# DCP 1206: Probability

# Lecture 25 — Central Limit Theorem and Parameter Estimation

Ping-Chun Hsieh

December 18, 2019

# Announcements

- HW6-Part I is now on E3 (Due: 12/27 in class)
- Final Exam on 1/8 (Wednesday)
  - 10:10am - 12pm
  - Coverage: Lec 1 - Lec 29 (focus on Lec 14-29)
  - You are allowed to bring a cheat sheet (A4 size, 2-sided)

# This Lecture

1. Central Limit Theorem (CLT)

2. Parameter Estimation: MLE and MAP

- Reading material: Chapter 11.5

# 1. Central Limit Theorem (CLT)

# Beyond SLLN

- **The Strong Law of Large Numbers**: Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean $\mu$. Define $S_n = (X_1 + \cdots + X_n)$. Then, we have

$$P\left(\left\{\omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu\right\}\right) = 1$$

- Question: What does SLLN say about $S_n(\omega)$?

For almost every $\omega$, we have $\lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu \Rightarrow S_n(\omega) \simeq n\cdot\mu$ for large $n$

- Question: Do we have $S_n(\omega) \overset{\lim}{=} n\mu + o(n)$? $\Rightarrow \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \lim_{n\to\infty} \mu + \frac{o(n)}{n} = \mu$

Sublinear terms ??

# Review (Lecture 11): Binomial and Normal

- Example: $X_1, X_2, \cdots$ are i.i.d. Bernoulli r.v.s with mean $\mu$ and variance $\sigma^2 = \mu(1-\mu)$
  - Define $S_n = X_1 + X_2 + \cdots + X_n$
  - Question: What type of r.v. is $S_n$? $E[S_n]$ = ? Var$[S_n]$ = ?

Binomial $\quad n\mu \quad n\mu(1-\mu) = n\cdot\sigma^2$

- Question: How to find the distribution of $\dfrac{S_n - n\mu}{\sigma\sqrt{n}}$?

① make the mean = 0 by shifting

② variance $= \dfrac{n\sigma^2}{(\sigma\sqrt{n})^2} = 1$

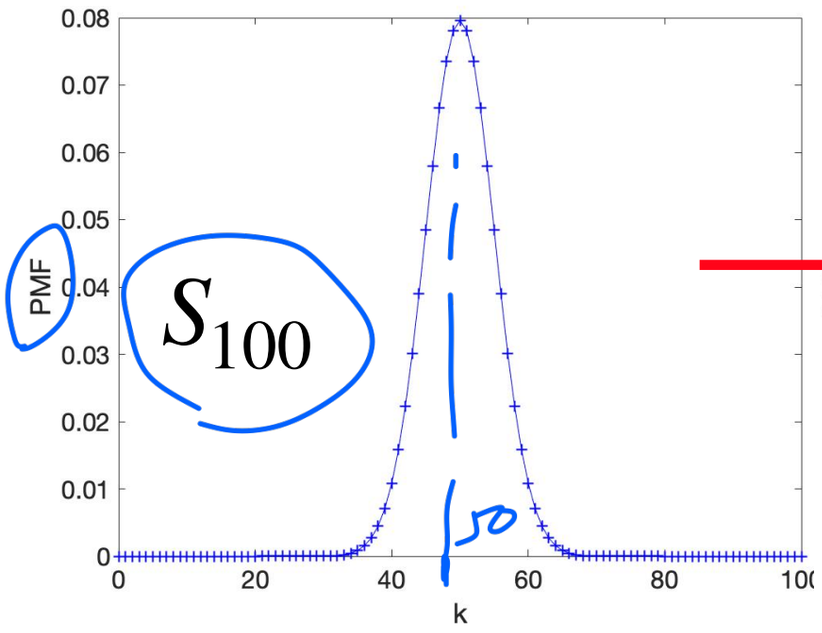
# Plotting $Z_n = (S_n - n\mu)/(\sigma\sqrt{n})$

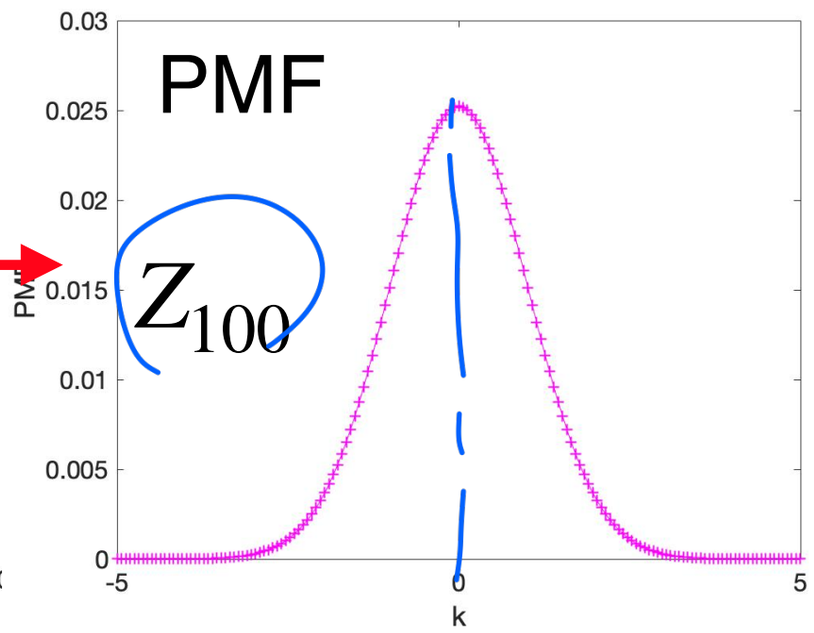
(shifted, normalized version of $S_n$)

