MA 781: Final Notes

Benjamin Draves

December 18, 2017

1 Preliminaries

- **Definition**: A family of densities is called an exponential family if we can write it as

$$f(x,\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

- **Definition**: The family of densities

$$\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$$

is called a location-scale family. The $X \sim \frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ iff there exists $Z \sim f(z)$ such that $X = \sigma Z + \mu$

- Some common inequalities
  1. (Markov Inequality) $P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$
  2. (Generalized Markov Inequality) For an increasing function $g(\cdot)$ then $P(X \geq a) \leq \frac{1}{g(a)}\mathbb{E}[g(x)]$
  3. (Chebyshev) $P(|X - E[X]| \geq a) \leq \frac{1}{a^2}Var(x)$
  4. (Jensen) If $g(\cdot)$ is convex $\mathbb{E}[g(x)] \geq g[\mathbb{E}(X)]$. If $g(\cdot)$ is concave then $\mathbb{E}[g(x)] \leq g[\mathbb{E}(X)]$

2 Properties of a Random Sample

2.1 Order Statistics

- Let $Y_i = X_{(i)}$ for $i = 1, 2, \ldots, n$. Then we say Y_i is the ith order statistic.
- Some useful distributions are given by
  1. $g(\mathbf{y}) = n!\prod_{i=1}^{n} f_X(y_i)$
  2. $G_1(y) = 1 - [1 - F_X(y)]^n \quad \& \quad G_n(y) = [F_X(y)]^n$
  3. $g_1(y) = n[1 - F_X(y)]^{n-1}f_x(y) \quad \& \quad G_n(y) = n[F_X(y)]^{n-1}f_x(y)$

2.2 Convergence Topics

- **Theorem**: (**Continuous Mapping**) If $X_n \to X$ in any mode and $g(\cdot)$ is continuous then $g(X_n) \to g(X)$ in the same mode.
- **Definition**: Suppose that $F_n(x) \to F(x)$. That is $X_n \xrightarrow{D} X$ then we say that X_n has limiting distribution $F(x)$
- **Definition**: X_n has asymptotic distribution (μ, σ^2) denoted $X_n \sim AN(\mu, \sigma^2)$ iff

$$\frac{X_n - \mu}{\sigma^2} \xrightarrow{D} Z$$

- **Theorem**: (**CLT**) Let $\mathbf{x}$ be a random sample from $X \sim f$. Then for $Z_n := \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} \xrightarrow{D} Z$

- **Theorem**: (**Delta Method 1**) If $X_n \sim AN(\mu, \sigma^2)$ and $g(\cdot)$ is differentiable with $g'(\mu) \neq 0$ then

$$g(X_n) \sim AN(g(\mu), [g'(\mu)]^2\sigma^2)$$

- **Theorem**: (**Delta Method 2**) If $X_n \sim AN(\mu, \sigma^2)$ and $g(\cdot)$ is differentiable with $g'(\mu) = 0$ and $g''(\mu) \neq 0$ then

$$\sqrt{n}[g(X_n) - g(\mu)] \xrightarrow{D} \frac{g''(\mu)\sigma^2}{2}\chi^2(1)$$

- **Theorem**: (**Variance Stabilizing Transformation**) By the Delta method one can write

$$\sqrt{n}(g(\overline{x}) - g(\mu)) \xrightarrow{D} N(0, [g'(\mu)]^2\sigma^2)$$

  Our goal, *to stabilize the variance*, we look to find a function $g(\cdot)$ such that $[g'(\mu)]^2\sigma^2 = k^2$ where k is a constant. Then by solving this ODE, we can find g such that variance is stabilized.

3 Principles of Data Reduction

3.1 The Sufficiency Principle

- The entire idea around sufficiency is to attain a more simple form of a sample. With large samples, we want a simple summary that still maintains all of the information inherent in a sample $\mathbf{x}$.
- Motivating question: Is there a function of our data (a *statistic*) $T(\mathbf{x})$ with $T : \mathcal{X} \to \mathbb{R}$ such that the information in $\mathbf{x}$ is equivalent to the information in $T(\mathbf{x})$. That is $T(\mathbf{x})$ is sufficient.
- If $p < n$, we achieve *data reduction*. That is our statistic simplifies our inference by considering $T(\mathbf{x})$ instead of $\mathbf{x}$.
- The Sufficiency Principle: If $T(\mathbf{x})$ is a sufficient statistic for a parameter θ then any inference about θ should depend on $\mathbf{x}$ only through $T(\mathbf{x})$.

