Maximum Likelihood Estimate

Tutorial 18th February 2020

Preliminaries - Population and Sample

- A **population** includes all of the elements from a set of data.
- A **sample** consists one or more observations drawn from the population.

- A measurable characteristic of a population, such as a mean or standard deviation, is called a **parameter**; but a measurable characteristic of a sample is called a **statistic**.
  - Parameter is **fixed number**.
  - Statistic is a **random variable** as it depends upon the particular random sample. This is used to estimate the parameter.

Preliminaries - Estimator

An **estimator** is a statistic that estimates the value of some parameter of the population.

For example, the sample mean($\bar{x}$) is an estimator for the population mean, μ.

Since it is a statistic, it is a random variable.

f(D) and f(D=d) used interchangeably. Similarly, P(D) and P(D=d).

Preliminaries - Binomial Distribution

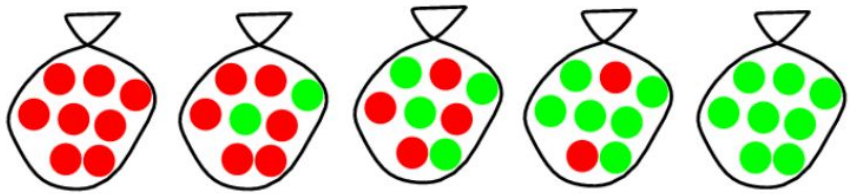
A binomial distribution can be thought of as simply the probability of a **success** or **failure** outcome in an experiment or survey that is repeated **multiple** times.

$$P(c \text{ successes}, (N-c)\text{failures}) = \binom{N}{c}\theta^c(1-\theta)^{N-c}$$

Binomial Model

Suppose there are five kinds of bags of lollies from Russell and Norvig):

1. 10% are h_1: 100% cherry lollies
2. 20% are h_2: 75% cherry lollies + 25% lime lollies
3. 40% are h_3: 50% cherry lollies + 50% lime lollies
4. 20% are h_4: 25% cherry lollies + 75% lime lollie
5. 10% are h_5: 100% lime lollies



Then we observe lollies drawn from some bag:

What kind of bag is it? What flavour will the next lolly be?

To answer these questions, we will first have to fit a model to the data

- Bags have a fraction θ of cherry lollies
- We are therefore dealing with binomial models (cherry vs lime lollies) in which we do not know θ. We will take this set of models to be characterised by the *parameter* θ
- Now we unwrap N lollies, and find c and $N - c$ limes. We will have to assume that these are i.i.d. (independent, identically distributed) observations
- What can we say about the probability of observed data, using the binomial distribution as our theoretical model. This is:

$$\text{Prob}(c \text{ cherries and } (N-c) \text{ limes } \propto \theta^c(1-\theta)^{(N-c)}$$

- Question: For what value of θ will this probability be highest?

Likelihood

Given the parameters, the probability that sample data is generated.

$$Likelihood \longrightarrow P(X_1, X_2, \ldots X_n|\theta)$$

$$Probability \longrightarrow P(\theta|X_1, X_2, \ldots X_n)$$

The Maximum Likelihood Estimator (MLE)

Let X_1, X_2, X_3, ..., X_n be a random sample from a distribution with a parameter θ. Given that we have observed $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$, a maximum likelihood estimate of θ, shown by $\hat{\theta}_{ML}$ is a value of θ that maximizes the likelihood function

$$L(x_1, x_2, \cdots, x_n; \theta).$$

A maximum likelihood estimator (MLE) of the parameter θ, shown by $\hat{\Theta}_{ML}$ is a random variable $\hat{\Theta}_{ML} = \hat{\Theta}_{ML}(X_1, X_2, \cdots, X_n)$ whose value when $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$ is given by $\hat{\theta}_{ML}$.

Why use MLE?

