

# 18: Central Limit Theorem

---

Jerry Cain  
May 6, 2022

## Table of Contents

2	iid Random Variables
7	Central Limit Theorem
19	Sample Statistics
24	Exercises



# iid Random Variables

# Independence of multiple random variables

Review

We have independence of  $n$  discrete random variables  $X_1, X_2, \dots, X_n$  if for all  $x_1, x_2, \dots, x_n$ :

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i)$$

$$p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p_{X_i}(x_i)$$

We have independence of  $n$  continuous random variables  $X_1, X_2, \dots, X_n$  if for all  $x_1, x_2, \dots, x_n$ :

$$P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \prod_{i=1}^n P(X_i \leq x_i)$$

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

# i.i.d. random variables

---

Consider  $n$  variables  $X_1, X_2, \dots, X_n$ .

$X_1, X_2, \dots, X_n$  are **independent and identically distributed** if

- $X_1, X_2, \dots, X_n$  are independent, and
- All have the same PMF (if discrete) or PDF (if continuous).
  - $\Rightarrow E[X_i] = \mu$  for  $i = 1, \dots, n$
  - $\Rightarrow \text{Var}(X_i) = \sigma^2$  for  $i = 1, \dots, n$

Same thing:

**i.i.d.**

**iid**

**IID**

# Quick check

---

Are  $X_1, X_2, \dots, X_n$  i.i.d. with the following distributions?





1.  $X_i \sim \text{Exp}(\lambda)$ ,  $X_i$  independent
2.  $X_i \sim \text{Exp}(\lambda_i)$ ,  $X_i$  independent
3.  $X_i \sim \text{Exp}(\lambda)$ ,  $X_1 = X_2 = \dots = X_n$
4.  $X_i \sim \text{Bin}(n_i, p)$ ,  $X_i$  independent



# Quick check

---

Are  $X_1, X_2, \dots, X_n$  i.i.d. with the following distributions?

1.  $X_i \sim \text{Exp}(\lambda)$ ,  $X_i$  independent 
2.  $X_i \sim \text{Exp}(\lambda_i)$ ,  $X_i$  independent  (unless  $\lambda_i$  equal)
3.  $X_i \sim \text{Exp}(\lambda)$ ,  $X_1 = X_2 = \dots = X_n$   dependent:  $X_1 = X_2 = \dots = X_n$
4.  $X_i \sim \text{Bin}(n_i, p)$ ,  $X_i$  independent  (unless  $n_i$  equal)  
Note underlying Bernoulli RVs are i.i.d.!



# Central Limit Theorem

# Central Limit Theorem

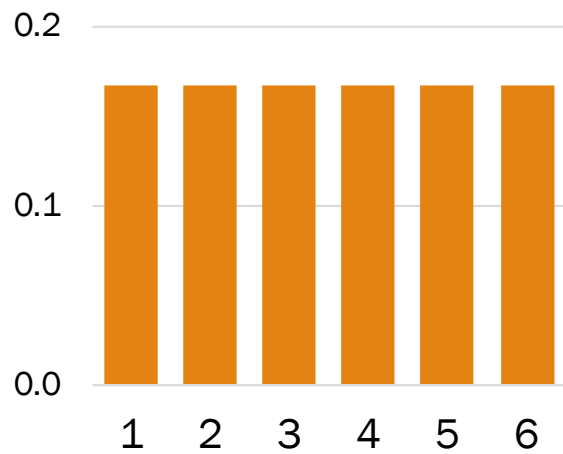
Consider  $n$  **independent and identically distributed (i.i.d.)** variables  $X_1, X_2, \dots, X_n$  with  $E[X_i] = \bar{\mu}$  and  $\text{Var}(X_i) = \bar{\sigma}^2$ .

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\bar{\mu}, n\bar{\sigma}^2) \quad \text{As } n \rightarrow \infty$$

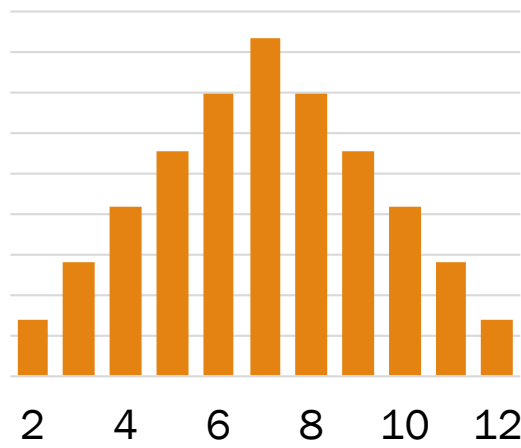
The sum of  $n$  **i.i.d.** random variables is normally distributed with mean  $n\bar{\mu}$  and variance  $n\bar{\sigma}^2$ .

