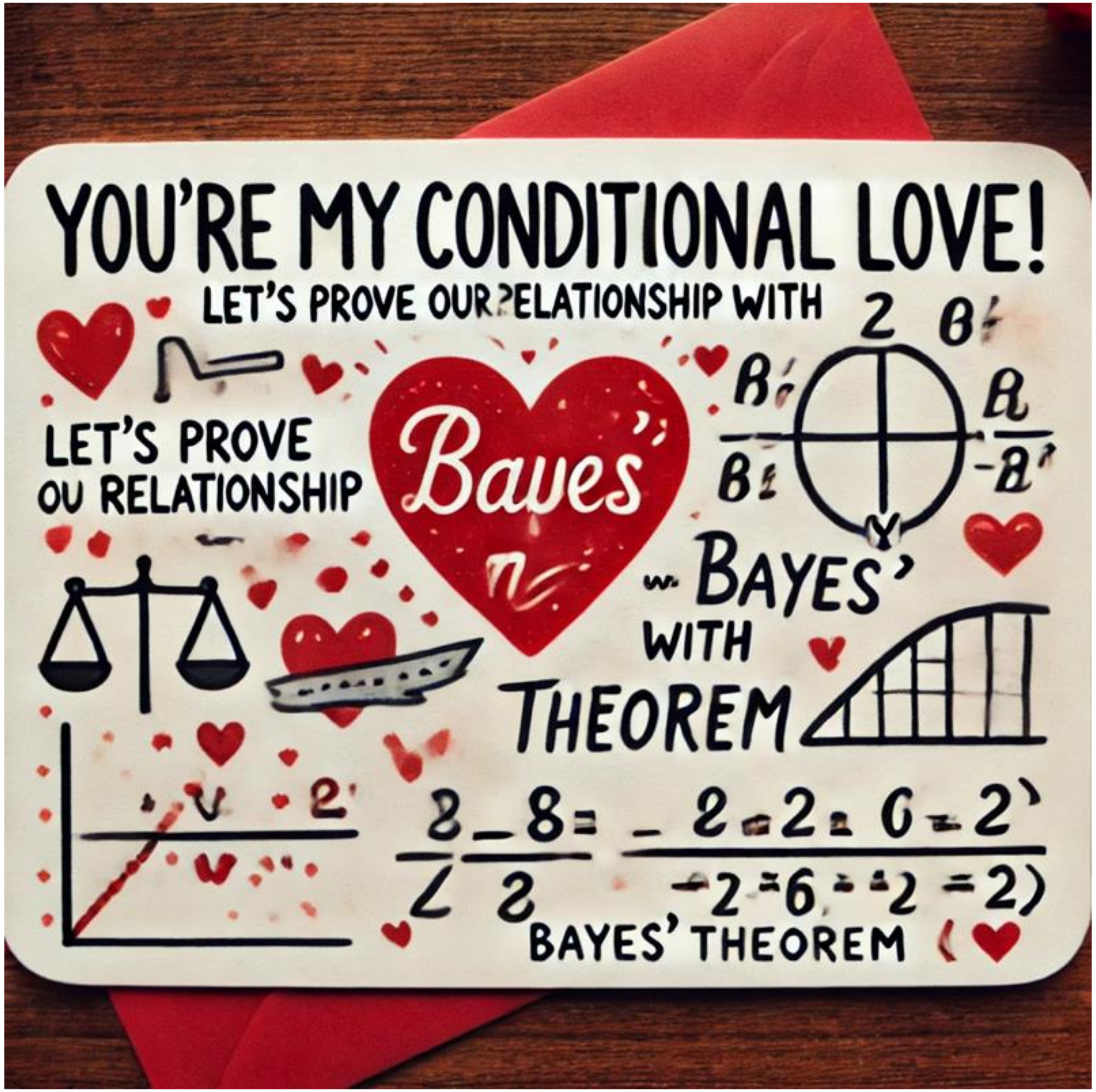
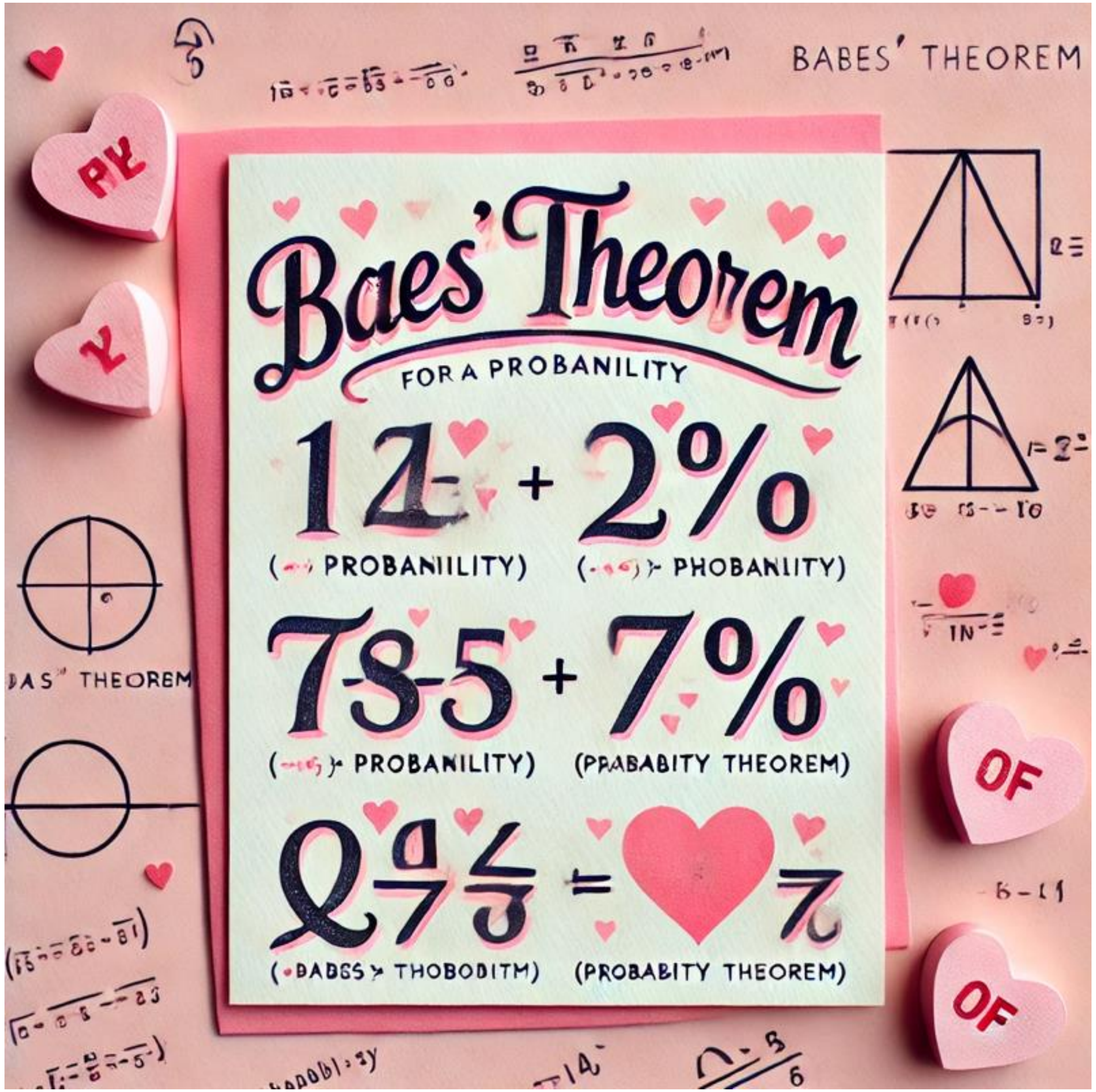
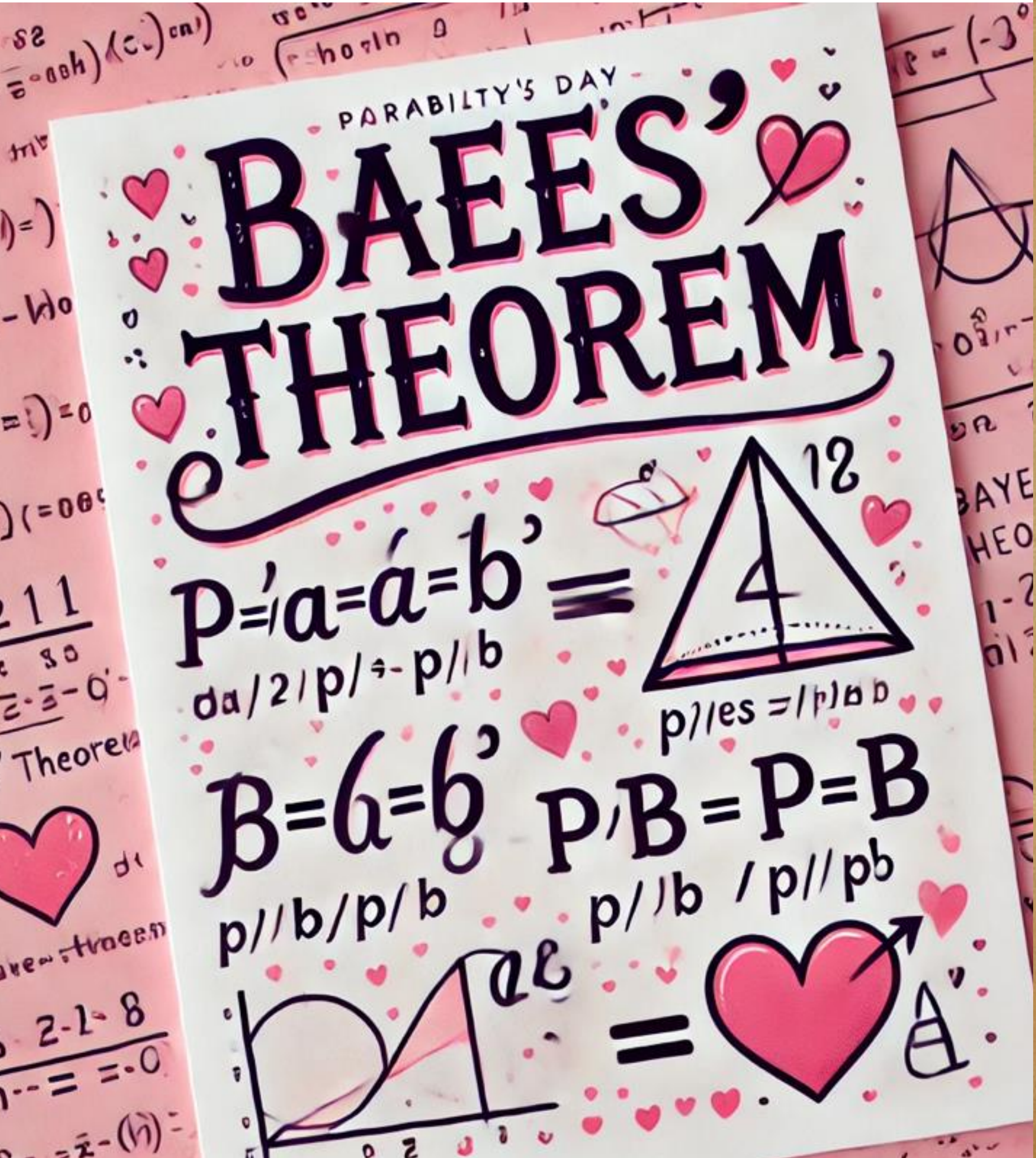
Adding Random Variables

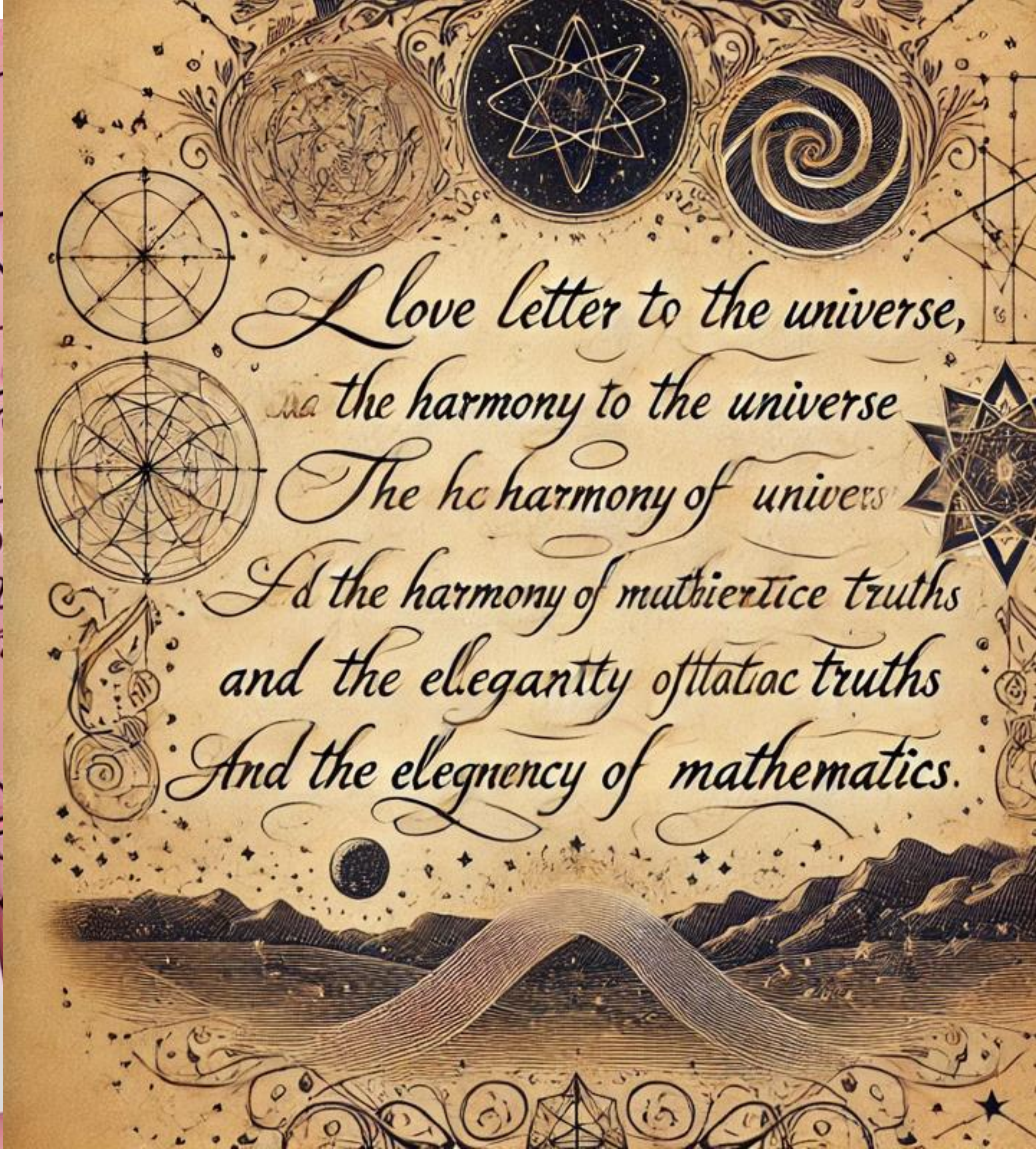
Chris Piech
CS109, Stanford University

MARCH 10TH

Diffusion Model creates a Bae’s theorem Valentines for you!


