

MOM estimators are *always* approximately normal. Here's two simple examples:

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

Then the CLT says

$$\hat{\pi} = \bar{X}_n \stackrel{\bullet}{\sim} N(\pi, \pi(1 - \pi)/n).$$

**Poisson data** :  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Pois}(\lambda)$ .

Then the CLT says

$$\hat{\lambda} = \bar{X}_n \stackrel{\bullet}{\sim} N(\lambda, \lambda/n).$$

In both cases the estimate was the sample mean.

How about

**Uniform data** :  $X_1, \dots, X_n \stackrel{iid}{\sim} U(0, b)$ .

Then  $\hat{b} = 2\bar{X}_n$ . Then the CLT and properties of normals tell us

$$\hat{b} = 2\bar{X}_n \overset{\bullet}{\sim} N(b, b^2/(3n)),$$

because  $\text{var}(X_i) = (b - 0)^2/12$ .

How about

**Exponential data** :  $X_1, \dots, X_n \stackrel{iid}{\sim} \exp(\lambda)$ .

Then  $\hat{\lambda} = 1/\bar{X}_n$ . The CLT doesn't work here...at least directly...

## The delta method

The delta method relies on two results:

- A linear function of a normal random variable is also normal.
- Taylor's theorem tells us that any continuous function can be approximated by a linear function over a small range of values.

### Specifically...

- $Y \sim N(\mu, \sigma^2)$  implies  $a + bY \sim N(a + b\mu, b^2\sigma^2)$  for any  $a$  and  $b > 0$ .
- For any continuous function  $g(x)$ , the mean value theorem (or Taylor's theorem, or a picture) tells us that

$$g(x) \approx g(a) + g'(a)(x - a),$$

for any fixed point  $a$ .

The central limit theorem tells us

$$\bar{X}_n \overset{\bullet}{\sim} N(\mu, \sigma^2/n),$$

where  $\mu = E(X_i)$  and  $\sigma^2 = \text{Var}(X_i)$ .

The mean value theorem gives us:

$$\begin{aligned} g(\bar{X}_n) &\approx g(\mu) + g'(\mu)(\bar{X}_n - \mu) \\ &= [g(\mu) - g'(\mu)\mu] + [g'(\mu)]\bar{X}_n \\ &= a + b\bar{X}_n \end{aligned}$$

where  $a = g(\mu) - g'(\mu)\mu$  and  $b = g'(\mu)$ .

**Finally:**

$$g(\bar{X}_n) \overset{\bullet}{\sim} N(g(\mu), g'(\mu)^2 \sigma^2 / n)$$

for any continuous  $g(x)$ .

For the case when  $\theta$  is univariate (exponential, Poisson, Bernoulli, etc.), the MOM solves

$$\bar{X}_n = E(X_i) = \mu = h(\theta),$$

yielding  $\hat{\theta} = h^{-1}(\bar{X}_n)$ . Just use  $g(x) = h^{-1}(x)$  in the previous result.

**Example:** exponential data

Let

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

Then using either MOM or MLE,  $\hat{\theta} = 1/\bar{X}_n = g(\bar{X}_n)$ .

Then  $g(x) = 1/x$  and  $g'(x) = -1/x^2$ .

So  $g(\mu) = g(1/\theta) = 1/(1/\theta) = \theta$  and similarly  
 $g'(1/\theta)^2 \text{Var}(X_i)/n = \theta^2/n$ .

So

$$\hat{\theta} \underset{\circ}{\sim} N(\theta, \theta^2/n).$$

Note then that as  $n$  gets large,  $E(\hat{\theta}) = \theta$  (that's nice!) and  $\text{Var}(\hat{\theta}) = \theta^2/n$ .

In fact packages like R or SAS use the approximation  $\hat{\theta} \overset{\bullet}{\sim} N(\theta, \theta^2/n)$  instead of the exact distribution of  $\hat{\theta} = 1/\bar{X}_n$  (which, for example, can only be positive).

