**Table of Contents**


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


# 18: Central Limit Theorem


Jerry Cain


February 21, 2024

Lecture Discussion on Ed

# 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 $x_1, x_2, \dots, x_n$:

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

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

# i.i.d. random variables

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

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

- $X_1, X_2, \ldots, X_n$ are independent, and
- All have the same PMF (if discrete) or PDF (if continuous).

  $\Rightarrow E[X_i] = \mu$ for $i = 1, \ldots, n$

  $\Rightarrow \text{Var}(X_i) = \sigma^2$ for $i = 1, \ldots, n$

Same thing: **i.i.d.** **iid** **IID**

# Quick check

Are $X_1, X_2, \dots, X_n$ iid with the following distributions?

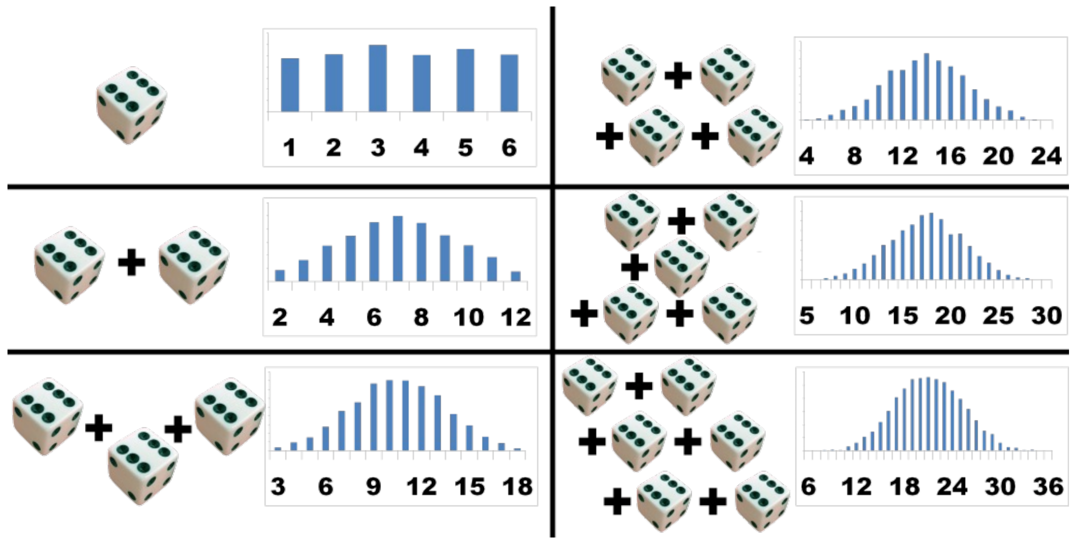
1. $X_i \sim \text{Exp}(\lambda)$, $X_i$ independent
2. $X_i \sim \text{Exp}(\lambda_i)$, $X_i$ independent
3. $X_i \sim \text{Exp}(\lambda)$, $X_1 = X_2 = \cdots = X_n$
4. $X_i \sim \text{Bin}(n_i, p)$, $X_i$ independent

# Quick check

Are $X_1, X_2, \dots, X_n$ iid 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. $X_i \sim \text{Exp}(\lambda),\ X_1 = X_2 = \cdots = X_n$ ❌ dependent: $X_1 = X_2 = \cdots = X_n$
4. $X_i \sim \text{Bin}(n_i, p),\ X_i$ independent ❌ (unless $n_i$ equal)
   Note underlying Bernoulli RVs are iid!