3.2 Sufficient Statistics

- **Definition**: A statistic is called a sufficient statistic for θ iff the conditional distribution of $\mathbf{x}|T(\mathbf{x}) = t$ does not depend on θ. That is

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n | T(\mathbf{x}) = t)$$

  is free from θ.
- **Theorem**: (**Neyman - Fisher**) $T(\mathbf{x})$ is a sufficient statistic iff $f(\mathbf{x}, \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$ for all possible $\mathbf{x}$ and θ.
- **Theorem**: (**Neyman - Fisher**) Let $q(T(\mathbf{x}), \theta)$ be the distribution of a statistic $T(\mathbf{x})$. $T(\mathbf{x})$ is a sufficient statistic iff

$$\frac{f(\mathbf{x}, \theta)}{q(T(\mathbf{x}), \theta)}$$

  is free from θ.
- Sufficient statistics need not be unique (order statistics and full sample for example)
- Any $1 - 1$ function of a sufficient statistic is also a sufficient statistic.
- **Theorem**: Let $\mathbf{x}$ be a sample from an exponential family. Then

$$T = (T_1, \ldots, T_k) = \left(\sum_{i=1}^{n} t_1(x_i), \ldots, \sum_{i=1}^{n} t_k(x_i)\right)$$

  is a sufficient statistic for $\theta = (\theta_1, \ldots, \theta_p)$.
- **Theorem**: (**N-S Conditions for SS**) For each $\theta_1 \neq \theta_2$ then

$$\frac{f(\mathbf{x}, \theta_1)}{f(\mathbf{x}, \theta_2)} = \frac{g(T(\mathbf{x}), \theta_1)}{g(T(\mathbf{x}), \theta_2)} = r(T(\mathbf{x}))$$

  is θ free.

- **Theorem**: Let $\theta_1 \neq \theta_2$ and $\mathbf{x}_1$ and $\mathbf{x}_2$ be two samples with $T(\mathbf{x}_1) = T(\mathbf{x}_2)$. If

$$\frac{f(\mathbf{x}_1, \theta_1)}{f(\mathbf{x}_1, \theta_2)} \neq \frac{f(\mathbf{x}_2, \theta_1)}{f(\mathbf{x}_2, \theta_2)}$$

then $T(\mathbf{x})$ is **not** a sufficient statistic.

3.3 Minimal Sufficient Statistics

- **Definition**: $T(\mathbf{x})$ is called a minimal sufficient statistic if for any other sufficient statistic $S(\mathbf{x})$ then there exists $\phi_S(\cdot)$ such that

$$T(\mathbf{x}) = \phi_S(S(\mathbf{x}))$$

- MSS provide the greatest data reduction (in a sense they are necessary statistics).
- **Theorem**: (**Lehman - Scheffe**) Suppose we have two samples $\mathbf{x}_1$,$\mathbf{x}_2$. Then if we have:

$$\frac{f(\mathbf{x}_1, \theta)}{f(\mathbf{x}_2, \theta)} \quad \text{free from } \theta \text{ iff} \quad T(\mathbf{x}_1) = T(\mathbf{x}_2)$$

then $T(\mathbf{x})$ is a minimal sufficient statistic for θ.

3.4 Ancillary Statistics

- **Definition**: A statistic $A(\mathbf{x})$ is called an ancillary statistic iff the distribution of $A(\mathbf{x})$ is free from θ.
- Basically, the statistic contains no information about the parameter in question.
- **Definition**: A statistic $A(\mathbf{x})$ is first order ancillary iff $\mathbb{E}[A(\mathbf{x})]$ is free from θ.
- **Theorem**: If a statistic is location and scale invariant, i.e.

$$T(aX_1 + b, \ldots, aX_n + b) = T(X_1, \ldots, X_n)$$

and $\mathbf{x} \sim f$ where f is a location scale model then $T(\mathbf{x})$ is an AS.

