Week 1

- **Theorem 1.5.2:**

  The function $F_X(x)$ is a CDF if and only if the following three conditions holds

  1. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.
  2. $F_X(x)$ is a non-decreasing function of x.
  3. $F_X(x)$ is right-continuous.

- **Theorem 1.6.1**:

  The PMF of a discrete random variable X is

  $$f_X(x) = P(X = x) \quad x \in \mathbb{R}$$

- **Theorem 1.6.5**:

  A function $f(x)$ is a PDF (or PMF) iff

  1. $f(x) \geq 0$ for all x;
  2. $\int_{-\infty}^{\infty} f(x)dx = 1$ (or $\sum_x f(x) = 1$).

- **Theorem 2.1.3, 2.1.5**

  Let X be a continuous random variable and $Y = g(X)$ with range $\mathcal{Y}$.

  1. If g is increasing, then $F_Y(y) = F_X(g^{-1}(y)), y \in \mathcal{Y}$.
  2. If g is decreasing, then $F_Y(y) = 1 - F_X(g^{-1}(y)), y \in \mathcal{Y}$.
  3. If f_X is continuous and g is continuously differentiable, then

  $$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right| & , y \in \mathcal{Y} \\ 0 & , \text{otherwise} \end{cases}$$

- **Theorem 2.1.8**

  Let X be a continuous random variable with PDF f_X. Suppose that there are disjoint $A_1, \ldots, A_k$ such that $P(X \in U_{t=0}^{k} A_t) = 1$. f_X is continuous on each $A_t, t = 1, \ldots, k$, and there are functions $g_1(x), \ldots, g_k(x)$ defined on $A_1, \ldots, A_k$, respectively, satisfying

  1. $g(x) = g_t(x)$ for $x \in A_t$;
  2. $g_t(x)$ is strictly monotone on A_t;
  3. The set $\mathcal{Y} = \{y : y = g_t(x)$ for some $x \in A_t\}$ is the same for each t.
  4. $g_t^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$ for each t.

Then Y has the PDF

$$f_Y(y) = \begin{cases} \sum_{t=1}^{k} f_X(g_t^{-1}(y))|\frac{d}{dy}g_t^{-1}(y)| & , y \in \mathcal{Y} \\ 0 & , \text{otherwise} \end{cases}$$

- **Geometric Series:**

$$\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}.$$

For $|r| < 1$, the sum convergences as $n \to \infty$, i.e.,

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$$

- **Integration by parts:**

$$\int f(x)g(x)dx = \int udv = uv - \int vdu$$

- **Stirling's approximation:**

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Week 2

- **Chebychev's Inequality**:

  Let X be a R.V. and let $g(x)$ be a nonnegative function. For any $r > 0$.

$$P(g(X) \geq r) \leq \frac{E[g(X)]}{r}$$

  - If g is nondecreasing, then another form of Chebychev's inequality is, for $\epsilon > 0$.

$$P(X \geq \epsilon) \leq \frac{E[g(X)]}{g(\epsilon)}$$

  - Suppose that X has expectation μ and variance σ^2. For $g(x) = (x-\mu)^2/\sigma^2$, we have

$$P(|X-\mu| \geq t\sigma) = P\left(\frac{(X-\mu)^2}{\sigma^2} \geq t^2\right) \leq \frac{1}{t^2}E\left[\frac{(X-\mu)^2}{\sigma^2}\right] = \frac{1}{t^2}$$

  - If X has a finite kth moment with an integer k, then for $t > 0$.

$$P(|X-\mu| \geq t) \leq \frac{E[X-\mu]^k}{t^k}$$

- **Cauchy-Schwarts's Inequality**:
  - If X and Y are random variables with $E[X^2] < \infty$ and $E[Y^2] < \infty$, then the following Cauchy-Schwartz's inequality holds:

$$E(XY)^2 \leq E(X^2)E(Y^2)$$

with equality holds iff $P(X=cY)=1$ for a constant c.

