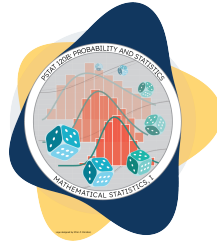


Topic 3: Estimation

Ethan P. Marzban University of California, Santa Barbara PSTAT 120B

Outline

1. Method of Moments

2. Likelihoods

Method of Moments

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Leadup

- Recall that, prior to Midterm 02, our primary concern was with determining what properties a "good" estimator should possess.
- We started off with the notions of unbiasedness and small MSE (both of which are good!), and then talked about consistency as yet another potential property we'd like an estimator to satisfy.
  - We'll actually return to this question of what makes a good estimator in a bit!
- But, for now, we'd like to turn our attention to actually *constructing* estimators.
- Today, we'll talk about the **Method of Moments**.

Method of Moments

- Allow me to introduce the Method of Moments (MoM) by way of an example.
- Say we have $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$ for some (unknown) $\theta > 0$, and suppose we want to construct an estimator for θ.
- Note that $\mathbb{E}[Y_1] = \theta/2$.
- Now, by the WLLN, it makes sense to assume that the sample mean ($\overline{Y}_n$) will be relatively close to the population mean ($\mu := \theta/2$).
- So, what if we set the sample mean to the population mean, solve for θ, and call the resulting quantity an estimator for θ?

Method of Moments

- That is, we assert our estimator $\widehat{\theta}_{\text{MM}}$ must satisfy

$$\overline{Y}_n = \frac{\widehat{\theta}_{\text{MM}}}{2}$$

- This in turn yields the estimator

$$\widehat{\theta}_{\text{MM}} = 2\overline{Y}_n$$

Method of Moments

- First and foremost, is this a "good" estimator?
- Well, note that it is unbiased:

$$\mathbb{E}[\widehat{\theta}_{\text{MM}}] = \mathbb{E}[2\overline{Y}_n] = 2 \cdot \mathbb{E}[\overline{Y}_n] = 2 \cdot \frac{\theta}{2} = \theta \checkmark$$

- Additionally, it is a consistent estimator for θ:
  - By the WLLN, $\overline{Y}_n \xrightarrow{p} (\theta/2)$
  - Hence, by the Continuous Mapping Theorem (CMT) with $g(x) = 2x$, we have that

$$\widehat{\theta}_{\text{MM}} = 2\overline{Y}_n \xrightarrow{p} 2 \cdot \frac{\theta}{2} = \theta \checkmark$$

Moments

- So, it seems we've constructed a pretty good estimator for θ!
- Before posing the general scheme for uing this strategy to construct estimators, let me quickly remind us of some terminology.
- If Y_i follows some distribution, recall that the **k^{th} population moment** of that distribution, notated μ_k, is given by $\mu_k := \mathbb{E}[Y_i^k]$.
- Given a sample $\{Y_i\}_{i=1}^n$, the **k^{th} sample moment**, notated $M_{k,n}$, is defined as

$$M_{k,n} := \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

Method of Moments, One-Parameter Case

- We're finally ready to pose the first version of the Method of Moments!
- Again, remember that this method is designed to yield an estimator $\widehat{\theta}_n$ for θ.

1. Equate the first sample moment (i.e. the sample mean) to the first population moment.
2. Replace all instances of θ with $\widehat{\theta}_{\text{MM}}$, and then solve for $\widehat{\theta}_{\text{MM}}$ in terms of the first sample moment.

Method of Moments, General Case

- What do we do if our distribution has multiple parameters?
- Well, we follow a very similar procedure:

1. Set $\mu_k = M_{k,n}$ for $k = 1, \cdots p$ where p is the number of parameters.
2. Replace all instances of θ_k with $\widehat{\theta}_{\text{MM},k}$ (again, for $k = 1, \cdots, p$), and then solve the **system** of equations for $\widehat{\theta}_{\text{MM},k}$ in terms of the first sample moment.

