Section 3: Asymptotics

Ricky Truong (rickytruong@college.harvard.edu),
Emily Xing (exing@college.harvard.edu)

1 Introduction

1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

1.2 Office Hours

- Mondays, 7:30 - 9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM - 12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30 – 11:30 AM in Cabot D-Hall (Emily).

2 Big Picture

Asymptotics describe behavior in the limit as the sample size approaches infinity. In the real world, we never have an infinite sample size, but asymptotic theory gives us approximations when n is sufficiently large. Recall the notion of *convergence* from calculus. Since random variables are different from usual numbers, we introduce different types of convergence for sequences of random variables. In the course, we'll be working with some asymptotic tools to describe *convergence in probability* and *convergence in distribution*.

We can build on likelihood theory to define the *score function* and *Fisher information*, which allow us to examine properties of the MLE when *regularity conditions* hold.

Related, we often examine different values of the parameter θ, but we can denote the *true* value as θ^*. E.g., suppose the true data-generating process is $\mathcal{N}(1000, 200^2)$. We may believe $\theta = 900$ based on our data, but actually, $\theta^* = 1000$. (Of course, if we're conducting inference, we wouldn't know this value—this is all theoretical!)

3 Convergence

Idea. For sequences of random variables, the three main types of convergence build upon the notion of convergence with usual numbers. From strongest to weakest, we have *almost sure convergence*, *convergence in probability*, and *convergence in distribution*.

3.1 Convergence for Numbers

Definition 1 (Sequence)**.** Informally, an ordered list of numbers, often denoted as $\{x_n\}_{n=1}^{\infty}$.

- Formally, a sequence on a set S is a function $f : \mathbb{N} \to S$.
- E.g., consider the sequence $\{\frac{1}{n^2}\}_{n=1}^{\infty} = 1, \frac{1}{4}, \frac{1}{9}, \ldots$

Definition 2 (Convergence of a sequence)**.** Informally, a sequence converges if its terms get progressively closer to a single, finite number (the limit) as the number of terms (n) increases to infinity, essentially "settling down" on the limit.

- Formally, let $\{x_n\}_{n=1}^{\infty}$ be a sequence of real numbers. It converges to a limit L if $\forall\ \varepsilon > 0$, $\exists\ N \in \mathbb{N}$ such that $|x_n - L| < \varepsilon\ \forall\ n \geq N$.
- E.g., the sequence $\{\frac{1}{n^2}\}_{n=1}^{\infty}$ converges to 0.

3.2 Convergence for Random Variables

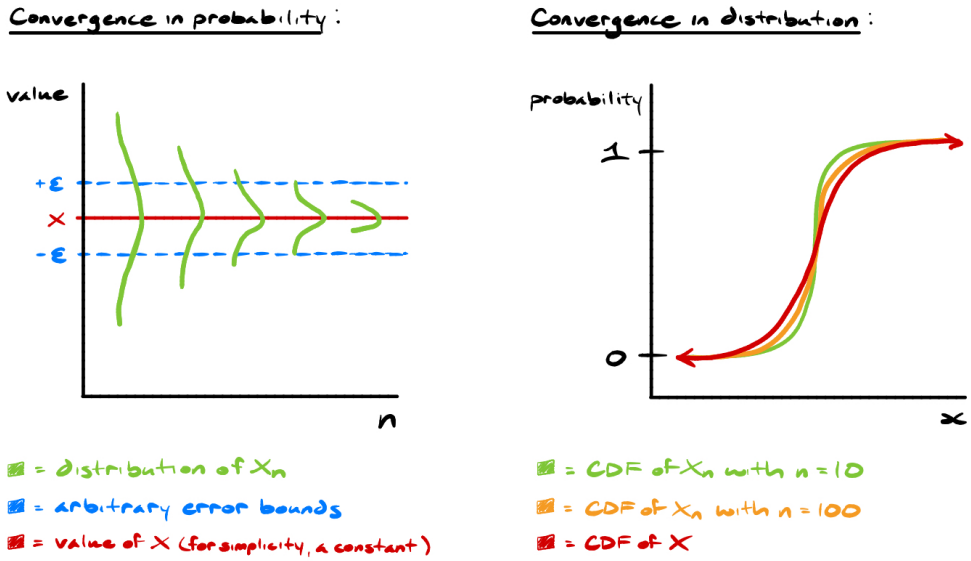

FIGURE 1: Convergence in probability vs. convergence in distribution.

Definition 3 (Almost sure convergence)**.** Let $X_n = X_1, X_2, \ldots$ be a sequence of random variables. X_n converges almost surely to a limit X if $P(\lim_{n\to\infty} X_n = X) = 1$, denoted $X_n \xrightarrow{a.s.} X$.

- We won't be working with almost sure convergence very much in the course.

Definition 4 (Convergence in probability)**.** Let $X_n = X_1, X_2, \ldots$ be a sequence of random variables. X_n converges in probability to a limit X if $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$, denoted $X_n \xrightarrow{p} X$.

- In Figure 1, notice the distribution gets "skinnier" around X as n increases, resulting in practically 0 probability mass/density away from X, no matter the distance ε.

