REVIEW

Likelihood Inference

Likelihood functions

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

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

where β 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:

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

- The **likelihood function** of β is equivalent to the probability function of the data:

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

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

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

- **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 β is the value, $\widehat{\beta}$, which maximizes the likelihood

$$L(\beta)$$

or the **log-likelihood**

$$\log L(\beta)$$

as a function of β, given the observed Y_i's.

- The value $\widehat{\beta}$ 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 $\widehat{\beta}$, i.e.,

$$\left[\frac{\partial^2 \log L(\beta)}{\partial \beta^2}\right]_{\beta=\widehat{\beta}} < 0$$

- 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

$$\begin{aligned} L(p) &= \Pi_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \\ &= p^{y}(1-p)^{n-y}, \end{aligned}$$

  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 $p = \widehat{p}$:

$$\begin{aligned}
\left(\frac{\partial^2 \log L(p)}{\partial p^2}\right)_{p=\widehat{p}} &= -\frac{y}{(y/n)^2} - \frac{(n-y)}{(1-(y/n))^2} \\
&= -\frac{n^2}{y} - \frac{n^2}{(n-y)} \quad < \quad 0
\end{aligned}$$

- When $0 < y < n$, the 2nd derivative at $\widehat{p}$ is negative, so $\widehat{p}$ is the maximum.

- When $y = 0$ or $y = n$, the estimate $\widehat{p} = 0$ or $\widehat{p} = 1$ is said to be on the 'boundary'.

Variance of the MLE

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

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

It is often estimated by the inverse of the **observed information**

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

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

$$\widehat{\beta} \,\dot{\sim}\, N[\beta, Var(\widehat{\beta})],$$

Example: Bernoulli (Binomial) data

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

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

- 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 $\mathrm{H}_0 : \beta = \beta^*$. Let $\hat{\beta}$ be the MLE.

- The following **Wald test** statistics can be used:

$$Z = \frac{\widehat{\beta} - \beta^*}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}} \overset{approx.}{\sim} N(0,1)$$

  under H_0.

- Since the square of a $N(0,1)$ r.v. follows a χ^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:

$$(\widehat{\beta} - \beta^*)' \widehat{\mathrm{Var}}(\widehat{\beta})^{-1} (\widehat{\beta} - \beta^*) \overset{approx.}{\sim} \chi^2_p$$

  under H_0, where p is the dimension of β.

B. Likelihood Ratio Test

In large samples, under the null hypothesis $\mathrm{H}_0 : \beta = \beta^*$, it can be shown that:

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

under H_0, where $\widehat{\beta}$ is the MLE of β.

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:

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

where β^* is the true value of β.

- It turns out that $E[U(\beta^*)] = 0$ under $\mathrm{H}_0 : \beta = \beta^*$. So

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

- Note also $\mathrm{Var}[U(\beta^*)] = I(\beta^*)$ the Fisher information (why?).

- In general, the **score test** statistic for testing $\mathrm{H}_0 : \beta = \beta^*$ is:

$$U(\beta^*)'\mathrm{Var}[U(\beta^*)]^{-1}U(\beta^*) \overset{approx.}{\sim} \chi^2_p$$

- Note that we don't need to estimate β here, so score test can be the simplest to compute among the three tests.