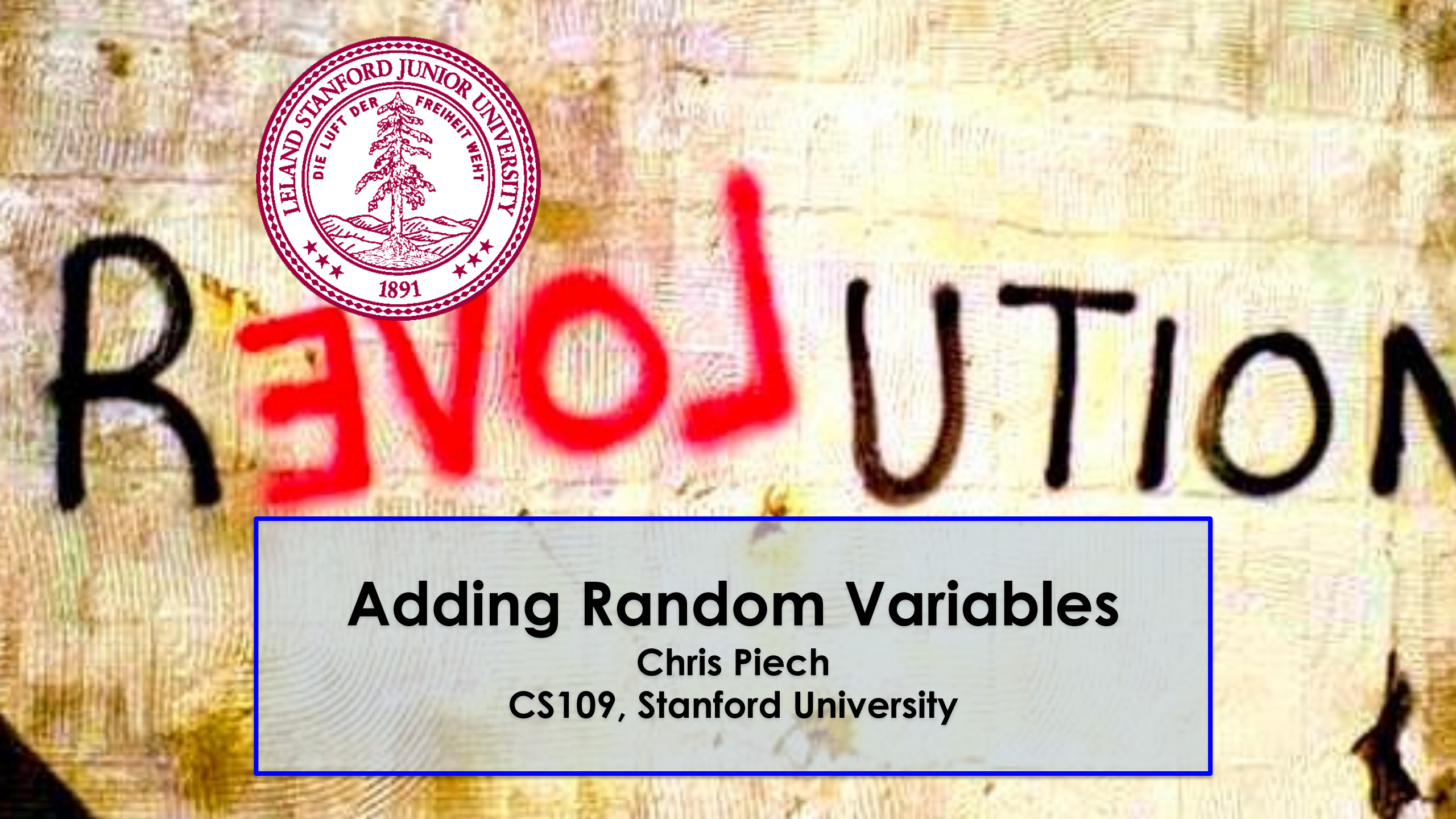
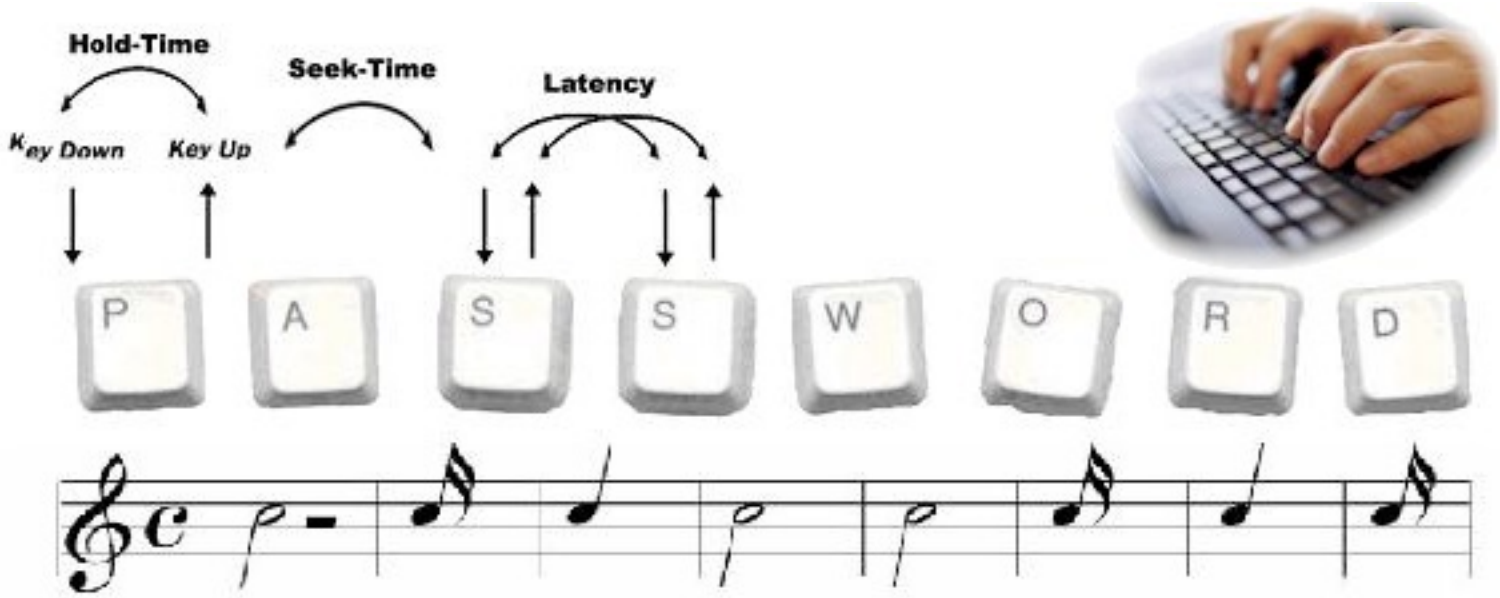


