Stats I review

J. Alexander Branham

Fall 2015

This document is intended as a review guide for a few topics we covered in the second half of the semester in Statistics I, Fall 2015.

1 MLE estimation

Maximum likelihood estimation (MLE) is a way of getting an estimator. In particular, MLE asks "What's the value for this parameter that makes my data the most likely to have occurred?" In order to get this, all we need to do is to write out the likelihood function then find its maximum. Oftentimes, we'll take the log of the likelihood function before finding the maximum because taking the derivative of the log of the likelihood function is oftentimes easier. It will give you the same value, though.

1.1 Step 1: Find the likelihood function

We can find the likelihood function $L(\theta)$ by simply multiplying together the probability of each X_i, if they're independent (which is what we generally assume). That's simple to do:

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

1.1.1 Step 1b

Mathematically, the product from the previous equation becomes difficult to work with for a variety of reasons. Therefore, we oftentimes take the log of the likelihood function which turns the product into summation. Note that for very simple examples, this isn't necessary. For more complicated examples, this makes your life easier. It is also how computers calculate MLE.

$$\log(L(x|\theta)) = \mathcal{L}(x|\theta) = \sum_{i=1}^{n} f(x_i|\theta) \tag{2}$$

Note that this step isn't absolutely necessary. It just makes the math easier (usually) and will always give the same values as if you didn't take the log.

1.2 Step 3: Find the maximum of the log-likelihood

Now that we have formally specified the likelihood of our data in terms of an unknown θ, we can find the value for θ that maximizes the likelihood of our data. We could do this by hand, plugging in all of the possible values for θ. But that would take a while and be a lot of work, so we can use *optimization* instead. This is the process of finding the maximum (or minimum) of a function.

Mechanically, this is pretty simple. We simply take the derivative of the (log) likelihood and set it equal to zero. To ensure we've found a maximum (instead of a minimum), we also need to check the second derivative to make sure it's negative.

1.3 Example

Suppose that X is a Bernoulli random variable and we observe 183 0's and 78 1's. What is the MLE for p?

1.3.1 Find likelihood function

If we let k represent the number of successes we have, then the likelihood function is then:

$$\begin{aligned} L(X|p) &= \prod_{i=1}^{n} Pr(X = x_i|p) \\ &= p^k(1-p)^{n-k} \end{aligned} \tag{3}$$

Which for our data is simply $L(X|p) = p^{78}(1-p)^{261-78}$

At this point, if the likelihood function looks like it's going to be a beast to maximize, you can take the log of it to make the math easier. This one won't give us a problem, though.

We can look at the likelihood function to see where we think our estimate might be. We can see from Figure 1 that the maximum likelihood of our data occurs somewhere around $\theta = 0.26$.

```
our_like<- function(p){
  p ^ 78 * (1 - p) ^ (261 - 78)
}

ggplot2::qplot(c(0,1), stat = "function",
               fun = our_like,
               geom = "line")
```

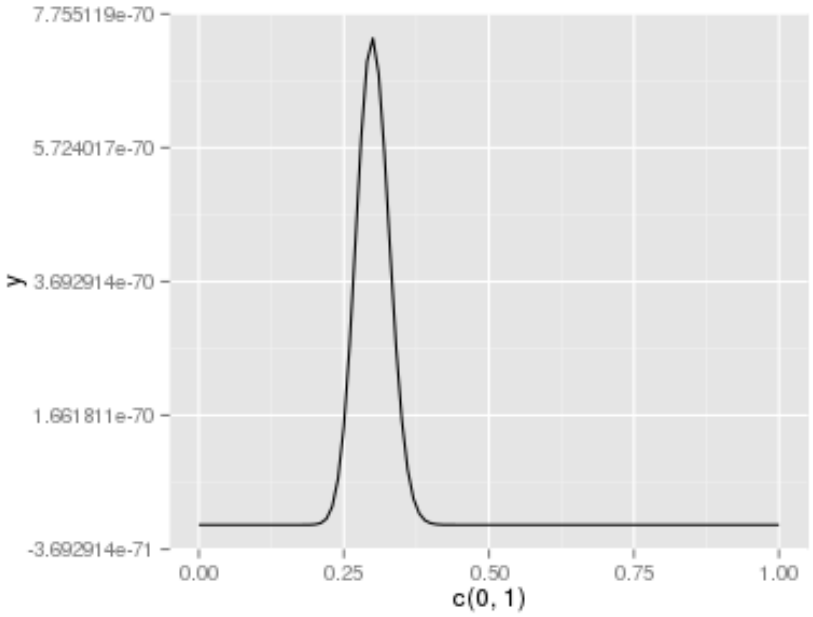

Figure 1: Graph of the log likelihood for varying thetas

1.3.2 Obtaining the MLE estimate

Although from the previous graph it's pretty obvious that the MLE estimate will be somewhere around 0.26, it would still be nice to formally know this. We start by finding the derivative:

$$\frac{dL(X|p)}{dp} = kp^{k-1}(1-p)^{n-k} + p^k(n-k)(1-p)^{n-k-1}(-1) \tag{4}$$

We can then set the derivative equal to zero and solve.

