**Stat 200: Introduction to Statistical Inference** **Winter 2023**

# Lecture 4: Moment wrap-up and likelihood beginning

*Lecturer: Dominik Rothenhäusler* *January 23*

**Note**: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

**Abstract**

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

## 4.1 Delta method

In the last notes we had the ***delta method***. Rice has this delta method in Chapter 4.6 "approximate methods" where it is called "propagation of error or $\delta$ method". Let $\hat{\theta} = g(\overline{X})$ where $\overline{X}$ has mean $\mu$ and variance $\sigma^2/n$. Expanding to first order and taking the variance yields

$$\hat{\theta} \approx g(\mu) + g'(\mu)(\overline{X} - \mu) \quad \text{and} \quad \mathrm{Var}(\hat{\theta}) \approx \mathrm{Var}(\overline{X}) \times g'(\mu)^2. \tag{4.1}$$

We used (4.1) in class to get the variance of the method of moments. Expanding to second order and taking the mean yields

$$\hat{\theta} \approx g(\mu) + g'(\mu)(\overline{X} - \mu) + \frac{1}{2}g''(\mu)(\overline{X} - \mu)^2 \quad \text{and} \quad \mathbb{E}(\hat{\theta}) \approx g(\mu) + \frac{1}{2}\mathrm{Var}(\overline{X}) \times g''(\mu). \tag{4.2}$$

From (4.2) we can see that the method of moments has a bias:

$$\mathbb{E}(\hat{\theta}) - \theta = \mathbb{E}(g(\bar{X})) - g(\mu) \approx \Big(g(\mu) + \frac{1}{2}\mathrm{Var}(\bar{X}) \times g''(\mu)\Big) - g(\mu) = \frac{\sigma^2}{2n}g''(\mu)$$

where $\mathrm{Var}(X_i) = \sigma^2$. So the bias becomes small as $n$ increases.

## 4.2 Some moment facts

You could get $\hat{\theta} = \bar{X} < 0$ in a problem where $\theta \geqslant 0$ is known. This might happen if for instance $X_i \sim N(\theta, \sigma^2)$ and we know $\theta \geqslant 0$ scientifically. In some more complicated settings, the method of moments can give a negative estimate of a parameter representing a variance. Also if $\theta \in \{0, 1, 2\}$ (number of copies of a certain gene that your puppy got) and for some $X$, $\mathbb{E}(X_i) = 10\theta$, then the moment estimator $\bar{X}/10$ might not be an integer.

On the plus side, suppose that you care about $\mathbb{E}(X)$. Then pick a model $X \sim f(x; \theta)$. Suppose that for that model, the method of moments estimator is $\hat{\theta} = \bar{X}$. If it turns out that you were very wrong about the distribution family $f$ you still have a consistent estimate of $\mathbb{E}(X)$.

## 4.3 Maximum likelihood estimation

Suppose that $X_i$ are IID $f(x;\theta)$ for $\theta \in \Theta$. Here $\Theta$ has all the 'legal' values of $\theta$. For instance in a Poisson model $\Theta = [0,\infty)$. There are some values of $\theta$ that make our data quite unusual. Others make our data more probable. The method of maximum likelihood involves picking whatever $\theta$ makes the given data most probable. This value is called the ***maximum likelihood estimate*** (MLE).

The likelihood function is

$$L(\theta) = \Pr(X_1 = x_1, \ldots, X_n = x_n; \theta).$$

Sometimes it is written $L(\theta; x_1, \ldots, x_n)$. What we are doing here is taking the same probability function but making it a function of $\theta$ for fixed $x_i$ instead of making it a function of $x_1, \ldots, x_n$ for fixed $\theta$. To keep track of the change we call it the ***likelihood function*** (of $\theta$) instead of the probability function (of $x_1, \ldots, x_n$).

Now consider independent $X_i \sim \mathrm{Poi}(\theta)$. We get

$$L(\theta) = \prod_{i=1}^{n} \Pr(X_i = x_i; \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!},$$

and we want to maximize it over $\theta \in \Theta = [0,\infty)$. The first thing we do here is work out $\ell(\theta) = \log(L(\theta))$, called, reasonably enough, ***the log likelihood***. If we maximize $\ell(\theta)$ then the same $\theta$ maximizes $L(\theta)$. Taking this step

$$\ell(\theta) = \sum_{i=1}^{n} -\theta + x_i \log(\theta) - \log(x_i!).$$

For IID data the log likelihood gives us a sum. We know a lot about sums from the LLN and CLT and we will use them.

