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

Lecture 5: Maximum likelihood theory

Lecturer: Dominik Rothenhäusler *January 25*

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.

Today's lecture was on some of our technically hardest material. We used a small number of tricks repeatedly. Thursday's lecture will be on Bayesian methods. That is some of the conceptually hardest material.

5.1 Some likelihood quick facts

No illegal values

The method of moments could give $\hat\theta < 0$ when we know $\theta \geqslant 0$. The MLE never gives an illegal value like that. For instance if we know that $0 \leqslant \theta \leqslant 1$ then we maximize $L(\theta)$ over the interval $0 \leqslant \theta \leqslant 1$. Legality is baked into the problem by definition. Sometimes this is called ***range respecting*** property of the MLE. If ρ is a correlation you never get $\hat\rho < -1$ or $\hat\rho > 1$. If σ^2 is a variance you never get $\hat\sigma^2 < 0$.

Transformations

If $\hat\theta$ is the maximum likelihood estimate of θ and g is some function of θ, then $g(\hat\theta)$ is the maximum likelihood estimate of $g(\theta)$. For instance the MLE of θ^3 is $\hat\theta^3$. The MLE of $\sin(\theta)$ is $\sin(\hat\theta)$, the MLE of $|\theta|$ is $|\hat\theta|$, the MLE of $1_{\theta>0}$ is $1_{\hat\theta>0}$ and so on.

Why is it so easy? Think of the most probable explanation θ that has $g(\theta) = \eta$. Then maximize that probability over all η. Letting η be anything at all means allowing θ to take any legal value. So $\hat\eta = g(\hat\theta)$.

5.2 Fisher information and the Cramer-Rao bound

In this material, we get a lot done based on the CLT, the LLN and a bunch of manipulation of expectations and derivatives. It is well worth reading the sections of Rice Chapter 8 multiple times to make sure that you get it. We illustrate for continuous random variables but the same holds for discrete random variables with probability mass function $p(x;\theta)$.

Let X, X_1, X_2, $\cdots$, X_n be IID random variables with PDF $f(x;\theta)$. The Fisher information in X for θ is

$$I(\theta) = \mathbb{E}\Big(\big[\frac{\partial}{\partial\theta}\log f(X;\theta)\big]^2\Big).$$

It is $\geqslant 0$ by construction. We see later why it is called 'Information'. Staring into it we see it is an expected squared slope of log likelihood. If the slope is large then small changes in θ change the log likelihood a lot. That should help separate likely from unlikely values. If that slope were zero then we get no effect of changing θ. This intution does not replace the formal derivations but it helps to see why they turn out as they do.

There are three different kinds of θ in this setting. First there is θ the parameter which has a range of legal values. Then there is θ_0. When we need to single out the one true value of θ it is θ_0. In practice we don't know which value is true. Our MLE is $\hat{\theta}$.

In class we proved the three things. First under smoothness conditions on f,

$$I(\theta) = -\mathbb{E}\Big(\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\Big). \tag{5.1}$$

Those conditions are described in Section 5.4. We can think of this version of $I(\theta)$ in terms of curvature. If $\ell(\theta)$ has large (negative) curvature then it drops off fast as we move away from the MLE $\hat{\theta}$. Ou ℓ is an average of n observed $\log(f(X_i;\theta))$ so it will be like (minus) the expectation above. Picture the transamerica pyramid and then picture mount Fuji. That pyramid is pointier so the location of the peak is more sharply defined, hence more informative. Again, this is intuition. The details are in the math. If a homework or exam question asks for a mathematical proof the intution is not the answer, though it might help you check whether your answer makes sense.

We proved this result by writing $\int f(x;\theta)\,\mathrm{d}x = 1$ and differentiating two times with respect to θ. We pulled the derivative past the integral both times. We also used

$$\frac{\partial}{\partial\theta}f(x;\theta) = \frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)} \times f(x;\theta) = \frac{\partial}{\partial\theta}\log(f(x;\theta)) \times f(x;\theta).$$

That is non-intuitive at first. The point of it was to turn an integral expression back into an expectation. Then we get to use the rules of expectation. It also gave us an expression with log f in it. Re-read Rice's derivation to see that method used.

The first time we took the derivative we found that

$$0 = \int \frac{\partial}{\partial\theta}\log(f(x;\theta))f(x;\theta)\,\mathrm{d}x \quad \text{i.e.,} \quad \mathbb{E}\Big(\frac{\partial}{\partial\theta}\log(f(X;\theta))\Big) = 0. \tag{5.2}$$

So we found that this expected derivative has mean zero.