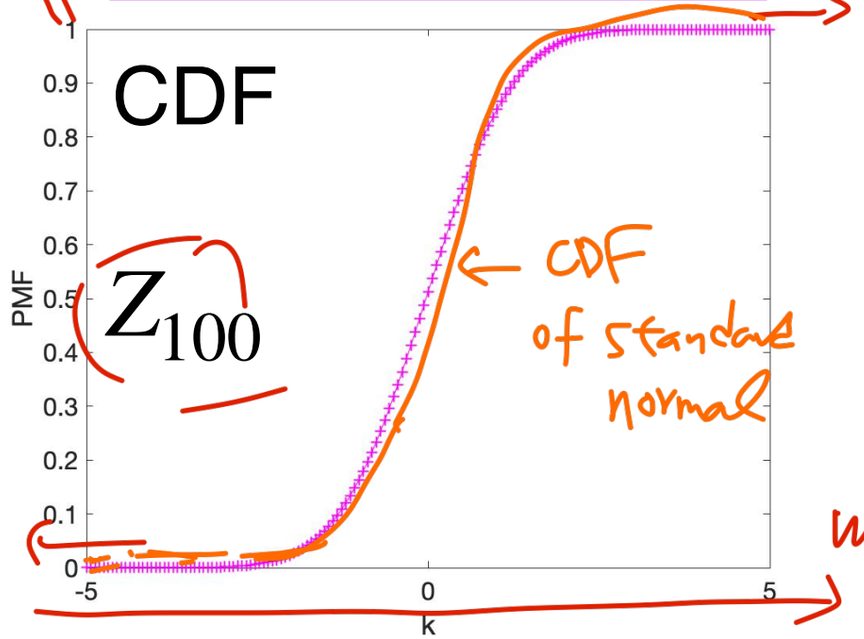
- Example: $\mu = 0.5$

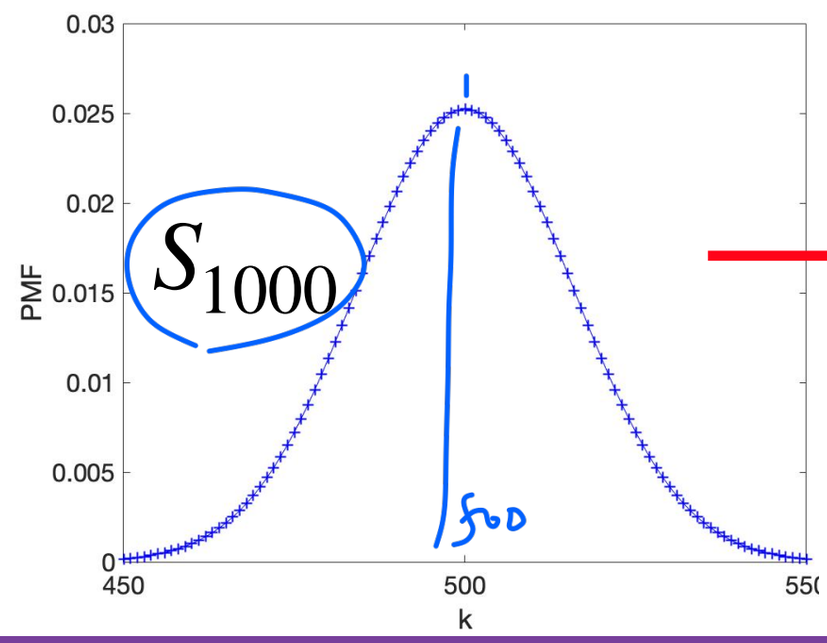
$P(Z_{100} \le w)$

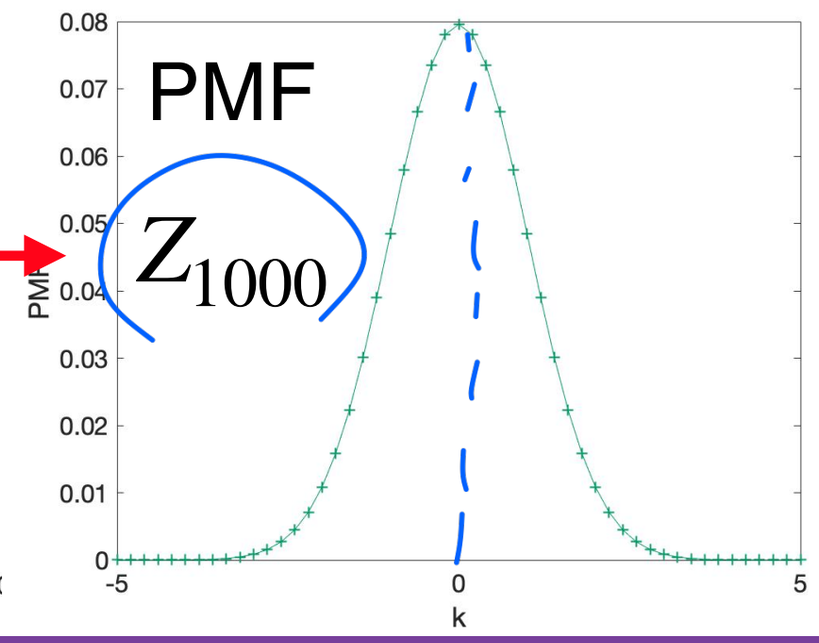
**≈ CDF of a standard normal!**

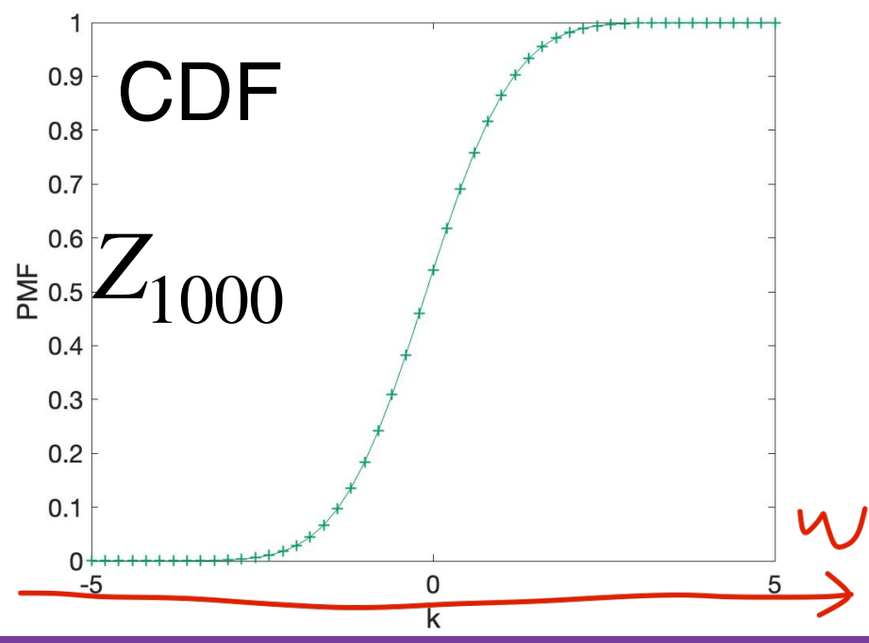


# Central Limit Theorem (Formally)

- **Central Limit Theorem (CLT)**: Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean $\mu$ and variance $\sigma^2$. Define $S_n = (X_1 + \cdots + X_n)$. Then, we have