Definition 5 (Convergence in distribution)**.** Let $X_n = X_1, X_2, \ldots$ be a sequence of random variables with CDF $F_{X_n}(x)$. X_n converges in distribution to a limit X if $F_{X_n}(x)$ converges to a limiting CDF $F_X(x)$ (i.e., $\forall\ x$ where $F_X(x)$ is continuous, $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$), denoted $X_n \xrightarrow{d} X$.

- In Figure 1, notice the CDFs get closer to $F_X(x)$ as n increases.
- Importantly, this is the "weakest" type of convergence as $\xrightarrow{a.s.} \implies \xrightarrow{p} \implies \xrightarrow{d}$, but generally, $\xrightarrow{d} \not\implies \xrightarrow{p} \not\implies \xrightarrow{a.s.}$. However, $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$ for constant c.

• For some intuition, if X is the number of heads and Y is the number of tails in n coin flips, X and Y share the same distribution—i.e., $\text{Bin}(n, 0.5)$—but are different quantities!

Example 1 (Sample mean)**.** For n random variables $X_1, \ldots, X_n$, let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be their sample mean.[1] Notice $\bar{X}_n$ is a sequence of random variables.

• For $n = 1$, $\bar{X}_1 = X_1$, which is a random variable.

• For $n = 2$, $\bar{X}_2 = \frac{1}{2}(X_1 + X_2)$, which is a random variable.

• For $n = 3$, $\bar{X}_3 = \frac{1}{3}(X_1 + X_2 + X_3)$, which is a random variable.

• As a sequence, we have $\bar{X}_n = X_1, \frac{1}{2}(X_1 + X_2), \frac{1}{3}(X_1 + X_2 + X_3), \ldots$

Concept Checker 1. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with $\mathbb{E}[Y_1] = \mu$ and $\text{Var}[Y_1] = \sigma^2$. Let $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ be the sample mean. What are $\mathbb{E}[\bar{Y}_n]$ and $\text{Var}[\bar{Y}_n]$?

Solution

An important result is $\mathbb{E}[\bar{Y}_n] = \mu$ and $\text{Var}[\bar{Y}_n] = \frac{\sigma^2}{n}$ with this setup. A detailed solution is provided in "Section 1: Statistics."

4 Asymptotic Tools

Idea. Now that we understand the different types of convergence, we can proceed with our main asymptotic tools, which allow us to make powerful statements about the convergence of sequences of random numbers.

Definition 6 (Strong Law of Large Numbers)**.** For n i.i.d. random variables $X_1, \ldots, X_n$ with finite expectation $\mathbb{E}[X_i] = \mu$, finite variance $\text{Var}[X_i] = \sigma^2$, and sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{X}_n \xrightarrow{a.s.} \mu$.

Definition 7 (Weak Law of Large Numbers)**.** For n i.i.d. random variables $X_1, \ldots, X_n$ with finite expectation $\mathbb{E}[X_i] = \mu$, finite variance $\text{Var}[X_i] = \sigma^2$, and sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{X}_n \xrightarrow{p} \mu$.

• In the course, we often just say Law of Large Numbers (LLN) and use convergence in probability.

Definition 8 (Central Limit Theorem)**.** For n i.i.d. random variables $X_1, \ldots, X_n$ with finite expectation $\mathbb{E}[X_i] = \mu$, finite variance $\text{Var}[X_i] = \sigma^2$, and sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

• This implies $\bar{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$ for large n.

• ☣: It is incorrect to say $\bar{X}_n \xrightarrow{d} \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Why? Recall convergence in distribution is in the limit as n approaches infinity, but here, n is on the right side of the equation! Thus, for each n, the "thing being converged to" keeps changing.

[1] You've probably seen sample mean denoted as simply $\bar{X}$. The notation $\bar{X}_n$ is equivalent, with the subscript n to emphasize this is a function of n.

Proof. Freeze at a sufficiently large n such that $\sqrt{n}(\bar{X}_n - \mu) \dot{\sim} \mathcal{N}(0, \sigma^2)$ by the Central Limit Theorem and definition of convergence in distribution. Treat n as a constant such that $\bar{X}_n \dot{\sim} \mathcal{N}(\mathbb{E}[\bar{X}_n], \text{Var}[\bar{X}_n])$ by shifting/scaling properties of the Normal.

$$\begin{aligned}
\sqrt{n}(\bar{X}_n - \mu) \dot{\sim} \mathcal{N}(0, \sigma^2) &\implies \text{Var}\left[\sqrt{n}(\bar{X}_n - \mu)\right] \approx \sigma^2 && \text{by variance of Normal} \\
&\implies n\text{Var}\left[\bar{X}_n - \mu\right] \approx \sigma^2 && \text{by bilinearity} \\
&\implies n\text{Var}\left[\bar{X}_n\right] \approx \sigma^2 && \text{by bilinearity} \\
&\implies \text{Var}\left[\bar{X}_n\right] \approx \frac{\sigma^2}{n} && \text{by algebra}
\end{aligned}$$

$$\begin{aligned}
\sqrt{n}(\bar{X}_n - \mu) \dot{\sim} \mathcal{N}(0, \sigma^2) &\implies \mathbb{E}\left[\sqrt{n}(\bar{X}_n - \mu)\right] \approx 0 && \text{by expectation of Normal} \\
&\implies \sqrt{n}\left(\mathbb{E}\left[\bar{X}_n - \mu\right]\right) \approx 0 && \text{by linearity} \\
&\implies \mathbb{E}\left[\bar{X}_n - \mu\right] \approx 0 && \text{by algebra} \\
&\implies \mathbb{E}\left[\bar{X}_n\right] - \mu \approx 0 && \text{by linearity} \\
&\implies \mathbb{E}\left[\bar{X}_n\right] \approx \mu && \text{by algebra}
\end{aligned}$$

