## Likelihood functions

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

$$Y_i \sim Bernoulli(p)$$

  i.e.,

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

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

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

- The **Likelihood function** of $\beta$ given the data are equivalent to the probability function of the data given $\beta$:

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

- 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 $\beta$ -

  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.

- 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, $\frac{-\,\partial^2 \log L(\beta)}{d\beta^2}$, is called the "**information**". This quantity plays an important part in likelihood theory.

## Example: Bernoulli data

- The likelihood is

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

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 $\widehat{p}$, you get

$$\widehat{p} = \frac{y}{n}$$

- 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’.

## Properties of MLE's

Any two likelihoods, $L_1(\beta)$ and $L_2(\beta)$, that are proportional, i.e.,

$$L_2(\beta) = \alpha\, L_1(\beta)$$

(where $\alpha$ is a constant that does not depend on $\beta$)

yield the same **maximum likelihood estimator**

- **Example:** If we had started with the Binomial distribution of $Y = \sum_{i=1}^{n} Y_i$ rather than $n$ independent Bernoulli r.v.'s, then the likelihood would be:

$$f(y) = \left( \begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y},$$

- For this distribution, the log-likelihood is

$$\log L(p) = \log \left( \begin{array}{c} n \\ y \end{array} \right) + y \log p + (n-y) \log(1-p),$$

- The first derivative of the binomial log-likelihood is

$$\begin{aligned}
&\frac{\partial \log L(p)}{\partial p} \\
&= \quad \frac{\partial}{\partial p}\left[\log\begin{pmatrix} n \\ y \end{pmatrix}\right] + \frac{\partial[y\log p + (n-y)\log(1-p)]}{\partial p} \\
&= \quad 0 + \frac{\partial[y\log p + (n-y)\log(1-p)]}{\partial p} \\
&= \quad \frac{\partial[y\log p + (n-y)\log(1-p)]}{\partial p}
\end{aligned}$$

  This is exactly the same as the first derivative of the log-likelihood for $n$ independent Bernoulli's.

- Therefore, we get the same MLE for independent Bernoulli data and Binomial data (both are based on $Y = \sum_{i=1}^{n} Y_i$ )

- The likelihood, $L_2(p)$ of the Binomial data is proportional to the likelihood, $L_1(p)$ based on the original Bernoulli data, since

$$
\begin{aligned}
L_2(p) &= \alpha L_1(p) \\
\text{where} \qquad \alpha &= \binom{n}{y}
\end{aligned}
$$

## Asymptotic Properties of MLE's

- The exact distribution of the MLE can be very complicated, so we often have to rely on large sample methods instead.
- Using a Taylor series expansion and the Delta Method, the following properties can be shown as $n \rightarrow \infty$ :

(1) $\hat{\beta}$ is **asymptotically unbiased**$^*$:

$$E(\hat{\beta}) \rightarrow \beta$$

(2) $\hat{\beta}$ is **consistent**:

$$\text{pr}\{|\hat{\beta} - \beta| > \epsilon\} \rightarrow 0,$$

(3) $\hat{\beta}$ is **asymptotically efficient**

(it achieves the minimum variance among all asymptotically unbiased estimators)

* - (Note that it may be biased in small samples)

- In addition, using the Central Limit Theorem, it can be shown that **MLE's are asymptotically normally distributed**, i.e,

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

where $Var(\hat{\beta})$ is the inverse of the expected value of the information:

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

(Note, however, that $Var(\hat{\beta})$ is itself a function of $\beta$. We will come back to examine this issue later)

## Example: Bernoulli data (continued)

- From the above MLE theory, we know that, for large $n$,

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

  where

$$Var(\widehat{p}) = -\left\{ E\left( \frac{d^2 \log L(p)}{dp^2} \right) \right\}^{-1},$$

- Recall, the second derivative of the log-likelihood is

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

  so that the 'information' equals

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

- The expected value of the information is

$$\begin{aligned} E\left(-\frac{\partial^2 \log L(p)}{\partial p^2}\right) &= \frac{E(Y)}{p^2} + \frac{E(n-Y)}{(1-p)^2} \\ &= \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} \\ &= \frac{n}{p} + \frac{n}{(1-p)} \\ &= \frac{n}{p(1-p)} \end{aligned}$$

- To get the asymptotic variance, we now take the inverse:

$$Var(\widehat{p}) = \left(\frac{n}{p(1-p)}\right)^{-1} = \frac{p(1-p)}{n}$$

- This confirms what we already derived using the CLT:

$$\widehat{p} \overset{\cdot}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

## MLE's of functions

The MLE of a function is the function of the MLE, i.e.,

The MLE of $g(\beta)$ is $g(\hat{\beta})$

**Variance of $g(\hat{\beta})$:**

Two possible methods for calculating the variance are:

(1) Apply the **Delta Method** to $g(\hat{\beta})$ According to the Delta method the variance of the function $g(\hat{\beta})$ is

$$Var[g(\hat{\beta})] = [g'(\beta)]^2 Var(\hat{\beta})$$

(2) **Rewrite the likelihood** in terms of $\theta = g(\beta)$, then take the second derivative of the corresponding log-likelihood with respect to $\theta$.