Adding Random Variables

Chris Piech
CS109, Stanford University

Biometric Keystrokes

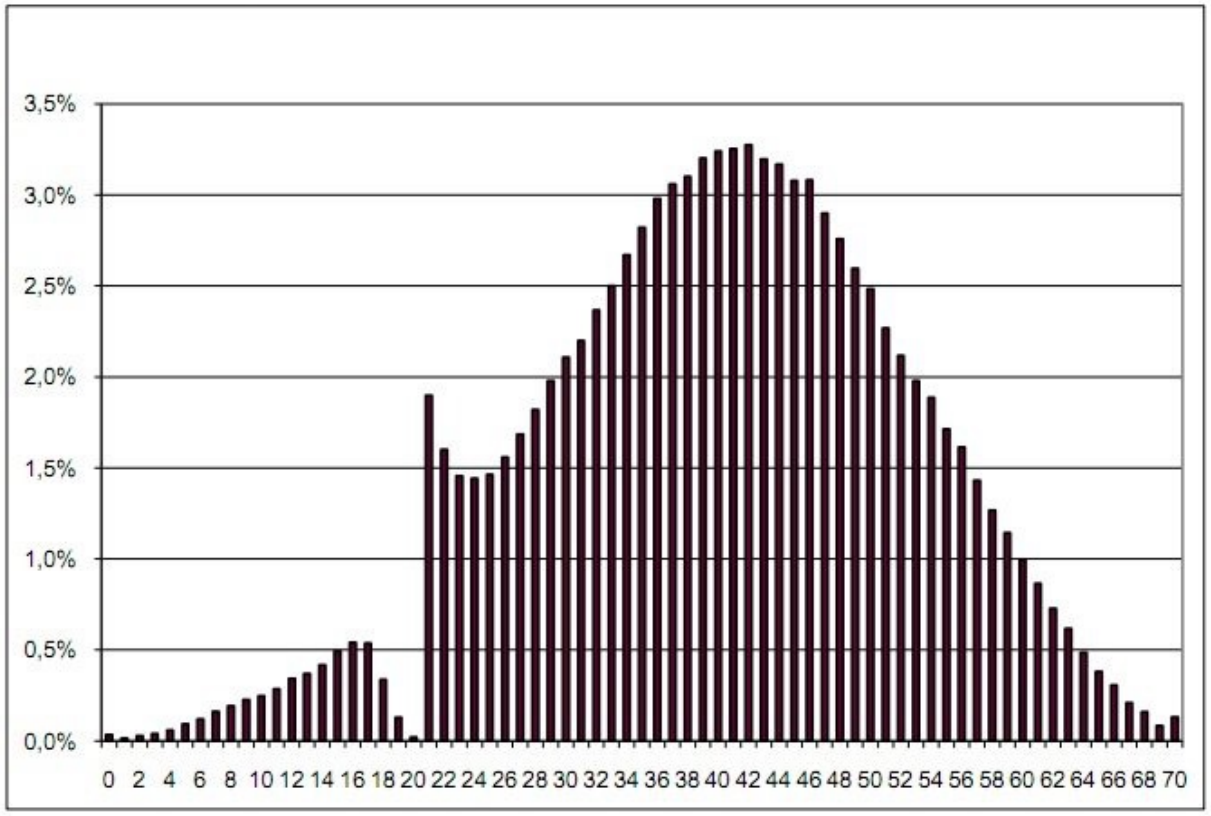


Altruism?

Scores for a standardized test that students in Poland are required to pass before moving on in school

See if you can guess the minimum score to pass the test.

