Stats 511 Notes

1.1 Fisher Information

Remark 1.1.1. Notes on Notation
- Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ.

- The probability function for a random variable X is a function

$$f(X;\theta)$$

- It is the probability mass (or probability density) of the random variable X conditional on the value of θ. The likelihood function for a parameter θ is a function

$$L(x;\theta)$$

- It is the likelihood of the parameter θ given an outcome x.
- When observing $X = x$,

$$f(X = x;\theta) = L(x;\theta)$$

- The expression $\mathbb{E}[...|\theta]$ denotes the conditional expectation over the values for X with respect to the probability function $f(X;\theta)$ given θ.

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} log f(X;\theta)\right] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} log f(X;\theta)\right)|\theta\right]$$

Motivation:
- Intuitively, if $\mathbb{P}(A)$ is small, then the occurrence of this event brings us much information.
- For a random variable $X \sim f(X;\theta)$, if θ_0 were the true value of the parameter, the likelihood function should be really large, that is, $L(x;\theta_0)$ is large, or
- equivalently, the derivative of log-likelihood function should be close to zero, and this is the basic principle of MLE.
- Score function is the derivative of log-likelihood function.
- If score function $\ell'(X|\theta)$ is close to zero, then MLE is close to θ_0, meaning that the random variable X does not provide much information about θ.
- If score function $|\ell'(X|\theta)|$ or $[\ell'(X|\theta]^2$ is large, then MLE is far away from θ, meaning that the random variable provides much information about θ.
- Thus, we can use $[\ell'(X|\theta]^2$ to measure the amount of information provided by X.
- Since X is random, we consider the average amount of information provided by X about the parameter θ.

Lemma 1.1.2. ***Fisher Information***

If $f(x;\theta)$ satisfies (the regularity condition of interchange of derivative and integral)

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta}logf(X;\theta)\Big] = \frac{\mathrm{d}}{\mathrm{d}\theta}\int\Big(\frac{\partial}{\partial\theta}logf(x;\theta)\Big)f(x;\theta)\mathrm{d}x = \int\frac{\partial}{\partial\theta}\Big(\frac{\partial}{\partial\theta}logf(x;\theta)f(x;\theta)\Big)\mathrm{d}x$$

(true for an exponential family), and $logf(X;\theta)$ is twice differentiable with respect to θ, then

$$\mathcal{I}(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}logf(X;\theta)\Big)^2\Big] = -\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta^2}logf(X;\theta)\Big]$$

1.2 Asymptotics: Point Estimation

1.2.1 Regularity Conditions (CB Ch10 Misc)

1 - 4 are sufficient to prove consistency of MLEs.

1. We observe $X_1, ...X_n$, where $X_i \sim f(x|\theta)$ are iid.
2. The parameter is *identifiable*; that is, if $\theta \neq \theta'$, then $f(x|\theta) \neq f(x|\theta')$.
3. The densities $f(x|\theta)$ have common support, and $f(x|\theta)$ is differentiable in θ.
4. The parameter space Ω contains an open set ω of which the true parameter value θ_0 is an interior point.

1 - 6 are sufficient to prove asymptotic normality and efficiency of MLEs.

5. For every $x \in \mathcal{X}$, the density $f(x|\theta)$ is three times differentiable with respect to θ, the third derivative is continuous in θ and $\int f(x|\theta)\mathrm{d}x$ can be differentiated three times under the integral sign. That is, interchange of integration and differentiation is allowed.
6. For any $\theta_0 \in \Omega$, there exists a positive number c and a function $M(x)$ (both of which depend on θ_0) such that
$$\frac{\partial^3}{\partial\theta^3}logf(x|\theta) \leq M(x) \quad \text{for all } x \in \mathcal{X}, \quad \theta_0 - c \leq \theta \leq \theta_0 + c$$
with $\mathbb{E}_{\theta_0} < \infty$

1.2.2 UMVUE and Cramer-Rao (CB 7.3.2)

Definition 1.2.1. UMVUE
An estimator W is a *best unbiased estimator/UMVUE* of $\tau(\theta)$ if