## Example: Binomial data

- The MLE of $p$ is $\widehat{p} = \frac{Y}{n}$. What is the MLE of logit$(p)$?

- Using the above result, the MLE of logit$(p)$ is logit$(\frac{Y}{n})$

- Calculating $Var(\text{logit}(\widehat{p}))$:
  - **Method (1):** already shown

  - **Method (2):**

    Let $\theta = \text{logit}(p) = \log\left(p/(1-p)\right)$. After some algebra, you can show that

$$p = \frac{e^{\theta}}{1 + e^{\theta}}$$

Substitute $e^\theta/(1+e^\theta)$ for $p$ in the likelihood:

$$\begin{aligned} f(y) &= \left(\begin{array}{c} n \\ y \end{array}\right) p^y (1-p)^{n-y} \\ &= \left(\begin{array}{c} n \\ y \end{array}\right) \left(\tfrac{e^\theta}{1+e^\theta}\right)^y \left(1 - \tfrac{e^\theta}{1+e^\theta}\right)^{n-y} \end{aligned}$$

Then take 2nd derivatives of the log-likelihood with respect to $\theta$ to find the information, and compute the inverse.

## Confidence Intervals and Hypothesis Testing

### I. Confidence Intervals

- From MLE theory, we know that for large $n$,

$$\widehat{p} \;\dot{\sim}\; N\left(p, \frac{p(1-p)}{n}\right)$$

- A 95% confidence interval for $p$ can thus be constructed as:

$$\widehat{p} \;\pm\; 1.96\sqrt{\frac{p(1-p)}{n}}$$

However, we do not know $p$ in the variance.

- Since $\widehat{p}$ is a consistent estimate of $p$, we can replace

  $$p(1-p)$$

  in the variance by

  $$\widehat{p}(1-\widehat{p}),$$

  and still get 95% coverage (in large samples).

- Therefore, a large sample confidence interval for $p$ is:

  $$\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

## Large Sample Confidence Interval for $\beta$

- In general, suppose we want a 95% confidence interval for $\beta$. For large $n$, we know that

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

where $Var(\hat{\beta})$ is the inverse of the expected value of the information, **and is a function of** $\beta$.

- If we knew $\beta$ in $Var(\hat{\beta})$, we could form an asymptotic 95% confidence interval for $\beta$ with

$$\hat{\beta} \pm 1.96\sqrt{Var(\hat{\beta})}.$$

(But.... if we knew $\beta$, we wouldn't need a confidence interval in the first place!)

- Since we do not know $\beta$, we have to estimate $Var(\hat{\beta})$ by replacing $\beta$ with its consistent estimator $\hat{\beta}$ :

$$\widehat{Var}(\hat{\beta}) = [Var(\hat{\beta})]_{\beta=\hat{\beta}}$$

- Then the confidence interval

$$\hat{\beta} \pm 1.96\sqrt{\widehat{Var}(\hat{\beta})},$$

will have coverage of 95% in large samples.

## Confidence Interval for a Function of $\beta$

A large sample 95% confidence interval for $g(\beta)$ is

$$g(\hat{\beta}) \pm 1.96\sqrt{\widehat{Var}[g(\hat{\beta})]}$$

where

$$\widehat{Var}[g(\hat{\beta})] = \{Var[g(\hat{\beta})]\}_{\beta=\hat{\beta}}$$

**Some motivation for this result:**

- Using the Delta method, we know that for large samples,

  $$g(\hat{\beta}) \dot{\sim} N\{g(\beta), Var[g(\hat{\beta})]\}.$$

  where

  $$Var[g(\hat{\beta})] = [g'(\beta)]^2[Var(\hat{\beta})]\},$$

  is a function of $\beta$.

- Since $\hat{\beta}$ is a consistent estimate of $\beta$ for large samples, we can substitute $\hat{\beta}$ for $\beta$ in our estimate of the variance, i.e., $Var(\hat{\beta})$.

