Parameter Estimation

- Sufficiency and Likelihood Principle (Ch. 6)
  - Sufficient statistics, Complete statistics
  - Ancillary statistics. (Ch 6.2)
  - Likelihood function and Factorization Theorem
- Exponential family of distributions (Ch. 3.4)
- Methods of Finding Estimators (Ch. 7.2)
- Methods of Evaluating Estimators (Ch. 7.3)
  - Cramer-Rao Lower Bound
  - Rao-Blackwell vs. Lehmann-Scheffe
- Interval Estimators (Ch. 9)
Sufficient statistics

- Factorization Theorem
- Exponential family of distributions
- Minimal sufficient statistics
- Complete sufficient statistics
- Ancillary statistics
- Basu’s Theorem
Sufficient statistics and Factorization Theorem

- $T(X)$ is a SS for $\theta$ if the conditional distribution of $X$ given $T(X)$ does not depend on $\theta$.

- $X = (X_1, X_2, \ldots, X_n)$, $x = (x_1, x_2, \ldots, x_n)$

**Factorization Theorem:** $T(X)$ is a SS for $\theta$ if and only if

$$f(x|\theta) = g(T(x), \theta) \cdot h(x)$$
Sufficient statistics and Exponential family

- Exponential family [1-parameter case]
  \[ f(x; \theta) = a(x) \ b(\theta) \ \exp(c(\theta) \ d(x)) \]
- How to identify \( a, b, c, d \) in exponential family? (e.g. 3.4.1)
- General notation used in the textbook:
  \[ f(x; \Theta) = h(x) \ c(\Theta) \ \exp(\sum w_i(\Theta) \ t_i(x)) \]
  \[ \Theta = (\theta_1, \theta_2, \ldots, \theta_k) \]
K-parameter exponential family

- \( f(x; \Theta) = h(x) \ c(\Theta) \exp(\sum w_i(\Theta) \ t_i(x)) \)
  - \( w_i(\Theta) \) are functions of \( \Theta \) alone, \( i=1,2,\ldots,k \).
  - \( t_i(x) \) does not depend on \( \Theta \), \( i=1,2,\ldots,k \).
  - \( k \) = number of parameters.
  - \( \Theta = (\theta_1, \theta_2, \ldots, \theta_k) \)
  - (see page 111, Ch 3.4)

- Connection with other topics?
Curved Exponential family

- Def 3.4.7 [page 115]: d=dim of the parameter space $\theta \in \Omega$ for k-parameter exponential family. If $d<k$, then it is called curved exponential family. Otherwise, it is full exponential family.

- E.g. $X \sim N(\theta, \theta^2)$

- Connection? Complete sufficient statistics. [Thm. 6.2.25, page 288]
Complete sufficient statistics

- \( T = T(X) \) is a CSS for \( \theta \), if
  \[ E_\theta(g(T)) = 0 \text{ for all } \theta \in \Omega \implies g(T) = 0. \]
- Why it is `complete’?
  - Parameter space \( \Omega \) must be large enough
- What are applications? (UMVUE)
- How to show CSS?
  - Exponential family [Thm 6.2.25, page 288]
  - Direct proof.
Minimal sufficient statistics

- The SS for $\theta$ is not unique. Many statistics are sufficient for $\theta$.
  - Whole data set is SS
  - Order statistics is also SS
- $T$ is minimal sufficient statistic (MSS) if it is SS and it is a function of other SS. [Def. 6.2.11, page 280]
- If MSS exists, then any CSS is also MSS. [Thm. 6.2.28, page 289]
Finding MSS

- [Thm 6.2.13, page 281] Let $f(x|\theta)$ be the pdf/pmf of random sample $\mathbf{X}$ and $f(x|\theta)/f(y|\theta)$ is a constant of $\theta$ if and only if $T(x)=T(y)$. Then $T(\mathbf{X})$ is MSS.
  - e.g. 6.2.14 (page 281),
  - e.g. 6.2.15 (page 282).
Ancillary statistics and Basu’s Theorem

- \( S(\mathbf{X}) \) is an ancillary statistic for \( \theta \), if its distribution does not depend on \( \theta \).

- Basu’s Theorem [Thm. 6.2.24, page 287]: If \( T(\mathbf{X}) \) is a CSS and MSS for \( \theta \), then \( T(\mathbf{X}) \) is independent of any ancillary statistic.