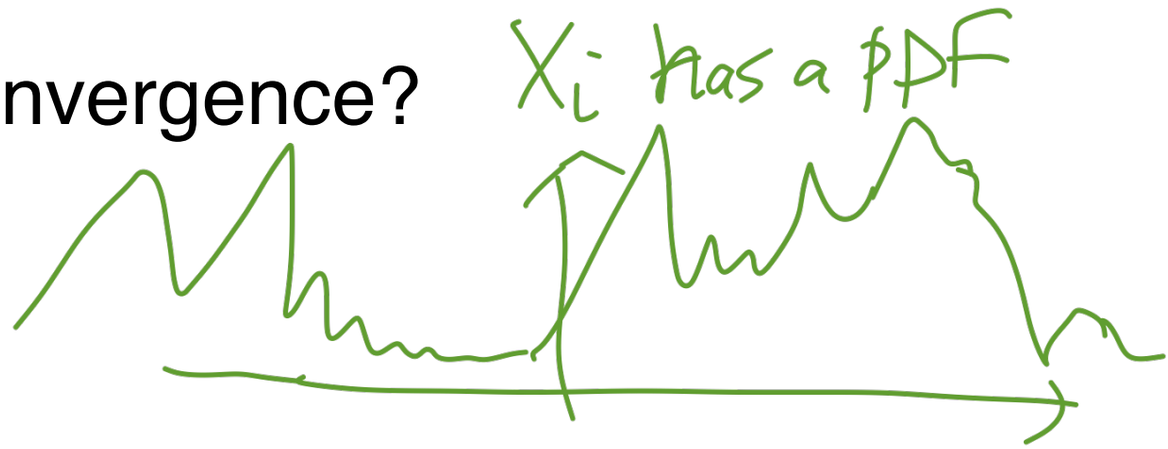
$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) = \Phi(z)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{v^2}{2}} dv$$

$\Phi(z)$ = the CDF of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$

where $\Phi(z)$ is the CDF of standard normal

- Question: How to interpret such convergence?



