Estimation

Owen G. Ward

2021-05-24

Introduction

We have now seen many examples where we can use a statistical distribution to describe how data are generated. Suppose we observe some data. How do we then estimate a statistical distribution to describe how this data was generated?

Suppose we observe a sequence of coin flips from a coin which we believe isn't fair, but we want to know how unfair it is.

To do this, we need to come up with an estimate of the true probability, p. We will call our estimate $\hat{p}$.

Or we have count data which we think comes from a Poisson distribution, and we want an estimate $\hat{\lambda}$ for the unknown λ.

We call $\hat{\theta}$ an **estimator** for some parameter θ is a function of data which returns an estimate for θ.

For the coin we want $\hat{p}$ to be based on the data we observe and so it will be a random variable. What properties would we like our $\hat{p}$ to have? What properties of random variables have we seen already?

Constructing a function of our random data which gives us a way to estimate the parameters of interest.

Unbiased Estimation

One property we have seen previously is the **expectation** of a random variable. If we are trying to estimate p with some $\hat{p}$, then it seems reasonable to desire that

$$\mathbb{E}(\hat{p}) = p,$$

that our estimator is equal in expectation to the parameter we are using it to estimate.

We say an estimator $\hat{\theta}$ is an **unbiased** estimator for θ if $\mathbb{E}\hat{\theta} = \theta$. If an estimator $\hat{\theta}$ is such that $\mathbb{E}\hat{\theta} \neq \theta$ we say the estimator is **biased** and we call $\mathbb{E}(\hat{\theta}) - \theta$ the **bias** of the estimator.

Considering again coin tosses, suppose we observe $X_1, \ldots, X_n$ coin tosses, where $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. Lets consider $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. In this case, we have

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i).$$

Each X_i is Bernoulli random variable, so we know $\mathbb{E}(X_i) = p$. We we use that above we get

$$\mathbb{E}(\widehat{p}) = \frac{1}{n}\sum_{i=1}^{n} p = p.$$

So in this case, the estimator we have chosen is unbiased. Suppose here we actually have $n = 5$ giving X_1, X_2, X_3, X_4, X_5 and our estimator is

$$\widehat{p} = \frac{1}{5}\sum_{i=1}^{5} X_i.$$

What happens if we use a slightly different estimator, say

$$\widehat{p}_1 = \frac{1}{6}\sum_{i=1}^{5} X_i$$

The expectation of this estimator is $\mathbb{E}\widehat{p}_1 = \frac{5}{6}p \neq p$. So this estimator is said to be **biased** for estimating p. Based on this, it makes sense to say that $\widehat{p}$ is a better estimator that $\widehat{p}_1$.

However, what about if we consider

$$\widehat{p}_2 = \frac{1}{2}\sum_{i=1}^{2} X_i,$$

which only uses the first two coin flips (out of the 5 total) to estimate p.

Now when we look at the expectation of $\widehat{p}_2$ we see that it is unbiased. So based on bias, there is no reason to say $\widehat{p}$ is a better estimator than $\widehat{p}_2$.

We need to compare further properties. What have we seen?

Like previously, a closely related property to the expectation of a random variable is its variance.

Variance of an Estimator

Like we computed the expectation of an estimator, we can also compute the variance of an estimator. An estimator with a smaller variance will show less uncertainty about it's estimate.

For example, we can compare the variance of $\widehat{p}$ and $\widehat{p}_2$.

For a general n we have that

$$Var(\widehat{p}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right)$$

Now we can compare the variance of $\widehat{p}$ when $n = 5$ and the variance of $\widehat{p}_2$.

$$Var(\widehat{p}) = \frac{p(1-p)}{5},\ Var(\widehat{p}_2) = \frac{p(1-p)}{2}.$$

We see that $Var(\widehat{p}) < Var(\widehat{p}_2)$, so $\widehat{p}$ provides an estimator with less variance than $\widehat{p}_2$.

So we know something about the type of properties we want estimators to have. But how do we construct them?

Constructing Estimators

We will consider two methods,

- Method of Moments Estimate (MME)
- Maximum Likelihood Estimate (MLE)

In many cases these will be the same.

Method of Moments

This is a simpler method. Essentially, match the sample mean to the expected value and solve for the parameter.

For $X_1, X_2, \dots, X_n$ i.i.d samples from a $\mathcal{N}(\mu, 1)$, we want to estimate μ.

Example

Suppose $X_1, X_2, \dots, X_n$ are i.i.d samples from an exponential distribution with parameter λ.

For each X_i we have $\mathbb{E}(X_i) = \frac{1}{\lambda}$.

So the method of moments estimator for λ is found by solving $\bar{X} = \mathbb{E}(X_i)$ for λ which gives

$$\hat{\lambda} = \frac{1}{\bar{X}},$$

our MME estimator.

Suppose we have $X_1, \dots, X_n$ samples from $Binom(m, p)$ where we know m but need to estimate p.

For $X \sim Binom(m, p)$ we know $\mathbb{E}(X) = mp$. $\bar{X}$ is the average of n samples from this Binomial distribution.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

So our MME estimate for p solves

$$mp = \bar{X},$$

giving $\hat{p} = \frac{\bar{X}}{m}$.

