Eco311 Optional Reading: Maximum Likelihood Method

(Jing Li, Miami University)

1. Maximum Likelihood (ML) is a statistical method used to estimate a population parameter. For instance, suppose we are interested in a Bernoulli distribution that has only two outcomes (1/0, yes/no, head/tail, Trump/Biden, vaccinated/unvaccinated...), and the unknown parameter is the population proportion θ : we want to know the proportion of persons voting for Trump, or the probability of seeing head after flipping a coin.

2. θ is not known unless we flip the coin *infinity* times, or we ask *all* voters who they vote. Using a sample means that we flip the coin finite times, or we only ask some voters who they vote.

3. Suppose the outcomes of tossing the coin 10 times are

   $$\{``H", ``T", ``H", ``T", ``T", ``H", ``H", ``H", ``T", ``H"\}$$

   where "H" denotes head; "T" denotes tail. We can convert those strings into numeric values:

   $$\{1, 0, 1, 0, 0, 1, 1, 1, 0, 1\}$$

   where 1 denotes head; 0 denotes tail

4. Let y be the outcome variable that equals either 1 or 0. Our goal is estimating the unknown parameter $\theta = P(y = 1)$, which represents the *population* proportion. Based on the fact that we observe 6 heads out of 10 tossing, we compute the *sample* proportion as

   $$\hat{\theta} = \frac{\texttt{number of heads}}{\texttt{number of tossing}} = \frac{6}{10} = 0.6$$

   We put a hat above θ to emphasize it is a *sample* estimate.

5. Intuitively, the unknown population proportion should be close to 0.6, $\theta \approx 0.6$. We say "close" because we may get different sample proportion if we toss the coin ten times again. There is inherent uncertainty associated with using samples. A key issue of statistics is accounting for the sampling uncertainty.

6. The number of heads we see after tossing a coin ten times follows *Binomial* distribution:

   $$P(\texttt{seeing k heads after 10 tossing}) = C_{10}^{k}\theta^{k}(1-\theta)^{10-k}, \quad (k = 0, 1, 2, ...10) \quad (1)$$

   where θ is the probability of seeing head when the coin is tossed once: $\theta = P(y = 1)$. For given θ, $C_{10}^{k}\theta^{k}(1-\theta)^{10-k}$ is a function of k, called probability mass function. For instance, if $\theta = 0.1$, the probability of seeing 6 heads after tossing the coin 10 times is

   $$C_{10}^{6}0.1^{6}0.9^{4} = 0.000137781$$

   The R code to get that probability is

   ```
   > choose(10, 6)*0.1^6*0.9^4
   [1] 0.000137781
   ```

   Equivalently, we can use this R function

   ```
   > dbinom(6,10,0.1)
   [1] 0.000137781
   ```

   This small probability makes sense because if the true probability of getting head is 0.1, then it is extremely rare to see 6 heads after tossing the coin 10 times. It is much more likely to see only one or two heads after ten tossing. For example, the probability of seeing only one head is

   ```
   > dbinom(1,10,0.1)
   [1] 0.3874205
   ```

   which is much greater than 0.000137781.

7. More generally, we can list the probability of seeing 6 heads assuming θ equals 0.1, 0.2, 0.3,...0.9, and see which θ value produces the greatest probability

```
> theta = 0.1*seq(1,9,1)
> probability = dbinom(6,10,theta)
> cbind(theta,probability)
     theta probability
[1,]   0.1 0.000137781
[2,]   0.2 0.005505024
[3,]   0.3 0.036756909
[4,]   0.4 0.111476736
[5,]   0.5 0.205078125
[6,]   0.6 0.250822656
[7,]   0.7 0.200120949
[8,]   0.8 0.088080384
[9,]   0.9 0.011160261
```

We see that the greatest probability 0.250822656 comes from $\theta = 0.6$, which implies that it is mostly likely to see 6 heads after 10 tossing if $\theta = 0.6$. Notice that the probabilities of seeing 6 heads when $\theta = 0.5$ or 0.7 are about 20 percent, so those two θ values are also plausible. Put differently, we are not confident to rule out $\theta = 0.5$ or 0.7 if seeing 6 heads from 10 tossing. The most likely θ value is 0.6, but 0.5 and 0.7 are also compatible with the reality.

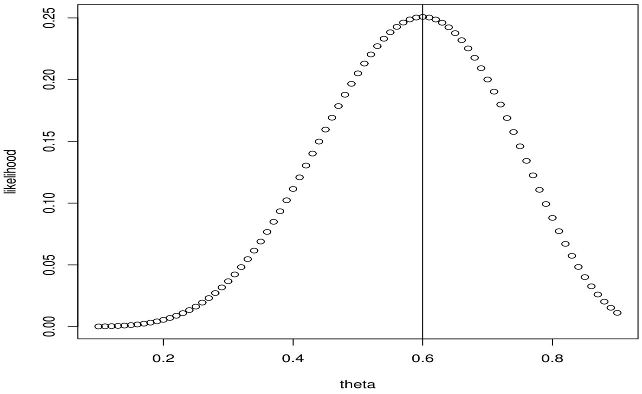
8. For given $k = 6$, the function in (1) is a function of θ, called <u>likelihood function</u>

$$C_{10}^{6}\theta^{6}(1-\theta)^{10-6}, \;\; \theta \in (0,1), \quad (\texttt{likelihood function}) \tag{2}$$

   Do not confuse it with the probability mass function, for which we hold θ constant and let k vary. Here for function (2), we hold $k = 6$ constant and let θ vary.