$$\begin{aligned} 0 &= kp^{k-1}(1-p)^{n-k} + p^k(n-k)(1-p)^{n-k-1}(-1) \\ kp^{k-1}(1-p)^{n-k} &= p^k(n-k)(1-p)^{n-k-1} \\ k(1-p) &= p(n-k) \\ k - kp &= pn - pk \\ p &= \frac{k}{n} \end{aligned}$$

So here $\hat{p}_{MLE} = \frac{k}{n} = 78261 \approx 0.299$

1.4 Example 2

Suppose that X is a discrete random variable with the following probability mass function where $0 \leq \theta \leq 1$:

X	Pr(X)
0	$\frac{2\theta}{3}$
1	$\frac{\theta}{3}$
2	$\frac{2(1-\theta)}{3}$
3	$\frac{1-\theta}{3}$

We observe the following data: $(3, 0, 2, 1, 3, 2, 1, 0, 2, 1)$. What is the MLE of θ?

1.4.1 Find likelihood function

The likelihood function is then:

$$\begin{aligned} L(X|\theta) =& Pr(X=3)Pr(X+0)Pr(X=2)Pr(X=1)Pr(X=3) \\ & Pr(X=2)Pr(X=1)Pr(X=0)Pr(X=2)Pr(X=1) \end{aligned} \tag{5}$$

We can plug in from the pmf to find the probabilities:

$$L(X|\theta) = \left(\frac{2\theta}{3}\right)^2 \left(\frac{\theta}{3}\right)^3 \left(\frac{2(1-\theta)}{3}\right)^3 \left(\frac{1-\theta}{3}\right)^2 \tag{6}$$

Which is going to be a beast to maximize. So we'll follow the advice of 1.1.1 above and take the log of the function:

$$\begin{aligned}\mathcal{L}(X|\theta) &= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) \\ &\quad + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + 2\left(\log\frac{1}{3} + \log(1-\theta)\right) \\ &= C + 5\log\theta + 5\log(1-\theta)\end{aligned} \tag{7}$$

Where C is some constant that doesn't depend on θ. Taking the derivative of that will be much easier than the likelihood function above.

We can look at the log likelihood function to see where we think our estimate might be. We can see from Figure 2 that the maximum likelihood of our data occurs somewhere around $\theta = 0.5$.

```
our_log_like<- function(theta){
  5 * log(theta) + 5 * log(1 - theta)
}

ggplot2::qplot(c(0,1), stat = "function",
               fun = our_log_like,
               geom = "line")
```

1.4.2 Obtaining the MLE estimate

Although from the previous graph it's pretty obvious that the MLE estimate will be somewhere around 0.5, it would still be nice to formally know this. We start by finding the derivative:

$$\frac{d\mathcal{L}(X|\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} \tag{8}$$

We can then set the derivative equal to zero and solve.

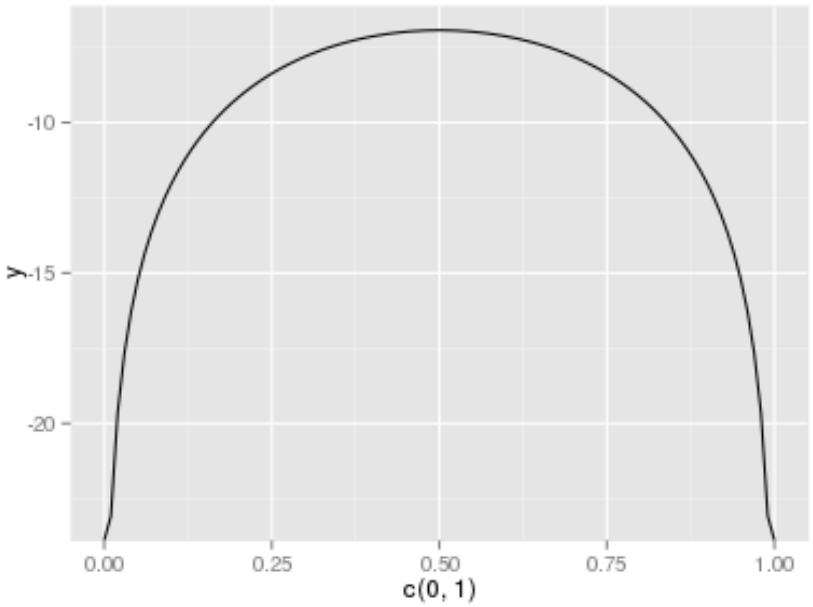

Figure 2: Graph of the log likelihood for varying thetas

$$
\begin{aligned}
0 &= \frac{5}{\theta} - \frac{5}{1-\theta} \\
\frac{5}{1-\theta} &= \frac{5}{\theta} \\
5\theta &= 5(1-\theta) \\
5\theta &= 5 - 5\theta \\
10\theta &= 5 \\
\hat{\theta}_{MLE} &= \frac{1}{2}
\end{aligned}
$$