- Applications [e.g. 6.2.26, page 288]
Using Basu’s Theorem - I

- e.g. 6.2.26, page 288
- \( X_1, X_2, \ldots, X_n \) i.i.d. \( \sim f(x; \theta) = \frac{1}{\theta} e^{-x/\theta} \)
- \( g(X) = \frac{X_n}{X_1 + X_2 + \ldots + X_n} \) is an ancillary statistics.
- \( T(X) = \sum X_i \) is CSS and MSS (why?).
- Hence, according to Basu’s Theorem, \( g(X) \) and \( T(X) \) are independent. (So?)
  - distribution of \( g(X) \) and \( T(X) \) ?
Using Basu’s Theorem - II

- e.g. 6.2.27, page 289
- $X_1, X_2, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$, fixed $\sigma^2$
- $g(\mathbf{X}) = S^2$ is an ancillary statistics for $\theta$.
- $T(\mathbf{X}) = \sum X_i$ is CSS and MSS for $\theta$.
- Hence, according to Basu’s Theorem, $S^2$ and $\sum X_i$ are independent. (So?)
  - distribution of $S^2$ and $\sum X_i$?
CSS and exponential family

\[ f(x; \Theta) = h(x) c(\Theta) \exp(\sum w_j(\Theta) \sum t_j(x)) \]

- If parameter space \( \Omega \) contains an open set in \( k \)-dim space, then
  \[ T(X) = [\Sigma t_1(X_i), \Sigma t_2(X_i), ..., \Sigma t_k(X_i)] \]
  is a CSS.

- [Thm. 6.2.25, page 288]
Sufficiency Principle

- If $T(\mathbf{X})$ is SS for $\theta$, and if $T(\mathbf{x}) = T(\mathbf{y})$ for two samples $\mathbf{x}$ and $\mathbf{y}$, inference about $\theta$ from $\mathbf{x}$ and from $\mathbf{y}$ should be the same.
  - (Section 6.2, page 272)
- Thm. 6.2.2 (p. 274) if $f(\mathbf{x}; \theta)/g(T(\mathbf{x}); \theta)$ does not involve $\theta$, then $T(\mathbf{X})$ is SS.
  - e.g. 6.2.5: (page 275) Order statistics are always a sufficient statistics.
Likelihood Principle

- likelihood function vs. joint p.d.f.
  - With $X=x$ given, $L(\theta|x) = f(x|\theta)$.
- likelihood principle: (Def. 6.3.1, p. 290)
  - if $L(\theta|x) = C(x, y) L(\theta|y)$ for all $\theta$, then inference about $\theta$ from $x$ and from $y$ should be the same.
- MLE (Ch. 7.2.2) obeys this principle while Moment Estimator (Ch. 7.2.1) may not.
Likelihood vs. joint p.d.f.

- $X_1, X_2, \ldots, X_n$ i.i.d. $\sim$ p.d.f. $f(x; \theta)$, joint p.d.f. is
  
  \[ f(x_1, \ldots, x_n; \theta) = \prod f(x_i; \theta). \]

- After data $(x_1, \ldots, x_n)$ observed, $f(x_1, \ldots, x_n; \theta)$
  is a function of $\theta$ alone. It is called the likelihood function.

  \[ L(\theta) = f(x_1, \ldots, x_n; \theta) = \prod f(x_i; \theta). \]

- MLE is the $\theta$ such to maximize $L(\theta)$. How?
Methods of Finding Estimators

- Method of Moments Estimator
- MLE
  - Likelihood function vs. joint p.d.f.
  - Invariance principle.
- Bayes Estimator
  - Prior and posterior distributions
  - Various loss functions.
Method of Moments Estimator

- [Ch 7.2] Find the first moment of a random variable $Y \sim$ p.d.f. $f(y; \theta)$:
  
  $$E(Y) = g(\theta)$$

- Let $\bar{Y} = \frac{1}{n} \sum Y_i$ be the sample mean. Solve
  
  $$\bar{Y} = E(Y) = g(\theta)$$

- MME is then found as
  
  $$\theta = g^{-1}(\bar{Y}).$$
More about MME

- What if we have k-parameters of interest?
- Match the first k population moments with corresponding sample moments. Then solve simultaneous equations.
  - e.g. 7.2.1. page 313.
  - e.g. 7.2.2. page 313.
Satterthwaite approximation

- e.g. 7.2.3 (pages 314-315)
- $Y_i$ independent $\sim \chi^2(r_i)$, $a_i$ are constants, $i=1,2,\ldots,k$.
- $Y = (\sum a_i Y_i) \sim \chi^2(r)/r$, (approx.)
- Q: Use the key idea of MME:
  - Restriction on $a_i$ ? ($\sum a_i = 1$)
  - Find (positive) estimator for $r$ ?
Maximum Likelihood Estimator

- Likelihood function vs. joint p.d.f.
- How to find MLE?
- Invariance principle.
- Properties of MLE:
  - a function of SS
  - BAN: best asymptotically normally distributed.
Likelihood vs. joint p.d.f.