YOU'RE MY CONDITIONAL LOVE!
LET'S PROVE OUR?ELATIONSHIP WITH
LET'S PROVE
OU RELATIONSHIP
Baves'
BAYES'
WITH
THEOREM
BAYES' THEOREM

BABES' THEOREM
Baes' Theorem
FOR A PROBANILITY
PROBANILITY)
PHOBANIITY)
PROBANILITY)
(PRABABITY THEOREM)
BAS' THEOREM
THOBODITM)
(PROBABITY THEOREM)
OF
OF


PARABILTY'S DAY
BAEES' THEOREM
P/B = P=B


love letter to the universe,
the harmony to the universe
The hc harmony of univers
the harmony of muthiertice truths
and the elegantty oftlatac truths
And the elegnency of mathematics.

Where are we in CS109?

You are here

Counting Theory

Core Probability

Random Variables

Probabilistic Models

Uncertainty Theory

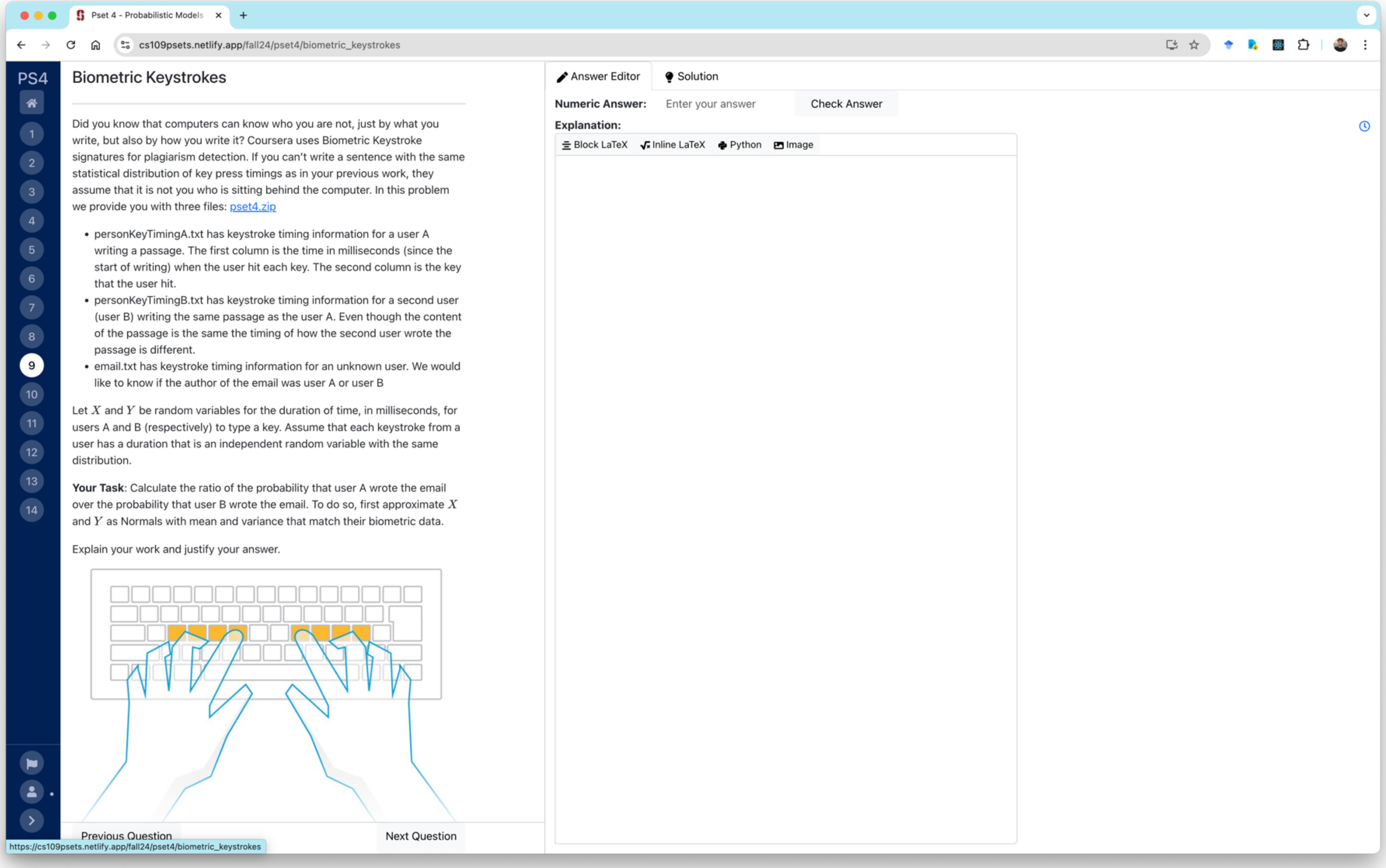
Machine Learning

Uncertainty Theory

- Beta Distributions
- Thompson Sampling
- Adding Random Vars
- Central Limit Theorem
- Sampling
- Bootstrapping
- Algorithmic Analysis
- Information Theory + Divergence

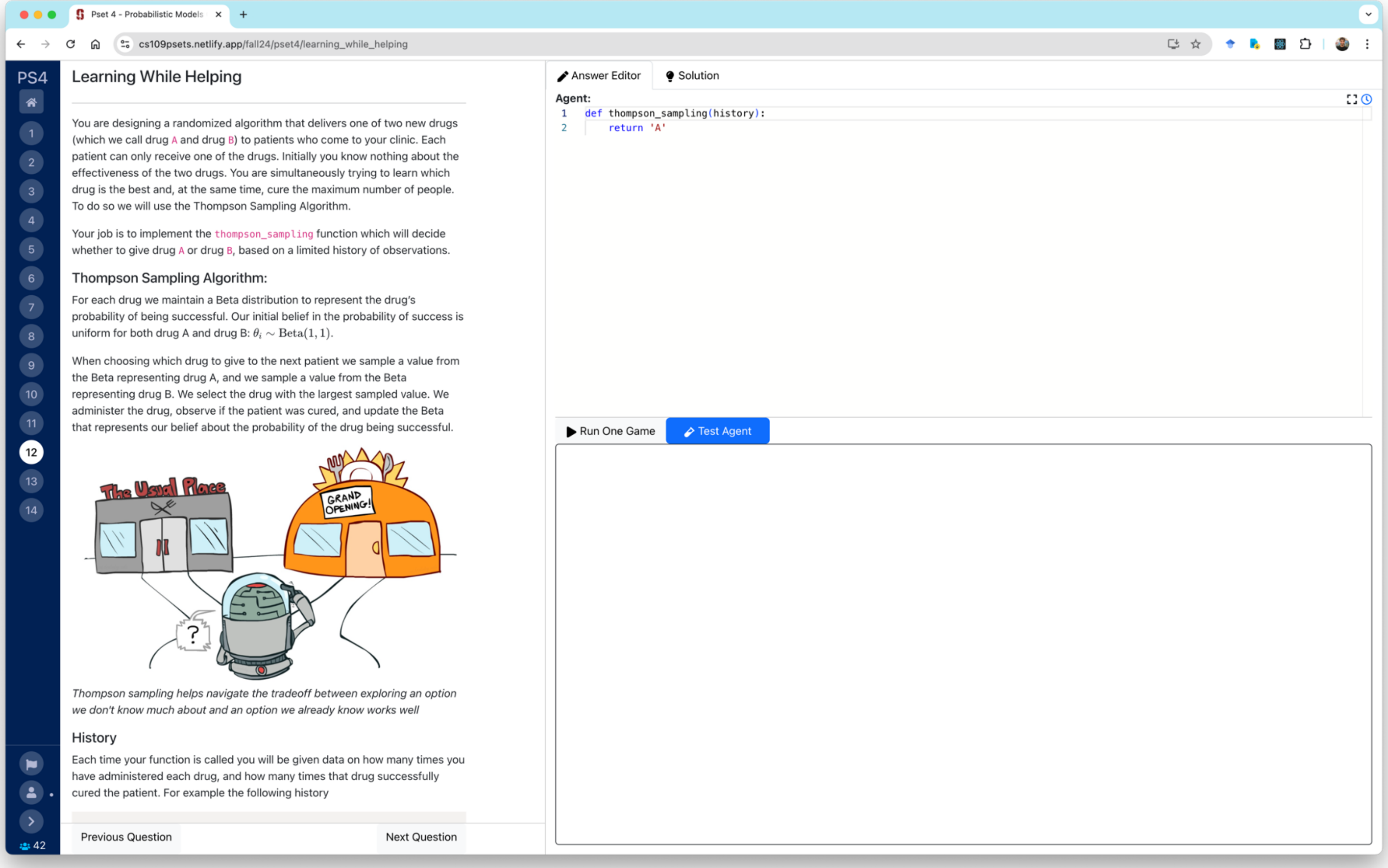
As requested by AI faculty

You have all you need for Pset #5



After today's lecture you can solve all of the problems

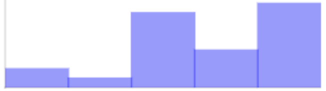
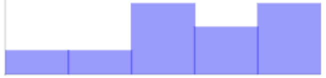
You have all you need for Pset #5



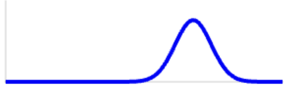
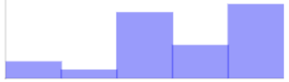
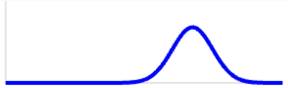
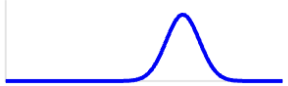
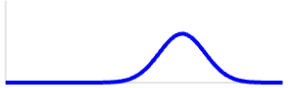
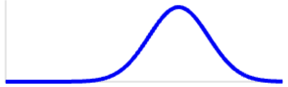
After today's lecture you can solve all of the problems

Which Songs are in the CS109 Top 16?

How Confident Are you that My Way is better than Yellow?

#	Song	Votes	NumVotes	SampleMean
1	My Way - Frank Sinatra		24	3.71
2	Yellow - Coldplay		23	3.7
3	Ma Meilleure Ennemie - Stromae, Pomme		16	3.56
4	Heavy - The Marias		20	3.55
5	Pluto Projector - Rex Orange County		10	3.5

How Confident Are you that My Way is better than Yellow?

#	Song	Sample Mean PDF	Votes	NumVotes	Pr(Top16)	Pr(Best)
1	My Way - Frank Sinatra			24	0.739	0.500
2	Yellow - Coldplay			23	0.709	0.490
3	Ma Meilleure Ennemie - Stromae, Pomme			16	0.530	0.336
4	Heavy - The Marias			20	0.507	0.352
5	Pluto Projector - Rex Orange County			10	0.460	0.338

New Definitions

Independent Random Variables

Recall the comma means "and"

$$P(X = i, Y = k - i)$$

$$= P(X = i) \cdot P(Y = k - i)$$

Because they are independent, "and" becomes multiplication

IID Random Variables

- Consider *n* random variables X_1, X_2, ... X_n
  - X_i are all independently and identically distributed (I.I.D.)
  - All have the same PMF (if discrete) or PDF (if continuous)
  - All have the same expectation
  - All have the same variance

IID   iid

Quick check

Are $X_1, X_2, \ldots, X_n$ 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. $X_i \sim \text{Exp}(\lambda)$, $X_1 = X_2 = \cdots = X_n$

4. $X_i \sim \text{Bin}(n_i, p)$, X_i independent

Quick check

Are $X_1, X_2, \ldots, X_n$ 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 λ_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 i.i.d.!

What happens when you add random variables?

Why should you care?

Zero Sum Games

What is the probability that the Warriors win?

How do you model zero sum games?

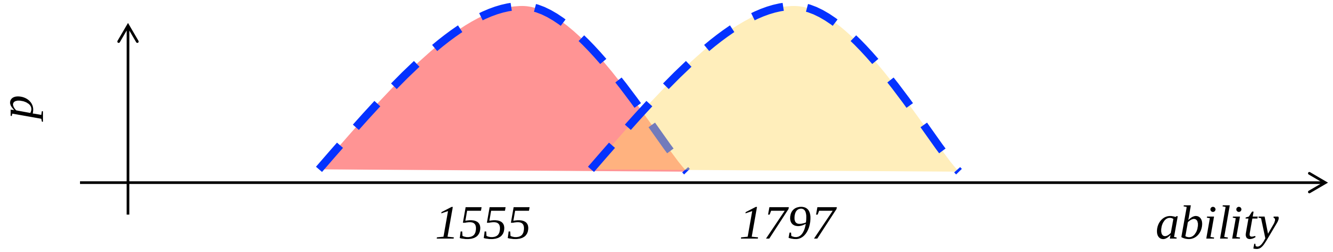
Motivating Idea: Zero Sum Games

How it works:

- Each team has an "ELO" score *S*, calculated based on their past performance.
- Each game, the team has ability $A \sim N(S, 200^2)$
- The team with the higher sampled ability wins.

Arpad Elo

$$A_B \sim \mathcal{N}(1555, 200^2) \qquad A_W \sim \mathcal{N}(1797, 200^2)$$



$$P(\text{Warriors win}) = P(A_W > A_B)$$

Motivating Idea: Zero Sum Games

$$A_W \sim \mathcal{N}(1797, 200^2)$$

$$A_B \sim \mathcal{N}(1555, 200^2)$$

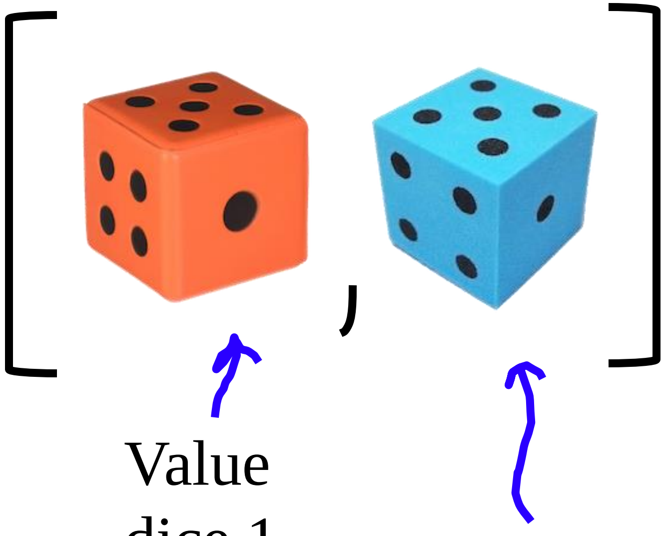
$$P(\text{Warriors win}) = P(A_W > A_B)$$

How do we do this???

Sum of Two Die?

Roll two 6-sidex dice. What is P(sum = 7)?