The worst looking part of $\ell(\theta)$ is $\log(x_i!)$. But that does not involve $\theta$ so for our purposes it is just a constant. We can write

$$\ell(\theta) = c - n\theta + \log(\theta) \sum_{i=1}^{n} x_i$$

where $c$ takes up all of the $\log(x_i!)$s. We commonly find that some awkward part of a model ends up not depending on $\theta$ and can then be replaced by $c$.

To maximize $\ell$, we find the first derivative

$$\ell'(\theta) = -n + \frac{1}{\theta} \sum_{i=1}^{n} x_i.$$

Setting $\ell'(\theta) = 0$ produces

$$\theta = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

We will check details later, but for now lets suppose that the MLE is $\hat{\theta} = \bar{x}$. This $\bar{x}$ is just the observed value of $\bar{X}$ and we write $\hat{\theta} = \bar{X}$.

To check details note that if $\bar{x} > 0$ then $\ell''(\bar{x}) = -n\bar{x}^{-1} < 0$. Then $\bar{x}$ is at least a local maximum. If $\bar{x} = 0$ then it is clear that the MLE is 0 which is $\bar{x}$. It remains to consider the possibility that $\ell(\theta)$ reaches its maximum in the limit as $\theta \to \infty$. That does not happen. So $\hat{\theta} = \bar{X}$ is the MLE.

If $X_i$ have PDF $f(x;\theta)$ then $\Pr(X_i = x_i; \theta) = 0$ for any $\theta$. Then $L(\theta) = 0$ which is not very useful. Upon reflection we realize that $X_i$ is never measured to infinite precision. Instead we observe something like

$x_i \leqslant X_i \leqslant x_i + \mathrm{d}x_i$. Now

$$\Pr(x_i \leqslant X_i \leqslant x_i + \mathrm{d}x_i; \theta) = \int_{x_i}^{x_i+\mathrm{d}x_i} f(x;\theta)\,\mathrm{d}x \approx f(x_i)\mathrm{d}x_i.$$

The likelihood is then

$$L(\theta) \approx \prod_{i=1}^{n} f(x_i;\theta)\mathrm{d}x_i$$

and then the log likelihood is

$$\ell(\theta) \approx c + \sum_{i=1}^{n} \log(f(x_i;\theta))$$

where this $c$ hides all the $\log(\mathrm{d}x_i)$ terms which we quite reasonably assume do not depend on $\theta$. [That would be weird.] We will drop the $\approx$ from our continuous models and just work with $\sum_i \log(f(x_i;\theta))$.

## 4.4 More examples

For $X_i \sim N(\mu, \sigma^2)$ we found $\ell(\mu, \theta)$. Differentiating with respect to $\mu$ lead to $\hat{\mu} = \bar{X}$. Differentiating with respect to $\sigma$ lead to

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2,$$

the sample variance. These are the same as the moment estimates. That also happened for the Poisson. It does not always happen.

Here is a capture-recapture example. A lake has $N$ fish. This $N$ is our unknown parameter of interest. We sample $n$ fish and tag them. Later we sample $m$ fish and notice that $x$ of them have tags. We could reason in a method of moments style way that since $x/m$ of the later fish have tags that a proportion $x/m$ of the fish in the lake have tags. That proportion is $n/N$. So we could set $n/N = x/m$ and get $\hat{N} = nm/x$ which might not be an integer. Or we could use maximum likelihood. Let us assume that any subset of $m$ fish has the same chance of being caught in our net. Then we get a hypergeometric probability

$$L(N) = \Pr(X = x; N) = \frac{\binom{n}{x}\binom{N-n}{m-x}}{\binom{N}{m}}.$$

Suppose that $Y_i \sim f(y;\theta)$, but if $Y_i \geqslant t_i$ then we don't see $Y_i$ we only learn that $Y_i \geqslant t$. This setting commonly arises when $Y_i$ is the time to an event and the event has not happened yet by time $t$. This has medical applications (e.g., patient survival) and industrial applications (lifetime of a part). All is not lost here. The probability that $Y_i \geqslant t$ is $1 - F(t;\theta)$ where $F$ is the CDF. We can place that knowledge in the likelihood

$$L(\theta) = \prod_{i \mid Y_i \text{ observed}} f(y_i;\theta) \times \prod_{i \mid Y_i \text{ not observed}} (1 - F(t;\theta)),$$

assuming independence in order to get those products. Indicator notation helps us a lot. Let $\delta_i = 1$ if $Y_i$ was observed and $\delta_i = 0$ otherwise. Then because raising a nonzero number to the power 0 gives 1, we have