Example

Example

Consider $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown parameters. Find $\widehat{\mu}_{\text{MM}}$ and $\widehat{\sigma^2}_{\text{MM}}$, the Method of Moments estimators for μ and σ^2, respectively.

Example

- Since we have two parameters, we will end up with a system of two equations to solve.
- Furthermore, we'll need both the first and second population moments, so let's compute those.
- We know that $\mu_1 := \mathbb{E}[Y_1] = \mu$; additionally, $\mu_2 = \text{Var}(Y_1) + \mu_1^2 = \sigma^2 + \mu^2$.
- Therefore, $\widehat{\mu}_{\text{MM}}$ and $\widehat{\sigma^2}_{\text{MM}}$ must satisfy the system of equations

$$\begin{cases} \widehat{\mu}_{\text{MM}} & = \overline{Y}_n \\ \widehat{\mu}_{\text{MM}}^2 + \widehat{\sigma^2}_{\text{MM}} & = M_{2,n} \end{cases}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Example

- The first equation immediately gives us the Method of Moments estimator for μ:

$$\widehat{\mu}_{\text{MM}} = \overline{Y}_n$$

- Substituting this into the second equation yields the Method of Moments estimator for σ^2:

$$\begin{aligned}\widehat{\sigma^2}_{\text{MM}} &= M_{2,n} - \widehat{\mu}^2_{\text{MM}} \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\overline{Y}_n)^2\end{aligned}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Example

Example

An i.i.d. sample of 8 cats is taken from the population of all DSH cats; their observed weights (in lbs) are given by

$$(8.2, 9.2, 8.7, 10.2, 15.2, 7.2, 16.1, 9.5)$$

Based on this sample, what is an appropriate estimate for the true average weight of all DSH cats? What is an appropriate estimate for the true variance of weights among all DSH cats? Assume the weight of a randomly-selected DSH cat is normally distributed.

Example

- Let's use the Method of Moments estimator.
- Define Y_i to be the weight of a randomly-selected DSH cat; the problem then tells us that

$$Y_1, Y_2, \cdots, Y_8 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

  where μ denotes the true average weight of all DSH cats and σ^2 denotes the true variance among all DSH cat weights.
- All we need to do is to plug into our MoM estimators we derived in the previous example!

Example

$$\vec{\boldsymbol{y}} = (8.2, 9.2, 8.7, 10.2, 15.2, 7.2, 16.1, 9.5)$$

$$\bar{y}_n = \frac{1}{8}(8.2 + 9.2 + 8.7 + \cdots + 16.1 + 9.5) = 10.5375$$

$$m_{2,8} = \frac{1}{8}(8.2^2 + 9.2^2 + 8.7^2 + \cdots + 16.1^2 + 9.5^2) = 120.4938$$

$$\implies \widehat{\mu}_{\text{MM}} = \boxed{10.5375 \text{ lbs}}$$

$$\widehat{\sigma^2}_{\text{MM}} = 120.4938 - (10.5375)^2 \approx \boxed{9.4549 \text{ lbs}^2}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Zero-Moment

- If a particular moment is equal to zero, move to the next moment.
- For example, if $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[-\theta, \theta]$ then $\mathbb{E}[Y_1] = 0$. Naturally, $\widehat{\theta}_{\text{MM}}$ doesn't satisfy the equation $\widehat{\theta}_{\text{MM}} = 0$; rather, you should consider finding the *second* population moment $\mathbb{E}[Y_i^2]$ in terms of θ, replace θ with $\widehat{\theta}_{\text{MM}}$, and solve.
- You'll work through a problem like this on the next homework (to be released early Tomorrow [Friday]).

Result

Theorem (Consistency of MoM Estimators)

In general, method of moments estimators are consistent.

- There are certain exceptions, but we won't concern ourselves with those in this class.
- Please keep in mind: though you are welcome to use this result (unless a problem explicitly asks you not to), you can usually appeal to a combination of the WLLN and CMT to establish consistency for MoM estimators.