Each outcome



Value dice 1

Value dice 2

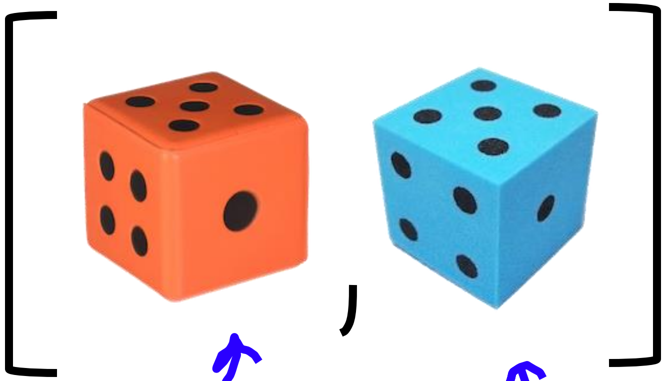
S = {

[1,1]	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]
[2,1]	[2,2]	[2,3]	[2,4]	[2,5]	[2,6]
[3,1]	[3,2]	[3,3]	[3,4]	[3,5]	[3,6]
[4,1]	[4,2]	[4,3]	[4,4]	[4,5]	[4,6]
[5,1]	[5,2]	[5,3]	[5,4]	[5,5]	[5,6]
[6,1]	[6,2]	[6,3]	[6,4]	[6,5]	[6,6]

}

Sum of Two Die = 7?

Roll two 6-sidex dice. What is P(sum = 7)?



S = {

[1,1]	[1,2]	[1,3]	[1,4]	[1,5]	**[1,6]**
[2,1]	[2,2]	[2,3]	[2,4]	**[2,5]**	[2,6]
[3,1]	[3,2]	[3,3]	**[3,4]**	[3,5]	[3,6]
[4,1]	[4,2]	**[4,3]**	[4,4]	[4,5]	[4,6]
[5,1]	**[5,2]**	[5,3]	[5,4]	[5,5]	[5,6]
[6,1]	[6,2]	[6,3]	[6,4]	[6,5]	[6,6] }

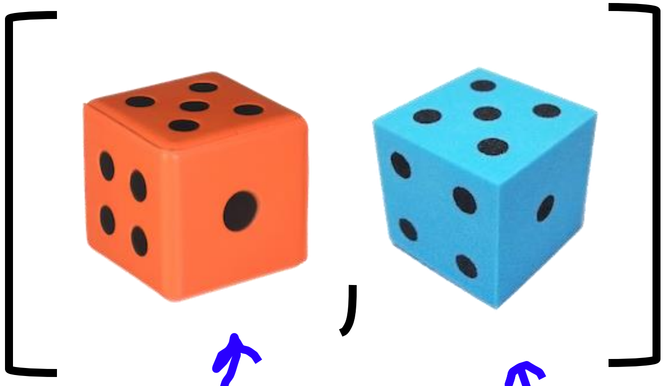
E = *in blue*

$$P(E) = \frac{|E|}{|S|} = \frac{6}{36} = 0.16\bar{6}$$

Sum of Two Die = 10?

Roll two 6-sidex dice. What is P(sum = 10)?

Each outcome



Value dice 1

Value dice 2

S = {

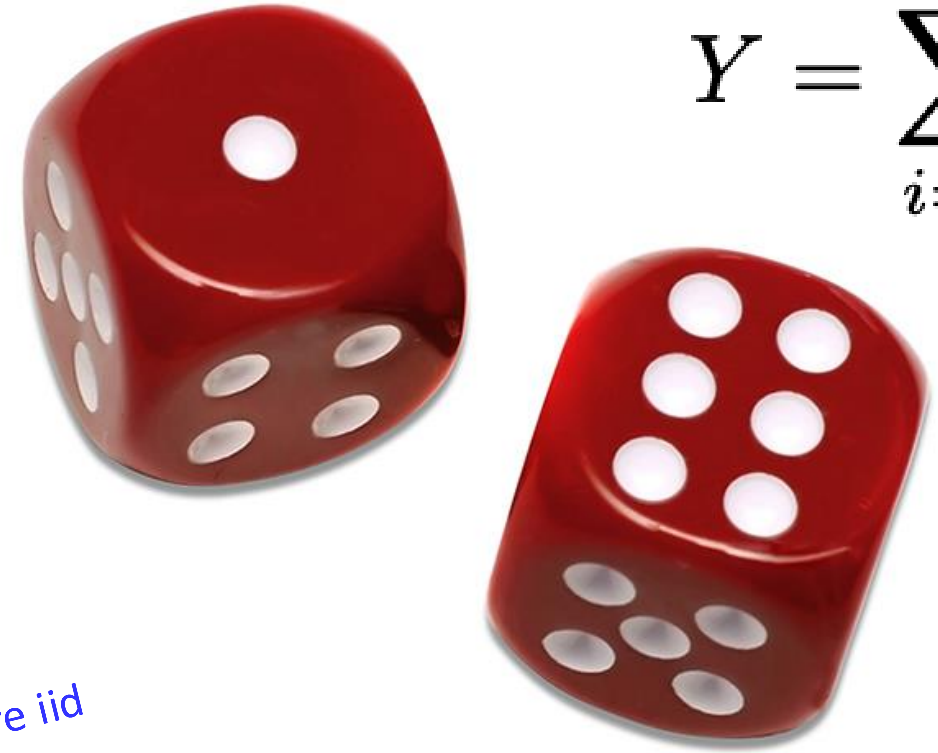
[1,1]	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]
[2,1]	[2,2]	[2,3]	[2,4]	[2,5]	[2,6]
[3,1]	[3,2]	[3,3]	[3,4]	[3,5]	[3,6]
[4,1]	[4,2]	[4,3]	[4,4]	[4,5]	**[4,6]**
[5,1]	[5,2]	[5,3]	[5,4]	**[5,5]**	[5,6]
[6,1]	[6,2]	[6,3]	**[6,4]**	[6,5]	[6,6] }

