Outline

1. Unbiased Estimation and Risk Inequalities
   - Unbiased Estimation
   - The Information Inequality
Comments on Unbiased Estimation

- Estimation decision problem:
  - \( X \sim P_{\theta}, \theta \in \Theta \)
  - \( \theta(P) = E[X \mid P_{\theta}] \)
  - Estimation: \( A = \times \)
  - Loss function: \( L : \Theta \times A \rightarrow R \).
  - Decision procedures: \( D = \{ \delta : X \rightarrow A \} \)

- Restrict estimation procedures to the subclass:
  \[
  D_0 = \{ \delta \in D : E[\delta(X) \mid \theta] = \theta, \text{ for all } \theta \in \Theta \}.
  \]

- Apply decision-theoretic principles to identify optimal procedures in \( D_0 \).

Choice of \( D_0 \) equivalent to choice of constraints:

- Unbiasedness
- Linearity (in \( X \))
- Computational algorithms (e.g., orthogonal polynomials in \( X \), Fourier series, generalized-basis series)
Comments on Unbiased estimation (continued)

- Significant role of *unbiasedness* in survey sampling.
- Bayes estimates are necessarily biased (Problem 3.4.20).
- Unbiasedness not preserved under non-linear re-parametrization (not equivariant).
- Asymptotic unbiasedness:

\[
\frac{\text{Bias}^2(\hat{\theta}_n)}{\text{Var}[\hat{\theta}_n \mid \theta]} \rightarrow 0.
\]
Outline

1. Unbiased Estimation and Risk Inequalities
   - Unbiased Estimation
   - The Information Inequality
Definition: Regular Problem A statistical inference problem with \( X \sim P_\theta, \theta \in \Theta \) which satisfies the following regularity conditions:

- \( \mathcal{X} = \{ x : p(x \mid \theta) > 0 \} \) does not depend on \( \theta \).
- \( \frac{\partial \log p(x \mid \theta)}{\partial \theta} \) exists and is finite for all \( x \in \mathcal{X} \) and \( \theta \in \Theta \).
- For any statistic \( T \) such that \( E[|T(X)| \mid \theta] < \infty \)
  \[ \int \frac{\partial}{\partial \theta} \left[ T(x) p(x \mid \theta) dx \right] = \int T(x) \frac{\partial}{\partial \theta} [p(x \mid \theta)] dx. \]

Definition: Efficient Score Function. For a fixed \( \theta_0 \in \Theta \), the efficient score for \( X \) is

\[
u(X; \theta_0) = \left. \frac{\partial \log p(x \mid \theta)}{\partial \theta} \right|_{\theta = \theta_0}
\]

Note: The magnitude of \( u(X; \theta_0) \) scales how far \( \theta_0 \) is from \( \hat{\theta}_{MLE} \).
Proposition The Efficient Score Function has the following properties:

\[
E[u(X; \theta_0) \mid \theta = \theta_0] = 0.
\]

\[
\text{Var}[u(X; \theta_0) \mid \theta = \theta_0] = E\left([u(X; \theta_0)]^2 \mid \theta = \theta_0\right) = I(\theta_0).
\]

\(I(\theta)\) is the *Fisher information* about \(\theta\) contained in \(X\) which satisfies the following identity

\[
I(\theta_0) = \text{Var}[(u(X; \theta_0) \mid \theta_0) = E\left[- \frac{\partial^2 \log p(X \mid \theta_0)}{\partial \theta^2} \mid \theta_0\right]
\]

Proof:

\[
\int p(x \mid \theta)dx = 1 \\
\implies \int \frac{\partial p(x \mid \theta)}{\partial \theta}dx = \frac{\partial}{\partial \theta}(1) = 0
\]

\[
\implies \int \left[\frac{\partial p(x \mid \theta)}{\partial \theta}/p(x \mid \theta)\right]p(x \mid \theta)dx = 0
\]

\[
\implies \int \left[\frac{\partial \log[p(x \mid \theta)]}{\partial \theta}\right]p(x \mid \theta)dx = 0
\]

\[
\implies E[u(X; \theta) \mid \theta] = 0
\]
Unbiased Estimation and Risk Inequalities

