STAT 346/446 Lecture 4

Methods of finding point estimators

CB Sections 7.1 - 7.2, DS Sections 7.5, 7.2 - 7.4

1. Method of Moments
2. Maximum Likelihood
   - examples
   - Invariance property
3. Bayes Estimators
   - Binomial-Beta model
   - Proportionality argument
   - Normal-Normal model

Note: We skip CB Section 7.2.4

Statistical Inference

- **Model:** Distribution of the population can be described with a distribution function (pmf or pdf) of a known form but with unknown parameters

$$f(x \mid \theta_1, \ldots, \theta_k)$$

  - So if we know the values of the parameters, we know all there is to know about the population.
  - Sometimes the parameter values θ_j have interpretable or physical meaning
    - E.g. population proportion of some trait of interest
  - Sometimes we are interested in a function of a parameter $\tau(\theta_j)$
- **Inference:** Have a sample $X_1, X_2, \ldots, X_n$ from $f(x \mid \theta)$ and want to use it to learn about the value of θ

Point estimation

Point estimator

A **point estimator** is any function $W(X_1, X_2, \ldots, X_n)$ of a sample

- Any *statistic* is a point estimator

A point estimate

A **point estimate** is the realized value of a point estimator

Example:

$$\textit{Estimator:} \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \textit{Estimate:} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- where $x_1, x_2, \ldots, x_n$ are the realized (observed) values of the sample $X_1, X_2, \ldots, X_n$

Method of Moments (MOM)

A sample $X_1, X_2, \ldots, X_n$ from a population with pmf/pdf $f(x \mid \theta_1, \ldots, \theta_k)$

- Idea: Match the first k sample moments with population moments and solve for parameters
- Define sample moments:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \ldots, \quad m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

- Population moments:

$$\mu_1' = E(X^1), \quad \mu_2' = E(X^2), \quad \ldots, \quad \mu_k' = E(X^k)$$

are usually functions of $\theta_1, \ldots, \theta_k$

Method of Moments (MOM)

A sample $X_1, X_2, \ldots, X_n$ from a population with pmf/pdf $f(x \mid \theta_1, \ldots, \theta_k)$

- Get k equations with k unknowns:

$$\begin{aligned} m_1 &= \mu_1'(\theta_1, \ldots, \theta_k) \\ m_2 &= \mu_2'(\theta_1, \ldots, \theta_k) \\ &\vdots \\ m_k &= \mu_k'(\theta_1, \ldots, \theta_k) \end{aligned}$$

  Solution will be functions of $m_1, \ldots, m_k$

$$\hat{\theta}_j = h_j(m_1, \ldots, m_k) \qquad j = 1, \ldots, k$$

If some population moments μ_1' are zero, continue with higher order moments until you have k equations.

Examples of MOMs

A sample $X_1, X_2, \ldots, X_n$ from a population with pmf/pdf $f(x \mid \theta_1, \ldots, \theta_k)$

Find the MOM estimators for these populations:

1. $\mathrm{N}(\mu, \sigma^2)$
2. $\mathrm{Poisson}(\lambda)$
3. $\mathrm{Gamma}(\alpha, \beta)$
4. t_ν

... on the board ...

Possible problems with MOM estimators

- Don't always exist
- May give estimates that are not in the parameter space
- Not always unique - can use higher moments or central moments

Maximum Likelihood Estimators

Likelihood Function

The joint pdf or pmf of a sample $X_1, \ldots, X_n$ is called a **Likelihood function** when considered a function of it's parameters

$$L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta) \qquad \theta \in \Theta$$

- When $X_1, \ldots, X_n$ is a random sample we have

$$L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- **log-likelihood function:** $\ell(\boldsymbol{\theta} \mid \mathbf{x}) = \log(L(\boldsymbol{\theta} \mid \mathbf{x}))$

Maximum Likelihood Estimators

Maximum likelihood estimator (MLE)

