Maximum Likelihood Estimates
Class 10, 18.05, Spring 2014
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to define the likelihood function for a parametric model given data.
2. Be able to compute the maximum likelihood estimate of unknown parameter(s).

2 Introduction

Suppose we have data consisting of values $x_1, \dots, x_n$ drawn from an exponential distribution. A question remains: which exponential distribution?!

We have casually referred to *the* exponential distribution or *the* binomial distribution or *the* normal distribution. In fact the exponential distribution $\exp(\lambda)$ is not a single distribution but rather a one-parameter family of distributions. Each value of λ defines a different distribution in the family, with pdf $f_\lambda(x) = \lambda e^{-\lambda x}$ on $[0, \infty)$. Similarly, a binomial distribution $\text{bin}(n, p)$ is determined by the two parameters n and p, and a normal distribution $N(\mu, \sigma^2)$ is determined by the two parameters μ and σ^2 (or equivalently, μ and σ). Parameterized families of distributions are often called *parametric distributions* or *parametric models*.

We are often faced with the situation of having random data which we know (or believe) is drawn from a parametric model, whose parameters we do not know. For example, in an election between two candidates, polling data constitutes draws from a Bernoulli(p) distribution with unknown parameter p. In this case we would like to use the data to estimate the value of the parameter p, as the latter determines the result of the election. Similarly, assuming gestational length follows a normal distribution, we would like to use the data of the gestational lengths from a random sample of pregnancies to draw inferences about the values of the parameters μ and σ^2.

Our focus so far has been on computing the *probability of data* arising from a parametric model with *known parameters*. Statistical inference flips this on its head: we will estimate the *probability of parameters* given a parametric model and *observed data* drawn from it. In the coming weeks we will see how parameter values are naturally viewed as hypotheses, so we are in fact estimating the probability of various hypotheses given the data.

3 Maximum Likelihood Estimates

There are many methods for estimated unknown parameters from data. We will first consider the *maximum likelihood estimate* (MLE), which answers the question:

For which parameter value does the observed data have the biggest probability?

The MLE is an example of a *point estimate* because it gives a single value for the unknown parameter (later our estimates will involve intervals and probabilities). Two advantages of

the MLE are that it is often easy to compute and that it agrees with our intuition in simple examples. We will explain the MLE through a series of examples.

Example 1. A coin is flipped 100 times. Given that there were 55 heads, find the maximum likelihood estimate for the probability p of heads on a single toss.

Before actually solving the problem, let's establish some notation and terms.

We can think of counting the number of heads in 100 tosses as an experiment. For a given value of p, the probability of getting 55 heads in this experiment is the binomial probability

$$P(55 \text{ heads}) = \binom{100}{55} p^{55}(1-p)^{45}.$$

The probability of getting 55 heads depends on the value of p, so let's include p in our notation using that of conditional probability:

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

You should read $P(55 \text{ heads} \,|\, p)$ as

'the probability of 55 heads given p,'

or more precisely as

'the probability of 55 heads given that the probability of heads on a single toss is p.'

Here are some standard terms we will use as we do statistics.

- *Experiment*: Flip the coin 100 times and count the number of heads.
- *Data*: The data is the result of the experiment. In this case it is '55 heads'.
- *Parameter(s) of interest*: We are interested in the value of the unknown parameter p.
- *Likelihood*, or *likelihood function*: this is $P(\text{data} \,|\, p)$. Note it depends on the data and the parameter p. In this case the likelihood is

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

Notes:
1. The likelihood $P(\text{data} \,|\, p)$ changes as the parameter of interest p changes.
2. Look carefully at the definition. One typical source of confusion is to mistake the likelihood $P(\text{data} \,|\, p)$ for $P(p \,|\, \text{data})$. We know from our earlier work with Bayes' theorem that $P(\text{data} \,|\, p)$ and $P(p \,|\, \text{data})$ are usually very different.

Definition: Given data the *maximum likelihood estimate* for the parameter p is the value of p that maximizes the likelihood $P(\text{data} \,|\, p)$.

That is, the MLE is the value of p for which the data is most likely.

answer: For the problem at hand, we saw above that the likelihood

$$P(55 \text{ heads} \,|\, p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

We'll use the notation $\hat{p}$ for the MLE. We find it by finding where the derivative of the likelihood function is 0.

$$\begin{aligned}\frac{d}{dp}P(\text{data }|p) &= \binom{100}{55}(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0\\ &\Rightarrow 55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44}\\ &\Rightarrow 55(1-p) = 45p\\ &\Rightarrow 55 = 100p\\ &\Rightarrow \text{the MLE is } \hat{p} = .55\end{aligned}$$

Note:
1. The MLE for p turned out to be exactly the fraction of heads we saw in our data.
2. The MLE is computed from the data. That is, it is a statistic.
3. Officially you should check that the critical point is indeed a maximum. You can do this with the second derivative test.