The Information Inequality

\[ E[u(X; \theta) | \theta] = 0 \]
\[ \iff \int \left( \frac{\partial \log[p(x | \theta)]}{\partial \theta} \right) p(x | \theta) dx = 0 \]
\[ \frac{\partial}{\partial \theta} \left( \int \left( \frac{\partial \log[p(x | \theta)]}{\partial \theta} \right) p(x | \theta) dx \right) = \frac{\partial}{\partial \theta} (0) \]
\[ \int \left( \frac{\partial^2 \log[p(x | \theta)]}{\partial \theta^2} \right) p(x | \theta) dx + \frac{\partial \log[p(x | \theta)]}{\partial \theta} \left( \frac{\partial p(x | \theta)}{\partial \theta} \right) dx = 0 \]

The last line can be written as:
\[ \int \left( \frac{\partial^2 \log[p(x | \theta)]}{\partial \theta^2} \right) p(x | \theta) dx + \int \left( \frac{\partial \log[p(x | \theta)]}{\partial \theta} \right)^2 p(x | \theta) dx = 0 \]

I.e.,
\[ E \left[ \frac{\partial^2 \log[p(x | \theta)]}{\partial \theta^2} \right | \theta] + E \left[ \left( \frac{\partial \log[p(x | \theta)]}{\partial \theta} \right)^2 \right | \theta] = 0 \]

So we have
\[ I(\theta) = E[(u(X; \theta))^2 | \theta] = -E \left[ \frac{\partial^2 \log[p(x | \theta)]}{\partial \theta^2} \right | \theta] \]
\[ = \text{Var}[u(X; \theta) | \theta] \]
Proposition 3.4.1 Suppose $P_\theta$ is a one-parameter exponential family with density/pmf function:

$$p(x \mid \theta) = h(x) \exp\{\eta(\theta) T(x) - B(\theta)\}$$

which has non-vanishing continuous derivative on $\Theta$. Then the statistical inference problem for $\theta$ given $X$ is a regular problem.
Theorem 3.4.1. Information Inequality

For a regular problem, let $T(X)$ be any statistic such that

$$E[T(X) \mid \theta] = \psi(\theta).$$
$$Var[T(X) \mid \theta] < \infty,$$ for all $\theta$.

Then for all $\theta$:

- $Var[T(X) \mid \theta] \geq \frac{[\psi'(\theta)]^2}{I(\theta)}$,

($\psi(\theta)$ is differentiable and $I(\theta)$ = Fisher Information of $P_\theta$).

**Proof:** By the conditions of a regular problem:

$$
\psi'(\theta) = \frac{\partial}{\partial \theta} \left( \int T(x)p(x \mid \theta)dx \right)
= \int \left( T(x) \frac{\partial}{\partial \theta} [p(x \mid \theta)] \right) dx
= \int \left( T(x) \frac{\partial}{\partial \theta} [\log p(x \mid \theta)]p(x \mid \theta) \right) dx
= E[T(X)U(X; \theta) \mid \theta] = Cov[T(X), U(X; \theta) \mid \theta]
$$

(the last equation follows since $E[U(X; \theta) \mid \theta] = 0.$)
The theorem follows from the Cauchy-Schwarz Inequality for two random variables:

\[(\text{Cov}[T(X), U(X; \theta) | \theta])^2 \leq \text{Var}[T(X) | \theta] \times \text{Var}[U(X; \theta) | \theta]\]

i.e.,

\[\left[\psi'(\theta)\right]^2 \leq \text{Var}[T(X) | \theta] \times I(\theta)\]

