Maximum Likelihood Estimation

Bruce M. Boghosian

Department of Mathematics

Tufts University

Outline

Motivation

Bruce M. Boghosian

- Suppose we know the results of a repeated random experiment.
- We have a priori knowledge of the *form* of the probability function.
- We do not know the values of the *parameter(s)* of the distribution.
- Can we use the experimental results to estimate the parameter(s)?
- In this lecture, we learn to do so by maximizing a quantity called *likelihood*.
- This method is called *maximum likelihood estimation*.
- It is best to learn this method by example, so we present examples using a variety of probability functions.

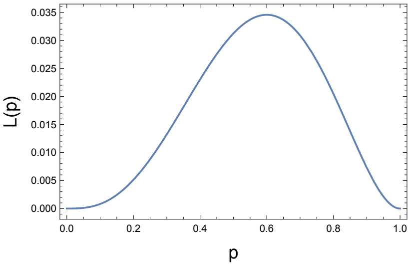
A random experiment

- A coin lands on heads with probability p, tails with probability $1 - p$.
- You are not given the value of p.
- You flip the coin five times and find the sequence $HTHHT$.
- Suppose this outcome is *all you know* about the coin.
- What is your best guess for p?

A posteriori calculation of probability of the outcome

Bruce M. Boghosian

- We have done the experiment and we know the result $HTHHT$.
- The coin flips may be assumed to have been independent.
- The *likelihood* of result $HTHHT$ is $L(p) = p(1-p)pp(1-p) = p^3(1-p)^2$.
- Note that $L(p)$ is not a probability density function!
- For which value of p is $L(p)$ maximized?



Finding the maximum likelihood

- We have $L(p) = p^3(1-p)^2$
- Derivative $L'(p) = p^2(1-p)(3-5p)$ has roots $p=0$, $p = {}^3\!/_5$ and $p=1$.
- Second derivative $L''(p) = 2p(10p^2 - 12p + 3)$, so $L''\left({}^3\!/_5\right) = -{}^{18}\!/_{25} < 0$
- Second derivative is negative at $p = {}^3\!/_5$, indicating a maximum at that point.
- *Maximum likelihood* occurs for $p = \boxed{p_e = {}^3\!/_5}$.

General Bernoulli trial

Bruce M. Boghosian

- Define random variable for each coin toss,

$$X := \begin{cases} 1 & \text{if toss results in heads (with probability } p\text{)} \\ 0 & \text{if toss results in tails (with probability } 1-p\text{).} \end{cases}$$

- Discrete probability function for one toss, where $k \in \{0,1\}$,

$$p_X(k) = \mathrm{Prob}(X = k) = p^k(1-p)^{1-k}$$

- Normalization: $\sum_{k=0}^{1} p_X(k) = (1-p) + p = 1$
- Mean: $\sum_{k=0}^{1} p_X(k)k = (1-p)0 + p1 = p$
- Variance: $\sum_{k=0}^{1} p_X(k)k^2 - p^2 = (1-p)0^2 + p1^2 - p^2 = p - p^2 = p(1-p)$

Defining the likelihood for n tosses

- *Likelihood* of n tosses with $\vec{X} = \langle X_1, X_2, \ldots, X_n \rangle$ equal to $\vec{k} = \langle k_1, k_2, \ldots, k_n \rangle$
- Product of discrete probability functions for observed data using parameter p,

$$
\begin{aligned}
L(p; \vec{k}) := \mathrm{Prob}(\vec{X} = \vec{k}) &= \prod_{j=1}^{n} p_{X_j}(k_j) \\
&= p^{k_1}(1-p)^{1-k_1} p^{k_2}(1-p)^{1-k_2} \cdots p^{k_n}(1-p)^{1-k_n} \\
&= \prod_{j=1}^{n} p^{k_j}(1-p)^{1-k_j} = p^K (1-p)^{n-K}
\end{aligned}
$$

where $K := \sum_{j=1}^{n} k_j$.

- We now wish to find the value of p that maximizes $L(p; \vec{k})$.

Maximizing the likelihood for n tosses

- Defining $K := \sum_{j=1}^{n} k_j$, the likelihood function is

$$L(p;\vec{k}) := p^K (1-p)^{n-K}.$$

- Note that $\vec{k}$ and hence K is known from the experimental outcome.
- We maximize the *log likelihood* with respect to p,

$$\ln L(p;\vec{k}) = K \ln p + (n-K)\ln(1-p)$$

- Setting derivative to zero yields

$$0 = \frac{d}{dp}\ln L(p;\vec{k}) = \frac{K}{p} - \frac{n-K}{1-p}.$$

- *Maximum likelihood* is $p = p_e := \frac{K}{n}$, so $\boxed{p_e = \frac{1}{n}\sum_{j=1}^{n} k_j}$