# Sum of dice rolls

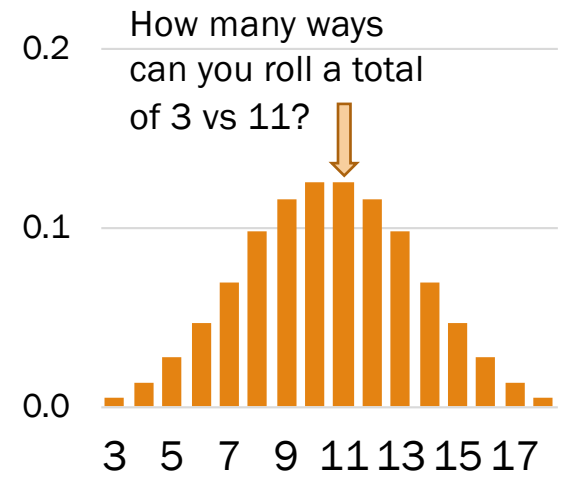
Roll  $n$  independent dice. Let  $X_i$  be the outcome of roll  $i$ .  $X_i$  are i.i.d.



$$\sum_{i=1}^1 X_i \quad \text{Sum of 1 die roll}$$



$$\sum_{i=1}^2 X_i \quad \text{Sum of 2 dice rolls}$$

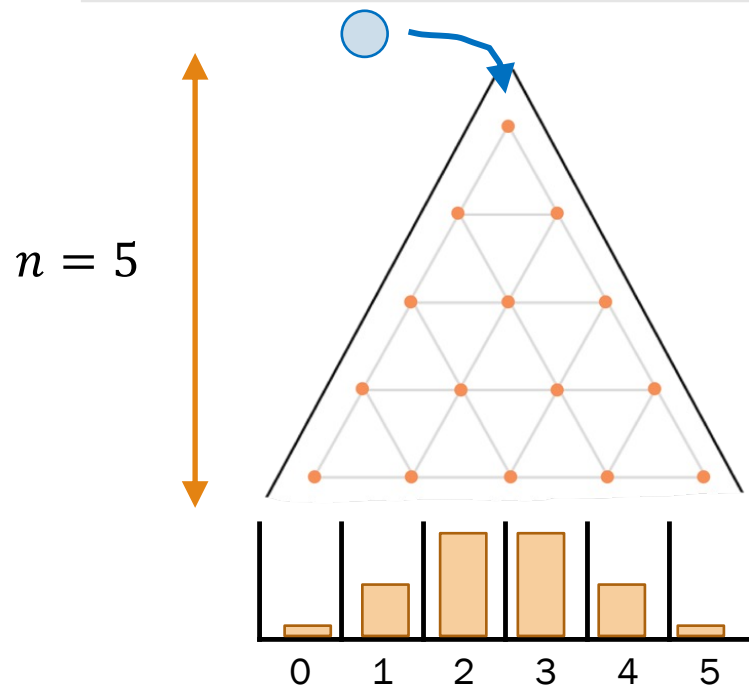


$$\sum_{i=1}^3 X_i \quad \text{Sum of 3 dice rolls}$$

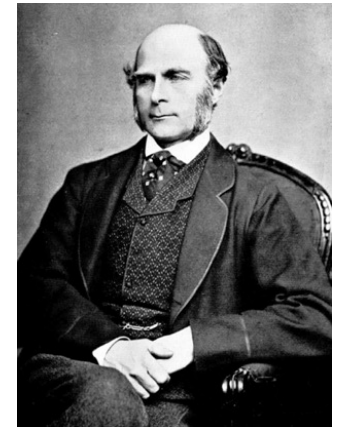
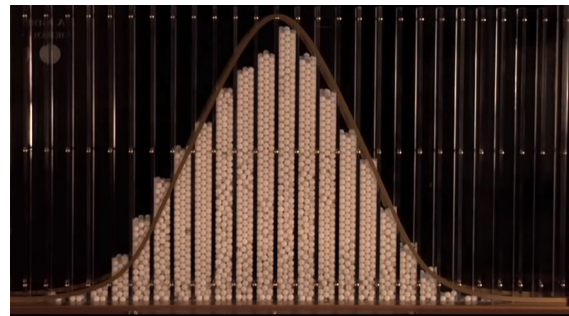
# CLT explains a lot

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \rightarrow \infty$$

The sum of  $n$  **i.i.d.** random variables is normally distributed with mean  $n\mu$  and variance  $n\sigma^2$ .