---

The CLT and the delta method imply that *all* MOM estimators are approximately normal. They may not be the *best* estimators, but they are often easy to find, and are always normal.

## Large sample normality of the MLE

Under certain conditions, the distribution of the MLE  $\hat{\boldsymbol{\theta}}$  is approximately normal. This is before we collect data  $\mathbf{X} = (X_1, \dots, X_n)$ , and so the data  $\mathbf{X}$  is random, implying the MLE, a function of the data, is random and has a distribution.

**prop:** Let  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x|\boldsymbol{\theta})$  where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ . If  $f(x|\boldsymbol{\theta})$  is sufficiently smooth and behaves well and if the range of  $X_i$  does not depend on  $\boldsymbol{\theta}$ , then each element of  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$  is approximately normal when  $n$  is large:

$$\hat{\theta}_j \stackrel{\bullet}{\sim} N(\theta_j, \text{Var}(\hat{\theta}_j)).$$

The proof of this proposition involves moment generating functions (Section 4.5) and Taylor's theorem. it's outlined in Section 8.5.2, pp. 274-279 for univariate  $\theta$  ( $p = 1$ ).

If  $\hat{\theta}_j$  has a closed form (e.g.  $\hat{\lambda} = 1/\bar{X}_n$  for exponential data) then it might be possible to compute  $\text{Var}(\hat{\theta}_j)$  explicitly. Otherwise, the *information matrix* is used to find  $\text{Var}(\hat{\theta}_j)$ .

For a one-dimensional  $\theta$  and *iid* data, the information is just a number  $I(\theta)$  that depends on  $\theta$ :

**def'n:** For one-dimensional  $\theta$  and  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x|\theta)$ ,

$$I(\theta) = E \left[ \frac{\partial}{\partial \theta} \log f(X_i|\theta) \right]^2 = -E \left[ \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta) \right].$$

The large sample result for MLE's of univariate  $\theta$  is, for large  $n$ ,

$$\hat{\theta} \underset{\bullet}{\sim} N \left( \theta, \frac{1}{nI(\theta)} \right).$$

Although not pretty, this result generalizes to multivariate  $\boldsymbol{\theta}$ , and to non-*iid* data, and is wildly useful and used over and over again in R, SAS, SPSS, Excel, Minitab, etc.

Let's compute  $I(\theta)$  for *iid*  $\exp(\theta)$  data.

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta) &= \frac{\partial}{\partial \theta} \left[ \frac{\partial}{\partial \theta} \log(\theta e^{-\theta X_i}) \right] \\ &= \frac{\partial}{\partial \theta} \left[ \frac{\partial}{\partial \theta} [\log \theta - \theta X_i] \right] \\ &= \frac{\partial}{\partial \theta} \left[ \frac{1}{\theta} - X_i \right] = -\frac{1}{\theta^2}. \end{aligned}$$

We have then

$$I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log f(X_i | \theta) \right] = -E \left[ -\frac{1}{\theta^2} \right] = \frac{1}{\theta^2}.$$

So

$$\hat{\theta} \overset{\bullet}{\sim} N \left( \theta, \frac{1}{nI(\theta)} \right) = N(\theta, \theta^2/n).$$

In fact, we can use moment generating functions to show that, *exactly*,

$$\bar{X}_n \sim \text{gamma}(n, n\theta).$$

This can be used to show that, *exactly*,  $E(\hat{\theta}) = E(1/\bar{X}_n) = \frac{n}{n-1}\theta$ . As  $n \rightarrow \infty$ ,  $E(\hat{\theta}) \approx \theta$  as required by the large sample normal result.

## Leukemia data (again!)

Assume the data are exponential

$$X_1, \dots, X_{33} \stackrel{iid}{\sim} \exp(\theta).$$

The MLE and the MOM of  $\theta$  is  $\hat{\theta} = 1/\bar{X}_n$ .

Recall  $\hat{\theta} \stackrel{\bullet}{\sim} N(\theta, \theta^2/n)$  for exponential data.

