MA 2B RECITATION 02/09/12

1. Introduction

Your midterms are now graded and available for pickup.

We're now moving onto statistics. It tends to have a slightly different flavor than probability, but we are approach statistics assuming that you know probability, unlike, say, your high school AP Statistics course. That's not to say that such knowledge is worthless, but we are approaching things from a slightly more sophisticated viewpoint.

2. Method of Maximum Likelihood/Maximum Likelihood Estimate

Suppose we're trying to measure the true value of some quantity x_T, like the average height of people in the United States. In many cases, like this example, it's impractical to try and find this quantity precisely. Instead, we would like to find a good approximation, which is usually good enough for our purposes.

One way to do this is just make repeated measurements of this quantity, call them $x_1, x_2, \ldots, x_n$. What is the most obvious way to estimate x_T given these measurements?

The most basic thing to do is to take the average, that is, set

$$x_T = \overline{x_i} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This is called the **sample mean.**

Does this procedure make sense? Sometimes yes, sometimes less so. For instance, it wouldn't help much if you're trying to measure the occurrence of a one-of-a-kind event, or if you don't know that your measurements are independent even identically distributed. However, before we learn the fancy stuff, we need to master the basic case.

When viewed from 10,000 feet, the Method of Maximum Likelihood is just a general method for estimating parameters of interest based on data. The way we estimate this is by finding a "likelihood" function that takes in parameters given some observed outcomes and spits out the probability of those observed outcomes given our parameter values. We then try and tweak our parameters to "fit" our data and get the parameters that provide the "maximum likelihood" of those outcomes occurring. It's most illustrative to see an example.

Example 1 (Discrete)**.** As usual, the discrete case is most intuitive. Suppose that you know that you have $X_1, \ldots, X_n$ IID variables that you know follow a Poisson distribution. We want to find the mean of the Poisson distribution.

Date: February 9, 2011.

Since the variables are independent, their joint density is a product of densities, in other words, our likelihood function is

$$L(\lambda) = \prod_i \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}.$$

We now want to maximize this as a function of λ. Since we have a product, it's usually easier to maximize the log of it, because log is monotonic and it's usually easier to work with sums than products:

$$\log(L(\lambda)) = \sum_i (-\lambda) + X_i \log(\lambda) - \log(X_i!).$$

By taking the derivative with respect to λ, we obtain

$$\frac{d}{d\lambda}\log(L(\lambda)) = \sum_i -1 + X_i \frac{1}{\lambda}.$$

As usual, to find the extrema, we want this number to be zero. With a little algebra, we see that

$$\lambda = \frac{1}{n}\sum_i X_i$$

is the value that works. We can check by taking a second derivative that this is indeed a maximum. This is not surprising, and behaves much like our motivating example above.

Here is the general approach to dealing with these problems, and it even extends to the continuous case:

(a) Find the density function representing the likelihood. It might even be a joint density function if you have many trials X_i as in the examples above.
(b) Consider the density function as a function of the parameter.
(c) Maximize the function with respect to the parameter.

Remark 2.1. Life isn't perfect or easy. Sometimes MLEs don't exist or are too unrealistic. We won't quite worry about this yet, but it's good to know, because it makes you think about ways around these shortcomings and anticipate things that we'll cover later in class.

Let's see another example.

Example 2. We want to find the MLE for the variance from n IID normal variables. Now, a normal distribution has two parameters, the mean μ and the variance σ^2. Suppose that we know the mean μ. Then the likelihood function is

$$L(\sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right).$$

Since we have a product, let's take logs

$$\log L(\sigma^2) = \sum_i -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X_i - \mu)^2.$$

The derivative of this is

$$\sum_i -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X_i - \mu)^2,$$

so we should set

$$\sigma^2 = \frac{1}{n}\sum_i (X_i - \mu)^2$$

as expected.

In practice, if you did not know the μ and σ^2 in the normal distribution, we would first try to find the MLE for μ because it doesn't depend σ^2, then use the information to find to σ^2.

Remark 2.2. If you want to maximize a function of one variable over $\mathbf{R}$ or $(0, \infty)$, make sure that it doesn't take its maximum at zero, or as tends to 0, or as it tends to ∞. Cover your bases.

3. Estimators

Suppose that you want to find the value of some function g of a parameter in your distribution. We can use estimator functions of random variables from that distribution to estimate the value of the function g.

In general, if we observe random variables $X_1, X_2, \ldots, X_n$ whose joint density function depends on unknown parameters $\theta_1, \ldots, \theta_k$, then a good estimator $T = T(x_1, \ldots, x_n)$ of, say, θ_1, based on the observed values $x_1, \ldots, x_n$, is a T such that the estimation error

$$T - \theta_1$$

is "small," that is, so that T_1 is close to θ most of the time. There are several ways to measure this closeness, and we will start with the most basic one.

3.1. **Definitions.** Since we are using an estimate, it's a good idea to have some sort of measurement of how accurate your estimator is. The most basic way of measuring this is to consider the **mean squared error (MSE)**, which is defined as follows.

