Review for Midterm-I

Chapter 5. Statistical Inference

§ 5.1 Sampling and Statistics

- **Random sample**: The random variables $X_1, \ldots, X_n$ constitute a random sample from a random variable X if they are independent and have identical distribution as X, which is denoted by

$$X_1, \ldots, X_n \text{ are iid } \sim F(x) \text{ or } f(x),$$

where $F(x)$ and $f(x)$ are the cdf and pdf of X respectively.

- **Statistic**: Function of the sample $\{X_1, \ldots, X_n\}$: $T = T(X_1, \ldots, X_n)$
- **Point estimator**: The statistic T is called a point estimator of the unknown parameter θ if the value of T can be used to estimate θ.
- **Unbiasedness**: T is an unbiased estimator of θ if $E(T) = \theta$.
- **Consistency**: T is a consistent estimator of θ if T converges to θ in probability, i.e. $T \xrightarrow{P} \theta$.
- **Confidence interval**: An interval based on the statistic T is called a $100\,(1-\alpha)\%$ confidence interval for θ if the probability of the event that the interval covers θ is $(1-\alpha)$.
- For practice: Example 5.1.1, Exercise 5.1.2

§ 5.2.1 Quantiles

- **Quantile**: For $0 < p < 1$, the pth quantile of a random variable X is $\xi_p = F^{-1}(p)$, where $F(x)$ is the cdf of X. Note: If $F(x)$ is not monotone, we may define $F^{-1}(p) = \min\{x : F(x) \geq p\}$.
- **Order statistics**: Let $X_1, \ldots, X_n$ denote a random sample. Rewrite the sample in ascending order and obtain $X_{(1)} \leq X_{(2)} \leq \cdots X_{(n)}$, which are called the order statistics of the sample.
- **Percentile**: Let $X_1, \ldots, X_n$ be a random sample and let $X_{(1)} \leq X_{(2)} \leq \cdots X_{(n)}$ be the corresponding order statistics. For $0 < p < 1$, let $k = [p(n+1)]$, where $[x]$ represents the greatest integer m such that $m \leq x$. Then $X_{(k)}$ is called the pth sample quantile or the $100p$th percentile of the sample.
- **Five number summary**: The minimum $\left(Min = X_{(1)}\right)$, the first quartile ($Q_1$), the median ($Q_2$), the third quartile ($Q_3$), and the maximum ($\text{Max} = X_{(n)}$).

- **Boxplot**: Box part (Q_1, Q_2, and Q_3), whisker part ($LF = Q_1 - -1.5 * IQR, UF = Q_3 + 1.5 * IQR$), where the interquartile range $IQR = Q_3 - Q_1$. Potential outliers (points that lie outside of (LF, UF)).
- For practice: Example 5.2.4, Example 5.2.5

§ 5.5 Hypothesis Testing

- Null hypothesis H_0 versus alternative hypothesis H_1. For example:
  $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ (one-sided test);
  $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (two-sided test)
- **Type I error** (or significance level, or size of test):
  $\alpha = P(\text{reject } H_0 | \theta = \theta_0)$.
- **Type II error**: $\beta(\theta_1) = P(\text{do not reject } H_0 | \theta = \theta_1), \theta_1 \neq \theta_0$
- **Power function** (or power of test at $\theta = \theta_1$):
  $1 - \beta(\theta_1) = P(\text{reject } H_0 | \theta = \theta_1)$
- For practice: Example 5.5.2, Example 5.5.4

§ 5.7 Chi-Square Test

- Chi-Square Test for Multinomial Distribution: $H_0 : p_i = p_{i0}, i = 1, \cdots, k$. Given that H_0 is true, then

  $$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - n \cdot p_{io})^2}{n \cdot p_{io}} \sim \chi^2 (k-1).$$

  Critical region is $\{Q_{k-1} > \chi^2_\alpha (k-1)\}$ if the significance level is α.
- Chi-Square Test for independence of the two variables in contingency table: $H_0 : p_{ij} = p_{i\cdot} \cdot p_{\cdot j}, \forall i = 1, \cdots, a, j = 1, \cdots, b$. Given that H_0 is true, then

  $$Q = \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{(X_{ij} - n \cdot \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j})^2}{n \cdot \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j}} \sim \chi^2 ((a-1)(b-1))$$

  where $\hat{p}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{b} X_{ij}$, $\hat{p}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{a} X_{ij}$.
  Critical region is $\{Q_{k-1} > \chi^2_\alpha ((a-1)(b-1))\}$
- For practice: Example 5.7.1

