

## 8.5 Maximum likelihood estimators

We are still looking at

$$X_1, \dots, X_n \stackrel{iid}{\sim} f(x|\boldsymbol{\theta}).$$

Maximum likelihood estimators do what their name implies: they find parameter values  $\boldsymbol{\theta}$  that make the data  $X_1, \dots, X_n$  as “likely” as possible.

**Example:** Flip an unfair coin 10 times; let  $X_i = 0$  if tails comes up on  $i^{th}$  toss and  $X_i = 1$  if heads comes up. Let  $\pi$  be the unknown probability of heads. So

$$X_1, \dots, X_{10} \stackrel{iid}{\sim} \text{Bern}(\pi).$$

Say we see  $\sum_{i=1}^{10} X_i = 8$  heads out of the 10 tosses.

If you had to pick between  $\pi = 0.1$  and  $\pi = 0.9$  as having generated the data, which would you pick? Why?

I would pick  $\pi = 0.9$  because seeing 8 heads out of 10 is more likely to have occurred when  $\pi = 0.9$  than when  $\pi = 0.1$ .

Formally:

$$p(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} | \pi) = \pi^{\sum_{i=1}^{10} x_i} (1 - \pi)^{10 - \sum_{i=1}^{10} x_i}.$$

So

$$P(1, 1, 1, 0, 1, 1, 0, 1, 1, 1 | \pi = 0.1) = 0.1^8 0.9^2 \approx 0.00000001,$$

$$P(1, 1, 1, 0, 1, 1, 0, 1, 1, 1 | \pi = 0.9) = 0.9^8 0.1^2 \approx 0.004.$$

If you had to pick between  $\pi = 0.5$  and  $\pi = 0.7$  which would you pick?

If you had to pick any number in  $(0, 1)$ , what would you pick?

Now let's consider seeing data  $X_1 = 4$  and  $X_2 = 6$ . If we have to choose between a  $N(0, 1^2)$  and a  $N(5, 1^2)$  as having generated these data, which would you pick?

How about between  $N(5, 1^2)$  and  $N(5, 4^2)$ ?

Plots will help us decide.