2 MOM estimation

Method of Moments estimation (MOM) is another way of getting estimators, just like MLE. It asks a slightly different question to get these estimators, though. Whereas MLE find the value of the parameter(s) that make your data the most likely to have occurred, MOM simply states that your sample "moments" are good estimators of the theoretical moments.

The general way to find the MOM estimators are to find the first K theoretical and sample moments, where K represents the number of equations you have. You then set them equal to one another and solve for your estimators.

2.1 Find the theoretical moments

The theoretical moments are simple. They're just $E(X^k)$ where k represents the theoretical moment. So if you want the first theoretical moment, that's just $E(X^1)$, or just $E(X)$. The second theoretical moment is just $E(X^2)$ and so on for higher-order moments.

2.2 Find the sample moments

The sample moments are just as easy to find as the theoretical moments. The k^{th} sample moment is just

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \tag{9}$$

Note that the first sample moment is $\frac{1}{n}\sum_{i=1}^{n} X_i$, which is simply $\bar{x}$

2.3 Set these equal and solve

2.4 Example 1

Let $x_1, x_2, ...x_n$ be random draws from a uniform distribution with an unknown lower bound but an upper bound of 100 (i.e. $x_i \sim U(a, 100)$)

Then the pdf of this is:

$$f(x) = \begin{cases} \frac{1}{100-a} & a \leq x \leq 100 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Find the method of moments estimator for a.

2.4.1 Theoretical moments

We are estimating one parameter, so we only need to find the first theoretical moment. For a uniform, this is:

$$E(X) = \int_a^b \frac{x}{100-a} dx = \frac{a+100}{2} \tag{11}$$

2.4.2 Sample moments

Again, we just need to find the first one, which is simply:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{12}$$

2.4.3 Solve for the estimator

We set these equal and solve for the MOM estimator:

$$\begin{aligned} \bar{x} &= \frac{a+100}{2} \\ 2\bar{x} &= a+100 \\ 2\bar{x}-100 &= a \end{aligned}$$

So $a_{MOM} = 2\bar{x} - 100$.

2.5 Example 2

Let $x_1, x_2, ...x_n$ be random draws from a uniform distribution (i.e. $X \sim U(a,b)$) and we need to calculate both of the bounds(a and b). Remember that the pdf of a uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

2.5.1 Theoretical moments

Since we have two unknown parameters, we need to calculate the first two theoretical moments:

$$E(X) = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2} \tag{14}$$

$$E(X^2) = \int_a^b \frac{x^2}{b-a} dx = \frac{a^2+ab+b^2}{3} \tag{15}$$

2.5.2 Sample moments

We need to find the first two sample moments:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{16}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = m_2 \tag{17}$$

2.5.3 Set theoretical and sample moments equal and solve

Now we just set the theoretical moments and sample moments equal to each other and solve to find our estimators.

$$\bar{x} = \frac{a+b}{2} \qquad m_2 = \frac{a^2+ab+b^2}{3} \tag{18}$$

When we solve for a and b, we get that (after some nasty algebra):

$$\hat{a} = \bar{x} - \sqrt{3\left(m_2 - \bar{x}^2\right)} \qquad \hat{b} = \bar{x} + \sqrt{3\left(m_2 - \bar{x}^2\right)} \tag{19}$$

3 Significance & Power

There are two kinds of errors we can make in hypothesis testing. A Type I error is committed when we reject H_0 when H_0 is actually true. We make a Type II error when we fail to reject a false null. Table 3 nicely summarizes this relationship.

H_0	Decision	
	Reject	Fail to Reject
True	Type I	✓
False	✓	Type II

There's an obvious tradeoff here between the frequency with which we commit either kind of error. In the limit, if we never reject a null, then we'll never commit a Type I error, but we'll never reject a false null either. We can formally define the probability of committing either kinds of error.

3.1 Significance

$$\alpha = Pr(\text{Type I error}|H_0) \tag{20}$$

Thus, α represents the probability of making a Type I error if the null is actually true. We use α such that there is $(1-\alpha)$ probability of being inside the critical region is our null is true. If we see a test statistic outside that critical region, then we know there is a less than α percent chance that that would happen purely due to randomness if the null were actually true. This is the **significance** of a test.