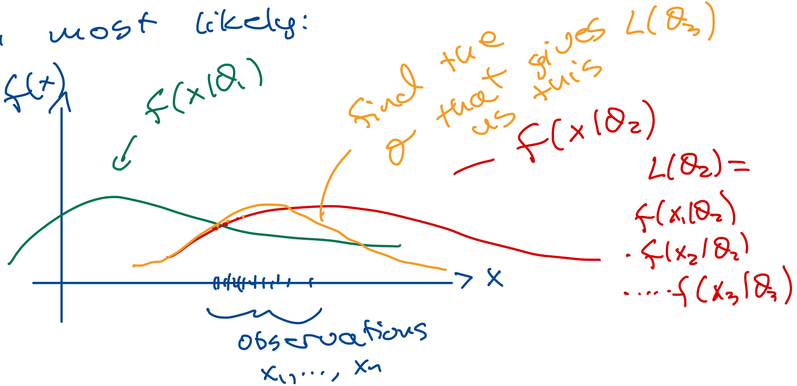
For each possible observation $\mathbf{x}$ let $\hat{\theta}(\mathbf{x})$ be the parameter value that maximizes $L(\theta \mid \mathbf{x})$. Then the statistic

$$\hat{\theta}(\mathbf{X})$$

is called the **Maximum Likelihood Estimator** of θ.

- Intuition: Given the data we observed, pick the value of θ that makes the likelihood (and therefore the joint pdf) the largest.
  - I.e. pick the parameter that makes the data "most likely"
- By restricting the optimization to the parameter space Θ we get estimates that are valid parameter values.
- Has the same problems as any optimization problem - global maximum may not exist or may be hard to find

Pick the θ that makes observed data most likely:



By changing θ we change the pdf $f(x|\theta)$

Finding Maximum Likelihood Estimators

- Differentiation
  1. Find *extreme points* in the *interior* of Θ by setting the first derivative equal to zero
  2. Check whether points give maximum (e.g. using second derivatives)
  3. Check boundary points
- Monotone functions
  - Often useful when domain depends on the parameter
- Often easier to maximize the log-likelihood - it gives the same result since

$$\hat{\theta}(\mathbf{x}) = \arg\max_{\theta \in \Theta} L(\boldsymbol{\theta} \mid \mathbf{x}) = \arg\max_{\theta \in \Theta} \ell(\boldsymbol{\theta} \mid \mathbf{x})$$

Finding maximum using one variable calculus

- If the likelihood function is differentiable (w.r.t θ) the MLE can be found easily: Set

$$\frac{d}{d\theta}L(\theta \mid \mathbf{x}) = 0 \qquad \text{or, if easier:} \qquad \frac{d}{d\theta}\ell(\theta \mid \mathbf{x}) = 0$$

  and solve for θ to find *extreme points* in the *interior* of Θ.
- We have a maximum if

$$\frac{d^2}{d\theta^2}L(\theta \mid \mathbf{x})\bigg|_{\theta=\hat{\theta}} < 0 \qquad \text{or:} \qquad \frac{d^2}{d\theta^2}\ell(\theta \mid \mathbf{x})\bigg|_{\theta=\hat{\theta}} < 0$$

  where $\hat{\theta}$ is the solution from above
- If Θ is bounded, check whether the boundary points of give a larger value of $L(\theta \mid \mathbf{x})$

Example 1 – Poisson

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{Poisson}(\lambda)$. Find the MLE of λ.

on the board...

Example 2: Discrete distribution

→ Only on observation, $L(\theta) = f(x \mid \theta)$

- Let X be a discrete random variable with pmf that depends on θ and $\theta \in \{0, 1, 2, 3\}$. The pmf, for different values of θ:

maximum

$f(x \mid \theta)$	x = 1	2	3	4
$\theta = 0$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\theta = 1$	$\frac{1}{2}$	0	0	$\frac{1}{2}$
$\theta = 2$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
$\theta = 3$	0	$\frac{1}{2}$	0	$\frac{1}{2}$