# How to Interpret CLT?

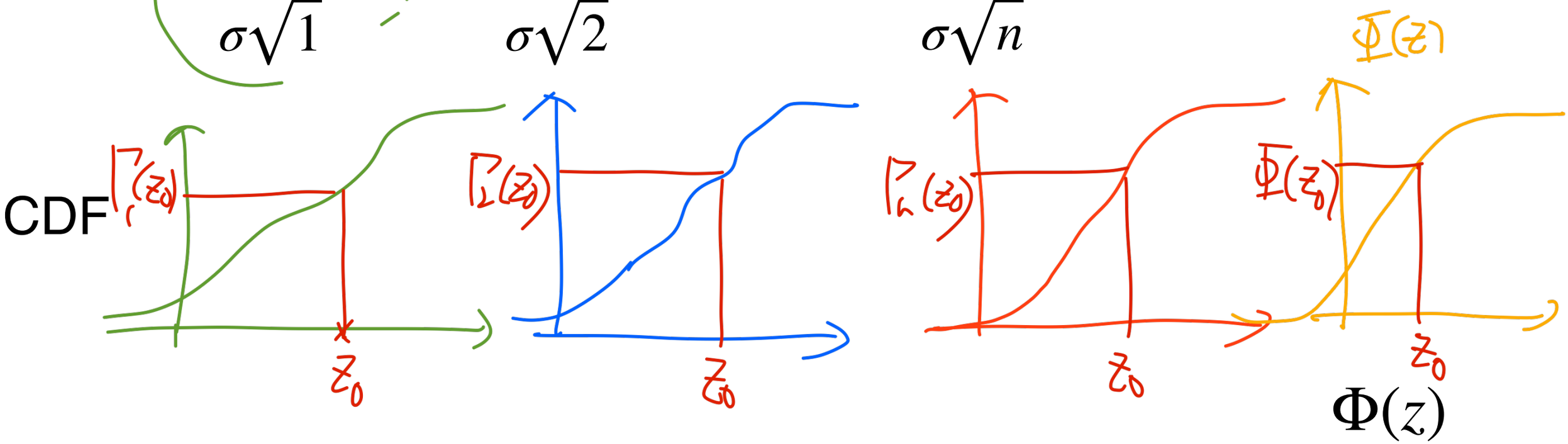
(Convergence in distribution)

- Let $X_1, X_2, \cdots$ be i.i.d. random variables with mean $\mu$ and variance $\sigma^2$
- Define $S_n = (X_1 + X_2 \cdots + X_n)$
- CLT: $\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) = \Phi(z)$

$F_n(z_0) \xrightarrow{?} \Phi(z_0)$ for every $z_0$.

$$\frac{S_1 - 1\cdot\mu}{\sigma\sqrt{1}} \quad \frac{S_2 - 2\cdot\mu}{\sigma\sqrt{2}} \quad \cdots \quad \frac{S_n - n\cdot\mu}{\sigma\sqrt{n}} \quad \cdots$$

(n=1) (n=2)



# Why is CLT Useful? Approximation!

- Recall that $S_n = (X_1 + X_2 \cdots + X_n)$
- CLT: $\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) = \Phi(z)$
- Idea: For large $n$, consider $P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) \approx \Phi(z)$ to find $P(S_n \le c)$ for any $c$

$$P(S_n \le c) = P(S_n - n\mu \le c - n\mu)$$

$$= P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{c - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{c - n\mu}{\sigma\sqrt{n}}\right)$$

$z$

# Example: Approximation via CLT

- Example: $X_1, \cdots, X_{20}$ are 20 i.i.d. continuous uniform r.v.s on (0,1)
  - Question: Find $P(\sum_{i=1}^{20} X_i \leq 8)$ using approximation?
  - Hint: $P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) \approx \Phi(z)$

$\mu = \frac{1}{2}$

$\sigma^2 = \frac{(1-0)^2}{12} = \frac{1}{12}$

$S_{20} = X_1 + X_2 + \cdots + X_{20}$

$$P(S_{20} \leq 8) = P\left(\frac{S_{20} - 20 \cdot \frac{1}{2}}{\sqrt{\frac{1}{12}} \cdot \sqrt{20}} \leq \frac{8 - 20 \cdot \frac{1}{2}}{\sqrt{\frac{1}{12}} \cdot \sqrt{20}}\right) = -2 \quad \sqrt{\frac{5}{3}}$$

$$\approx \Phi\left(\frac{-2}{\sqrt{\frac{5}{3}}}\right).$$

# Now let's prove CLT!

# Review: From MGF to Distributions

- Recall: Lecture 20 and HW5, Problem 2

- **MGF Uniqueness Theorem**: Let $X_1$ and $X_2$ be two random variables with MGFs $M_{X_1}(t)$ and $M_{X_2}(t)$, respectively. If $M_{X_1}(t) = M_{X_2}(t)$ for all $t$ in some interval $(-\alpha, \alpha)$, then $X_1$ and $X_2$ follow the same distribution, i.e.

