A brief introduction to maximum likelihood

The key idea behind the method of maximum likelihood is that we start with a probability distribution that we believe is appropriate for our data, then we change the focus from calculating a probability of an observation given model parameters, to finding the most likely parameter given a particular observation. To illustrate, consider the binomial distribution:

$$P(x|n,p) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

We can use this probability density function to answer questions like "if an event has probability $p = .6$, and we have $n = 10$ trials, what is the probability of the event occurring $x = 3$ times"? This was a probability question, but when we collect data we have a statistical question such as "in $n = 10$ trials I observed the event occur $x = 3$ times, so what is a good estimator of the success probability p"? To address this question, we consider essentially the same function but from a different point of view: now the data (x) are fixed, and we view the expression in terms of the parameter (p), and use it to obtain an estimator of the parameter (the success probability p). Thus we define the likelihood function for binomial data as:

$$L(p|n,x) = \binom{n}{x} p^x (1-p)^{(n-x)},$$

which is the same expression now viewed as a function of p instead of x. Some authors leave out the $\binom{n}{x}$ term in the definition of the likelihood function, since it does not affect subsequent calculations. The maximum likelihood estimator of p is defined as the value of p that maximizes the function $L(p|n,x)$. It is easier to maximize the log of the likelihood function, so we define the log-likelihood function as:

$$\log L(p|n,x) = \ell(p|n,x) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p).$$

We can use calculus to differentiate this function with respect to p, set the derivative to zero, solve for p, and confirm that our solution is a maximum value. It turns out the the maximum likelihood estimator for p in the binomial model is:

$$\widehat{p} = x/n.$$

With variance component estimates, maximum likelihood estimators can be biased for finite sample sizes, so a related method called REML (residual or restricted maximum likelihood) is used instead.