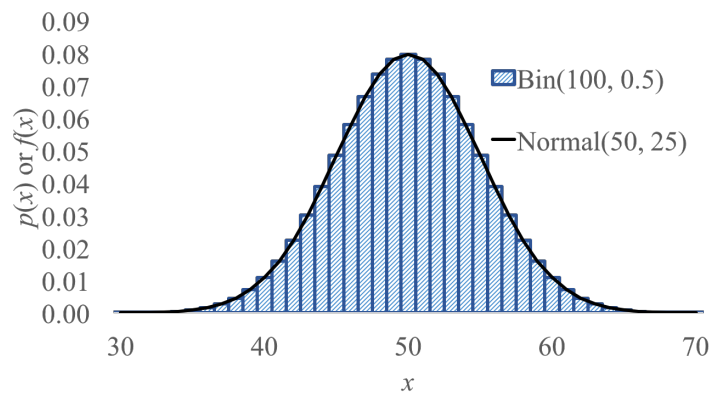
Galton Board, by Sir Francis Galton (1822-1911)



# CLT explains a lot

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \rightarrow \infty$$

The sum of  $n$  **i.i.d.** random variables is normally distributed with mean  $n\mu$  and variance  $n\sigma^2$ .



**Normal approximation of Binomial**  
Sum of i.i.d. Bernoulli RVs  $\approx$  Normal

Proof:

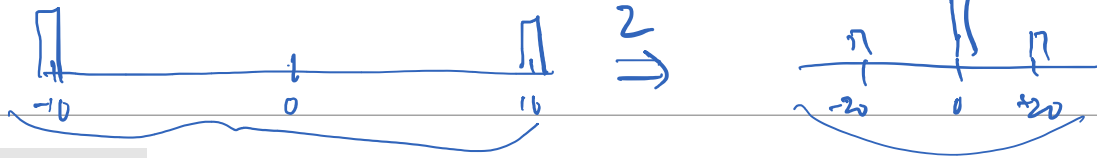
Let  $X_i \sim \text{Ber}(p)$  for  $i = 1, \dots, n$ , where  $X_i$  are i.i.d.  
 $E[X_i] = p$ ,  $\text{Var}(X_i) = p(1 - p)$

$$X = \sum_{i=1}^n X_i \quad (X \sim \text{Bin}(n, p))$$

$$X \sim \mathcal{N}(n\mu, n\sigma^2) \quad (\text{CLT, as } n \rightarrow \infty)$$

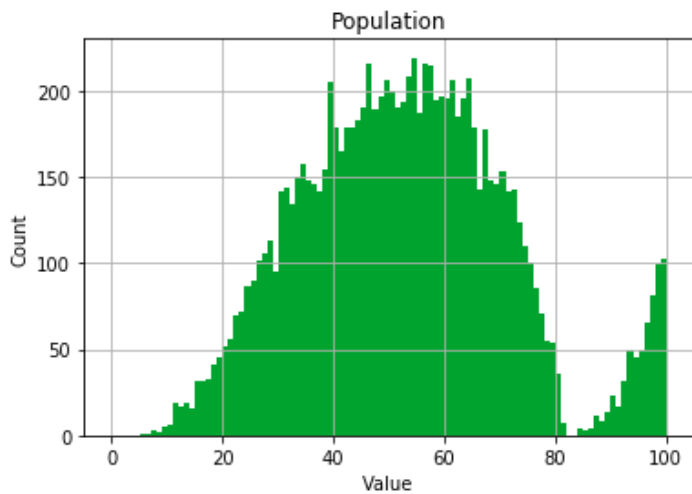
$$X \sim \mathcal{N}(\underline{np}, \underline{np(1 - p)}) \quad (\text{substitute mean, variance of Bernoulli})$$

# CLT explains a lot



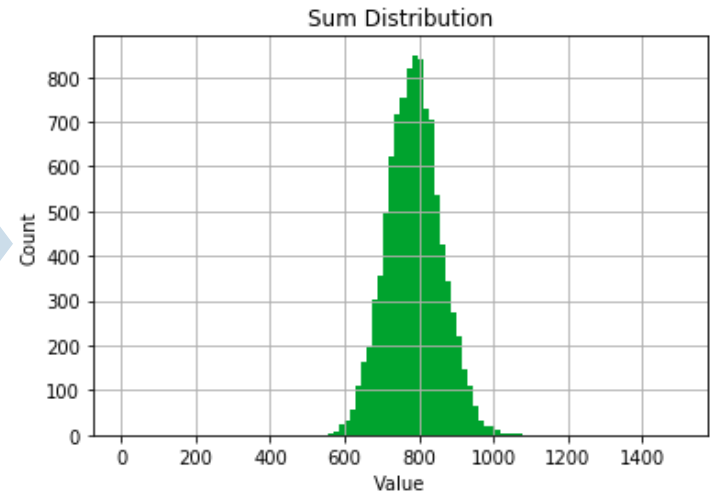
$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \rightarrow \infty$$

The sum of  $n$  **i.i.d.** random variables is normally distributed with mean  $n\mu$  and variance  $n\sigma^2$ .