- We also have

$$[E|XY|]^2 \leq E(X^2)E(Y^2)$$

- **Jensen's Inequality**:

  If g is a convex function on a convex $A \subset \mathcal{R}$ and X is a random variable with $P(X \in A)$, then

$$g(E(X)) \leq E[g(X)]$$

  provided that the expectations exist. If g is strictly convex, then $\leq$ in the previous inequality can be replaced by $<$ unless $P(g(X)=c)=1$ for a constant c.

- **Definition 2.3.6**: (Moment Generating Function)

  The moment generation function (mgf) of a random variable X is

$$M_X(t)=E(e^{tX})=\begin{cases}\sum_x e^{tx}f_X(x) & \text{, if } X \text{ has a PMF}\\ \int_{-\infty}^{\infty} e^{tx}f_X(x)dx & \text{, if } X \text{ has a PDF}\end{cases}$$

  provided that $E(e^{tX})$ exists. (Note that $M_X(0)=E(e^{tX})=1$ always exists.) Otherwise, we say that the MGF $M_X(t)$ does not exist at t.

- **Theorem**:

  If $M_X(t)$ exists at a neighborhood of $t=0$, then $E(X^n)$ exists for any positive integer n and

$$E(X^n)=M_X^{(n)}(0)=\frac{d^n}{dt^n}M_X(t)\bigg|_{t=0}$$

- **Theorem 2.3.15**:

  For any constants a and b, the MGF of the random variable $aX+b$ is

$$M_{aX+b}(t)=e^{bt}M_X(at)$$

- Useful distributions
  - Binomial
  - Poisson
  - Uniform
  - Normal
  - Gamma
  - Chi-square
  - Exponential

Lecture 4

- **Joint and conditional distributions**
- **Definition 4.1.10 Continuous joint PDF**
- **Definition 4.2.1 Conditional PMF**
- **Definition 4.2.3 Conditional PDF**
- **Conditional Expectations**
- **Properties of conditional expectation**

1. If $P(Y=c)=1$ for a constant c, then $E(Y|X)=c$.
2. If $Y \leq Z$, then $E(Y|X) \leq E(Z|X)$.
3. For constants a and b, $E(aY+bZ|X)=aE(Y|X)+bE(Z|X)$.
4. $E[E(Y|X)]=E(Y)$.
5. $Var(Y)=E[Var(Y|X)]+Var(E(Y|X))$, where $Var(Y|X)$ is the variance of the conditional distribution

- **Definition 4.5.1 Correlation and independence**

  The covariance of random variables X and Y is defined as

  $$\operatorname{Cov}(X,Y)=E\{[X-E(X)][Y-E(Y)]\}=E(XY)-E(X)E(Y)$$

  Provided that the expectation exists.

- **Definition 4.5.2**

  The correlation (coefficient) of random variables X and Y is defined as

  $$\rho_{X,Y}=\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

- **Theorem 4.5.6**

  If X and Y are random variables and a and b are constants, then

  $$\operatorname{Var}(aX+bY)=a^2\mathrm{X}+b^2\operatorname{Var}(Y)+2ab\operatorname{Cov}(X,Y)$$

- **Variance-covariance matrix**

  For an n-dimensional random vector $X=(X_1,\ldots,X_n)$, its mean is $E(X)$ and its variance-covariance matrix is

  $$\operatorname{Var}(X)=E\{[X-E(X)][X-E(X)]'\}=E(XX')-E(X)E(X')$$

  which is an $n\times n$ symmetric matrix whose i-th diagonal element is the variance $\operatorname{Var}(X_i)$ and (i,j)th off-diagonal element is the covariance $\operatorname{Cov}(X_i,X_j)$.