E = *in blue*

$$P(E) = \frac{|E|}{|S|} = \frac{3}{36} = 0.08\bar{3}$$

End Review

Sum of Two Dice



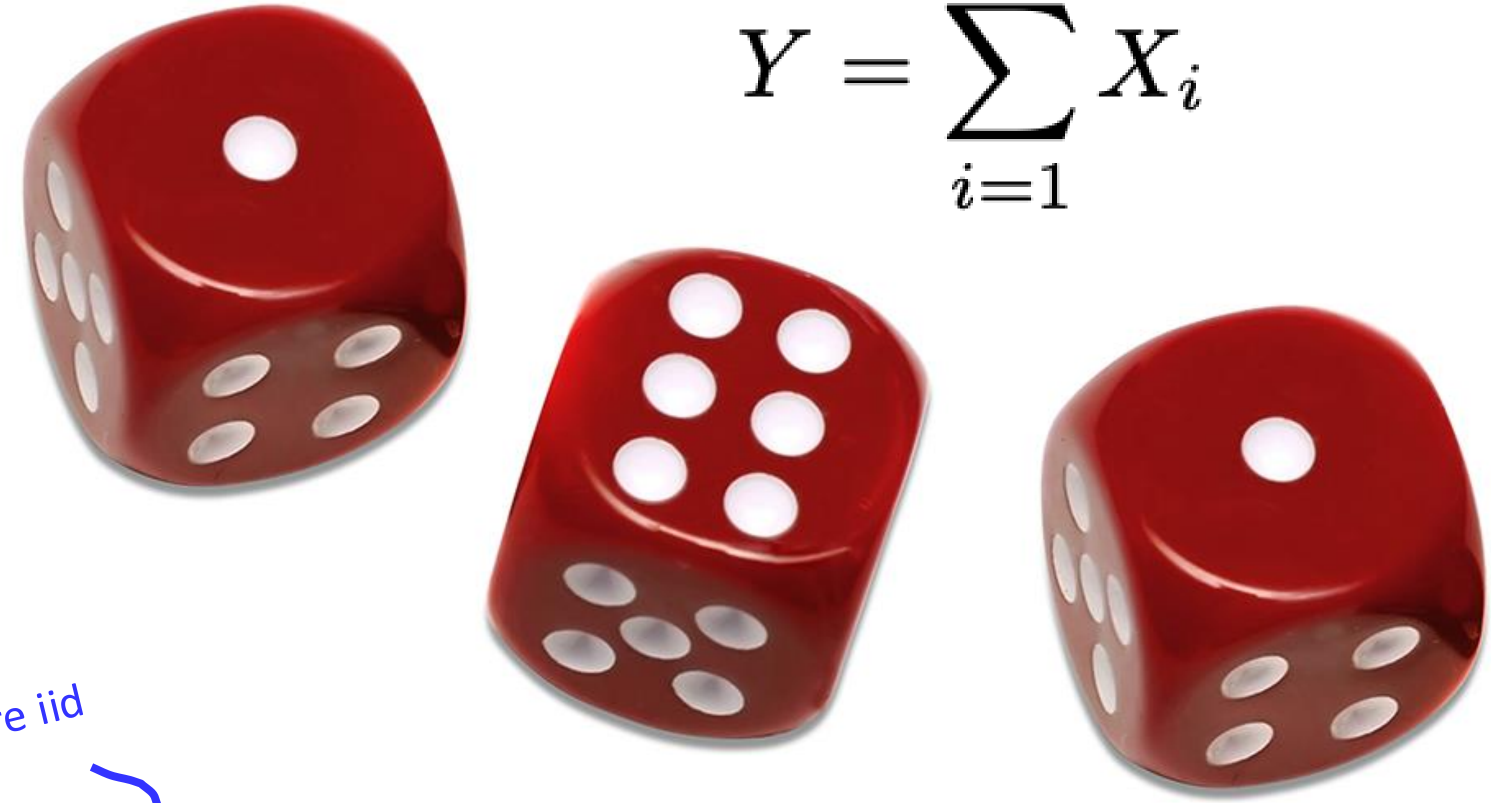
$$Y = \sum_{i=1}^{2} X_i$$

X_i s are iid

X_i is the outcome of dice roll i

Sum of Three Dice

$$Y = \sum_{i=1}^{3} X_i$$

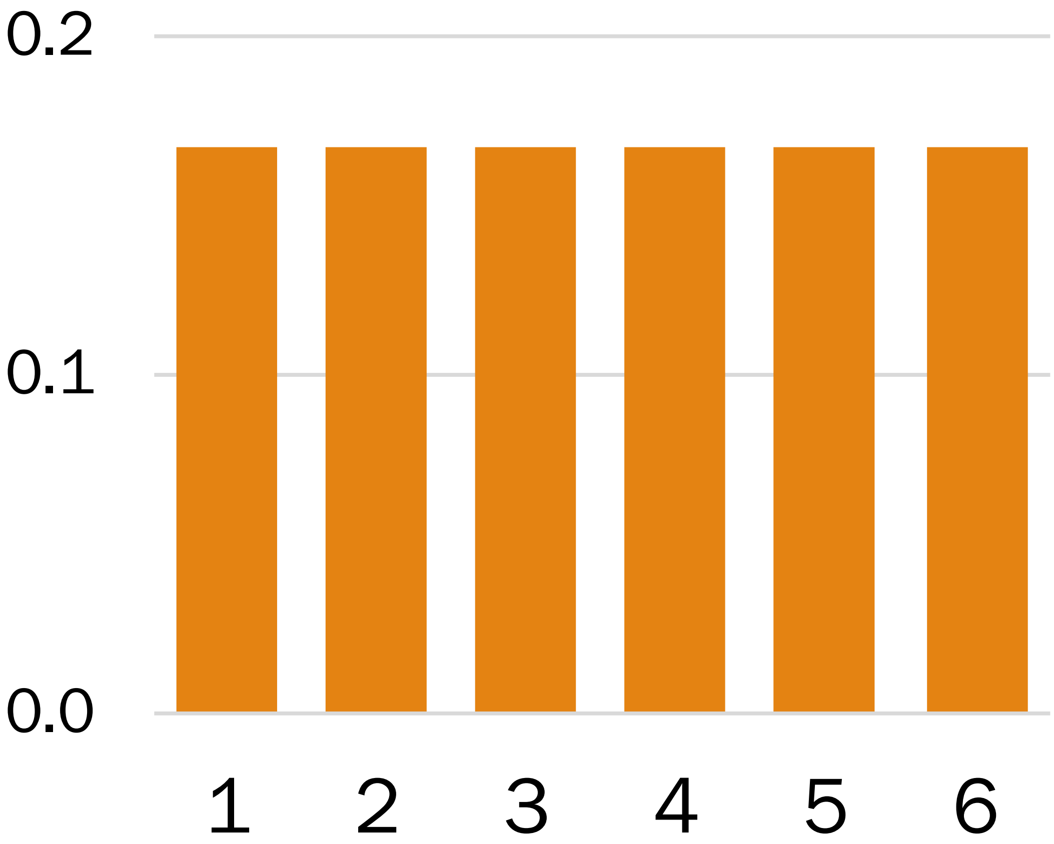


X_i s are iid

X_i is the outcome of dice roll i

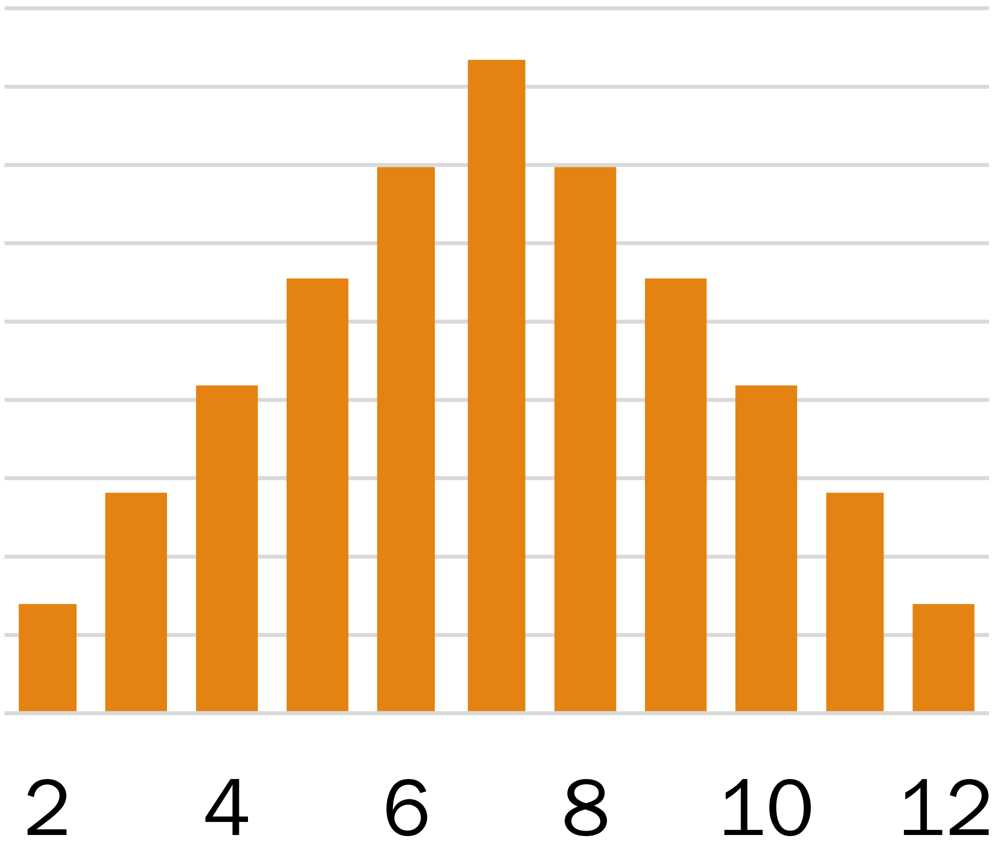
Sum of One Dice

This is the PMF of the sum of one dice



Sum of Two Dice

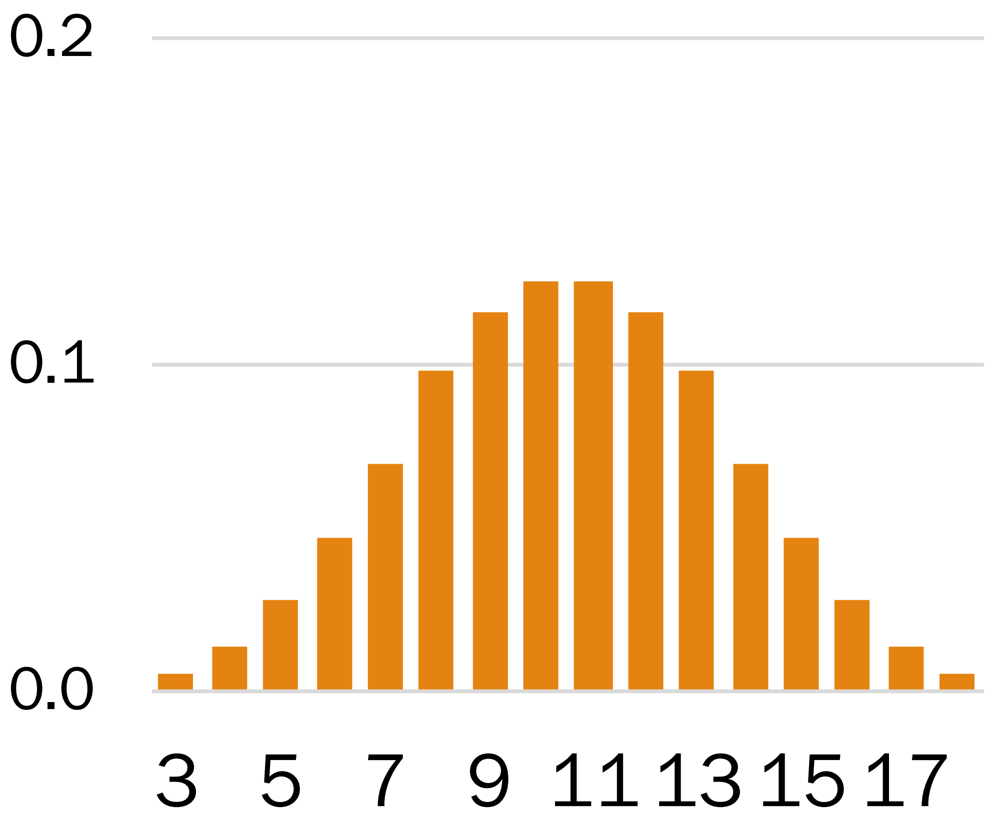
This is the PMF of the sum of two dice



Why is there more mass in the middle?

Sum of Three Dice

This is the PMF of the sum of three dice



Why is there more mass in the middle?

Sum of 50 dice?

Insight to Convolution

Imagine a game
where each player *independently* scores between 0 and 100 points:

Let X be the amount of points you score.
Let Y be the amount of points your opponent scores.
Let's say you know $P(X = x)$ and $P(Y = y)$.

What is the probability of a tie?

Note: these could be any distribution! As long as you know the PMFs

---

$$P(\text{tie}) = \sum_{i=0}^{100} P(X = i, Y = i)$$

$$= \sum_{i=0}^{100} P(X = i)P(Y = i)$$

What is the PMF for X + Y, $P(X+Y=n)$?

In English: What is the probability that X + Y = n?

Consider the case where X and Y are discrete and non-negative

X	Y	i	
0	n	0	$P(X=0, Y=n)$
1	n - 1	1	$P(X=1, Y=n-1)$
2	n - 2	2	$P(X=2, Y=n-2)$
	…		
n	0	n	$P(X=n, Y=0)$

$$P(X+Y=n) = \sum_{i=0}^{n} P(X=i, Y=n-i)$$

What is the PMF for X + Y, $P(X+Y=n)$?

Consider the case where X and Y are discrete and non-negative

In English: What is the probability that X + Y = n?

$$P(X+Y=n) = \sum_{k=0}^{n} P(X=k, Y=n-k)$$

Since this is the OR of mutually exclusive events

$$= \sum_{k=0}^{n} P(X=k)P(Y=n-k)$$

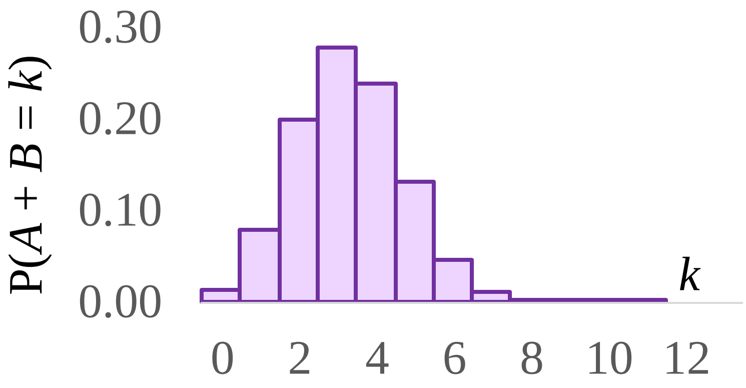
If the random variables are independent

Wildlife Disease Outbreak

Researchers are tracking a contagious disease in two distinct animal populations.

Population A has Bin(n = 5, p = 0.1) infections
Population B has Bin(n = 8, p = 0.5) infections.

Find the ***exact*** probability distribution of total infections across A and B.



$$P(A + B = k)$$
$$= \sum_{i=0}^{k} P(A = i) \cdot P(B = k - i)$$
$$= \sum_{i=0}^{k} \binom{5}{i} 0.1^i \cdot 0.9^{k-i} \cdot \binom{8}{k-i} 0.5^{k-i} \cdot 0.5^{8-k-i}$$

Sometimes the PMF is zero

```
def main():
  for k in range(0, 5+8+1):
    pr_k = get_prob_sum(k)
    print(f"{k},{pr_k}")

def get_prob_sum(k):
  A = stats.binom(5, 0.1)
  B = stats.binom(8, 0.5)
  pr = 0
  for i in range(0, k+1):
    pr += A.pmf(i) * B.pmf(k-i)
  return pr
```

Discrete Vs Continuous

Discrete

$$P(X+Y=a) = \sum_{y=-\infty}^{\infty} P(X=a-y)P(Y=y)\ dy$$

Continuous

$$f(X+Y=a) = \int_{y=-\infty}^{\infty} f(X=a-y)f(Y=y)\ dy$$

Infinity is necessary when the values can be negative

Convolution: The fanciest way to say “adding random variables”

Side Quest
Sometimes Adding is Easy:

Sum of Independent Binomials

- Let X and Y be independent binomials with the same value for *p*:
  - $X \sim Bin(n_1, p)$ and $Y \sim Bin(n_2, p)$
  - $X + Y \sim Bin(n_1 + n_2, p)$

- Intuition:
  - X has n_1 trials and Y has n_2 trials
    - Each trial has same "success" probability p
  - Define Z to be $n_1 + n_2$ trials, each with success prob. p
  - $Z \sim Bin(n_1 + n_2, p)$, and also $Z = X + Y$

Sum of Independent Poissons

- Let X and Y be independent random variables
  - $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$
  - $X + Y \sim \text{Poi}(\lambda_1 + \lambda_2)$

Sum of Independent Normals

- Let X and Y be independent random variables
  - $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$
  - $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

- Generally, have n independent random variables $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, ..., n$:

$$\left(\sum_{i=1}^{n} X_i\right) \sim N\left(\sum_{i=1}^{n} \mu_i,\ \sum_{i=1}^{n} \sigma_i^2\right)$$

Linear Transform

Thinking of Y as a linear transform

$$X \sim N(\mu, \sigma^2)$$

$$Y = X + X = 2 \cdot X$$

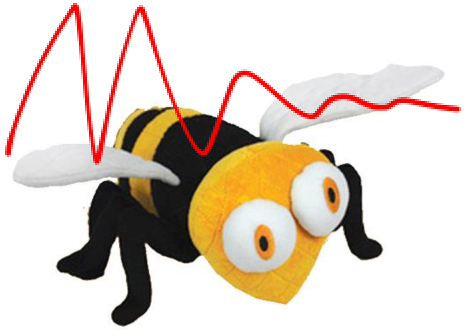
$$Y \sim N(2\mu, 4\sigma^2)$$

Thinking of Y as the sum of independent normals

$$Y = X + X = 2 \cdot X$$

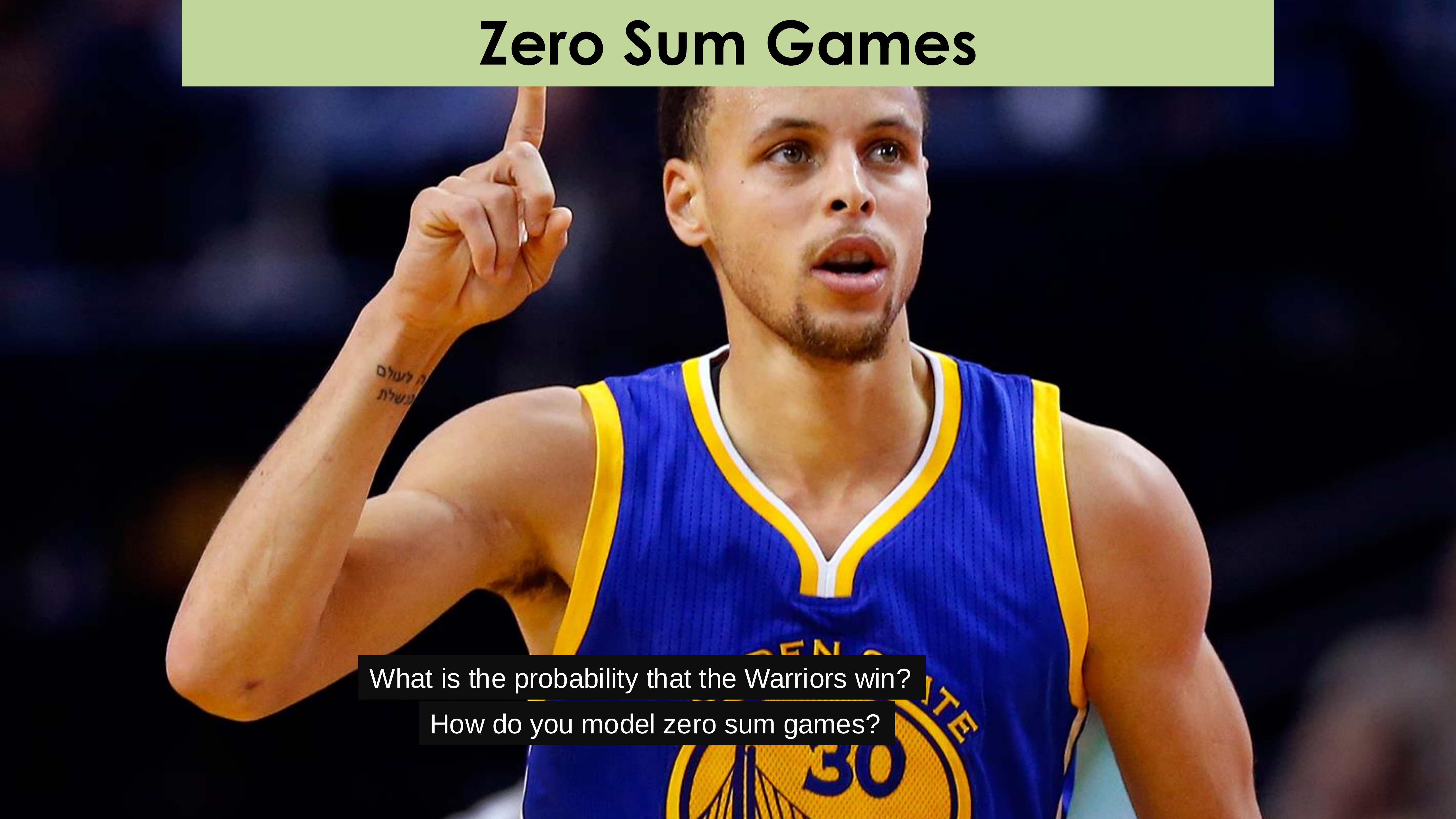
$$X + X \sim N(\mu + \mu, \sigma^2 + \sigma^2)$$

$$Y \sim N(2\mu, 2\sigma^2)$$



X is not independent of X

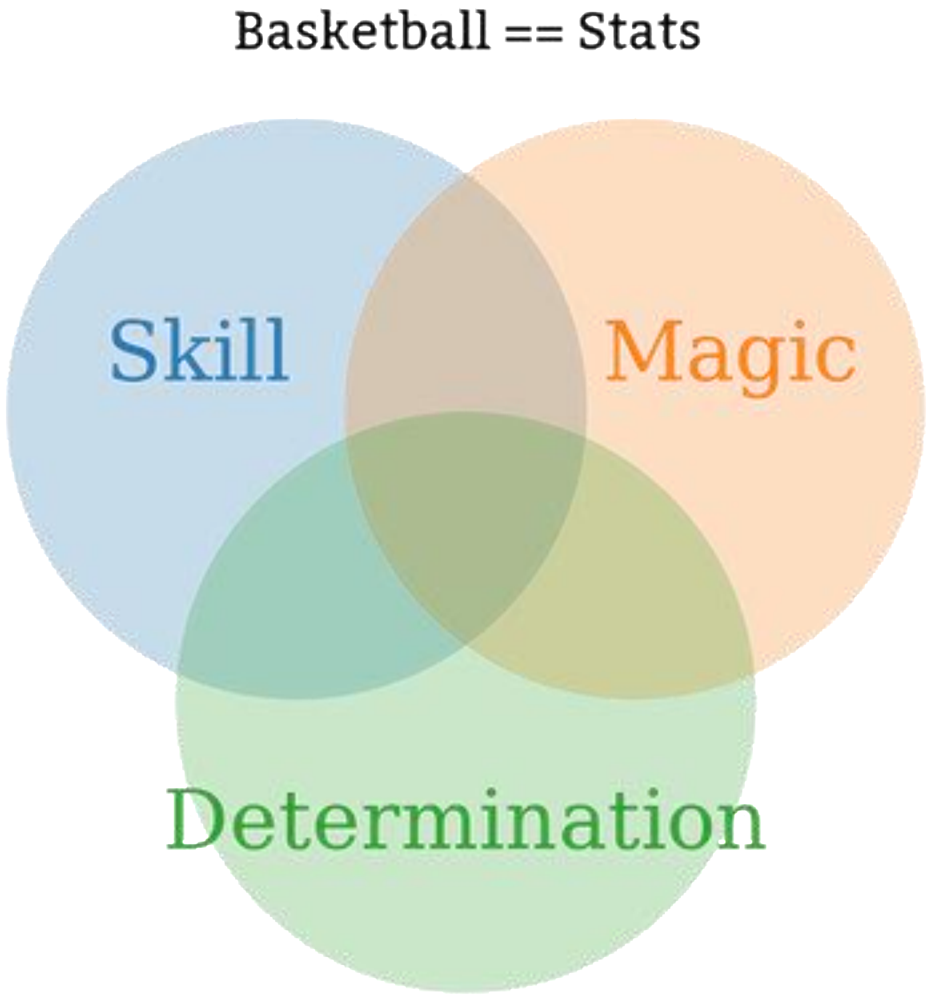
Zero Sum Games



What is the probability that the Warriors win?