# Central Limit Theorem




[source]

# Central Limit Theorem

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

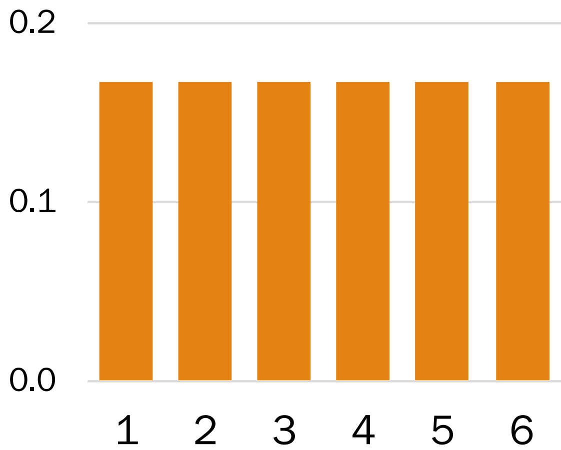
As $n \to \infty$

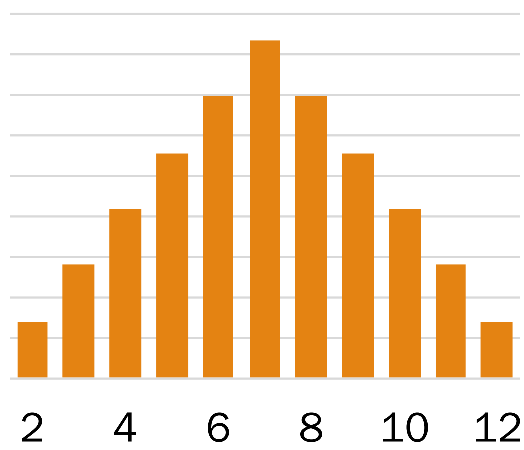
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

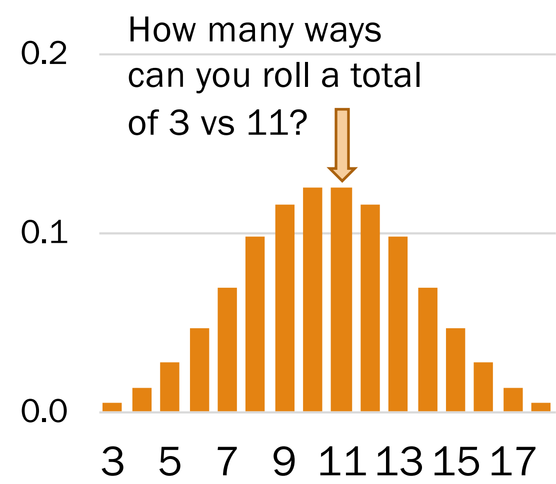
The sum of $n$ iid random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

# Sum of dice rolls

Roll $n$ independent dice. Let $X_i$ be the outcome of roll $i$. $X_\mathrm{i}$ are iid




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

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

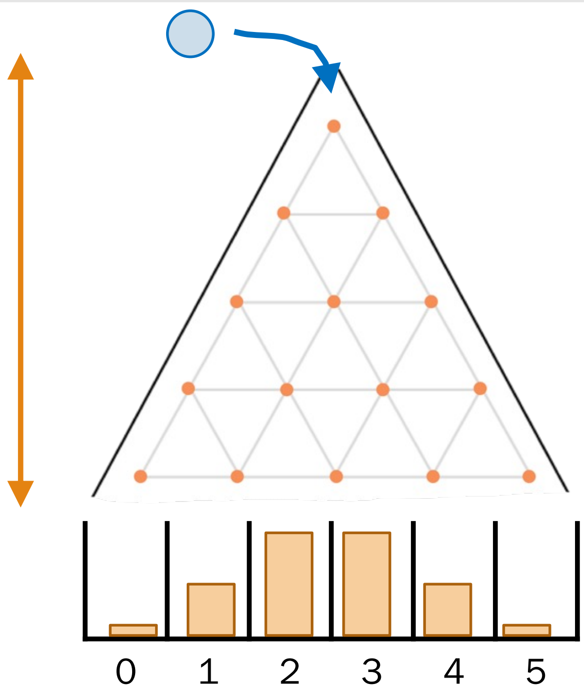
$$\sum_{i=1}^{3} X_i$$ Sum of 3 dice rolls