(i) (unbiasedness of W)
$\mathbb{E}_\theta W = \tau(\theta)$ for all θ

(ii) (unbiasedness of an arbitrary W)
$\mathbb{E}_\theta W' = \tau(\theta)$ for all θ

Then, (W is the best) $\mathrm{Var}_\theta W \leq \mathrm{Var}_\theta W'$

Theorem 1.2.2. ***Cramer-Rao Inequality***
Under the regularity condition 7 that allows interchange of integration and differentiation.
Let $X_1, ..., X_n$ be a sample with pdf $f(x;\theta)$.
*Let $W(X^n)$ be **any** estimator satisfying*

(i)

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta W(X^n) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[W(x^n)f(x^n;\theta)]\mathrm{d}x^n$$

(ii) $\mathrm{Var}_\theta W(X^n) < \infty$

Then,

$$\mathrm{Var}_\theta W(X^n) \geqslant \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta W(X^n)\right)^2}{\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta} log f(X^n;\theta)\right)^2\right)}$$

Corollary 1.2.3. ***Cramer-Rao Inequality, iid case***
If the assumptions of Theorem 1.2.2 are satisfied and, additionally, if $X_1, ..., X_n$ are iid with pdf $f(x;\theta)$, then

$$\begin{aligned}\mathrm{Var}_\theta W(X^n) &\geqslant \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta W(X^n)\right)^2}{n\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta} log f(X;\theta)\right)^2\right)} \\ &= \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta W(X^n)\right)^2}{n\mathcal{I}(\theta)} = \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta W(X^n)\right)^2}{\mathcal{I}_n(\theta)}\end{aligned}$$

Remark 1.2.4. Cramer-Rao Inequality, iid case + unbiased estimator
If the assumptions of Theorem 1.2.2 are satisfied and, additionally, if $X_1, ..., X_n$ are iid with pdf $f(x;\theta)$, and $\mathbb{E}_\theta W = \tau(\theta)$ for all θ, then

$$\begin{aligned}\mathrm{Var}_\theta W(X^n) &\geqslant \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta W(X^n)\right)^2}{n\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta} log f(X;\theta)\right)^2\right)} \\ &= \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\tau(\theta)\right)^2}{\mathcal{I}_n(\theta)} = \frac{(\tau'(\theta))^2}{\mathcal{I}_n(\theta)}\end{aligned}$$

Corollary 1.2.5. ***Cramer-Rao Attainment***
The CRLB is achieved by distributions of the exponential family.

1.2.3 Asymptotic Approximation for Large Samples

Theorem 1.2.6. ***Asymptotic Normality of MLE***
Let $X_1, ..., X_n$ be iid $f(x|\theta)$.
Let $\hat{\theta}_n$ denote the MLE for θ.

Let $\tau(\theta)$ be a continuous function of θ.
Under regularity conditions on $f(x|\theta)$ and hence $L(\theta|x^n)$,

$$\sqrt{n}\Big(\tau(\hat{\theta}_n) - \tau(\theta)\Big) \to N\Big(0, \upsilon(\theta)\Big), \tag{1.2.1}$$

where $\upsilon(\theta) = \dfrac{(\tau'(\theta))^2}{\mathcal{I}_n(\theta)}$ is the Cramer-Rao Lower Bound (iid case).
That is, $\tau(\hat{\theta}_n)$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$

Definition 1.2.7. Asymptotically efficient A sequence of estimators $W_n = W_n(X_1, ..., X_n)$ is *asymptotically efficient* for a parameter $\tau(\theta)$ if

$$\sqrt{n}\Big(W_n - \tau(\theta)\Big) \to N\Big(0, \upsilon(\theta)\Big) \text{ in distribution}$$

and it just happens that

$$\upsilon(\theta) = \frac{\Big(\tau'(\theta)\Big)^2}{\mathbb{E}_\theta\Big(\Big(\frac{\partial}{\partial\theta} log f(X^n|\theta)\Big)^2\Big)}$$

