Tufts UNIVERSITY

Bruce M. Boghosian

Maximum Likelihood Estimation:

Motivation and a First Example

Bruce M. Boghosian

Tufts UNIVERSITY | School of Arts and Sciences

Department of Mathematics

Tufts University

Outline

Parameter estimation

Bruce M. Boghosian

- Say we know results of repeated random experiment.
- We have a priori knowledge of *form* of probability function.
- We do not know values of *parameter(s)* of distribution.
- Can we use experimental results to estimate parameter(s)?

Examples

Name	Distribution	Type	Parameters
Poisson	$p_X(k) = \lambda^k e^{-\lambda}/k!$	Discrete	λ
Normal	$f_X(x) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-(x-\mu)^2/(2\sigma^2)}$	Continuous	μ and σ

- Can we use experimental data to estimate parameter(s)?
  - Use $\vec{k} = \langle k_1, \ldots, k_n \rangle$ to estimate λ in first case.
  - Use $\vec{x} = \langle x_1, \ldots, x_n \rangle$ to estimate μ and σ in second case.
- If so, we have *estimator functions* (denoted with a hat):
  - $\hat{\lambda}(\vec{k})$ is the estimate for λ in first case.
  - $\hat{\mu}(\vec{x})$ and $\hat{\sigma}(\vec{x})$ are estimators for μ and σ in second case.
- Estimated *parameter values* denoted with subscript e
  - e.g., $\hat{\lambda}(\vec{k}) = \lambda_e$.

Maximum likelihood estimation

- In this module, we learn to estimate parameters by maximizing a quantity called *likelihood*. This method is called *maximum likelihood estimation*.
- It is best to learn this method by example.
- We give examples using variety of probability distributions.
- Method works for discrete and continuous distributions.
- Method works for one or more parameters.
- Method can be used to calculate *estimator functions*.

A random experiment

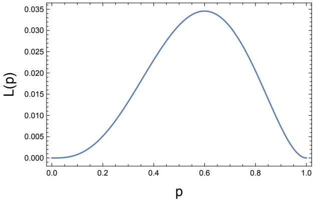
- A coin lands on
  - heads with probability p
  - tails with probability $1 - p$.
- You are not given the value of the parameter p.
- You flip the coin five times and find the sequence $HTHHT$.
- Suppose this outcome is *all you know* about the coin.
- What is your best guess for parameter p?

A posteriori calculation of probability of the outcome

- We have done experiment and we know result $HTHHT$.
- The coin flips are assumed to have been independent.
- *Likelihood* of $HTHHT$: $L(p) = p(1-p)pp(1-p) = p^3(1-p)^2$.
- Note $L(p)$ is not a probability density function!

$$\int_0^1 dp\ p^3(1-p)^2 \neq 1$$

- For which value of p is $L(p)$ maximized?



Finding the maximum likelihood

- We have
  - $L(p) = p^3(1-p)^2$
  - $L'(p) = p^2(1-p)(3-5p)$ has roots $p = 0, {}^3/_5, 1$.
  - $L''(p) = 2p(10p^2 - 12p + 3)$, 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 trials

- 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 distribution for one toss,

$$\boxed{p_X(k) = \text{Prob}(X = k) = p^k(1-p)^{1-k} \text{ where } k \in \{0, 1\}}$$

- 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: $\sum_{k=0}^{1} p_X(k)k^2 - p^2 = p - p^2 = p(1-p)$

Defining the likelihood for n tosses

- *Likelihood* of n tosses with $X_1 = k_1,\ X_2 = k_2,\ \ldots,\ X_n = k_n$
- Product of discrete probability functions for observed data using parameter p,

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

- Where $\overline{k} := \frac{1}{n}\sum_{j=1}^{n} k_j$ is the *average* of the data k_j.
- We now wish to find the value of p that maximizes $L(p;\vec{k})$.

Maximizing the likelihood for n tosses

Bruce M. Boghosian

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

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

- Note that $\vec{k}$ and hence $\overline{k}$ are known from experiment
- We maximize the *log likelihood* with respect to p,

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

- Setting derivative to zero yields

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

- Solving yields *maximum likelihood* at $\boxed{p_e := \overline{k} = \frac{1}{n}\sum_{j=1}^{n} k_j}$

Estimator for p

- Maximum likelihood occurs when $p = p_e = \overline{k} = \frac{1}{n}\sum_{j=1}^{n} k_j$.
- This is function of outcomes $\vec{k}$ that estimates parameter p.
- Considered as function $\vec{k} \mapsto p$, 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 *average* of experimental outcomes $\vec{k}$.
- We use a "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*.
- Estimates parameter(s) of known probability functions.
- There must be a priori knowledge of the form of $p_X(k)$.

Summary

Bruce M. Boghosian

- We have learned *estimation* for distribution parameter λ.
- We have learned what an *estimate* is, λ_e.
- We have learned what an *estimator* is, $\hat{\lambda}(\vec{x})$.
- There are many ways of finding estimators.
- We have learned *maximum likelihood estimation*.
- Allows estimation of parameters if form of the distribution is known a priori.
- We used the *log likelihood* to find a maximum likelihood estimate for iterated Bernoulli trials.