□

Definition 9 (Continuous Mapping Theorem)**.** Let $X_n = X_1, X_2, \ldots$ be a sequence of random variables and $g(x)$ be a continuous function. If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$. Additionally, if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

Definition 10 (Theorem 3.5.7.)**.** Let $X_n = X_1, X_2, \ldots$ and $Y_n = Y_1, Y_2, \ldots$ be sequences of random variables. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$, $X_n - Y_n \xrightarrow{p} X - Y$, and $X_n Y_n \xrightarrow{p} XY$. Additionally, if $P(Y_n = 0) = P(Y = 0) = 0$, then $\frac{X_n}{Y_n} \xrightarrow{p} \frac{X}{Y}$.

Definition 11 (Slutsky's Theorem)**.** Let $X_n = X_1, X_2, \ldots$ and $Y_n = Y_1, Y_2, \ldots$ be sequences of random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for any constant c, then $X_n + Y_n \xrightarrow{d} X + c$, $X_n - Y_n \xrightarrow{d} X - c$, and $X_n Y_n \xrightarrow{d} cX$. Additionally, if $c \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.

Definition 12 (Delta Method)**.** If $g(x)$ is a differentiable function and $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$, then $\sqrt{n}\left(g(\hat{\theta}) - g(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0, (g'(\theta))^2\omega^2\right)$.

• This implies $g(\hat{\theta}) \dot{\sim} \mathcal{N}\left(g(\theta), (g'(\theta))^2 \frac{\omega^2}{n}\right)$ for large n.

• **Tip**: Whenever you see an "ugly" expression that's a function of a known random variable, define it as a new random variable to pattern match to one of the asymptotic tools! E.g., if $Y_1, \ldots, Y_n$ are i.i.d., then $Y_1^4, \ldots, Y_n^4$ are also i.i.d., so let $A_i = Y_i^4$. From there, $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n A_i - \mathbb{E}[A_i]\right)$ matches the form of CLT.

Concept Checker 2. Let $\text{Var}[\hat{\theta}] = \sigma^2$ and $g(x)$ be a differentiable function such that $g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$ by Taylor approximation. Show $\text{Var}[g(\hat{\theta})] \approx (g'(\theta))^2\sigma^2$.

Solution

$$\begin{aligned}
\text{Var}[g(\hat{\theta})] &\approx \text{Var}[g(\theta) + g'(\theta)(\hat{\theta} - \theta)] && \text{by substituting} \\
&\approx (g'(\theta))^2 \text{Var}[\hat{\theta} - \theta] && \text{by bilinearity} \\
&\approx (g'(\theta))^2 \text{Var}[\hat{\theta}] && \text{by bilinearity} \\
&\approx (g'(\theta))^2 \sigma^2 && \text{by substituting}
\end{aligned}$$

Concept Checker 3. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with finite expectation $\mathbb{E}[Y_1] = \mu \neq 0$ and finite variance $\text{Var}[Y_1] = \sigma^2 > 0$. What do the following sequences converge to?[2]

1. $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i^4 - \mathbb{E}[Y_i^4]\right) \xrightarrow{d}$ _____ by _____
2. $\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^3 \xrightarrow{p}$ _____ by _____
3. $\frac{\sqrt{n}(\bar{Y}_n - \mu)}{(\bar{Y}_n)^5 + \mu^5} \xrightarrow{d}$ _____ by _____

Solution

1. $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i^4 - \mathbb{E}[Y_i^4]\right) \xrightarrow{d} \mathcal{N}(0, \text{Var}[Y_i^4])$ by CLT.
2. $\frac{1}{n}\sum_{i=1}^n Y_i^2 \xrightarrow{p} \mathbb{E}[Y_1^2] = \sigma^2 + \mu^2$ by LLN, so $\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right)^3 \xrightarrow{p} (\sigma^2 + \mu^2)^3$ by CMT with $g(x) = x^3$.
3. First, $\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by CLT. Next, $\bar{Y}_n \xrightarrow{p} \mu$ by LLN, so $(\bar{Y}_n)^5 + \mu^5 \xrightarrow{p} \mu^5 + \mu^5 = 2\mu^5$ by CMT with $g(x) = x^5 + \mu^5$. Since $2\mu^5 \neq 0$, $\frac{\sqrt{n}(\bar{Y}_n - \mu)}{(\bar{Y}_n)^5 + \mu^5} \xrightarrow{d} \frac{X}{2\mu^5}$ by Slutsky's, where $X \sim \mathcal{N}(0, \sigma^2)$. Notice $\frac{X}{2\mu^5} \sim \mathcal{N}(0, \frac{\sigma^2}{4\mu^{10}})$ by property of the Normal, so $\frac{\sqrt{n}(\bar{Y}_n - \mu)}{(\bar{Y}_n)^5 + \mu^5} \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2}{4\mu^{10}})$.