- **Theorem 4.5.7 Correlation measures linearity**
- **Independence of random variables**
- **Lemma 4.2.7 Check independence**
- **Bivariate Normal Distribution**
- **Theorem 4.6.12**
- **Definition (Conditional Independence)**

Lecture 5

- **PMF of sum of discrete random variables**

  PMF of

  $$f_{X+Y}(t)=\sum_{x+y\leq t}f(x,y)=\sum_{x}f(x,t-x)=\sum_{y}f(t-y,y)$$

and if X and Y are independent with marginal PMF's f_X and f_Y, then

$$f_{X+Y}(t)=\sum_x f_X(x)f_Y(t-x)=\sum_y f_X(t-y)f_Y(y)$$

- **PDF of sum of continuous random variables**

  f_{X+Y} is called a convolution.

$$f_{X+Y}(t)=\int_{-\infty}^{\infty} f(t-y,y)dy=\int_{-\infty}^{\infty} f(x,t-x)dx$$

  If X and Y are independent

$$f_{X+Y}(t)=\int_{-\infty}^{\infty} f_X(x)f_Y(t-x)dx=\int_{-\infty}^{\infty} f_X(t-y)f_Y(y)dy$$

  Example 4.3.1 Sum of Poisson R.V. is a Poisson R.V.[Lecture5-update]

- **Theorem 4.2.12**

  $T=X_1+\ldots+X_n$, X_i are independent.

$$M_T(t)=M_{X_1}(t)\ldots M_{X_n}(t)$$

- **Theorem 4.2.14**

  If $X\sim\mathcal{N}(\mu,\sigma^2)$ and $Y\sim\mathcal{N}(\gamma,\tau^2)$ are independent, then $T=X+Y\sim\mathcal{N}(\mu+\gamma,\sigma^2+\tau^2)$.

- **Additivity of the gamma distributions**

  If $X_i\sim\text{gamma}(\alpha_i,\beta), i=1,\ldots,k$, are independent, then the sum $T=X_1+\ldots+X_k\sim\text{gamma}(\alpha_1+\ldots+\alpha_k,\beta)$.

- **Additivity of the chi-square distributions**

  If $X_i\sim\mathcal{N}(\mu_i,\sigma_i^2), i=1,\ldots,k$, are independent, then the distribution of

$$Y=\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2+\ldots+\left(\frac{X_k-\mu_k}{\sigma_k}\right)^2$$

  is the chi-square distribution with degrees of freedom k.

- **Hierarchical models**

  Binomial-Poisson hierarchy [Lecture5-update]

- **Mixture distribution**

  Given a finite set of cumulative distribution functions and weights such that $w_k\geq 0$ and $\sum_{k=1}^K w_k=1$, the mixture distribution can be represented by the cumulative distribution function:

$$F(x)=\sum_{k=1}^K w_kF_j(x)$$

- **Definition 5.1.1 Random Sample**

  $X_1,\ldots,X_i$ are iid if:

  1. $X_1,\ldots,X_n$ are independent
  2. The CDF of X_i is F for all i.

- **Statistics and their distributions**

  Sample mean: $\overline{X} = \frac{X_1+...+X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

  Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

  Sample standard deviation: $S = \sqrt{S^2}$.

  Sample moment: $M_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, k = 1, 2, \ldots$

  Sample central moment: $\tilde{M}_j = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^j, k = 2, 3, \ldots$

  Empirical CDF: $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), x \in \mathbb{R}$.

- **Lemma 5.2.5**

  Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $E[g(X_1)]$ and $Var(g(X_1))$ exist. Then,

  $$E\left[\sum_{i=1}^{n} g(X_i)\right] = nE[g(X_1)] \quad and \quad Var\left(\sum_{i=1}^{n} g(X_i)\right) = nVar(g(X_1))$$

- **Theorem 5.2.6**

  Let $X_1, \ldots, X_n$ be a random sample from a population F on $\mathcal{R}$ with mean μ and variance σ^2. Then

  1. $E(\overline{X}) = \mu$.
  2. $Var(\overline{X}) = \sigma^2/n$.
  3. $E(S^2) = \sigma^2$.