**Corollary 3.4.1** Suppose \(T(X)\) is unbiased estimate of \(\theta\) in a regular problem, then

\[\text{Var}(T(X) | \theta) \geq \frac{1}{I(\theta)} \quad \text{(Cramer-Rao Lower Bound)}\]
Proposition 3.4.2 For a random sample \( \mathbf{X} = (X_1, \ldots, X_n) \) from a distribution \( P_\theta \) with density \( p(x | \theta) \) satisfying the conditions of a regular problem. If \( l_1(\theta) = E \left[ \left( \frac{\partial}{\partial \theta} \log p(x_1 | \theta) \right)^2 \right] \) then
\[
l(\theta) = nl_1(\theta) \quad \text{and} \quad \text{Var}[T(T) | \theta] \geq \frac{[\psi(\theta)]^2}{nl_1(\theta)}
\]

Proof: This follows directly from the results above upon noting that
\[
U(X; \theta) = \frac{\partial}{\partial \theta} \left[ \log p(X | \theta) \right] = \frac{\partial}{\partial \theta} \left[ \sum_{i=1}^{n} \log p(x_i | \theta) \right] = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \left[ \log p(x_i | \theta) \right] = \sum_{i=1}^{n} U(X_i; \theta)
\]
By the independence of the terms,
\[
\text{Var}[U(X; \theta) | \theta] = \sum_{i=1}^{n} \text{Var}[U(X_i; \theta)] = nl_1(\theta) = l(\theta).
\]
Theorem 3.4.2 Consider a regular problem with $X \sim P_\theta, \theta \in \Theta$, and $T^*(X)$ is an estimator of $\psi(\theta)$ which is

- Unbiased: $E[T^*(X) | \theta] = \psi(\theta)$, for all $\theta \in \Theta$.
- Achieves the Cramer-Rao Lower Bound:
  \[ \text{Var}(T^*(X) | \theta) = \frac{|\psi'(\theta)|^2}{I(\theta)}, \text{ for all } \theta \in \Theta. \]

Then $\{P_\theta\}$ is a one-parameter exponential family with density/pmf:
\[ p(x | \theta) = h(x) \exp\{\eta(\theta) T^*(x) - B(\theta)\} \]

**Proof:** From the proof of Theorem 3.4.1 for any unbiased estimator of $\psi(\theta)$,
\[
\psi(\theta) = E[T(x) | \theta] = \int T(x) p(x | \theta) dx
\]
\[
\implies \psi'(\theta) = \int T(x) U(x; \theta) p(x | \theta) dx
\]
where $U(x; \theta) = \partial \log p(x | \theta) / \partial \theta$
\[
= \text{Cov}(T(X), U(X; \theta) | \theta)
\]
\[
\implies |\psi'(\theta)| \leq \sqrt{\text{Var}(T(X) | \theta) \times \text{Var}(U(X; \theta) | \theta)}
\]
with equality if and only if $U(X; \theta) = a_1(\theta) + a_2(\theta) T(X)$ for some functions $a_1(\theta)$ and $a_2(\theta)$. 
Technical Details of Proof:

- For each $\theta_0 \in \Theta$, define
  \[ A_{\theta_0} = \{ x : U(x; \theta_0) = a_1(\theta_0) T^*(x) + a_2(\theta_0) \} \]
  Note: $P_{\theta_0}(A_{\theta_0}) = 1$
  (otherwise the absolute correlation would be less than 1)

- Define $\{ \theta_i, i = 1, 2, \ldots \}$ to be a denumerable dense subset of $\Theta$.

- Define $A^{**} = \cap_i A_{\theta_i}$. Then
  \[ P_{\theta_i}(A^{**}) = 1, \text{ for all } \theta_i. \]

- Fix any two values $x_1, x_2 \in A^{**}$, for which $T^*(x_1) \neq T^*(x_2)$.
  Solve the equations:
  \[
  U(x_1; \theta) = a_1(\theta) T^*(x_1) + a_2(\theta) \\
  U(x_2; \theta) = a_1(\theta) T^*(x_2) + a_2(\theta)
  \]
  to obtain equations for $a_1(\theta), a_2(\theta)$ as linear combinations of $U(x_1; \theta)$ and $U(x_2; \theta)$.
  Since $U(x; \theta)$ is continuous in $\theta$, so are $a_1(\theta)$ and $a_2(\theta)$. 
Technical Details of Proof (continued):

- **Since**
  \[ U(x; \theta) = a_1(\theta) T^*(x) + a_2(\theta), \]  
  for all \( \theta_i \in \{ \theta_i \} \)  
  and both \( U(x; \theta) \) and \( a_1(\theta) \) and \( a_2(\theta) \) are continuous,  
  this equation must hold for all \( \theta \).