Concept Checker 4. Let $X_1, \ldots, X_n$ be random variables where $X_1 \sim \mathcal{N}(0, 1)$ and $X_i = 2X_{i-1}$ for $i \in \{2, 3, \ldots, n\}$. Do we know what $\bar{X}_n$ converges to? If so, how?

Solution

In this case, the data are not i.i.d. since knowing one observation gives perfect information about another. Thus, we cannot apply our usual LLN.

5 Likelihood Theory

Idea. *Regularity conditions* are a set of requirements to ensure functions (e.g., likelihood) behave well enough for standard statistical theory to apply. If they hold (and they often will in the course), we enjoy some nice results.

The *score function* and *Fisher information* build onto log-likelihood. We've already seen the score function; we set it equal to 0 during the MLE derivation. Informally, a large

[2]Inspired by Problem 2 in "Stat 111 Homework 3, Spring 2025" by Joseph K. Blitzstein and Neil Shephard.

Fisher information means the data give us much information on what θ^* is, where θ^* is the *true* value of our estimand. There are special results when we evaluate score and Fisher information at θ^*.

5.1 Regularity Conditions

Definition 13 (Regularity conditions)**.** For data $Y_1, \ldots, Y_n \sim F_{\vec{Y};\theta}$, the following must be true:

- $\mathcal{L}(\theta; \vec{y})$ is a smooth function on Θ.
- $\mathbb{E}[s(\theta^*; \vec{Y})]$ and $\text{Var}[s(\theta^*; \vec{Y})]$ exist.
- The support of $\vec{Y}$ doesn't depend on θ.
- We can differentiate under the integral sign (i.e., DUThIS).

Concept Checker 5. Do regularity conditions hold for $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$?

Solution

Regularity conditions do NOT hold since $\text{supp}(Y_1) = [0, \theta]$, so $\text{supp}(\vec{Y}) = [0, \theta]^n$.

Concept Checker 6. Let $f_X(x)$ be the PDF for a random variable X. Assume we can DUThIS. How can we rewrite $\frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx$.

Solution

$$\begin{aligned}
\frac{d}{dt}\int_{-\infty}^{\infty} e^{tx} f_X(x)dx &= \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f_X(x)dx && \text{by DUThIS} \\
&= \int_{-\infty}^{\infty} x e^{tx} f_X(x)dx && \text{by chain rule} \\
&= \mathbb{E}[X e^{tX}] && \text{by LOTUS}
\end{aligned}$$

5.2 Score Function and Fisher Information

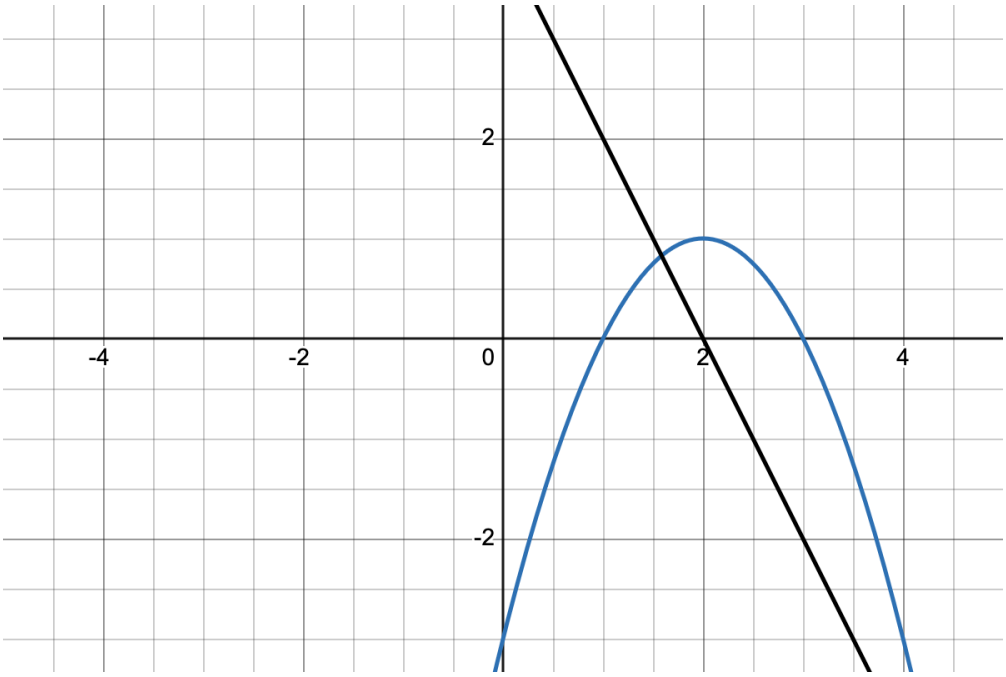

FIGURE 2: The graph for $\ell(\theta;\vec{y}) = -\theta^2 + 4\theta - 3$ (in blue) and $s(\theta;\vec{y}) = \frac{\partial}{\partial\theta}\ell(\theta;\vec{y}) = -2\theta + 4$ (in black).