- **Sampling Distribution**
- **Definition: Convergence in probability**
- **Theorem: Weak Law of Large Numbers (WLLN)**
- **Definition 5.5.6 Convergence almost surely**
- **Theorem 5.5.9 Strong Law of Large Numbers (SLLN)**
- **Definition: Convergence in distribution**
- **Theorem (Central Limit Theorem)**

  Let $X_1, X_2, \ldots$ be iid random variable with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$.

  Then, for any $x \in \mathbb{R}$,

  $$\lim_{n\to\infty} P(\sqrt{n}(\overline{X} - \mu)/\sigma \leq x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

  That is, $\sqrt{n}(\overline{X} - \mu)/\sigma$ converges in distribution to $Z \sim \mathcal{N}(0, 1)$.

- **The multivariate CLT**

  Let $X_1, X_2, \ldots$ be iid random vectors on $\mathbb{R}^k$ with $E(X_1) = \mu$ and finite covariance matrix Σ. Then $\sqrt{n}(\overline{X} - \mu)$ converges in distribution to a random vector $X \sim \mathcal{N}(0, \Sigma)$, the k-dimensional normal distribution with mean 0 and covariance matrix Σ.

Normal approximation to binomial [Lecture5_2]

- **Continuous mapping theorem**

  Let $X, X_1, X_2, \ldots$ be random k-vectors defined on a probability space S and g be a continuous function. Then,

  1. $X_n \xrightarrow{p} X$ implies $g(X_n) \xrightarrow{p} g(X)$;
  2. $X_n \xrightarrow{a.s} X$ implies $g(X_n) \xrightarrow{a.s} g(X)$;
  3. $X_n \xrightarrow{d} X$ implies $g(X_n) \xrightarrow{d} g(X)$.

- **Slutsky's theorem**

  Let $X, X_1, X_2, \ldots, Y_1, Y_2$ be random variables on a probability space. Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant.

  1. $X_n + Y_n \xrightarrow{d} X + c$;
  2. $Y_n X_n \xrightarrow{d} X$;
  3. $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$.

  Example 5.5.3 [Lecture 5_2]

Lecture 6

- **Method of Moments (MoM)**

  1. If the model has d parameters we compute the first d-th population moments,

     j-th population moment:

     $$\mu_j = E(x_1^j) = E(x_2^j) = \ldots E(x_n^j) = \begin{cases} \sum_x x^j f(x|\theta) \\ \int_x x^j f(x|\theta) dx \end{cases}$$
     $$\mu_j = g_j(\theta) = g_j(\theta_1, \ldots, \theta_d)$$

     The system of equations of population moments:

     $$\mu_j = \begin{cases} \mu_1 = g_1(\theta_1, \ldots, \theta_d) \\ \mu_2 = g_2(\theta_1, \ldots, \theta_d) \\ \vdots \\ \mu_d = g_d(\theta_1, \ldots, \theta_d) \end{cases}$$

  2. j-th sample moment:

     $$m_j = \frac{1}{n} \sum_{i=1}^{n} x_i^j$$

  3. **WLLN**

     $$m_j \xrightarrow[n \to \infty]{p} \mu_j \implies m_j = \mu_j$$

     Link the population moments to sample moments

$$m_j = \begin{cases} m_1 = g_1(\theta_1, \dots, \theta_d) \\ m_2 = g_2(\theta_1, \dots, \theta_d) \\ \vdots \\ m_d = g_d(\theta_1, \dots, \theta_d) \end{cases}$$

4. Solve for the d parameters as functions of sample moments.

$$\begin{cases} \tilde{\theta}_1 = h_1(m_1, \dots, m_d) \\ \tilde{\theta}_2 = h_2(m_1, \dots, m_d) \\ \vdots \\ \tilde{\theta}_d = h_d(m_1, \dots, m_d) \end{cases}$$

Example: MoM of normal distribution, binomial distribution, uniform distribution. [Lecture6-2]

- **Definition: Likelihood function**

  Let $X_1, .., X_n$ be an iid sample from a population with PDF or PMF $f(x|\theta)$. The likelihood function is defined by

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i|\theta).$$

  The log-likelihood function is defined by

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

- **Definition 7.2.4 MLE**

  A $\hat{\theta} = \Theta$ satisfying $\ell(\hat{\theta}) = \max_{\theta \in \Theta} \ell(\theta)$ is called a maximum likelihood estimator (MLE) of θ.

