7.2: Methods of Point Estimation

Taylor

University of Virginia

Motivation

Given a model, we've learned how to answer probabilistic questions. If I tell you your data is coming from a normal distribution with parameters $\mu = 10$ and $\sigma^2 = 2$, you can answer any question you want.

In real life, however, we won't be given a completely specified model. We'll have strong ideas about what types of procedures and distributions are appropriate, but that isn't enough. We need to figure the parameters. That's what we'll do in this section.

Definition

Let $X_1, \ldots, X_n$ be a random sample from the pmf or pdf $f(x)$. For $k = 1, 2, \ldots$, the **k-th population moment** is $E[X^k]$. We learned about these earlier already.

Contrast that with a **kth sample moment**, which is $\frac{\sum_i^n X_i^k}{n}$

Definitions

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x; \theta_1, \ldots, \theta_m)$, where $\theta_1, \ldots, \theta_m$ are parameters whose values are unknown. The **moment estimators** $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments.

Definitions

If there are m parameters $\theta_1, \ldots, \theta_m$, then we solve m equations:

$$\begin{pmatrix} E[X] \\ E[X^2] \\ \vdots \\ E[X^m] \end{pmatrix} \overset{set}{=} \begin{pmatrix} \frac{\sum_i^n X_i}{n} \\ \frac{\sum_i^n X_i^2}{n} \\ \vdots \\ \frac{\sum_i^n X_i^m}{n} \end{pmatrix}$$

Note: the left hand side will be a vector of functions of all the different thetas, and the right hand side will be functions of the data

Example (example 7.13 on page 351)

Assume $X_1, \ldots, X_n \overset{i.i.d.}{\sim}$ Exponential(λ). Estimate λ.

Example (example 7.13 on page 351)

Assume $X_1, \ldots, X_n \overset{i.i.d.}{\sim}$ Exponential(λ). Estimate λ.

There's only one parameter, so we only need one equation. Because $EX = 1/\lambda$, we get $\hat{\lambda} = \frac{1}{\bar{X}}$

Example 2

Assume $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimate μ and σ^2.

Example 2

Assume $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimate μ and σ^2.

Recall $EX = \mu$ and $E[X^2] = V[X] + (EX)^2 = \sigma^2 + \mu^2$. So we solve the following system of equations:

$$\begin{pmatrix} \mu \\ \sigma^2 + \mu^2 \end{pmatrix} \overset{set}{=} \begin{pmatrix} \frac{\sum_i^n X}{n} \\ \frac{\sum_i^n X^2}{n} \end{pmatrix}$$

which yields

$$\hat{\mu} = \bar{X}$$

and

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_i X_i^2}{n} - (\hat{\mu})^2 = \frac{\sum_i X_i^2}{n} - (\bar{X})^2 \\ &= \frac{\sum_i X_i^2 - n\bar{X}^2}{n} = \frac{1}{n}\sum_i (X_i - \bar{X})^2 \end{aligned}$$

Definitions

We will also learn another way to estimate parameters. This way usually has nicer properties. It's called **maximum likelihood estimation**. First, we have to set this up a bit.

Without loss of generality, let's talk about cts rvs. So we have a density function $f(x;\theta)$. Until now, we've assumed we're given the parameters θ, so it's a function in x. However, what if we look at it after we get data (so the x is known), and we're thinking about it like a function in θ?

It's the same function, but we're thinking about it two different ways. We used to hold θ fixed, but now we're fixing the data, x.

Example 7.16 on page 352 (ten people and their bike helmets)

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim}$ Bernoulli(p). Then

$$f(x_1, \ldots, x_n; p) = \prod_i^n p(x_i; p) = \prod_i^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

If our data set is $1, 0, 1, 0, 0, 0, 0, 0, 0, 1$, then

$$f(x_1, \ldots, x_n; p) = p^3(1-p)^7$$

This is a function of p now. If we plug in a value of p, this function will tell us how *likely* that value is. We don't say how *probable* it is, because technically, there is nothing random going on if we've already observed the data.

A word on notation

The likelihood function has nothing to do with probability, since there is nothing random going on. That's why we say *likely* instead of *probable*. These are synonyms in real life, but in statistics they are jargon.

So even though densities/pmfs have the same formulas as likelihood functions, they have different interpretations in this chapter. That's why in many textbooks, to emphasize this, they will write the likelihood as $L(\theta; \mathbf{X})$ instead of $f(\mathbf{X}; \theta)$.

However, the authors of this textbook seem to favor consistency in notation. I'm going to stick with the book and keep writing $f(x_1, \ldots, x_n; \theta)$, so that my slides cohere.

And another thing

Instead of maximizing the likelihood, we will often maximize the log of the likelihood. $f' = 0$ if and only if $(\log f)' = 0$, so these functions have the same maximizers.