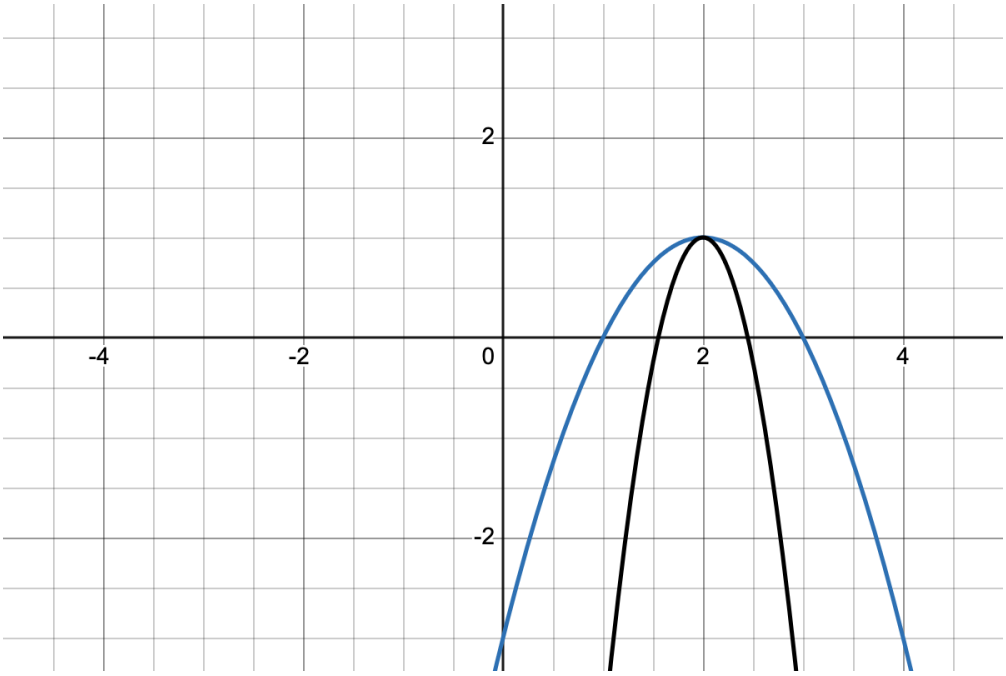

FIGURE 3: The graph for $\ell_1(\theta;\vec{y}) = -\theta^2 + 4\theta - 3$ (in blue) and $\ell_2(\theta;\vec{y}) = -5\theta^2 + 20\theta - 19$ (in black).

Definition 14 (Score function)**.** $s(\theta;\vec{y}) = \frac{\partial}{\partial\theta}\ell(\theta;\vec{y})$, where $\ell(\theta;\vec{y})$ is the log-likelihood.

• Equivalently, $s(\theta;\vec{y}) = \frac{1}{\mathcal{L}(\theta;\vec{y})}(\frac{\partial}{\partial\theta}\mathcal{L}(\theta;\vec{y}))$ by the chain rule.

• Previously, we've regarded $s(\theta;\vec{y})$ as a function of θ and set it to 0. E.g., during the MLE derivation, we fix $\vec{y}$ and ask which possible value for θ achieves maximum likelihood.

However, we can also consider θ as fixed at its true value θ^* and regard $s(\theta^*; \vec{Y})$ as a function of the random data $\vec{Y}$.

• For θ^* under regularity conditions, $\mathbb{E}[s(\theta^*; \vec{Y})] = 0$, where expectation is with respect to $\vec{Y}$.[3]

• For θ^* under regularity conditions, $\text{Var}[s(\theta^*; \vec{Y})] = \mathbb{E}[(s(\theta^*; \vec{Y}))^2] - (\mathbb{E}[s(\theta^*; \vec{Y})])^2 = \mathbb{E}[(s(\theta^*; \vec{Y}))^2]$ by definition of variance, where expectation and variance are with respect to $\vec{Y}$.

• As illustrated in Figure 2, θ^* should maximize likelihood, so at $\theta = \theta^*$, we expect $s(\theta; \vec{Y}) = \frac{\partial}{\partial\theta}\ell(\theta; \vec{Y})$ to be 0.

Definition 15 (Information equality)**.** For θ^* under regularity conditions, $\text{Var}[s(\theta^*; \vec{Y})] = -\mathbb{E}[\frac{\partial}{\partial\theta}s(\theta^*; \vec{Y})]$, where expectation and variance are with respect to $\vec{Y}$.[4]

Definition 16 (Fisher information)**.** $\mathcal{I}_{\vec{Y}}(\theta) = \mathbb{E}[(s(\theta; \vec{Y}))^2]$, where $\mathcal{I}_{\vec{Y}}(\theta)$ is the Fisher information for θ from $\vec{Y}$.

• For θ^* and under regularity conditions, $\mathcal{I}_{\vec{Y}}(\theta^*) = \text{Var}[s(\theta^*; \vec{Y})]$ by definition of variance.

• For θ^* and under regularity conditions, $\mathcal{I}_{\vec{Y}}(\theta^*) = -\mathbb{E}[\frac{\partial}{\partial\theta}s(\theta^*; \vec{Y})]$ by information equality.

• In this form, Fisher information is related to the second derivative of $\ell(\theta^*; \vec{y})$. As illustrated in Figure 3, a larger $\mathcal{I}_{\vec{Y}}(\theta^*)$ corresponds to a more concave-down $\ell(\theta^*; \vec{y})$ and thus more information about θ^*.

