, like the variance.
- The reason is that the MSE is a symmetric criterion: both small and large errors are penalized the same.

MSE IV

- However, scale parameters like σ^2 must be positive, so they have a natural lower bound $\sigma^2 > 0$, so the estimation problem is *not* symmetric.

MSE V

- The MSE_θ uses the subscript θ to denote that it is a function of θ.
- Often, there is not a uniformly best estimator for θ, and some estimators will perform better for some values of θ and worse for others.

MSE VI

MSE of MLE and Bayes: Binomial Model

(Example 7.3.5 Casella and Berger, 2024) Let X_i be iid Bernoulli(p), such that $X = \sum_i X_i$ is Binomial(n, p). The MLE $\hat{p} = X/n$ is unbiased for p, and has MSE:

$$\text{MSE}_p(\hat{p}) = \text{Var}_p(\hat{p}) = \frac{p(1-p)}{n}.$$

Compare this to the posterior mean estimate, where the prior $P \sim U(0, 1)$.

$\text{Var}_p(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot n(p)(1-p)$

$\text{MSE}_p(\hat{p}_{MLE}) = \frac{p(1-p)}{n}$

$$p|x \sim \text{Beta}(x+1, n-x+1)$$

$$\hat{p}_B = E[p|x] = \frac{x+1}{n+2}$$

$$MSE_p(\hat{p}_B) = Var_p(\hat{p}_B) + \left(Bias_p(\hat{p}_B)\right)^2$$

$$\hat{p}_B = \frac{x}{n+2} + \frac{1}{n+2}$$

$$E[\hat{p}_B] = \frac{1}{n+2}(np) + \frac{1}{n+2} = \frac{1}{n+2}(np+1)$$

$$\text{Bias}_p(\hat{p}_B) = \left(\frac{np+1}{n+2} - p \right)$$

$$\text{Var}_p(\hat{p}_B) = \frac{1}{(n+2)^2} np(1-p)$$

$$\text{MSE}_p(\hat{p}_B) = \frac{np(1-p)}{(n+2)^2} + \left(\frac{np+1}{n+2} - p \right)^2$$

Uniformly Best Unbiased Estimators

- As the previous example demonstrates, there is almost *never* a uniformly best estimate in terms of MSE.
- This is because the class of estimators is "too large".
- Example: consider estimating θ, using $\hat{\theta} = 17$. If $\theta = 17$, this will always beat any other estimator. However, it's a very poor estimate if $\theta \neq 17$.
- This leads us to thinking of ways of limiting the possible estimators to certain classes of estimators.
- A natural constraint to consider is the class of unbiased estimators. In this class, the MSE = Variance of the estimator.

Uniformly Best Unbiased Estimators II

Definition: Uniformly best unbiased estimators

An estimator θ^* is called the *best unbiased estimator* of θ if $E[\theta^*] = \theta$, for any other estimator $\tilde{\theta}$ such that $E[\tilde{\theta}] = \theta$, then $\mathrm{Var}_\theta(\theta^*) \leq \mathrm{Var}_\theta(\tilde{\theta})$ for all θ.

θ^* is also called a uniform minimum variance unbiased estimator (UMVUE) of θ.

- Initially, it can be extremely difficult to find a UMVUE estimator without adidtional supporting theory.

Uniformly Best Unbiased Estimators III

Poisson unbiased estimation

Let X_i be iid Poisson(λ) random variables, and let $\bar{X}$ and S^2 be the sample mean and variance, respectively. Since $E[X_i] = \mathrm{Var}(X_i) = \lambda$, it can be shown that $E[\bar{X}] = E[S^2] = \lambda$, and therefore both $\bar{X}$ and S^2 are unbiased estimators of λ. Which is a better estimate?

- Since they are unbiased, we just need to find the lower variance of the two. Evaluating $\mathrm{Var}_\lambda\left(S^2\right)$ is actually a very lengthy calculation.

Uniformly Best Unbiased Estimators IV

- However, even if we do the calculation and find $\mathrm{Var}_\lambda(\bar{X}) \leq \mathrm{Var}_\lambda(S^2)$, this doesn't immediately imply that $\bar{X}$ is the UMVUE. Consider for instance:

$$\hat{\lambda} = a\bar{X} + (1-a)S^2.$$

- This estimate is also unbiased, so from this alone we have an infinite number of unbiased estimators to consider.
- The next Theorem provides a lower bound on the variance of an unbiased estimator.

Uniformly Best Unbiased Estimators V

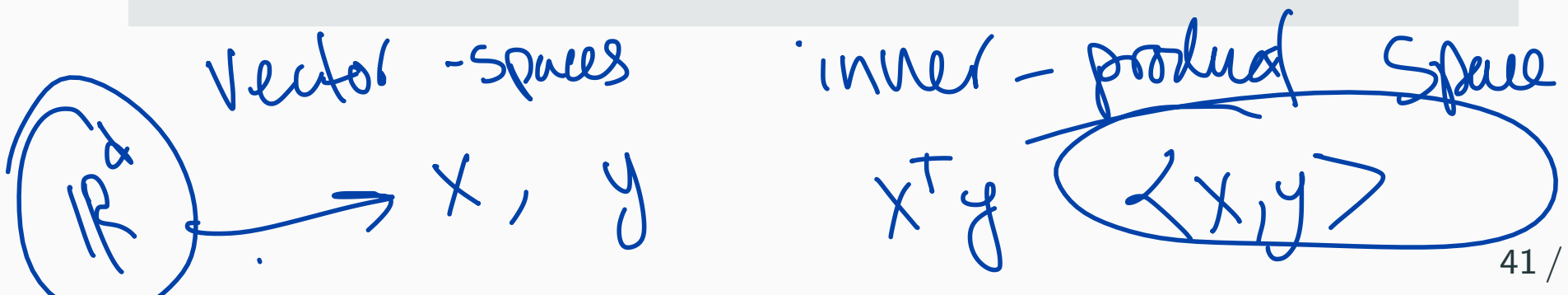
Theorem: Cramér-Rao Inequality

Let $X_1, \ldots, X_n$ be a sample with joint pdf $f_\theta(X;\theta)$, and let θ^* be any estimator of θ (biased or unbiased), that satisfies the smoothness condition

$$\frac{d}{d\theta} E_\theta[\theta^*] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} \theta^* f_\theta(X;\theta)\, dx.$$