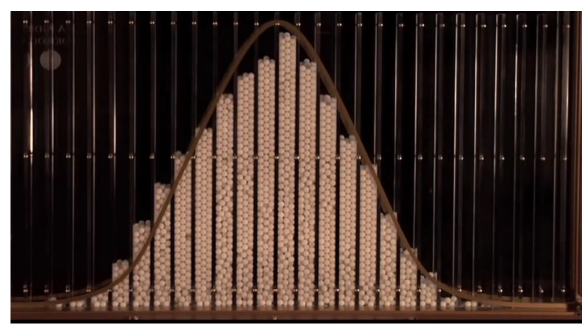
# CLT explains a lot

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

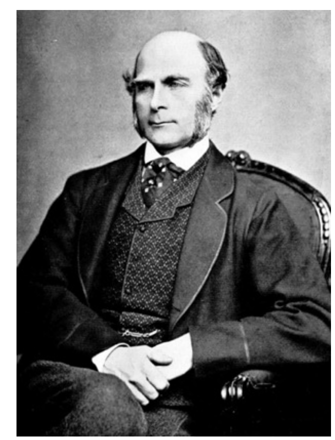
The sum of $n$ **iid** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

$n = 5$



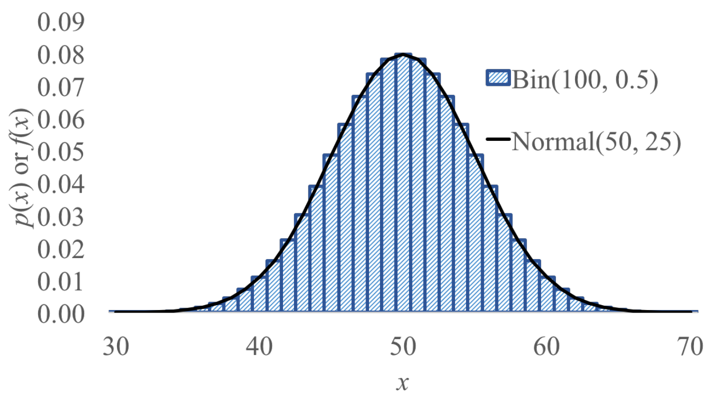
Galton Board, by Sir Francis Galton (1822-1911)



# CLT explains a lot

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

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




**Normal approximation of Binomial**

Sum of iid Bernoulli RVs ≈ Normal

Proof:

Let $X_i \sim \text{Ber}(p)$ for $i = 1, \dots, n$, where $X_i$ are iid

$E[X_i] = p$, $\text{Var}(X_i) = p(1-p)$

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

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

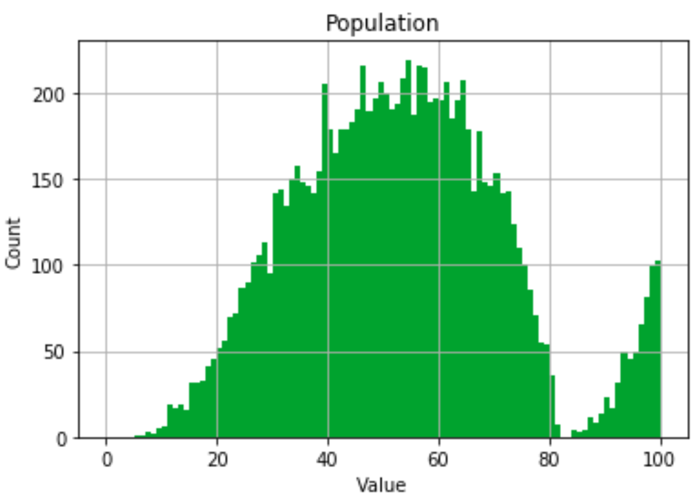
$$X \sim \mathcal{N}(np, np(1-p)) \qquad (\text{substitute mean, variance of Bernoulli})$$