- \( X_1, X_2, \ldots, X_n \) i.i.d. \( \sim \) p.d.f. \( f(x;\theta) \), joint p.d.f. is 
  \[ f(x_1, \ldots, x_n; \theta) = \prod f(x_i; \theta). \]
- After data \( (x_1, \ldots, x_n) \) observed, \( f(x_1, \ldots, x_n; \theta) \) is a function of \( \theta \) alone. It is called the likelihood function.
  \[ L(\theta) = f(x_1, \ldots, x_n; \theta) = \prod f(x_i; \theta). \]
- MLE is the \( \theta \) such to maximize \( L(\theta) \). How?
How to find MLE?

- Write the likelihood function:
  \[ L(\theta) = f(x_1, \ldots, x_n; \theta) = \prod f(x_i; \theta). \]

- How to maximize \( L(\theta) \)?
  - If \( L(\theta) \) is differentiable, solve \( \frac{dL(\theta)}{d\theta} = 0 \).
    - e.g. \( X_1, X_2, \ldots, X_n \) i.i.d. \( \sim \) Poisson(\( \theta \)).
  - What if \( L(\theta) \) is NOT differentiable? E.g.,
    - e.g. \( X_1, X_2, \ldots, X_n \) i.i.d. \( \sim \) U(0, \( \theta \)).
More examples of finding MLE

- $X_1, X_2, \ldots, X_n$ i.i.d. $\sim B(k,p)$.
  - e.g. 7.2.9 (page 318) where $k$ is unknown parameter and $p$ is a known constant.

- $X_1, X_2, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$.
  - e.g. 7.2.11 (page 321-322) where $\mu$ and $\sigma^2$ are unknown parameters.
  - e.g. 7.2.8 (page 318) where $\mu$ is unknown parameter $\geq 0$ and $\sigma^2 = 1$. (restricted range MLE).
Bayes Estimator

- Difference between traditional approach and Bayesian approach.
- Prior and posterior distribution.
- Method to find Bayes Estimator.
- Loss functions.
- [Ch. 7.2.3, page 324]
Traditional vs. Bayesian

- Classical approach: $X_1, X_2, \ldots, X_n$ i.i.d. 
  $\sim p.d.f. \ f(x;\theta)$, where $\theta$ is a fixed unknown parameter.

- Bayesian Approach: $\theta$ is a random variable with some known prior distribution.

- $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p.d.f. \ f(x|\theta)$, where $\theta \sim p.d.f. \ \pi(\theta)$. 
Prior and posterior distribution

\( X_1, X_2, \ldots, X_n \) i.i.d. \( \sim p.d.f. \ f(x|\theta) \), where \( \theta \sim p.d.f. \ p(\theta) \).

- Likelihood function:
  \[
  L(\theta) = f(x_1, \ldots, x_n|\theta) = \prod f(x_i|\theta).
  \]

- Joint p.d.f. = product of \( L(\theta) \) and prior \( p(\theta) \).

- Posterior:
  \[
  f(\theta|x_1, \ldots, x_n) = \text{Const} \times L(\theta) \times p(\theta).
  \]
Conjugate family

- Def. 7.2.15 (page 325) $F = \text{family of pdf/pmf } f(x|\theta)$. $\Pi = \text{family of prior distributions for } \theta$. $\Pi$ is a conjugate family for $F$, if its posterior distribution is also in $\Pi$.
- Why consider conjugate prior?
- e.g. $X \sim \text{Poisson}(\theta)$, its conjugate prior is $\theta \sim \text{gamma}(a,b)$. [show]
Examples for conjugate family

- e.g. 7.2.14 (page 324-325) $X \sim B(n, \theta)$, its conjugate prior is $\theta \sim \text{beta}(a,b)$. [show]

- e.g. 7.2.16 (page 326) $X \sim N(\theta, \sigma^2)$, $\sigma^2$ is assumed known. The conjugate prior is $\theta \sim N(\mu,\tau^2)$. [show]
Step to find Bayes Estimator