Then,

$$\mathrm{Var}_\theta(\theta^*) \geq \frac{\left(\frac{d}{d\theta} E_\theta[\theta^*]\right)^2}{E_\theta\left[\left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2\right]} = \frac{\left(\frac{d}{d\theta} E_\theta[\theta^*]\right)^2}{\mathcal{I}(\theta)}.$$



X, Y r.v.'s

$$E[XY] = \langle X, Y \rangle$$

$$\star \quad |E[XY]|^2 \leq E[X^2]\,E[Y^2]$$

$$|\langle X, Y \rangle|^2 \leq \langle X, X \rangle \langle Y, Y \rangle$$

$$(\mathrm{Cov}(Z, Y))^2 \leq (\mathrm{Var}(Z))(\mathrm{Var}(Y))$$

Uniformly Best Unbiased Estimators VI

- In the Cramér-Rao inequality, consider what happens if θ^* is an unbiased estimator of θ:

$$\left(\frac{d}{d\theta}E_\theta[\theta^*]\right)^2 = \left(\frac{d}{d\theta}\theta\right)^2 = 1.$$

- This leads to the common "Cramér-Rao lower bound", for unbiased estimators:

$$\mathrm{MSE}_\theta(\theta^*) = \mathrm{Var}_\theta(\theta^*) \geq \frac{1}{\mathcal{I}(\theta)}.$$

Uniformly Best Unbiased Estimators VII

- The Cramér-Rao Inequality can at times be used to find UMVUEs (which are not necessarily unique).

Example: UMVUE for Poisson model

Let X_i be iid from a Poisson(λ) distribution. Both $\bar{X}$ and S^2 (the sample mean and variance, respectively) are unbiased estimators of λ. This implies that any estimator $a\bar{X} + (1-a)S^2$ with $a \in [0,1]$ is an unbiased estimator, among other possibilities.

If θ^* is *some* unbiased estimator, find the minimal possible variance of θ^*, and use that to find a UMVUE estimator of θ.

Uniformly Best Unbiased Estimators VIII

- One weakness of the Cramér-Rao inequality is that the lower bound, $\frac{1}{\mathcal{I}(\theta)}$, might not be tight.
- That is, it gives us a minimum quantity to compare to, but there may not exist an unbiased estimator that attains the lower-bound.
- There are attainment theorems, stating when the Cramér-Rao inequality is attainable and when it is not. We won't present any of these in this class, but it's closely related to *exponential families*.
- Instead, we can use some of our existing theorems to provide asymptotic results.

Consistency

- Before continuing our discussion about asymptotic results, we will introduce a few new definitions and principles of point estimation.

Consistency II

Definition: Consistency

Estimators of θ are really sequences of random variables: θ_n^*.
Example: estimate $E[X_i] = \mu$ with $\bar{X}_n = \frac{1}{n}\sum_i X_i$.

One desirable feature of a sequence of estimators θ_n^* is that as n grows large, θ_n^* converges in some way to the truth θ.

We say that an estimator θ_n^* is consistent if it converges in probability to the true value θ:

$$\lim_{n\to\infty} P_\theta\big(|\theta_n^* - \theta| \geq \epsilon\big) = 0.$$

Consistency III

- When trying to show convergence in probability last semester, we often used Chebychev's inequality:

$$P_\theta\big(|\theta_n^* - \theta| \geq \epsilon\big) \leq \frac{E_\theta\Big[(\theta_n^* - \theta)^2\Big]}{\epsilon^2}.$$

- Thus, if for every value of θ, we have:

$$\lim_{n\to\infty} E_\theta\Big[(\theta_n^* - \theta)^2\Big] = 0,$$

  then we know that θ_n^* is a consistent estimator of θ.

- Note that this condition is our definition of MSE!

$$\mathsf{MSE}_\theta\big(\theta_n^*\big) = E_\theta\Big[(\theta_n^* - \theta)^2\Big] = \mathrm{Var}_\theta\big(\theta_n^*\big) + \Big(\mathsf{Bias}_\theta\big(\theta_n^*\big)\Big)^2.$$

Consistency IV

- From this, we get an immediate result:

Theorem: Convergence in MSE

If θ_n^* is a sequence of estimators for θ, then if for every value of θ,

1. $\lim_{n\to\infty} \mathrm{Var}_\theta\left(\theta_n^*\right) = 0$
2. $\lim_{n\to\infty} \mathrm{Bias}_\theta\left(\theta_n^*\right) = 0$

then θ_n^* is a consistent estimator for θ.

Consistency V

Consistency of the MLE

Under suitable conditions on the likelihood function, the MLE $\hat{\theta}$ is a consistent estimator of θ.

A simple application of the Bernstein-von Misses Theorem implies that the Bayesian posterior mean is also a consistent estimator of θ.

$$\sqrt{I_n(\theta)}\,(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,1), \text{ or}$$

$$(\hat{\theta}) \xrightarrow{d} N\left(\theta_0, \frac{1}{I_n(\theta)}\right)$$

$$I_n(\theta) = nI(\theta)$$

Principles of parameter est.

- Invariance: transformations.

$\psi = g(\theta)$, and $\hat{\psi} = g(\hat{\theta})$

→ MLE has invariance property.