Distribution of  $X_i$

Sample of  
size 15,  
**sum** values

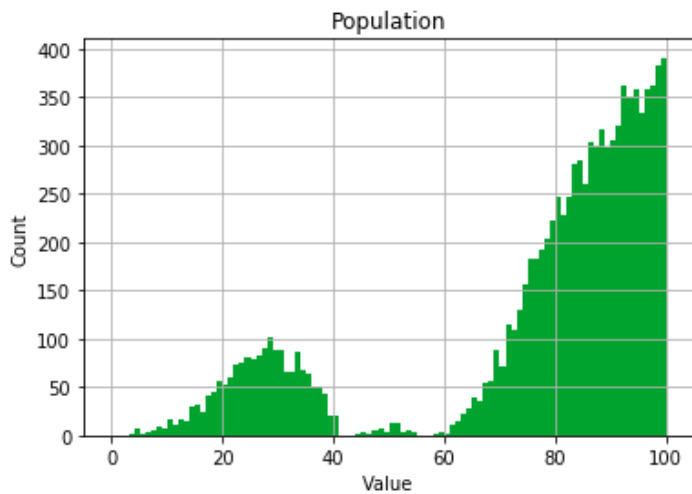


Distribution of  $\sum_{i=1}^{15} X_i$

# CLT explains a lot

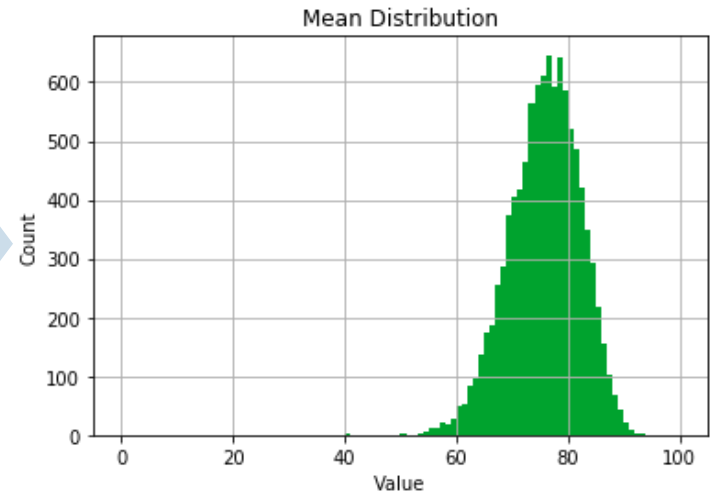
$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \rightarrow \infty$$

The sum of  $n$  **i.i.d.** random variables is normally distributed with mean  $n\mu$  and variance  $n\sigma^2$ .



Distribution of  $X_i$

Sample of  
size 15,  
**average** values



Distribution of  $\frac{1}{15} \sum_{i=1}^{15} X_i$

# Proof of CLT

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \rightarrow \infty$$

The sum of  $n$  **i.i.d.** random variables is normally distributed with mean  $n\mu$  and variance  $n\sigma^2$ .

Proof:

- The Fourier Transform of a PDF is called a **characteristic function**.
- Take the characteristic function of the probability mass of the sample distance from the mean, divided by standard deviation
- Show that this approaches an exponential function in the limit as  $n \rightarrow \infty$ :  $f(x) = e^{-\frac{x^2}{2}}$
- This function is in turn the characteristic function of the Standard Normal,  $Z \sim \mathcal{N}(0,1)$ .

(this proof is beyond the scope of CS109)

# Sum of $n$ independent Uniform RVs

Let  $X = \sum_{i=1}^n X_i$  be sum of i.i.d. RVs, where  $X_i \sim \text{Uni}(0,1)$ .  $\mu = E[X_i] = 1/2$   
 $\sigma^2 = \text{Var}(X_i) = 1/12$

For different  $n$ , how close is the CLT approximation of  $P(X \leq n/3)$ ?