§ 5.8 The Method of Monte Carlo (not required)

Chapter 6. Maximum Likelihood Methods

§ 6.1 Maximum Likelihood Estimation

- **Likelihood function**: $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$, if $X_1,\ldots,X_n$ are iid $\sim f(x;\theta)$.
- **Log likelihood function**: $l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta)$
- **M.L.E.**: The value of θ which maximizes $L(\theta)$ or $l(\theta)$ is called the maximum likelihood estimator of θ, denoted by $\hat{\theta}$ or $\hat{\theta}_{MLE}$.
- **Theorem 6.1.1**: Let θ_0 be the true value of θ. Under regularity conditions,
$$\lim_{n\to\infty} P_{\theta_0}[L(\theta_0;\mathbf{X}) > L(\theta;\mathbf{X})] = 1, \text{ for all } \theta \neq \theta_0$$
- **Theorem 6.1.2**: Suppose $\hat{\theta}$ is the mle of θ, then $g(\hat{\theta})$ is the mle of $g(\theta)$.
- **Theorem 6.1.3**: Under regularity conditions, the likelihood equation $\frac{\partial}{\partial\theta} l(\theta) = 0$ has a solution $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{P} \theta_0$.
- **Corollary 6.1.1**: Under regularity conditions, if the likelihood equation has a unique solution $\hat{\theta}_n$, then $\hat{\theta}_n$ is a consistent estimator of θ_0 .
- For practice: Example 6.1.1, Example 6.1.2, Example 6.1.5, Example 6.1.6

§ 6.2 Rao-Cramér Lower Bound and Efficiency

- Let X be a random variable with pdf $f(x;\theta)$, $\theta \in \Omega$. Under regularity conditions,
$$E\left[\frac{\partial \log f(X;\theta)}{\partial\theta}\right] = 0, \quad E\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right]$$
- **Fisher information of a single random variable** X:
$$I(\theta) = E\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right] = \mathrm{Var}\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right) = -E\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right]$$
- **Fisher information of a random sample** $X_1,\ldots,X_n$**:** $nI(\theta)$
- **Theorem 6.2.1** (Rao-Cramér Lower Bound): Let $X_1,\ldots,X_n$ be iid $\sim f(x;\theta)$. Let $Y = u(X_1,\ldots,X_n)$ be a statistic with mean $k(\theta)$. Under regularity conditions,
$$\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)}$$

- **Corollary 6.2.1**: Under the assumptions of Theorem 6.2.1, if Y is an unbiased estimator of θ, then $\text{Var}(Y) \geq \frac{1}{nI(\theta)}$.
- **Efficient estimator**: Let Y be an unbiased estimator of θ. Y is called an efficient estimator of θ if $\text{Var}(Y)$ attains the Rao-Cramér lower bound.
- **Efficiency**: Let Y be an unbiased estimator of θ. Then $\frac{1}{nI(\theta)}[\text{Var}(Y)]^{-1}$ is called the efficiency of Y.
- **Theorem 6.2.2**: Assume $X_1, \ldots, X_n$ are iid with pdf $f(x;\theta_0)$. Suppose $0 < I(\theta_0) < \infty$ and $\hat{\theta}_n$ is a consistent estimator of θ_0 such that $\frac{\partial}{\partial\theta} l(\hat{\theta}_n) = 0$. Under regularity conditions,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N\left(0,\ \frac{1}{I(\theta_0)}\right)$$

- **Corollary 6.2.2**: Suppose $g(x)$ is differentiable at θ_0 and $g'(\theta_0) \neq 0$. Under the assumptions of Theorem 6.2.2,

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{D} N\left(0,\ \frac{[g'(\theta_0)]^2}{I(\theta_0)}\right)$$