2.1. Poziom podstawowy



Wykres 1. Rozkład wyników na poziomie podstawowym

http://freakonomics.com/2011/07/07/another-case-of-teacher-cheating-or-is-it-just-altruism/comment-page-2/

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

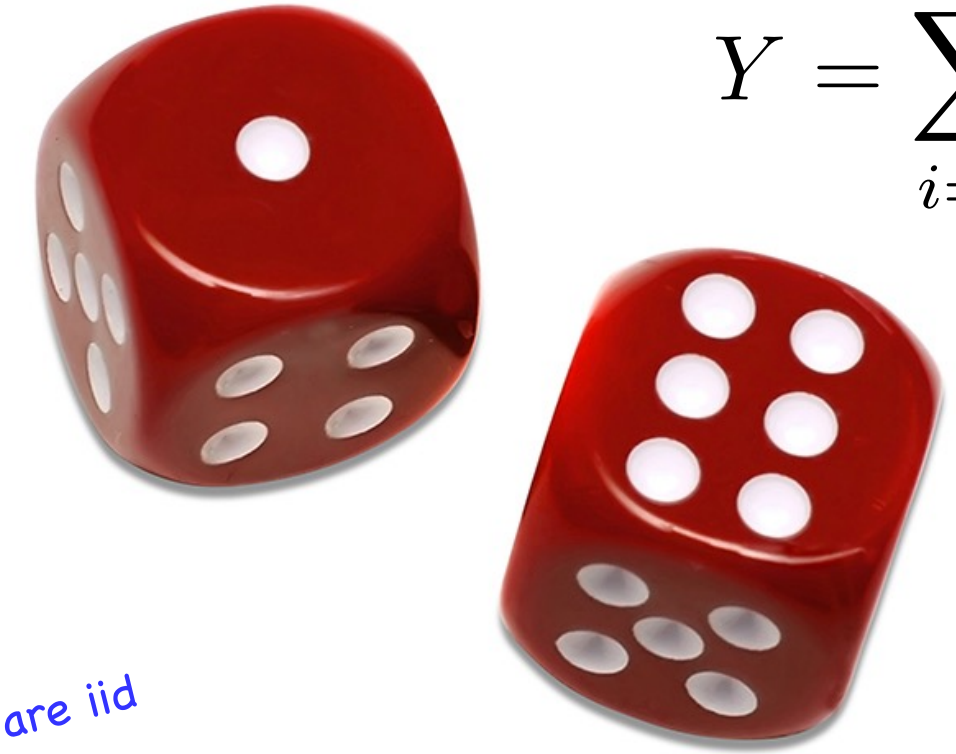
Quick check

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?