$n = 2$ :

Exact

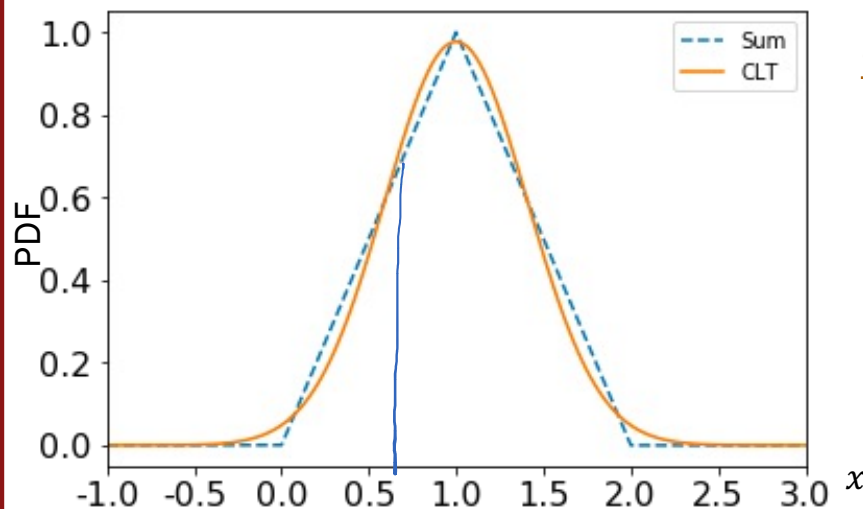
$$P(X \leq 2/3) \approx 0.2222$$

CLT approximation

$$X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow Y \sim \mathcal{N}(1, 1/6)$$

$$P(X \leq 2/3) \approx P(Y \leq 2/3)$$

$$= \Phi\left(\frac{2/3 - 1}{\sqrt{1/6}}\right) \approx 0.2071$$



# Sum of $n$ independent Uniform RVs

Let  $X = \sum_{i=1}^n X_i$  be sum of i.i.d. RVs, where  $X_i \sim \text{Uni}(0,1)$ .  $\mu = E[X_i] = 1/2$   
 $\sigma^2 = \text{Var}(X_i) = 1/12$

For different  $n$ , how close is the CLT approximation of  $P(X \leq n/3)$ ?

$n = 5$ :

Exact

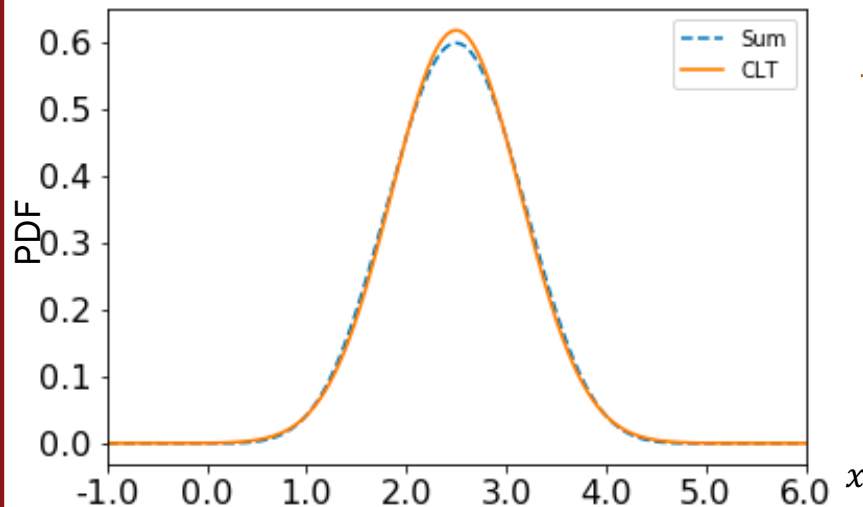
$$P(X \leq 5/3) \approx 0.1017$$

CLT approximation

$$X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow Y \sim \mathcal{N}(5/2, 5/12)$$

$$P(X \leq 5/3) \approx P(Y \leq 5/3)$$

$$= \Phi\left(\frac{5/3 - 5/2}{\sqrt{5/12}}\right) \approx 0.0984$$



# Sum of $n$ independent Uniform RVs

Let  $X = \sum_{i=1}^n X_i$  be sum of i.i.d. RVs, where  $X_i \sim \text{Uni}(0,1)$ .  $\mu = E[X_i] = 1/2$   
 $\sigma^2 = \text{Var}(X_i) = 1/12$

For different  $n$ , how close is the CLT approximation of  $P(X \leq n/3)$ ?

$n = 10$ :