- Identify the Posterior distribution:
  \[ f(\theta | x_1, \ldots, x_n) = C \times L(\theta) \times p(\theta). \]
- Bayes estimator for loss functions chosen: (page 349)
  - \( L(\theta, a) = (\theta - a)^2 \), \( \theta_B = E(\theta | X) \).
  - \( L(\theta, a) = w(\theta) (\theta - a)^2 \),
    \[ \theta_B = \frac{E(\theta \ w(\theta) | X)}{E(w(\theta) | X)}. \]
  - \( L(\theta, a) = |\theta - a| \), \( \theta_B = \text{posterior median} \).
Methods of Evaluating Estimators

- Unbiased estimator
- Mean Squared Error
- Cramer-Rao Lower Bound
- Rao-Blackwell Theorem
- Lemann-Scheffe Theorem
- UMVUE
Mean Squared Error (MSE)

- Def. 7.3.1 (page 330), W be an estimator of $\theta$
  \[ \text{MSE} = E_\theta(W - \theta)^2 \]
- MSE is common to compare two or more estimators.
  \[ \text{MSE} = E_\theta(W - \theta)^2 = \text{Var}_\theta(W) + \text{Bias}_\theta(W)^2 \]
Examples of using MSE

- e.g. 7.3.4 (page 331) \(X_1, X_2, \ldots, X_n\) i.i.d. \(\sim \mathcal{N}(\mu, \sigma^2)\), unknown \(\mu, \sigma^2\)
  - use MSE to compare \(S^2\) and MLE of \(\sigma^2\)
  - which one is better?

- e.g. 7.3.5 (page 332) \(X_1, X_2, \ldots, X_n\) i.i.d. \(\sim \text{Ber}(p)\) unknown \(p\).
  - use MSE to compare MLE of \(p\) and its Bayes estimator with beta\((a,b)\) prior (e.g. 7.2.14)
  - how to choose \(a\) and \(b\) (to minimize its MSE)?
  - which one is better?
Fisher’s Information

- Ways to compute Fisher’s Information
  1. $I(\theta) = E[\partial \log f(X| \theta)/\partial \theta]^2$
  2. $I(\theta) = \text{Var}[\partial \log f(X| \theta)/\partial \theta]$
  3. $I(\theta) = -E[\partial^2 \log f(X| \theta)/\partial \theta^2]$

Unbiased estimator and Cramer-Rao Lower Bound

- $I(\theta) = \text{Fisher’s Information}$
- For any unbiased estimator $S$ of $\theta$:
  $\text{Var}(S) \geq 1/[n \times I(\theta)]$.
- For any unbiased estimator $S$ of $\phi(\theta)$:
  $\text{Var}(S) \geq [\phi'(\theta)]^2 /[n \times I(\theta)]$.
  - Corollary 7.3.10 (page 337)
  - Theorem 7.3.9 (page 335) [general case]
Example of failure of CRLB

- CRLB requires a regularity condition on the pdf/pmf $f(x; \theta)$ and $E_\theta(W)$. See page 335: $\text{Var}_\theta(W) < \infty$ and

$$[E_\theta(W)]' = \int [W(x) f(x; \theta)]' \, dx$$

- If the regularity condition does not hold, then CRLB may not hold.

- e.g. 7.3.13 (page 339) $X_1, X_2, ..., X_n$ i.i.d. $\sim U(0, \theta)$. [show!!]
UMVUE

- Uniformly Minimum Variance Unbiased Estimator for \( \theta \). (Def. 7.3.7, page 334)
- UMVUE and Cramer-Rao Lower Bound
  - Cramer-Rao Inequality [pages 335-337]
  - If CRLB is achieved, then it is UMVUE.
    - Condition ? [Corollary 7.3.15, page 341]
    - e.g. 7.3.14-16 (pages 340-341) [show!]
  - Common ways to find UMVUE ?
Rao-Blackwell Theorem

- Let S be an unbiased estimator of $\varphi(\theta)$. Therefore, $h(T) = E(S|T)$ is unbiased.
- (Rao-Blackwell) [Thm. 7.3.17, page 342] If T is a SS for $\theta$.
  \[ \text{Var}(h(T)) \leq \text{Var}(S) \]
- That is, $h(T)$ is a uniformly better unbiased estimator of $\varphi(\theta)$ than S.
Rao-Blackwell Theorem

