I02 - Likelihood

STAT 587 (Engineering)
Iowa State University

March 30, 2021

Statistical modeling

A statistical model is a pair $(\mathcal{S}, \mathcal{P})$ where $\mathcal{S}$ is the set of possible observations, i.e. the sample space, and $\mathcal{P}$ is a set of probability distributions on $\mathcal{S}$.

Typically, assume a parametric model

$$p(y|\theta)$$

where

- y is our data and
- θ is unknown parameter vector.

The

- allowable values for θ determine $\mathcal{P}$ and
- the support of $p(y|\theta)$ is the set $\mathcal{S}$.

Binomial model

Suppose we will collect data were we have

- the number of success y
- out of some number of attempts n
- where each attempt is independent
- with a common probability of success θ.

Then a reasonable statistical model is

$$Y \sim Bin(n, \theta).$$

Formally,

- $\mathcal{S} = \{0, 1, 2, \ldots, n\}$ and
- $\mathcal{P} = \{Bin(n, \theta) : 0 < \theta < 1\}$.

Normal model

Suppose we have one datum

- real number,
- has a mean μ and variance σ^2, and
- uncertainty is represented by a bell-shaped curve.

Then a reasonable statistical model is

$$Y \sim N(\mu, \sigma^2).$$

Marginally,

- $\mathcal{S} = \{y : y \in \mathbb{R}\}$
- $\mathcal{P} = \{N(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ where $\theta = (\mu, \sigma^2)$.

Normal model

Suppose our data are

- n real numbers,
- each has a mean μ and variance is σ^2,
- a histogram is reasonably approximated by a bell-shaped curve, and
- each observation is independent of the others.

Then a reasonable statistical model is

$$Y_i \overset{ind}{\sim} N(\mu, \sigma^2).$$

Marginally,

- $\mathcal{S} = \{(y_1, \ldots, y_n) : y_i \in \mathbb{R}, i \in \{1, 2, \ldots, n\}\}$
- $\mathcal{P} = \{N_n(\mu, \sigma^2 \mathrm{I}) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ where $\theta = (\mu, \sigma^2)$.

Likelihood

The likelihood function, or simply likelihood, is the joint probability mass/density function for fixed data when viewed as a function of the parameter (vector) θ. Generically, let $p(y|\theta)$ be the joint probability mass/density function of the data and thus the likelihood is

$$L(\theta) = p(y|\theta)$$

but where y is fixed and known, i.e. it is your data.

The log-likelihood is the (natural) logarithm of the likelihood, i.e.

$$\ell(\theta) = \log L(\theta).$$

Intuition: The likelihood describes the relative support in the data for different values for your parameter, i.e. the larger the likelihood is the more consistent that parameter value is with the data.

Binomial likelihood

Suppose $Y \sim Bin(n, \theta)$, then

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

where θ is considered fixed (but often unknown) and the argument to this function is y.

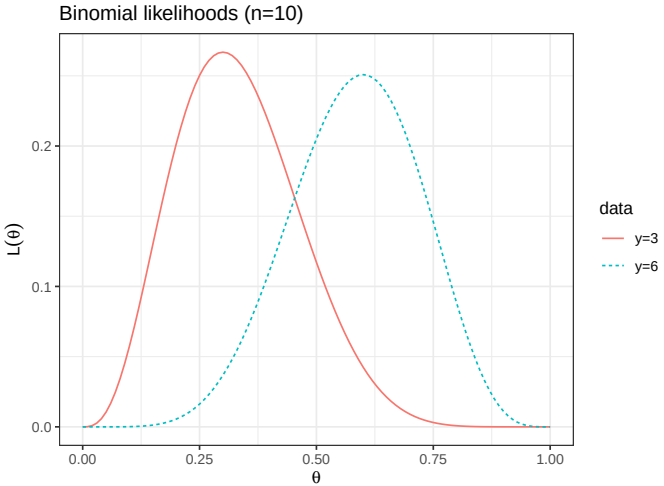
Thus the likelihood is

$$L(\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

where y is considered fixed and known and the argument to this function is θ.

Note: I write $L(\theta)$ without any conditioning, e.g. on y, so that you don't confuse this with a probability mass (or density) function.