We estimate  $\theta$  by  $\hat{\theta} = 1/\bar{y}_{33} = 1/40.9 = 0.0245$ .

The approximate variance of  $\hat{\theta}$ , given by  $\theta^2/n$ , is unknown. We can estimate it by  $\hat{\theta}^2/n = 0.0245^2/33 = 1.82 \times 10^{-5}$ .

Before the lifetimes were actually observed we approximate the distribution of  $\hat{\theta}$  based on  $n = 33$  observations by

$$\hat{\theta} \stackrel{\bullet}{\sim} N(\theta, 1.82 \times 10^{-5}).$$

This is subtle.

Before the data are collected,  $\hat{\theta}$  is a *random function of the data*.

After data are collected we actually see  $\hat{\theta} = 0.0245$ .

We use this to estimate the variance of the *random estimator*  $\hat{\theta}$  *before the survival times are collected*.

We will use this later to construct a hypothesis test.

## A basic property of estimators $\hat{\theta}$ , unbiasedness

**def'n:** The estimator  $\hat{\theta}$  is unbiased for  $\theta$  if  $E(\hat{\theta}) = \theta$ . Examples:

- For  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(\pi)$ ,  $E(\bar{X}_n) = \pi$  so  $\hat{\pi} = \bar{X}_n$  is unbiased for  $\pi$ .
- For  $X_1, \dots, X_n \stackrel{iid}{\sim} \exp(\theta)$ ,  $E(1/\bar{X}_n) = \frac{n}{n-1}\theta$  so  $1/\bar{X}_n$  is biased for  $\lambda$ . What happens to the bias when  $n$  gets large?
- For  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ , the sample variance  $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  is biased for  $\sigma^2$ ,  $E(S_n^2) = \frac{n-1}{n}\sigma^2$ . What happens to the bias when  $n$  gets large? ( $S_n^2$  is the MLE for  $\sigma^2$ ).

## MLE's for general models

Any data,  $\mathbf{X} = (X_1, \dots, X_n)$ , independent, dependent, not identically distributed, whatever, has a distribution

$$\mathbf{X} = (X_1, \dots, X_n) \sim f(x_1, \dots, x_n | \boldsymbol{\theta}).$$

The MLE of  $\boldsymbol{\theta}$  is the value  $\hat{\boldsymbol{\theta}}$  that maximizes this distribution function with the  $(x_1, \dots, x_n)$  fixed.

There are also “generalized method of moments” estimators, but we wont discuss them here.

**def'n:** Let

$$\mathbf{X} = (X_1, \dots, X_n) \sim f(x_1, \dots, x_n | \boldsymbol{\theta}).$$

The MLE of  $\boldsymbol{\theta}$ , denoted  $\hat{\boldsymbol{\theta}}$ , is the value of  $\boldsymbol{\theta}$  which makes  $f(x_1, \dots, x_n | \boldsymbol{\theta})$  as large as possible keeping the  $(X_1, \dots, X_n) = (x_1, \dots, x_n)$  fixed.

**def'n:** The *standard error* of an estimator  $\hat{\theta}$ , denoted  $\text{se}(\hat{\theta})$  is an estimate of the unknown standard deviation of the estimator  $\text{sd}(\hat{\theta})$  obtained by plugging in the MLE or MOM estimate  $\hat{\theta}$ .

**Examples:**

For exponential data,

$$\text{se}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}^2}{n}} = \frac{1}{\bar{X}_n \sqrt{n}}.$$

For Bernoulli data,

$$\text{se}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} = \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}}.$$

We often have the important result

$$\hat{\theta} \bullet \sim N(\theta, \text{se}(\hat{\theta})).$$

## Models I: One-sample Bernoulli data

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

Examples:

- Polling for Obama versus McCain. Who will win? With what ‘confidence?’
- Testing that proportion of those taking new medication that get side effects is less than 1%.

We know both MLE and MOM of  $\pi$  is  $\hat{\pi} = \bar{X}_n$  which is the sample proportion of ‘events’ out of the  $n$  sampled.