Sum of Two Dice

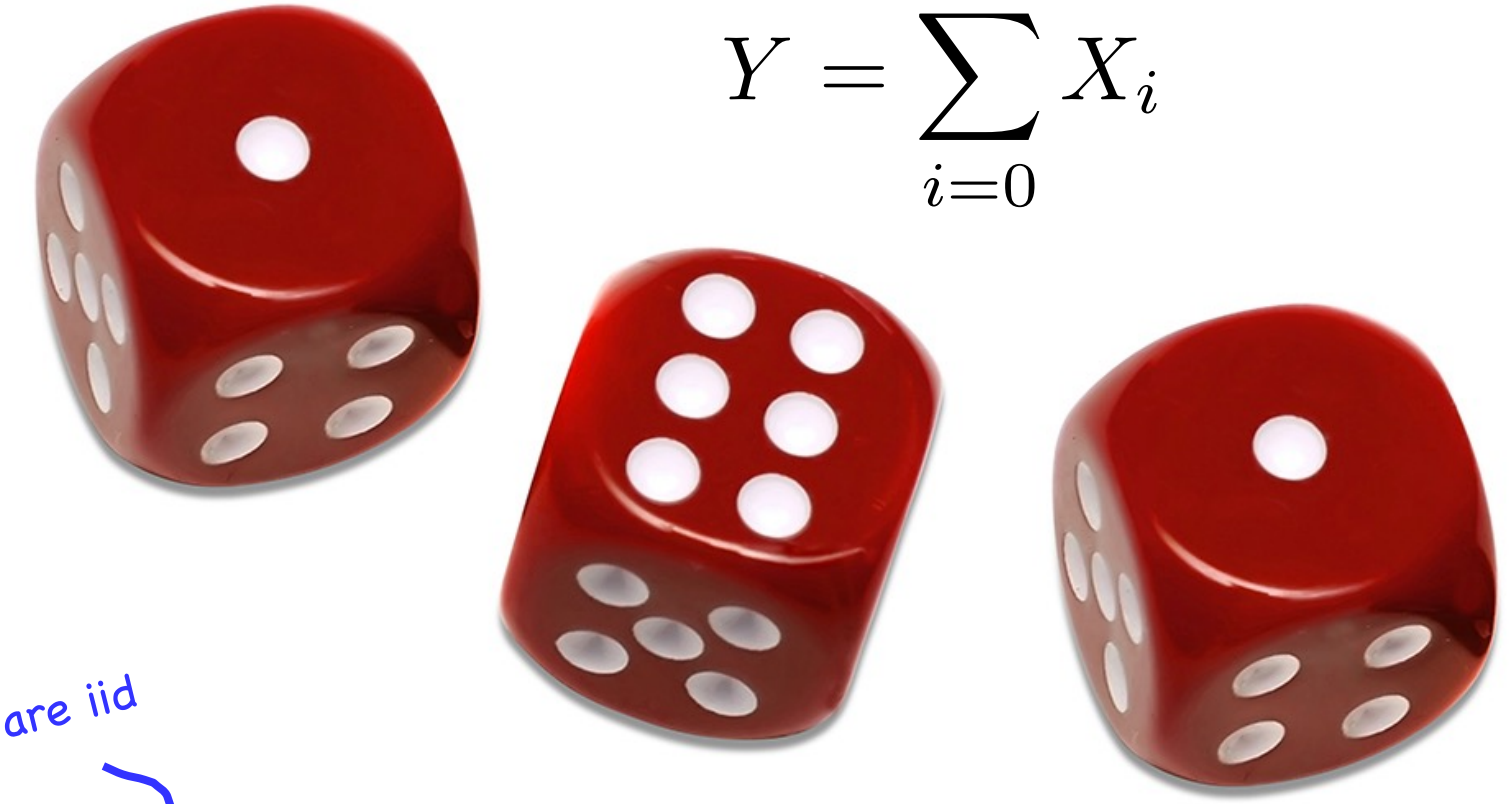


$$Y = \sum_{i=0}^{2} X_i$$

X_i s are iid

X_i is the outcome of dice roll i

Sum of Three Dice

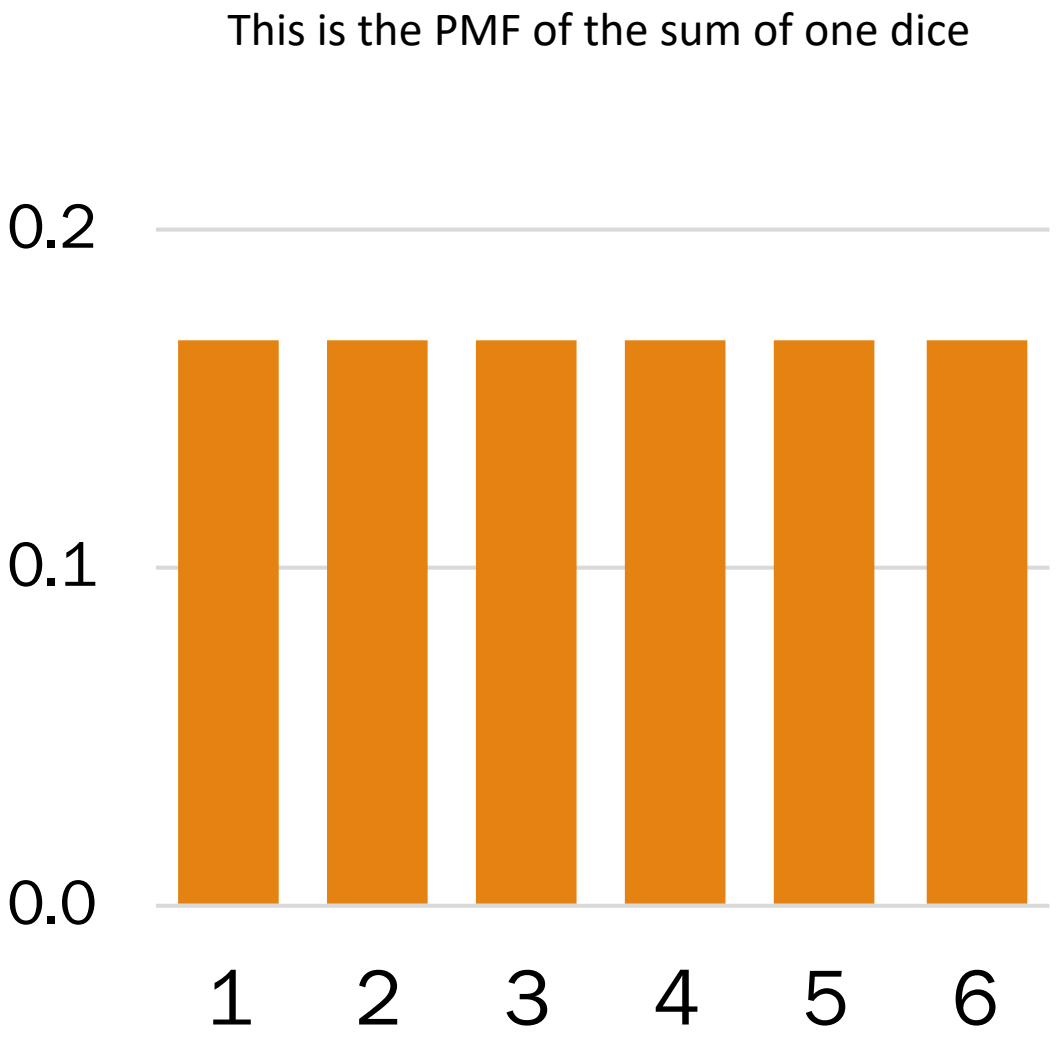


$$Y = \sum_{i=0}^{3} X_i$$

X_i s are iid

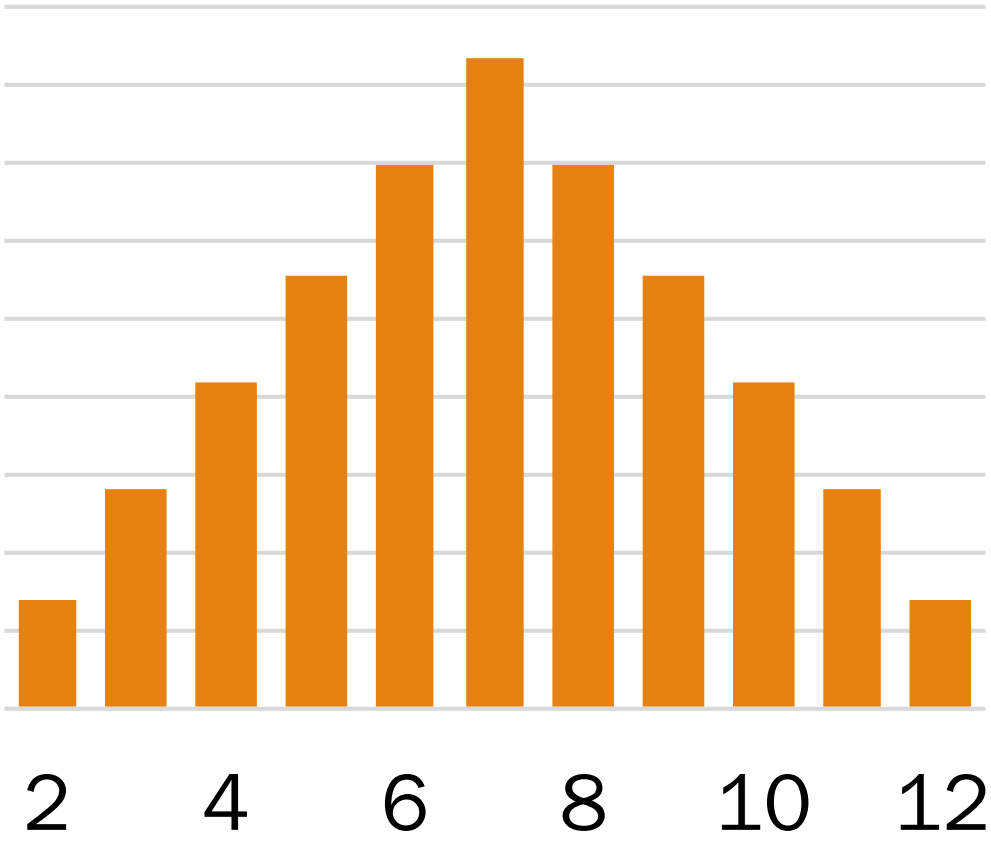