  To find MLE in one parameter case:

  1. Solve likelihood equation by 1st derivative
  2. For all candidates, check 2nd derivative
  3. Still need to calculate boundary see if it's global, but if there is only one candidates, we don't need step 3.

  Example: MLE of Poisson distribution, Normal distribution, Bernoulli distribution, ... [Lecture6-2].

- **Theorem 7.2.10 Invariance Property of MLE**

  If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta), \tau(\hat{\theta})$ is the MLE for $\tau(\theta)$.

- **Numerical Method: Newton-Raphson algorithm**

  If we want to find $\hat{\theta}$ such that $U(\hat{\theta}) = 0$ start with $\hat{\theta}^0$ and create a linear Taylor series approximation to $u(\theta)$ at $\hat{\theta}^0$.

$$u(\theta) \approx u(\hat{\theta}^0) + u'(\hat{\theta}^0)(\theta - \hat{\theta}_0) = 0$$
$$\implies \hat{\theta}' = \hat{\theta}_0 - \frac{u(\hat{\theta}^0)}{u'(\hat{\theta}^0)}$$

- **Bayesian Approach**

Lecture 7

- **Bias**

  If W is an estimator of a parameter θ, then the bias is

$$Bias(W) = E(W) - \theta$$

  An estimator is called unbiased if $E(W) = \theta$.

  Example of the bias of MLE [Lecture7-1].

- **Variance**

$$Var(W) = E\{W - E(W)\}^2$$

- **Bias-Variance Trade-offs from Mean Square Error**

$$E\{(W-\theta)^2\} = E\{W - E(W)\}^2 + (E(W)-\theta)^2 = Var(W) + Bias(W)^2$$

> **Example of variance and bias of S^2:**
>
> - The sample variance S^2 is defined as
>
> $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$
>
>   Since $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, from the properties of χ^2 we have
>
> $$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1 \implies E(S^2) = \sigma^2$$
>
>   and
>
> $$Var\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1) \implies Var(S^2) = \frac{2\sigma^4}{n-1}$$
>
> - Let $X_1, \ldots, X_n$ be iid from $\mathcal{N}(\mu, \sigma^2)$ with expected value μ and variance σ^2, then $\overline{X}$ is an unbiased estimator for μ, and S^2 is an unbiased estimator for σ^2.
>
>   We have
>
> $$E(\overline{X}) = \mu, \quad Var(\overline{X}) = \frac{\sigma^2}{n}$$
>
>   The MSE for S^2 is
>
> $$MSE_{S^2} = E(S^2 - \sigma^2) = Var(S^2) = \frac{2\sigma^4}{n-1}$$
>
>   The MLE of MoM method gives estimator for σ^2 is $\hat{\sigma}^2$.
>
> $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}S^2$$

> So it's a biased estimator for σ^2. The variance of $\hat\sigma^2$ is
>
> $$Var(\hat\sigma^2) = Var(\frac{n-1}{n}S^2) = \frac{(n-1)^2}{n^2}Var(S^2) = \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$$
>
> Hence the MSE of $\hat\sigma^2$ is given by
>
> $$E(\hat\sigma^2-\sigma^2)^2 = Var(\hat\sigma^2) + Bias^2 = \frac{2(n-1)\sigma^4}{n^2} + (\frac{n-1}{n}\sigma^2-\sigma^2)^2 = \frac{2n-1}{n^2}\sigma^4$$
>
> Comparing the two MSEs:
>
> $$MSE_{\hat\sigma^2} = \frac{2n-1}{n^2}\sigma^4 < \frac{2n}{n^2}\sigma^4 = \frac{2\sigma^4}{n} < \frac{2\sigma^4}{n-1} = MSE_{S^2}$$
>
> This shows that $\hat\sigma^2$ has smaller MSE than S^2.

