STAT 504 | Analysis of Discrete Data

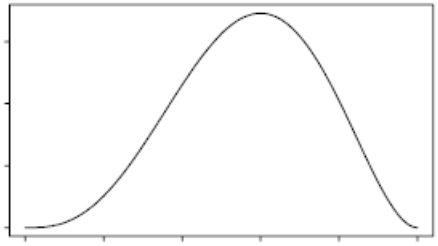
1.5 - Maximum-likelihood (ML) Estimation

Printer-friendly version (../../print/book/export/html/28/)

Suppose that an experiment consists of $n = 5$ independent Bernoulli trials, each having probability of success p. Let X be the total number of successes in the trials, so that $X \sim Bin(5, p)$. If the outcome is $X = 3$, the likelihood is

$$\begin{aligned} L(p; x) &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \frac{5!}{3!(5-3)!} p^3 (1-p)^{5-3} \\ &\propto p^3 (1-p)^2 \end{aligned}$$

where the constant at the beginning is ignored. A graph of $L(p; x) = p^3(1-p)^2$ over the unit interval $p \in (0, 1)$ looks like this:



It's interesting that this function reaches its maximum value at $p = .6$. An intelligent person would have said that if we observe 3 successes in 5 trials, a reasonable estimate of the long-run proportion of successes p would be 3/5 = .6.

This example suggests that it may be reasonable to estimate an unknown parameter θ by the value for which the likelihood function $L(\theta ; x)$ is largest. This approach is called *maximum-likelihood (ML) estimation*. We will denote the value of θ that maximizes the likelihood function by $\hat{\theta}$, read "theta hat."$\hat{\theta}$ is called the *maximum-likelihood estimate (MLE)* of θ.

Finding MLE's usually involves techniques of differential calculus. To maximize $L(\theta ; x)$ with respect to θ:

- first calculate the derivative of $L(\theta ; x)$ with respect to θ,
- set the derivative equal to zero, and
- solve the resulting equation for θ.

These computations can often be simplified by maximizing the *loglikelihood function*,

$$l(\theta; x) = \log L(\theta; x),$$

where "log" means natural log (logarithm to the base e). Because the natural log is an increasing function, maximizing the loglikelihood is the same as maximizing the likelihood. The loglikelihood often has a much simpler form than the likelihood and is usually easier to differentiate.

In Stat 504 you will not be asked to derive MLE's by yourself. In most of the probability models that we will use later in the course (logistic regression, loglinear models, etc.) no explicit formulas for MLE's are available, and we will have to rely on computer packages to calculate the MLE's for us. For the simple probability models we have seen thus far, however, explicit formulas for MLE's are available and are given next.

ML for Bernoulli trials

If our experiment is a single Bernoulli trial and we observe $X = 1$ (success) then the likelihood function is $L(p\ ; x) = p$. This function reaches its maximum at $\hat{p} = 1$. If we observe $X = 0$ (failure) then the likelihood is $L(p\ ; x) = 1 - p$, which reaches its maximum at $\hat{p} = 0$. Of course, it is somewhat silly for us to try to make formal inferences about θ on the basis of a single Bernoulli trial; usually multiple trials are available.

Suppose that $X = (X_1, X_2, \ldots, X_n)$ represents the outcomes of n independent Bernoulli trials, each with success probability p . The likelihood for p based on X is defined as the joint probability distribution of $X_1, X_2, \ldots, X_n$. Since $X_1, X_2, \ldots, X_n$ are iid random variables, the joint distribution is

$$L(p; x) \approx f(x; p) = \prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^x (1-p)^{1-x}$$

Differentiating the log of $L(p\ ; x)$ with respect to p and setting the derivative to zero shows that this function achieves a maximum at $\hat{p} = \sum_{i=1}^{n} x_i / n$. Since $\sum_{i=1}^{n} x_i$ is the total number of successes observed in the n trials, $\hat{p}$ is the observed proportion of successes in the n trials. We often call $\hat{p}$ the sample proportion to distinguish it from p , the "true" or "population" proportion. Note that in some textbooks the authors may use π instead of p. For repeated Bernoulli trials, the MLE $\hat{p}$ is the sample proportion of successes.

ML for Binomial

Suppose that X is an observation from a binomial distribution, $X \sim \text{Bin}(n, p\ )$, where n is known and p is to be estimated. The likelihood function is

$$L(p; x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

which, except for the factor $\frac{n!}{x!(n-x)!}$, is identical to the likelihood from n independent Bernoulli trials with $x = \sum_{i=1}^{n} x_i$. But since the likelihood function is regarded as a function only of the parameter p, the factor $\frac{n!}{x!(n-x)!}$is a fixed constant and does not affect the MLE. Thus the MLE is again $\hat{p} = x/n$, the sample proportion of successes.

You get the same value by maximizing the *binomial loglikelihood function*

$$l(p; x) = k + x \log p + (n - x) \log (1 - p)$$

where k is a constant that does not involve the parameter p. In the future we will omit the constant, because it's statistically irrelevant.

The fact that the MLE based on n independent Bernoulli random variables and the MLE based on a single binomial random variable are the same is not surprising, since the binomial is the result of n independent Bernoulli trials anyway. In general, whenever we have repeated, independent Bernoulli trials with the same probability of success p for each trial, the MLE will always be the sample proportion of successes. This is true regardless of whether we know the outcomes of the individual trials $X_1, X_2, \ldots, X_n$, or just the total number of successes for all trials $X = \sum_{i=1}^{n} X_i$.

Suppose now that we have a sample of iid binomial random variables. For example, suppose that $X_1, X_2, \ldots, X_{10}$ are an iid sample from a binomial distribution with $n = 5$ and p unknown. Since each X_i is actually the total number of successes in 5 independent Bernoulli trials, and since the X_i's are independent of one another, their sum $X = \sum_{i=1}^{10} X_i$ is actually the total number of successes in 50 independent Bernoulli trials. Thus $X \sim Bin(50, p)$ and the MLE is $\hat{p} = x/n$, the observed proportion of successes across all 50 trials. Whenever we have independent binomial random variables with a common p , we can always add them together to get a single binomial random variable.

Adding the binomial random variables together produces no loss of information about *p if the model is true*. But collapsing the data in this way may limit our ability to diagnose model failure, i.e. to check whether the binomial model is really appropriate.

ML for Poisson

Suppose that $X = (X_1, X_2, \ldots, X_n)$ are iid observations from a Poisson distribution with unknown parameter λ. The likelihood function is:

$$\begin{aligned} L(\lambda; x) &= \prod_{i=1}^{n} f(x_i; \lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{x_1! x_2! \cdots x_n!} \end{aligned}$$

By differentiating the log of this function with respect to λ, that is by differentiating the *Poisson loglikelihood function*

$$l(\lambda; x) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda$$

ignoring the constant terms that do not depend on λ, one can show that the maximum is achieved at $\hat{\lambda} = \sum_{i=1}^{n} x_i/n$. Thus, for a Poisson sample, the MLE for λ is just the sample mean.

Next: Likelihood-based confidence intervals and tests.

‹ 1.4 - Likelihood & LogLikelihood (../27/) up (../4/) 1.6 - Likelihood-based Confidence Intervals & Tests › (../29/)