How do you model zero sum games?

Gaussian Sampling and ELO ratings



What is the probability that the Warriors win?
How do you model zero-sum games?

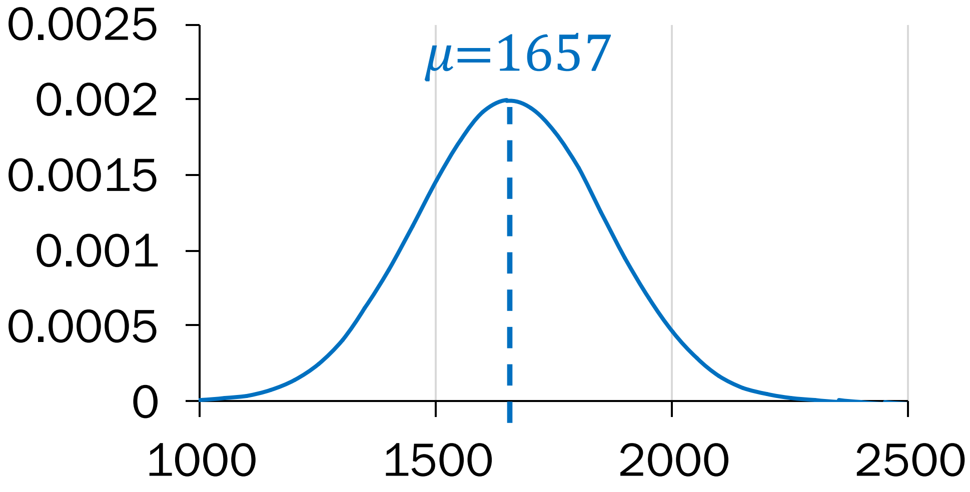
Gaussian Sampling and ELO ratings

Each team has an ELO score S, calculated based on its past performance.

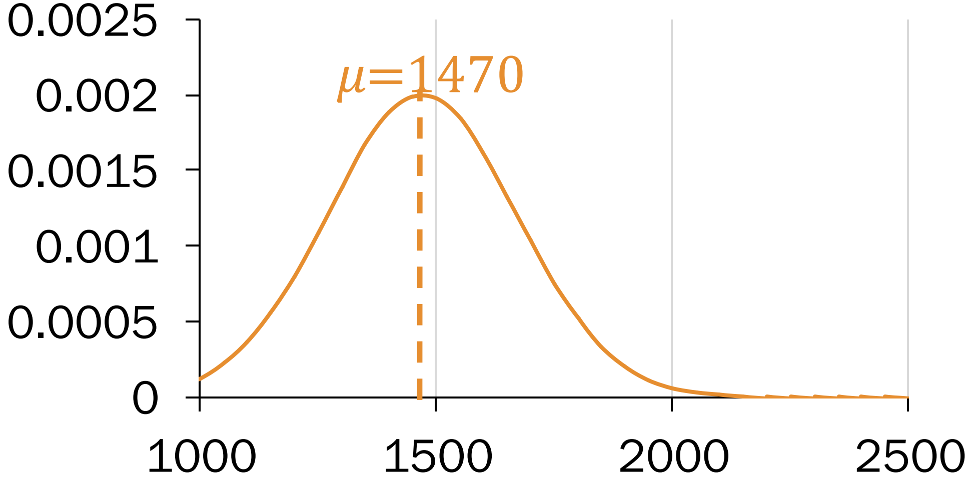
- Each game, a team has ability $A \sim \mathcal{N}(S, 200^2)$.
- The team with the higher sampled ability wins.

Arpad Elo

Warriors' $A_W \sim \mathcal{N}(S = 1657, 200^2)$



Opponent's $A_O \sim \mathcal{N}(S = 1470, 200^2)$



What is the probability that Warriors win this game?

Want: $P(\text{Warriors win}) = P(A_W > A_O)$

Probability of Winning a Game

$$A_W \sim N(1797, 200^2)$$

$$A_O \sim N(1555, 200^2)$$

$$P(\text{Warriors win}) = P(A_W > A_O)$$

$$P(\text{Warriors win}) = P(A_W - A_O > 0)$$

$$-A_O \sim N(-1555, 200^2)$$

$$D = A_W + (-A_O)$$

$$D \sim N(242, 2 \cdot 200^2)$$

$$P(D > 0) = 1 - F_D(0) \approx 0.804$$

Virus Infections Revisited

- Say you are working with the WHO to plan a response to a the initial conditions of a virus:
  - Two exposed groups
  - P1: 50 people, each independently infected with $p = 0.1$
  - P2: 100 people, each independently infected with $p = 0.4$
  - Question: Probability of more than 40 infections?

Virus Infections Revisited

- Say you are working with the WHO to plan a response to a the initial conditions of a virus:
  - Two exposed groups
  - P1: 50 people, each independently infected with p = 0.1
  - P2: 100 people, each independently infected with p = 0.4
  - A = # infected in P1 A ~ Bin(50, 0.1) ≈ X ~ N(5, 4.5)
  - B = # infected in P2 B ~ Bin(100, 0.4) ≈ Y ~ N(40, 24)
  - What is P(≥ 40 people infected)?
  - P(A + B ≥ 40) ≈ P(X + Y ≥ 39.5)
  - X + Y = W ~ N(5 + 40 = 45, 4.5 + 24 = 28.5)

$$P(W > 39.5) = 1 - P(W < 39.5)$$
$$= 1 - F_W(39.5)$$

$$= 1 - \Phi\left(\frac{39.5 - 45}{\sqrt{28.5}}\right) \approx 0.8485$$

End Side Quest

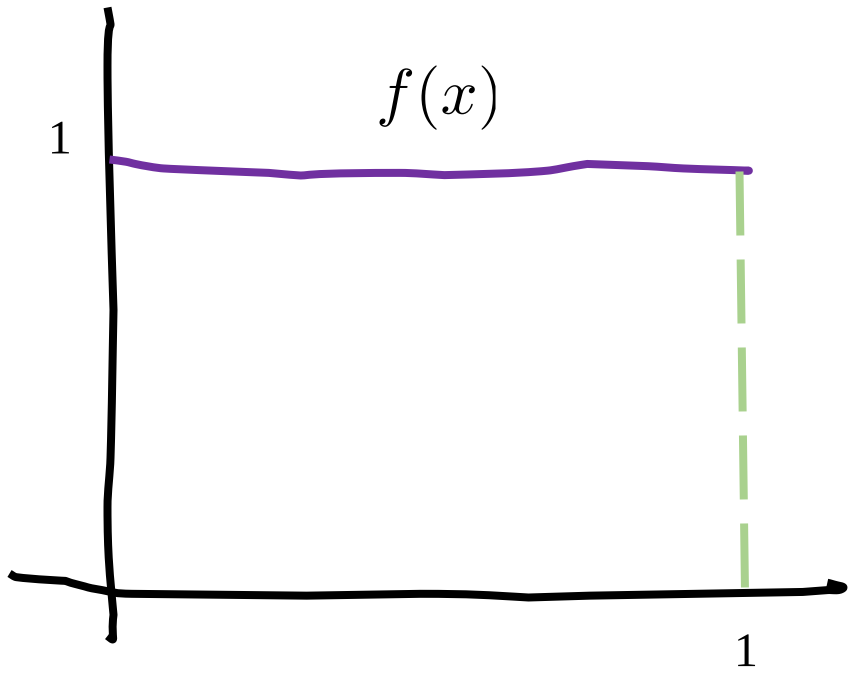
Sometimes Adding is Easy:

We talked about sum of Binomial, Normal and Poisson…who’s missing from this party?

Uniform.

Sum of Independent Uniforms

- Let X and Y be independent random variables
  - X ~ Uni(0, 1) and Y ~ Uni(0, 1) → $f(x) = 1$ for $0 \leq x \leq 1$



For both X and Y

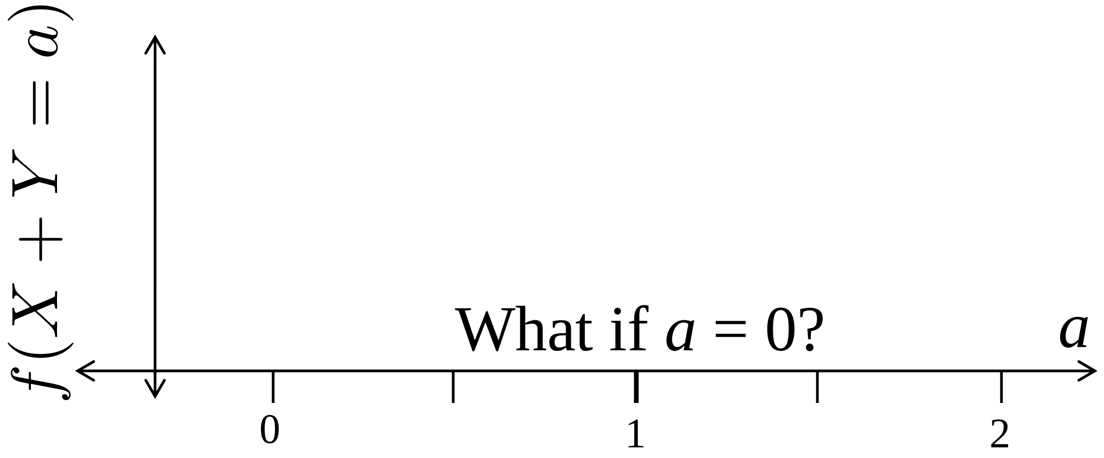
$f(X + Y = a)?$ $X \sim \text{Uni}(0, 1)$ $Y \sim \text{Uni}(0, 1)$

$f(X + Y = a)?$ $X \sim \text{Uni}(0, 1)$ $Y \sim \text{Uni}(0, 1)$

$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$

$f(X+Y=a)$? $X \sim \text{Uni}(0,1)$ $Y \sim \text{Uni}(0,1)$

$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$

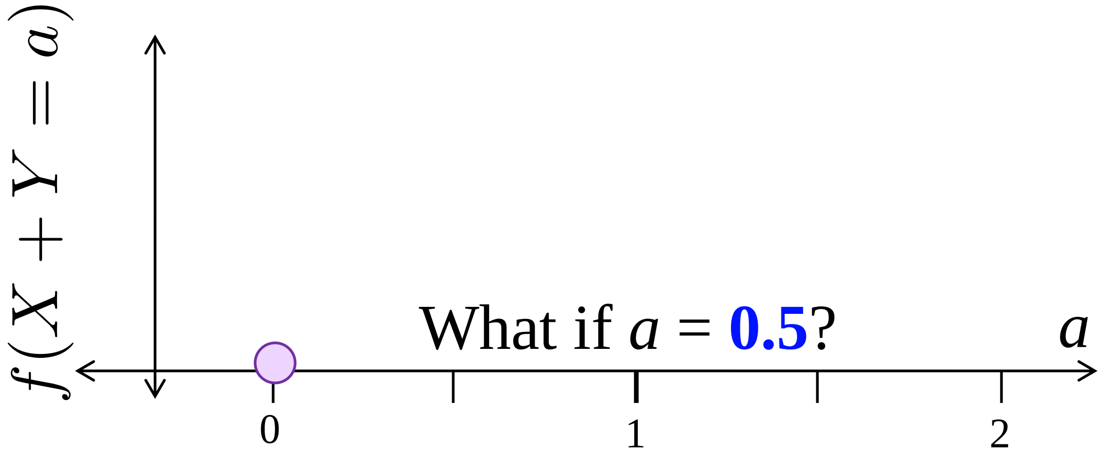


$f(X+Y=a)?$ $X \sim \text{Uni}(0,1)$ $Y \sim \text{Uni}(0,1)$

$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$

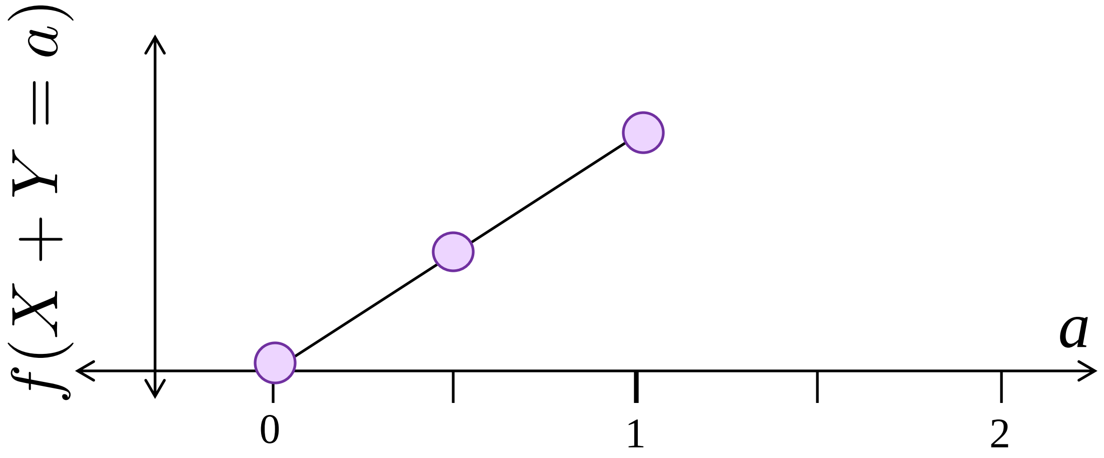
$f(X + Y = a)$? $X \sim \text{Uni}(0, 1)$ $Y \sim \text{Uni}(0, 1)$

$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$



$f(X+Y=a)$? $X \sim \text{Uni}(0,1)$ $Y \sim \text{Uni}(0,1)$

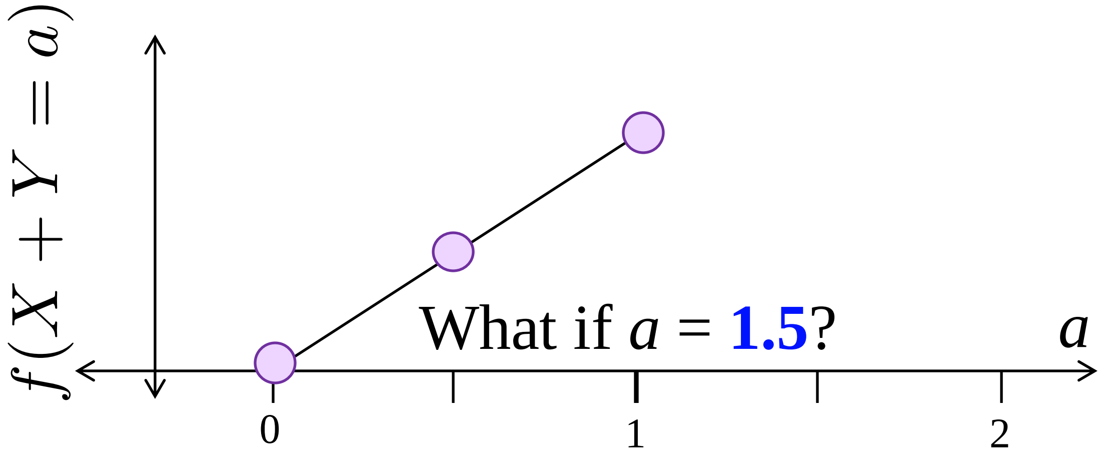
$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$



$f(X+Y=a)$?