Clicker Question

Clicker Question 1

Are Method of Moments estimators always unbiased?

(A) Yes
(B) No

UC SANTA BARBARA
Department of Statistics
and Applied Probability

Example Word Problem 1

Example

A coin is tossed 27 times. Among these 27 tosses, 13 heads are observed. Based on this data, what is an ideal estimate for the probability that this coin lands heads on any given toss?

Example Word Problem 2

Example

A company wants to know what proportion of their stock is defective. To that end, they hire 3 workers (Abhi, Biyonka, and Cameron) who each keep randomly selecting products until they find a defective product. Abhi observes his first defective product on his third draw, Biyonka observers her first defective product on her fourth draw, and Cameron observes their first defective product on their fourth draw.

Based on this data, what is a good estimate for the true proportion of products that are defective?

Leadup

- The Method of Moments is certainly a very popular technique for constructing estimators.
- It is, however, not the only technique that can be used to construct estimators.
- Before going too far, we'll need to actually take a step back and try and gain some new perspective on some old concepts...

Likelihoods

Leadup

- Let's consider a discrete random vector $\vec{\boldsymbol{X}} := (X_1, \cdots, X_n)$. In other words, consider a collection of discrete random variables $X_1, \cdots, X_n$.
- To make things even more explicit, assume $X_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$, so that X_i simply models the outcome of a $p-$coin flip [recall that a "$\theta-$coin" is a coin that lands "heads" with probability θ].
- What does the joint PMF $p_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}})$ actually represent?
- Well, by definition (since we are assuming a *discrete* random vector)

$$\begin{aligned} p_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}}) &= p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) \\ &= \mathbb{P}(X_1 = x_1,\ X_2 = x_2, \cdots,\ X_n = x_n) \end{aligned}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Leadup

- For example, if $n = 3$ then

$$p_{X_1,X_2,X_3}(1,0,0) = \mathbb{P}(X_1 = 1,\ X_2 = 0,\ X_3 = 0)$$

  represents the probability that the first coin landed heads, the second coin landed tails, and the third coin landed tails.
- So, a joint PMF is really a specification of *probabilities*; i.e. a way of quantifying our *beliefs*.
- Of course, $p_{\vec{X}}(\vec{x})$ will, in general, involve our population parameter (θ in this example). For example,

$$p_{X_1,X_2,X_3}(1,0,0) = \theta(1-\theta)^2$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Notation

- As such, let's modify our notation a bit to make it clear that these joint PMFs depend on θ.
- Specifically, we'll write $p_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}}; \theta)$.
- Now, if $\vec{\boldsymbol{X}}$ is a *continuous* random vector, we have instead a joint *PDF* (as opposed to a joint *PMF*)

$$f_{\vec{\boldsymbol{X}}}(\vec{\boldsymbol{x}}) = f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n)$$

- Of course, the value of a joint PDF at any particular input is not a probability, but we can still view the joint PDF as a way of quantifying our joint beliefs on $\vec{\boldsymbol{X}}$ (and, indirectly, θ).

Likelihood

- This brings us to the notion of a **likelihood**.

Definition (Likelihood)

Let $\vec{\boldsymbol{y}} := \{y_i\}_{i=1}^n$ be an observed instance of a random sample $\vec{\boldsymbol{Y}} := \{Y_i\}_{i=1}^n$, whose distribution depends on some parameter θ. The **likelihood** of the sample is simply the joint PMF/PDF of $\vec{\boldsymbol{Y}}$.

- To avoid having to constantly separate the discrete and continuous cases, we adopt the notation

$$\mathcal{L}_{\vec{\boldsymbol{y}}}(\theta) \qquad \text{or} \qquad \mathcal{L}(y_1, \cdots, y_n; \theta)$$

to mean the likelihood.