Binomial likelihood



Likelihood for independent observations

Suppose Y_i are independent with marginal probability mass/density function $p(y_i|\theta)$.

The joint distribution for $y = (y_1, \ldots, y_n)$ is

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

The likelihood for θ is

$$L(\theta) = p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

where we are thinking about this as a function of θ for fixed y.

Normal model

Suppose $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

$$p(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

and

$$\begin{aligned} p(y|\mu, \sigma^2) &= \prod_{i=1}^n p(y_i|\mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2} \end{aligned}$$

where μ and σ^2 are fixed (but often unknown) and the argument to this function is $y = (y_1, \ldots, y_n)$.

Normal likelihood

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

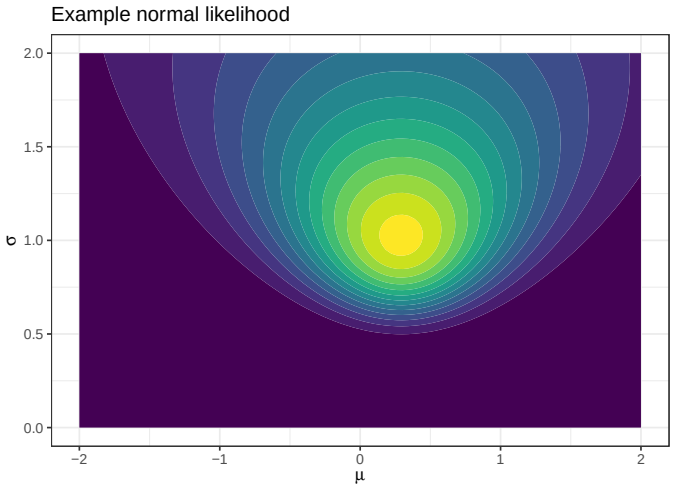
$$p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

The likelihood is

$$L(\mu, \sigma) = p(y|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

where y is fixed and known and μ and σ^2 are the arguments to this function.

Normal likelihood - example contour plot



Maximum likelihood estimator (MLE)

Definition

The maximum likelihood estimator (MLE), $\hat{\theta}_{MLE}$ is the parameter value θ that maximizes the likelihood function, i.e.

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}\, L(\theta).$$

When the data are discrete, the MLE maximizes the probability of the observed data.

Binomial MLE - derivation

If $Y \sim Bin(n, \theta)$, then

$$L(\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

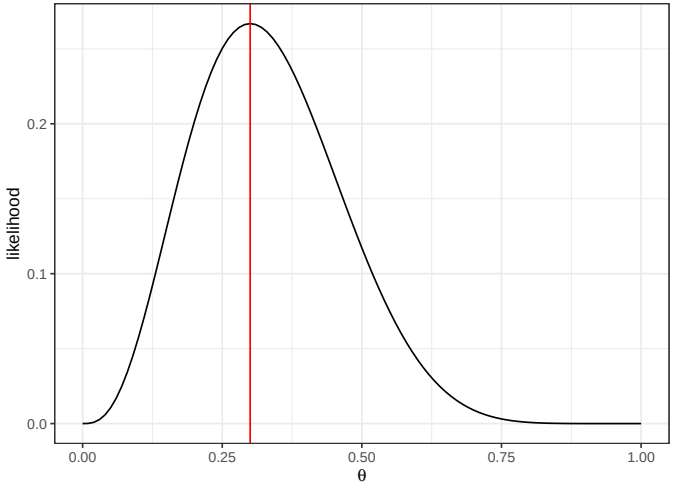
To find the MLE,

1. Take the derivative of $\ell(\theta)$ with respect to θ.
2. Set it equal to zero and solve for θ.

$$\begin{aligned} \ell(\theta) &= \log \tbinom{n}{y} + y \log(\theta) + (n-y) \log(1-\theta) \\ \tfrac{d}{d\theta}\ell(\theta) &= \tfrac{y}{\theta} - \tfrac{n-y}{1-\theta} \overset{set}{=} 0 \implies \\ \hat{\theta}_{MLE} &= y/n \end{aligned}$$

Take the second derivative of $\ell(\theta)$ with respect to θ and check to make sure it is negative.