Exact

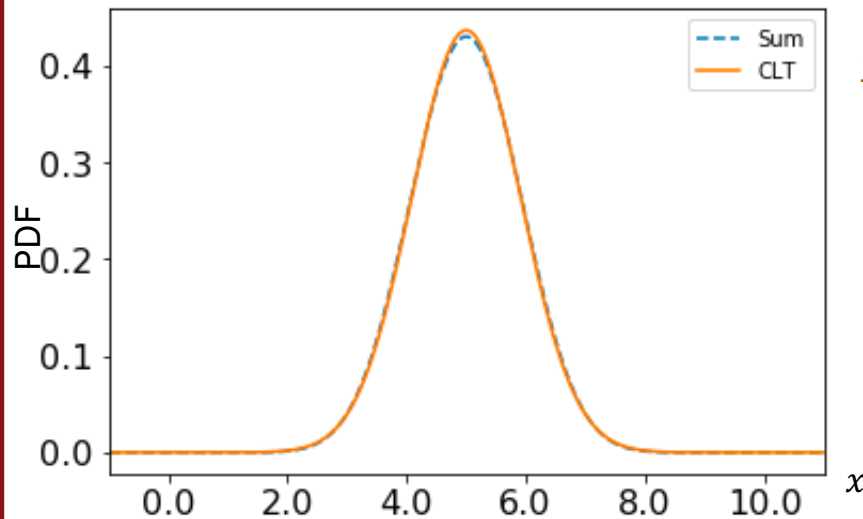
$$P(X \leq 10/3) \approx 0.0337$$

CLT approximation

$$X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2) \Rightarrow Y \sim \mathcal{N}(5, 5/6)$$

$$P(X \leq 10/3) \approx P(Y \leq 10/3)$$

$$= \Phi\left(\frac{10/3 - 5}{\sqrt{5/6}}\right) \approx 0.0339$$

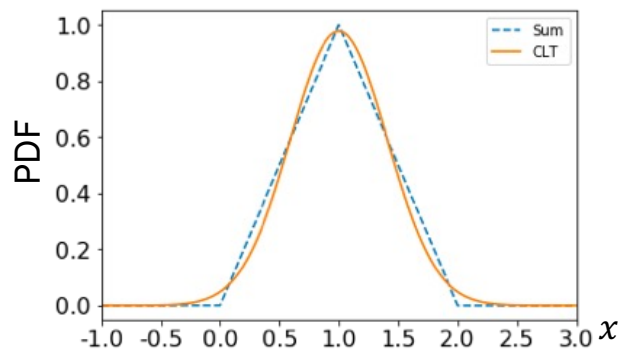


# Sum of $n$ independent Uniform RVs

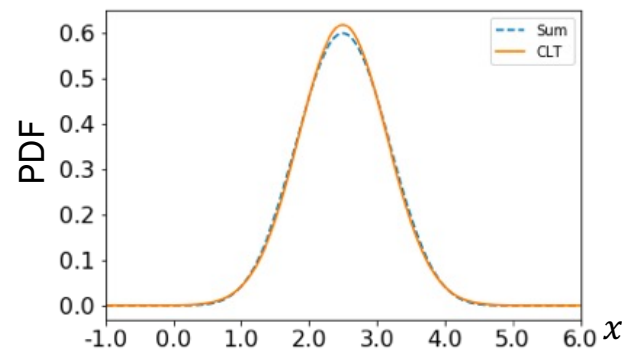
Let  $X = \sum_{i=1}^n X_i$  be sum of i.i.d. RVs, where  $X_i \sim \text{Uni}(0,1)$ .  $\mu = E[X_i] = 1/2$   
 $\sigma^2 = \text{Var}(X_i) = 1/12$

For different  $n$ , how close is the CLT approximation of  $P(X \leq n/3)$ ?

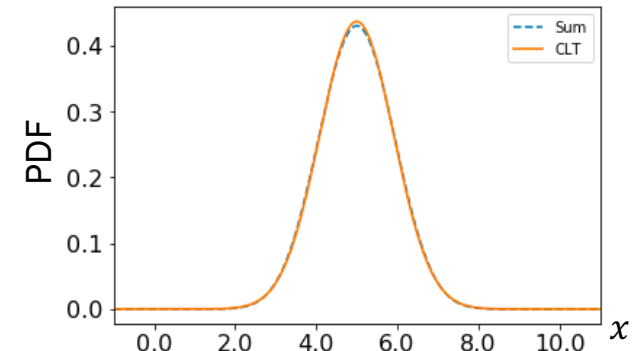
$n = 2$ :



$n = 5$ :



$n = 10$ :



Most books will tell you that CLT holds if  $n \geq 30$ , but it can hold for smaller  $n$  depending on the distribution of your i.i.d.  $X_i$ 's.



# Sample Statistics

# What about other functions?

Let  $X_1, X_2, \dots, X_n$  be i.i.d., where  $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$ . As  $n \rightarrow \infty$ :

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Sum of i.i.d. RVs

?

Average of i.i.d. RVs  
(sample mean)

?

Max of i.i.d. RVs



# What about other functions?

---

Let  $X_1, X_2, \dots, X_n$  be i.i.d., where  $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$ . As  $n \rightarrow \infty$ :

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Sum of i.i.d. RVs

?

Average of i.i.d. RVs  
(sample mean)

?