- **UMVUE**

  An estimator W^* is called a UMVUE of $\tau(\theta)$ if it satisfies $E(W)=\tau(\theta)$ for all θ (It's unbiased), and, for any other unbiased estimator W, we have $Var(W^*) \le Var(W)$ for all θ.

  We can identify a lower bound (**Cramer-Rao Lower Bound**) on the variance of unbiased estimators. If an unbiased estimator has variance equal to the bound, then we know we have a UMVUE.

- **Cramer-Rao Inequality**

  Let $X_1,\ldots,X_n$ be a sample with PDF $f(X|\theta)$, and let $W(X)$ be any estimator satisfying

  $$\frac{d}{d\theta}E\{W(X)\} = \int \frac{\partial}{\partial\theta}\{W(x)f(x|\theta)\}dx$$

  and $Var(W(X)) < \infty$. Then

  $$Var(W(X)) \ge \frac{[\frac{d}{d\theta}E\{W(X)\}]^2}{E[\frac{\partial}{\partial\theta}\log f(X|\theta)]^2}$$

  Prove it by Cauchy-Swartz's Inequality [Lecture7-1].

- **Cramer-Rao Lower-Bound**

  Let $X_1,\ldots,X_n$ be iid sample with PDF $f(x|\theta)$. Suppose W is an **unbiased** estimator for θ, we have

  $$Var(W) \ge \frac{1}{nE\left[\{\frac{\partial}{\partial\theta}\log f(X_1|\theta)\}^2\right]}$$

  where $nE\left[\{\frac{\partial}{\partial\theta}\log f(X_1|\theta)\}^2\right]$ is known as the **Fisher Information**.

  Let $X_1,\ldots,X_n$ be iid sample with PDF $f(x|\theta)$. Suppose W is an unbiased estimator for $\tau(\theta)$, we have

  $$Var(W) \ge \frac{\{\tau'(\theta)\}^2}{nE\left[\{\frac{\partial}{\partial\theta}\log f(X_1|\theta)\}^2\right]}$$

  Prove the CRLB [Lecture7-1].

- **Fisher information** can be calculated from log likelihood

$$E\left[\left\{\frac{\partial}{\partial\theta}\log f(x_1,\ldots,x_n|\theta)\right\}^2\right]=E\left[\left\{\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_1|\theta)\right\}^2\right]=nE\left[\left\{\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right\}^2\right]$$

Calculate Fisher information: Take log of the PDF => take derivative => square => take expectation

- Fisher information is defined as follows for a random variable X.

$$I(\theta)=E\left[\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}^2\right]$$

- The fisher information can be rewritten as

$$I(\theta)=E\left[\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}^2\right]=Var\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]$$

because $E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]=0$.

Also, we have

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

- Finally, the fisher information can also be written as

$$I(\theta)=-E\left[\frac{\partial^2}{\partial^2\theta}\log f(X|\theta)\right]=-\int\left[\frac{\partial^2}{\partial^2\theta}\log f(X|\theta)\right]f(x|\theta)dx$$

- **Fisher Information Matrix**

Fisher information contained in the random variable X following distribution $f(x|\theta)$ is defined as

$$E\left[\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}^2\right]$$

First Bartlett identity

$$E\left\{\frac{\partial\log f(x|\theta)}{\partial\theta}\right\}=0$$

Second Bartlett identity

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

Prove the second Bartlett identity [Lecture7-1].

Example: fisher information matrix for normal distribution.