X_i is the outcome of dice roll i

Sum of One Dice



Sum of Two Dice

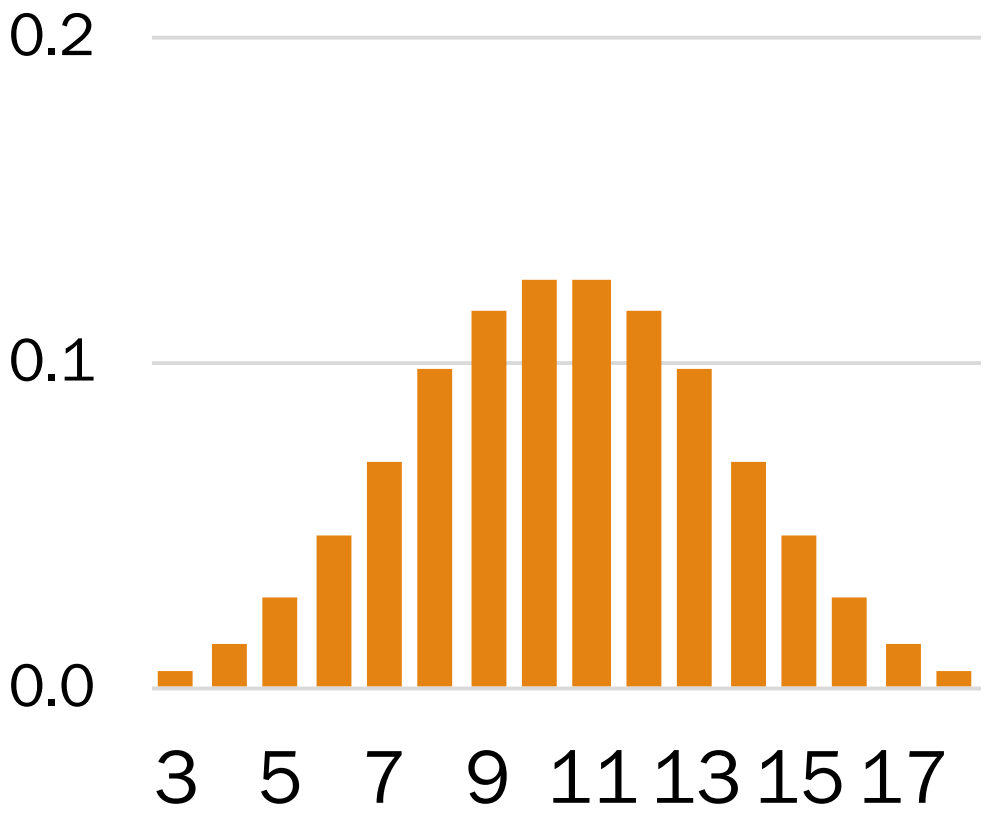
This is the PMF of the sum of two dice



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?

The 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?