**Example: 95% confidence interval for $g(p) = \textbf{logit}(p)$:**

- From MLE theory and the Delta method, we know that

$$\text{logit}(\widehat{p}) \dot{\sim} N\left\{\text{logit}(p), \left[\frac{1}{np(1-p)}\right]\right\}$$

- We can obtain a 95% confidence interval for logit$(p)$ by replacing $p$ in the variance with $\widehat{p}$,

$$\text{logit}(\widehat{p}) \pm 1.96\sqrt{\frac{1}{n\widehat{p}(1-\widehat{p})}},$$

## II. Hypothesis Testing

### A. Wald Tests

- Suppose we want to test $\mathrm{H}_0 : \beta = \beta^*$.

- For example, for data that follows a Binomial distribution, we may be interested in testing

$$\mathrm{H}_0 : p = 0.5$$

- Under the null hypothesis, the large sample distribution of $\widehat{p}$ is:

$$\widehat{p} \dot{\sim} N[0.5, 0.5(1 - 0.5)/n]$$

- To test the null, you can use

$$Z_1 = \frac{(\widehat{p} - 0.5)}{\sqrt{\widehat{p}\,(1 - \widehat{p})/n}} \sim N(0,1),$$

  in which $p$ in $Var(\widehat{p})$ is estimated by replacing $p$ by $\widehat{p}$

- Alternatively, you can use

$$Z_2 = \frac{(\widehat{p} - 0.5)}{\sqrt{0.5\,(1 - 0.5)/n}} \sim N(0,1),$$

  in which $p$ in $Var(p)$ is determined by replacing $p$ by $p = 0.5$ (its value under the null).

## Wald Tests, cont'd

In general, the following Wald Test statistics can be used to test the null hypothesis $\mathrm{H}_0 : \beta = \beta^*$

$$Z_1 = \frac{\hat{\beta} - \beta^*}{\sqrt{\widehat{Var}(\hat{\beta})}} \dot{\sim} N(0, 1)$$

or

$$Z_2 = \frac{\hat{\beta} - \beta^*}{\sqrt{[Var(\hat{\beta})]_{\beta=\beta^*}}} \dot{\sim} N(0, 1)$$

## Motivation for these test statistics:

- Based on large sample properties of MLE's:

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

- We already saw that for constructing confidence intervals, we can replace $\beta$ by its consistent estimator $\hat{\beta}$ in $Var(\hat{\beta})$. This motivates use of $Z_1$.

- Under the null hypothesis, an alternative is to rely on the assumption that $\beta = \beta^*$, and replace $\beta$ by $\beta^*$ in $Var(\hat{\beta})$. This motivates use of $Z_2$.

- Since the square of a $N(0,1)$ r.v. follows a $\chi^2_1$ distribution, we can also use the test statistics $Z_1^2$ or $Z_2^2$.

## B. Likelihood Ratio Tests

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

$$2\log\left\{\frac{L(\hat{\beta}|\mathrm{H}_A)}{L(\beta^*|\mathrm{H}_0)}\right\} = 2[\log L(\hat{\beta}|\mathrm{H}_A) - \log L(\beta^*|\mathrm{H}_0)] \dot{\sim} \chi^2_1$$

where

$$L(\hat{\beta}|\mathrm{H}_A)$$

is the likelihood after replacing $\beta$ by its estimate, $\hat{\beta}$, under the alternative ($\mathrm{H}_A$), and

$$L(\beta^*|\mathrm{H}_0)$$

is the likelihood after replacing $\beta$ by its specified value, $\beta^*$, under the null ($\mathrm{H}_0$).

### Example: Binomial Data

- Suppose we are interested in testing $\mathrm{H}_0 : p = 0.5$.

  For Binomial data, recall that the log-likelihood equals

$$\log L(p) = \log \begin{pmatrix} n \\ y \end{pmatrix} + y \log p + (n - y) \log(1 - p),$$

- Under the alternative,

$$\log L(\widehat{p}|\mathrm{H}_A) = \log \begin{pmatrix} n \\ y \end{pmatrix} + y \log \widehat{p} + (n - y) \log(1 - \widehat{p})$$

- Under the null,

$$\log L(0.5|\mathrm{H}_0) = \log \left( \begin{array}{c} n \\ y \end{array} \right) + y \log(0.5) + (n-y)\log(1-0.5)$$

- Then the likelihood ratio statistic is

$$2\left[\log \left( \begin{array}{c} n \\ y \end{array} \right) + y \log \widehat{p} + (n-y)\log(1-\widehat{p})\right]$$
$$-2\left[\log \left( \begin{array}{c} n \\ y \end{array} \right) + y \log(0.5) + (n-y)\log(1-0.5)\right] =$$
$$2\left[y \log\left(\tfrac{\widehat{p}}{0.5}\right) + (n-y)\log\left(\tfrac{1-\widehat{p}}{1-0.5}\right)\right]$$

which is approximately $\chi^2_1$.