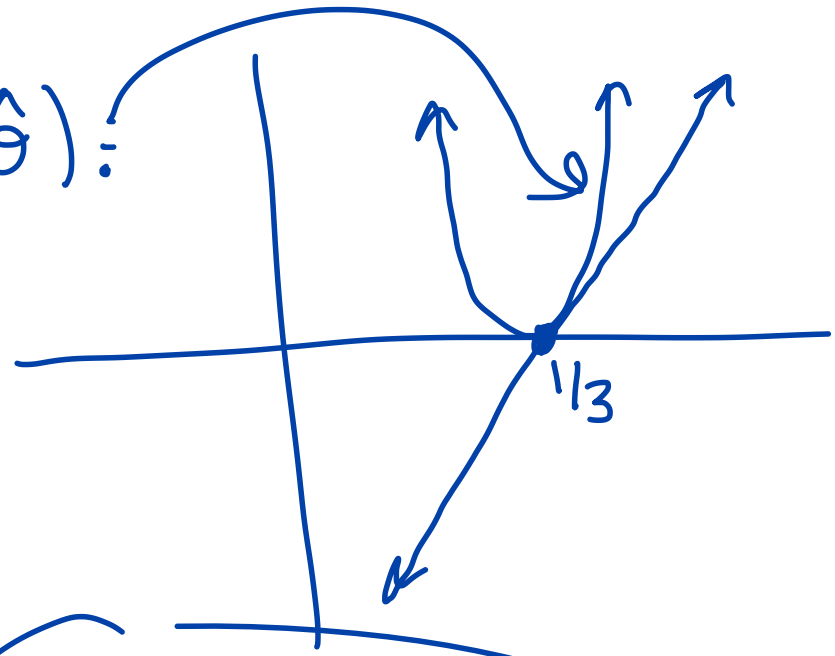
$\hat{\theta} = \frac{\sum x_i}{n}$, $\psi = \log\left(\frac{\theta}{1-\theta}\right)$

$\hat{\psi} = \log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right)$

Bias: $\hat{\theta}$, $E_\theta[\hat{\theta}] - \theta = \text{Bias}_\theta(\hat{\theta})$.

$\hat{\theta} = \frac{1}{3}$ $\quad \text{Bias}_\theta(\hat{\theta})$:

$\theta = 1/3$



MSE: $\text{MSE}_\theta(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$

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

Consistency: $\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$

$$\lim_{n\to\infty} P_\theta\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$$

$\hat{\theta}_n = \frac{1}{n}\sum X_i$

Chebychev's ineq.

$$P_\theta\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \leq \frac{E_\theta\left[(\hat{\theta} - \theta)^2\right]}{\varepsilon^2}$$

Efficiency

Definition: relative efficiency

If θ^* and $\tilde{\theta}$ are two estimators of θ, then the relative efficiency of θ^* with respect to $\tilde{\theta}$ is defined as:

$$\text{RE}\left(\theta^*, \tilde{\theta}\right) = \frac{\text{Var}\left(\tilde{\theta}\right)}{\text{Var}\left(\theta^*\right)}.$$

- If RE > 1, then θ^* is said to be a more efficient estimator than $\tilde{\theta}$
- If RE < 1, then θ^* is said to be less efficient than $\tilde{\theta}$.
- If θ^* is more efficient than any other estimator, we call θ^* an efficient estimator.

Efficiency II

Definition: Limiting variance and relative efficiency

An estimator θ_n^* of θ depends on the data, and the number of observations. If ~~$\lim_{n\to\infty} k_n \theta_n^* = \sigma^2$~~ for some sequence k_n and constant σ^2, then σ^2 is called the limiting variance of the estimator θ^*.

$$\lim_{n\to\infty} k_n \operatorname{Var}(\theta_n^*) = \sigma^2$$

If $\tilde{\theta}_n^*$ is a different estimate of θ with limiting variance τ^2, then the relative asymptotic efficiency of θ_n^* with respect to $\tilde{\theta}_n$ is defined as:

$$\operatorname{ARE}(\theta_n^*, \tilde{\theta}) = \frac{\tau^2}{\sigma^2}$$

$X_i \sim \text{pois}(\lambda)$, $\hat{\lambda} = \frac{1}{n}\sum X_i$

$\operatorname{Var}(\hat{\lambda}) = \frac{\lambda}{n}$, $k_n = n$,

$n \operatorname{Var}(\hat{\lambda}) = \lambda$

Efficiency III

Theorem: asymptotic efficiency of MLE and posterior mean

Under suitable regularity conditions, MLE $\hat{\theta}$ for parameter vector θ is asymptotically efficient, achieving the Cramér-Rao lower bound $\mathcal{I}(\theta)$.

By the Bernstein-von Misses Theorem, posterior distribution in a Bayesian analysis asymptotically converges to a normal distribution with mean $\hat{\theta}$, and variance $1/\mathcal{I}(\theta)$. Consequently, the posterior mean is also asymptotically efficient.

