# STA 114: Statistics

## Notes 3. Maximum Likelihood

### Inference from the Likelihood Function

The likelihood function $L_x(\theta)$ (or the log-likelihood function $\ell_x(\theta)$) gives scores to theories $\theta \in \Theta$ as to how well they explain the observed data $X = x$. There are two ways we can report the message from the likelihood function. First, we can split the parameter space $\Theta$ into two subsets: a subset of well supported theories and the other with theories not so well supported by the observed data, with $L_x(\theta)$ for any $\theta$ in the first subset being larger than $L_x(\theta)$ for every $\theta$ in the second. This approach leads to maximum likelihood (ML) inference. An alternative is to convert the likelihood function $L_x(\theta)$ into a pdf/pmf on $\Theta$ and then report visual or numerical summaries of this pdf/pmf through density plot, mean, variance, quantiles etc. This approach leads to the Bayesian inference. For now we focus on ML inference.

### Maximum Likelihood Estimate

A set of well supported theories can be identified by considering all $\theta$ such that

$$L_x(\theta) \geq k \times \max_{\theta \in \Theta} L_x(\theta),$$

for some fraction $k \in [0, 1]$ of our choice. If $A$ denotes the subset of $\theta$ where this happens, then clearly $L_x(\theta_1) > L_x(\theta_2)$ for any $\theta_1 \in A$ and $\theta_2 \in \Theta \setminus A$. See Figure 1.

An extreme case of this obtains if we choose $k = 1$, so that we only consider theories that are best supported by data:

$$L_x(\theta) = \max_{\theta \in \Theta} L_x(\theta).$$

Any point $\theta$ that satisfies the above is called a maximum likelihood estimate (MLE), and is denoted $\hat{\theta}_{\text{MLE}}(x)$. In many cases there is a single point where this happens, so the MLE is unique, and we can talk about *the* MLE. Note that since log is a monotone transform, we also have $\ell_x(\hat{\theta}_{\text{MLE}}(x)) = \max_{\theta \in \Theta} \ell_x(\theta)$, i.e., the MLE maximizes the log-likelihood function over $\Theta$.

In ML inference, a summary of the likelihood function begins with the reporting of the MLE (provided it exists and is unique). The popularity of ML inference is partly due to the fact that finding the MLE is an optimization problem, which is both elegant and well understood. Optimization is one mathematical problem that can be routinely solved either analytically or with the help of a computer, thanks to an array of very powerful numerical algorithms that have been developed over centuries (starting from Newton).

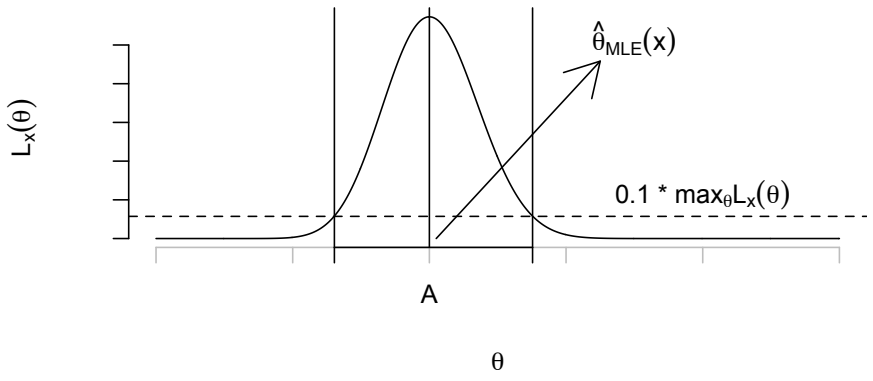


Figure 1: A schematic representation of ML inference. Solid curve is the likelihood function. Horizontal dashed line gives separation by $0.1 \times \max_{\theta \in \Theta} L_x(\theta)$. The corresponding set $A$ of well-supported $\theta$ are marked by the black segment on the horizontal axis. The MLE is marked with the arrow.

**Finding the MLE**

A standard technique to find the MLE relies on the following observation. If $L_x(\theta)$, or equivalently, $\ell_x(\theta)$ is a differentiable function over $\Theta$ with a unique maxima inside $\Theta$, then its first derivative vanishes at the maximum. Thus, if $\theta$ is a $p$-dimensional vector $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ then the MLE $\hat{\theta}_{\text{MLE}}(x)$ can be found by solving the simultaneous equations

$$\frac{\partial}{\partial \theta_j} \ell_x(\theta) = 0, \quad j = 1, 2, \cdots, p,$$

in $\theta$. In many cases these equations can be solved analytically, and we'd see some examples shortly. In many other cases, these equations can be solved by running a suitable computer algorithm.