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

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

The Insight to Convolution Proofs

What is the probability that X + Y = n?

$$P(X+Y=n)?$$

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

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	$P(X=n, Y=0)$

The Insight to Convolution Proofs

What is the probability that X + Y = n?

$$P(X+Y=n)?$$

Since this is the OR or mutually exclusive events

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

If the random variables are independent

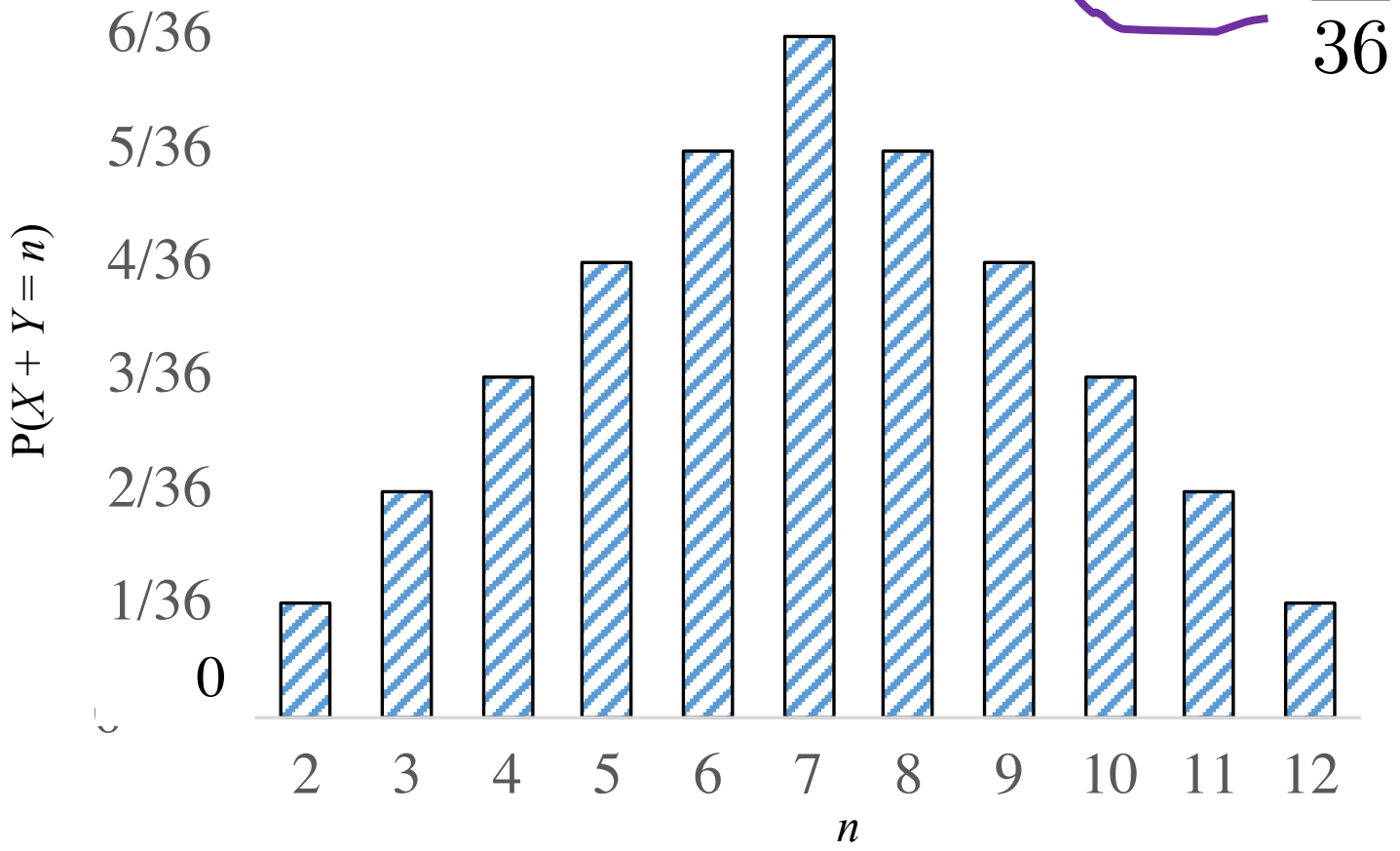
$$=\sum_{k=0}^{n} P(X=k)P(Y=n-k)$$

Sum of Two Dice

Let $X+Y$ be the value of the sum of two dice
(aka two independent random variables)

$$P(X+Y=n) = \sum_{i=1}^{n-1} \mathrm{P}(X=i, Y=n-i)$$

$$\frac{1}{36}$$



Why should you care?

Many Processes
(think binomial)

Comparing random variables

Zero Sum Games

What is the probability that the Warriors win?

How do you model zero sum games?

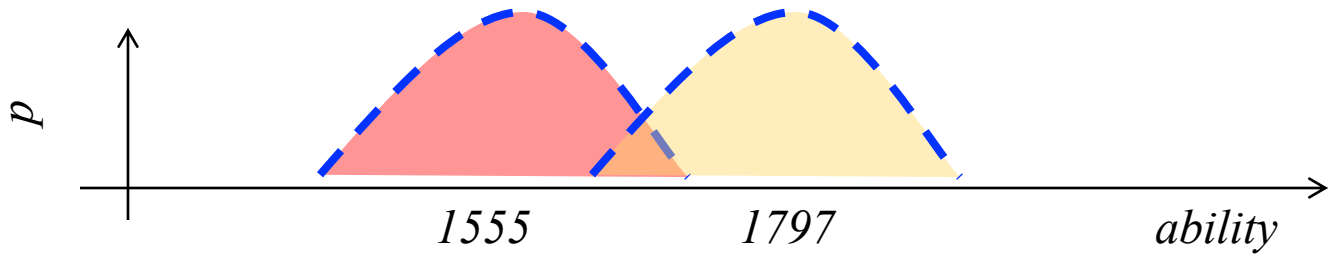
Motivating Idea: Zero Sum Games

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

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

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

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

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

In class we solved this by sampling. But that is a bit of a "cheat" and is computationally expensive.