The second thing we proved was that for large n

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \approx N(0,1).$$

We can interpret this colloquially as

$$\hat{\theta} \approx N\Big(\theta_0, \frac{1}{nI(\theta_0)}\Big).$$

Now we see why it is called information. More information means less variance, so we have learned θ with greater accuracy. By the same token, more variance means less information.

To do this one, we used Taylor expansion, just at Rice does and arrive at

$$\hat{\theta} - \theta_0 \approx -\frac{\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i;\theta_0)}{\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(x_i;\theta_0)}.$$

Inserting the $1/n$'s in numerator and denominator gives us averages. We use the LLN on the denominator and the CLT on the numerator. By the LLN and (5.1), the denomator converges to $-I(\theta_0)$. We could apply the LLN to the numerator but we would get 0 (from (5.2)). That's good. It means that $\hat\theta$ is getting close to the true value θ_0. We want a sharper answer though. The variance of $\frac{\partial}{\partial\theta}\log f(X_i;\theta_0)$ is $I(\theta_0)$. So the CLT gives us a numerator approximately $N(0, I(\theta_0)/n)$. Now

$$-\frac{\text{NUM.}}{\text{DEN.}} \approx \frac{N(0, I(\theta_0)/n)}{I(\theta_0)} = N\Bigl(0, \frac{I(\theta_0)/n}{I(\theta_0)^2}\Bigr) = N\Bigl(0, \frac{1}{nI(\theta_0)}\Bigr).$$

The derivations above are at the same formal level as the ones in Rice. If you take a more advanced theory class you'll see more rigorous derivations.

The third thing was the Cramer-Rao lower bound. A statistic $T = T(X_1, \dots, X_n)$ is unbiased for θ if $\mathbb{E}(T) = \theta$. By this we mean $\mathbb{E}(T;\theta) = \theta$. It has to hold for all legal values of θ not just one of them. Notice the notation: the random variable T is random because it depends on $X_1, \dots, X_n$. We can use T by itself to get shorter expressions. Or we can use the full $T(X_1, \dots, X_n)$ in places where we want to remind ourselves about the randomness in T coming from the X_i values.

The Cramer-Rao lower bound is as follows. If T is an unbiased statistic constructed from IID observations $X_1, \dots, X_n$ with PDF $f(x;\theta)$ then

$$\text{Var}(T(X_1, \dots, X_n); \theta) \geqslant \frac{1}{nI(\theta)}.$$

There are two takeaways from this result. First it means that no unbiased estimator can be better, as measured by variance than $1/(nI(\theta))$. It is not a matter of waiting on new methodology like a better algorithm or for a statistical super hero to appear on the scene. No possible unbiased statistic can be better than this. There is a limit to how much information is in the data and you cannot get more information out of it than it contains. There is $I(\theta_0)$ information in one observation and we get $nI(\theta_0)$ information from n IID observations.

The second takeway is the the MLE is (for large n) pretty good. It is almost $N(\theta_0, 1/(nI(\theta_0)))$.

What could beat the MLE? A biased estimator could have less variance and maybe come out better overall. We will see that with some Bayesian estimators. The MLE might also have trouble with small n. The LLN and CLT are about large n.

The proof of the Cramer-Rao bound introduced the score statistic

$$S = \sum_{i=1}^n \frac{\partial}{\partial\theta}\log f(x_i;\theta),$$

which we know has mean 0. We started with $\theta = \mathbb{E}(T;\theta)$, differentiated both sides with respect to θ and after some manipulations (also in Rice) got

$$1 = \mathbb{E}(T \times S) = \text{Cov}(T, S) + \mathbb{E}(T)\mathbb{E}(S) = \text{Cov}(T, S) \leqslant \sqrt{\text{Var}(T)\text{Var}(S)}.$$

The last inequality is because the correlation between T and S can be at most 1. That does it.

5.3 Examples

Suppose that $X_i \sim N(\mu, \sigma^2)$ and we actually know σ but not μ. Let's find the Fisher information. Let's do it in small steps because plugging everything in to the formula at once could be unweildy. First

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}.$$

So

$$\log(f(x;\mu)) = -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(x-\mu)^2.$$

The parameter of interest, usually θ, is μ in this example. We differentiate and get

$$\frac{\partial}{\partial\mu}\log(f(x;\mu)) = -\frac{1}{2\sigma^2}\times 2(x-\mu)(-1) = \frac{x-\mu}{\sigma^2}.$$

At this point we could use square this derivative and take its expectation. That would be

$$I(\mu) = \mathbb{E}\Bigl(\Bigl(\frac{X-\mu}{\sigma^2}\Bigr)^2\Bigr) = \frac{\mathbb{E}((X-\mu)^2)}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

This example reinforces the pattern where information is inverse to variance.