Ans: Find maximum by differentiating and setting first differential to 0. Actually easier to differentiate $\log(P)$ and set that to 0:

$$\log(P) \;=\; L(P) \;=\; c\log\theta \;+\; (N-c)\log(1-\theta)$$

Differentiating w.r.t. θ and setting this to zero:

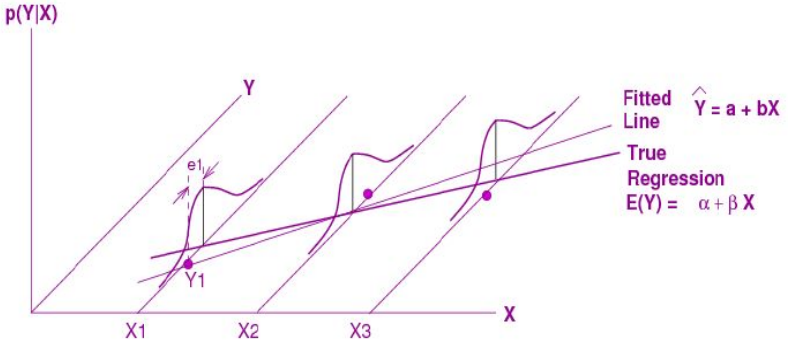
$$\frac{dL(P)}{d\theta} \;=\; \frac{c}{\theta} - \frac{N-c}{1-\theta} \;=\; 0$$

which gives $\theta = c/N$

This is the "Maximum Likelihood Estimate" for θ ($L(P)$ is called the likelihood function)
(Seems sensible, but causes problems with 0 counts! But more on that later.)

Linear Gaussian Model

Recall the regression model:



The probability model being assumed is:

$$Y_i = \alpha + \beta X_i + e_i$$

where e_i are distributed with mean 0 and variance σ^2. In addition, we are further assuming that the frequency distribution of the e_i can be approximated using a Gaussian distribution

That is, we are assuming that $P(Y_i|X_i)$ is a Gaussian distribution with mean $\alpha + \beta X_i$ and variance σ^2:

$$P(Y_i|X_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - f(X_i))^2}{2\sigma^2}}$$

(where $f(X_i) = \alpha + \beta X_i$)

Assume we are given a set of points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$. Then the probability of obtaining these points is:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_1{}^n (y_i - f(x_i))^2}$$

This is the likelihood function. Maximising this, will require minimising $\sum_1{}^n (y_i - f(x_i))^2$, which is the same as finding the least squares estimate

So, the least squares estimators for the regression line are the same as the maximum likelihood estimators for that linear Gaussian model (with i.i.d. data, and fixed variance)

The MLE of the population mean is the sample mean. The sample mean is statistically unbiased, so the ML principle results in an unbiased estimate of the population mean

However, the MLE of the variance is not unbiased (that is, the ML estimator is biased). So, it is not always the case that the ML principle results in an estimate with 0 bias. So, what can we say about ML estimators?

As the sample size gets large, the variance of the MLE tends to the CR bound v_{min}. So, for all unbiased estimators (that is, all estimators that have $b = 0$), the MLE will have the lowest MSE (for large samples)

Poisson Distribution

Let $x_1, x_2, \ldots, x_n$ be a sample of observations from a Poisson distribution with parameter λ. Find the maximum likelihood estimate of λ in terms of the x_i and n.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$L(\lambda) = P(D|\lambda) = \Pi_{i=1}^{n} P(X_i|\lambda)$$

$$\lambda_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Exercise

Let $x_1, x_2, \ldots, x_n$ be a sample from an exponential distribution, which has a density function $f(X = x) = \lambda e^{-\lambda x}$ $(x > 0)$. Derive a maximum likelihood estimate of λ in terms of the x_i and n.

$$L(\lambda) = \lambda^n \exp^{-\lambda \sum_{i=1}^{n} x_i}$$

$$l(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{dl}{d\lambda}(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\lambda_{MLE} = \frac{n}{\sum x_i}$$

Exercise

Let $x_1, x_2, \ldots, x_n$ be observations from a normal distribution with parameters μ and σ^2. Derive maximum likelihood estimates of μ and σ^2.