3.1 Log likelihood

If is often easier to work with the natural log of the likelihood function. For short this is simply called the *log likelihood*. Since ln is an increasing function, the maxima of the likelihood and log likelihood coincide.

Example 2. Redo the previous example using log likelihood.

answer: We had the likelihood $P(55 \text{ heads } | \, p) = \binom{100}{55}p^{55}(1-p)^{45}$. Therefore the log likelihood is

$$\ln(P(55 \text{ heads } | \, p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p).$$

Maximizing likelihood is the same as maximizing log likelihood. We check that calculus gives us the same answer as before:

$$\begin{aligned}\frac{d}{dp}(\text{log likelihood}) &= \frac{d}{dp}\left[\ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p)\right]\\ &= \frac{55}{p} - \frac{45}{1-p} = 0\\ &\Rightarrow 55(1-p) = 45p\\ &\Rightarrow \hat{p} = .55\end{aligned}$$

3.2 Maximum likelihood for continuous distributions

For continuous distributions, we use the probability density function to define the likelihood.

Example 3. Light bulbs
Suppose that the lifetime of *Badger* brand light bulbs is modeled by an exponential distribution with (unknown) parameter λ. We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ?

answer: We need to be careful with our notation. With five different values it is best to use subscripts. Let X_j be the lifetime of the i^{th} bulb and let x_i be the value X_i takes. Then each X_i has pdf $\lambda e^{-\lambda x}$. We assume the lifetimes of the bulbs are independent, so the joint pdf is

$$f(x_1, x_2, x_3, x_4, x_5 \,|\, \lambda) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}.$$

Note that we write this as a conditional density, since it depends on λ. Viewing the data as fixed and λ as variable, this density is the likelihood function. Our data had values

$$x_1 = 2,\ x_2 = 3,\ x_3 = 1,\ x_4 = 3,\ x_5 = 4.$$

So the likelihood and log likelihood functions with this data are

$$f(2, 3, 1, 3, 4 \,|\, \lambda) = \lambda^5 e^{-13\lambda}, \quad \ln(f(2, 3, 1, 3, 4 \,|\, \lambda) = 5\ln(\lambda) - 13\lambda$$

Finally we use calculus to find the MLE:

$$\frac{d}{d\lambda}(\text{log likelihood}) = \frac{5}{\lambda} - 13 = 0 \ \Rightarrow\ \boxed{\hat{\lambda} = \frac{5}{13}}.$$

Note:
1. In this example we used an uppercase letter for a random variable and the corresponding lowercase letter for the value it takes. This will be our usual practice. 2. The MLE for λ turned out to be the reciprocal of the sample mean $\bar{x}$, so $X \sim \exp(\hat{\lambda})$ satisfies $E(X) = \bar{x}$.

We can use the method of maximum likelihood to estimate multiple parameters at once.

Example 4. Normal distributions
Suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a $N(\mu, \sigma^2)$ distribution, where μ and σ are unknown. Find the maximum likelihood estimate for the pair (μ, σ^2).

answer: Let's be precise and phrase this in terms of random variables and densities. Let uppercase $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables, and let lowercase x_i be the value X_i takes. The density for each X_i is

$$\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Since the X_i are independent their joint pdf is the product of the individual pdf's:

$$f(x_1, \ldots, x_n \,|\, \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

For the fixed data $x_1, \ldots, x_n$, the likelihood and log likelihood are

$$f(x_1, \ldots, x_n|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}, \quad \ln(f(x_1, \ldots, x_n|\mu, \sigma)) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Since $\ln(f(x_1, \ldots, x_n|\mu, \sigma))$ is a function of the two variables μ, σ we use partial derivatives to find the MLE. The easy value to find is $\hat{\mu}$:

$$\frac{\partial f(x_1, \ldots, x_n|\mu, \sigma)}{\partial \mu} = \sum_{i=1}^n \frac{(x_i - \mu)}{\sigma^2} = 0 \ \Rightarrow\ \sum_{i=1}^n x_i = n\mu \ \Rightarrow\ \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}.$$

To find $\hat{\sigma}$ we differentiate and solve for σ:

$$\frac{\partial f(x_1,\dots,x_n|\mu,\sigma)}{\partial\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^3} = 0 \;\Rightarrow\; \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}.$$

We already know $\hat{\mu} = \overline{x}$, so we use that as the value for μ in the formula for $\hat{\sigma}$. We get the maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \overline{x} && = \text{the mean of the data} \\ \hat{\sigma}^2 &= \sum_{i=1}^{n}\frac{1}{n}(x_i-\hat{\mu})^2 = \sum_{i=1}^{n}\frac{1}{n}(x_i-\overline{x})^2 && = \text{the variance of the data.} \end{aligned}$$