9. We use these R codes to plot the likelihood function

```
theta = 0.1*seq(1,9,0.1)
likelihood = dbinom(6,10,theta)
plot(theta,likelihood)
abline(v=0.6)
```



where a vertical line is drawn for $\theta = 0.6$, the value maximizing the likelihood function.

10. (A technical issue) Likelihood function is not probability function since it does not sum up to 1

```
> sum(likelihood)
[1] 9.071025
```

We can make it "look" like a probability function by rescaling

```
> s.likelihood = likelihood/sum(likelihood)
> sum(s.likelihood)
[1] 1
```

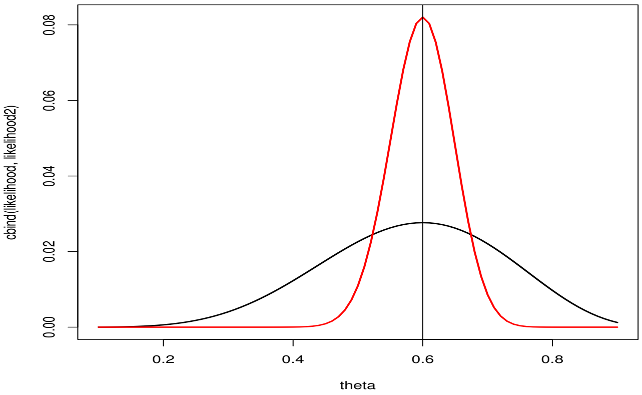
11. By definition, the maximum likelihood (ML) estimate of a population parameter is the value that maximizes the likelihood function (the value that corresponds to the peak of hump). For this example, the ML estimate is $\hat{\theta} = 0.6$. For any other value, the chance of seeing 6 heads from 10 tossing gets smaller. In light of that, $\hat{\theta} = 0.6$ is the estimate that is mostly supported by the observed reality.

12. Next, suppose we toss the coin 100 times, and see 60 heads. We use R code

```
dbinom(60,100,theta)/sum(dbinom(60,100,theta))
```

to compute the new rescaled likelihood (red color), and compare it to the old one (black)

```
theta = 0.1*seq(1,9,0.1)
likelihood = dbinom(6,10,theta)/sum(dbinom(6,10,theta))
likelihood2 = dbinom(60,100,theta)/sum(dbinom(60,100,theta))
matplot(theta,cbind(likelihood,likelihood2), lwd  = 2, type="l", lty=c(1,1),col=c("
abline(v=0.6)
```



We see that the ML estimate is still $\hat{\theta} = 0.6$. However, the red curve is *narrower* than the black one, indicating *greater* confidence or *less* uncertainty. In other words, getting 60 heads from 100 tossing is more convincing than getting 6 heads from 10 tossing in terms of justifying the estimate $\hat{\theta} = 0.6$.

13. Another way to understand the implication of narrower likelihood function is comparing probabilities of seeing 6 heads from 10 tossing and seeing 60 heads from 100 tossing assuming the true probability is $\theta = 0.5$:

```
> dbinom(6,10,0.5)
[1] 0.2050781
> dbinom(60,100,0.5)
[1] 0.01084387
```

The second probability is much less than the first one. According to the first probability 0.2050781, which is from the black likelihood, we cannot rule out that the true probability is 0.5 when seeing 6 heads out of 10 tossing. But according to the second probability 0.01084387, which is from the red likelihood, we can rule out $\theta = 0.5$ with confidence since 0.01084387 is close to 0. In short, a narrower likelihood allows us to rule out more θ values, or equivalently, we can have more confidence on the values that we cannot rule out.

14. We can also compare the *slopes* of the two likelihoods. A narrow likelihood is also a *steep* one, which means that the deduction in probability is greater than a wide likelihood when moving away from the peak of hump

    ```
    > dbinom(6,10,0.5)-dbinom(6,10,0.6)
    [1] -0.04574453
    > dbinom(60,100,0.5)-dbinom(60,100,0.6)
    [1] -0.07037528
    ```

    So when θ changes from 0.6 to 0.5, the probability of seeing 6 heads from 10 tossing reduces by 0.04574453. By contrast, the probability of seeing 60 heads from 100 tossing reduces by 0.07037528, a bigger change. The greater reduction in probability adds more confidence to the ML estimate.

15. Lesson 1: the value that maximizes the likelihood function is the maximum likelihood estimate

16. Lesson 2: the *width* of likelihood function measures confidence—a narrow or steep likelihood (from a big sample) leads to more confidence in the estimate than a wide or flat likelihood (from a small sample).

17. Obtaining an interval estimate from the rescaled likelihood is easy:

    ```
    > theta[as.logical((cumsum(likelihood2)>0.025)*(cumsum(likelihood2)<0.975))]
     [1] 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 0.60 0.61 0.62 0.63 0.64 0.6
    ```

    The 95 percent confidence interval based on the sample of seeing 60 heads out of 100 tossing is $(0.5, 0.68)$. By contrast, the 95 percent confidence interval based on the sample of seeing 6 heads out of 10 tossing is $(0.31, 0.82)$.

    ```
    > theta[as.logical((cumsum(likelihood)>0.025)*(cumsum(likelihood)<0.975))]
     [1] 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44 0.45 0.4
    [21] 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 0.60 0.61 0.62 0.63 0.64 0.65 0.6
    [41] 0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82
    ```

    The *wider* confidence intervals indicate *greater uncertainty*.