Lecture 11: Point Estimation

Ye Tian

Department of Statistics, Columbia University
Calculus-based Introduction to Statistics (S1201)

July 27, 2022
Where we are

- **Goal of statistics**: To make inferences on parameters of a population, which are assumed to be **fixed but unknown**. Examples: population proportion, mean, standard deviation, median etc.
- **Two types of estimation**:
  - **Point estimation**: use a **single value** as the estimate of the parameter
  - **Confidence interval**: point out a **range (interval)** which is very likely to cover the parameter
Two types of estimation

Two types of estimation:
- **Point estimation**: use a single value as the estimate of the parameter
- **Confidence interval**: point out a range which covers the parameter with high probability

An example: If we want to estimate the acceptance rate of Columbia next year...
- Mike: It might be 7%. → a point estimation
- Lee: It could be between 6% and 8%. → a confidence interval
This week's goal

- Understand two methods of point estimation and know how to calculate them (Today)
  - Method of moments
  - Maximum likelihood estimation (MLE)
- Understand confidence intervals and know how to construct them (Tuesday and Wednesday)
Some Concepts
Some concepts

- Category of "models":
  - **Parametric**: the form of population distribution (i.e. which distribution the sample follows) is **known** (but the true parameter is still unknown)
  - **Nonparametric**: the form of population distribution is **unknown**

- In this course, we focus on **parametric models**.
  1. Assume (Or have already known) the distribution of data
  2. Estimate the parameters of interest based on (1)
  3. Check the assumption you make in (1)
Some concepts

(Sample) Statistic: A function of sample $X_1, \ldots, X_n$, which is calculable given the samples. It does NOT depend on unknown parameters.

Estimator: A statistic used to estimate the parameter.

Estimate: A numerical value of the estimator based on current sample.

Remark: Estimator is a random variable! (Why?)

An "old" example: The grades of final exams of this course last summer:
99, 70, 74, 55, 60, 60, 80, 88, 85, 92, 98, 100, 86, 85, 74, 90, 72, 92, 88, 87, 81, 100, 79, 90, 68, 89, 91, 90, 96, 85

We can have many different estimators for the true mean score $\theta$:

- $\hat{\theta}_1 = \bar{X}$ (note that this is a r.v.), estimate $= \bar{x} = 83.47$
- $\hat{\theta}_2 = \text{sample median}$ (note that this is a r.v.), estimate $= \bar{x} = 86.5$
- $\hat{\theta}_3 = \frac{1}{2} \left( \min_{1 \leq i \leq n} X_i + \max_{1 \leq i \leq n} X_i \right)$ (note that this is a r.v.), estimate $= 77.5$
- $\hat{\theta}_4 = \bar{X} + 10$ (note that this is a r.v.), estimate $= \bar{x} = 93.47$

Question: How to evaluate them? Which ones are better?
Bias and unbiased estimator

Suppose we are interested in a population parameter $\theta$ and considering estimator $\hat{\theta}$.

**Estimation bias:** $\mathbb{E}\hat{\theta} - \theta$

**Unbiased estimator:** The estimator $\hat{\theta}$ that satisfies $\mathbb{E}\hat{\theta} = \theta$

**Examples of unbiased estimator:**

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$. Then $\hat{\theta} = \bar{X}$ is an unbiased estimator of $\theta = \mathbb{E}X$
  - $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$. $\hat{p} = \bar{X}$ is an unbiased estimator of $p$.
  - $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$. $\hat{\mu} = \bar{X}$ is an unbiased estimator of $\mu$.

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$. Then $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator of $\text{Var}(X)$. (You will prove this in HW4!)
Method of Moments
Method of moments

**Definition:** Let $X_1, \ldots, X_n$ be an independent random sample from a pmf or a pdf $f(x)$. For $k = 1, 2, 3, \ldots$, the $k$-th population moment is $\mathbb{E}(X^k)$. The $k$-th sample moment is $\frac{1}{n} \sum_{i=1}^{n} X_i^k$.

- Sample mean $\bar{X}$ is the first sample moment and $\mathbb{E}X$ is the first population moment.
- In many cases, sample moments are good estimators for population moments (by Weak Law of Large Number!)