Notation

- A quick note on notation: I will use the notations $\mathcal{L}_{\vec{\boldsymbol{y}}}(\theta)$ and $\mathcal{L}(y_1, \cdots, y_n; \theta)$ interchangeably [though the second notation makes the sample values clearer, it is clunkier than the first].
- Just be aware that the textbook always uses $\mathcal{L}(y_1, \cdots, y_n; \theta)$.
  - Technically the textbook writes $\mathcal{L}(y_1, \cdots, y_n \mid \theta)$, but so as to avoid confusion with conditional distributions I will avoid using this notation for the purposes of this class.
- And, again, to reiterate - the likelihood is nothing more than the joint PMF/PDF of a random sample, evaluated at a particular observed instance $\vec{\boldsymbol{y}}$.

Simplification

- Now, if we assume an i.i.d. sample, we can expand things a bit.
- For instance, if $Y_1, \cdots, Y_n$ are i.i.d. discrete random variables from a distribution with mass function $p(y; \theta)$, then

$$\begin{aligned}\mathcal{L}_{\vec{\boldsymbol{y}}}(\theta) &= p_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) \\ &= p_{X_1}(x_1; \theta) \times p_{X_2}(x_2; \theta) \times \cdots \times p_{X_n}(x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)\end{aligned}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Simplification

- Similarly, if $Y_1, \cdots, Y_n$ are i.i.d. continuous random variables from a distribution with density function $f(y;\theta)$, then

$$\begin{aligned}\mathcal{L}_{\vec{\boldsymbol{y}}}(\theta) &= f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) \\ &= f_{X_1}(x_1;\theta) \times f_{X_2}(x_2;\theta) \times \cdots \times f_{X_n}(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)\end{aligned}$$

UC SANTA BARBARA
Department of Statistics and Applied Probability

Example

Example

The weight of a randomly-selected DSH cat is assumed to be normally distributed about some unknown mean μ and with some known standard deviation $\sigma = 2$ lbs. An i.i.d. random sample of 3 cats is taken; their weights are 8.2 lbs, 16.2 lbs, and 14.1 lbs. What is the likelihood of this sample? (Remember that this will be a function of μ!)

Solution

- Let Y_i denote the weight of a randomly-selected DSH cat; then $Y_1, Y_2, \dots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 4)$.
- Hence, the density of Y_i at a point y_i is given by the density of a $\mathcal{N}(\mu, 4)$ distribution, evaluated at y_i:

$$f(y_i; \mu) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(y_i - \mu)^2}$$

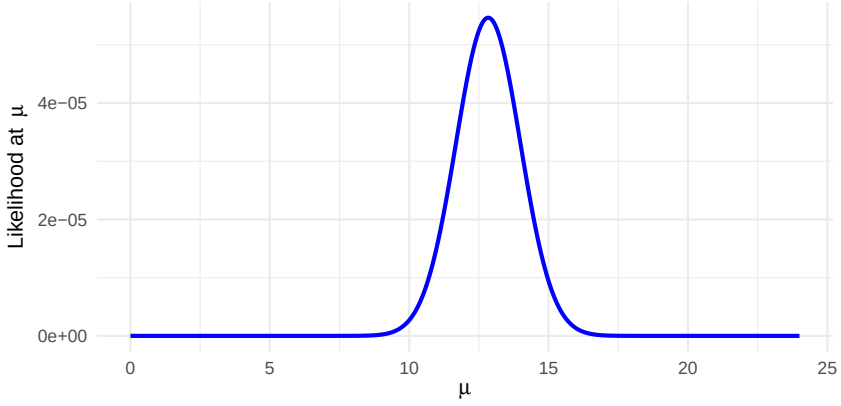
Solution

- Therefore,

$$\begin{aligned}\mathcal{L}_{(8.2,16.2,14.1)}(\mu) &= f(8.2;\mu) \times f(16.2;\mu) \times f(14.1;\mu) \\ &= \left(\frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(8.2-\mu)^2}\right) \times \left(\frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(16.2-\mu)^2}\right) \times \\ &\qquad \left(\frac{1}{2\sqrt{2\pi}} e^{-\frac{1}{8}(14.1-\mu)^2}\right) \\ &= \left(\frac{1}{2\pi}\right)^3 \exp\left\{-\frac{1}{8}[(8.2-\mu)^2 + (16.2-\mu)^2 + (14.1-\mu)^2]\right\}\end{aligned}$$