Max of i.i.d. RVs

# Distribution of sample mean

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

Let  $X_1, X_2, \dots, X_n$  be i.i.d., where  $E[X_i] = \mu$ ,  $\text{Var}(X_i) = \sigma^2$ . As  $n \rightarrow \infty$ :

Define:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  (sample mean)       $Y = \sum_{i=1}^n X_i$  (sum)

$$Y \sim \mathcal{N}(n\mu, n\sigma^2) \quad (\text{CLT, as } n \rightarrow \infty)$$

$$\bar{X} = \frac{1}{n} Y$$

$$\bar{X} \sim \mathcal{N}(?, ?)$$

(Linear transform of a Normal)

$$\frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The average of i.i.d. random variables (i.e., **sample mean**) is normally distributed with mean  $\mu$  and variance  $\sigma^2/n$ .

Demo: [http://onlinestatbook.com/stat\\_sim/sampling\\_dist/](http://onlinestatbook.com/stat_sim/sampling_dist/)

# What about other functions?

Let  $X_1, X_2, \dots, X_n$  be i.i.d., where  $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$ . As  $n \rightarrow \infty$ :

$$\sum_{i=1}^n X_i \sim \mathcal{N}(\underline{n\mu}, \underline{n\sigma^2})$$

Sum of i.i.d. RVs CLT

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$\mathcal{N}(n\mu, n\sigma^2)$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Average of i.i.d. RVs  
(sample mean)

Max of i.i.d. RVs

sample mean  
RV

$$\frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\underline{\mu}, \underline{\frac{\sigma^2}{n}})$$

Gumbel

(see Fisher-Tippett Gnedenko Theorem)



# Exercises

# Dice game

$$\text{As } n \rightarrow \infty: \sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

You will roll 10 6-sided dice( $X_1, X_2, \dots, X_{10}$ ).

- Let  $X = X_1 + X_2 + \dots + X_{10}$ , the total value of all 10 rolls.
- You win if  $X \leq 25$  or  $X \geq 45$ .



[To the demo!](#)



# Dice game

$$\text{As } n \rightarrow \infty: \sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

You will roll 10 6-sided dice ( $X_1, X_2, \dots, X_{10}$ ).

- Let  $X = X_1 + X_2 + \dots + X_{10}$ , the total value of all 10 rolls.
- You win if  $X \leq 25$  or  $X \geq 45$ .



And now the truth (according to the CLT)...

1. Define RVs and state goal.

$$E[X_i] = 3.5,$$
$$\text{Var}(X_i) = 35/12$$

Want:  $P(X \leq 25 \text{ or } X \geq 45)$

Approximate:

?

2. Solve.



# Dice game

$$\text{As } n \rightarrow \infty: \sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

You will roll 10 6-sided dice( $X_1, X_2, \dots, X_{10}$ ).

- Let  $X = X_1 + X_2 + \dots + X_{10}$ , the total value of all 10 rolls.
- You win if  $X \leq 25$  or  $X \geq 45$ .



And now the truth (according to the CLT)...

1. Define RVs and state goal.

$$E[X_i] = 3.5, \\ \text{Var}(X_i) = 35/12$$

Want:  $P(X \leq 25 \text{ or } X \geq 45)$

Approximate:

$$X \approx Y \sim \mathcal{N}(10(3.5), 10(35/12))$$

2. Solve.

$$P(Y \leq 25.5) + P(Y \geq 44.5) \quad \text{or}$$

$$1 - P(25.5 \leq Y \leq 44.5)$$



continuity  
correction

# Dice game

$$\text{As } n \rightarrow \infty: \sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

You will roll 10 6-sided dice ( $X_1, X_2, \dots, X_{10}$ ).

- Let  $X = X_1 + X_2 + \dots + X_{10}$ , the total value of all 10 rolls.
- You win if  $X \leq 25$  or  $X \geq 45$ .



And now the truth (according to the CLT)...

1. Define RVs and state goal.

$$E[X_i] = 3.5, \\ \text{Var}(X_i) = 35/12$$

Want:  $P(X \leq 25 \text{ or } X \geq 45)$

Approximate:

$$X \approx Y \sim \mathcal{N}(10(3.5), 10(35/12))$$

2. Solve.

$$P(Y \leq 25.5) + P(Y \geq 44.5) = \Phi\left(\frac{25.5 - 35}{\sqrt{10(35/12)}}\right) + \left(1 - \Phi\left(\frac{44.5 - 35}{\sqrt{10(35/12)}}\right)\right)$$

$$\approx \Phi(-1.76) + (1 - \Phi(1.76)) \approx (1 - 0.9608) + (1 - 0.9608) = \mathbf{0.0784}$$

# Dice game

$$\text{As } n \rightarrow \infty: \sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

