Supplement 1:
Some Basic Statistical Theory

Suppose $X$ is a random variable with density $f(x)$ (or $f(x; \theta)$ if the distribution of $X$ depends on a parameter $\theta$).

Mean of $X$:

$$\mu_X = E\{X\} = \int xf(x)dx.$$  \hspace{1cm} (1)

Variance of $X$:

$$\sigma^2_X = E\{(X - \mu_X)^2\}.$$  \hspace{1cm} (2)

If $X$ and $Y$ are two random variables, the covariance is

$$\sigma_{XY} = E\{(X - \mu_X)(Y - \mu_Y)\}.$$  \hspace{1cm} (3)
For a vector $X$, say $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{pmatrix}$, the mean is

$$
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}, \text{ where } \mu_i = E\{X_i\}.
$$

The covariance matrix is $\Sigma = (\sigma_{ij})$, where

$$
\sigma_{ij} = E\{(X_i - \mu_i)(X_j - \mu_j)\}.
$$

The vector $X$ of dimension $d$ is said to have a multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$ \{X $\sim N_d[\mu, \Sigma]$\} if its density is

$$
f(x; \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \cdot \exp\left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}.
$$
Suppose $X_1, X_2, ...$ are independent random vectors of dimension $d$ with mean $\mu$ and covariance matrix $\Sigma$. Let $S_n = X_1 + ... + X_n$.

**Law of Large Numbers.** If $\mu$ is finite,

$$\frac{S_n}{n} \xrightarrow{p} \mu \text{ as } n \to \infty. \quad (6)$$

**Technical aside:** The mode of convergence here is convergence in probability, which means that for any $\epsilon > 0$,

$$\Pr \left\{ \left| \frac{S_n}{n} - \mu \right| > \epsilon \right\} \to 0. \quad (7)$$

There is a stronger form of convergence, called convergence almost surely or convergence with probability one, but we shall not need this.

**Central Limit Theorem.** If $\Sigma$ is finite,

$$\frac{1}{\sqrt{n}} (S_n - n\mu) \xrightarrow{d} \mathcal{N}_d[0, \Sigma]. \quad (8)$$

Convergence in distribution.
Estimation of Parameters

General setting: Independent observations $X_1, ..., X_n$ from some density $f(\cdot; \theta)$.

Choose an estimator $\hat{\theta}_n$ to minimize some function $Q_n(\theta)$. Assume it is consistent, that is, $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$, where $\theta_0$ is the true value of $\theta$.

Example: If $Q_n(\theta) = -\sum_i \log f(X_i; \theta)$, then $Q_n$ is called the negative log likelihood (NLLH) function and $\hat{\theta}_n$ is called the maximum likelihood estimator (MLE).
Suppose, as $n \to \infty$,
\[ \frac{1}{n} \nabla^2 Q_n(\theta) \xrightarrow{p} H(\theta) \] (9)
uniformly over $\theta$ in some neighborhood of the true value $\theta_0$, and
\[ \frac{1}{\sqrt{n}} \nabla Q_n(\theta_0) \xrightarrow{d} \mathcal{N}_d[0, V(\theta_0)]. \] (10)

By a Taylor expansion,
\[ \hat{\theta}_n = \theta_0 - \left\{ \nabla^2 Q_n(\theta^*_n) \right\}^{-1} \nabla Q_n(\theta_0) \] (11)
for some $\theta^*_n$ between $\theta$ and $\hat{\theta}_n$.

Then
\[ \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}_d[0, H(\theta_0)^{-1}V(\theta_0)H(\theta_0)^{-1}]. \] (12)

This is the information sandwich formula.
For the MLE, two nice things happen:

- $H(\theta) = V(\theta)$. (*Proof:* Calculus.)

- For any unbiased estimator $\tilde{\theta}_n$, the variance of $\tilde{\theta}_n$ is at least $\frac{V(\theta)}{n}$ (Cramér-Rao lower bound).

Therefore, the asymptotic variance of $\hat{\theta}_n$ is $\frac{V(\theta)}{n}$ and the MLE is asymptotically efficient.

$V$ (or $H$) is called the *Fisher information matrix*.

In practice, $\hat{\theta}_n$ is found by a Newton or quasi-Newton optimization and $H$ is estimated from the (exact or approximate) Hessian matrix of $Q_n$ at $\theta = \hat{\theta}_n$ — the *observed information matrix*. 
Likelihood Ratio Tests

Suppose $M_0$ and $M_1$ are two possible models, which are nested, that is, $M_0 \subset M_1$. We want to test whether $M_0$ is the right model.

For $j = 0, 1$, let $\ell_j$ be the NLLH under $M_j$.

Let

$$T_n = 2(\ell_0 - \ell_1) \tag{13}$$

Then in large samples ($n \to \infty$),

$$T_n \overset{d}{\to} \chi_q^2, \tag{14}$$

where $q$ is the difference in the number of parameters between $M_0$ and $M_1$.

The large-sample results for the MLE and likelihood ratio test depend on so-called regularity conditions which we shall usually ignore.