Efficiency IV

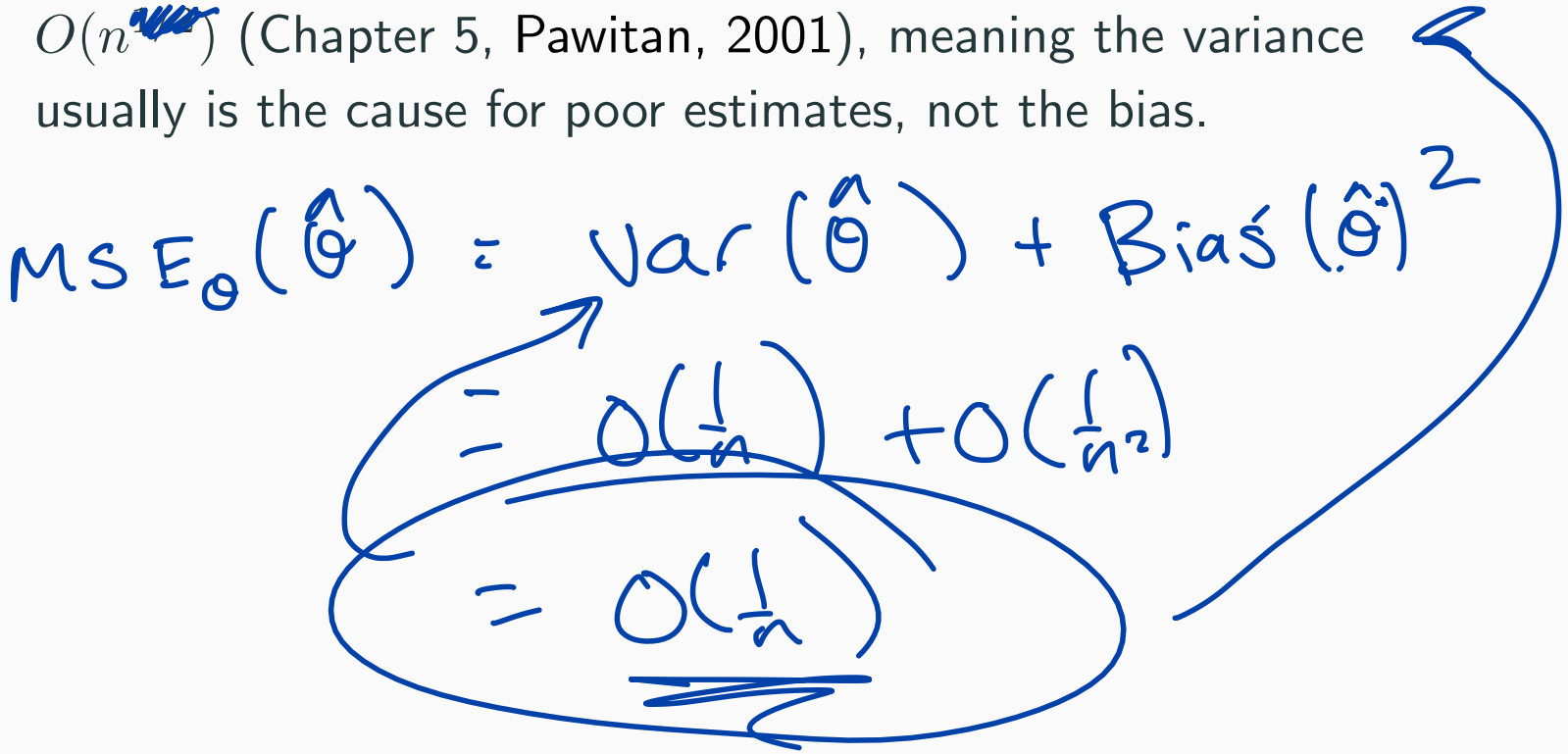
- Brief Summary:
- We have argued that under very mild conditions, the MLE and Bayesian posterior mean are both consistent (i.e., converges in probability to the true parameter value), **and** asymptotically efficient.
- This largely explains why these two approaches dominate the literature on parameter estimation:
- even if closed-form expressions for the sampling distributions of parameter estimates are unavailable, we are effectively guaranteed to have well-behaved estimates if we estimate using the MLE or Bayesian Posterior, and we have sufficient data.

Efficiency V

- Much of the statistical theory of the 19/20th centuries was focused on finding UMVUE estimators. This includes our discussion on consistency and asymptotic efficiency.
- Unbiasedness is nice feature, but strictly requiring unbiasedness sometimes results in a few issues (Chapter 5, Pawitan, 2001).
  - Finding an unbiased estimate can be extremely difficult.
  - Not all parameter *have* an unbiased estimator.
  - Requiring unbiasedness can produce terrible estimates.
  - Regardless of the prior and likelihood, all estimates using the posterior means are biased (Section 7.5.2, Casella and Berger, 2024).

Efficiency VI

- The bias of good estimators (like the MLE) typically shrinks at a rate $O(n)$, where the variance typically shrinks at a rate $O(n^{-2})$ (Chapter 5, Pawitan, 2001), meaning the variance usually is the cause for poor estimates, not the bias.



Sufficiency

Sufficiency

- One desirable trait of an estimator θ^* is that it it captures all of the information available in the data X about the parameter θ.
- Example: estimating $E[X_i] = \mu$, using: $\theta_1^* = X_1$, vs $\theta_2^* = \bar{X}_n$.
- We argued that both θ_1^* and θ_2^* are unbiased estimators, and that $\mathrm{Var}(\theta_2^*) < \mathrm{Var}(\theta_1^*)$. An additional way of viewing this, however, is that θ_1^* is *not capturing all of the information available for* θ
- This idea is known as the principle of sufficiency. This will be formally defined on the next slide.

Sufficiency II

Definition: Sufficiency

A statistic $T(X) = T(X_1, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$, given $T = t$, does not depend on θ for any value of t.

- If we take some statistic sufficient T (could be an estimator), then after accounting for T, there is no more information available about θ in the available data.
- Formally, a sufficient statistic suggests that we could throw away all of the data, except the sufficient statistic, and not lose any information about the parameter θ.
- If we reduce the data + model to a point estimate, we want this to not result in information loss.

Idea:

$$X_1, X_2, \ldots, X_n \sim F_\theta$$

$$T(X_1, X_2, \ldots, X_n) = T$$

$$(X_1, X_2, \ldots, X_n) \mid T \sim G_\eta$$

$X_1, X_2, \ldots, X_n,$

$\sum X_i = 625.$

Sufficiency III

Binomial sufficient statistic

Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameter θ. Show that $T(X) = \sum_i X_i$ is a sufficient statistic for θ.

$T(x) = \sum X_i = t$

$(X_1, X_2, \ldots, X_n) \mid T \sim \theta_\psi$

$f_X(x \mid T) = \frac{f_1(x)}{f(T)}$

$T(X) \sim \text{Bin}(n, \theta)$

$$f_{X|T}(x_1, x_2, \ldots, x_n | T = t)$$

$$= \frac{\prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{\cancel{\theta^t}(1-\theta)^{n-t}}{\binom{n}{t}\cancel{\theta^t}\cancel{(1-\theta)^{n-t}}}$$

$$= \frac{1}{\binom{n}{t}}.$$ ← nothing to do w/ θ!