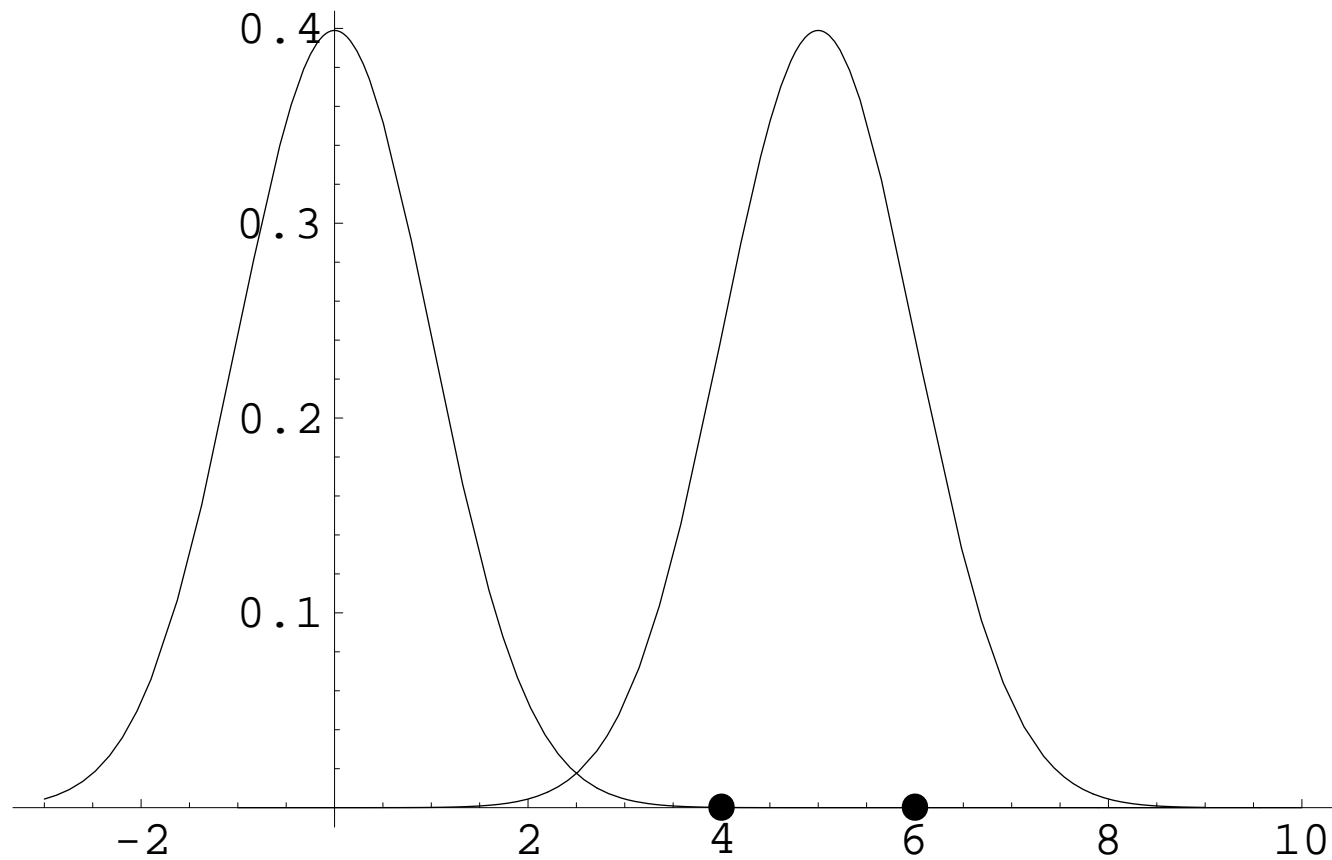


Figure 1: Data  $x_1 = 4$ ,  $x_2 = 6$ , and  $N(0, 1^2)$ ,  $N(5, 1^2)$  densities.