3.5 Complete Sufficient Statistics

- Ideally, a sufficient statistic and an ancillary statistic should be independent. Unfortunately they aren't.
- One useful example: Consider $Unif(\theta, \theta + 1)$. Then $(X_{(1)}, X_{(n)})$ is MSS and $T(\mathbf{x}) := (X_{(n)} - X_{(1)}, \frac{X_{(n)} - X_{(1)}}{2})$ is MSS. But $A(\mathbf{x}) := X_{(n)} - X_{(1)}$ is AS so $T(\mathbf{x}) \not\perp A(\mathbf{x})$
- Motivation: Are there sufficient statistics that are independent to ancillary statistics? If so, what additional properties do we require?
- **Definition**: A family of distributions $\mathcal{F}$ is complete iff for any measurable function $g(\cdot)$ with $\mathbb{E}(g(x)) = 0$ for all $\theta \in \Theta$ then $P(g(x) = 0) = 1$.
- **Definition**: A statistic $T(\mathbf{x})$ is complete iff $\mathcal{F}_T = \{f_T(\mathbf{x}, \theta)\}$ is complete.
- Any $1-1$ function of a CSS is also complete.
- **Theorem**: (**CSS for Exponential Families**) Suppose $\mathbf{x} \sim f$ where f is an exponential family. Then

$$T(\mathbf{x}) = \left(\sum_{i=1}^{n} t_1(x_i), \ldots, \sum_{i=1}^{n} t_k(x_i) \right)$$

is a CSS provided that $(w_1(\theta), \ldots, w_k(\theta))$ contains an open set in $\mathbb{R}^k$.
- **Theorem**: (**Basu**) If $T(\mathbf{x})$ is CSS and $A(\mathbf{x})$ is AS then $T(\mathbf{x}) \perp A(\mathbf{x})$
- **Theorem**: (**Wackerly**) If $T(\mathbf{x})$ is CSS then $T(\mathbf{x})$ is MSS.

4 Point Estimation

4.1 Methods for Finding Estimators

4.1.1 Substitution Method

- Motivation: Suppose we have some distribution F and we want to estimate a parameter based on F (e.g. μ, σ^2,ξ_p). If we can find a good estimator $\hat{F}$, then we simply plug-in $\hat{F}$ into our functional to provide an estimate
- Questions that arise: Can we find a good estimator of F? Can every parameter of interest be written as $\theta(F)$?
- One possible estimator of F is given by the *empirical distribution function* defined as

$$\hat{F}(x) = \begin{cases} 0 & x < X_{(1)} \\ k/n & X_{(k)} < x < X_{(k+1)} \\ 1 & X_{(n)} < x \end{cases}$$

- **<u>Theorem</u>**: $n\hat{F}(x) \sim Binom(n, F(x))$
- $\hat{F}$ is consistent and strongly consistent for F
- **<u>Theorem</u>**: $\hat{F} \sim AN(F(x), \frac{1}{n}F(x)(1-F(x)))$
- **<u>Theorem</u>**: (**Continuity Property of Plug-in Estimator**) Let $h(\cdot)$ be continuous and $g(\cdot)$ is Borel. Then

$$h\left(\sum_{i=1}^{n} g(x_i)\right) \xrightarrow{a.s.} h\left(\int_{\mathbb{R}} g(x)dF(x)\right)$$

4.1.2 Method of Moments

- **<u>Definition</u>**: The <u>population moments</u> of a parametric distribution F are given by

$$\mu_k := \mathbb{E}(X^k) = \int x^k dF(x)$$

- **<u>Definition</u>**: The <u>sample moments</u> are given by

$$m_k := \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

- **<u>Definition</u>**: Suppose we have a parameter $\theta = (\theta_1, \ldots, \theta_p)$. Then the <u>method of moment estimators</u> are given by the solutions to the system of equations given by $\{m_k = \mu_k\}_{k=1}^{t}$ for $t \geq p$.