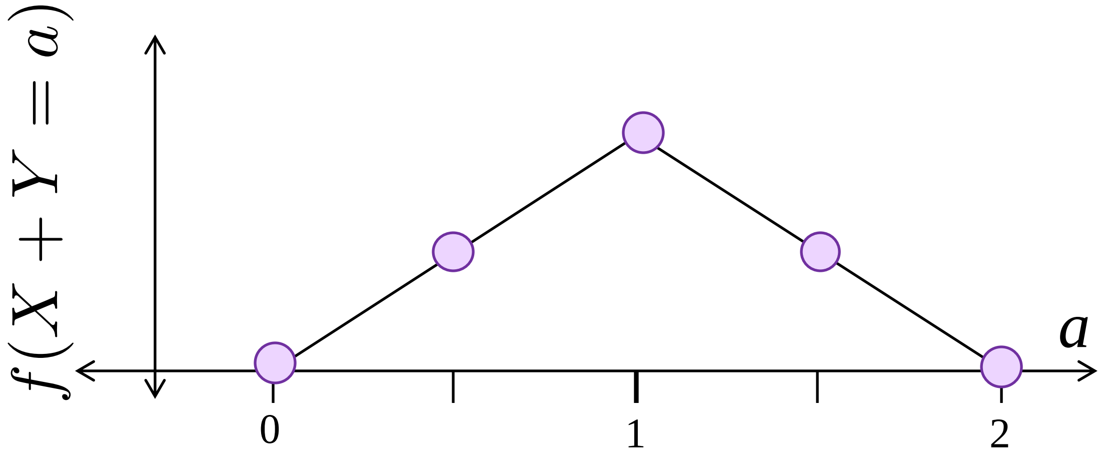
$X \sim \text{Uni}(0,1)$ $\quad$ $Y \sim \text{Uni}(0,1)$

$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$



$f(X+Y=a)$? $X \sim \text{Uni}(0,1)$ $Y \sim \text{Uni}(0,1)$

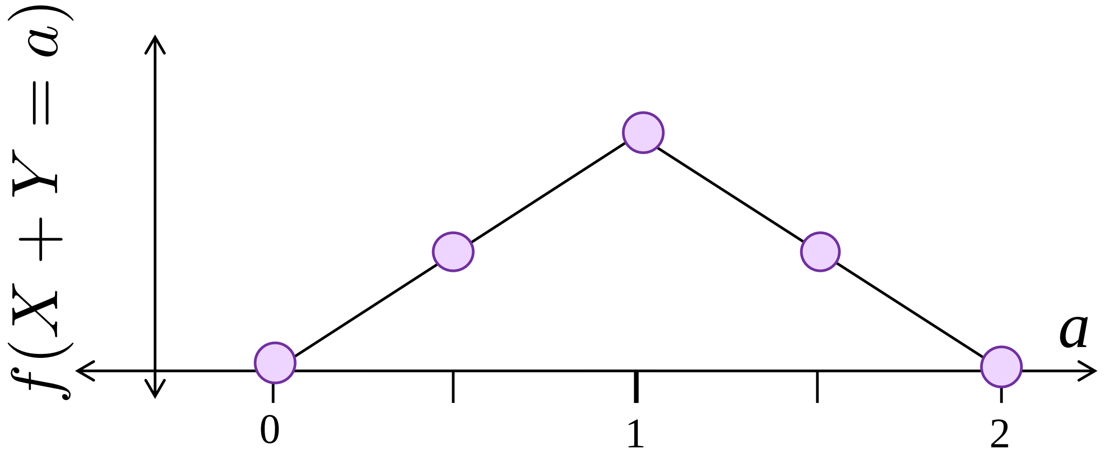
$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$



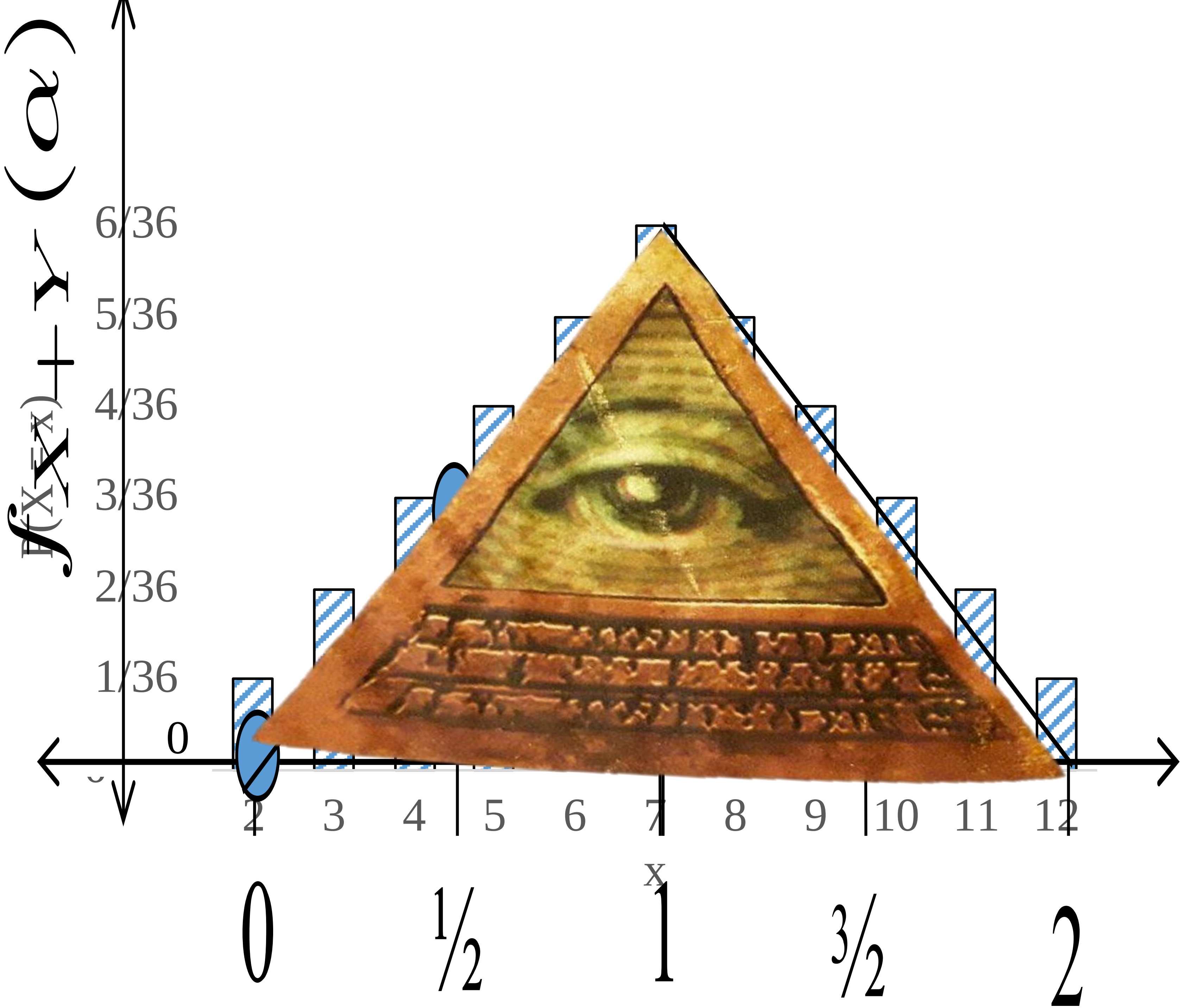
$$f(X+Y=a) = \begin{cases} a & 0 < a < 1 \\ 2-a & 1 < a < 2 \\ 0 & \text{otherwise} \end{cases}$$

$f(X + Y = a)$? $X \sim \text{Uni}(0, 1)$ $Y \sim \text{Uni}(0, 1)$

$$f(X+Y=a) = \int_{x=0}^{a} f(X=x)f(Y=a-x)\ \partial x$$



$$f(X+Y=a) = \begin{cases} a & 0 < a < 1 \\ 2-a & 1 < a < 2 \\ 0 & \text{otherwise} \end{cases}$$



$f_{X+Y}(\alpha)$
$f_{(X=x)}$
6/36
5/36
4/36
3/36
2/36
1/36
0
2 3 4 5 6 7 8 9 10 11 12
x
0 ½ 1 3/2 2

Gotta care about summing more than two things….

Sum of 100 uniforms???

Were talking about the sum of uniforms

sum.py

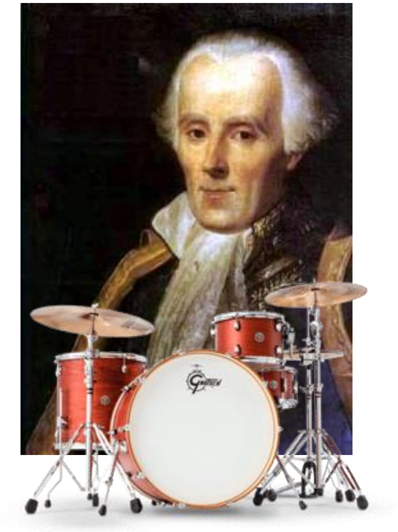
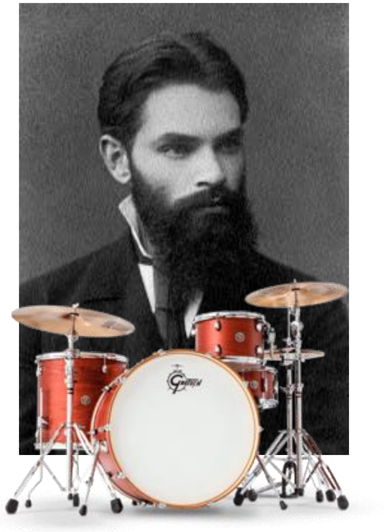
```python
import random

def main():
    x = random.random()
    y = random.random()
    z = x + y
    print(z)

if __name__ == '__main__':
    main()
```

Sum of 100 poissons???

And now a moment of silence...

...before we present...

...a beautiful result of probability theory!

(silent drumroll)

Central Limit Theorem

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

As $n \to \infty$

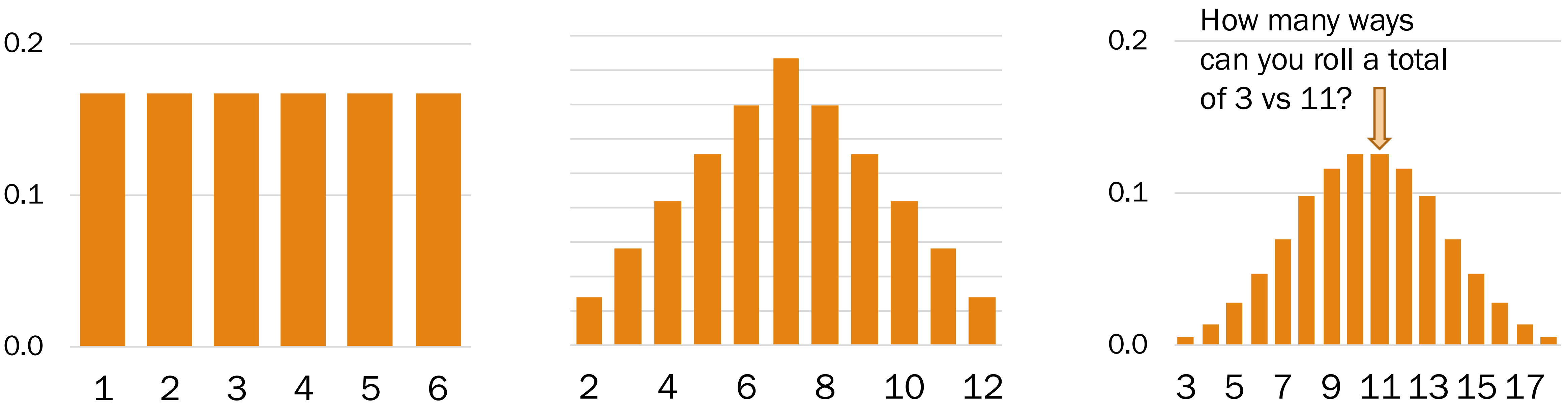
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

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

True happiness

Sum of dice rolls

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



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

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

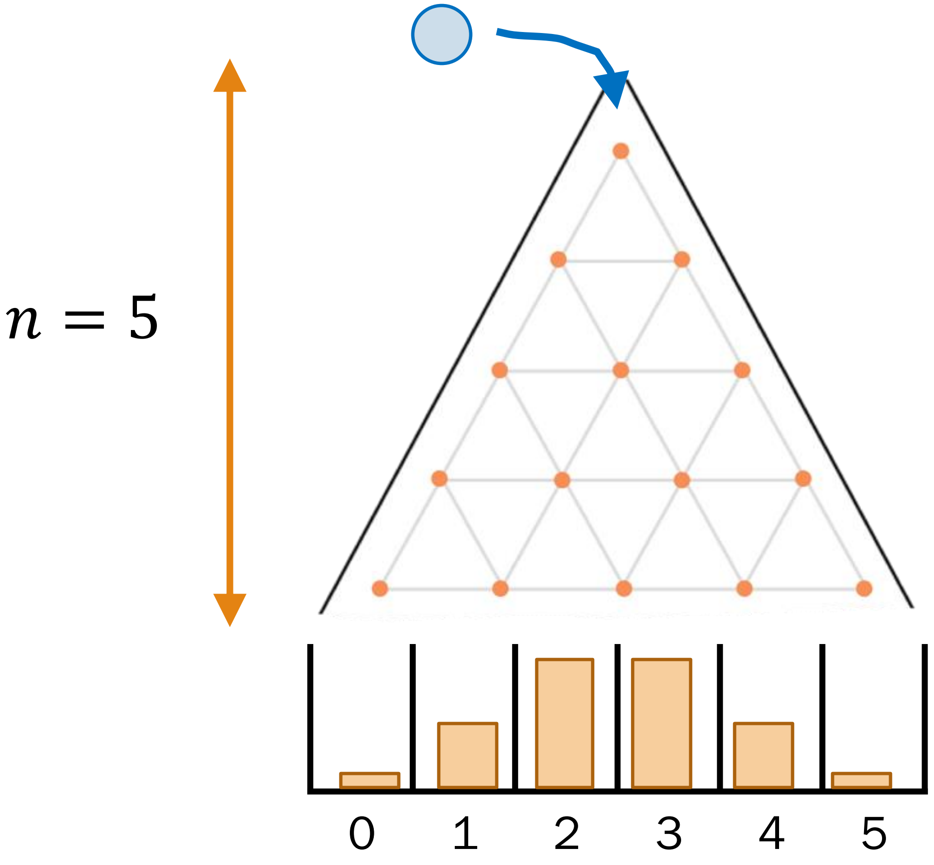
$\sum_{i=1}^{3} X_i$ Sum of 3 dice rolls