Example 5. Uniform distributions
Suppose our data $x_1,\dots x_n$ are independently drawn from a uniform distribution $U(a,b)$. Find the MLE estimate for a and b.

answer: This example is different from the previous ones in that we won't use calculus to find the MLE. The density for $U(a,b)$ is $\frac{1}{b-a}$ on $[a,b]$. Therefore our likelihood function is

$$f(x_1,\dots,x_n \,|\, a,b) = \left(\frac{1}{b-a}\right)^n$$

if all x_i are in the interval $[a,b]$, and 0 otherwise. This is maximized by making $b-a$ as small as possible. The only restriction is that the interval $[a,b]$ must include all the data. Thus the MLE for the pair (a,b) is

$$\hat{a} = \min(x_1,\dots,x_n) \qquad \hat{b} = \max(x_1,\dots,x_n).$$

Example 6. Capture/recapture method

The capture/recapture method is a way to estimate the size of a population in the wild. The method assumes that each animal in the population is equally likely to be captured by a trap.

Suppose 10 animals are captured, tagged and released. A few months later, 20 animals are captured, examined, and released. 4 of these 20 are found to be tagged. Estimate the size of the wild population using the MLE for the probability that a wild animal is tagged.

answer: Our unknown parameter n is the number of animals in the wild. Our data is that 4 out of 20 recaptured animals were tagged (and that there are 10 tagged animals). The likelihood function is

$$P(\text{data} \,|\, n \text{ animals}) = \frac{\binom{n-10}{16}\binom{10}{4}}{\binom{n}{20}}$$

(The numerator is the number of ways to choose 16 animals from among the $n-10$ untagged ones times the number of was to choose 4 out of the 10 tagged animals. The denominator is the number of ways to choose 20 animals from the entire population of n.) We can use R to compute that the likelihood function is maximized when $n = 50$. This should make some sense. It says our best estimate is that the fraction of all animals that are tagged is 10/50 which equals the fraction of recaptured animals which are tagged.

Example 7. Hardy-Weinberg. Suppose that a particular gene occurs as one of two alleles (A and a), where allele A has frequency θ in the population. That is, a random copy of the gene is A with probability θ and a with probability $1-\theta$. Since a diploid genotype consists of two genes, the probability of each genotype is given by:

genotype	AA	Aa	aa
probability	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

Suppose we test a random sample of people and find that k_1 are AA, k_2 are Aa, and k_3 are aa. Find the MLE of θ.

answer: The likelihood function is given by

$$P(k_1, k_2, k_3 \,|\, \theta) = \binom{k_1+k_2+k_3}{k_1}\binom{k_2+k_3}{k_2}\binom{k_3}{k_3}\theta^{2k_1}(2\theta(1-\theta))^{k_2}(1-\theta)^{2k_3}.$$

So the log likelihood is given by

$$\text{constant} + 2k_1\ln(\theta) + k_2\ln(\theta) + k_2\ln(1-\theta) + 2k_3\ln(1-\theta)$$

We set the derivative equal to zero:

$$\frac{2k_1+k_2}{\theta} - \frac{k_2+2k_3}{1-\theta} = 0$$

Solving for θ, we find the MLE is

$$\hat{\theta} = \frac{2k_1+k_2}{2k_1+2k_2+2k_3},$$

which is simply the fraction of A alleles among all the genes in the sampled population.

4 Appendix: Properties of the MLE

For the interested reader, we note several nice features of the MLE. These are quite technical and will not be on any exams.

The MLE behaves well under transformations. That is, if $\hat{p}$ is the MLE for p and g is a one-to-one function, then $g(\hat{p})$ is the MLE for $g(p)$. For example, if $\hat{\sigma}$ is the MLE for the standard deviation σ then $(\hat{\sigma})^2$ is the MLE for the variance σ^2.

Furthermore, the MLE is *asymptotically unbiased* and has *asymptotically minimal variance*. To explain these notions, note that the MLE is itself a random variable since the data is random and the MLE is computed from the data. Let $x_1, x_2, \ldots$ be an infinite sequence of samples from a distribution with parameter p. Let $\hat{p}_n$ be the MLE for p based on the data $x_1, \ldots, x_n$.

Asymptotically unbiased means that as the amount of data grows, the mean of the MLE converges to p. In symbols: $E(\hat{p}_n) \to p$ as $n \to \infty$. Of course, we would like the MLE to be close to p with high probability, not just on average, so the smaller the variance of the MLE the better. Asymptotically minimal variance means that as the amount of data grows, the MLE has the minimal variance among all unbiased estimators of p. In symbols: for any unbiased estimator $\tilde{p}_n$ and $\epsilon > 0$ we have that $\text{Var}(\tilde{p}_n) + \epsilon > \text{Var}(\hat{p}_n)$ as $n \to \infty$.

18.05 Introduction to Probability and Statistics

Spring 2014