Binomial MLE - graphically



Binomial MLE - Numerical maximization

```
log_likelihood <- function(theta) {
  dbinom(3, size = 10, prob = theta, log = TRUE)
}

o <- optim(0.5, log_likelihood,
           method='L-BFGS-B',            # this method to use bounds
           lower = 0.001, upper = .999,  # cannot use 0 and 1 exactly
           control = list(fnscale = -1)) # maximize

o$convergence # 0 means convergence was achieved

[1] 0

o$par         # MLE

[1] 0.3000006

o$value       # value of the likelihood at the MLE

[1] -1.321151
```

Normal MLE - derivation

If $Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$, then

$$
\begin{aligned}
L(\mu, \sigma^2) &= \tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2} \\
&= \tfrac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\overline{y}+\overline{y}-\mu)^2} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n \left[(y_i-\overline{y})^2 + 2(y_i-\overline{y})(\overline{y}-\mu) + (\overline{y}-\mu)^2\right]\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\overline{y})^2 + -\tfrac{n}{2\sigma^2}(\overline{y}-\mu)^2\right) \quad \text{since } \textstyle\sum_{i=1}^n (y_i-\overline{y}) = 0 \\
\ell(\mu, \sigma^2) &= -\tfrac{n}{2}\log(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\overline{y})^2 - \tfrac{1}{2\sigma^2} n(\overline{y}-\mu)^2 \\
\tfrac{\partial}{\partial\mu}\ell(\mu, \sigma^2) &= \tfrac{n}{\sigma^2}(\overline{y}-\mu) \overset{set}{=} 0 \implies \hat{\mu}_{MLE} = \overline{y} \\
\tfrac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) &= -\tfrac{n}{2\sigma^2} + \tfrac{1}{2(\sigma^2)^2}\sum_{i=1}^n (y_i-\overline{y})^2 \overset{set}{=} 0 \\
&\implies \hat{\sigma}^2_{MLE} = \tfrac{1}{n}\sum_{i=1}^n (y_i-\overline{y})^2 = \tfrac{n-1}{n} S^2
\end{aligned}
$$

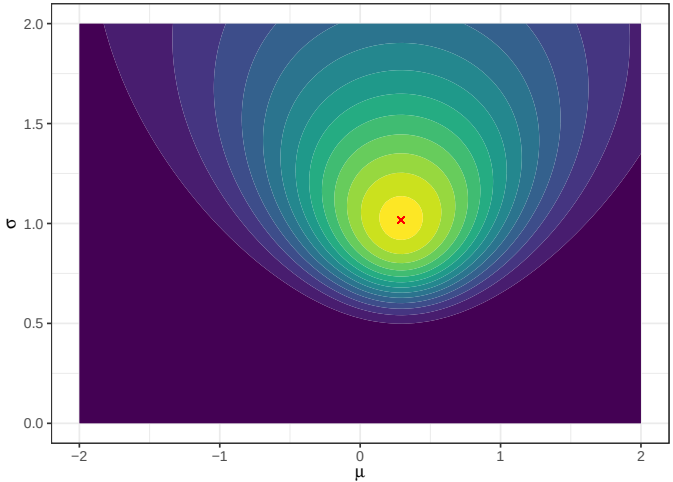
Thus, the MLE for a normal model is

$$\hat{\mu}_{MLE} = \overline{y}, \quad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2$$

Normal MLE - numerical maximization

```
x

[1] -0.8969145  0.1848492  1.5878453

log_likelihood <- function(theta) {
  sum(dnorm(x, mean = theta[1], sd = exp(theta[2]), log = TRUE))
}

o <- optim(c(0,0), log_likelihood,
           control = list(fnscale = -1))
c(o$par[1], exp(o$par[2])^2)                # numerical MLE

[1] 0.2918674 1.0344601

n <- length(x); c(mean(x), (n-1)/n*var(x)) # true MLE

[1] 0.2919267 1.0347381
```

Normal likelihood - graph



Summary

- For independent observations, the joint probability mass (density) function is the product of the marginal probability mass (density) functions.
- The likelihood is the joint probability mass (density) function when the argument of the function is the parameter (vector).
- The maximum likelihood estimator (MLE) is the value of the parameter (vector) that maximizes the likelihood.