4.1.3 Maximum Likelihood

- In the likelihood setting, we consider the joint density $f(\mathbf{x})$ parameterized by θ as a two dimensional function $f(\mathbf{x}, \theta)$. The density measures the probability density of the sample, so *given the data* we want to maximize the probability density as a function of θ.
- We define a function $\mathcal{L}(\theta|\mathbf{x}) := f(\mathbf{x}, \theta)$ and we look to *maximize the likelihood.*
- **<u>Definition</u>**: The <u>maximum likelihood estimate</u> is given by

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max} L(\theta|\mathbf{x})$$

- We can find these through calculus methods (check second derivatives!) or through direct arguments
- **<u>Theorem</u>**: (**Invariance Principle of MLE**) If $\hat{\theta}_{MLE}$ is MLE for θ then for any measurable function $g(\cdot)$, we have

$$\widehat{g(\theta)}_{MLE} = g(\hat{\theta}_{MLE})$$

- MLE needs not be unique - we can have uncountably many. Consider the example Unif$(\theta - 1/2, \theta + 1/2)$.
- If the MLE is unique then $\hat{\theta}_{MLE} = \phi(T(\mathbf{x}))$ for any sufficient statistic $T(\mathbf{x})$.

4.1.4 Minimization (M) Estimators

- Motivation: In MLE we look to maximize $\mathcal{L}(\theta|\mathbf{x})$ or $\ell(\theta|\mathbf{x}) := \log(\mathcal{L}(\theta|\mathbf{x}))$. Which is equivalent to minimizing $-\ell(\theta|\mathbf{x})$. Why only $\log(\cdot)$? Are there other functions that provide nice properties?
- **Definition**: Suppose we have a nonparametric family $\mathcal{F}$ and we have this function $\psi(x,t)$. Then the M estimator is given by $\hat{T} = T(\hat{F})$; the solution to

$$\int \psi(x, T(\hat{F}))d\hat{F}(x) = \sum_{i=1}^{n} \psi(x_i, T(\hat{F})) = 0$$

- MLE is a special case of M estimators with $\psi(x,\theta) = -\frac{\partial}{\partial\theta}\log f(x,\theta)$.
- Least squares estimation is given by $\psi(x,\theta) = (x-\theta)^2$
- **Definition**: The minimum distance estimator for θ and distance function $\mathbf{d}$ is given by

$$\hat{\theta}_{MDE} = \arg\min_{\theta\in\Theta} \mathbf{d}(F(\mathbf{x},\theta), \hat{F}(\mathbf{x}))$$

- One popular choice of distance measures is given by the Kullback-Leibler Divergence

$$KL(f||g) = \int_{\mathcal{X}} g(x)\log\left(\frac{g(x)}{f(x)}\right)dx$$

- Maximizing the likelihood is equivalent to minimizing the KL divergence

4.1.5 Bayes Estimators & Minimax Estimators

- In the Bayesian framework, we assume that θ is a random variable with distribution $\pi(\theta)$.
- **Definition**: We say θ has prior distribution $\pi(\theta)$, $f(\mathbf{x}|\theta)$ is the conditional likelihood, with marginal distribution $f(\mathbf{x})$, and posterior distribution is written as $\pi(\theta|\mathbf{x})$.
- Through Bayes Theorem we have the relation

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{f(\mathbf{x})}$$

- **Definition**: Let $\mathcal{F}$ be a collection of parametric distributions and Π be a family of prior distributions. Then Π is a conjugate family for $\mathcal{F}$ iff $\pi(\mathbf{x}|\theta) \in \Pi$.
- **Definition**: Let ℓ be a loss function and $\hat{\theta}$ be a point estimator of θ. Then the classical risk is defined as

$$R(\hat{\theta},\theta) = \mathbb{E}[\ell(\hat{\theta},\theta)] = \int_{\mathcal{X}} \ell(\hat{\theta},\theta)f(\mathbf{x},\theta)d\mathbf{x}$$

- **Definition**: The Bayes Risk for an estimator δ, loss function ℓ, and prior π is given by

$$R(\pi,\delta) := \int_{\Theta} R(\delta,\theta)\pi(\theta)d\theta = \int_{\mathcal{X}} f(\mathbf{x})\left\{\int_{\Theta}\pi(\theta|\mathbf{x})\ell(\theta,\hat{\theta})d\theta\right\}d\mathbf{x}$$