- **Optimality and Decision Theory**
- **UMVUE**
  - If an estimator is unbiased and the asymptotic variance reaches the CRLB, it is the UMVUE of the parameter.
- **Bayes Optimality**
- **Minimax Optimality**

Lecture 8

- **Consistency**

  Let $W(X_1, \ldots, X_n)$ be an estimator for θ based on random sample $X_1, \ldots, X_n$. W is an consistent estimator if for any $\epsilon > 0$.

  $$\lim_{n\to\infty} P(|W - \theta| > \epsilon) \to 0$$

  If W is consistent, then

  - Asymptotic bias: $\lim_{n\to\infty} E(W) - \theta \to 0$.
  - Variance: $\lim_{n\to\infty} Var(W) \to 0$.

  Example: Consistency of MoM, MLE.

- **Consistency of MLE**

  If $X_1, \ldots, X_n$ are iid from a density $f(x|\theta)$ and the below conditions 1,2, and 3 hold, then the solution $\hat{\theta}$ of the likelihood equation $\partial l(\theta)/\partial\theta = 0$ is a consistent estimator for θ.

  1. Identifiability: for $\theta_1, \theta_2 \in \Theta$, $f(x;\theta_1) = f(x;\theta_2)$ for all x implies $\theta_1 = \theta_2$.
  2. θ_0 is a interior point of Θ (Not on the boundary).
  3. Common support and differentiable: $f(x|\theta)$ have a common support for all θ, and $\log f(x|\theta)$ is differentiable in θ (Does not apply if support depends on θ).

- **Definition 10.1.7: Limiting Variance**
- **Asymptotic normality**

  For estimator W, if $\lim_{n\to\infty} \sqrt{n}\{W - \tau(\theta)\} \to \mathcal{N}(0, v(\theta))$, we say W is asymptotically normal with asymptotic variance $v(\theta)$. Suppose $X_1, \ldots, X_n$ are iid samples with mean μ and variance σ^2. By **CLT**, we have

  $$\sqrt{n}(\overline{X} - \mu) \to \mathcal{N}(0, \sigma^2)$$

  Thus, $\overline{X}$ is asymptotically normal with asymptotic variance σ^2.

- **Asymptotic efficiency**

  Estimator $W(X_1, \ldots, X_n)$ is asymptotic efficient for $\tau(\theta)$ where $\tau'(\theta) \neq 0$ if

  $$\lim_{n\to\infty} \sqrt{n}\{W - \tau(\theta)\} \xrightarrow{d} \mathcal{N}(0, v(\theta))$$

  and

  $$v(\theta) = \frac{\{\tau'(\theta)\}^2}{E\left[\{\frac{\partial}{\partial\theta} \log f(X_1|\theta)\}^2\right]}$$

  That is, the asymptotic variance reaches the Cramer-Rao Lower Bound.

- **Prove MLE normality**

  - Taylor Expansion

  $$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!} + \ldots + \frac{f^{(k-1)}(a)}{(k-1)!}(x-a)^{k-1} + \frac{f^{(k)}(a')}{k!}(x-a)^k$$

for some a' between a and x, satisfying $|a'-a| \leq |x-a|$ (So $a < a' < x$).

  - Prove: [Lecture8-4 & Lecture8_updated2]

- **Condition for MLE normality**

  1. Identifiability: for $\theta_1, \theta_2 \in \Theta$, $f(x;\theta_1) = f(x;\theta_2)$ for all x implies $\theta_1 = \theta_2$.
  2. θ_0 is a interior point of Θ.
  3. Common support and differentiable: $f(x|\theta)$ have a common support for all θ, and $\log f(x|\theta)$ is differentiable in θ.
  4. Concavity: $I(\theta) = -E\{d^2 \log f(X_i|\theta)/d\theta^2\} > \rho > 0$.
  5. Integratable third derivative: $|\partial^3 \log f(x|\theta)/\partial\theta^3| \leq g(x)$, for some function $g(x)$ such that $Eg(x) < \infty$.