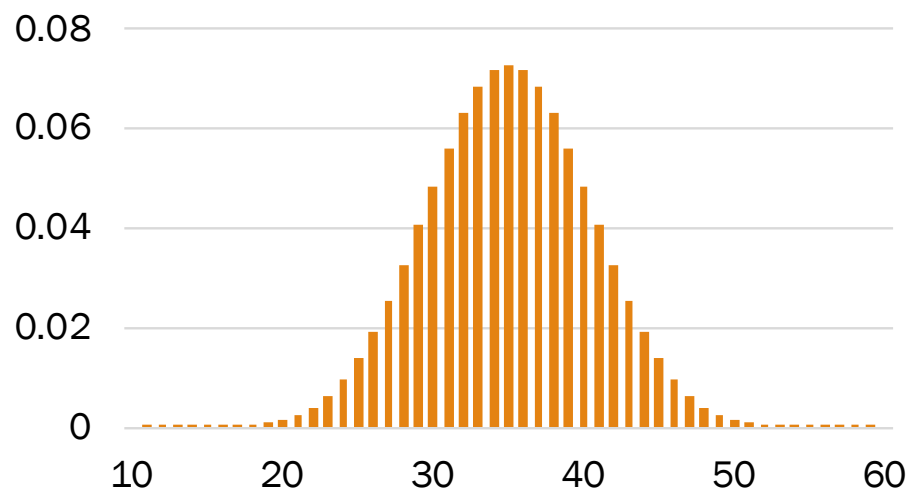
You will roll 10 6-sided dice ( $X_1, X_2, \dots, X_{10}$ ).

- Let  $X = X_1 + X_2 + \dots + X_{10}$ , the total value of all 10 rolls.
- You win if  $X \leq 25$  or  $X \geq 45$ .



And now the truth (according to the CLT)...

Check out  
the [code!](#)



(by CLT)

$$\approx P(X \leq 25.5) + P(X \geq 44.5) \\ \approx 0.0786$$

(exact, by computer)

$$P(X \leq 25 \text{ or } X \geq 45) = 0.0780$$

(exact, by computer)

$$P(X \leq 25 \text{ or } X \geq 45) \approx 0.0776$$

# Summary: Working with the CLT

Let  $X_1, X_2, \dots, X_n$  i.i.d., where  $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$ . As  $n \rightarrow \infty$ :

$$\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Sum of i.i.d. RVs

$$\frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Average of i.i.d. RVs  
(sample mean)



If  $X_i$  is discrete:  
Use the **continuity correction** on  $Y$ !

# Crashing website

- Let  $X$  = number of visitors to a website, where  $X \sim \text{Poi}(100)$ .
- The server crashes if there are  $\geq 120$  requests/minute.

What is  $P(\text{server crashes in next minute})$ ?

Strategy:

Poisson (exact)

$$P(X \geq 120) = \sum_{k=120}^{\infty} \frac{(100)^k e^{-100}}{k!} \approx 0.0282$$

*per minute*

$P_{\text{Poi}}(100) = \sum_{i=1}^n \text{Poi}(100/n)$

$E = 100$   
 $\text{Var} = 100$

$E = 100$   
 $\text{Var} = 100$

Strategy:

CLT

(approx.)

How would we involve CLT here?

(Hint: Is there a way to represent  $X$  as a sum of i.i.d. RVs?)



# Crashing website

- Let  $X$  = number of visitors to a website, where  $X \sim \text{Poi}(100)$ .
- The server crashes if there are  $\geq 120$  requests/minute.

What is  $P(\text{server crashes in next minute})$ ?

Strategy:

Poisson (exact)

$$P(X \geq 120) = \sum_{k=120}^{\infty} \frac{(100)^k e^{-100}}{k!} \approx \underline{0.0282}$$

Strategy:

CLT

(approx.)

State  
approx.  
goal

*forget*

$$\text{Poi}(100) \sim \sum_{i=1}^n \text{Poi}(100/n)$$

$$X \approx Y \sim \mathcal{N}(n\mu, n\sigma^2)$$

$$P(X \geq 120) \approx P(Y \geq 119.5)$$

Check out  
the [code!](#)

Solve

$$P(Y \geq 119.5) = 1 - \Phi\left(\frac{119.5 - 100}{\sqrt{100}}\right) = 1 - \Phi(1.95) \approx \underline{0.0256}$$

# Next time

---

Central Limit Theorem:

- Sample mean  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- If we know  $\mu$  and  $\sigma^2$ , we can compute probabilities on sample mean  $\bar{X}$  of a given sample size  $n$

In real life:

- Yes, the CLT still holds....
- But we **often don't know**  $\mu$  or  $\sigma^2$  of our original distribution
- However, we can collect data (a sample of size  $n$ )!
- How can we **estimate** the values  $\mu$  and  $\sigma^2$  from our sample?

...until next time!