What if we observed $x=1$
→ $\theta=1$ gives highest $L(\theta)$

$x=2 \Rightarrow \theta=3$

$x=3 \Rightarrow \theta=2$

$x=4 \Rightarrow$

$$\hat{\theta}^{MLE} = \begin{cases} 1 & \text{if } x=1 \\ 3 & \text{if } x=2 \\ 2 & \text{if } x=3 \\ 1 \text{ or } 3 & \text{if } x=4 \end{cases}$$

- Find the MLE for θ

Example 2: Discrete distribution

- Let X be a discrete random variable with pmf that depends on θ and $\theta \in \{0, 1, 2, 3\}$. The pmf, for different values of θ:

$f(x \mid \theta)$	x = 1	2	3	4
$\theta = 0$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\theta = 1$	$\frac{1}{2}$	0	0	$\frac{1}{2}$
$\theta = 2$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
$\theta = 3$	0	$\frac{1}{2}$	0	$\frac{1}{2}$

2 obs? $L(\theta) = f(x_1 \mid \theta) f(x_2 \mid \theta)$

x_1	x_2	0	1	2	3	$\hat{\theta}$
2	3	$\left(\frac{1}{4}\right)^2$	0	$\left(\frac{1}{3}\right)^2$	0	2
etc.						
$\vdots$						

- Find the MLE for θ

Example 3: Uniform

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{Uniform}(0, \theta)$, $\theta > 0$. Find the MLE for θ

on the board

Example 4: Another uniform distribution

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{Uniform}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$
  Find the MLE for θ

on the board

Example 5: Normal distribution

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$
  Find the MLE for μ and σ^2

on the board

Finding maximum using two variable calculus

- To verify that a function $L(\theta_1, \theta_2 \mid \mathbf{x})$ has a *local* maimum at $(\hat{\theta}_1, \hat{\theta}_2)$ the following three conditions must hold

1. First order partial derivatives are zero

$$\left.\frac{\partial}{\partial \theta_1} L(\theta_1, \theta_2 \mid \mathbf{x})\right|_{\substack{\theta_1 = \hat{\theta}_1, \\ \theta_2 = \hat{\theta}_2}} = 0 \quad \text{and} \quad \left.\frac{\partial}{\partial \theta_2} L(\theta_1, \theta_2 \mid \mathbf{x})\right|_{\substack{\theta_1 = \hat{\theta}_1, \\ \theta_2 = \hat{\theta}_2}} = 0$$

2. At least one second-order partial derivatives is negative

$$\left.\frac{\partial^2}{\partial \theta_1^2} L(\theta_1, \theta_2 \mid \mathbf{x})\right|_{\substack{\theta_1 = \hat{\theta}_1, \\ \theta_2 = \hat{\theta}_2}} < 0 \quad \text{or} \quad \left.\frac{\partial^2}{\partial \theta_2^2} L(\theta_1, \theta_2 \mid \mathbf{x})\right|_{\substack{\theta_1 = \hat{\theta}_1, \\ \theta_2 = \hat{\theta}_2}} < 0$$

3. The Jacobian of the second order partial derivatives is positive

$$\left|\begin{array}{cc} \frac{\partial^2}{\partial \theta_1^2} L(\theta_1, \theta_2 \mid \mathbf{x}) & \frac{\partial^2}{\partial \theta_1 \theta_2} L(\theta_1, \theta_2 \mid \mathbf{x}) \\ \frac{\partial^2}{\partial \theta_2 \theta_1} L(\theta_1, \theta_2 \mid \mathbf{x}) & \frac{\partial^2}{\partial \theta_2^2} L(\theta_1, \theta_2 \mid \mathbf{x}) \end{array}\right|_{\theta_1 = \hat{\theta}_1, \theta_2 = \hat{\theta}_2} > 0$$

MLE for the Normal distribution

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$
- We saw that the MLE for μ and σ^2 are

$$\hat{\mu}^{\text{MLE}} = \overline{X} \qquad \text{and} \qquad (\widehat{\sigma^2})^{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Note that

$$(\widehat{\sigma^2})^{\text{MLE}} = (\widehat{\sigma^2})^{\text{MOM}} = \frac{n-1}{n}S^2 \quad \text{where} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- MLE for the standard deviation σ?