- **Definition**: The Bayes Estimator δ_* is given by

$$\delta_* = \arg\min_{\delta} R(\pi,\delta)$$

- **Theorem**: Using quadratic loss, then the Bayes estimator is given by the posterior mean

$$\delta_* = E(\theta|\mathbf{x})$$

- **Definition**: A minimax estimator is one that satisfies

$$\hat{\delta}_{MM} := \min_{\delta}\max_{\theta\in\Theta} R(\delta,\theta)$$

- **Theorem**: Suppose there Bayes estimator δ_* such that $R(\theta, \delta_*)$ is free from θ. Then $\hat{\delta}_{MM} = \delta_*$.
- **Theorem**: Let $\{\delta_*^k\}_{k=1}^{\infty}$ be a sequence of Bayes estimators with Bayes risk $\{R(\pi_k, \delta_*^k)\}_{k=1}^{\infty}$. If

$$\lim_{n\to\infty} R(\pi_k, \delta_*^k) = r^* < \infty$$

and there exists δ such that $\sup_\theta R(\theta, \delta) \le r^*$ then δ is minimax.

- **Theorem**: (**Lehman**) If δ_* is an unbiased Bayes estimator then necessarily

$$\mathbb{E}\Big[(\delta_* - \theta)^2\Big] \equiv 0$$

4.2 Methods for Evaluating Estimators

- The best risk estimator is given by

$$\hat{\theta} := \arg\min_{\theta\in\Theta} R(\hat{\theta}, \theta)$$

- In general, this problem has no solution. So we reduce the problem into two subproblems (1) Reduce Θ to the class of unbiased estimators (2) Reduce some function of the risk
- We already solved (2) using Bayes & minimax. Here we focus on (1).

4.2.1 Fisher Efficiency

- If we work with quadratic loss, with $\hat{\theta}$ unbiased then

$$R(\theta, \hat{\theta}) = MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2 = Var(\hat{\theta})$$

so we simply want to minimize variance

- **Definition**: We can directly compare estimators by considering relative efficiency which is give by

$$eff(\hat{\theta}_1, \hat{\theta}_2) := \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$$

- **Definition**: $\hat{\theta}$ is a uniform minimum variance unbiased estimator (UMVUE) if $\hat{\theta}$ is unbiased and for any other estimator $\hat{\theta}'$ we have $Var(\hat{\theta}) \le Var(\hat{\theta}')$ for all $\theta \in \Theta$.
- **Definition**: The Fisher Information is given by

$$I_n(\theta) := \mathbb{E}\Big[\frac{\partial}{\partial\theta} \log f(\mathbf{x}, \theta)\Big]^2$$

- **Theorem**: (**Cramer-Rao**) Let $\hat{\theta}$ be a statistic. Under the following regularity conditions
  1. $\mathcal{X}$ does not depend on θ
  2. $\frac{\partial}{\partial\theta} f(\mathbf{x}, \theta)$ exists and is finite
  3. For $h(\mathbf{x})$ with $\mathbb{E}[h(\mathbf{x})] < \infty$ then $\frac{\partial}{\partial\theta} \int h(\mathbf{x}) f(\mathbf{x}, \theta) dx = \int h(\mathbf{x}) \frac{\partial}{\partial\theta} f(\mathbf{x}, \theta) dx$

  we have

$$Var(\hat{\theta}) \ge \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}[\hat{\theta}]\right)^2}{I_n(\theta)}$$

- Notice that if $\mathbb{E}(\hat{\theta}) = \theta$ then $Var(\hat{\theta}) \ge 1/I_n(\theta)$
- If $\mathbf{x}$ are iid then $I_n(\theta) = nI_1(\theta)$.
- **Lemma**: The fisher information can also be written as

$$I_n(\theta) = -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{x}, \theta)\Big]$$

- **Corollary**: If $\mathbf{x}$ are iid and $\hat{\theta}$ is unbiased then the CRLB is attained iff

$$a(\theta)[\hat{\theta} - \theta] = \frac{\partial}{\partial\theta} \log f(\mathbf{x}, \theta)$$

- **Definition**: The Fisher Efficiency of $\hat{\theta}$ is given by

$$eff(\hat{\theta}) = \frac{CRLB}{Var(\hat{\theta})}$$

  and we say a statistic is efficient iff $eff(\hat{\theta}) = 1$.