CLT explains a lot

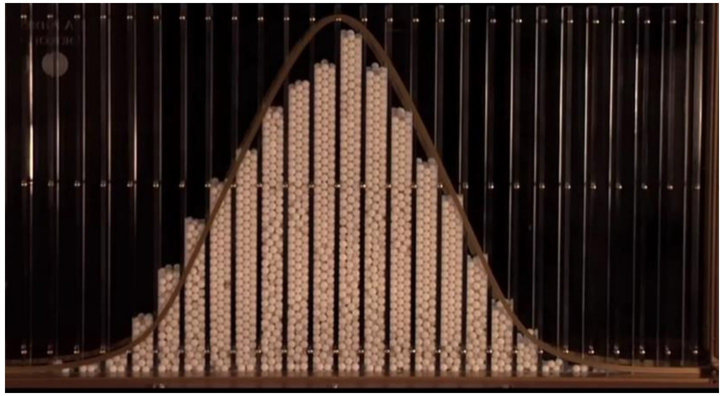
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

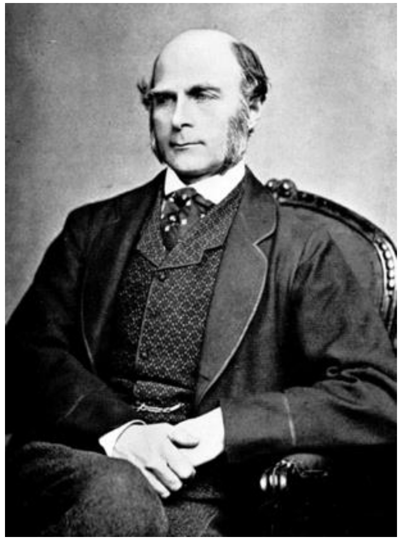
As $n \to \infty$

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)

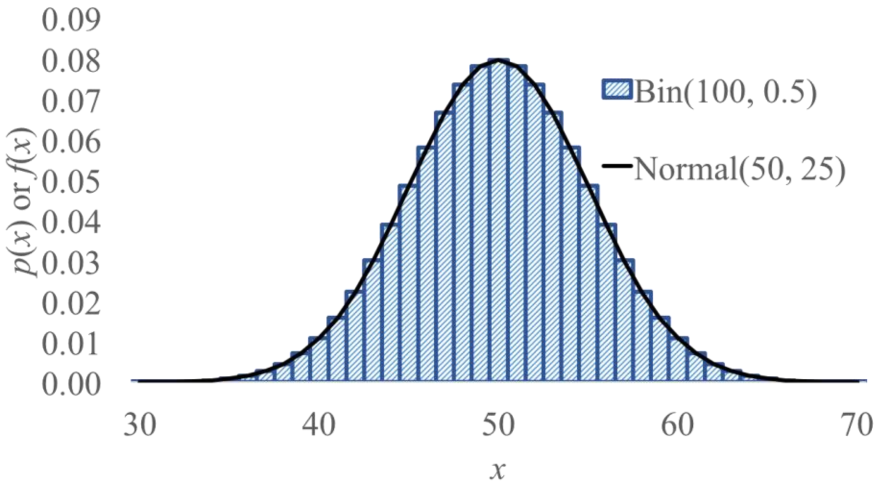




CLT explains a lot

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

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 ≈ Normal

Proof:

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

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

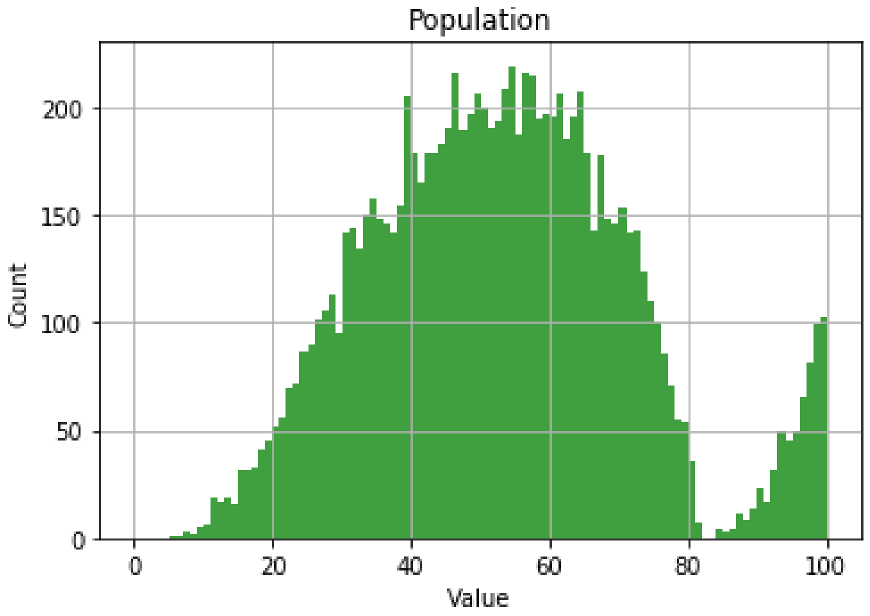
$$X \sim \mathcal{N}(n\mu, n\sigma^2)$$ (CLT, as $n \to \infty$)

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

CLT explains a lot

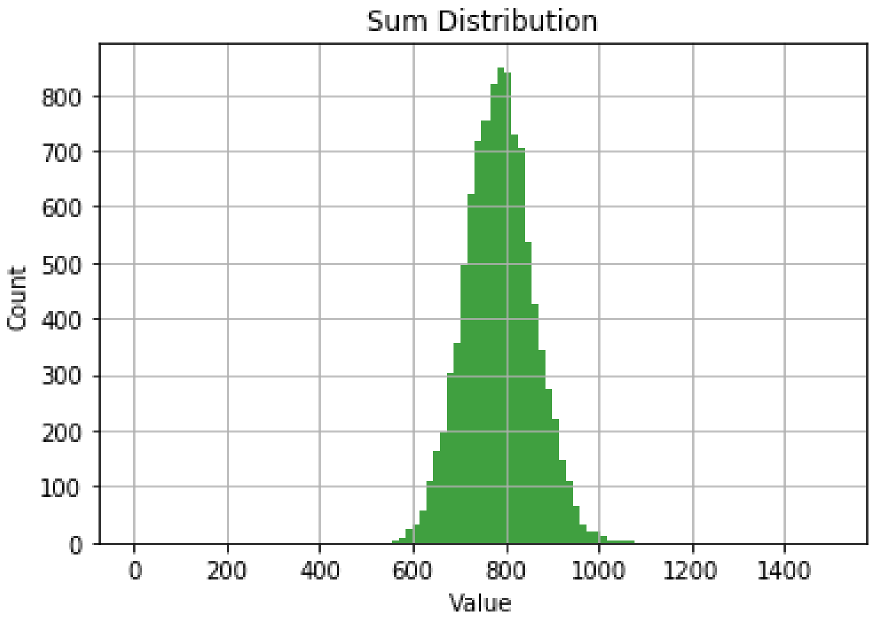
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \to \infty$$

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



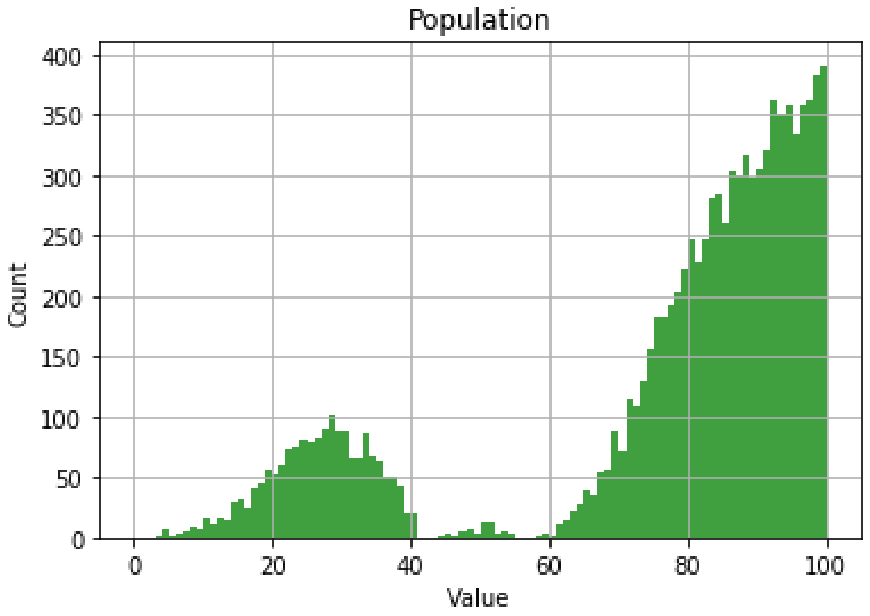
Distribution of $\sum_{i=1}^{15} X_i$

CLT explains a lot

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

As $n \to \infty$

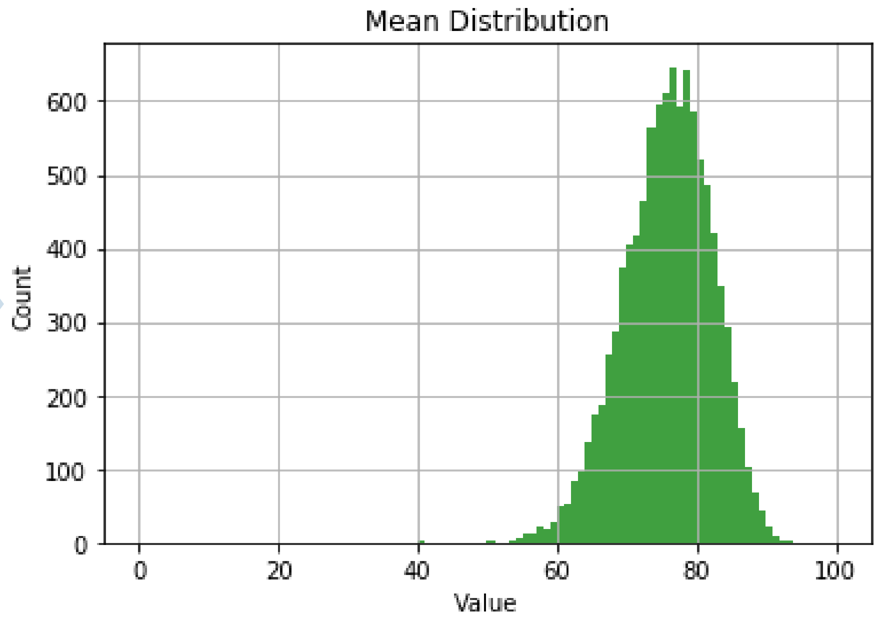
The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.



Distribution of X_i

Sample of size 15, **average** values

(sample mean)



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

Proof Outline of CLT

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

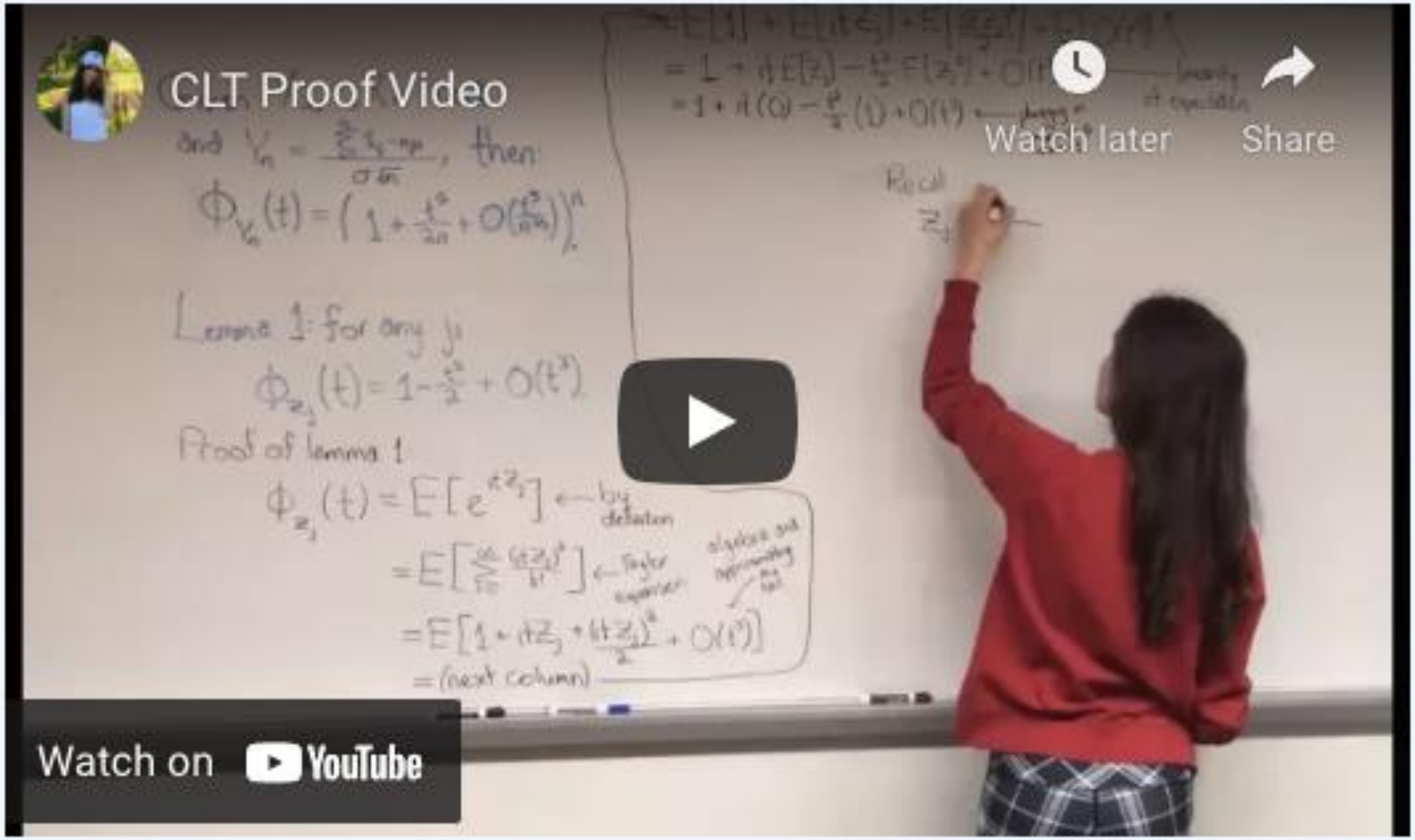
The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.