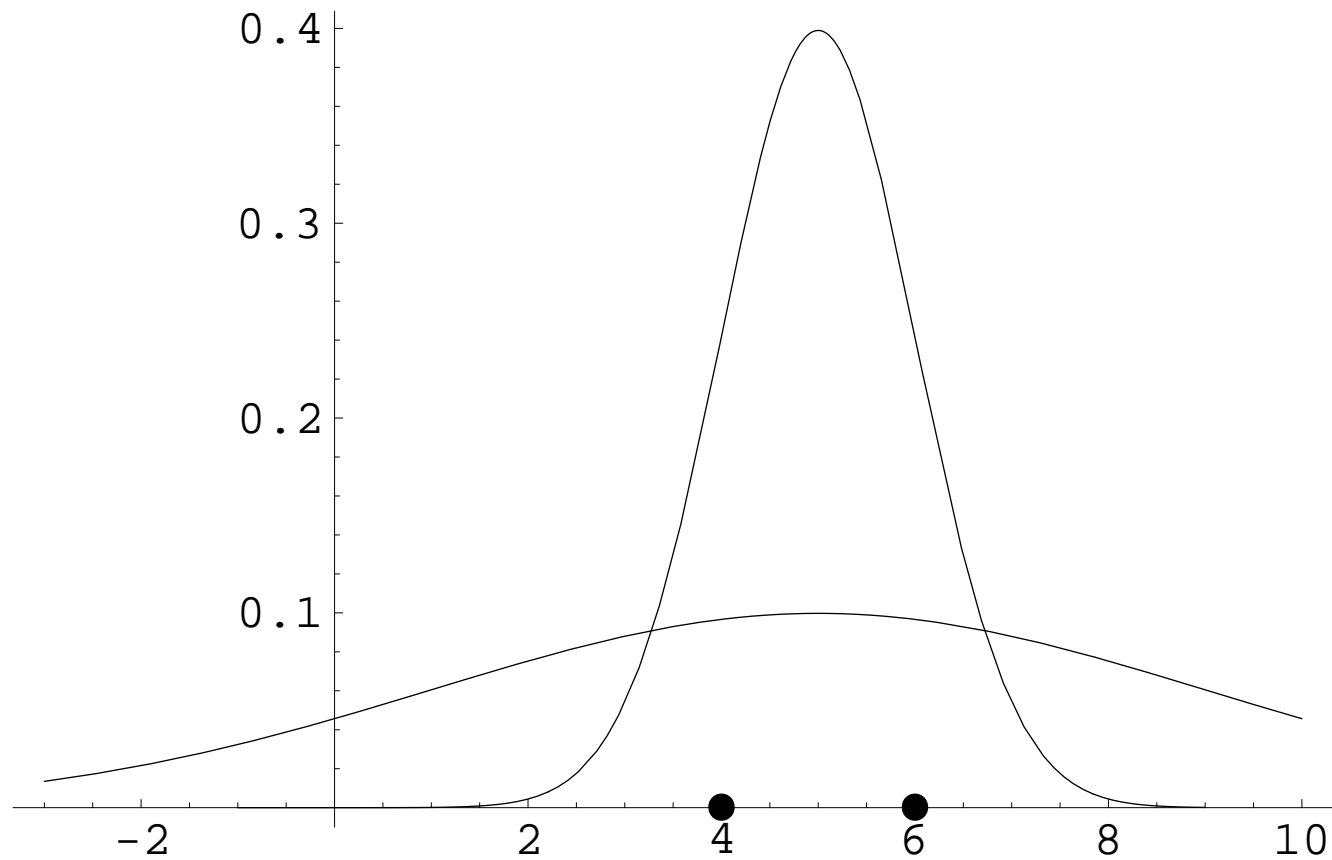


Figure 2: Data  $x_1 = 4$ ,  $x_2 = 6$ , and  $N(5, 1^2)$ ,  $N(5, 4^2)$  densities.

**def'n:** the maximum likelihood estimate (MLE) of  $\boldsymbol{\theta}$ , denoted  $\hat{\boldsymbol{\theta}}$ , is the value of  $\boldsymbol{\theta}$  that maximizes the likelihood of the data

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} f(x_1, \dots, x_n | \boldsymbol{\theta}),$$

where  $f(x_1, \dots, x_n | \boldsymbol{\theta})$  is the joint density of the data  $\mathbf{X} = \mathbf{x}$ . When we treat the joint density as a function of  $\boldsymbol{\theta}$  given the  $(x_1, \dots, x_n)$  we call it a likelihood and denote it  $\mathcal{L}(\boldsymbol{\theta}) = f(x_1, \dots, x_n | \boldsymbol{\theta})$ .

If the data are *iid* then

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i | \boldsymbol{\theta}).$$

For discrete  $X_1, \dots, X_n$ , we replace the pdf  $f(x_1, \dots, x_n | \boldsymbol{\theta})$  with the pmf  $p(x_1, \dots, x_n | \boldsymbol{\theta})$ .

In finding MLE's by hand, we can use a useful result:

**Result:** For any function  $g(x)$ ,  $a$  is a maximum of  $g(x)$  if and only if  $a$  is a maximum of  $\log g(x)$ .

The reason this is useful is that finding an MLE often involves taking derivatives of  $\mathcal{L}(\boldsymbol{\theta})$  and setting to zero. The result says we could instead take derivatives of  $\log \mathcal{L}(\boldsymbol{\theta})$  and set to zero to find the MLE.

Define  $l(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta})$ .

**Bernoulli data:** Let

$$X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(\pi).$$

Note that we can write the pmf of  $X_i$  as

$$p(x_i|\pi) = P(X_i = x_i|\pi) = \pi^{x_i}(1 - \pi)^{1-x_i},$$

for  $x_i \in \{0, 1\}$ . To find the MLE  $\hat{\pi}$  given  $\mathbf{X} = \mathbf{x}$  first note that since the data are *iid*,

$$\mathcal{L}(\pi) = \prod_{i=1}^n \pi^{x_i} (1 - \pi)^{1-x_i} = \pi^{\sum_{i=1}^n x_i} (1 - \pi)^{n - \sum_{i=1}^n x_i}.$$

Given data  $\mathbf{X} = \mathbf{x}$ , this *is a function of  $\pi$  only*. We maximize it by taking its derivative and setting equal to zero. It turns out to be much easier to take the derivative of the log-likelihood  $l(\pi) = \log \mathcal{L}(\pi)$  and setting that to zero instead.