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta)^{\delta_i}(1 - F(t;\theta))^{1-\delta_i}.$$

In class we had $f(y;\theta) = \theta^{-1}e^{-y/\theta}$ on $0 < y < \infty$. Then $F(t;\theta) = 1 - e^{-t/\theta}$. This is the exponential distribution with mean $\theta$. We worked out the MLE of $\theta$ by differentiating $\ell(\theta)$.

The phenomenon above is called 'censoring'. Instead of seeing $Y$ all you are told is that $Y > t$. Another kind of censoring arises when a pollutant $Y$ is below a minimum detectable quantity $\Delta$. We know it wasn't negative so we know that $0 \leqslant Y \leqslant \Delta$ and that observation contributes a factor of $\Pr(0 \leqslant Y \leqslant \Delta; \theta)$ to the likelihood. For a continously distributed $Y$ this probability is $F(\Delta;\theta) - F(0;\theta)$. Since that $Y$ cannot be negative $F(0;\theta) = 0$ and the likelihood contribution is $F(\Delta;\theta)$.

We also considered an historical example of French military recruits' heights. If the populations heights are $H_i \sim N(\mu, \sigma^2)$ but only people with $H_i \geqslant \tau$ were accepted then the data won't have anybody shorter than $\tau$ and we miss that part of the population. It would be fine if we cared about the recruits' height but not fine if we cared about the population from which they were drawn. We can put this rule into our likelihood if we can figure out the distribution of $X_i$ where $X_i$ is the height of a recruit. Let's get the CDF of $X$. It is

$$F(x;\mu,\sigma) = \begin{cases} 0, & x < \tau \\ \Pr(X \leqslant x \mid H \geqslant \tau; \mu, \sigma), & x \geqslant \tau. \end{cases}$$

The reasoning is as follows. First, if $H < \tau$ then the person is not recruited. We never see $X < \tau$ and so for sure $F(x) = 0$ when $x < \tau$. Next for $x > \tau$, the fraction of recruits with $X \leqslant x$ is the fraction of people with $H > \tau$ that also satisfy $X \leqslant x$. Now for $x > \tau$,

$$\begin{aligned} \Pr(X \leqslant x \mid H \geqslant \tau; \mu, \sigma) &= \frac{\Pr(H \leqslant x \;\&\; H \geqslant \tau; \mu, \sigma)}{\Pr(H \geqslant \tau; \mu, \sigma)} \\ &= \frac{\Pr(\tau \leqslant H \leqslant x; \mu, \sigma)}{\Pr(H \geqslant \tau; \mu, \sigma)} \\ &= \frac{\int_\tau^x f(y;\mu,\sigma)\,\mathrm{d}y}{\int_\tau^\infty f(y;\mu,\sigma)\,\mathrm{d}y} \end{aligned}$$

where $f(y;\mu,\sigma)$ is the $N(\mu,\sigma^2)$ PDF. We just got the CDF of $X$ but for a likelihood we want the PDF. We use $f$ for the $N(\mu,\sigma^2)$ PDF of $H$. Lets use $g(x;\mu,\sigma^2)$ for the PDF of $X$.

Differentiating with respect to $x$ we get

$$g(x;\mu,\sigma) = \begin{cases} \dfrac{f(x;\mu,\sigma)}{\int_\tau^\infty f(y;\mu,\sigma)\,\mathrm{d}y}, & x \geqslant \tau \\ 0, & \text{else.} \end{cases}$$

The likelihood is then

$$L(\mu,\sigma) = \prod_{i=1}^n \frac{f(x_i;\mu,\sigma)}{\int_\tau^\infty f(y;\mu,\sigma)\,\mathrm{d}y}.$$

It then requires numerical work to get MLEs.

The situation with heights is different from the one about minimal detection. If a height would have been below $\tau$ then we would not have even known about that individual. If a chemical is measured below $\Delta$ then we know there was a value and we know it was in the interval $[0, \Delta]$. The distinction in our knowledge is captured by the different forms the likelihood functions take.