$$
\begin{aligned}
L(\mu,\sigma^2;x_1,\ldots,x_n) &= \prod_{j-1}^{n} f_X(x_j;\mu,\sigma^2) \\
&= \prod_{j-1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2}\frac{(x_j-\mu)^2}{\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2\right)
\end{aligned}
$$

$$
\begin{aligned}
l(\mu,\sigma^2;x_1,\ldots,x_n) &= \ln\left(L(\mu,\sigma^2;x_1,\ldots,x_n)\right) \\
&= \ln\left(\left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2\right)\right) \\
&= \ln\left(\left(2\pi\sigma^2\right)^{-n/2}\right)+\ln\left(\exp\left(-\frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2\right)\right) \\
&= -\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2 \\
&= -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2
\end{aligned}
$$

We need to solve the following maximization problem

$$\max_{\mu,\sigma^2} l\left(\mu,\sigma^2;x_1,\ldots,x_n\right)$$

The first order conditions for a maximum are

$$\frac{\partial}{\partial\mu} l\left(\mu,\sigma^2;x_1,\ldots,x_n\right) = 0$$

$$\frac{\partial}{\partial\sigma^2} l\left(\mu,\sigma^2;x_1,\ldots,x_n\right) = 0$$

$$\frac{\partial}{\partial \mu} l\left(\mu, \sigma^2; x_1, \ldots, x_n\right)$$

$$= \frac{\partial}{\partial \mu}\left(-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j - \mu)^2\right)$$

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

$$= \frac{1}{\sigma^2}\left(\sum_{j-1}^{n} x_j - n\mu\right)$$

$$\frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2;x_1,\ldots,x_n)$$

$$= \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2 \right)$$

$$= -\frac{n}{2\sigma^2} - \left[\frac{1}{2}\sum_{j-1}^{n}(x_j-\mu)^2 \right]\frac{d}{d\sigma^2}\left(\frac{1}{\sigma^2}\right)$$

$$= -\frac{n}{2\sigma^2} - \left[\frac{1}{2}\sum_{j-1}^{n}(x_j-\mu)^2 \right]\left(-\frac{1}{(\sigma^2)^2}\right)$$

$$= -\frac{n}{2\sigma^2} + \left[\frac{1}{2}\sum_{j-1}^{n}(x_j-\mu)^2 \right]\frac{1}{(\sigma^2)^2}$$

$$= \frac{1}{2\sigma^2}\left[\frac{1}{\sigma^2}\sum_{j-1}^{n}(x_j-\mu)^2 - n \right]$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{j-1}^{n} x_j$$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{j-1}^{n} (x_j - \hat{\mu})^2$$

Exercise - Biased Coin

- **N = 10**
- **7 heads and 3 tails in 10 tosses**
- **Assume data comes from Binomial(N, p)**

Find MLE of p.

Exercise - Multiple Experiment with a biased coin

- **Three experiments each of 10 trials**
- **1st Experiment: 7 heads and 3 tails**
- **2nd Experiment: 6 heads and 4 tails**
- **3rd Experiment: 8 heads and 2 tails**

Find MLE of p.

Exercise - Biased Die

- **$\{X_i\}$ be N i.i.d trail outcomes s.t, $N = n_1 + n_2 \ldots + n_6$.**
- **Assume $X_i \sim$ Multinomial$(N, \theta_1 + \theta_2 \ldots + \theta_6)$**

Find MLE of θ.

$$P(n_1, n_2, \ldots, n_k) = \frac{N!}{n_1! n_2! \ldots n_k!} \theta_1^{n_1} \theta_2^{n_2} \ldots \theta_k^{n_k}$$

Thank You

Resources

- StatQuest: Probability vs Likelihood
- StatQuest: Maximum Likelihood
- StatQuest: Maximum Likelihood For the Normal Distribution
- MIT OCW Maximum Likelihood Estimates*
- UWashington Maximum Likelihood Estimates*

* - read only what is relevant