VectorBiTE Methods Training
Introduction to the Likelihood

The VectorBiTE Team
(Leah R. Johnson, Virginia Tech)

Summer 2021

Assumed Background

In this workshop, we expect that you are familiar with:

- axioms of probability and their consequences.
- conditional probability and Bayes theorem
- definition of a random variable (discrete and continuous)
- the idea of a probability distribution and likelihood

Pre-workshop reading and exercises were assigned to help you review and get you ready.

We'll do a VERY fast review of likelihoods and then practice building them and finding the MLEs analytically and with R.

Finding estimates of parameters

When we fit lines using least squares and similar techniques, we defined a metric to measure distance between a prediction and our data, and then found parameters that made that distance as small as possible.

Likelihoods are another way of defining a distance between our prediction (probability distribution) and data and allow us to find parameter values that are consistent with the data under the constraint of a particular probability distribution.

Method of Moments

Before we review likelihoods, let's review an easy alternative to finding consistent parameters that assumes a probability distribution: method of moments.

Consider an *iid* sample of n observations of a random variable $\{x_1, \ldots, x_n\}$. You can calculate sample values of the moments of the RV from these, i.e.:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
$$s^2 = \frac{1}{n} \sum (x - \bar{x})^2$$

You estimate the parameters of a probability distribution by “matching” up the sample moments with the analytical values of the moments for your probability distribution.

Example: The Poisson distribution has only one parameter λ. Since the expected value of the Poisson $\mathrm{E}[x] = \lambda$ we set:

$$\lambda = \mathrm{E}[x] = \bar{x}$$

Then the MoM estimator is:

$$\Rightarrow \hat{\lambda} = \bar{x}$$

Likelihoods

Recall that $f(Y_i)$ is the pmf (pdf), and it tells us the probability (density) of some yet to be observed datum Y_i given a probability distribution and its parameters.

If we make many observations, $\mathbf{Y} = y_1, y_2, \ldots, y_n$, we are interested how probable it was that we obtained these data, jointly. We call this the **likelihood** of the data, and denote it as

$$\mathcal{L}(\theta; Y) = f_\theta(Y)$$

where $f_\theta(Y)$ is the pdf (or pmf) of the data interpreted as a function of θ.

For instance, for *binomial* data:

$$\Pr(Y_i = k|\theta = p) = \binom{N}{k} p^k (1-p)^{N-k}.$$

If we have data $\mathbf{Y} = y_1, y_2, \ldots, y_n$ that are i.i.d. as binomial RVs, the probabilities multiply, and the likelihood is:

$$\mathcal{L}(\theta; Y) = \prod_{i=1}^{n} \binom{N}{y_i} p^{y_i} (1-p)^{N-y_i}.$$

Likelihoods vs. probability

"Likelihood is the hypothetical probability [density] that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes." (1)

Further, the likelihood is a function of θ (the parameters), assuming fixed data.

1. Weisstein, Eric W. "Likelihood.'' From MathWorld–A Wolfram Web Resource.

http://mathworld.wolfram.com/Likelihood.html

We are usually interested in relative likelihoods – e.g., is it more likely that the data we observed came from a distribution with parameters θ_1 or θ_2? Thus we only worry about the likelihood up to a constant.

Further, it is often easier to work with the log-likelihood:

$$L(\theta; Y) = \ell(\theta; Y) = \log(\mathcal{L}(\theta; Y))$$

where $\log(\cdot)$ is the natural log.

Maximum Likelihood Estimators (MLEs)

We can find the parameters that are most likely to have generated our data – the maximum likelihood estimate (MLE) of the parameters. To do this we maximize the likelihood (or equivalently minimizing the negative log-likelihood) by taking its derivative and setting it equally to zero:

$$\frac{\partial \mathcal{L}}{\partial \theta_j} = 0 \qquad \text{or} \qquad -\frac{\partial L}{\partial \theta_j} = 0$$

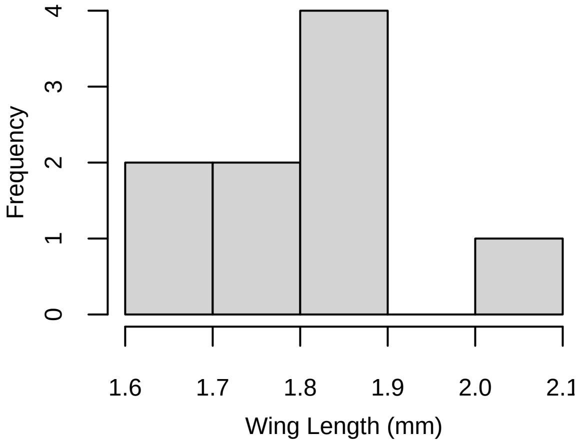
where j denotes the j^{th} parameter.

We usually denote the MLE as $\hat{\theta}_j$.

The likelihood **DOES NOT** tell you the probability that parameters have a certain value, given the data.

To obtain that quantity, usually called the "posterior probability of the parameters" in Bayesian statistics, you have to use Bayes Theorem (later lectures).

A Simple Example: MLE for mean midge wing lengths

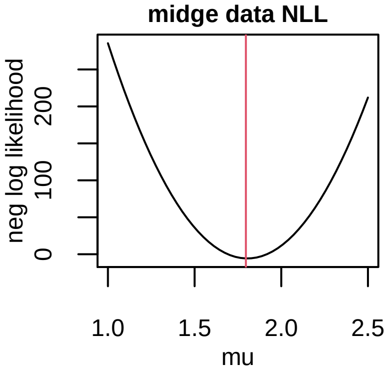


Likelihood profile in R

We interpret the negative log-likelihood (NLL) as a function of the parameters assuming that the data are constant. We visualize the NLL with a profile → evaluate the NLL for many possible values of a parameter. The **best** estimate has the lowest NLL value.

```
N<-50
sigma<-0.1
mus<-seq(1, 2.5, length=N)
mynll<-rep(NA, length=50)

for(i in 1:N){
  mynll[i]<- nll.norm(
    par=mus[i],
    dat=midgedat$WingLength,
    sigma=sigma
    )
}
```



Next Steps

There are two sets of tasks in the likelihood practical to help you get comfortable with likelihoods:

1. Mathematical Practice (using Binomial Distribution Example)
2. Coding Practice (maximum likelihood for SLR using R)