**Example** (Opinion poll)**.** In our opinion poll example data $X$ is modeled by $\mathsf{Binomial}(n, p)$, $p \in [0, 1]$ and the likelihood function based on observation $X = x$ is given by

$$L_x(p) = \binom{n}{p} p^x (1-p)^{n-x}$$

and so the log-likelihood function is given by

$$\ell_x(p) = \text{const} + x \log p + (n - x) \log(1 - p),$$

with $p \in [0, 1]$. To find the MLE we set up the equation

$$0 = \frac{\partial}{\partial p} \ell_x(p) = \frac{x}{p} - \frac{n-x}{1-p}$$

which is solved at $p = x/n$. Hence $\hat{p}_{\text{MLE}}(x) = x/n$. For $n = 500$ and observed data $X = 200$, the MLE is 0.40. This is the researcher's "estimate", based on the ML approach, of the unknown proportion of supporters in the entire college.

**Example** (Lactic acid in cheese)**.** A cheese manufacturer wants to quantify the lactic acid concentration in cheese produced at one of his factories. He gets $n$ randomly chosen pieces of cheese measured and their lactic acid concentrations $X_1, \cdots, X_n$ recorded. Suppose these data are to be modeled as $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, i.e., each $X_i$ is described by the $\mathsf{Normal}(\mu, \sigma^2)$ pdf $g(x_i|\mu, \sigma^2) = (2\pi\sigma)^{-1/2} \exp\{-(x_i - \mu)^2/(2\sigma^2)\}$, and they are described independently of each other. So the pdf of the data $X$ at any $x = (x_1, \cdots, x_n)$ is given by:

$$f(x|\mu, \sigma^2) = \prod_{i=1}^{n} g(x_i|\mu, \sigma^2).$$

Here $\mu$ gives the overall concentration of lactic acid, which is the quantity of interest, and $\sigma$ explains the variability from one piece to another. Suppose, it is known that the variability $\sigma = 1/3$. So our statistical model is $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2), \mu \in (-\infty, \infty), \sigma = 1/3$.

Let's start by writing the log-likelihood function

$$\begin{aligned}
\ell_x(\mu) = \log f(x|\mu, \sigma^2) &= \sum_{i=1}^{n} \log g(x_i|\mu, \sigma^2) \\
&= \sum_{i=1}^{n} \left[ -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(x_i - \mu)^2}{2\sigma^2} \right] \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}
\end{aligned}$$

which is a quadratic function in $\mu$ (here $\sigma = 1/3$ is known, but we retain the symbol $\sigma$ to keep the calculations general and adaptable to other values of $\sigma$). At this stage we use the identity that for any $n$ numbers $x_1, \cdots, x_n$ and another number $a$,

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - a)^2$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ is the average of $x_1, \cdots, x_n$. Using this above we see

$$\begin{aligned}
\ell_x(\mu) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2} \\
&= \text{const} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}
\end{aligned}$$

where "const" absorbs all additive terms that do not involve the argument $\mu$ of the log-likelihood function.

To find the MLE we now set up the equation

$$0 = \frac{\partial}{\partial\mu}\ell_x(\mu) = \frac{n(\bar{x} - \mu)}{\sigma^2}$$

which is solved at $\mu = \bar{x}$ and hence $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$. If the manufacturer had $n = 10$ pieces measured to have concentrations $(0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58)$, then his MLE of the overall concentration is the average of these 10 numbers, 1.379.

**Example** (Lactic acid concentration in cheese, Cond.)**.** Now consider the case where the variability is also unknown along with the overall concentration $\mu$. Now our statistical model is $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2), (\mu, \sigma^2) \in (-\infty, \infty) \times (0, \infty)$. Working as before we get

$$\ell_x(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}.$$

To find the MLE we set up the equations:

$$\begin{aligned} 0 = \frac{\partial}{\partial \mu}\ell_x(\mu, \sigma^2) &= \frac{n(\bar{x} - \mu)}{\sigma^2} \\ 0 = \frac{\partial}{\partial \sigma^2}\ell_x(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2(\sigma^2)^2} + \frac{n(\bar{x} - \mu)^2}{2(\sigma^2)^2} \end{aligned}$$

which are solved at $\mu = \bar{x}$, $\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$. Hence $\hat{\mu}_{\text{MLE}}(x) = \bar{x}$ and $\hat{\sigma}^2_{\text{MLE}}(x) = \sum_{i=1}^n (x_i - \bar{x})^2/n$. We shall denote the latter quantity, which is a measure of the variability among $\{x_1, \cdots, x_n\}$, by $v_x$. Going back to the 10 measurements reported above, we see the MLE of $\mu$ remains 1.379 and now we also have the MLE of $\sigma^2$ equal to $0.096 = 0.31^2$. So the manufacturer's ML estimates of concentration and variability are 1.379 and 0.31.