# CLT explains a lot

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

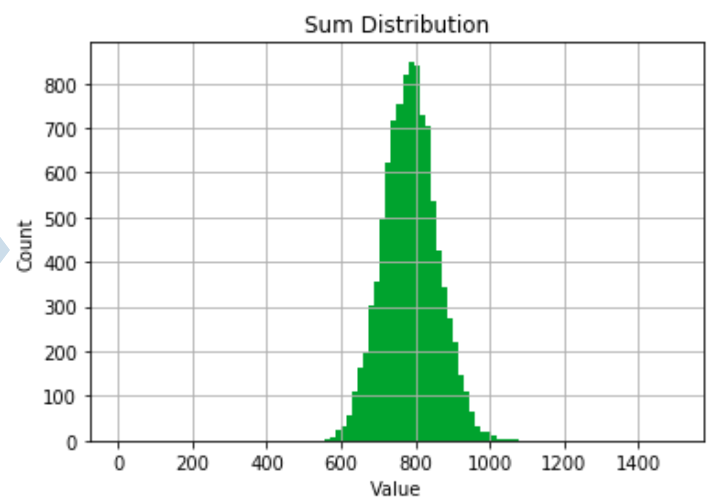
As $n \to \infty$

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




Distribution of $X_i$

Sample of size 15, sum values

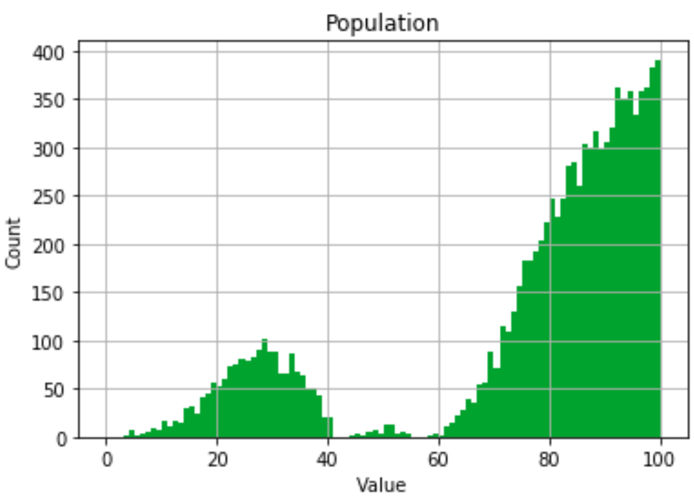



Distribution of $\sum_{i=1}^{15} X_i$

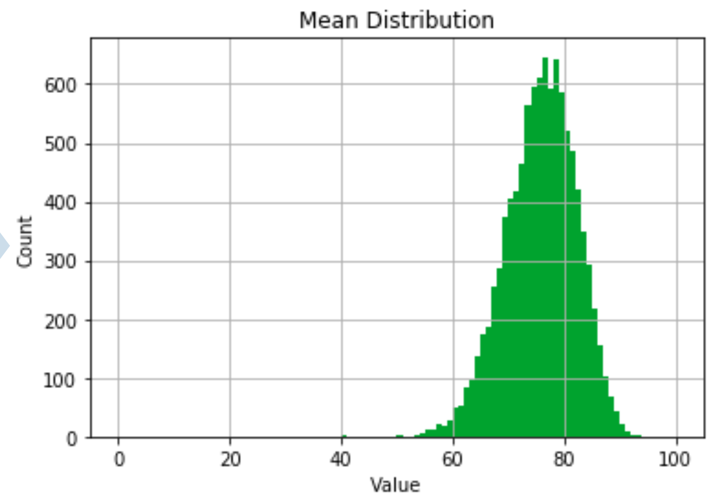
# CLT explains a lot

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

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




Sample of size 15, **average** values




Distribution of $X_i$

(sample mean)

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

# Proof of CLT

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

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

Proof:

- The Fourier Transform of a PDF is its **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 $n \to \infty$: $f(x) = e^{-\frac{x^2}{2}}$
- 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)