• Fisher information is additive, so for i.i.d. data, $\mathcal{I}_{\vec{Y}}(\theta) = n\mathcal{I}_{Y_1}(\theta)$, where $\mathcal{I}_{Y_1}(\theta)$ is the Fisher information for θ from Y_1.

• Fisher information is not invariant under reparameterization, but if $\tau = g(\theta)$ where g is a differentiable function with $g'(\theta) \neq 0$, then $\mathcal{I}_{\vec{Y}}(\tau) = \frac{\mathcal{I}_{\vec{Y}}(\theta)}{(g'(\theta))^2}$.

Definition 17 (Cramér–Rao lower bound)**.** For any unbiased estimator $\hat{\theta}_{\text{UB}}$ under regularity conditions, $\text{Var}[\hat{\theta}_{\text{UB}}] \geq \frac{1}{\mathcal{I}_{\vec{Y}}(\theta)}$.

• For some intuition, variance decreases as we gain more information about θ, which is illustrated in the inverse relationship.

Definition 18 (Maximum likelihood estimator)**.** $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta; \vec{Y})$. Under regularity conditions, the MLE has the following properties:

• MLE is invariant, meaning if $\hat{\theta}_{\text{MLE}}$ is the MLE of θ, then $g(\hat{\theta}_{\text{MLE}})$ is the MLE of $g(\theta)$.

• MLE is consistent, meaning $\hat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$.

• MLE is asymptotically unbiased, meaning $\lim_{n\to\infty} \text{Bias}[\hat{\theta}_{\text{MLE}}] = 0$.

• MLE is asymptotically efficient, meaning no other asymptotically unbiased estimator has a lower asymptotic variance.

• MLE is asymptotically Normal, meaning for i.i.d. $Y_1, \ldots, Y_n$, $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_{Y_1}(\theta^*)}\right)$.

• ☠: Notice the Fisher information in the Normal distribution is from Y_1, not $\vec{Y}$. For i.i.d. data, $\mathcal{I}_{\vec{Y}}(\theta^*) = n\mathcal{I}_{Y_1}(\theta^*)$. Keep the notation clear and consistent!

[3]The proof for this result, on pages 119 and 120 of the textbook, relies on DUThIS and evaluating at θ^*.
[4]The proof for this result, on pages 119 and 120 of the textbook, relies on DUThIS and evaluating at θ^*.

6 Practice Problems

Problem 1. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables, with both $\mathbb{E}[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma^2 < \infty$ unknown. Suppose we are interested in σ^2.[5]

(a) Find the method of moments estimator: $\hat{\sigma}^2_{\text{MOM}}$.

(b) Is $\hat{\sigma}^2_{\text{MOM}}$ a consistent estimator for σ^2?

(c) Find the asymptotic distribution of $\hat{\sigma}^2_{\text{MOM}}$. Assume $\text{Var}[(Y_i - \mu)^2] < \infty$.

(d) Show $\hat{\sigma}^2_{\text{MOM}} = \hat{\sigma}^2_{\text{MLE}}$ if $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown. Now suppose we're interested in standard deviation $\sigma = \sqrt{\text{Var}[Y_i]}$. Use the previous result to find the maximum likelihood estimator: $\hat{\sigma}_{\text{MLE}}$.

(e) Though MLE has nice properties, there is a major problem. Show $\hat{\sigma}^2_{\text{MLE}}$ is biased. However, show $\hat{\sigma}^2_{\text{MLE}}$ is asymptotically unbiased.[6]

Solution

First, let's find $\hat{\sigma}^2_{\text{MOM}}$ and show it's consistent for σ^2. We know $\sigma^2 = \text{Var}[Y_i] = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2]$ by definition of variance, so $\hat{\sigma}^2_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$ by replacing the theoretical quantities with their sample analogues. Now, $\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n}(\sum_{i=1}^n Y_i^2 - n(\bar{Y})^2)$ by the sum of squares identity with $c = 0$. This can be rewritten as $\frac{1}{n}\sum_{i=1}^n Y_i^2 - (\bar{Y})^2$ by distributing. By LLN, $\frac{1}{n}\sum_{i=1}^n Y_i^2 \xrightarrow{p} \mathbb{E}[Y_i^2]$. By LLN, $\bar{Y} \xrightarrow{p} \mathbb{E}[Y_i]$, so by CMT with $g(x) = x^2$, $(\bar{Y})^2 \xrightarrow{p} (\mathbb{E}[Y_i])^2$. By Theorem 3.5.7., $\frac{1}{n}\sum_{i=1}^n Y_i^2 - (\bar{Y})^2 \xrightarrow{p} \mathbb{E}[Y_i^2] - (\mathbb{E}[Y_i])^2$. We know $\mathbb{E}[Y_i^2] - (\mathbb{E}[Y_i])^2 = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] = \text{Var}[Y_i]$, so we conclude $\hat{\sigma}^2_{\text{MOM}} \xrightarrow{p} \sigma^2$.

Next, let's find its asymptotic distribution. We begin by trying to "pattern-match" to the CLT. First, $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \mu)^2 - n(\bar{Y} - \mu)^2$ by the sum of squares identity with $c = \mu$.

[5] Inspired by Problem 1 in "Section 3: Asymptotics, MLE, & Fisher Information" by Danielle Paulson.