The CLT tells us immediately that

$$\hat{\pi} \overset{\bullet}{\sim} N(\pi, \pi(1 - \pi)/n).$$

Note that exactly,

$$n\hat{\pi} = X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n, \pi),$$

but this is harder to work with. However, this forms the basis of exact tests for  $\pi$ .

The odds of success are  $O = \pi/(1 - \pi)$ . The MLE of the odds is just  $\hat{O} = \hat{\pi}/(1 - \hat{\pi})$ . What is the approximate distribution of the odds?

Let's use the delta method. Let  $g(x) = x/(1 - x)$ . Then  $g'(x) = 1/(1 - x)^2$ . So then

$$\hat{O} \underset{\bullet}{\approx} N \left( O, \frac{\pi(1 - \pi)}{(1 - \pi)^4 n} \right) = N \left( O, \frac{\pi}{(1 - \pi)^3 n} \right).$$

The variability 'blows up' when  $\pi \approx 1$ , i.e. for large odds.

Often the log-transform is used to make estimation more symmetric – i.e. less variable for  $\pi \approx 1$ . Let's see how this works.

The MLE of  $\log O$  is  $\log \hat{O} = \log \hat{\pi} - \log(1 - \hat{\pi})$ . The derivative of  $g(x) = \log x - \log(1 - x) = 1/[x(1 - x)]$ . So then

$$\log \hat{O} \overset{\bullet}{\sim} N \left( \log O, \frac{\pi(1 - \pi)}{\pi^2(1 - \pi)^2 n} \right) = N \left( \log O, \frac{1}{\pi(1 - \pi)n} \right).$$

The variability blows up for either  $\pi \approx 0$  or  $\pi \approx 1$ , but not as fast as before. This should be valid for a larger range of  $\pi$  values and is actually used to construct confidence intervals for the odds. More on this later.

**Example:** 13 out of 20 respondents reported that they went dancing at some point within the last year. So  $\hat{\pi} = \frac{13}{20} = 0.65$  is our MLE and MOM estimate of the unknown population proportion. We estimate the variability of the estimator with the *standard error*

$$\text{se}(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} = \sqrt{\frac{0.65(1 - 0.65)}{20}} = 0.107.$$

The CLT tells us, roughly,

$$\hat{\pi} \overset{\bullet}{\sim} N(\pi, 0.107^2),$$

before the experiment took place.

The odds of having gone dancing are estimated to be  $\hat{O} = 0.65/0.35 = 1.86$ . The standard error of the log-odds is

$$\text{se}(\log \hat{O}) = \sqrt{\frac{1}{0.65 \times 0.35 \times 20}} = 0.47.$$

Later we'll see this implies a 'plausible interval' for  $\log O$  of  $\log \hat{O} \pm 1.96(0.47)$ , i.e.  $0.62 \pm 1.96(0.47)$  giving the interval  $(-0.30, 1.54)$ . Exponentiating gives an interval for the odds (not the log-odds), which is  $(0.74, 4.7)$ .

## Old homework problem:

Large sample properties of MLE and MOM estimator  $\hat{\theta} = 1/\bar{X}_n$  for

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

1. The R source code `MLE1a.txt` creates histograms of  $m = 10000$  MLE's  $\hat{\theta}_1, \dots, \hat{\theta}_{10000}$  from *iid*  $\exp(10)$  data. Run it from the R command prompt as, for example,

```
> source("c:/biostat/MLE1a.txt").
```

2. As the sample sizes  $n$  going into each MLE  $\hat{\theta}$  increases from 5 to 25 to 125 to 625, what in general happens to the distribution of  $\hat{\theta}$ ? Does it concentrate around the true parameter value  $\theta = 10$ ? Does it become more normal looking?
3. The large sample result for MLE's says  $\hat{\theta} \sim N(\theta, \theta^2/n)$ . Since we know  $\theta = 10$ , let's superimpose a  $N(10, 100/n)$  curve on top of each histogram and see how well the large sample result works. The is code is in `MLE1b.txt`.