- Note, we would get the same likelihood ratio statistic if we had also used the likelihood associated with $n$ independent Bernoulli r.v.'s. The log-likelihood would not contain the term

$$\log \begin{pmatrix} n \\ y \end{pmatrix}$$

but this term subtracts out in the likelihood ratio statistic, since it is not a function of $p$, and thus is the same under the null and alternative.

- In other words, any two likelihoods that are proportional will yield the same likelihood ratio statistic.

## C. Score Tests

- The SCORE TEST statistic is based on the first derivative of the log-likelihood evaluated under the null hypothesis.

- The first derivative of the log-likelihood is often referred to as the "**score**", 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.

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

  $$U(\beta^*) \dot{\sim} N(E[U(\beta^*)], Var[U(\beta^*)])$$

However, it turns out that $E[U(\beta^*)]$ always equals 0. So the asymptotic distribution can be simplified to:

$$U(\beta^*) \dot{\sim} N(0, Var[U(\beta^*)])$$

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

$$Z = \frac{U(\beta^*)}{\sqrt{Var[U(\beta^*)]}} \dot{\sim} N(0,1)$$

## Example: Test for Binomial Data:

- For Binomial data, we showed that the first derivative of the log-likelihood with respect to $p$ is

$$\begin{aligned}\frac{\partial \log L(p)}{\partial p} &= \frac{y - np}{p(1-p)} \\ &= \sum_{i=1}^{n} \frac{y_i - p}{p(1-p)}\end{aligned}$$

- The score test statistic for $\mathrm{H}_0 : p = p^*$ is:

$$\begin{aligned} Z &= \frac{U(p^*) - E[U(p^*)]}{\sqrt{Var[U(p^*)]}} \\ &= \frac{\left[\frac{y-np^*}{p^*(1-p^*)}\right] - E\left[\frac{y-np^*}{p^*(1-p^*)}\right]}{\sqrt{Var\left[\frac{y-np^*}{p^*(1-p^*)}\right]}} \end{aligned}$$

and $Z \dot{\sim} N(0,1)$.

- Next, we need to find the MEAN and VARIANCE of $U(p)$ under the null hypothesis

  Under the null $p = p^*$

$$E[U(p^*)] = \frac{E(Y - np^*)}{p^*(1 - p^*)} = 0$$

and

$$\begin{aligned} Var[U(p^*)] &= \tfrac{Var(Y-np^*)}{[p^*(1-p^*)]^2} \\ &= \tfrac{np^*(1-p^*)}{[p^*(1-p^*)]^2} \\ &= \tfrac{n}{p^*(1-p^*)}, \end{aligned}$$

- Given this mean and variance, the score statistic is

$$\begin{aligned} Z &= \frac{\left[\frac{y-np^*}{p^*(1-p^*)}\right] - E\left[\frac{y-np^*}{p^*(1-p^*)}\right]}{\sqrt{Var\left[\frac{y-np^*}{p^*(1-p^*)}\right]}} \\ &= \frac{\left[\frac{y-np^*}{p^*(1-p^*)}\right] - 0}{\sqrt{\left[\frac{n}{p^*(1-p^*)}\right]}} \\ &= \frac{y - np^*}{\sqrt{n[p^*(1-p^*)]}} \end{aligned}$$

and $Z \dot{\sim} N(0,1)$.

- Suppose we are interested in testing $\mathrm{H}_0 : p = 0.5$ then,

$$\begin{aligned} Z &= \frac{u(0.5)}{\sqrt{Var[u(0.5)]}} \\ &= \frac{y-0.5n}{\sqrt{n[0.5(1-0.5)]}} \\ &= \frac{[y-0.5n]/n}{[\sqrt{n[0.5(1-0.5)]}]/n} \\ &= \frac{\widehat{p}-0.5}{\sqrt{0.5(1-0.5)/n}} \end{aligned}$$

and $Z \dot{\sim} N(0,1)$.

## Notes on Score Tests vs Wald and LR Tests:

- Note that the SCORE statistic and the Wald statistic $Z_2$ (with the variance calculated under the null) are identical for this particular example. This is **not** usually the case.
- In more complicated problems, Score test statistics are often the easiest to calculate since you only need $\beta^*$ under the null, whereas the Likelihood Ratio and Wald statistics both use estimates of $\beta$ under the alternative. Thus, Score statistics are often popular for their simplicity.
- In large samples, all three test statistics (Score, Wald, Likelihood ratio) are numerically almost identical if the null hypothesis is true. However, if the alternative is true, the power of the three may be different.