Of course we could also have taken the second derivative and worked out minus the expectation. Let's do it that way. The second derivative is

$$\frac{\partial}{\partial\mu}\frac{x-\mu}{\sigma^2} = -\frac{1}{\sigma^2}.$$

That is not even random, so its expectation is easy to get. It is just $-1/\sigma^2$. So minus this expectation is $1/\sigma^2$, same as we had before.

We have these two formulas to choose from. One uses a second moment of a first derivative and the other uses a first moment of a second derivative. Either one might prove to be easier.

If $X \sim \mathrm{Poi}(\lambda)$ then $p(x;\lambda) = e^{-\lambda}\lambda^x/x!$ for $x = 0, 1, \dots$. Then

$$\log(p(x;\lambda)) = -\lambda + x\log(\lambda) - \log(x!).$$

The first derivative is $-1 + x/\lambda$. The second derivative is $-x/\lambda^2$. Either way we get Fisher information $1/\lambda$.

5.4 Smoothness

We passed the derivative under the integral repeatedly. Suppose that we have a function $h(x,\theta)$ where, in our likelihood problems, h might be $f(x;\theta)$ or that f times some other things involving x and θ.

Here is Liebniz's rule. For finite $a(\theta) < b(\theta)$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{a(\theta)}^{b(\theta)} h(x,\theta)\,\mathrm{d}x = h(b(\theta),\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}b(\theta) - h(a(\theta),\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\mathrm{d}h(x,\theta)}{\mathrm{d}\theta}\,\mathrm{d}x.$$

Our first regularity condition is that the set of x with $f(x;\theta) > 0$ does not depend at all on θ. That rules out $U(0,\theta)$ for instance. Then the expectations we needed are integrals over x between values a and b that don't depend on θ. In that case, we get

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_a^b h(x,\theta)\,\mathrm{d}x = \int_a^b \frac{\mathrm{d}h(x,\theta)}{\mathrm{d}\theta}\,\mathrm{d}x.$$

In some of our problems $a = -\infty$ and/or $b = \infty$. Then an additional technical condition is required to differentiate under the integral sign.

For this course we will check that the range of legal x values must not depend on θ for our likelihood theorems to apply, but we assume otherwise that our derivatives can be passed under the integral.

5.5 Sufficiency

This is in Rice Ch 8.8.

A statistic $T = T(X_1, \dots, X_n)$ is sufficient for θ if the distribution of $X_1, \dots, X_n$ given $T = t$ does not depend on θ.

We could imagine the data arriving in two steps. First we are told that $T = t$. Then we are told that $X_1 = x_1, \dots, X_n = x_n$ for some values x_i where $T(x_1, \dots, x_n) = t$. If T is sufficient, then it means the second step gives us no more information about θ. It is just a random variable unrelated to θ.

For an extreme version suppose that step one gives us $X_1 \sim N(\mu, 1)$ and then step two gives us $X_2 \sim N(0, 1)$. Because X_2 has a distribution that does not depend on μ it tells us nothing about μ. Everything we would learn about μ from (X_1, X_2) is in X_1 and none of it is in X_2.

In class we saw that if $X_i \sim N(\mu, 1)$ then $\bar{X}$ is sufficient for μ. Instead of keeping all n X values we could just keep $\bar{X}$.

We saw the example of coin tosses (that Rice does too). The total number of Heads observed is sufficient. If there are 3 heads in 4 tosses learning whether it was $HHHT$, $HHTH$, $HTHH$ or $THHH$ tells us nothing more about the coin's probability θ.

There is a factorization theorem that tells us about sufficiency. If we can write

$$f(x_1, x_2, \dots, x_n; \theta) = g\bigl(T(x_1, \dots, x_n), \theta\bigr) \times h(x_1, \dots, x_n)$$

for some functions g and h, then T is sufficient for θ. See the proof in Rice. We used that to prove that $\bar{X}$ is sufficient for μ in the $N(\mu, 1)$ example. We also saw there that replacing $(x_i - \mu)^2$ by $(x_i - \bar{X} + \bar{X} - \mu)^2$ is a useful tactic. When we sum over i we get $\sum_i (X_i - \bar{X}) = 0$