That is, the asymptotic variance of W_n achieves the Cramer-Rao Lower Bound.
Comments:
- Calculate the asymptotic variance of W_n by Delta method, and obtain $\upsilon(\theta) = Var(X_i)\Big(\tau'(\theta)\Big)^2$
- Calculate the CRLB of $\mathrm{Var}(W_n|\theta)$, and obtain $\mathrm{Var}(W_n) \geq CRLB = \dfrac{\Big(\tau'(\theta)\Big)^2}{\mathbb{E}_\theta\Big(\Big(\frac{\partial}{\partial\theta} log f(X^n|\theta)\Big)^2\Big)}$
- Compare Delta method $\upsilon(\theta)$ with CRLB. - If asymptotic variance obtained using Delta method $\upsilon(\theta)$ is the same as CRLB, then W_n is asymptotically efficient.

Theorem 1.2.8. ***Delta Method (A generalization of CLT)***
Let Y_n be a sequence of random variables ($\mathbb{E}(Y_i) = \theta$ and $\mathrm{Var}(Y_i) = \sigma^2$) that satisfies

$$\sqrt{n}(Y_n - \theta)) \to N\Big(0, \sigma^2\Big) \quad \textit{in distribution}$$

For a given function τ and a specific value of θ, suppose that τ' exists and is not 0. Then,

$$\sqrt{n}\Big(\tau(Y_n) - \tau(\theta)\Big) \to N\Big(0, \sigma^2\Big(\tau'(\theta)\Big)^2\Big)$$

Comments:
True variance of Y_n is $\mathrm{Var}(Y_n)$
Limiting variance of Y_n is $\lim_{n\to\infty} \sqrt{n}\mathrm{Var}(Y_n)$
Asymptotic variance of Y_n is σ^2

True variance of $\tau(Y_n)$ is $Var(\tau(Y_n))$
Limiting variance of $\tau(Y_n)$ is $\lim_{n\to\infty} \sqrt{n}\mathrm{Var}(\tau(Y_n))$
Asymptotic variance of $\tau(Y_n)$ is $\sigma^2\Big(\tau'(\theta)\Big)^2\Big)$

1.2.4 Approximate true variances of MLEs through asymptotic formulas

- $\mathrm{Var}(\hat{\theta}_n)$ is the true variance of MLE.

- Under regularity conditions, the theorems for asymptotic distribution of MLEs can be used to approximate the true variances of MLEs, $\mathrm{Var}(\hat{\theta}_n)$, for large samples, as $n \to \infty$.

- If an MLE, $\hat{\theta}_n$ is asymptotically efficient, then

(i) asymptotic variance $\upsilon(\theta)$ obtained from Delta method achives CRLB.

(ii) then CRLB can be approximated by evaluating at $\theta = \hat{\theta}_n$

(iii) then true variance can be approximated by the approximated CRLB

Remark 1.2.9. Method of Approximation of $\mathrm{Var}(\tau(\hat{\theta}_n))$

$$\begin{aligned}\mathrm{Var}(\tau(\hat{\theta}_n)) \approx CRLB = \frac{\Big(\tau(\theta)\Big)^2}{\mathcal{I}(\theta))} &= \frac{\Big(\tau'(\theta)\Big)^2}{\mathbb{E}_\theta\Big(-\dfrac{\partial^2}{\partial\theta^2} logf(X^n|\theta)\Big)} \\ \approx \frac{\Big(\tau(\theta)\Big)^2}{\mathcal{I}(\theta))} &= \frac{\Big(\tau'(\theta)\Big)^2}{-\dfrac{\partial^2}{\partial\theta^2} logf(X^n|\theta)|_{\theta=\hat{\theta}}} = \hat{\mathrm{Var}}(\tau(\hat{\theta}_n))\end{aligned}$$

Example 1.2.10. (Violation of regularity conditions – scale uniform $\mathrm{Uniform}(0,\theta)$)

In general, if the range of the pdf depends on the parameter, the Cramer-Rao Theorem does not apply (due to the inability to differentiate under the integral sign)

1.3 UMP

1.4 Interval Estimation

1.4.1 Inverting a Test Statistic

Example 1.4.1. Inverting a normal test
Let $X_1, ..., X_n$ be iid $N(\mu, \sigma^2)$.
Consider testing $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

1.5 Asymptotic Interval Estimation