$T(X) = \sum X_i$, is sufficient for θ.

□

Sufficiency IV

Normal sufficient statistic

Let $X = X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where σ^2 is known. Show that the mean, $T(X) = \bar{X}_n$ is a sufficient statistic for μ.

$$T(X) = \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

$$f_T(t;\theta) = \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{\frac{-n}{2\sigma^2}(t-\mu)^2}$$

$$f_X(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-1}{2\sigma^2}(x_i-\mu)^2\right)$$

$$= (2\pi\sigma^2)^{-n/2} e^{\frac{-1}{2\sigma^2}\sum (x_i-\mu)^2}$$

$$f_{X|T}(x|t) = \frac{f_X(x)}{f_T(t)}$$

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

$$\star: \exp\left(\frac{-1}{2\sigma^2}\sum (x_i - t) + (t-\mu)^2\right)$$

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

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

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

$T(x) = \bar{x}$ is sufficient for μ

pdf of $N\left(t, \sigma^2\left(1 - \frac{1}{n}\right)\right)$

Sufficiency V

- Using the definition to find a sufficient statistic is inconvenient. It requires knowing a candidate $T(X)$ ahead of time, knowing the distribution of the statistic, and computing conditional pmf / pdf.
- The following theorem provides us a way of finding sufficient statistics.

4/14 Review

Sufficiency:

→ idea: we don't want to lose information.

$X_1, X_2, \ldots, X_n \sim f(x;\theta)$

$\hat{\theta} = g(X_1, X_2, \ldots, X_n) = \bar{X}_n$

$T(X)$ ← Statistic

$\bar{X}$, $\text{med}(X)$, S^2,

$X \mid T(X) = G_\psi$, does not depend on θ.

The Factorization Theorem

Let $f_X(x;\theta)$ be the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t;\theta)$ and $h(x)$ such that, for all sample points x and parameter values θ,

$$f_X(x;\theta) = g\big(T(X);\theta\big)h(x).$$

proof: Assume $T(X)$ is sufficient,

Discrete case:

$$f_X(x|\theta) = P_\theta(X = x)$$

$$= P_\theta(X = x \text{ and } T(X) = T(x))$$

$X / T(X)$ is ind. of θ.

$$= P_\theta\left(X = x \text{ and } T(X) = T(x)\right)$$

$$= P_\theta\left(X = x \mid T(X) = T(x)\right) P_\theta\left(T(X) = T(x)\right)$$

$$= P\left(X = x \mid T(X) = T(x)\right) P_\theta\left(T(X) = T(x)\right)$$

$$f(x;\theta) = h(x)\, g(T(x);\theta)$$

Now assume $\exists\ g(t;\theta),\ h(x)$

s.t.:

$$f(x;\theta) = g(T(x);\theta)\,h(x)$$

suppose $T(x) = \sum x$

$X = (-1, 1) \Rightarrow T(x) = 0$

$X = (0, 0) \Rightarrow T(x) = 0$

$$A_t = \{x : T(x) = t\}$$

$$A_0 = \{x : T(x) = 0\}$$

$X|T(X)$ was no θ.

$q \sim$ pmf of $T(X)$.

$$\frac{f(x;\theta)}{q(T(x);\theta)} = \frac{g(T(x);\theta)\,h(x)}{q(T(x);\theta)}$$

$$= \frac{g(T(x);\theta)\,h(x)}{\sum_{y \in A_{T(x)}} f_X(y;\theta)}$$

$$= \frac{g(T(x);\theta)\,h(x)}{\sum_{y \in A_{T(x)}} g(T(y);\theta)\,h(y)}$$

$$A_t = \{x : T(x) = t\} \Rightarrow A_{T(x)} = \{y : T(y) = T(x)\}$$

$$= \frac{g(T(x);\theta)\,h(x)}{\sum_{y \in A_{T(x)}} g(T(x);\theta)\,h(x)}$$

$$= \frac{g(T(x);\theta)h(x)}{g(T(x);\theta)\sum_y h(y)}$$

$$= \frac{h(x)}{\sum_y h(y)} \Bigg] \rightarrow \text{does not include } \theta!$$

$X|T(X)$ does not include θ,

so $T(X)$ is sufficient. $\square$

Example: Bernoulli

$(X_1, X_2, \ldots, X_n) \overset{iid}{\sim} Ber(\theta)$

$$f_X(x;\theta) = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} \cdot 1$$

$$h(x) = 1$$

$$g(t;\theta) = \theta^t (1-\theta)^{n-t}$$

$$f_X(x;\theta) = g\left(\sum x_i ; \theta\right) \cdot h(x)$$

$\rightarrow$ $\sum x_i$ is sufficient

$$T_1(X) = \frac{1}{n}\sum(X_i)$$

$$g_1(t;\theta) = \theta^{nt}(1-\theta)^{n-nt}$$

$$f_X(x;\theta) = g_1\left(\frac{1}{n}\sum X_i;\ \theta\right) h(x)$$

$T(X)$ is sufficient,

$cT(X)$ is sufficient

$\frac{1}{n}\sum X_i$

Sufficient

Sufficiency VII

- The factorization theorem gives us a way to find sufficient statistics:
- Factor the joint pdf / pmf into two parts: one depending on θ, and one part that does not depend on θ. The part without θ becomes the function $h(x)$. In the remaining part of the function, find the remaining function of the data $T(X)$.

