Parameter Estimation

Given a representative sample of data from some population, we may need to estimate the parameters for a distribution characterizing the population. We discus properties of estimators and the following estimation methods below:

1. Method of Moments.
2. Maximum Likelihood Estimation (MLE).
3. Maximum a posteriori probability estimate (MAP).

General references:

- Pattern Recognition and Machine Learning (9780387310732) Bishop, Christopher M.
- Statistical Inference (9780534243128): Casella, George, Berger, Roger L.
- Probability Theory and Statistical Inference: Empirical Modeling with Observational Data (9781107185142): Spanos, A.
- Bayesian Models: A Statistical Primer for Ecologists (9780691159287): Hobbs, N. Thompson, Hooten, Mevin B.
- A First Course in Bayesian Statistical Methods (0387922997): Hoff, Peter D.

Contents

Properties of estimators

Consistency

Bias

Efficency

Mean squared error

Method of moments

Maximum Likelihood Estimation

Maximum a posteriori estimate

GRADED EVALUATION (15 mins)

Properties of estimators

What makes a good estimator? We likely want an estimator that points to the parameter we are estimating and does not vary much around that value. Usually, there is a tradeoff between these two desires. A few definitions:

Consistency

Consistency is

$$P(|\theta_n - \theta| > 0) \rightarrow 0 \text{ as } n \rightarrow \infty$$

In words, this is stating that as the sample gets large, the estimator converges in probability to θ.

Bias

θ_n is unbiased if

$$E(\theta_n) = \theta$$

basically, the estimator is unbiased if it is centered on the true.

Efficency

An estimator that has the lowest possible variance among all unbiased estimators is considered efficent.

Mean squared error

Combining the variance and bias, we get a measure of the quality of the estimator called mean squared error (MSE):

$$MSE = variance + bias^2$$

MSE is a measure of the trade off between accuracy (spread) and precision (location).

Method of moments

The method of moments amounts to matching population moments to sample moments. Basically, we are using the finite approximation given by:

$$E[f] = \int f(x)^r p(x)dx \approx \frac{1}{N}\sum f(x)^r p(x)$$

where $f(x) = x$ and $r = 1$, this amounts to

$$\mu \approx \bar{x}$$

It is up to us to chose which moment to use, however, the number of moments required will be equal to the number of parameters we are looking to estimate. As an example, suppose we are looking to estimate the probability of success for a coin toss experiment given by the Bernoulli distribution. We know:

$X_i \sim Bern(\theta)$, where we have coded x = 1 for heads:

$$P_X(x;\theta) = \begin{cases} \theta, \text{for x} = 1 \\ 1-\theta, \text{ for x} = 0 \end{cases}$$

Alternatively, we can write this as:

$$f(x;\theta) = \theta^x(1-\theta)^{(1-x)}$$

For this experiment, let us assume we are collecting N=20 tosses of the dime and end up with data: {1,1,0,1,1,1,1,0,1,0,1,0,1,1,0,0,1,1,1,0} (13 heads).

We are looking to estimate both the mean and variance.

For the population:

$$E[X] = \sum_x xf(x) = \sum_x x[\theta^x(1-\theta)^{(1-x)}] = 0*(1-\theta)+1*\theta = \theta$$

$$Var(X) = \theta(1-\theta)$$

From this, we see we only need to estimate one parameter as the variance is a function of the mean.

So, we can match the first sample moment to the population moment to get our estimate:

$$\hat{\theta} = \frac{1}{N}\sum_N x_i = \frac{13}{20}$$

from which the variance can also be computed.

Method of Moment estimators can be show to be consitent, but not necessarily efficent and can give estimates that are outside the parameter space.

Maximum Likelihood Estimation

Another approach to parameter estimation follows from an assumption that our data results from independent and identically distributed observations from a population. Our goal is to find a θ that maximizes the likelihood of us observing our data.

The likelihood function is defined as the joint probability of observing the data:

The likelihood function is defined as the joint probability of observing the data.

$$\mathcal{L}(\theta|x_1...x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Our job is then to solve:

$\frac{d}{d\theta}\mathcal{L}(\theta|x) = 0$ make sure it is a max and not on the boundary.

If we return to the coin toss example with 13 heads in 20 tosses, we start by setting up the likelihood function and differentiating wrt θ.

$\mathcal{L}(\theta|x_1 \ldots x_n) = \prod_{i=1}^{n} \theta^x (1-\theta)^{(1-x)}$

Note, it is often necessary to convert the likelihood to log likelihood to avoid computational difficulties arising from having lots of data.

$ln\mathcal{L}(\theta|x_1 \ldots x_n) = (\sum_{i=1}^{n} x_i)ln\theta + (\sum_{i=1}^{n} (1-x_i))ln(1-\theta)$

differentiating wrt to θ and setting to zero, we get

$(\sum_{i=1}^{n} x_i)\frac{1}{\hat{\theta}} - (\sum_{i=1}^{n} (1-x_i))\frac{1}{1-\hat{\theta}}$ = 0

solving for $\hat{\theta}$, we end up with

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Which matches the previous result of $\frac{13}{20}$.

MLE can be shown to be a consistent estimator, but may be biased. Operationally, it can be computationally expensive to calculate, but offers a useful fact that any function of the parameters is also a function of the MLE, ie invariant to transformations.

Maximum a posteriori estimate

We will discuss MAP estimates in more detail when talking through priors, for now, we can leave this as the MAP estimate is an augmented MLE using prior, or additional information. The procedure is the same as in finding the MLE, however, we add additional information via:

$$\hat{\theta}_{MAP} = argmax_{\theta}\mathcal{L}(\theta|x_1...x_n) * \pi(\theta)$$

$\pi(\theta)$ is our prior or additional information.

GRADED EVALUATION (15 mins)

1. For a variable $X_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1 \ldots n$, find the MLE for (μ, σ^2)

$(\hat{\mu}, \hat{\sigma^2})$ =

a. $\overline{X}, \frac{\sum(X_i - \overline{X})^2}{n}$

b. $\overline{X}, \frac{\sum(X_i - \overline{X})^2}{n-1}$

1. Does the MLE of the last question agree with the MoM estimator?

a. Yes

b. No

1. The MLE estimator is invariant to transformation. If you want an estimate for μ^2 and have the estimate for μ, you can simply take the estimate $\hat{\mu}^2$.

a. False

b. True

By Srijith Rajamohan, Ph.D.