**Example:** Suppose $X_1, \ldots, X_n$ are from the following pmf:

$$
P(X = 1) = a, \quad P(X = 2) = b, \quad P(X = 4) = 1/4, \quad P(X \neq 1, 2, 4) = 0.
$$

Construct estimators of parameters $a$ and $b$.

First by definition of pmf: sum of probabilities $= a + b + 1/4 = 1$

And $\mathbb{E}X = a + 2b + 1$. 
Method of moments

**Example:** Suppose $X_1, \ldots, X_n$ are from the following pmf:

$$P(X = 1) = a, \quad P(X = 2) = b, \quad P(X = 4) = 1/4, \quad P(X \neq 1, 2, 4) = 0.$$  

Construct estimators of parameters $a$ and $b$.

First by definition of pmf: sum of probabilities $= a + b + 1/4 = 1$

And $\mathbb{E}X = a + 2b + 1$.

Population-level equations

$$\begin{cases} a + b = \frac{3}{4}, \\ a + 2b + 1 = \mathbb{E}X \end{cases}$$

Sample-level equations

$$\begin{cases} a + b = \frac{3}{4}, \\ a + 2b + 1 = \bar{X} \end{cases}$$

Finally, motivated by the equation system on RHS, we can use the following estimators

$$\hat{a} = \frac{5}{2} - \bar{X}, \quad \hat{b} = \bar{X} - \frac{7}{4}.$$
A general statement of method of moments

Suppose $X_1, \ldots, X_n$ are from some pmf or pdf. $\theta_1, \ldots, \theta_m$ are unknown parameters. We want to construct estimators of $\theta_1, \ldots, \theta_m$.

**Method of moments:**

1. Calculate $m$ population moments and express them as functions of $\theta_1, \ldots, \theta_m$ (e.g.: $f_1, \ldots, f_m$ below)

   $$
   \begin{align*}
   \mathbb{E}X &= f_1(\theta_1, \ldots, \theta_m), \\
   \mathbb{E}(X^2) &= f_2(\theta_1, \ldots, \theta_m), \\
   \vdots \\
   \mathbb{E}(X^m) &= f_m(\theta_1, \ldots, \theta_m).
   \end{align*}
   $$

2. Replace the *population* moments by *sample* moments:

   $$
   \begin{align*}
   \frac{1}{n} \sum_{i=1}^{n} X_i &= f_1(\theta_1, \ldots, \theta_m), \\
   \frac{1}{n} \sum_{i=1}^{n} X_i^2 &= f_2(\theta_1, \ldots, \theta_m), \\
   \vdots \\
   \frac{1}{n} \sum_{i=1}^{n} X_i^m &= f_m(\theta_1, \ldots, \theta_m).
   \end{align*}
   $$

3. Solve the $m$ equations (w.r.t. $\theta_1, \ldots, \theta_m$) to get the *moment estimators* $\hat{\theta}_1, \ldots, \hat{\theta}_m$
More examples

\(X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Exp}(\lambda)\). Find an estimator of \(\lambda\).

\[
\mathbb{E}X = 1/\lambda \Rightarrow \bar{X} = 1/\hat{\lambda}
\]

\[
\Rightarrow \hat{\lambda} = 1/\bar{X}
\]

Actually we can also use the second moment:

\[
\mathbb{E}X^2 = \text{Var}(X) + (\mathbb{E}X)^2 = 2/\lambda^2 \Rightarrow \frac{1}{n} \sum_{i=1}^{n} X_i^2 = 2/\hat{\lambda}^2
\]

\[
\Rightarrow \hat{\lambda} = \sqrt{\frac{2n}{\sum_{i=1}^{n} X_i^2}}
\]

They are both moment estimators!
Maximum Likelihood Estimation
Review: Independence between random variables

Suppose the joint pmf or pdf of $X_1, \ldots, X_n$ is $f(x_1, \ldots, x_n)$. And the marginal pmf or pdf of $X_i$ is $f_i(x_i)$. Then $X_1, \ldots, X_n$ are independent if and only if

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i),$$

for any numbers $x_1, \ldots, x_n$. 
An example

10 individuals with Columbia email accounts are selected, and the first, third, and tenth individuals are using a "strong" password, whereas the others do not. If the probability that each person uses a strong password is Bernoulli$(p)$. Estimate $p$.