Example: Normal (known variance)

Let $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, with σ^2 known. Find a sufficient statistic for μ.

$$f_X(x;\theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right\}$$

$$+\bar{x} \quad -\bar{x}$$

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

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

$$= \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma^2}\sum(x_i-\bar{x})^2\right\}}_{h(x)} \underbrace{\exp\left\{\frac{-n}{2\sigma^2}(\bar{x}-\mu)^2\right\}}_{g(t;\theta),}$$

$$T(x) = \bar{x}$$

$$g(t;\theta) = \exp\left\{\frac{-n}{2\sigma^2}(t-\mu)^2\right\}$$

Sufficiency VIII

- The Factorization Theorem gives us a way to find sufficient statistics, but these statistics are not unique.
- For instance, if $T(X)$ is a sufficient statistic, than $cT(X)$ is also sufficient (for any constant c).
- Furthermore, the factorization theorem states that the entire vector of data X is itself a sufficient statistic:

$$f_X(x;\theta) = f_X(x;\theta)h(x),$$

  where $h(x) = 1$. This satisfies all conditions of the Factorization Theorem, so $T(X) = X$ is sufficient for θ.
- This leads us to the idea of minimal sufficiency.

$(X_1, X_2, \ldots, X_n)$ $\theta \in \mathbb{R}^d$

$T(X) = X$ is sufficient, $\mathbb{R}^n$

$T^*(X_1, X_2, \ldots, X_n) \in \mathbb{R}^d$

Sufficiency IX

Definition: Minimal Sufficient Statistic

A sufficient statistic $T(X)$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(X)$, $T(x)$ is is a function of $T'(x)$.

- One way of thinking about this is that a minimal sufficient statistic is the *coarsest* possible partition for a sufficient statistic, and thus achieves the greatest possible data reduction for a sufficient statistic.
- In the normal model, for instance, both $T(X) = \bar{X}$ and $T'(X) = X$ were shown to be sufficient. $T'(X) = X$ cannot be minimally sufficient, because it cannot be recovered from the mean alone (i.e., X is not a function of $\bar{X}$).

$\bar{X}$ is sufficient for $\mathcal{U}$.

$\Updownarrow$

$(X_1, X_2, \ldots, X_n)$ is suffient forall.

$$\bar{X} = \frac{1}{n}\sum X_i$$

$\theta \in \mathbb{R}^d$

$\bar{X} \longleftrightarrow c\bar{X}$

$T(X) \in \mathbb{R}^d$

Sufficiency X

- Finding a minimally sufficient statistic can be difficult, but there are theorems that can be used to help. See, for instance, Theorem 6.2.13 from Casella and Berger (2024).
- We'll finish discussion about sufficieny with one last example.

Joint sufficiency: IID normal model

Let X_i be iid $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Find a sufficient statistic for $\theta = (\mu, \sigma^2)$.

$$f(x;\theta) = (\sigma^2 2\pi)^{-n/2} \exp\left\{\frac{-1}{2\sigma^2}\sum (x_i-\mu)^2\right\}$$

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

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

$$= \underbrace{(2\pi)^{-n/2}}_{h(x)} \underbrace{\sigma^{-n} \exp\left\{ \frac{-1}{2\sigma^2} \left[T_1 - 2\mu T_2 + n\mu^2 \right] \right\}}_{g(t;\theta)}$$

$$T = (T_1, T_2) = \left(\sum x_i^2, \sum x_i \right)$$

$\left(\sum x_i^2, \sum x_i \right)$ is Sufficient for $\theta = (\sigma^2, \mu)$

Rao-Blackwell Theorem

- In addition to having desireable proprties, sufficient statistics can be used to improve estimators. This is known as Rao-Blackwellization.

The Rao-Blackwell Theorem

Let $\hat{\theta}$ be an estimator of θ, with $E[\hat{\theta}] < \infty$. Suppose that $T(X)$ is a sufficient statistic for θ, and define $\tilde{\theta} = E\big[\hat{\theta}|T(X)\big]$. Then, for all θ,

$$E\big[(\tilde{\theta} - \theta)^2\big] \leq E\big[(\hat{\theta} - \theta)^2\big].$$

The inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

$X_i \overset{iid}{\sim} \text{Pois}(\lambda)$. $T(X) = \sum(X_i)$

$\hat{\lambda}_{MLE} = \frac{1}{n}\sum(X_i)$. $E\left[\hat{\lambda}_{MLE} \middle| T(X)\right]$

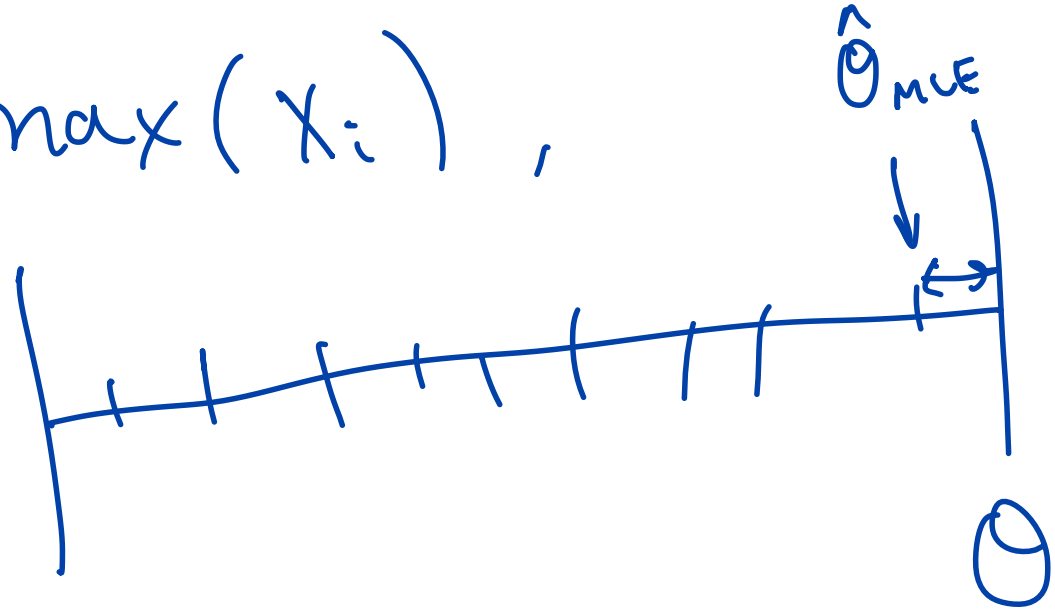
$$E\left[\frac{1}{n}\sum(x_i) \mid \sum(x_i)\right]$$