$$P(X_1 \leq u) = P(X_2 \leq u), \text{ for all } u \in \mathbb{R}$$

**Problem 2 (Use MGF to find distributions)** (6+6+6=18 points)

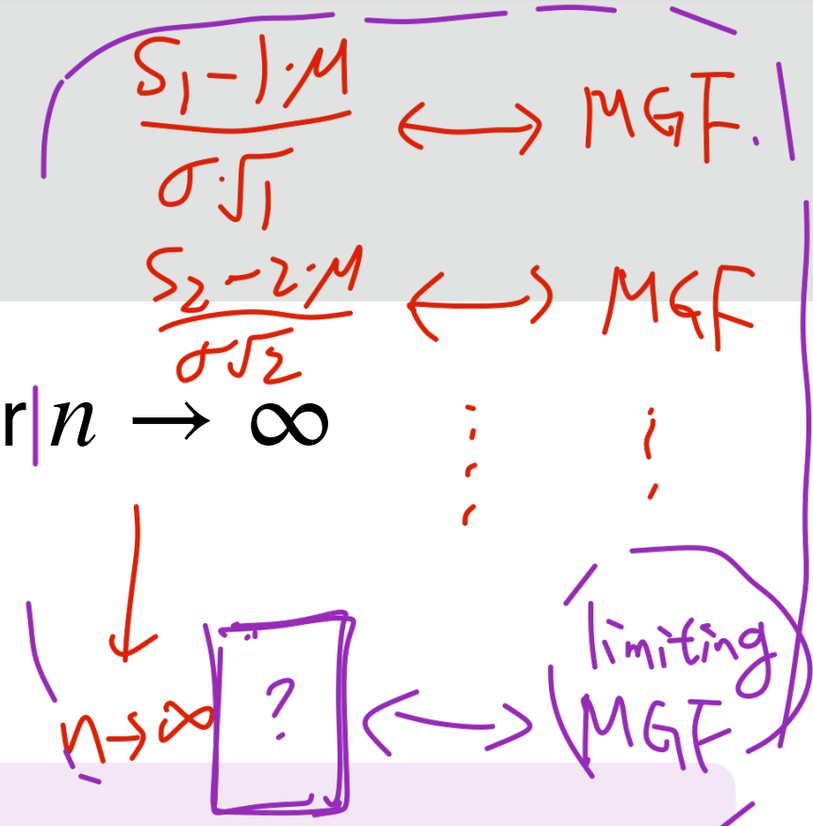
In each of the following cases, $M_X(t)$, the moment generating functions of $X$, is given. Please determine the distribution of $X$. (You could use the MGF table in the slides or the one in the textbook)

**(a)** $M_X(t) = \left(\frac{1}{4}e^t + \frac{3}{4}\right)^7$.

**(b)** $M_X(t) = e^t/(2 - e^t)$.

**(c)** $M_X(t) = \exp\left[3(e^t - 1)\right]$.

# Use MGF to Show CLT




- Idea: Suppose we find the MGF of $\dfrac{S_n - n\mu}{\sigma\sqrt{n}}$ for $n \to \infty$
  - Question: Can we find its distribution?

- **Levy Continuity Theorem**: Let $V_1, V_2 \cdots$ be a sequence of random variables with CDFs $F_1, F_2, \cdots$ and MGFs $M_{V_1}(t), M_{V_2}(t) \cdots$. Let $V$ be a random variable with CDF $F$ and MGF $M_V(t)$. If for every $t \in \mathbb{R}$, $\lim_{n\to\infty} M_{V_n}(t) = M_V(t)$, then the CDFs $F_n$ converge to $F$.

- Remark: MGF of $\mathcal{N}(\mu, \sigma^2)$ is $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ → $\mathcal{N}(0,1)$ MGF $= e^{\frac{t^2}{2}}$

# Use MGF to Show CLT (Cont.)

- Example: $X_1, X_2, \cdots$ are i.i.d. r.v.s with mean $\mu$ and variance $\sigma^2$
  - Define $S_n = X_1 + X_2 + \cdots + X_n$ and $Y_i = X_i - \mu$
  - Question: $E[Y_i]$ = 0 ? Var$[Y_i]$ = $\sigma^2$ ?
  - Question: What is the MGF of $\dfrac{S_n - n\mu}{\sigma\sqrt{n}}$ (in terms of MGF of $Y_i$)?

MGF of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$

$$= E\left[\exp\left(t \cdot \frac{S_n - n\mu}{\sigma\sqrt{n}}\right)\right] = E\left[\exp\left(t \cdot \frac{Y_1 + \cdots + Y_n}{\sigma\sqrt{n}}\right)\right]$$

$$= E\left[\exp\left(t \cdot \frac{Y_1}{\sigma\sqrt{n}}\right)\right] \cdot E\left[\exp\left(t \cdot \frac{Y_2}{\sigma\sqrt{n}}\right)\right] \cdots E\left[\exp\left(t \cdot \frac{Y_n}{\sigma\sqrt{n}}\right)\right]$$

$$= \left(E\left[\exp\left(t \cdot \frac{Y_1}{\sigma\sqrt{n}}\right)\right]\right)^n$$

$$\approx \left(M\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$$

# Use MGF to Show CLT (Cont.)

$$M(0) = E[e^{tY_i}]\Big|_{t=0} = 1$$

- Question: When $n \to \infty$, what is the MGF of $\dfrac{S_n - n\mu}{\sigma\sqrt{n}}$?

$$\lim_{n\to\infty} \left(M\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n = e^{\frac{t^2}{2}}$$

$$k = \frac{t}{\sigma\sqrt{n}} \iff n = \frac{t^2}{\sigma^2 k^2}$$

(Equivalently):

$$\lim_{n\to\infty} n \cdot \ln\left(M\left(\frac{t}{\sigma\sqrt{n}}\right)\right) = \lim_{k\to 0} \frac{t^2}{\sigma^2 k^2} \cdot \ln M(k)$$

$$= \frac{t^2}{\sigma^2} \lim_{k\to 0} \frac{\ln M(k)}{k^2} = \frac{t^2}{\sigma^2} \lim_{k\to 0} \frac{M'(k)}{2k \cdot M(k)}$$

$$= \frac{t^2}{\sigma^2} \lim_{k\to 0} \frac{M''(k) \to \sigma^2}{2 \cdot M(k) + 2k \cdot M'(k)} = \frac{t^2}{\sigma^2} \cdot \frac{\sigma^2}{2}$$

($M(k) \to 1$, $k \to 0$)

$$= \frac{t^2}{2}$$

L'Hospital rule:

$$\lim \frac{f(x)}{g(x)} = \lim \frac{f'(x)}{g'(x)}$$

# 2. Parameter Estimation From Data / Sampling

# How to Find a Good Estimator Under Small $n$?

- 2 possible outcomes: Yes / No-Laughing
- $p = P(\text{outcome is "Yes"})$ is unknown
- Each toss is independent from other tosses

- Question: Suppose we observe "Yes, No-L, Yes, Yes, No-L"
  - How to estimate $p$ in a principled way?
  - Is "sample mean" a good estimator?