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 $n \to \infty$: $f(x) = e^{-\frac{x^2}{2}}$
- 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)

For Proof, See Video



CLT example

Sum of n independent Uniform RVs

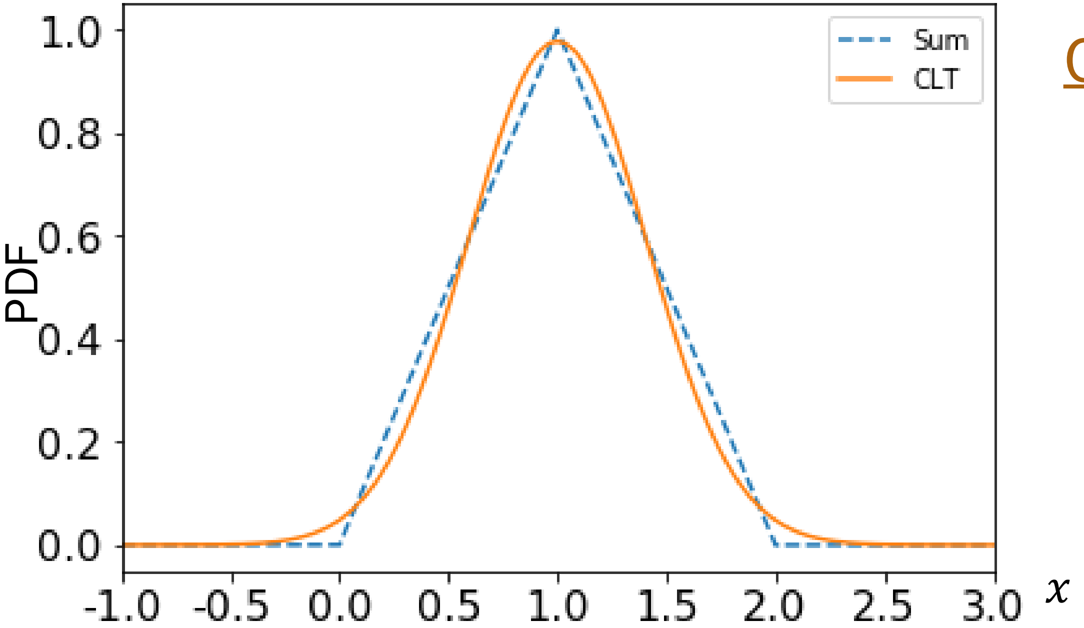
Let $X = \sum_{i=1}^{n} X_i$ be sum of i.i.d. RVs, where $X_i \sim \text{Uni}(0,1)$.

$\mu = E[X_i] = 1/2$

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

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

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

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

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

Sum of n independent Uniform RVs

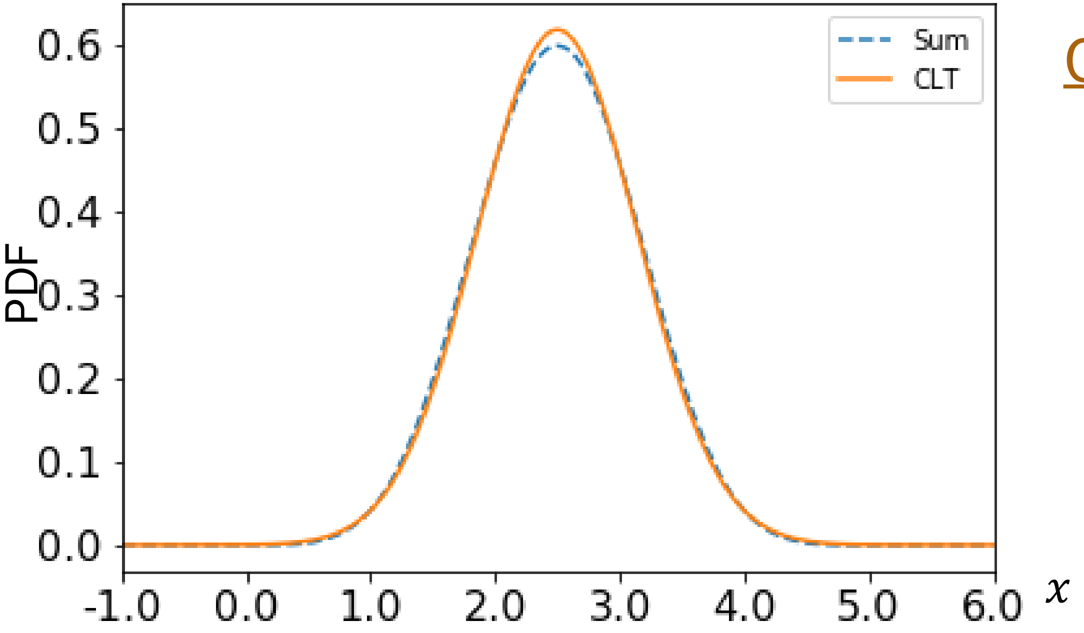
Let $X = \sum_{i=1}^{n} X_i$ be sum of i.i.d. RVs, where $X_i \sim \text{Uni}(0,1)$.

$\mu = E[X_i] = 1/2$

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

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

$n = 5$:



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

CLT approximation

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

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

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

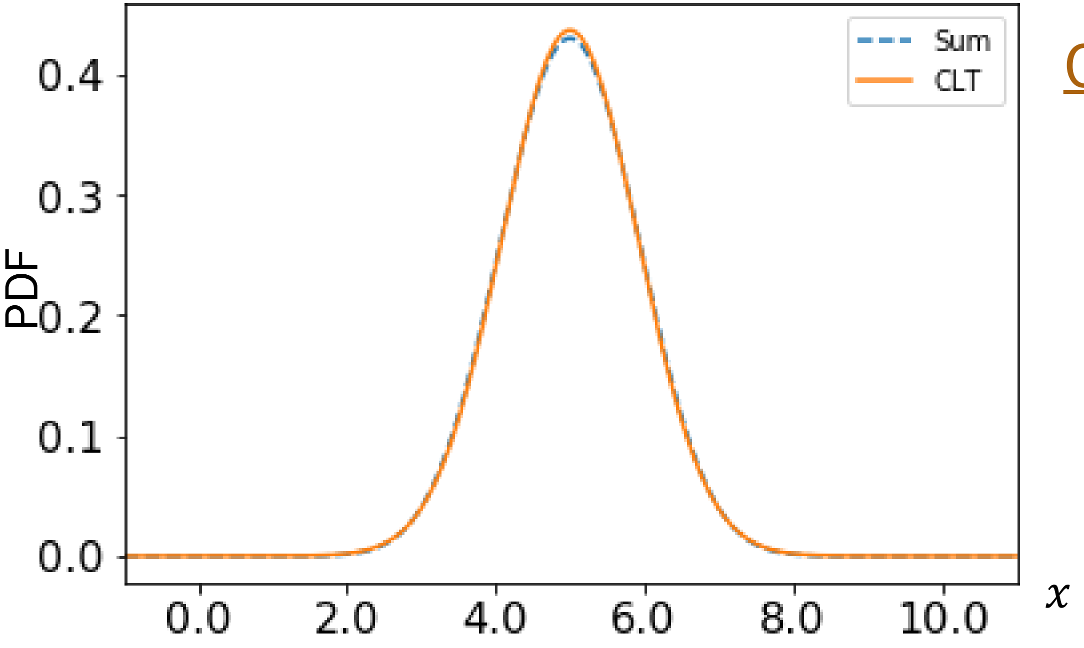
Sum of n independent Uniform RVs

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

CLT approximation

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

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

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

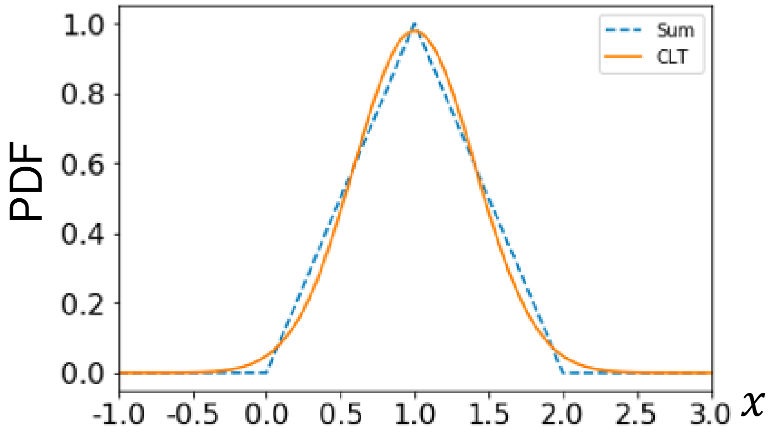
Sum of n independent Uniform RVs

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

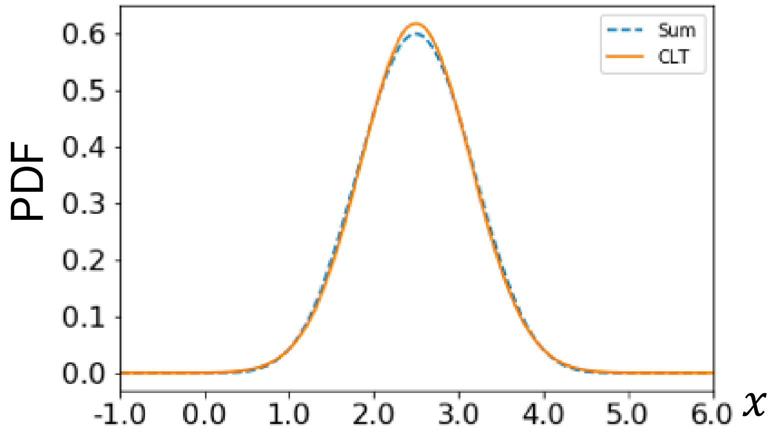
$\mu = E[X_i] = 1/2$
$\sigma^2 = \text{Var}(X_i) = 1/12$

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

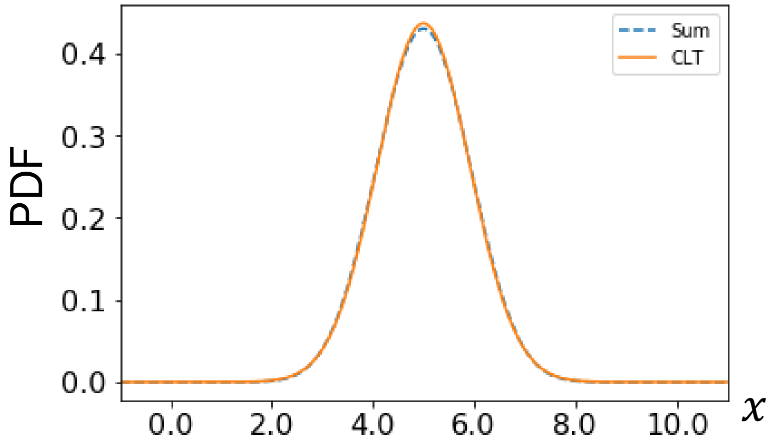
$n = 2$:



$n = 5$:



$n = 10$:



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 i.i.d. X_i's.

The sum of independent, identically distributed variables:

$$Y = \sum_{i=0}^{n} X_i$$

Is normally distributed:

$$Y \sim N(n\mu, n\sigma^2)$$

---

where $\mu = E[X_i]$

$$\sigma^2 = \text{Var}(X_i)$$

What about other functions?

Sum of iid? Normal

Average of iid?

Max of iid?

Average of iid?

Demo

http://onlinestatbook.com/stat_sim/sampling_dist/

It’s play time!

Sum of Dice

- You will roll 10 6-sided dice (X_1, X_2, …, X_{10})
  - X = total value of all 10 dice = $X_1 + X_2 + \ldots + X_{10}$
  - Win if: $X \leq 25$ or $X \geq 45$
  - Roll!

- And now the truth (according to the CLT)…

Sum of Dice

- You will roll 10 6-sided dice (X_1, X_2, …, X_{10})
  - X = total value of all 10 dice = $X_1 + X_2 + \ldots + X_{10}$
  - Win if: $X \leq 25$ or $X \geq 45$

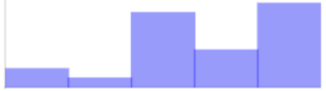
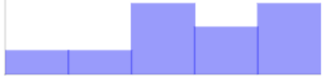
- Recall CLT: $X = \sum_{i}^{n} X_i \to N(n\mu, n\sigma^2)$ As $n \to \infty$
  - Determine $P(X \leq 25 \text{ or } X \geq 45)$ using CLT:

$$\mu = E[X_i] = 3.5 \qquad \sigma^2 = \text{Var}(X_i) = \frac{35}{12} \qquad X \approx N(35, 29.2)$$

$$1 - P(25.5 < X < 44.5) = 1 - P\left(\frac{25.5 - 35}{\sqrt{29.2}} < Z < \frac{44.5 - 35}{\sqrt{29.2}}\right)$$

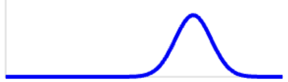
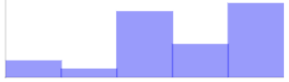
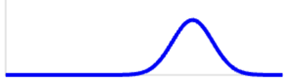
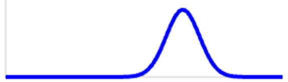
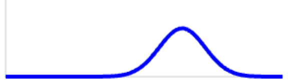
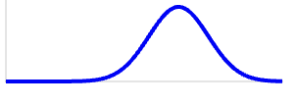
$$\approx 1 - (2\Phi(1.76) - 1) \approx 2(1 - 0.9608) = 0.0784$$

How Confident Are you that My Way is better than Yellow?

#	Song	Votes	NumVotes	SampleMean
1	My Way - Frank Sinatra		24	3.71
2	Yellow - Coldplay		23	3.7
3	Ma Meilleure Ennemie - Stromae, Pomme		16	3.56
4	Heavy - The Marias		20	3.55
5	Pluto Projector - Rex Orange County		10	3.5

Because of the CLT we can be a lot more thoughtful than this

How Confident Are you that My Way is better than Yellow?

#	Song	Sample Mean PDF	Votes	NumVotes	Pr(Top16)	Pr(Best)
1	My Way - Frank Sinatra			24	0.739	0.500
2	Yellow - Coldplay			23	0.709	0.490
3	Ma Meilleure Ennemie - Stromae, Pomme			16	0.530	0.336
4	Heavy - The Marias			20	0.507	0.352
5	Pluto Projector - Rex Orange County			10	0.460	0.338

Wonderful Form of Cosmic Order

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the ”[Central limit theorem]". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

- Sir Francis Galton

-