Sums (or subtractions) of random variables show up all the time. But we have no explicit tools for dealing with them!

Challenge: try and come up with the way to solve this by the end of class

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 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 $i = 1, 2, \ldots, n$:

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

Sum of Independent Poissons

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

Virus Infections

- 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?

Sanity check: Should we use the Binomial Sum-of-RVs shortcut?

A. YES!
B. NO!
C. Other/none/more

Virus Infections

- 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 $\quad A \sim Bin(50, 0.1) \approx X \sim N(5, 4.5)$
  - B = # infected in P2 $\quad B \sim Bin(100, 0.4) \approx Y \sim N(40, 24)$
  - What is P(≥ 40 people infected)?
  - $P(A + B \geq 40) \approx P(X + Y \geq 39.5)$
  - $X + Y = W \sim N(5 + 40 = 45, 4.5 + 24 = 28.5)$

$$P(W \geq 39.5) = P\left(\frac{W - 45}{\sqrt{28.5}} > \frac{39.5 - 45}{\sqrt{28.5}}\right) = 1 - \Phi(-1.03) \approx 0.8485$$

Linear Transform

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

By Linear Transform

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

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

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

By Sum of Independent Normals

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

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

X is not independent of X

Sum of Independent Poissons

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

Binomial theorem:

$$(\lambda_1 + \lambda_2)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}$$

$$(x+y)^0 = 1,$$

$$(x+y)^1 = x+y,$$

$$(x+y)^2 = x^2 + 2xy + y^2,$$

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3,$$

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4,$$

https://en.wikipedia.org/wiki/Binomial_theorem

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

What is X + Y?

Poisson PMF

$$\mathrm{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Binomial Theorem

$$(\lambda_1 + \lambda_2)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}$$

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

Uniform.

Dance, Dance Convolution

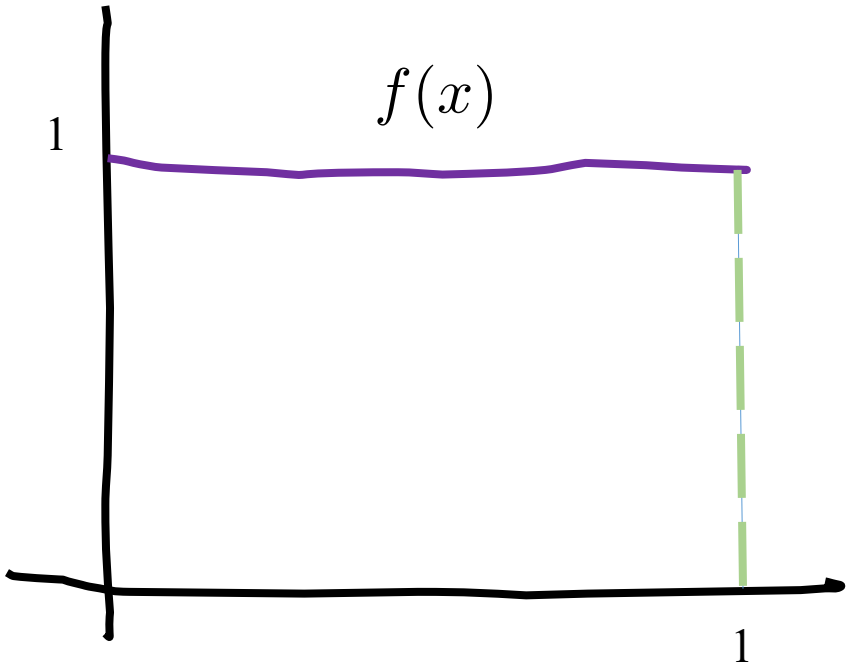
- Let X and Y be independent random variables
  - Probability Density Function (PDF) of X + Y:

$$f_{X+Y}(a) = \int_{y=-\infty}^{\infty} f_X(a-y)\, f_Y(y)\, dy$$

  - In discrete case, replace $\int_{y=-\infty}^{\infty}$ with $\sum_{y}$, and $f(y)$ with $p(y)$