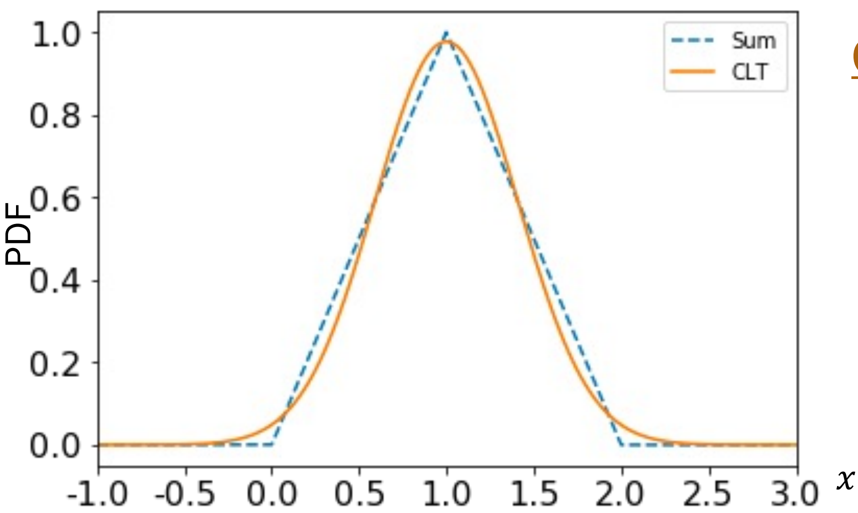
# Sum of $n$ independent Uniform RVs

Let $X = \sum_{i=1}^{n} X_i$ be sum of **iid** 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 $\quad P(X \leq 2/3) \approx 0.2222$

CLT approximation

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

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

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

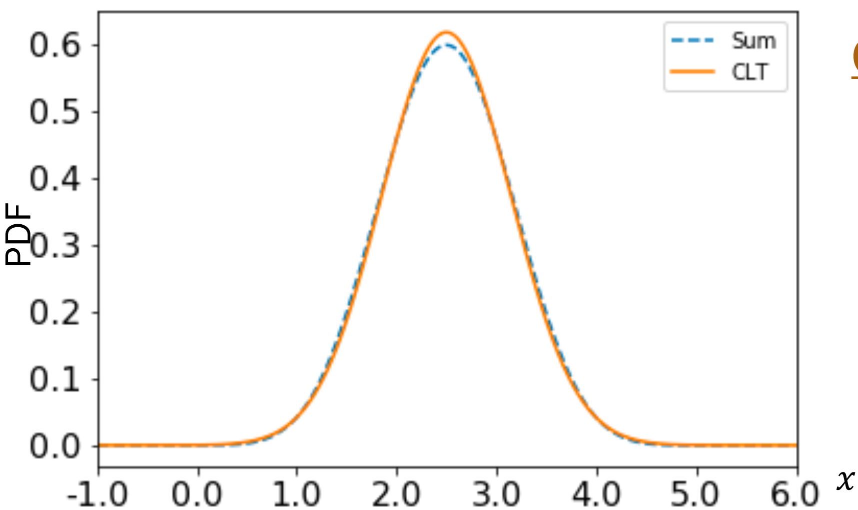
# Sum of $n$ independent Uniform RVs

Let $X = \sum_{i=1}^{n} X_i$ be sum of **iid** 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 $\quad P(X \leq 5/3) \approx 0.1017$