[6] You don't have to show this, but it is good to know the ordinary least-squares estimator—i.e., $\hat{\sigma}^2_{\text{OLS}} = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$—is not only unbiased but also has lower MSE than $\hat{\sigma}^2_{\text{MLE}}$.

$$\begin{aligned}
\sqrt{n}\left(\hat{\sigma}^2_{\text{MOM}} - \sigma^2\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sigma^2\right) && \text{by substituting} \\
&= \sqrt{n}\left(\frac{1}{n}(\sum_{i=1}^{n}(Y_i - \mu)^2 - n(\bar{Y} - \mu)^2) - \sigma^2\right) && \text{by substituting} \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - (\bar{Y} - \mu)^2 - \sigma^2\right) && \text{by distributing} \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - \sigma^2 - (\bar{Y} - \mu)^2\right) && \text{by rearranging} \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - \sigma^2\right) - \sqrt{n}(\bar{Y} - \mu)^2 && \text{by distributing}
\end{aligned}$$

This is starting to look better! Recall in the CLT setup, we want something minus its expectation. Notice $(Y_i - \mu)^2$ is i.i.d. with $\mathbb{E}[(Y_i - \mu)^2] = \sigma^2$ by definition of variance. Thus, $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - \sigma^2\right) \xrightarrow{d} \mathcal{N}(0, \text{Var}[(Y_i - \mu)^2])$ by CLT.
As for the second term, notice $\sqrt{n}(\bar{Y} - \mu)^2 = \frac{1}{\sqrt{n}}(\sqrt{n}(\bar{Y} - \mu))^2$ by algebra. By CLT, $\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, so by CMT with $g(x) = x^2$, $(\sqrt{n}(\bar{Y} - \mu))^2 \xrightarrow{d} Z^2$, where $Z \sim \mathcal{N}(0, \sigma^2)$. Now, $\frac{1}{\sqrt{n}}$ is a "degenerate" random variable such that $\frac{1}{\sqrt{n}} \xrightarrow{p} 0$. Thus, by Slutsky's, $\sqrt{n}(\bar{Y} - \mu)^2 = \frac{1}{\sqrt{n}}(\sqrt{n}(\bar{Y} - \mu))^2 \xrightarrow{d} (0)Z^2 = 0$. Since this is convergence in distribution to a constant, we have $\sqrt{n}(\bar{Y} - \mu)^2 \xrightarrow{p} 0$.
Together, $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - \sigma^2\right) - \sqrt{n}(\bar{Y} - \mu)^2 \xrightarrow{d} X - 0$ by Slutsky's, where $X \sim \mathcal{N}(0, \text{Var}[(Y_i - \mu)^2])$. Notice $X - 0 \sim \mathcal{N}(0, \text{Var}[(Y_i - \mu)^2])$ by property of the Normal, so $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2 - \sigma^2\right) - \sqrt{n}(\bar{Y} - \mu)^2 \xrightarrow{d} \mathcal{N}(0, \text{Var}[(Y_i - \mu)^2])$. We conclude $\sqrt{n}\left(\hat{\sigma}^2_{\text{MOM}} - \sigma^2\right) \xrightarrow{d} \mathcal{N}(0, \text{Var}[(Y_i - \mu)^2])$.
Next, we want to show $\hat{\sigma}^2_{\text{MOM}} = \hat{\sigma}^2_{\text{MLE}}$ if $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Again, $\hat{\sigma}^2_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Let's find the maximum likelihood estimator.

$$
\begin{aligned}
\mathcal{L}(\mu,\sigma;\vec{Y}) &= f_{\vec{Y}}(\vec{Y};\mu,\sigma) && \text{by def. of likelihood}\\
&= \prod_{i=1}^{n} f_{Y_i}(Y_i;\mu,\sigma) && \text{by i.i.d.}\\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{Y_i-\mu}{\sigma})^2) && \text{by PDF}\\
&= \prod_{i=1}^{n} \frac{1}{\sigma} \exp(-\frac{1}{2}(\frac{Y_i-\mu}{\sigma})^2) && \text{by mult. constants}\\
&= \frac{1}{\sigma^n} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2) && \text{by product}\\
\ell(\mu,\sigma;\vec{Y}) &= \log\left(\frac{1}{\sigma^n} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2)\right) && \text{by def. of log-likelihood}\\
&= -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2 && \text{by log}\\
&= -n\log(\sigma) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2 + n(\bar{Y}-\mu)^2\right) && \text{by substituting}
\end{aligned}
$$

Notice as a function of μ, we're subtracting a constant by $\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2 + n(\bar{Y}-\mu)^2\right)$, and as a square, $(\bar{Y}-\mu)^2$ must be non-negative. To maximize, let $\mu = \bar{Y}$ such that we subtract by as little as possible. Thus, $\hat{\mu}_{\text{MLE}} = \bar{Y}$. To find $\hat{\sigma}_{\text{MLE}}$, notice $\max_{(\mu,\sigma)\in\mathbb{R}\times[0,\infty)} \ell(\mu,\sigma) = \max_{\sigma\in[0,\infty)}(\max_{\mu\in\mathbb{R}} \ell(\mu,\sigma))$.

