Notes

Maximum-Likelihood Estimation: Basic Ideas

- The *method of maximum likelihood* provides estimators that have both a reasonable intuitive basis and many desirable statistical properties.
- The method is very broadly applicable and is simple to apply.
- Once a maximum-likelihood estimator is derived, the general theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for statistical inference.
- A disadvantage of the method is that it frequently requires strong assumptions about the structure of the data.

1. An Example

- We want to estimate the probability π of getting a head upon flipping a particular coin.
  - We flip the coin 'independently' 10 times (i.e., we sample $n = 10$ flips), obtaining the following result: $HHTHHHTTHH$.
  - The probability of obtaining this sequence — in advance of collecting the data — is a function of the unknown parameter π:

$$\begin{aligned}\Pr(\text{data}|\text{parameter}) &= \Pr(HHTHHHTTHH|\pi)\\ &= \pi\pi(1-\pi)\pi\pi\pi(1-\pi)(1-\pi)\pi\pi\\ &= \pi^7(1-\pi)^3\end{aligned}$$

  - But the data for our particular sample are *fixed*: We have already collected them.
  - The parameter π also has a fixed value, but this value is unknown, and so we can let it vary in our imagination between 0 and 1, treating the probability of the observed data as a function of π.

- This function is called the likelihood function:

$$\begin{aligned} L(\text{parameter}|\text{data}) &= L(\pi|HHTHHHTTHH) \\ &= \pi^7(1-\pi)^3 \end{aligned}$$

▶ The probability function and the likelihood function are given by the same equation, but the probability function is a function of the data with the value of the parameter fixed, while the likelihood function is a function of the parameter with the data fixed.

- Here are some representative values of the likelihood for different values of π:

π	$L(\pi\|\text{data}) = \pi^7(1-\pi)^3$
0.0	0.0
.1	.0000000729
.2	.00000655
.3	.0000750
.4	.000354
.5	.000977
.6	.00179
.7	.00222
.8	.00168
.9	.000478
1.0	0.0

- The complete likelihood function is graphed in Figure 1.
- Although each value of $L(\pi|\text{data})$ is a notional probability, the function $L(\pi|\text{data})$ is not a probability or density function — it does not enclose an area of 1.
- The probability of obtaining the sample of data that we have in hand, $HHTHHHTTHH$, is small regardless of the true value of π.
  - This is usually the case: *Any specific* sample result — including the one that is realized — will have low probability.
- Nevertheless, the likelihood contains useful information about the unknown parameter π.
- For example, π *cannot* be 0 or 1, and is 'unlikely' to be close to 0 or 1.

▶ Reversing this reasoning, the value of π that is most supported by the data is the one for which the likelihood is largest.
- This value is the *maximum-likelihood estimate (MLE)*, denoted $\widehat{\pi}$.
- Here, $\widehat{\pi} = .7$, which is the sample proportion of heads, 7/10.

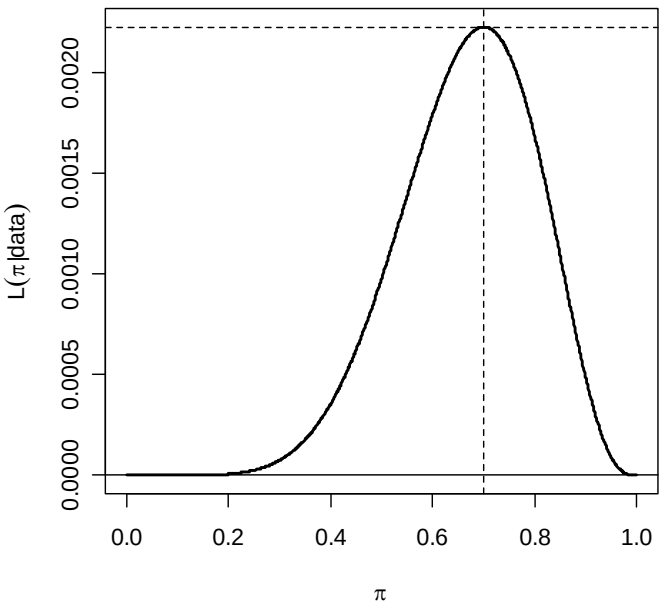

Figure 1. Likelihood of observing 7 heads and 3 tails in a particular sequence for different values of the probability of observing a head, π.