**Example** (Annual TC counts)**.** As a final example consider inference on trend of tropical cyclone counts, where annual counts $X_1, \cdots, X_n$ from $n$ consecutive years are modeled as $X_t \overset{\text{IND}}{\sim} \mathsf{Poisson}(\exp\{\alpha + \beta(t-1)\})$, $(\alpha, \beta) \in (-\infty, \infty)^2$. Here the expected counts $\mu_t = \exp\{\alpha + \beta(t-1)\}$ in years $t$ satisfy the growth equations: $\mu_1 = e^\alpha$, $\mu_t = \mu_{t-1}e^\beta$. Based on observations $X_t = x_t$, $t = 1, \cdots, n$, the log-likelihood function is:

$$\begin{aligned} \ell_x(\alpha, \beta) &= \sum_{t=1}^n \log(e^{-\mu_t}\mu_t^{x_t}/x_t!) = -\sum_{t=1}^n \mu_t + \sum_{t=1}^n x_t \log\mu_t - \sum_{t=1}^n \log x_t! \\ &= -\sum_{t=1}^n \exp\{\alpha + \beta(t-1)\} + (\alpha - \beta)\sum_{t=1}^n x_t + \beta\sum_{t=1}^t t x_t + \text{const.} \end{aligned}$$

To find the MLE, we could set $\frac{\partial}{\partial\alpha}\ell_x(\alpha, \beta) = 0$, $\frac{\partial}{\partial\beta}\ell_x(\beta) = 0$ and solve. But it is difficult to derive the solution analytically. Instead we can use iterative optimization routines to directly maximize $\ell_x(\alpha, \beta)$. The `R` package has such a routine called `optim()`.

As observed data, we use the recorded TC counts in the north Atlantic between 1908 and 2007. For these $n = 100$ records, we have $\sum_{t=1}^n x_t = 932$ and $\sum_{t=1}^n t x_t = 51884$. For these observed data, `optim()` computes (see table 1) the MLE to be $(1.9, 6.2 \times 10^{-3})$. Accordingly, the estimated expected annual counts $\mu_t$ satisfy $\hat{\mu}_1 = \exp(1.9) = 6.7$, with a growth rate $e^{\hat{\beta}} - 1 = 0.63\%$ (see Figure 2).

```
## Code for computing MLE for the TC counts.
## Note: optim() performs minimization, so work with negative log-likelihood.

> n <- 100; t <- 1:n; sum.x <- 932; sum.tx <- 51884
> neg.log.lik <- function(par){
+   alpha <- par[1]; beta <- par[2]; mu <- exp(alpha + beta * (t - 1))
+   return(sum(mu) - (alpha - beta) * sum.x - beta * sum.tx)
+ }
> o <- optim(c(0,0), neg.log.lik)
> print(o)
$par
[1] 1.9068 0.0062

$value
[1] -1163

$counts
function gradient
      89       NA

$convergence
[1] 0

$message
NULL

> alpha.hat <- o$par[1]; beta.hat <- o$par[2];
> print(alpha.hat); print(beta.hat)
[1] 1.9
[1] 0.0062
```

Table 1: R code for computing the MLE of $(\alpha, \beta)$ for the annual TC count model. You are **not required** to learn this code. This is presented here only for those curious to know how MLE can computed numerically, using computer programs.

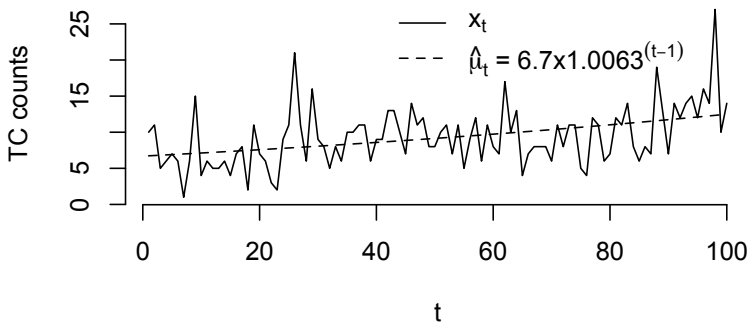


Figure 2: Observed annual north-Atlantic TC counts and ML estimate of expected counts for the log-linear Poisson model described in the text. Data are from 1908 through 2007.

**Sufficient statistics**

It is interesting to note that for the annual TC count model, the observed data $(x_1, \cdots, x_n)$ affect the log-likelihood function through only two summaries $\sum_t x_t$ and $\sum_t t x_t$. Any summary of the data is called a **statistic**. For a statistical model $\{f(x|\theta) : \theta \in \Theta\}$ , if there is a vector of statistics $T(x) = (T_1(x), \cdots, T_m(x))$ such that for some functions $h(x)$ and $g(t, \theta)$,

$$\ell_x(\theta) = h(x) + g(T(x), \theta), \theta \in \Theta$$

for every possible observation $X = x$, then $T(x)$ is called a vector of sufficient statistics for the model. For the TC counts model $(\sum_t x_t, \sum_t t x_t)$ is a vector of sufficient statistics. For the lactic acid concentration model with known variability, $\bar{x}$ is a sufficient statistic. For the lactic acid concentration model with unknown concentration and variability, $(\bar{x}, v_x)$ is a vector of sufficient statistics. It should be obvious that any likelihood based inference of $\theta$ depends on data only through the sufficient statistics.