- **So** \( A^{**} = \cap_i A_{\theta_i} \) must equal  
  \[ A^* = \{ x : U(x; \theta) = a_1(\theta) T^*(x) + a_2(\theta), \text{ for all } \theta \in \Theta \} \]  
  and \( P(A^*) = 1 \).

- **With**
  \[ U(x; \theta) = \frac{\partial \log p(x|\theta)}{\partial \theta} = a_1(\theta) T^*(x) + a_2(\theta) \]
  Define: \( \eta(\theta) = \int_{\theta_0}^{\theta} a_1(t)dt \) and \( B(\theta) = -\int_{\theta_0}^{\theta} a_2(t)dt \),  
  Then  
  \[ \log \left[ \frac{p(x|\theta)}{p(x|\theta_0)} \right] = \int_{\theta_0}^{\theta} \left[ \frac{\partial \log p(x|\theta)}{\partial \theta} \right] d\theta = T^*(x) \eta(\theta) - B(\theta), \]
  and we have:
  \[ p(x \mid \theta) = h(x) \exp \{ \eta(\theta) T^*(x) - B(\theta) \}, \ x \in A^* \]
  where \( h(x) = p(x \mid \theta_0) \) (for a fixed value \( \theta_0 \)).
Definition: Regular Problem A statistical inference problem with $X \sim P_{\theta}, \theta \in \Theta$ which satisfies the following regularity conditions:

- $\mathcal{X} = \{x : p(x | \theta) > 0\}$ does not depend on $\theta$.
- $\frac{\partial \log p(x | \theta)}{\partial \theta}$ exists and is finite for all $x \in \mathcal{X}$ and $\theta \in \Theta$.
- For any statistic $T$ such that $E[|T(X)| | \theta] < \infty$
  $$\frac{\partial}{\partial \theta} \left[ \int T(x)p(x | \theta)dx \right] = \int T(x)\frac{\partial}{\partial \theta}[p(x | \theta)]dx.$$

Definition: Efficient Score Function. For a fixed $\theta_0 \in \Theta$, the efficient score for $X$ is
$$u(X; \theta_0) = \frac{\partial \log p(x | \theta)}{\partial \theta} \bigg|_{\theta=\theta_0}$$

Note: The magnitude of $u(X; \theta_0)$ scales how far $\theta_0$ is from $\hat{\theta}_{MLE}$.

The definitions extend to vector-valued $\theta$ immediately
Proposition (I). The Efficient Score Function has the following properties:

\[ E[u(X; \theta_0) \mid \theta = \theta_0] = 0. \]
\[ \text{Cov}[u(X; \theta_0) \mid \theta = \theta_0] = E(\left[u(X; \theta_0)\right]\left[u(X; \theta_0)\right]^T \mid \theta = \theta_0) = I(\theta_0). \]

(II). \( I(\theta) \) is the \((d \times d)\) Fisher information matrix whose elements satisfy the following identities

\[ [I(\theta_0)]_{i,j} = [\text{Cov}[u(X; \theta_0) \mid \theta_0]]_{i,j} \]
\[ = E[[u(X; \theta)]_i[u(X; \theta)]_j \mid \theta = \theta_0] \]
\[ = E\left[\frac{\partial \log p(X \mid \theta)}{\partial \theta_i} \frac{\partial \log p(X \mid \theta)}{\partial \theta_j} \mid \theta = \theta_0\right] \]
\[ = -E\left[\frac{\partial^2 \log p(X \mid \theta)}{\partial \theta_i \partial \theta_j} \mid \theta = \theta_0\right] \]