We can simplify this a bit by noting  $\sum_{i=1}^n x_i = n\bar{x}_n$ . Then the log-likelihood is  $l(\pi) = n\bar{x}_n \log \pi + n(1 - \bar{x}_n) \log(1 - \pi)$ . Take the first derivative, set equal to zero, and solve for  $\pi$ :

$$l'(\pi) = \frac{n\bar{x}_n}{\pi} - \frac{n(1 - \bar{x}_n)}{1 - \pi} \stackrel{\text{set}}{=} 0.$$

The solution is  $\pi = \bar{x}_n$  and so the MLE is  $\hat{\pi} = \bar{x}_n$ , the sample proportion of 1's out of  $n$  trials. For Bernoulli data, the MLE is the same as the MOM estimator.

In general, one needs to check that the solution is a local maximum by checking second derivatives. We won't worry about that here.

## Normal data

$$X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

The log-likelihood is

$$\begin{aligned} l(\mu, \sigma^2) &= \sum_{i=1}^n \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \right\} \\ &= -0.5n \log(2\pi) - 0.5n \log(\sigma^2) - \frac{0.5}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\ &= -0.5n \log(2\pi) - 0.5n \log(\sigma^2) - \frac{0.5}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - \mu)^2 \\ &= -0.5n \log(2\pi) - 0.5n \log(\sigma^2) - \frac{0.5}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2 \\ &= -0.5n \log(2\pi) - 0.5n \log(\sigma^2) - \frac{0.5}{\sigma^2} n s_n^2 + n(\bar{x}_n - \mu)^2 \end{aligned}$$

This is because

$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - \mu)^2 &= \sum_{i=1}^n [(x_i - \bar{x}_n)^2 + 2(x_i - \bar{x}_n)(\bar{x}_n - \mu) + (\bar{x}_n - \mu)^2] \\
 &= \sum_{i=1}^n (x_i - \bar{x}_n)^2 + 2 \sum_{i=1}^n (x_i - \bar{x}_n)(\bar{x}_n - \mu) + \sum_{i=1}^n (\bar{x}_n - \mu)^2 \\
 &= ns_n^2 + 2(\bar{x}_n - \mu) \sum_{i=1}^n (x_i - \bar{x}_n) + n(\bar{x}_n - \mu)^2 \\
 &= ns_n^2 + 2(\bar{x}_n - \mu)(n\bar{x}_n - n\bar{x}_n) + n(\bar{x}_n - \mu)^2 \\
 &= ns_n^2 + n(\bar{x}_n - \mu)^2
 \end{aligned}$$

We need to solve the simultaneous equations:

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = 2n(\bar{x}_n - \mu)(-1) \stackrel{set}{=} 0$$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = \frac{-0.5n}{\sigma^2} + \frac{0.5n}{\sigma^4} s_n^2 \stackrel{set}{=} 0$$

This leads to  $\hat{\mu} = \bar{X}_n$  and  $\hat{\sigma}^2 = S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ .

Same as MOM.

## Comments:

- Sometimes it is easy to find the MOM estimator, but not the MLE, e.g. gamma data, beta data.
- MLE's always exist, not so for MOM.
- The method of maximum likelihood immediately generalizes to independent but not identically distributed data (e.g. two sample problem, regression) or even not independent nor identically distributed data (e.g. time series data).
- In class we found MOM's for normal data, gamma data, Poisson data,  $U(a, b)$  data, and exponential data. In your notes we found MLE's for Bernoulli and normal data. We will find MLE's for exponential data,  $U(0, b)$  data (tricky! can't use derivative), and Poisson data.

## Example: leukemia survival

The lifetimes in weeks of  $n = 33$  patients who died of acute myelogenous leukemia are 156, 65, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 5, 1, 1, 65, 56, 65, 17, 7, 16, 22, 3, 4, 2, 3, 8, 4, 3, 30, 4, 43.

Assuming

$$X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$$

gives MLEs  $\hat{\mu} = \bar{x}_{33} = 40.9$  and  $\hat{\sigma}_{33} = 46.0$ .

Assuming

$$X_1, \dots, X_n \stackrel{iid}{\sim} \exp(\lambda)$$

gives MLE  $\hat{\lambda} = 1/\bar{x}_{33} = 1/40.91 = 0.0245$

The next slide shows a normal distribution provides comparatively worse fit to these data.