$$= \frac{1}{n} E\left[\sum(x_i) \mid \sum(x_i)\right]$$

$$= \frac{1}{n} \sum(x_i) = \hat{\lambda}_{MLE}$$

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} U(0, \theta)$

$\hat{\theta}_{MLE} = \max(X_i)$,



$E[X_i] = \frac{\theta}{2}$

$2\bar{X} = \hat{\theta}$

$E[2\bar{X}] = 2E[\bar{X}] = \frac{2}{n} \cdot n\frac{\theta}{2}$

$= \theta.$

R.B.

→ Find sufficient statistic.

$X_i \overset{iid}{\sim} U(0, \theta)$,

$$f(x;\theta) = \prod_{i=1}^{n} f_{x_i}(x_i;\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\theta} \cdot \mathbb{1}(0 \le x_i \le \theta)$$

$$= \frac{1}{\theta^n} \cdot \mathbb{1}(\max x_i < \theta)$$

$$= \underbrace{\frac{1}{\theta^n} \mathbb{1}(\max x_i \le \theta)}_{g(T(x);\theta)} \cdot \underbrace{1}_{h(x)}$$

$$T(x) = \max(x_i)$$

RB:

$$\hat{\theta} = 2\bar{x}.$$

$$\tilde{\theta} = E[\hat{\theta} \mid T(x)]$$

$$= E[2\bar{x} \mid \max(X_i)]$$

$$= \frac{2}{n} E\left[\sum X_i \mid X_{(n)}\right]$$

$$= \frac{2}{n} \sum E\left[X_i \mid X_{(n)}\right]$$

$$= \frac{2}{n} \cdot n \, E\left[X_1 \mid X_{(n)}\right]$$

$$= \underline{2 E\left[X_1 \mid X_{(n)}\right]}$$

2 Scenarios:

(1) X_1 is max

(2) X_1 is not max.

$$I = I(X_1 = X_{(n)})$$

$$P(I=1 \mid X_{(n)}) = \frac{1}{n}$$

$$P(I=0 \mid X_{(n)}) = \frac{n-1}{n},$$

$$E[X_1 \mid X_{(n)}] = E_I\left[E[X_1 \mid X_{(n)}, I]\right]$$

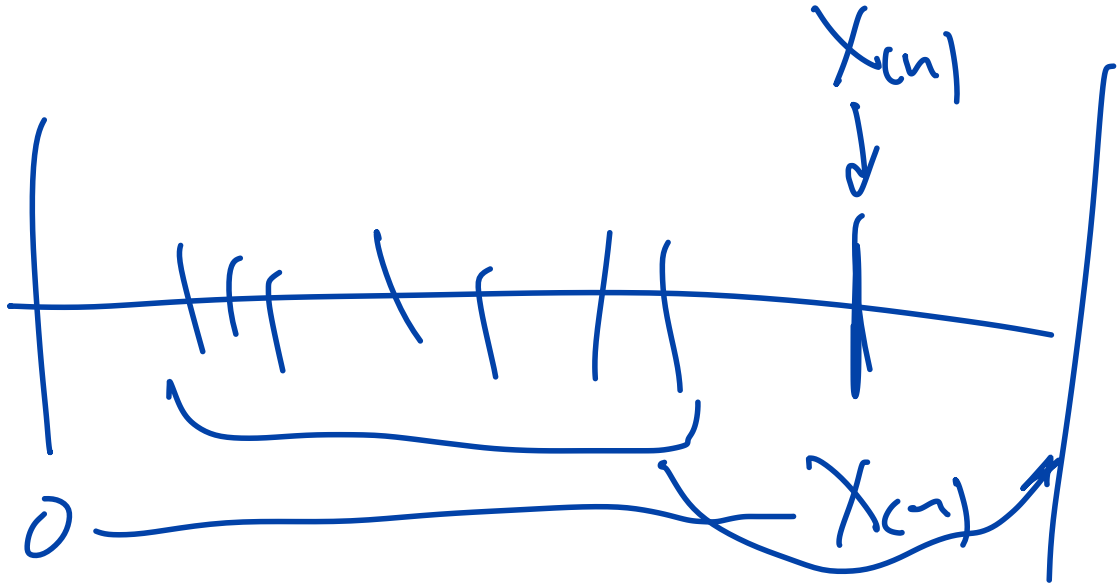
$$= E\left[X_1 \mid X_{(n)}, I=1\right] P(I=1 \mid X_{(n)})$$

$$+ E\left[X_1 \mid X_{(n)}, I=0\right] P(I=0 \mid X_{(n)})$$

$$= X_{(n)} \cdot \frac{1}{n} + E\left[X_1 \mid X_{(n)}, I=0\right] \frac{n-1}{n}$$

$I=0$, X_1 is not max.

$$X_1 \mid X_{(n)}, I=0 \sim U(0, X_{(n)})$$



$$X_i \mid X_{(n)} \text{ and } X_i \neq X_{(n)} \sim U(0, X_{(n)})$$

$$= X_{(n)} \cdot \frac{1}{n} + E\left[X_1 \mid X_{(n)}, I=0\right] \frac{n-1}{n}$$

$$= \frac{X_{(n)}}{n} + \frac{X_{(n)}}{2} \cdot \frac{n-1}{n}$$

$$E\left[X_1 \mid X_{(n)}\right] = (n+1) X_{(n)} / (2n)$$

$$\tilde{\theta} = E[\hat{\theta} \mid T(x)]$$

$$= 2\, E[X_1 \mid X_{(n)}]$$

$$= 2\, \frac{(n+1)\, X_{(n)}}{2n}$$

UMVUE

$$= \frac{(n+1)}{n} \cdot X_{(n)}$$

$$E[\tilde{\theta}] = \theta$$

Towards Lehmann-Scheffé

completeness.