Invariance property of MLEs

Theorem: Invariance property of MLE

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\cdot)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Example

- $X_1, X_2, \ldots, X_n$ i.i.d. $\mathrm{N}(\mu, \sigma^2)$
- Since MLE for σ^2 is

$$(\widehat{\sigma^2})^{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

we get that the MLE for $\sigma = \sqrt{\sigma^2}$ is

$$(\widehat{\sigma})^{\text{MLE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Bayes Estimators

1. $X_1, X_2, \ldots, X_n$ is a random sample from $f(x \mid \theta)$
2. θ is an unknown constant
3. Use observed data $x_1, x_2, \ldots, x_n$ to learn about θ
   - Via estimators and sampling distributions

Bayesian inference

1. $X_1, X_2, \ldots, X_n$ is a random sample from $f(x \mid \theta)$
   - But the "likelihood" $f(\mathbf{x} \mid \theta)$ is viewed as the joint *conditional* pmf/pdf of $X_1, X_2, \ldots, X_n$ given θ.
2. θ is a random variable with a **prior distribution** $p(\theta)$
3. Use observed data $x_1, x_2, \ldots, x_n$ to learn about θ
   - By obtaining the **posterior distribution** $p(\theta \mid \mathbf{x})$

Bayesian Statistics

STAT 448: Bayesian Theory with Applications

- A Bayesian model specifies both
  - a likelihood $f(\mathbf{x} \mid \theta)$
  - and a prior distribution $p(\theta)$
- Note that together they define the joint distribution of the data and the parameters, $(\mathbf{x}, \theta)$ since

$$p(\mathbf{x}, \theta) = f(\mathbf{x} \mid \theta) p(\theta)$$

- Bayesian inference is all contained in the posterior $p(\theta \mid \mathbf{x})$
- A point estimator in this setting is just a on-number summary of the posterior distribution (mean, median or mode)

Posterior distribution

- The **posterior distribution** follows from Bayes theorem

$$p(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)p(\theta)}{m(\mathbf{x})}$$

  where

$$m(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)p(\theta)d\theta$$

- The posterior mean of θ is called the **Bayes estimator** of θ

$$\hat{\theta}^B = E(\theta \mid \mathbf{X} = \mathbf{x})$$

  - Like any other point estimator, $\hat{\theta}^B$ is a function of $X_1, \ldots, X_n$ (a statistic)

Binomial-Beta model

- Suppose $Y \sim \text{Binomial}(n, \theta)$. Then

$$f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad y = 0, 1, \ldots, n$$

- Goal: Learn about (estimate) the population proportion, θ.
  - Need to pick a prior distribution for θ, with support on $(0, 1)$.

- Popular prior distribution: $\theta \sim \text{Beta}(\alpha, \beta)$

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \, \theta^{\alpha-1} \, (1-\theta)^{\beta-1} \qquad \text{for } 0 < \theta < 1$$

where α and β are known

Posterior for Binomial-Beta model

- First find the marginal pdf

$$
\begin{aligned}
m(y) &= \int f(y \mid \theta)\, p(\theta)\, d\theta \\
&= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}
\end{aligned}
$$

$= \ldots$ on the board

$$= \frac{\binom{n}{y} \Gamma(\alpha+\beta)\, \Gamma(\alpha+y)\, \Gamma(\beta+n-y)}{\Gamma(\alpha)\Gamma(\beta)\, \Gamma(\alpha+\beta+n)}$$

Posterior for Binomial-Beta model

- The posterior pdf is

$$p(\theta \mid y) = \frac{f(y \mid \theta)\, p(\theta)}{m(y)}$$

$$= \frac{\binom{n}{y} \theta^{y} (1-\theta)^{n-y} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\dfrac{\binom{n}{y} \Gamma(\alpha+\beta)\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha)\Gamma(\beta)\, \Gamma(\alpha+\beta+n)}}$$

$$= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)} \theta^{\alpha+y-1} (1-\theta)^{\beta+n-y-1}$$