$$
\begin{aligned}
\ell(\mu,\sigma;\vec{Y}) &= -n\log(\sigma) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2 + n(\bar{Y}-\mu)^2\right) && \text{from before}\\
\ell(\sigma;\vec{Y}) &= -n\log(\sigma) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right) && \text{by substituting}\\
\ell'(\sigma;\vec{Y}) &= \frac{-n}{\sigma} + \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sigma^3} && \text{by derivative}\\
0 &= \frac{-n}{\sigma} + \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sigma^3} && \text{by setting equal to 0}\\
\frac{n}{\sigma} &= \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sigma^3} && \text{by algebra}\\
\sigma &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2} && \text{by algebra}
\end{aligned}
$$

Thus, $\sigma^* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$ is a critical point. We want to show it is a global max to

conclude it is the MLE. Notice $\ell''(\sigma;\vec{Y}) = \frac{n}{\sigma^2} - \frac{3\sum_{i=1}^n (Y_i-\bar{Y})^2}{\sigma^4}$, and $\ell''(\sigma;\vec{Y}) < 0$ evaluated at $\sigma = \sigma^*$. Thus, σ^* is a local maximum. Since $\Theta = [0,\infty)$, notice $\ell(\sigma;\vec{Y}) \to -\infty$ as $\sigma \to 0^+$ and $\ell(\sigma;\vec{Y}) \to -\infty$ as $\sigma \to \infty$. Thus, $\sigma^* = \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2}$ is a global maximum, so $\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2}$ by definition of MLE, which means $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$ by invariance.
Finally, let's derive the bias.

$$
\begin{aligned}
\mathbb{E}[\hat{\sigma}^2_{\text{MLE}}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2\right] && \text{by substituting} \\
&= (\frac{1}{n})\mathbb{E}\left[\sum_{i=1}^n (Y_i - \bar{Y})^2\right] && \text{by linearity} \\
&= (\frac{1}{n})\mathbb{E}\left[\sum_{i=1}^n (Y_i^2) - n(\bar{Y})^2\right] && \text{by substituting} \\
&= (\frac{1}{n})\left(\sum_{i=1}^n (\mathbb{E}[Y_i^2]) - n\mathbb{E}[(\bar{Y})^2]\right) && \text{by linearity} \\
&= (\frac{1}{n})\left(\sum_{i=1}^n (\text{Var}[Y_i] + (\mathbb{E}[Y_i])^2) - n(\text{Var}[\bar{Y}] + (\mathbb{E}[\bar{Y}])^2)\right) && \text{by substituting} \\
&= (\frac{1}{n})\left(\sum_{i=1}^n (\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)\right) && \text{by substituting} \\
&= (\frac{1}{n})\left(n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)\right) && \text{by simplifying} \\
&= (\sigma^2 + \mu^2) - (\frac{\sigma^2}{n} + \mu^2) && \text{by simplifying} \\
&= \sigma^2(1 - \frac{1}{n}) && \text{by simplifying} \\
\text{Bias}[\hat{\sigma}^2_{\text{MLE}}] &= \mathbb{E}[\hat{\sigma}^2_{\text{MLE}}] - \sigma^2 && \text{by definition} \\
&= \sigma^2(1 - \frac{1}{n}) - \sigma^2 && \text{by substituting} \\
&= -\frac{\sigma^2}{n} && \text{by simplifying}
\end{aligned}
$$

Notice $\lim_{n\to\infty} \text{Bias}[\hat{\sigma}^2_{\text{MLE}}] = 0$. We conclude $\hat{\sigma}^2_{\text{MLE}}$ is biased but asymptotically unbiased.

Problem 2. For n i.i.d. random variables $X_1, \ldots, X_n$ with finite expectation $\mathbb{E}[X_1] = \mu$, finite variance $\text{Var}[X_1] = \sigma^2$, and sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{X}_n \xrightarrow{p} \mu$ by the Weak Law of Large Numbers.

(a) **Challenge**: Prove this result using Chebyshev's Inequality: For any random variable X, $P(|X - \mathbb{E}[X]| \geq c) \leq \frac{\text{Var}[X]}{c^2}$ if $c > 0$.

Hint: A similar proof is provided in "Section 2: Estimators."

Solution

First, $\bar{X}_n \xrightarrow{p} \mu \iff \forall\, \varepsilon > 0$, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$ by definition of convergence in probability, so let's prove the limit.

Let $\varepsilon > 0$ be any positive number. Recall $\bar{X}_n$ is a random variable with $\mathbb{E}[\bar{X}_n] = \mu$ and $\text{Var}[\bar{X}_n] = \frac{\sigma^2}{n}$. Thus, $P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2}$ by Chebyshev's Inequality. Additionally, since the event $|\bar{X}_n - \mu| \geq \varepsilon$ implies the event $|\bar{X}_n - \mu| > \varepsilon$, $P(|\bar{X}_n - \mu| > \varepsilon) \leq P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2}$. We can rewrite all this as $P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2/n}{\varepsilon^2}$.

Now, $\lim_{n\to\infty} \frac{\sigma^2/n}{\varepsilon^2} = 0$, and $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2/n}{\varepsilon^2}$ by applying the limit to both sides, so $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) \leq 0$ by substituting. Since probability cannot be negative, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$.

Notice ε was arbitrary. We conclude $\forall\, \varepsilon > 0$, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$, so $\bar{X}_n \xrightarrow{p} \mu$.