- More generally, for n independent flips of the coin, producing a particular sequence that includes x heads and $n - x$ tails,

$$L(\pi|\text{data}) = \Pr(\text{data}|\pi) = \pi^x(1-\pi)^{n-x}$$

  - We want the value of π that maximizes $L(\pi|\text{data})$, which we often abbreviate $L(\pi)$.
  - It is simpler — and equivalent — to find the value of π that maximizes the log of the likelihood

$$\log_e L(\pi) = x \log_e \pi + (n-x) \log_e (1-\pi)$$

  - Differentiating $\log_e L(\pi)$ with respect to π produces

$$\begin{aligned} \frac{d \log_e L(\pi)}{d\pi} &= \frac{x}{\pi} + (n-x)\frac{1}{1-\pi}(-1) \\ &= \frac{x}{\pi} - \frac{n-x}{1-\pi} \end{aligned}$$

  - Setting the derivative to 0 and solving produces the MLE which, as before, is the sample proportion x/n.
  - The maximum-likelihood *estimator* is $\widehat{\pi} = X/n$.

2. Properties of Maximum-Likelihood Estimators

Under very broad conditions, maximum-likelihood estimators have the following general properties:

- Maximum-likelihood estimators are consistent.
- They are asymptotically unbiased, although they may be biased in finite samples.
- They are asymptotically efficient — no asymptotically unbiased estimator has a smaller asymptotic variance.
- They are asymptotically normally distributed.
- If there is a sufficient statistic for a parameter, then the maximum-likelihood estimator of the parameter is a function of a sufficient statistic.
  - A sufficient statistic is a statistic that exhausts all of the information in the sample about the parameter of interest.

- The asymptotic sampling variance of the MLE $\widehat{\alpha}$ of a parameter α can be obtained from the second derivative of the log-likelihood:

$$\mathcal{V}(\widehat{\alpha}) = \frac{1}{-E\left[\dfrac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]}$$

  - The denominator of $\mathcal{V}(\widehat{\alpha})$ is called the *expected* or *Fisher information*

$$\mathcal{I}(\alpha) \equiv -E\left[\frac{d^2 \log_e L(\alpha)}{d\alpha^2}\right]$$

  - In practice, we substitute the MLE $\widehat{\alpha}$ into the equation for $\mathcal{V}(\widehat{\alpha})$ to obtain an *estimate* of the asymptotic sampling variance, $\widehat{\mathcal{V}(\widehat{\alpha})}$.

- $L(\widehat{\alpha})$ is the value of the likelihood function at the MLE $\widehat{\alpha}$, while $L(\alpha)$ is the likelihood for the true (but generally unknown) parameter α.
  - The *log likelihood-ratio statistic*

$$G^2 \equiv -2\log_e \frac{L(\alpha)}{L(\widehat{\alpha})} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha)]$$

    follows an asymptotic chisquare distribution with one degree of freedom.
    - Because, by definition, the MLE maximizes the likelihood for our particular sample, the value of the likelihood at the true parameter value α is generally smaller than at the MLE $\widehat{\alpha}$ (unless, by good fortune, $\widehat{\alpha}$ and α happen to coincide).

3. Statistical Inference: Wald, Likelihood-Ratio, and Score Tests

These properties of maximum-likelihood estimators lead directly to three common and general procedures for testing the statistical hypothesis H_0: $\alpha = \alpha_0$.

1. *Wald Test:* Relying on the asymptotic normality of the MLE $\widehat{\alpha}$, we calculate the test statistic

$$Z_0 \equiv \frac{\widehat{\alpha} - \alpha_0}{\sqrt{\widehat{\mathcal{V}(\widehat{\alpha})}}}$$

   which is asymptotically distributed as $N(0, 1)$ under H_0.