Maximum Likelihood Estimation

# Example: Likelihood of Bernoulli Experiments

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. Bernoulli random variables with mean $p_*$
  - $P(X_1 = 1, X_2 = 0, X_3 = 0) = ?$ $p_* \cdot (1-p_*) \cdot (1-p_*)$
  - How about $P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$? $x_i \in \{0,1\}$

$$= \underbrace{(1-p_*)^{1-x_1} \cdot p_*^{x_1}}_{\text{Success}} \underbrace{(1-p_*)^{1-x_2} \cdot p_*^{x_2}}_{\text{Failure}} \cdots \underbrace{(1-p_*)^{1-x_n} \cdot p_*^{x_n}}$$

$$= p_*^{(x_1+x_2+\cdots+x_n)} \cdot (1-p_*)^{(n-(x_1+x_2+\cdots+x_n))}$$

"Likelihood"

# Example: Maximum Likelihood for Bernoulli

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. Bernoulli random variables with unknown mean
  - Suppose we observe $X_1 = 1, X_2 = 0, X_3 = 0$
  - Question: Under a guess $\theta$, $P(X_1 = 1, X_2 = 0, X_3 = 0; \theta)$ = ?
  - Question: Under what $\theta$ is $P(X_1 = 1, X_2 = 0, X_3 = 0; \theta)$ maximized?

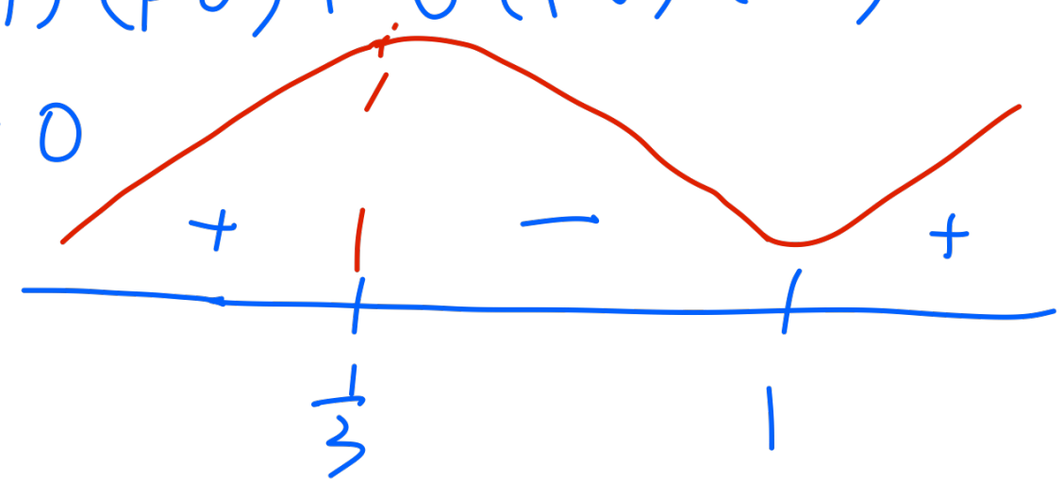
$$P(X_1=1, X_2=0, X_3=0; \theta) = \theta \cdot (1-\theta) \cdot (1-\theta) \triangleq f(\theta)$$

Maximize likelihood:

$$\frac{df(\theta)}{d\theta} = (1-\theta)(1-\theta) + \theta \cdot (-1) \cdot (1-\theta) + \theta \cdot (1-\theta) \cdot (-1)$$

$$= (1-\theta) \cdot (1-3\theta) = 0$$

$\theta = \frac{1}{3}$ or $\theta = 1$



# Example: Maximum Likelihood for Bernoulli

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. Bernoulli random variables with unknown mean
  - Suppose we observe $X_1 = x_1, \cdots, X_n = x_n \ \ (x_i \in \{0,1\})$
  - Question: Under a guess of mean = $\theta$, $P(X_1 = x_1, \cdots, X_n = x_n; \theta)$ = ?
  - Question: Under what $\theta$ is $P(X_1 = x_1, \cdots, X_n = x_n; \theta)$ maximized?