**Method 1: Method of moments.** Define

$$X_i = \begin{cases} 1, & \text{i-th individual uses a strong password,} \\ 0, & \text{otherwise} \end{cases}$$

Then $X_i \sim \text{Bernoulli}(p)$. Thus, $\mathbb{E}X_i = p \Rightarrow \hat{p} = \bar{X}$.

The estimate $\bar{x} = 3/10$. 
An example

10 individuals with Columbia email accounts are selected, and the first, third, and tenth individuals are using a "strong" password, whereas the others do not. If the probability that each person uses a strong password is Bernoulli($p$). Estimate $p$.

**Method 2:** Joint pmf $P(X_1 = X_3 = X_{10} = 1, \text{others} = 0) = p^3(1 - p)^7$.

- This joint pmf represents the possibility of observing the current samples under a specific value of $p$
- We call the joint pmf value under current observations a **likelihood function**, which is a function of parameter $p$.
- Is the likelihood the larger the better? Why?
Maximum likelihood estimation

$X_1, \ldots, X_n$ have a joint pmf or pdf $f(x_1, \ldots, x_n; \theta)$, where $\theta$ is an unknown parameter. Suppose $x_1, \ldots, x_n$ are the observed sample values.

- Then $f(x_1, \ldots, x_n; \theta)$ can be seen as a function of $\theta$, which is called the likelihood function.
- The maximum likelihood estimate (MLE) is the value of $\theta$ that maximize $f(x_1, \ldots, x_n; \theta)$. In general, the maximum likelihood estimator (MLE) is the value of $\theta$ that maximize $f(X_1, \ldots, X_n; \theta)$.
- Equivalently, we can find the MLEs by finding the maximizer of log-likelihood $\ell(\theta) = \log f(X_1, \ldots, X_n; \theta)$.\(^1\)
- Ways to solve MLEs:
  - Finite $\theta$ choices: try all possible values or by graph of $\ell(\theta)$
  - A range of $\theta$: take the derivative of $\ell(\theta)$ and set it to be 0 i.e. solve the equation $\frac{d\ell(\theta)}{d\theta} = 0$.

\(^1\)Can you tell why log-likelihood makes things easier compared to the likelihood?
The previous example

Log-likelihood

\[ \ell(p) = \log[P(X_1 = X_3 = X_{10} = 1, \text{others} = 0)] \]
\[ = \log(p^3(1 - p)^7) \]
\[ = 3 \log p + 7 \log(1 - p). \]

Set \( \ell'(\hat{p}) = \frac{3}{\hat{p}} - \frac{7}{1 - \hat{p}} = 0 \Rightarrow \hat{p}_{\text{MLE}} = 0.3 \)
More examples

Example: \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Exp}(\lambda) \). Find the MLE of \( \lambda \).

The likelihood function equals the joint pdf

\[
f(X_1, \ldots, X_n; \lambda) = \prod_{i=1}^{n} (\lambda e^{-\lambda X_i}) \quad \text{(by independence)}
\]

\[
= \lambda^n \exp \left\{ - \lambda \sum_{i=1}^{n} X_i \right\}.
\]

Log-likelihood \( \ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} X_i \). Let \( \ell'(\hat{\lambda}) = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} X_i = 0 \),
we get \( \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = 1/\bar{X} \).

- Here the MLE is the same as the method of moments estimator (by 1st moment)
- But the MLE is NOT always equal to the method of moments estimator
More examples

**Example:** $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$. Find the maximum likelihood estimator of $\mu$.

The likelihood function equals the joint pdf

$$f(X_1, \ldots, X_n; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (X_i - \mu)^2 \right\} \quad \text{(by independence)}$$

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

Log-likelihood $\ell(\mu) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2$. Let $\ell'(\hat{\mu}) = -\sum_{i=1}^{n} (\mu - X_i) = 0$, we get $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$. 
Reading list (optional)

- "Probability and Statistics for Engineering and the Sciences" (9th edition):
  - Chapter 6.1 (only read the part before "Estimators with Minimum Variance")
  - Chapter 6.2 (skip the part "Large Sample Behavior of the MLE")
- "OpenIntro statistics" (4th edition, free online, download [here]):
  - Chapter 5.1
Many thanks to
- Yang Feng
- Joyce Robbins
- Chengliang Tang
- Owen Ward
- Wenda Zhou
- And all my teachers in the past 25 years