$= \text{pdf of Beta}(\alpha+y, \beta+n-y)$

Bayes estimator for Binomial-Beta model

- Recall that the mean of the $\mathrm{Beta}(\alpha, \beta)$ distribution is $\frac{\alpha}{\alpha+\beta}$
- So the Bayes estimator of θ is

$$\hat{\theta}^B = \frac{\alpha + Y}{\alpha + Y + \beta + n - Y} = \frac{\alpha + Y}{\alpha + \beta + n} \approx Y \text{ if}$$

α, β are small compared to Y and n.

- **Conjugacy:** Prior and Posterior are from the same family of distributions
  - E.g. $\theta \sim \mathrm{Beta}(\alpha, \beta)$ and $\theta \mid y \sim \mathrm{Beta}(\tilde{\alpha}, \tilde{\beta})$

Proportionality argument

- Note that

$$p(\theta \mid \mathbf{x}) = \frac{f(y \mid \theta)\, p(\theta)}{m(y)} \propto f(y \mid \theta)\, p(\theta)$$

proportional to

more on the board...

- Sometimes we can recognize the functional form (of θ) as the kernel of a known pdf

- Example: Binomial-Beta model

$$\begin{aligned} p(\theta \mid y) &\propto \binom{n}{y} \theta^{y} (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &\propto \text{ pdf of the } \mathrm{Beta}(\alpha+y, \beta+n-y) \text{ distribution.} \end{aligned}$$

Normal-Normal model

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{N}(\theta, \sigma^2)$, where σ^2 is known. Joint likelihood:

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\left[-2\theta n\overline{x} + n\theta^2\right]\right)$$

Normal-Normal model

- Prior distribution on θ: $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$

$$p(\theta) = \frac{1}{\sqrt{2\pi}\,\tau_0} \exp\left(-\frac{(\theta-\mu_0)^2}{2\tau_0^2}\right)$$

...

$$\propto \exp\left(-\frac{1}{2}\left[\frac{\theta^2}{\tau_0^2} - \frac{2\theta\mu_0}{\tau_0^2}\right]\right)$$

- In general the kernel of a normal distribution:

$$\exp\left(-\frac{1}{2}\left[\frac{\text{variable}^2}{\text{variance}} - \frac{2\cdot\text{variable}\cdot\text{mean}}{\text{variance}}\right]\right)$$

Posterior for Normal-Normal model

- Posterior distribution:

$$
\begin{aligned}
p(\theta \mid \mathbf{x}) &\propto \exp\left(-\frac{1}{2\sigma^2}\left[-2\theta n\bar{x} + n\theta^2\right]\right) \exp\left(-\frac{1}{2}\left[\frac{\theta^2}{\tau_0^2} - \frac{2\theta\mu_0}{\tau_0^2}\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[\theta^2\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu_0}{\tau_0^2} + \frac{n\bar{x}}{\sigma^2}\right)\right]\right) \\
&\propto \text{ pdf of } \mathrm{N}(\mu_1, \tau_1^2)
\end{aligned}
$$

Posterior for Normal-Normal model

- The posterior distribution of θ is $\mathrm{N}(\mu_1, \tau_1^2)$ where

$$\tau_1^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1} = \frac{\sigma^2/n\,\tau_0^2}{\sigma^2/n + \tau_0^2}$$

$$\mu_1 = \tau_1^2\left(\frac{\mu_0}{\tau_0^2} + \frac{n\overline{x}}{\sigma^2}\right) = \frac{\tau_0^2}{\sigma^2/n + \tau_0^2}\overline{x} + \frac{\sigma^2/n}{\sigma^2/n + \tau_0^2}\mu_0$$

- Posterior mean is a weighted average of data average ($\overline{x}$) and prior mean (μ_0).
  - Weights depend on σ^2, n, and τ_0^2