We do this because it's easier to handle mathematically. Because we are assuming independence all the time, our likelihoods are going to be in the form of a big product a lot:

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_n) = \Pi_{i=1}^{n} f_{X_i}(x_i).$$

When you take the log of this product, it turns into a big sum. And sums are easier to take the derivative of!

$$\log f_{X_1,\ldots,X_N}(x_1,\ldots,x_n) = \log \Pi_{i=1}^{n} f_{X_i}(x_i) = \sum_{i=1}^{n} \log f_{X_i}(x_i).$$

Example 7.16 again

Recall that we had

$$f(x_1, \ldots, x_n; p) = p^3(1-p)^7.$$

You can try setting the derivative of that thing equal to 0, or you can set the derivative of the log likelihood equal to 0:

$$\log f(x_1, \ldots, x_n; p) = 3 \log p + 7 \log(1-p).$$

We set the derivative of this equal to 0. $3/p - 7/(1-p) \overset{set}{=} 0$ gives us $\hat{p} = 3/10$.

Definitions

For a joint pmf or pdf $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$, the **maximum likelihood estimates** $\hat{\theta}_1, \ldots, \hat{\theta}_m$ for the parameters are such that

$$f(x_1, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_m) \geq f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_m)$$

for any m-tuple $\theta_1, \ldots, \theta_m$.

The **maximum likelihood estimator** is the same as above but when we replace each x_i with X_i.

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Gamma$(1, \theta)$. Each rv has a density $f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$. For this density, $\theta > 0$, as well as $x > 0$. Find the MLE of θ.

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(1, \theta)$. Each rv has a density $f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$. For this density, $\theta > 0$, as well as $x > 0$. Find the MLE of θ.

$$f(x_1, \ldots, x_n; \theta) = \prod_i \left[\frac{1}{\theta} e^{-x_i/\theta} \right] = \theta^{-n} e^{-\sum_i x_i/\theta}.$$

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(1, \theta)$. Each rv has a density $f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$. For this density, $\theta > 0$, as well as $x > 0$. Find the MLE of θ.

$$f(x_1, \ldots, x_n; \theta) = \prod_i \left[\frac{1}{\theta} e^{-x_i/\theta}\right] = \theta^{-n} e^{-\sum_i x_i/\theta}.$$

$$\log f(x_1, \ldots, x_n; \theta) = -n \log \theta - \theta^{-1} \sum_i x_i$$

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Gamma}(1, \theta)$. Each rv has a density $f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}$. For this density, $\theta > 0$, as well as $x > 0$. Find the MLE of θ.

$$f(x_1, \ldots, x_n; \theta) = \prod_i \left[\frac{1}{\theta} e^{-x_i/\theta} \right] = \theta^{-n} e^{-\sum_i x_i/\theta}.$$

$$\log f(x_1, \ldots, x_n; \theta) = -n \log \theta - \theta^{-1} \sum_i x_i$$

$$\frac{d \log f(x_1, \ldots, x_n; \theta)}{d\theta} = -\frac{n}{\theta} + \theta^{-2} \sum_i x_i$$

More examples!

Let $X_1, \dots, X_n \overset{iid}{\sim}$ Gamma$(1, \theta)$. Each rv has a density $f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}$. For this density, $\theta > 0$, as well as $x > 0$. Find the MLE of θ.

$$f(x_1, \dots, x_n; \theta) = \prod_i \left[\frac{1}{\theta}e^{-x_i/\theta}\right] = \theta^{-n}e^{-\sum_i x_i/\theta}.$$

$$\log f(x_1, \dots, x_n; \theta) = -n\log\theta - \theta^{-1}\sum_i x_i$$

$$\frac{d\log f(x_1, \dots, x_n; \theta)}{d\theta} = -\frac{n}{\theta} + \theta^{-2}\sum_i x_i$$

Setting the last expression equal to zero and solving yields $\hat{\theta} = \bar{X}$.

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $f(x;\theta) = (\theta + 1)x^{\theta}$.

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $f(x;\theta) = (\theta+1)x^\theta$.

$$f(x_1,\ldots,x_n;\theta) = \prod_i \left[(\theta+1)x_i^\theta\right] = (\theta+1)^n(\prod_i x_i)^\theta.$$

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $f(x;\theta) = (\theta+1)x^{\theta}$.

$$f(x_1, \ldots, x_n; \theta) = \prod_i \left[(\theta+1)x_i^{\theta}\right] = (\theta+1)^n (\prod_i x_i)^{\theta}.$$

$$\log f(x_1, \ldots, x_n; \theta) = n \log(1+\theta) + \theta \sum_i \log(x_i)$$

More examples!

Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $f(x;\theta) = (\theta+1)x^{\theta}$.

$$f(x_1, \ldots, x_n; \theta) = \prod_i \left[(\theta+1)x_i^{\theta}\right] = (\theta+1)^n (\prod_i x_i)^{\theta}.$$

$$\log f(x_1, \ldots, x_n; \theta) = n \log(1+\theta) + \theta \sum_i \log(x_i)$$

$$\frac{d \log f(x_1, \ldots, x_n; \theta)}{d\theta} = \frac{n}{1+\theta} + \sum_i \log(x_i).$$

More examples!