- **Asymptotic normality of MLE**

  If $X_1, \ldots, X_n$ are iid from a density $f(x|\theta)$ and the conditions in the previous slide hold, then the MLE $\hat{\theta}$ which satisfies the likelihood equation $\partial l(\theta)/\partial\theta = 0$ is asymptotically normal, and

  $$\sqrt{n}(\hat{\theta} - \theta) \rightarrow \mathcal{N}(0, I(\theta)^{-1})$$

  where $I(\theta) = E[\{\frac{\partial}{\partial\theta} \log f(X_1|\theta\}^2]$ is the Fisher Information.

  Example of exact and limiting distribution of MLE of binomial distribution, normal distribution, ... [Lecture8-4]

- **Estimating the asymptotic variance of MLEs**

  Since

  $$\sqrt{n}(\hat{\theta} - \theta) \rightarrow \mathcal{N}(0, I^{-1}(\theta))$$

  The reasonable estimator for $I^{-1}(\theta)$ could be

  1. Close-form expression of $I(\theta)$.
  2. $I(\theta) = -E\{\frac{\partial^2}{\partial\theta^2} \log f(x_i|\theta)\}$,
     $\widehat{I}(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log f(x_i|\theta) \xrightarrow{p} -E\{\frac{\partial^2}{\partial\theta^2} \log f(x_i|\theta)\} = I(\theta)$.
     $\widehat{I}(\hat{\theta}) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log f(x_i|\hat{\theta})$
  3. $I(\theta) = E[\{\frac{\partial}{\partial\theta} \log f(x_i|\theta)\}^2]$
     $\check{I}(\theta) = \frac{1}{n}\sum_{i=1}^n \{\frac{\partial}{\partial\theta} \log f(x_i|\theta)\}^2$
     $\check{I}(\hat{\theta}) \xrightarrow{p} I(\theta)$

- **Asymptotic relative efficiency (ARE)**

  Estimators W_1 and W_2 satisfy

  $\lim_{n\rightarrow\infty} \sqrt{n}\{W_1 - \tau(\theta)\} \xrightarrow{d} \mathcal{N}(0, v_1)$ and $\lim_{n\rightarrow\infty} \sqrt{n}\{W_2 - \tau(\theta)\} \xrightarrow{d} \mathcal{N}(0, v_2)$.

  Then the asymptotic relative efficiency of W_2 with respect to W_1 is

  $$ARE(W_2, W_1) = \frac{v_1}{v_2}$$

  Example of ARE of Poisson estimator [Lecture8-4].

- **Parameter transformation and Delta Method**

If W is asymptotic normal, i.e., $\sqrt{n}(W-\theta) \xrightarrow{d} \mathcal{N}\left(0, v(\theta)\right)$, then for a differentiable function $h(\theta)$, such that $h'(\theta) \neq 0$, we have

$$\sqrt{n}(h(W)-h(\theta)) \xrightarrow{d} \mathcal{N}(0, \{h'(\theta)\}^2 v(\theta))$$

Prove Delta Method using Taylor expansion [Lecture8-4].

Example of estimating odds [Lecture8-4].

- **Second-order Delta Method**

  If W is asymptotic normal, i.e., $\sqrt{n}(W-\theta) \xrightarrow{d} \mathcal{N}(0, v(\theta))$, then for a differentiable function $h(\theta)$, if $h'(\theta) = 0$ and $h''(\theta) \neq 0$, we have

$$n(h(W)-h(\theta)) \xrightarrow{d} \frac{h''(\theta)v(\theta)}{2}\chi_1^2$$

  Prove it by Taylor expansion [Lecture8-4].

Resources

- Wiki: Summation Identities
- Wiki: List of Math Series
- Wiki: List of Convolutions of Probability Distribution
- Examples for midterms

  https://people.missouristate.edu/songfengzheng/Teaching/MTH541F21.htm
- Examples for midterms

  https://www2.econ.iastate.edu/classes/econ671/hallam/
- Relevant lecture notes

  https://www.stat.cmu.edu/~larry/=stat705/
- MIT OCW lecture notes and assignment

  https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-spring-2015/lecture-notes/