Estimator for p

Bruce M. Boghosian

- Maximum likelihood occurs when p is $p_e = \frac{1}{n}\sum_{j=1}^{n} k_j$
- Note that this is a function of the outcomes $\vec{k}$ that estimates the parameter p.
- Considered as a function of $\vec{k}$ yielding p, this is called an *estimator*,

$$\boxed{\hat{p}(\vec{k}) = \frac{1}{n}\sum_{j=1}^{n} k_j.}$$

- In this case, $\hat{p}(\vec{k})$ is just the *average* of the experimental outcomes $\vec{k}$.
- Here and henceforth, we use the "hat" to denote estimator functions.
- More generally, $L(p;\vec{k})$ is maximized for $p = \hat{p}(\vec{k})$.
- This approach is called *maximum likelihood estimation*.
- It estimates one or more parameters of known probability functions.
- Note that there must be a priori knowledge of the form of $p_X(k)$.

The Poisson distribution

- Sample space S is nonnegative integers.
- Poisson random variable $X \in S = \{0, 1, 2, \ldots\}$
- We have

$$p_X(k) = \text{Prob}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- Normalization

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{+\lambda} = 1.$$

- Mean

$$E(X) = \sum_{k=0}^{\infty} p_X(k) k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} k = \lambda$$

The Poisson distribution (continued)

- Variance

$$E(X^2) - E(X)^2 = \sum_{k=0}^{\infty} p_X(k)k^2 - \lambda^2 = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} k^2 - \lambda^2 = \lambda$$

- Again, note that this is a one-parameter probability function.
- The mean and variance are both equal to the parameter λ.

- Now, suppose we are given $n = 50$ samples from this distribution

$$3, 4, 3, 5, 3, 4, 5, 0, 5, 5, 5, 4, 2, 2, 3, 2, 2, 5, 5, 5, 5, 6, 5, 2, 5,$$
$$2, 2, 0, 6, 3, 6, 7, 2, 4, 5, 1, 3, 3, 5, 1, 7, 2, 0, 8, 4, 5, 8, 2, 3, 5$$

- Can we estimate λ using maximum likelihood estimation?

Defining the likelihood for n samples

Bruce M. Boghosian

- Given the results of an experiment $\vec{k} = \langle k_1, k_2, \ldots \rangle$ where $k_j \in \{0, 1, 2, \ldots\}$
- Define the likelihood

$$L(\lambda; \vec{k}) := \text{Prob}(\vec{X} = \vec{k}) = \prod_{j=1}^{n} p_{X_j}(k_j) = \prod_{j=1}^{n} e^{-\lambda} \frac{\lambda^{k_j}}{k_j!}$$

- Easier to maximize the log likelihood

$$\ln L(\lambda; \vec{k}) = \sum_{j=1}^{n} \left[-\lambda + k_j \ln \lambda - \ln\left(k_j!\right)\right] = -n\lambda + K \ln \lambda - \sum_{j=1}^{n} \ln\left(k_j!\right)$$

where, once again, $K := \sum_{j=1}^{n} k_j$.

Maximizing the likelihood for n samples

- Log likelihood

$$\ln L(\lambda; \vec{k}) = -n\lambda + K \ln \lambda - \sum_{j=1}^{n} \ln\left(k_j!\right)$$

- Log likelihood maximized for

$$0 = \frac{d}{d\lambda} \ln L(\lambda; \vec{k}) = -n + \frac{K}{\lambda}$$

- Result is $\lambda = \lambda_e := \frac{K}{n}$, or

$$\boxed{\lambda_e = \frac{1}{n} \sum_{j=1}^{n} k_j}$$

MLE for Poisson-distributed data

- Maximum likelihood estimator for the Poisson distribution is

$$\boxed{\hat{\lambda}(\vec{k}) = \frac{1}{n}\sum_{j=1}^{n} k_j}$$

- For the 50 points shown earlier, $\hat{\lambda}(\vec{k}) = 3.78$.
- Actual value of λ used to sample the points was 4.

$$3, 4, 3, 5, 3, 4, 5, 0, 5, 5, 5, 4, 2, 2, 3, 2, 2, 5, 5, 5, 5, 6, 5, 2, 5,$$
$$2, 2, 0, 6, 3, 6, 7, 2, 4, 5, 1, 3, 3, 5, 1, 7, 2, 0, 8, 4, 5, 8, 2, 3, 5$$

Some discussion

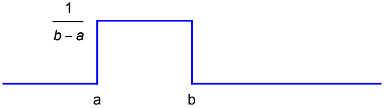
- So far, these results are not terribly surprising.
- The mean of the Bernoulli trials is p.
- The mean of the Poisson distribution is λ.
- Both MLE analyses estimate the parameter to be the sample mean.