$X \longmapsto X$ is sufficient,

"ancillary" information.

$X_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

$T(X) = \sum X_i$,

sufficient for μ.

$T_2(X) = \sum X_i^2$

T_2 has no information about μ.

$$T_3(x) = \left(\sum x_i , \sum x_i^2 \right)$$

Definition: Completeness.

We say $T(X)$ is complete for θ, if any function g that satifies

$E_\theta[g(T)] = 0$ Then

$P_\theta(g(T) = 0) = 1.$

(Example)

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bern}(\theta)$

$T = \sum X_i$ is sufficient for θ.

T is also complete.

$$E_\theta[g(T)] = 0$$

$$= \sum_{t=0}^{n} g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = 0$$

$(1-\theta)^n \cdot (1-\theta)^{-t}$

$$= (1-\theta)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t = 0$$

$$\sum g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t = 0$$

$$\frac{\theta}{1-\theta} > 0$$

$$\binom{n}{t} > 0$$

$$\sum g(t) \cdot A(t) = 0$$

$$g(t) = 0.$$

$$P_\theta(g(t)=0) = 1.$$

T is complete.

Lehmann-Scheffé

Theorem:

$X_1, X_2, \ldots, X_n$ are from $f(x;\theta)$.

If $T(X)$ complete, and sufficient, and $E[T(X)] = \theta$, then

$T(X)$ is UMVUE.

$\rightarrow E[\hat{\theta} \mid T(X)]$ = UMVUE. if $\hat{\theta}$ unbiased.

$X_1, \ldots, X_n \overset{iid}{\sim} Ber(\theta)$.

$E[X_1] = \theta$. X_1 unbiased

$T(x) = \sum X_i$ is complete sufficient

$E[X_1 \mid T(x)]$ ← UMVUE.

proof.

Suppose θ^* is some unbiased estimator. Apply RB:

$$\tilde{\theta}_1 = E[\theta^* | T(X)], \text{ unbiased,}$$

$$Var(\tilde{\theta}_1) \leq Var(\theta^*).$$

- $\hat{\theta}$ be an alternative unbiased estimator. Apply R.B.

$$\tilde{\theta}_2 = E[\hat{\theta} | T(X)].$$

Because both are unbiased,

$$E[\tilde{\theta}_1 - \tilde{\theta}_2] = 0$$

$$\tilde{\theta}_1 = E[\theta^* | T] = g_1(T)$$

$$\tilde{\theta}_2 = E[\hat{\theta} | T] = g_2(T).$$

$$g(T) = g_1(t) - g_2(t).$$

$$\therefore E[\tilde{\theta}_1 - \tilde{\theta}_2] = 0$$

$$= E\left[g_1(T) - g_2(T) \right] = 0$$

$$= E\left[g(T) \right] = 0 .$$

T is complete, so

$$P_\theta\left(g(T) = 0 \right) = 1 .$$

$$P_\theta\left(\tilde{\theta}_1 = \tilde{\theta}_2 \right) = 1$$

Rao-Blackwell Theorem II

- The Rao-Blackwell theorem implies that any estimator that is not a function of a sufficient statistic can be improved by conditioning on a sufficient statistic.
- This provides strong rationale for basing estimators on sufficient statistics (if they exist), and provides strong support for the MLE as an estimator.
- This theorem can be expanded to find UMVUE estimators, using the Lehmann–Scheffé Theorem: if $T(X)$ is a *complete* (definition not given), and $\hat{\theta}$ is unbiased, then $\tilde{\theta} = E[\hat{\theta}|T(X)]$ is UMVUE.

Rao-Blackwell Theorem III

Lemma: Sufficiency of the MLE

If $T(X)$ is sufficient for θ, then the MLE is a function of $T(X)$. Thus, the MLE is a minimal sufficient statistic.

Let $f(x;\theta)$ be model.

$T(X)$ is any sufficient statistic.

$$L(\theta) = f(x;\theta) = g(T(x),\theta)\, h(x)$$

$$\propto g(T(x),\theta).$$

maximize $g(T(x),\theta)$

Rao-Blackwell Theorem IV

Lemma: The likelihood function is sufficient

If $T(X)$ is sufficient for θ, then the likelihood function $L(\theta)$ is a function of T. Thus, the likelihood function is minimal sufficient.

→ Theorem 3.2 of Pawitan (2001).

$$T(X) = \frac{L(\theta)}{L(\theta_0)}$$

$$g(t,\theta) = \frac{L(t)}{L(\theta_0)}$$

$h(x)$

$$L(\theta) = f(x;\theta) = g(t;\theta) \cdot L(\theta_0)$$

Rao-Blackwell Theorem V

- The last few results explain why likelihood-based inference is so important: if your estimator does not involve the likelihood function, then the estimator can be improved.
- Don't forget! The likelihood depends on the chosen model.

→ Ways of point estimation

→ MoM

★. MLE

★. Bayesian posterior

→ Principles of estimation

- unbiasedness
- Invariant.
- minimize MSE
- sufficiency / completeness

→ RB / L-S.

References and Acknowledgements

Casella G, Berger R (2024). *Statistical inference*. Chapman and Hall/CRC.

Keener RW (2010). *Theoretical statistics: Topics for a core course*. Springer Science & Business Media.

Pawitan Y (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Rice JA (2007). *Mathematical statistics and data analysis*, volume 371. 3 edition. Thomson/Brooks/Cole Belmont, CA.

References and Acknowledgements II

- Compiled on March 24, 2026 using R version 4.5.3.
- Licensed under the Creative Commons Attribution-NonCommercial license. Please share and remix non-commercially, mentioning its origin.
- We acknowledge students and instructors for previous versions of this course / slides.