CLT approximation

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

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

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

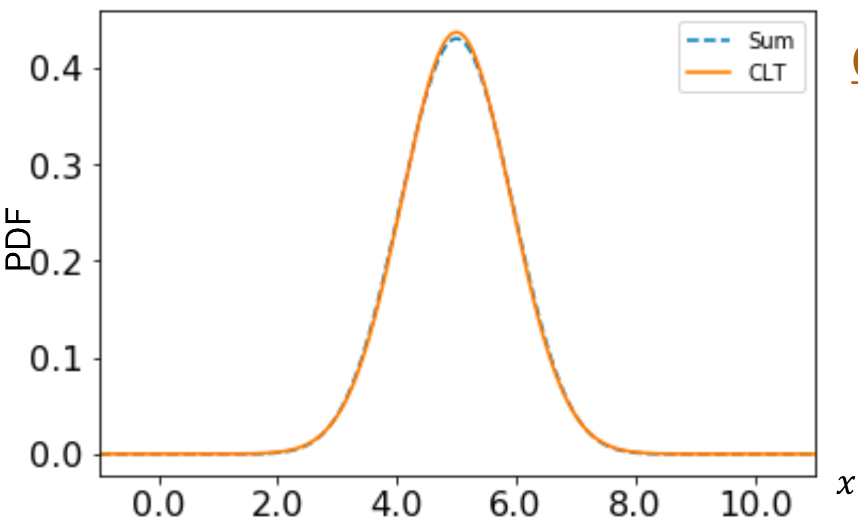
# Sum of $n$ independent Uniform RVs

Let $X = \sum_{i=1}^{n} X_i$ be sum of **iid** 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 $P(X \le n/3)$?

$n = 10$:




Exact $\quad P(X \le 10/3) \approx 0.0337$

CLT approximation

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

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

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

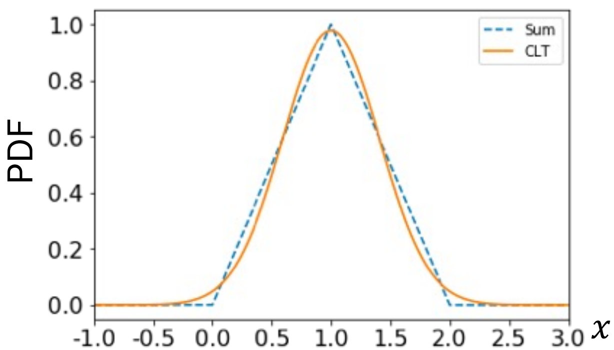
# Sum of $n$ independent Uniform RVs

Let $X = \sum_{i=1}^{n} X_i$ be sum of iid RVs, where $X_i \sim \text{Uni}(0,1)$.

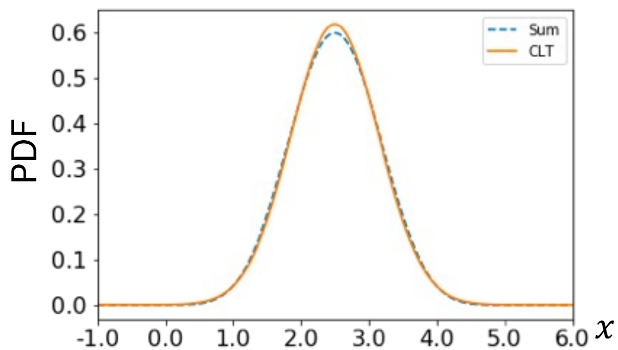
$\mu = E[X_i] = 1/2$
$\sigma^2 = \text{Var}(X_i) = 1/12$

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

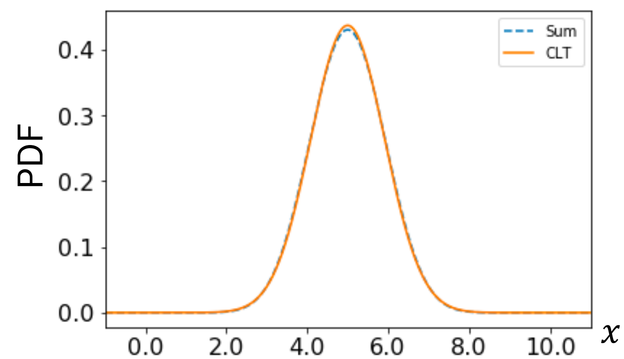
$n = 2$:




$n = 5$:




$n = 10$:




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

# Sample Statistics

# What about other functions?

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

| | |
|---|---|
| $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$ | Sum of iid RVs |
| ? | Average of iid RVs (sample mean) |
| ? | Max of iid RVs |

# What about other functions?

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

| | |
|---|---|
| $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$ | Sum of iid RVs |
| ? | Average of iid RVs (sample mean) |
| ? | Max of iid RVs |

# Distribution of sample mean

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

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

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

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

$\bar{X} \sim \mathcal{N}(\,?\,,\ ?\,)$ (Linear transform of a Normal)

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

The average of iid 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/

# What about other functions?

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

| | |
|---|---|
| $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$ | Sum of iid RVs |
| $\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ | Average of iid RVs<br>(sample mean) |
| Gumbel | Max of iid RVs |

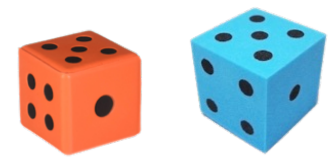
(see Fisher-Tippett Gnedenko Theorem)

# Exercises

# Dice game

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

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



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

[To the demo!]