$P(X_1 = x_1, \cdots, X_n = x_n; \theta) = \theta^{x_1 + x_2 + \cdots + x_n} \cdot (1-\theta)^{n - (x_1 + x_2 + \cdots + x_n)}$

$= \theta^{\alpha} \cdot (1-\theta)^{n-\alpha} = g(\theta)$

Maximize Likelihood:

$\frac{d\, g(\theta)}{d\theta} = \cdots$

$\text{maximizer} = \frac{\alpha}{n}$ sample mean

# Maximum Likelihood Estimation (Formally)

- **Maximum Likelihood Estimation (MLE)**: Given observed data $D$, choose $\theta$ that maximizes the probability of observed data:

$$\theta_{\text{MLE}} = \arg\max_{\theta \in \Theta} P(D; \theta) = \arg\max_{\theta \in \Theta} \log P(D; \theta)$$

- Question: What if $D = \{X_i\}_{i=1}^{N}$ and the data samples are independent?

$$\theta_{\text{MLE}} =$$

- Question: What if $D = \{X_i\}_{i=1}^{N}$ and the data samples are i.i.d. Bernoulli?

$$\theta_{\text{MLE}} =$$

# What about continuous random variables?

Use density for MLE!

# Example: MLE for Normal RVs

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. normal random variables with unknown mean and variance
  - Suppose we observe $X_1 = x_1, \cdots, X_n = x_n$
  - Question: Under a guess of mean = $\theta$, $p(X_1 = x_1, \cdots, X_n = x_n; \theta)$ = ?

# Example: MLE for Normal RVs (Cont.)

- Example: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. normal random variables with unknown mean and variance
  - Suppose we observe $X_1 = x_1, \cdots, X_n = x_n$
  - Question: Under what $\theta$ is $p(X_1 = x_1, \cdots, X_n = x_n; \theta)$ maximized?

# MLE for Normal Random Variables (Formally)

- **MLE for Normal** : Given observed data $D = \{X_i\}_{i=1}^n$ of i.i.d. normal random variables , the MLE estimators $\theta_{\text{MLE}} = (\mu_{\text{MLE}}, \sigma^2_{\text{MLE}})$ are:

$$\mu_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sigma^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu_{\text{MLE}})^2$$

# Why is MLE a Good Estimator?

- Question: What kind of property do we want for MLE?

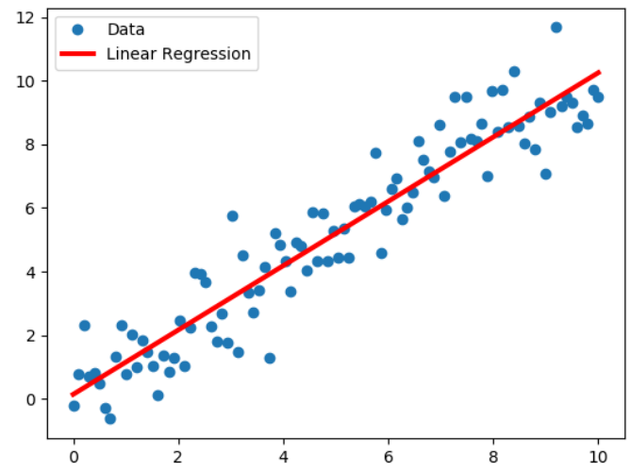
- **Consistency of MLE**: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. random variables with model parameters $\theta \in \Theta$ (where the model is identifiable and $\Theta$ is assumed to be finite). Then, MLE converges to the true $\theta$ in probability:

$$\theta_{\text{MLE}} \xrightarrow{\text{p}} \theta$$

- Remark: For the proof, please see the material on E3

# Application: MLE for Linear Regression

- MLE is widely used in machine learning problems.
- Example: Linear regression
  - $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. samples and assume $y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$
  - Question: Likelihood $p(\{y_1, \cdots, y_n\} \mid \{x_1, \cdots, x_n\}, \theta, \sigma)$ = ?



# MLE for Linear Regression (Cont.)

- Example: Linear regression
  - $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. samples and assume $y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$
  - Question: What is the MLE $\theta_{\text{MLE}}$?

# MLE for Linear Regression (Formally)

- **MLE for Linear Regression** : Given i.i.d. data samples $\{(x_i, y_i)\}_{i=1}^n$ with $y_i \sim \mathcal{N}(\theta^T x_i, \sigma^2)$, the MLE is