Let $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$, where $f(x;\theta) = (\theta+1)x^{\theta}$.

$$f(x_1, \dots, x_n; \theta) = \prod_i \left[(\theta+1)x_i^{\theta}\right] = (\theta+1)^n (\prod_i x_i)^{\theta}.$$

$$\log f(x_1, \dots, x_n; \theta) = n \log(1+\theta) + \theta \sum_i \log(x_i)$$

$$\frac{d \log f(x_1, \dots, x_n; \theta)}{d\theta} = \frac{n}{1+\theta} + \sum_i \log(x_i).$$

Setting the last expression equal to zero and solving yields $\hat{\theta} = -\frac{n}{\sum_i \log x_i} - 1$.

Some Complications

Sometimes the maximum isn't where the derivative is 0.

Suppose we have $X_1, \ldots, X_n \overset{iid}{\sim} \text{Uniform}([0,\theta])$. Then $f(x;\theta) = 1/\theta,\ \ 0 \leq x \leq \theta$. That means our joint density is

$$f(x_1, \ldots, x_n; \theta) = \theta^{-n}, \qquad 0 \leq x_1, \ldots, x_n \leq \theta$$

or equivalently

$$f(x_1, \ldots, x_n; \theta) = \theta^{-n}, \qquad 0 \leq \max(x_1, \ldots, x_n) \leq \theta$$

Another complication

Sometimes the you can't take the deriative at all.

Recall a hypergeometric distribution. We have N, the size of the total population, M, which is the number of things in the population that have some characteristic, and n, our sample size.

$$p(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

In a capture/recapture experiment, the objective is to estimate N. You start of by capturing and tagging M animals. You release them into the wild, and wait a bit for them to mix back in with their species. After that, you collect your data. X is the number of animals out of a sample size of n that you tagged.

Another complication (continued)

Since N is an integer, $p(\mathbf{x}; N)$ isn't defined on a subset of the real line; it isn't differentiable. This is how the book does it. Recall $p(x; N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

$$
\begin{aligned}
&p(\mathbf{x}; N) \text{ is increasing} \\
&p(\mathbf{x}; N) > p(\mathbf{x}; N-1) \\
&\frac{p(\mathbf{x}; N)}{p(\mathbf{x}; N-1)} > 1 \\
&\frac{\binom{N-M}{n-x}}{\binom{N}{n}} \Bigg/ \frac{\binom{N-1-M}{n-x}}{\binom{N-1}{n}} > 1 \\
&\frac{(N-M)(N-n)}{N(N-M-n+x)} > 1 \\
&N < Mn/x
\end{aligned}
$$

This means $\hat{N} = \lfloor Mn/x \rfloor$

Proposition

Here's a cool property of MLE estimators called the **invariance principle**:

Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ be the mles of the parameters $\theta_1, \ldots, \theta_m$. Then the mle of any function $h(\theta_1, \ldots, \theta_m)$ of these parameters is the function $h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$. In other words:

$$h(\widehat{\theta_1, \ldots, \theta_m}) = h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$$

Example

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. We (will) know $\hat{\mu} = \bar{X}$ and $\hat{\sigma^2} = \sum_i (X_i - \bar{X})^2/n$. Estimate the **coefficient of variation**: σ/μ.

Example

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. We (will) know $\hat{\mu} = \bar{X}$ and $\hat{\sigma^2} = \sum_i (X_i - \bar{X})^2/n$. Estimate the **coefficient of variation**: σ/μ.

The mle estimator of this quantity is then

$$\sqrt{\hat{\sigma}^2}/\hat{\mu} = \frac{\sqrt{\sum_i (X_i - \bar{X})^2/n}}{\bar{X}}$$

Another Example

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimate $P(X > 100) = 1 - \Phi\left(\frac{100-\mu}{\sigma/\sqrt{n}}\right)$

Another Example

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. Estimate $P(X > 100) = 1 - \Phi\left(\frac{100-\mu}{\sigma/\sqrt{n}}\right)$

Just plug in $\hat{\mu}$ for μ and $\hat{\sigma}^2$ for σ^2:

$$1 - \Phi\left(\frac{100 - \hat{\mu}}{\sqrt{\hat{\sigma}^2/n}}\right)$$

where $\hat{\mu} = \bar{X}$ and $\hat{\sigma^2} = \sum_i (X_i - \bar{X})^2/n$.

Large Sample Behavior of the MLE

Under very general conditions on the joint distribution of the sample, when the sample size is large, the MLE of any parameter θ

1. is “close to” θ (consistency)
2. is approximately unbiased ($E[\hat{\theta}] \approx \theta$)
3. has variance as nearly as small as possible

So it’s asymptotically the MVUE.