- For practice: Example 6.2.1, Example 6.2.2, Example 6.2.3

§ 6.3 Maximum Likelihood Tests

- **Likelihood ratio test statistic**: $\Lambda = L(\theta_0)/L(\hat{\theta})$, where $\hat{\theta}$ is the mle of θ.
- **Likelihood ratio test (two-sided)**: Reject H_0 in favor of H_1 if $\Lambda \leq c$.
- **Theorem 6.3.1**: Assume the same regularity conditions as for Theorem 6.2.2 . Under the null hypothesis $H_0 : \theta = \theta_0$, $-2\log\Lambda \xrightarrow{D} \chi^2(1)$.
- For practice: Example 6.3.1, Example 6.3.2

§ 6.4 Multiparameter Case: Estimation

- **Parameters**: Let $X_1, \ldots, X_n$ be iid $\sim f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$.
  Likelihood function: $L(\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i;\boldsymbol{\theta})$;
  Log likelihood function: $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(x_i;\boldsymbol{\theta})$.
  MLE: The value of $\boldsymbol{\theta}$ which maximizes $L(\boldsymbol{\theta})$ or $l(\boldsymbol{\theta})$ is called the maximum likelihood estimator (mle) of $\boldsymbol{\theta}$ and denoted by $\hat{\boldsymbol{\theta}}$.
- Under regularity conditions, the gradient is

$$\nabla \log f(x,\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1}\log f(X;\boldsymbol{\theta}), \cdots, \frac{\partial}{\partial\theta p}\log f(X;\boldsymbol{\theta})\right)$$

$$E\left[\nabla \log f(x,\boldsymbol{\theta})\right]=0, E\left[\frac{\partial}{\partial\theta_j}\log f(X;\boldsymbol{\theta})\right]=0, \quad \text{for } j=1,\ldots,p\,;$$

$$\text{Cov}\left(\frac{\partial}{\partial\theta_j}\log f(X;\boldsymbol{\theta}),\ \frac{\partial}{\partial\theta_k}\log f(X;\boldsymbol{\theta})\right)=-E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X;\boldsymbol{\theta})\right]$$

- **Fisher information matrix**:

$$\mathbf{I}(\boldsymbol{\theta})=Cov\left(\nabla \log f(x,\boldsymbol{\theta})\right)=(I_{jk})_{p\times p},$$

  where

$$I_{jk}=-E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(X;\boldsymbol{\theta})\right]$$

- **Theorem**: Under regularity conditions, if $Y=u(X_1,\ldots,X_n)$ is an unbiased estimate of θ_j, then

$$\text{Var}(Y)\geq\frac{1}{n}[\mathbf{I}^{-1}(\boldsymbol{\theta})]_{jj}$$

  .

- **Theorem 6.4.1**: Let $X_1,\ldots,X_n$ be iid $\sim f(x;\boldsymbol{\theta})$. Under regularity conditions, any consistent solution sequence $\hat{\theta}_n$ of the likelihood equation $\frac{\partial}{\partial\boldsymbol{\theta}}l(\boldsymbol{\theta})=\mathbf{0}$

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})\xrightarrow{D}N_p(\mathbf{0},\ \mathbf{I}^{-1}(\boldsymbol{\theta}))$$

- **Corollary**. Let $\mathbf{g}$ be a transformation $\mathbf{g}(\boldsymbol{\theta})=(g_1(\boldsymbol{\theta}),\cdots,g_k(\boldsymbol{\theta}))^{\ T}$, such that $1\leq k\leq p$. and the $k\times p$ matrix of partial derivatives:

$$\mathbf{B}=\left[\frac{\partial g_i}{\partial\theta_j}\right], i=1,\cdots,k; j=1,\cdots,p$$

  has continuous elements and does not vanish in a neighborhood of $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\eta}}=\mathbf{g}\left(\hat{\boldsymbol{\theta}}\right)$, then $\hat{\boldsymbol{\eta}}$ is the mle of $\boldsymbol{\eta}=\mathbf{g}(\boldsymbol{\theta})$, and

$$\sqrt{n}(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta})\xrightarrow{D}N\left(\mathbf{0},\mathbf{B}\mathbf{I}^{-1}(\boldsymbol{\theta})\ \mathbf{B}^T\right)$$

  and the information matrix for $\boldsymbol{\eta}$ is $\mathbf{I}(\boldsymbol{\eta})=\left(\mathbf{B}\mathbf{I}^{-1}(\boldsymbol{\theta})\ \mathbf{B}^T\right)^{-1}$

- For practice: Example 6.4.1, Example 6.4.3, Example 6.4.4, Example 6.4.6