REVIEW

Likelihood Inference

Likelihood functions

- Setting: Let $Y_1, \ldots, Y_n$ be independent random variables, with $Y_i$ having probability (or density) function

$$f(y_i; \beta),$$

where $\beta$ is some unknown parameter.

- For example, in the Bernoulli distribution, all the $Y_i$’s are i.i.d. with distribution depending on the parameter $\beta = p$:

$$Y_i \sim \text{Bernoulli}(p)$$

i.e.,

$$f(y_i; p) = p^{y_i}(1 - p)^{1-y_i},$$

- In general, for $n$ independent random variables, the joint probability function of the data is the product of the individual probability distributions:

$$f(y_1, \ldots, y_n; \beta) = \prod_{i=1}^{n} f(y_i; \beta)$$
• The **likelihood function** of $\beta$ is equivalent to the probability function of the data:

$$L(\beta) = L(\beta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i; \beta).$$

The idea is to find the $\beta$ value that maximizes this likelihood (probability of observing such data). This is the $\beta$ value most 'coherent' with the data.

• Once you take the random sample of size $n$, the $Y_i$'s are known, but $\beta$ is not – in fact, the only unknown in the likelihood is the parameter $\beta$.

• **Example:** The likelihood function of $p$ for a sample of $n$ Bernoulli r.v.'s is:

$$L(p) = \prod_{i=1}^{n} p^{y_i}(1 - p)^{(1-y_i)} = p^{\sum_{i=1}^{n} y_i}(1 - p)^{n - \sum_{i=1}^{n} y_i}.$$
• **Maximum Likelihood Estimator (MLE)** of \( \beta \) is the value, \( \hat{\beta} \), which maximizes the likelihood

\[
L(\beta)
\]

or the **log-likelihood**

\[
\log L(\beta)
\]

as a function of \( \beta \), given the observed \( Y_i \)'s.

• The value \( \hat{\beta} \) that maximizes \( L(\beta) \) also maximizes \( \log L(\beta) \), since the latter is a monotone function of \( L(\beta) \).

• It is usually easier to maximize \( \log L(\beta) \), (why?) so we focus on the log-likelihood.

• Most of the estimates we will discuss in this class will be MLE’s. This is because they have optimal properties:
  
  – consistent: as \( n \to \infty \), \( \hat{\beta} \to \beta \) in probability
  
  – efficient: achieves minimum variance
• For most distributions, the maximum is found by solving
\[ \frac{\partial \log L(\beta)}{\partial \beta} = 0 \]

• Technically, we need to verify that we are at a maximum (rather than a minimum) by seeing if the second derivative is negative at \( \hat{\beta} \), i.e.,
\[ \left[ \frac{\partial^2 \log L(\beta)}{\partial \beta^2} \right]_{\beta=\hat{\beta}} < 0 \]

• The opposite of the second derivative,
\[ -\frac{\partial^2 \log L(\beta)}{\partial \beta^2}, \]

is called the **Fisher information**. It plays an important role in the likelihood theory.
Example: Bernoulli (Binomial) data

- The likelihood is

\[
L(p) = \prod_{i=1}^{n} p^{y_i} (1 - p)^{1-y_i}
\]

\[= p^y (1 - p)^{n-y},\]

where

\[y = \sum_{i=1}^{n} y_i = \text{number of successes}\]

- The log-likelihood is

\[
\log L(p) = y \log p + (n - y) \log(1 - p),
\]

- The first derivative is

\[
\frac{\partial \log L(p)}{\partial p} = \frac{y}{p} - \frac{n - y}{1 - p} = \frac{y - np}{p(1 - p)}
\]

Setting this to 0 and solving for \(\hat{p}\), you get

\[\hat{p} = \frac{y}{n},\]

i.e. proportion of successes.
• The second derivative of the log-likelihood is

\[
\frac{\partial^2 \log L(p)}{\partial p^2} = \frac{-y}{p^2} - \frac{(n-y)}{(1-p)^2}
\]

• Evaluating at \( p = \hat{p} \):

\[
\left( \frac{\partial^2 \log L(p)}{\partial p^2} \right)_{p=\hat{p}} = -\frac{y}{(y/n)^2} - \frac{(n-y)}{(1-(y/n))^2}
\]

\[
= -\frac{n^2}{y} - \frac{n^2}{(n-y)} < 0
\]

• When \( 0 < y < n \), the 2nd derivative at \( \hat{p} \) is negative, so \( \hat{p} \) is the maximum.

