CLARK UNIVERSITY

Maximum likelihood estimators
Math 218, Mathematical Statistics

D Joyce, Spring 2016

We'll look at maximum likelihood estimators in section 15.1.

Likelihood functions for discrete distributions and maximum likelihood estimators. The setting is that we have a family of distributions parametrized by θ, and we run an experiment to get outcome values $\mathbf{x} = (x_1, \ldots, x_n)$. From this data we want to decide what the value of θ is. The idea of the maximum likelihood estimator is that the best value of θ is the one makes the probability of the outcome the greatest.

In some cases, it's pretty easy to see what the maximum likelihood estimator is. Let's take the Bernoulli case where the unknown parameter of success is p. (So, in this case the parameter θ is p.) Suppose we get a sample $\mathbf{x}$ of $n = 100$ trials, and 47 of them turn out success while 53 are failures. What value of p maximizes the probability

$$P(X = \mathbf{x} \,|\, p)?$$

Intuitively, it's $p = 0.47$, and that's the maximum likelihood estimator. We'll verify that guess is correct after stating some general definitions and principles.

The likelihood of a parameter θ for a given random sample $\mathbf{X} = \mathbf{x}$, that is, $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, is the probability

$$P(\mathbf{X} = \mathbf{x} \,|\, \theta)$$

but it's denoted

$$L(\theta \,|\, \mathbf{x}).$$

Thus, likelihood is not a probability, but the reverse of a conditional probability. That probability can also be written as the product of values of the probability mass function f as

$$f(x_1 \,|\, \theta) f(x_2 \,|\, \theta) \cdots f(x_n \,|\, \theta).$$

Then the maximum likelihood estimator $\hat{\theta}$ for the unknown parameter θ is just that value of θ that has the highest probability of that outcome, that is, has the greatest likelihood.

So in our Bernoulli example with a particular outcome $\mathbf{x}$ having 47 successes and 53 failures,

$$L(p|\mathbf{x}) = p^{47}(1-p)^{53}).$$

We want to find the value $\hat{p}$ which maximizes $p^{47}(1-p)^{53}$. We can use calculus to do that. Take the derivative with respect to p, $\dfrac{d}{dp}p^{47}(1-p)^{53}$, and set that derivative to 0 to find the critical points. There is an easier way. The likelihood function $L(P|\mathbf{x})$ has its maximum at the same place as its natural log $\ln L(P|\mathbf{x})$ does since the log function is an increasing function, and it's easier to take the derivative of the log

$$\begin{aligned} & \frac{d}{dp} \ln(p^{47}(1-p)^{53}) \\ = \; & \frac{d}{dp}(47 \ln p + 53 \log(1-p)) \\ = \; & \frac{47}{p} - \frac{53}{1-p} \end{aligned}$$

If we set that to 0 to find the critical points, and simplify the equation, we get

$$\frac{47}{p} = \frac{53}{1-p},$$

so $47(1-p) = 53p$, and so $p = \dfrac{47}{100}$. Thus, the maximum likelihood estimator is $\hat{p} = 0.47$ just as we expected.

In more complicated cases, we can't see right off what value of θ will maximize the likelihood $L(\theta|\mathbf{x})$, and we'll have to resort to the method above where we take logarithmic derivatives to find $\hat{\theta}$.

Likelihood functions for continuous distributions. For continuous distributions, we can't use probability, because the probability of any particular outcome is 0. But we can use the density function. Thus, for a continuous distribution, the likelihood of a parameter θ for a given random sample $X_1 = x_1$, $X_2 = x_2$, ..., $X_n = x_n$, also denoted $L(\theta|x_1, x_2, \ldots, x_n)$ is the product of values of the density function f as

$$f(x_1 \,|\, \theta) f(x_2 \,|\, \theta) \cdots f(x_n \,|\, \theta).$$

(So, the same formula, but the symbol f now denotes a probability density function instead of a probability mass function.)

The likelihood function for the normal distribution and its maximum likelihood estimators. Since the probability density function for a normal$(\mu, \sigma^)$ distribution is

$$f(x \,|\, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

the likelihood function is

$$\begin{aligned} & L(\mu, \sigma^2 | x_1, x_2, \ldots, x_n) \\ = & \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ = & \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \end{aligned}$$

Next, to find the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$ for the parameters μ and σ^2, we just have to find those values of μ and σ^2 that maximize the function $L(\mu, \sigma^2|x_1, x_2, \ldots, x_n)$. We need to compute its derivative to find the critical points so we can find where the maximum occurs. But, since the function $L(\mu, \sigma^2)$ at the same places where its log does, that is, where $\ln L(\mu, \sigma^2)$ does, we'll use its log instead, because its easier to find the derivative of its log. Its log is

$$\ln L(\mu, \sigma^2) = -n \ln \sqrt{2\pi} - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Here we have two parameters, μ and σ^2, so we need to set both derivatives of $\ln L(\mu, \sigma^2)$ to 0. First, the derivative with respect to μ

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$

Set that to 0. Since $\displaystyle\sum_{i=1}^{n}(x_i - \mu) = 0$, therefore the critical value for μ is $\frac{1}{n} sum_{i=1}^{n} x_i$, which is the sample mean $\overline{x}$. This is the only critical value, so it maximizes $L(\mu, \sigma^2)$. Therefore, the maximum likelihood estimator $\hat{\mu}$ for the mean μ is the sample mean $\overline{x}$.

Second, the derivative with respect to σ^2

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Set that to 0. We can simplify the resulting equation a bit to get

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = n$$

Therefore

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

and, as we're solving these equations simultaneously, we've already determined the solution has $\mu = \overline{x}$, so we can rewrite that as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Again, as this is the only critical value for σ^2, it maximizes $L(\mu, \sigma^2)$. Therefore, the maximum likelihood estimator $\hat{\sigma}^2$ for the population variance σ^2 is the statistic

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

As we discussed before, sometimes this statistic is called the sample variance, but our text uses $n - 1$ in the denominator for the sample variance.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`