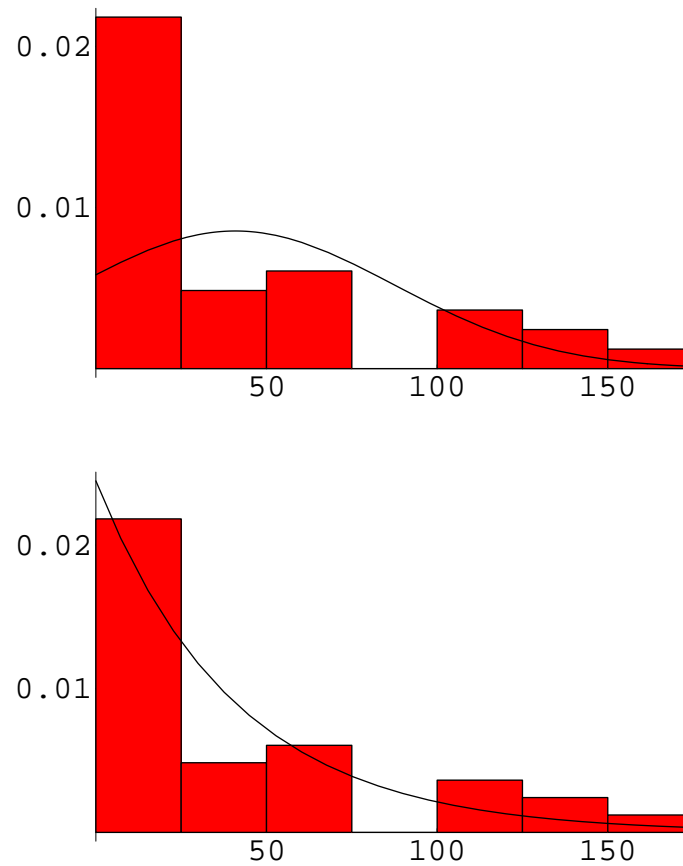


Figure 3: Histogram of leukemia survival times and MLE fits of normal and exponential models.

**Important result:** the MLE of any function of all the model parameters  $\theta$  is just the function evaluated at  $\hat{\theta}$ .

That is, the MLE of  $g(\theta)$  is  $g(\hat{\theta})$ , just plug in the MLE.

**Leukemia example:** Say we are interested in is  $x_{99}$ , the survival time in weeks that 99% of the patients do not reach. This is given by  $x_{0.99} = F^{-1}(0.99) = -\log(1 - 0.99)/\lambda$  for exponential data.

So here  $g(\lambda) = -\log(1 - 0.99)/\lambda$ .

The MLE is given by

$$\hat{x}_{0.99} = -\log(1 - 0.99)/\hat{\lambda} = -\log(1 - 0.99)/0.0245 = 188.0 \text{ weeks,}$$

about 3.6 years.

## Asymptotic normality

Under very general conditions MOM and MLE estimators  $\hat{\boldsymbol{\theta}}$  are both approximately normal when  $n$  is large. This is easier to show for MOM estimators using the CLT and an a technique known as the “delta method.” Showing that MLE estimators are approximately normal takes more work (e.g. Section 8.5.2 see pp. 274-279).

The conditions under which the elements of MLE  $\hat{\boldsymbol{\theta}}(\mathbf{X})$  are normal boil down to two things: (1)  $f(x|\boldsymbol{\theta})$  or  $p(x|\boldsymbol{\theta})$  must be sufficiently smooth as a function of  $\boldsymbol{\theta}$ , and (2) The range of each  $X_i$  must not depend on  $\boldsymbol{\theta}$ . This rules out, for example,  $U(a, b)$  data.

MOM estimators, when they exist, have asymptotic normal distributions through the central limit theorem and the *delta method*, which we'll talk about next time.

**Example:** For  $U(0, \theta)$  data, the MOM is  $\hat{\theta} = 2\bar{X}_n$  which is obviously approximately normal by the CLT, but the MLE is  $\hat{\theta} = \max\{X_1, X_2, \dots, X_n\}$ , which is not.

**Example:** For Poisson( $\theta$ ) data, both the MLE and the MOM are  $\hat{\theta} = \bar{X}_n$  which is approximately normal by the CLT.