Likelihood and Maximum Likelihood Estimation

A more complicated but more powerful estimation method is maximum likelihood estimation.

This consists of constructing a **likelihood** function of the data in terms of the parameter, and maximising that function with respect to the parameter.

To maximise a function, we find where the derivative is zero and confirm the solution is a max, not a min.

Likelihood Functions

Recall a random variable X has a pmf/pdf $f_X(x)$. We use slightly different notation $f_\theta(x)$, where θ is the parameter of the distribution.

We define the **likelihood function** as the product of the pdf evaluated at each observed data point. For data $X_1, \dots, X_n$

$$L = \prod_{i=1}^{n} f_\theta(x_i).$$

We want to find the value of θ which maximises L. Maximising L is the same as maximising $l = \log L$, for natural logarithms.

Let's see some examples.

MLE for Exponetial Data

We have $X_1, \dots, X_n \sim Exp(\lambda)$, with

$$f_\lambda(x_i) = \lambda e^{-\lambda x_i}.$$

So this gives

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

$$L = \lambda^n e^{-\lambda \sum x_i}.$$

We take the log,

$$l = \log L = n \log \lambda - \lambda \sum x_i.$$

Then we differentiate l with respect to λ and set this derivative equal to 0.

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum x_i.$$

$$\frac{\partial l}{\partial \lambda} = 0 \Rightarrow \frac{n}{\lambda} - \sum x_i = 0$$

Solving this for λ gives

$$\frac{n}{\lambda} = \sum x_i \Rightarrow \lambda = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}.$$

MLE for Binomial Data

When $X_1, \dots, X_n \sim Binom(m, p)$ then

$$f_p(x_i) = \binom{m}{x_i} p^{x_i} (1-p)^{m-x_i}.$$

This gives the likelihood for all data as

$$L = \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i} (1-p)^{m-x_i}$$

$$= \left(\prod_{i=1}^{n} \binom{m}{x_i} \right) p^{\sum x_i} (1-p)^{nm-\sum x_i}$$

Taking the log of this gives

$$l = \log \left(\prod_{i=1}^{n} \binom{m}{x_i} \right) + \sum x_i \log(p) + (nm - \sum x_i) \log(1-p).$$

When we differentiate this with respect to p,

$$\frac{\partial l}{\partial p} = \frac{\sum x_i}{p} + \frac{nm - \sum x_i}{1-p}.$$

We set this equal to zero, cross multiply to get

$$\hat{p} = \frac{\sum x_i}{mn} = \bar{x}.$$

MLE for Normal Mean

Suppose we have n data points $X_1, \dots, X_n$ from $\mathcal{N}(\mu, 1)$. We want to find the MLE for μ.

The likelihood function for a single data point x_i, in terms of unknown μ, is

$$f_\theta(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2 \right)$$

When we get the likelihood of n observations this is

$$L = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2 \right)$$

The product of exponents is the exponent of the sum, giving

$$L = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

We take the log of this,

$$l = \log L = n \log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

and differentiate l with respect to μ.

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2}2\sum_{i=1}^{n}(x_i - \mu)(-1) = \sum_{i=1}^{n}(x_i - \mu).$$

Then we solve

$$\frac{\partial l}{\partial \mu} = 0,$$

and confirm this is a maximum of the log likelihood function.

Variability of an Estimator

Whatever estimate we come up with, it will be a function of some data from a distribution, which has some inherent randomness built in.

If we compute the sample mean from different datasets we will get different value.

We want to quantify how much our estimate can vary. This is much like computing the variance of a random variable, because our estimate is a random variable.

Suppose we want to estimate the proportion of graduate students who supported the contract agreed. We have a survey of 500 students where 225 supported the agreement.

If this was a survey of all students, we want to use it to estimate p, the proportion who support the agreement.

Each person like a coin flip, with probability p of supporting the agreement. If $X_i = 1$ if person i supported, $X_i = 0$ if against, then $f_p(x_i) = p^{x_i}(1-p)^{1-x_i}$.

The MME/MLE estimate for p is $\hat{p} = \bar{X}$, which is $\frac{225}{500} = 0.45$.

How much variability is there in this estimate? If we asked another 500 students, how different do we think the answer could be?

A standard way to estimate the uncertainty of a estimate is to compute it's **standard error**.

This is given by the standard deviation of $\hat{p} = \bar{X}$, where

$$se(\hat{p}) = \sqrt{Var(\bar{X})} = \sqrt{\frac{p(1-p)}{n}},$$

from earlier in these slides. The standard error involves p, which we don't know!

But we know how to estimate it, giving

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

So for our example with $\hat{p} = \frac{225}{500} = 0.45$ the standard error of this estimator is

$$se(\hat{p}) = \sqrt{\frac{0.45(0.55)}{500}} = 0.022.$$

What does this standard error tell us?

How many standard errors do we need to go from our estimate to get to 0.5?

How can we decrease the standard error of our estimate?

Data Quality

This assumes that the 500 students we surveyed are a **representative** sample of all graduate students. If we only chose students from Humanities programs, can we say this estimate is a good estimate for the proportion of **all** graduate students who support the union?

While the theory will go through, you need to be aware of how the data was collected, which can be often overlooked.