Continuous random variable: The uniform distribution

Bruce M. Boghosian

- Suppose $X \in \mathbb{R}$ has the continuous *probability density function*,

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } X \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

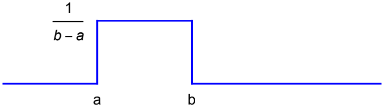


- Normalization: $\int_{\mathbb{R}} dx\ f_X(x) = \int_a^b dx\ \frac{1}{b-a} = \frac{b-a}{b-a} = 1$
- Mean: $\int_{\mathbb{R}} dx\ f_X(x)x = \int_a^b dx\ \frac{x}{b-a} = \frac{b+a}{2}$
- Variance: $\int_{\mathbb{R}} dx\ f_X(x)x^2 - \left(\frac{b+a}{2}\right)^2 = \int_a^b dx\ \frac{x^2}{b-a} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}$
- This is a two-parameter distribution, with parameters a and b.
- Given experimental results $\vec{x} = \{x_1, x_2, \ldots, x_n\}$, can we estimate a and b?

Defining the likelihood for n samples

- Suppose $X \in \mathbb{R}$ has the continuous *probability density function*,

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } X \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$



- Likelihood is

$$L(a,b;\vec{x}) = \prod_{j=1}^{n} f_X(x_j) = \begin{cases} \left(\frac{1}{b-a}\right)^n & \text{if } x_j \in [a,b] \text{ for all } j = 1,\ldots,n \\ 0 & \text{if } x_j \notin [a,b] \text{ for any } j = 1,\ldots,n \end{cases}$$

- Choose $a \leq \min_j x_j$ and $b \geq \max_j x_j$ so result is $\left(\frac{1}{b-a}\right)^n$.
- Maximize result by choosing $a_e = \min_j x_j$ and $b_e = \max_j x_j$.

Maximum likelihood estimation of a and b

Bruce M. Boghosian

- Maximize $L(a, b; \vec{x})$ by choosing $a_e = \min_j x_j$ and $b_e = \max_j x_j$.
- Maximum likelihood estimators for a and b are

$$\boxed{\hat{a}(\vec{x}) = \min_j x_j}$$

$$\boxed{\hat{b}(\vec{x}) = \max_j x_j}$$

Continuous random variable: The standard normal distribution

- Suppose $X \in \mathbb{R}$ has the continuous *probability density function*,

$$f_X(x) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-\mu)^2}{2v}\right],$$

  which we recognize to be normalized, with mean μ and variance v.
- This is a two-parameter distribution, with parameters μ and v.
- Given experimental results $\vec{x} = \{x_1, x_2, \ldots, x_n\}$, can we estimate μ and v?

Defining the likelihood for n samples

- Likelihood is

$$L(\mu, v; \vec{x}) = \prod_{j=1}^{n} f_X(x_j) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x_j-\mu)^2}{2v}\right]$$

- Log likelihood is

$$\ln L(\mu, v; \vec{x}) = \sum_{j=1}^{n}\left[-\frac{1}{2}\ln(2\pi v) - \frac{(x_j-\mu)^2}{2v}\right]$$

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

- We must find maximum with respect to both μ and v.

Maximum likelihood estimation of μ and v

Bruce M. Boghosian

- Log likelihood is $\ln L(\mu, v; \vec{x}) = -\frac{n}{2} \ln(2\pi v) - \frac{1}{2v} \sum_{j=1}^{n} (x_j - \mu)^2$
- Set partial derivatives to zero

$$0 = \frac{\partial}{\partial \mu} \ln L(\mu, v; \vec{x}) = \frac{1}{v} \sum_{j=1}^{n} (x_j - \mu) = \frac{1}{v} \left(\sum_{j=1}^{n} x_j - n\mu \right)$$

$$0 = \frac{\partial}{\partial v} \ln L(\mu, v; \vec{x}) = -\frac{n}{2v} + \frac{1}{2v^2} \sum_{j=1}^{n} (x_j - \mu)^2$$

- Solving for location of maximum (μ_e, v_e) yields

$$\mu_e = \frac{1}{n} \sum_{j=1}^{n} x_j \quad \text{and} \quad v_e = \frac{1}{n} \sum_{j=1}^{n} \left[x_j - \left(\frac{1}{n} \sum_{k=1}^{n} x_k \right) \right]^2$$

Maximum likelihood estimation of μ and v

- Maximum likelihood estimators for a and b are

$$\boxed{\hat{\mu}(\vec{x}) = \frac{1}{n}\sum_{j=1}^{n} x_j}$$

$$\boxed{\hat{v}(\vec{x}) = \frac{1}{n}\sum_{j=1}^{n}\left[x_j - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)\right]^2}$$

Summary

Bruce M. Boghosian

- We have learned the method of *maximum likelihood estimation*.
- Allows estimation of parameters if form of the distribution is known a priori.
- We have seen examples with discrete and continuous probability functions.
- We have seen examples with one and two parameters.
- We have seen examples where using the log likelihood is useful and not useful.
- In all cases, we have calculated *estimator functions*.