Let $T(X)$ be an estimator for $g(\theta)$ where θ is some unkown parameter. Then the mean squared error is simply

$$E[(T(X) - g(\theta))^2].$$

It makes sense to take the expected value of this, because it's a function of the random variable X. We can also easily generalize this to multiple observation by replacing X with a vector $(X_i)_i$.

Remark 3.1. Why do we take the square? (And not just the number or a cube or the exponential?) It's the same reason that we take a square to find the variance. It's the simplest way to measure change as some sort of positive quantity.

That said, we are also interested in the quantity

$$E[T(X) - g(\theta)],$$

called the **bias**, for obvious reasons. If this quantity is zero, regardless of θ, we say that the estimator is **unbiased**. If we have an unbiased operator, it doesn't mean that our estimator guess correctly every time, but that in guessing, it doesn't favor a lower or higher value.

Remark 3.2. A warning! The best estimator is not always unbiased. This is fairly counterintuitive, so it is something you need to watch out for. This might not make sense now, but this concept should be clear after you work through some calculations. To reiterate, sometimes a *biased* estimator gives you better estimates than an unbiased one!

Example 3. It is always true that the sample mean $\overline{X}_i$ is an unbiased estimator of the true mean. This is because

$$\begin{aligned} E[(1/n)\sum_i X_i - \mu] &= (1/n)\sum_i E[X_i] - \mu \\ &= \mu - \mu = 0. \end{aligned}$$

However, the (MSE) of this estimator can change based on the distribution. For example, if they follow the Poisson distribution as in our example above, we get

$$\begin{aligned} E\left[\left(\frac{1}{n}\sum_i X_i - \mu\right)^2\right] &= E\left[\frac{1}{n^2}\left(\sum_i X_i\right)^2\right] - 2\cdot\frac{1}{n}\cdot\mu\sum_i E[X_i] + \mu^2. \\ &= E\left[\frac{1}{n^2}\left(\sum_i X_i\right)^2\right] - 2\mu^2 + \mu^2 \\ &= E\left[\left(\frac{1}{n}\sum_i X_i\right)^2\right] - \mu^2 \\ &= E[Y^2] - E[Y]^2 \\ &= \mathrm{Var}(Y) \\ &= \frac{\lambda}{n}. \end{aligned}$$

where $Y = (1/n)\sum_i X_i$.

Note the variance of a sum of independent variables is the sum of the variances, and multiplying by a constant multiplies the variance by the square, so we are left with a $1/n$. This is good, since clearly the MSE of the sample mean should decrease with n, because we should get better estimates with larger samples. Also note that the MSE changes as the true value of λ changes.

Example 4. Suppose that we have n IID variables from the uniform distribution on $[0, \theta]$. It turns out that the MLE for θ is $\max_i X_i$. To see this, note that the likelihood function is $L(\theta) = (1/\theta)^n$ if $\theta > \max_i X_i$, and 0 otherwise. Since maximizing the likelihood means minimizing θ, we make this as small as possible.

Remark 3.3. Note that this estimator is **not** unbiased, because it always guesses high!

It turns out you can make our estimator unbiased by multiplying the max by $n/(n+1)$, but let's take an approach that you might find useful on the homework. Consider all estimators of the form $c\max_i X_i$, and try to minimize the MSE.

The relevant expected value is

$$E[c^2(\max_i X_i)^2 - 2c\theta\max_i X_i + \theta^2] = c^2E[(\max_i X_i)^2] - 2c\theta E[\max_i X_i] + \theta^2.$$

Finding the first quantity requires finding the density of $\max_i X_i$. This is

$$\frac{d}{dt}P(\max_i X_i < t) = \frac{d}{dt}\left(\frac{t}{\theta}\right)^n = \left(\frac{n}{\theta}\right)\left(\frac{t}{\theta}\right)^{n-1}.$$

Now, this only makes sense up to θ, but it turns to work out OK:

$$\begin{aligned} E[(\max_i X_i)^2] &= \int_0^\theta t^2 \frac{nt^{n-1}}{\theta^n}\, dt \\ &= \frac{n}{\theta^n}\int_0^\theta t^{n+1}\, dt \\ &= \frac{n}{\theta^n}\cdot\frac{\theta^{n+2}}{n+2} \\ &= \frac{n}{n+2}\theta^2. \end{aligned}$$

Similarly, $E[\max_i X_i] = \frac{n}{n+1}\theta$. Therefore, the MSE of the estimator $c\max_i X_i$ is

$$c^2\frac{n}{n+2}\theta^2 - 2c\theta\frac{n}{n+1}\theta + \theta^2 = \left(c^2\frac{n}{n+2} - 2c\frac{n}{n+1} + 1\right)\theta^2.$$

Therefore, we just minimize this with respect to c to get the estimator with the smallest MSE, which you can check is $\frac{n+2}{n+1}$. Since we can't get rid of the multiplication by θ in our MSE, we can just try and to minimize the remaining factor.