• When \( y = 0 \) or \( y = n \), the estimate \( \hat{p} = 0 \) or \( \hat{p} = 1 \) is said to be on the ‘boundary’.
Variance of the MLE

The asymptotic variance of the MLE \( \hat{\beta} \) is

\[
Var(\hat{\beta}) = - \left\{ E \left( \frac{\partial^2 \log L(\beta)}{\partial \beta^2} \right) \right\}^{-1}.
\]

It is often estimated by the inverse of the observed information

\[
\left\{ - \frac{\partial^2 \log L(\beta)}{\partial \beta^2} \right\}^{-1}
\left| _{\beta = \hat{\beta}} \right.
\]

In addition, MLE’s are asymptotically normally distributed, i.e.

\[
\hat{\beta} \sim N[\beta, Var(\hat{\beta})],
\]

Example: Bernoulli (Binomial) data

- \( Var(\hat{p}) \) is estimated by

\[
\left\{ - \frac{\partial^2 \log L(p)}{\partial p^2} \right\}^{-1} \bigg| _{p = \hat{p}} = \left\{ \frac{n^2}{y} + \frac{n^2}{n - y} \right\}^{-1} = \frac{y(n - y)}{n^3} = \frac{\hat{p}(1 - \hat{p})}{n}
\]

- Note that

\[
Var(\hat{p}) = \frac{p(1 - p)}{n}.
\]

(why?)
Test Statistics Associated with the Likelihood
(see Section 12.4 of Lehmann and Romano book ‘Testing Statistical Hypotheses’)

A. Wald Test

- Suppose we want to test $H_0 : \beta = \beta^*$. Let $\hat{\beta}$ be the MLE.

- The following Wald test statistics can be used:

$$Z = \frac{\hat{\beta} - \beta^*}{\sqrt{\text{Var}(\hat{\beta})}} \approx N(0, 1)$$

under $H_0$.

- Since the square of a $N(0, 1)$ r.v. follows a $\chi^2_1$ distribution, we can also use the test statistics $Z^2$.

- The advantage of the chi-squared form is that it can be extended to higher dimensions:

$$\frac{(\hat{\beta} - \beta^*)'\text{Var}(\hat{\beta})^{-1}(\hat{\beta} - \beta^*)}{\text{approx.}} \sim \chi^2_p$$

under $H_0$, where $p$ is the dimension of $\beta$. 
B. Likelihood Ratio Test

In large samples, under the null hypothesis $H_0 : \beta = \beta^*$, it can be shown that:

$$2 \log \left\{ \frac{L(\hat{\beta})}{L(\beta^*)} \right\} = 2[\log L(\hat{\beta}) - \log L(\beta^*)] \approx \chi^2_p$$

under $H_0$, where $\hat{\beta}$ is the MLE of $\beta$. 
C. Score Test

- The first derivative of the log-likelihood is often referred to as the **score function**, and is denoted by

\[
U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial \log L_i(\beta)}{\partial \beta}
\]

where \( L_i(\beta) \) is the likelihood from the \( i \)-th observation.

- Recall that the MLE, \( \hat{\beta} \), is obtained by setting the score \( U(\beta) = 0 \).

- Since the score can also be written as a sum of i.i.d. observations, we can apply the Central Limit Theorem to show that it is approximately normal:

\[
U(\beta^*) \overset{\text{approx.}}{\sim} N(E[U(\beta^*)], \text{Var}[U(\beta^*)])
\]

where \( \beta^* \) is the true value of \( \beta \).

- It turns out that \( E[U(\beta^*)] = 0 \) under \( H_0 : \beta = \beta^* \). So

\[
U(\beta^*) \overset{\text{approx.}}{\sim} N(0, \text{Var}[U(\beta^*)])
\]

- Note also \( \text{Var}[U(\beta^*)] = I(\beta^*) \) the Fisher information (why?).
• In general, the **score test** statistic for testing $H_0 : \beta = \beta^*$ is:

$$U(\beta^*)'\text{Var}[U(\beta^*)]^{-1}U(\beta^*) \approx \chi_p^2$$

• Note that we don’t need to estimate $\beta$ here, so score test can be the simplest to compute among the three tests.