- Recall that
  \[ \text{Var}(X) = \text{Var}[E(X|Y)] + E[\text{Var}(X|Y)] \]
- For any \( S, h(T) = E(S|T) \), we have
  \[ \text{Var}(h(T)) \leq \text{Var}(S) \]
  \[ E(h(T)) = E(S) = \varphi(\theta) \]
- If \( T \) is SS, then \( h(T) \) is a statistics not involve (why not ?) on the unknown parameter \( \theta \).
  - Hence, \( h(T) \) is uniformly better unbiased estimator of \( \varphi(\theta) \) than \( S \).
  - How about comparison of \( h(T) \) with other \( S^* \) ?
Lemann-Scheffe Theorem

- Let $S$ be an unbiased estimator of $\phi(\theta)$. Therefore, $h(T) = E(S|T)$ is unbiased.
- (Lemann-Scheffe) [Thm. 7.3.23, page 347] If $T$ is a CSS for $\theta$.
  \[ \text{Var}(h(T)) \leq \text{Var}(S^*), \]
  for any $S^*$ unbiased estimator of $\phi(\theta)$.
- e.g. 7.3.24 (page 347) [show!]
Two common methods to find UMVUE

1. Let S be any SS unbiased for $\theta$ and T is a CSS, then $h(T) = E(S|T)$ is UMVUE.
   - e.g. 7.3.22 (page 346)

2. Let T be a CSS for $\theta$ and find $h(T)$ such that $E(h(T)) = \theta$, then $h(T)$ is UMVUE.
   - e.g. 7.3.24 (page 347) [show!]
Example 7.3.22 (page 346)

- $X_1, X_2, \ldots, X_n$ i.i.d. $\sim \text{U}(0, \theta)$.
- $Y = X_{(n)}$ is CSS for $\theta$
- It is easy to show $E(Y) = \frac{n}{n+1} \theta$
- Therefore, $(n+1)Y/n$ is UMVUE of $\theta$. 
Example 7.3.24 (page 347)

- X₁, X₂, ..., Xₙ i.i.d. ~B(k, θ).
- T = Σ Xᵢ is CSS for θ, parameter of interest is τ(θ) = P(X=1) = k θ(1− θ)ᵏ⁻¹
- First, we find an unbiased estimator
  S=1, if X₁ =1; otherwise S=0.
- Second, compute E(S|T) which is the UMVUE. (page 348).
Interval Estimators

Methods of Finding Interval Estimators
- Inverting a Test Statistics
- Pivotal Quantities
- Bayes Confidence Interval

Methods of Evaluating Interval Estimators
- Size and coverage probability
- Shortest interval
Methods of Finding Interval Estimators

- Inverting a Test Statistic
- Pivotal Quantities Method
- Bayes Confidence Interval
- Sampling from Normal Population:
  - Mean and variance for one sample.
  - Difference of two means, Ratio of variances for one samples.
Pivotal Quantities Method

- Start with a test statistic or pivotal quantity $T(X, \theta)$ whose distribution does not depend on $\theta$.
- Identify the distribution of $T(X, \theta)$ and find its lower $q_1$ and upper percentile $q_2$.
  \[ P[q_1 < T(X, \theta) < q_2] = 1- \alpha. \]
- Solve the inequality in terms of $\theta$.
  \[ P[C_1(X) < \theta < C_2(X)] = 1- \alpha. \]
Bayesian Intervals

- [Ch 9.2.4, page 435] \( \pi(\theta|x) \) posterior distribution of \( \theta \), \( A \) is a subset of \( \Omega \). The credible probability of \( A \) is \( P(\theta \in A|x) = \int_A \pi(\theta|x) \, d\theta \).

- [Cor. 9.3.10, p. 448] \( A=\{\theta: \pi(\theta|x) \geq k\} \) is \((1-\alpha)\%\) HPD (highest posterior density) region, where 

\[
P(\theta \in A|x) = \int_A \pi(\theta|x) \, d\theta = 1-\alpha.
\]
Sampling from Normal Population

- One sample:
  - Mean $\mu$, ($\sigma^2$ known/unknown)
  - Variance $\sigma^2$.

- Two samples:
  - Difference $\mu_1 - \mu_2$ ($\sigma^2_1 = \sigma^2_2$, $\sigma^2_1 \neq \sigma^2_2$)
  - Ratio of variances $\sigma^2_1/\sigma^2_2$. 