2. *Likelihood-Ratio Test:* Employing the log likelihood ratio, the test statistic

$$G_0^2 \equiv -2\log_e \frac{L(\alpha_0)}{L(\widehat{\alpha})} = 2[\log_e L(\widehat{\alpha}) - \log_e L(\alpha_0)]$$

   is asymptotically distributed as χ_1^2 under H_0.

3. *Score Test:* The 'score' is the slope of the log-likelihood at a particular value of α, that is, $S(\alpha) \equiv d\log_e L(\alpha)/d\alpha$.
   - At the MLE, the score is 0: $S(\widehat{\alpha}) = 0$. It can be shown that the *score statistic*
   $$S_0 \equiv \frac{S(\alpha_0)}{\sqrt{\mathcal{I}(\alpha_0)}}$$
   is asymptotically distributed as $N(0, 1)$ under H_0.

- Unless the log-likelihood is quadratic, the three test statistics can produce somewhat different results in specific samples, although the three tests are asymptotically equivalent.
- In certain contexts, the score test has the practical advantage of not requiring the computation of the MLE $\widehat{\alpha}$ (because S_0 depends only on the null value α_0, which is specified in H_0).
- The Wald and likelihood-ratio tests can be 'turned around' to produce confidence intervals for α.

- Figure 2 compares the three test statistics.
- Maximum-likelihood estimation and the Wald, likelihood-ratio, and score tests, extend straightforwardly to simultaneous estimation of several parameters.
- When the log-likelihood function is relatively flat at its maximum, as opposed to sharply peaked, there is little information in the data about the parameter, and the MLE will be an imprecise estimator: See Figure 3.

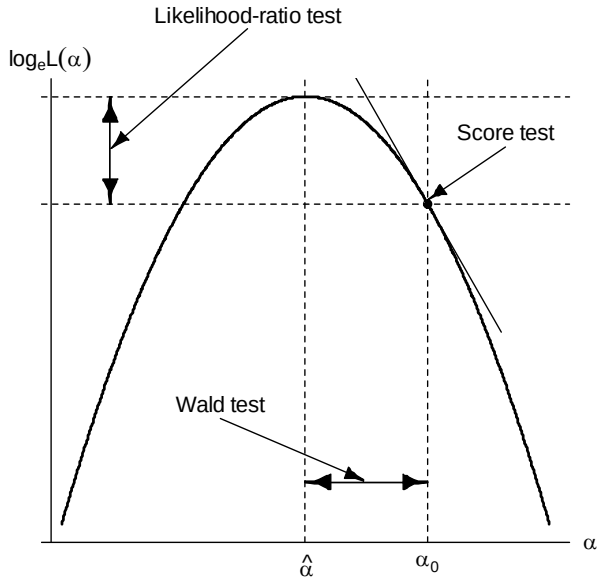

Figure 2. Likelihood-ratio, Wald, and score tests.

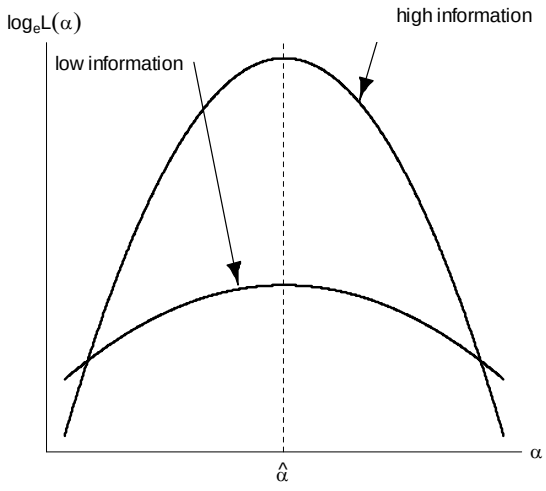

Figure 3. Two imagined log likelihoods: one strongly peaked, providing high information about the the parameter α; and the other flat, providing low information about α.