# Dice game

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

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

- Let $X = X_1 + X_2 + \cdots + X_{10}$, 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: ?

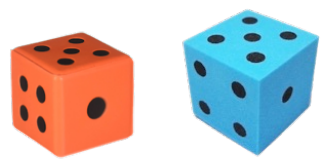
2. Solve.

# Dice game

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

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

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



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: $P(X \le 25 \text{ or } X \ge 45)$

Approximate:

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

2. Solve.

$$P(Y \le 25.5) + P(Y \ge 44.5) \quad \text{or} \quad 1 - P(25.5 \le Y \le 44.5)$$

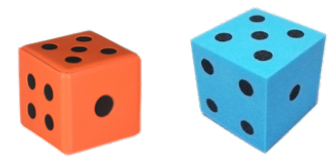
⚠ continuity correction

# Dice game

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

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

- Let $X = X_1 + X_2 + \cdots + X_{10}$, 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.

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

$$\approx \Phi(-1.76) + \left(1 - \Phi(1.76)\right) \approx (1 - 0.9608) + (1 - 0.9608) = 0.0786$$

# Dice game

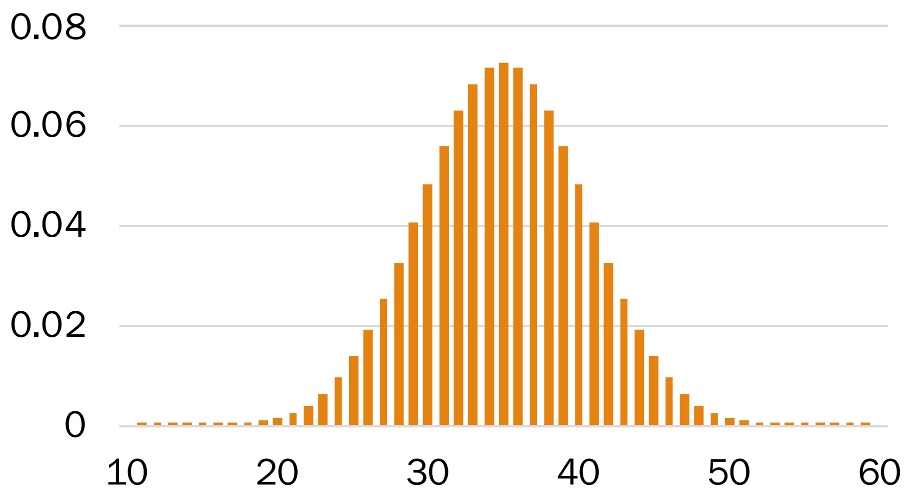
As $n \to \infty$: $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$

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

- Let $X = X_1 + X_2 + \cdots + X_{10}$, 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)

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

(exact, by computer)

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

(sampling via computer)

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

# Summary: Working with the CLT

Let $X_1, X_2, \ldots, X_n$ iid, where $E[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2$. As $n \to \infty$:

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

Sum of iid RVs

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

Average of iid 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 $P$(server crashes in next minute)?

Strategy:
Poisson (exact)

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

Strategy:
CLT
(approx.)

How would we involve CLT here?

(Hint: Is there a way to represent $X$ as a sum of iid 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 $P$(server crashes in next minute)?

Strategy:
Poisson (exact)

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

Strategy:
CLT
(approx.)

State approx. goal

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

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

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

Solve

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

# Clock running time

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

Want to find the mean (clock) runtime of an algorithm, $\mu = t$ sec.