$$\theta_{\text{MLE}} = (X^T X)^{-1} X^T y$$

where $y = [y_1, y_2, \cdots, y_n]^T$

$X = [x_1, x_2, \cdots, x_n]^T$

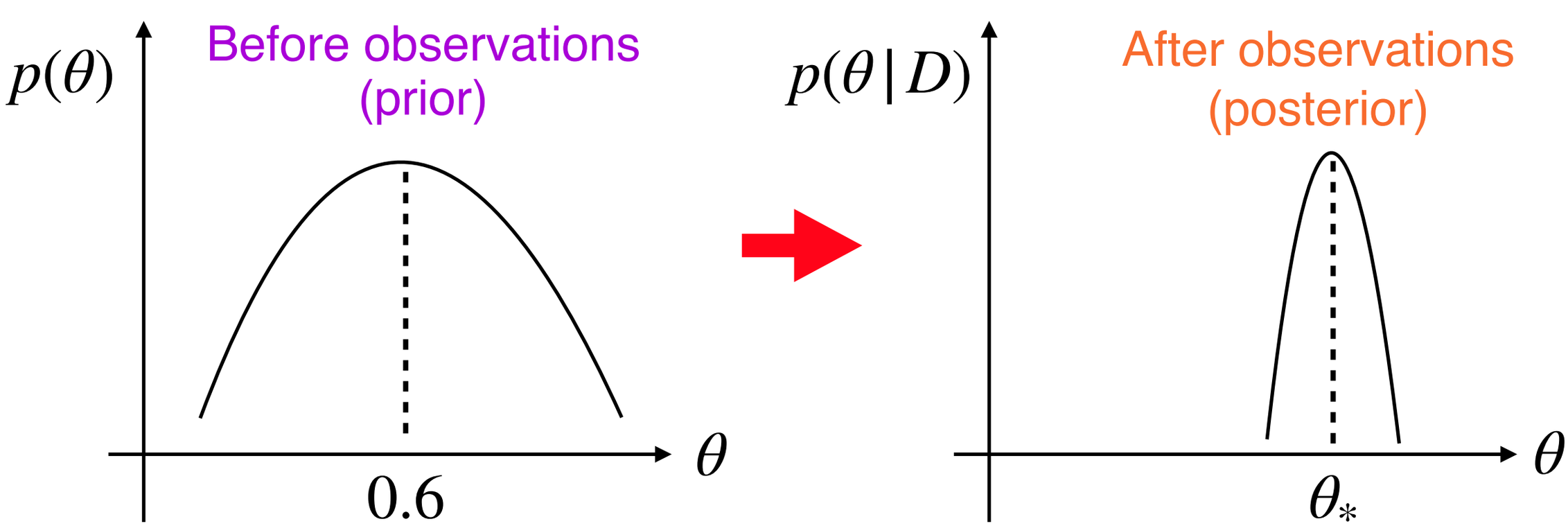
What if we have some prior knowledge about the possible value of $\theta$?

# Maximum a Posteriori Estimation

# What if We Have Some Prior Knowledge?

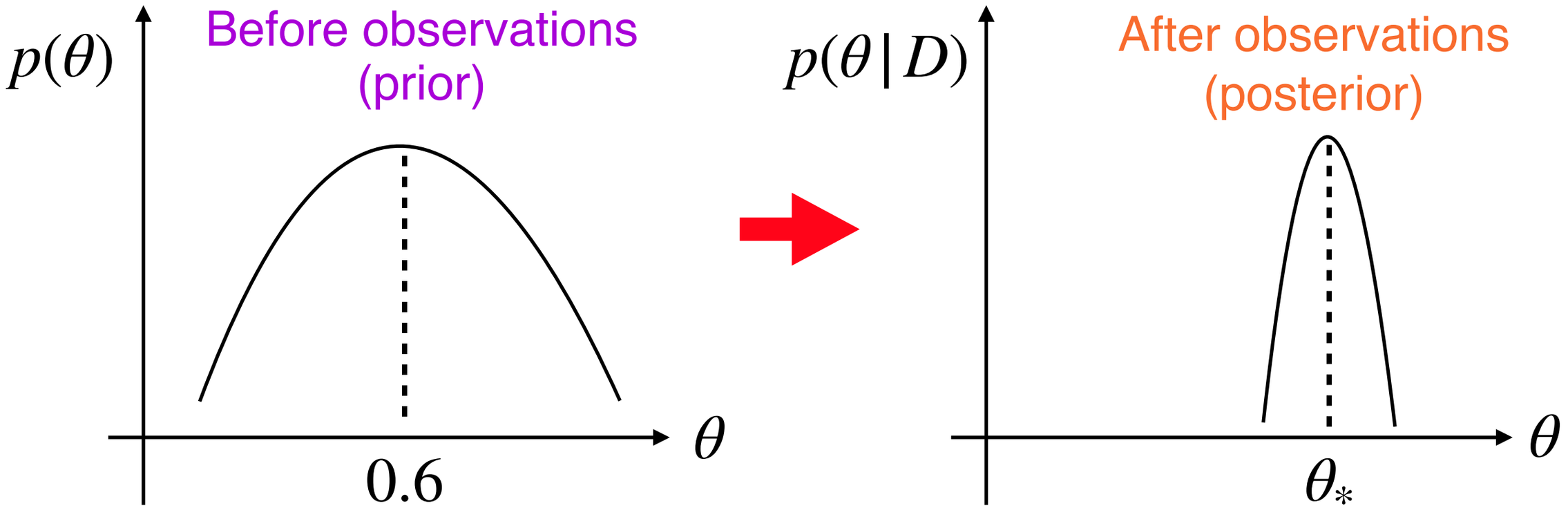
- 2 possible outcomes: Yes / No-Laughing
- $\theta_* = P(\text{outcome is “Yes”})$ is unknown
- Each toss is <u>independent</u> from other tosses

- Prior knowledge: Suppose we know $\theta_*$ is "close" to 0.6




- Idea: Instead of estimating a single $\theta$, obtain a <u>distribution</u> over possible values of $\theta$

# How to Obtain Posterior Distribution?




$$p(\theta \mid D) =$$

# Next Lecture

- Parameter Estimation
  - Bayesian approach: Maximum a Posteriori (MAP)

# 1-Minute Summary

## 1. Central Limit Theorem (CLT)

- CLT: $\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) = \Phi(z)$
- CDF approximation for large $n$
- Proof by MGF

## 2. Parameter Estimation: MLE and MAP

- MLE: $\theta_{\text{MLE}} = \arg\max_{\theta\in\Theta} P(D;\theta) = \arg\max_{\theta\in\Theta} \log P(D;\theta)$
- Bernoulli / Normal / Linear Regression