(III). If \( X = (X_1, \ldots, X_n) \) is an iid sample from \( X \sim P_\theta \) with Information \( I_1(\theta) \), then

\[ I(X) = nI_1(\theta). \]
Theorem 3.4.3 For a regular problem with non-singular information matrix \( I(\theta) \), consider a scalar-valued statistic \( T(X) \) estimating the scalar \( \psi(\theta) \), and suppose
\[
E[T(X) \mid \theta] = \psi(\theta)
\]
\[
\dot{\psi}(\theta) = \nabla \psi(\theta) = \left[ \frac{\partial \psi(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \psi(\theta)}{\partial \theta_1} \right]^T
\]
Then
\[
\text{Var}[T(X) \mid \theta] \geq [\dot{\psi}(\theta)]^T[I(\theta)]^{-1}[\dot{\psi}(\theta)]
\]
Proof. For a random variable \( Y \), and a random \( d \)-vector \( Z \), recall the minimum MSPE linear predictor \( \mu_L(Z) \) of \( Y \) is given by:
\[
\mu_L(Z) = \mu_Y + (Z - \mu_Z)^T\Sigma^{-1}_{Z,Z}\Sigma_{Z,Y}
\]
where \( \mu_Y = E[Y] \), \( \mu_Z = E[Z] \),
\[
\Sigma_{Z,Z} = \text{Cov}(Z) \ (d \times d), \text{ and } \Sigma_{Z,Y} = \text{Cov}(Z, Y) \ (d \times 1).
\]
The variance of \( \mu_L(Z) \) satisfies
\[
\text{Var}(\mu_L(Z)) = [\Sigma_{Z,Y}]^T\Sigma^{-1}_{Z,Z}\Sigma_{Z,Y} \leq \text{Var}(Y),
\]
with equality only if \( Y = \mu_L(Z) \).
The Theorem follows setting \( Y = T(X) \) and \( Z = u(X; \theta) \).
Theorem 3.4.4 For a regular problem as in Theorem 3.4.3 suppose:

\[ T(X) = (T_1(X), \ldots, T_d(X))^T \in \mathbb{R}^d \]

\[ E[T(X) \mid \theta] = \psi(\theta) \quad (d \times 1) \text{ vector} \]

\[ \dot{\psi}(\theta) = \nabla \psi(\theta) = \begin{bmatrix} \frac{\partial \psi(\theta)}{\partial \theta_1} & \cdots & \frac{\partial \psi(\theta)}{\partial \theta_d} \end{bmatrix} \quad (d \times d) \text{ matrix} \]

Then

\[ \text{Var}[T(X) \mid \theta] \geq [\dot{\psi}(\theta)][I(\theta)]^{-1}[\dot{\psi}(\theta)]^T \]

where \( A \geq B \) means \( (A - B) \) is positive semi-definite:

\[ a^T(A - B)a \geq 0, \text{ for all } a \in \mathbb{R}^d. \]

Proof. Problem 3.4.21

Note: For \( \hat{\theta} : E[\hat{\theta} \mid \theta] = \theta, \)

\[ \psi(\theta) = \theta, \text{ and } \dot{\psi}(\theta) = I_d, \text{ the } (d \times d) \text{ identity matrix.} \]

and

\[ \text{Var}(\hat{\theta} \mid \theta) \geq [I(\theta)]^{-1} \]
Preview:

- When \( \mathbf{X} = (X_1, \ldots, X_n) \) corresponds to a random sample from a population whose distribution has information \( I_1(\theta) \) for a single observation, the information in a sample of size \( n \) is
  \[
  I(\mathbf{X}) = nI_1(\theta)
  \]
- As the sample size grows large such samples, optimal estimators of parameters \( q(\theta) \) are sought.
- The Cramer-Rao Lower Bound defines the golden standard of performance for estimators which are unbiased asymptotically.
- Such estimators will be called \textit{asymptotically efficient}. 