- Suppose variance of runtime is $\sigma^2 = 4$ sec².

Run algorithm repeatedly (iid trials):

- $X_i$ = runtime of $i$-th run (for $1 \le i \le n$)
- Estimate runtime to be **average** of $n$ trials, $\bar{X}$

How many trials do we need s.t. estimated time $= t \pm 0.5$ with **95% certainty**?

1. Define RVs and state goal.
2. Solve.

(CLT) $\bar{X} \sim \mathcal{N}\left(t, \frac{4}{n}\right)$

Want: $P(t - 0.5 \le \bar{X} \le t + 0.5) = 0.95$

(linear transform of a normal) $\bar{X} - t \sim \mathcal{N}\left(0, \frac{4}{n}\right)$

$P(-0.5 \le \bar{X} - t \le 0.5) = 0.95$

# Clock running time

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

Want to find the mean (clock) runtime of an algorithm, $\mu = t$ sec.

- Suppose variance of runtime is $\sigma^2 = 4$ sec².

Run algorithm repeatedly (iid trials):

- $X_i$ = runtime of $i$-th run (for $1 \le i \le n$)
- Estimate runtime to be **average** of $n$ trials, $\bar{X}$

How many trials do we need s.t. estimated time $= t \pm 0.5$ with 95% certainty?

1. Define RVs and state goal.

$$\bar{X} - t \sim \mathcal{N}\left(0, \frac{4}{n}\right)$$

$$0.95 = P(-0.5 \le \bar{X} - t \le 0.5)$$

2. Solve.

$$0.95 = F_{\bar{X}-t}(0.5) - F_{\bar{X}-t}(-0.5)$$

$$= \Phi\left(\frac{0.5 - 0}{\sqrt{4/n}}\right) - \Phi\left(\frac{-0.5 - 0}{\sqrt{4/n}}\right) = 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1$$

# Clock running time

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

Want to find the mean (clock) runtime of an algorithm, $\mu = t$ sec.

- Suppose variance of runtime is $\sigma^2 = 4$ sec².

Run algorithm repeatedly (iid trials):

- $X_i$ = runtime of $i$-th run (for $1 \le i \le n$)
- Estimate runtime to be **average** of $n$ trials, $\bar{X}$

How many trials do we need s.t. estimated time $= t \pm 0.5$ with **95% certainty**?

1. Define RVs and state goal.

$$\bar{X} - t \sim \mathcal{N}\left(0, \frac{4}{n}\right)$$

$$0.95 = P(-0.5 \le \bar{X} - t \le 0.5)$$

2. Solve.

$$0.95 = F_{\bar{X}-t}(0.5) - F_{\bar{X}-t}(-0.5)$$

$$= \Phi\left(\frac{0.5 - 0}{\sqrt{4/n}}\right) - \Phi\left(\frac{-0.5 - 0}{\sqrt{4/n}}\right) = 2\Phi\left(\frac{\sqrt{n}}{4}\right) - 1$$

$$0.975 = \Phi(\sqrt{n}/4)$$

$$\sqrt{n}/4 = \Phi^{-1}(0.975) \approx 1.96 \quad \Rightarrow \quad n \approx 62$$

# Clock running time

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

Want to find the mean (clock) runtime of an algorithm, $\mu = t$ sec.

- Suppose variance of runtime is $\sigma^2 = 4$ sec².

Run algorithm repeatedly (iid trials):

- $X_i =$ runtime of $i$-th run (for $1 \le i \le n$)
- Estimate runtime to be **average** of $n$ trials, $\bar{X}$

How many trials do we need s.t. estimated time $= t \pm 0.5$ with **95% certainty**?

$$n \approx 62$$

**Interpret:** As we increase $n$ (the size of our sample):

- The variance of our sample mean, $\sigma^2/n$ decreases
- The probability that our sample mean $\bar{X}$ is *close* to the true mean $\mu$ increases

# 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. It always 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? And how reliable are those estimates?

...until next time!