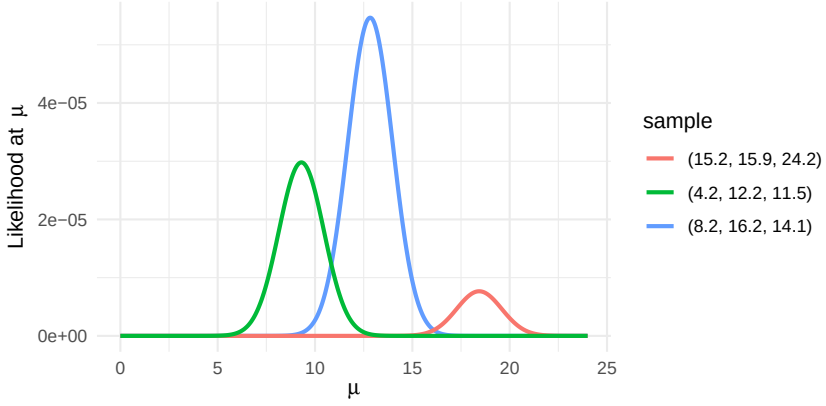
UC SANTA BARBARA
Department of Statistics and Applied Probability

Example



UC SANTA BARBARA
Department of Statistics
and Applied Probability

Example



UC SANTA BARBARA
Department of Statistics and Applied Probability

Example

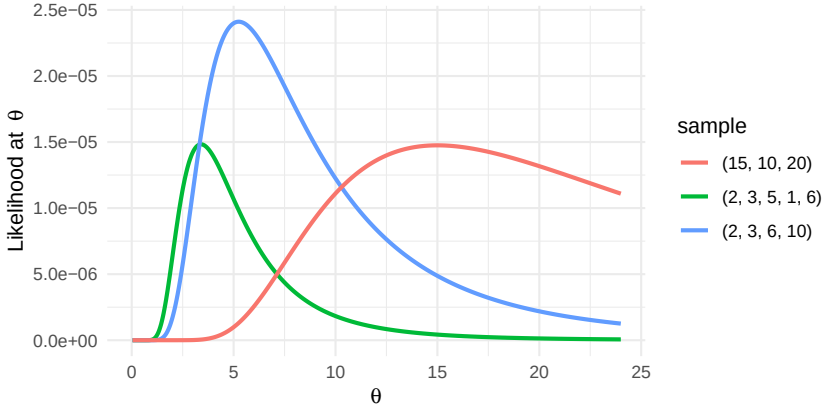
Example

The wait time of a randomly-selected person at the DMV follows an exponential distribution with unknown parameter θ. Assuming an i.i.d. sample $\{Y_i\}_{i=1}^n$ of wait times and their corresponding observed instances $\{y_i\}_{i=1}^n$, what is the likelihood as a function of θ and $\{y_i\}_{i=1}^n$?

- Let's do this one on the board.

UC SANTA BARBARA
Department of Statistics and Applied Probability

Example



UC SANTA BARBARA
Department of Statistics and Applied Probability

Leadup

- Alright, so that's what a likelihood is. Why do we care?
- Well, again - I think of the likelihood as, well, the *likelihood* of obtaining a particular observation $\vec{\boldsymbol{y}}$ of a sample $\vec{\boldsymbol{Y}}$.
- So, here's the clever idea of how to leverage this to construct an estimator for θ - why don't we choose θ to maximize the likelihood of observing the sample that we actually observed?
- This is the idea behind **maximum likelihood estimation**, which we will begin next lecture.