3.2 Power

$$\beta = Pr(\text{Type II error}) \tag{21}$$

β, on the other hand, represents the probability of committing a Type II error. This is impossible to mathematically calculate most of the time, though. It's not enough just to say that the null isn't true - we need to specify what the true parameter is equal to in order to calculate β. We refer to $(1-\beta)$ as the **power** of a test. Usually, we'll look at how power varies as a function of unknown parameters or n.

3.2.1 Example

You've designed an experiment to test the effect of disgust on attitudes towards the incumbent. From the results of a pilot study, you believe that attitudes toward the incumbent are normally distributed with a mean of 50 and standard deviation of 6. You believe that your treatment will increase the mean by 4 points. How many participants do you need in order to detect this with 90 percent probability? Use a two-tailed test and $\alpha = 0.05$.

Note here that $H_0 : \mu = 50$ and $H_A : \mu \neq 50$. For this example, we'll assume that we know the standard deviation is 6. Relaxing that assumption is pretty straightforward, though. So note that under the null, our estimator $\bar{x} \sim N(50, \frac{6^2}{n})$ and that if our guess about the effect size is true, then $\bar{x} \sim N(54, \frac{6^2}{n})$.

We can calculate the critical values as a function of n:

$$50 \pm 1.96 \left(\sqrt{\frac{6^2}{n}} \right) \tag{22}$$

So we'll reject if we see a value lower than that when we subtract or great than that when we add. Now we just need to figure out the probability of

that happening if the true effect is to lower the mean by 6 points. That's pretty easy to do - we know that if we subtract off the mean and divide by the standard deviation, then we've standardized our variable and can look up probabilities using the standard normal table. So to find the probability of being *less* than the critical value, we:

$$\Phi\left(\frac{\left(50 - 1.96\left(\sqrt{\frac{6^2}{n}}\right)\right) - 54}{\sqrt{\frac{6^2}{n}}}\right) \tag{23}$$

And then we add that to the probability of being *greater* than our other critical value:

$$1 - \Phi\left(\frac{\left(50 + 1.96\left(\sqrt{\frac{6^2}{n}}\right)\right) - 54}{\sqrt{\frac{6^2}{n}}}\right) \tag{24}$$

So Equation 23 plus Equation 24 gives us the probability of rejecting the null hypothesis if the true mean is actually 54 instead of 50.

Now we can actually answer the question that we're interested in. We want to know the number of participants needed in order to detect this effect with a probability of 0.90.

```
our_power_test <- function(n){
  left <- pnorm(((50 - 1.96 * sqrt(6 ^ 2 / n)) - 54) / sqrt(6 ^ 2 / n))
  right <- 1 - pnorm(((50 + 1.96 * sqrt(6 ^ 2 / n)) - 54) / sqrt(6 ^ 2 / n))
  left + right
}

library(ggplot2)

ggplot(data.frame(n=c(0, 100)), aes(n)) +
  stat_function(fun = our_power_test) +
  geom_hline(yintercept = .9, linetype = "dashed")
```

So from Figure 3, we can see that we'd need about 25 people in order to detect this with 90 percent probability.

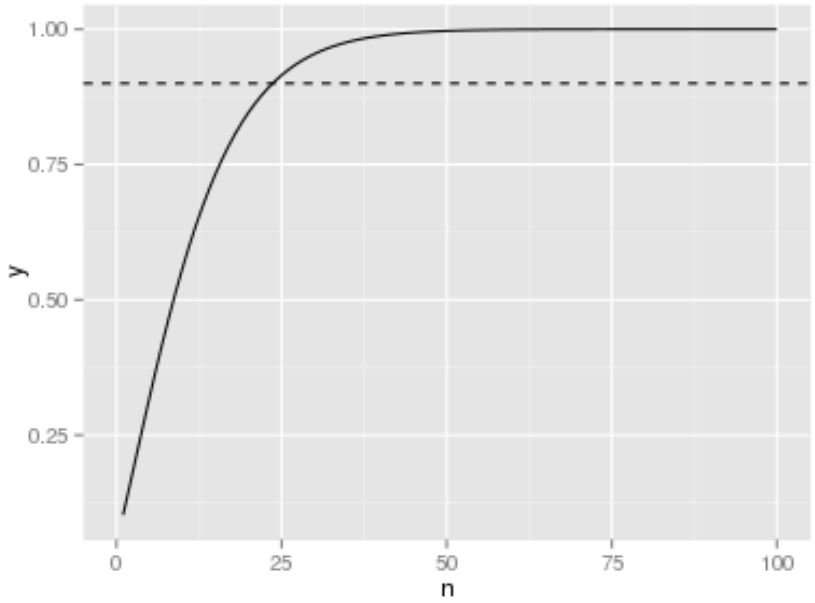

Figure 3: Power tests