- With this, we see the UMVUE $\iff$ Unbiased + Fisher Efficient

4.2.2 Sufficiency Approaches

- Oftentimes the CRLB is not sufficient in evaluating estimators. First it is not defined for several models and simply gives a lower bound. Instead we turn to sufficiency based methods to find UMVUE's.
- **Theorem**: (**Rao-Blackwell**) Let W be an unbiased estimator of θ and let $T(\mathbf{x})$ be a sufficient statistic. Then $\phi(T) := \mathbb{E}[W|T]$ is a UMVUE for θ.
- "Unbiased conditioned on SS is UMVUE"
- **Theorem**: (**Lehman-Scheffe**) Let $T(\mathbf{x})$ be a complete sufficient statistic. Let $\phi(T)$ be a statistic relying only on $T(\mathbf{x})$. Then $\phi(T)$ is UMVUE for $\mathbb{E}[\phi(T)]$.
- If $\mathbb{E}[\phi(T)] = \theta$ then "unbiased function of CSS is UMUVE"
- **Theorem**: (**Necessary-Sufficient Conditions**) Let $\mathcal{U}$ be the class of unbiased estimators, $\mathcal{U}_0 \subseteq \mathcal{U}$ be the the class of unbiased estimators for zero, and $\mathcal{U}_0(T) \subseteq \mathcal{U}_0$ be the class of unbiased estimators of zero that can be written as $h(T)$. Then we have
  1. $W \in \mathcal{U}$ is UMUVE iff $Cov(W, X) = 0$ for all $X \in \mathcal{U}_0$
  2. $W = \phi(T)$ for sufficient statistic T is UMUVE iff $Cov(W, Y) = 0$ for all $Y \in \mathcal{U}_0(T)$

5 Asymptotic Evaluations

- While we have a notion of asymptotic evaluations for means and distributions, to compare estimators in this sense we wish to have some formal notion of asymptotic variance.
- **Definition**: For a sequence of estimators $\{T_n\}_{n=1}^{\infty}$, the asymptotic variance is given by

$$\sigma^2(\theta) := \lim_{n\to\infty} k_n Var(T_n) < \infty$$

- **Definition**: A sequence of estimators $\{T_n\}_{n=1}^{\infty}$, is called asymptotically normal with limiting variance $\sigma^2(\theta)$ iff
  1. $\lim_n nVar(T_n) = \sigma^2(\theta)$
  2. $\sqrt{n}(T_n - \theta) \xrightarrow{D} V \sim N(0, \sigma^2(\theta))$
- **Definition**: Let $T_2 \sim AN(\theta, \sigma_1^2(\theta)/n)$ and $T_2 \sim AN(\theta, \sigma_2^2(\theta)/n)$. Then the asymptotic relative efficiency is given by

$$ARE(T_2, T_2) := \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}$$

- **Definition**: An estimator T is called asymptotically efficient iff $T \sim AN(\theta, \sigma^2(\theta)/n)$ where $\sigma^2(\theta) = 1/I_1(\theta)$
- The Fisher program was an attempt to show that MLE estimates are also asymptotically efficient. This would show that in a sense MLE are the best estimators under Fisher's framework. Unfontunately this is not the case in general.
- **Theorem**: Under the following regularity conditions, MLE's are asympotitically efficient.

1. Indentifiability
2. All estimators in the sequence have a common support
3. Differentiable density with respect to θ
4. Θ contains an open set
5. $f(\mathbf{x}, \theta)$ is three times differentiable
6. $|\partial^3/\partial\theta^3 \log f(x, \theta)| \leq M(x)$ with $\mathbb{E}|M(x)| < \infty$

- Under 1 - 4 MLE is consistent. Under 1 - 6 MLE is asymptotically efficient.
- **Theorem**: If $\{\hat{\theta}_k\}_{k=1}^{\infty}$ is asymptotically normal then $\{\hat{\theta}_k\}_{k=1}^{\infty}$ is consistent.
- **Definition**: If there exists a statistic M for θ such that $M \sim AN(\mu, \sigma^2(\theta))$ (note $\mu \neq \theta$) and we have $\sigma^2(\theta) \leq CRLB$ and there exists θ' such that $\sigma^2(\theta') < CRLB$ then we say M is super efficient.