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

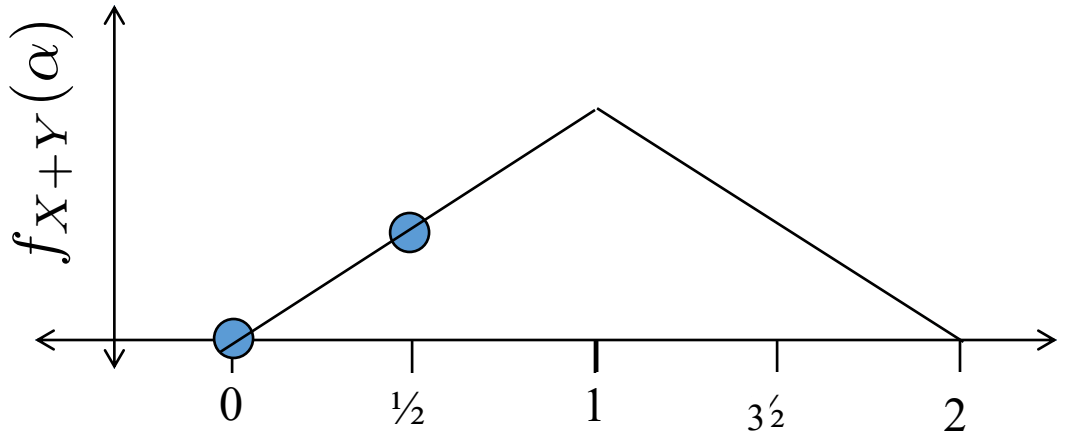
$1 < \alpha < 2$

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

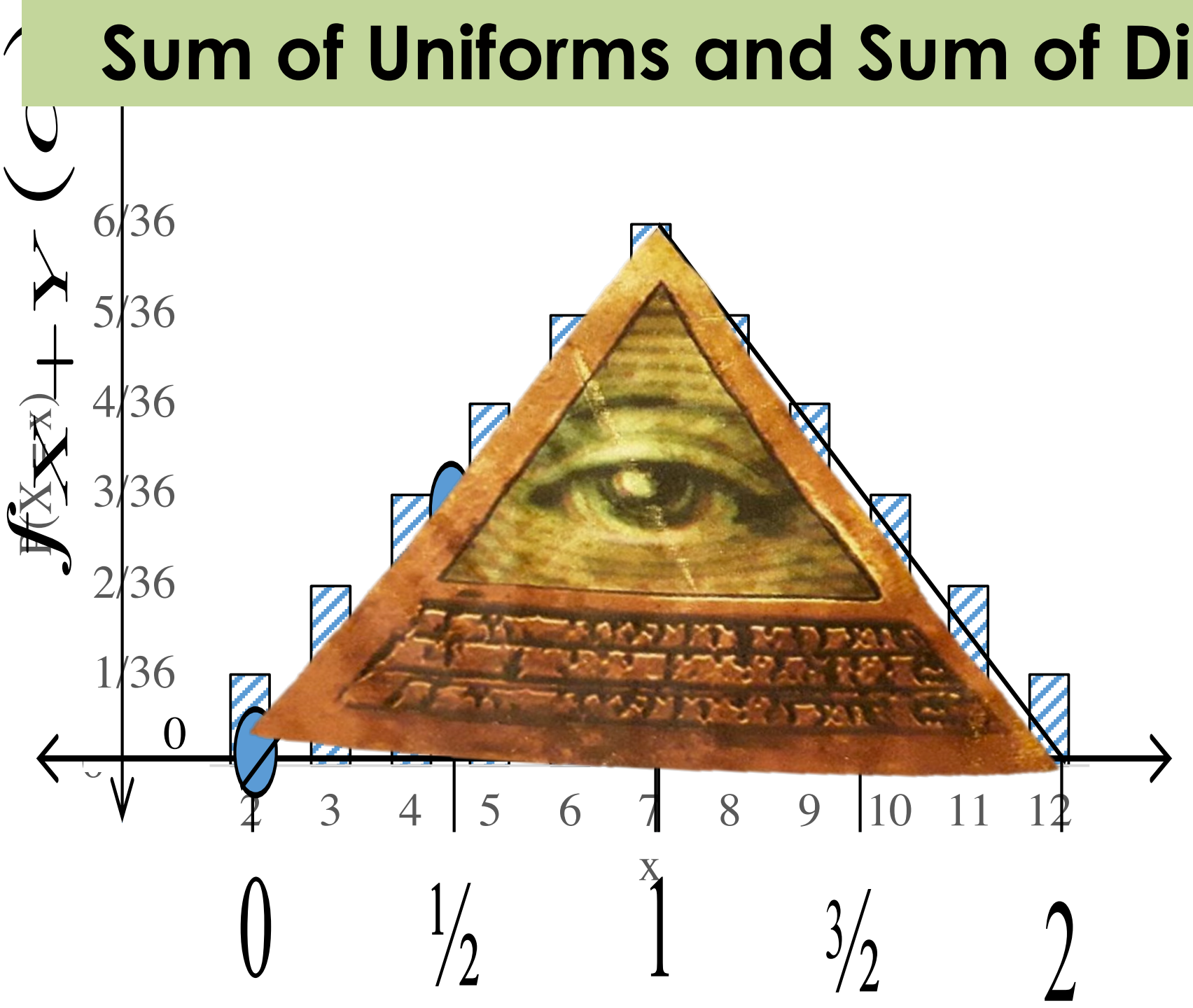
X and Y are independent

$f_{X+Y}(\alpha)?$

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



Sum of Uniforms and Sum of Dice



Sum of 100 uniforms???

Were talking about the sum of uniforms

sum.py

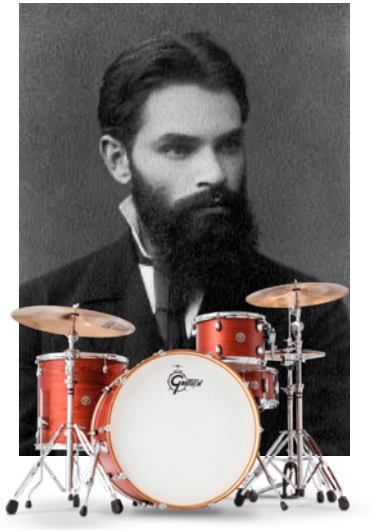
```
1  import random
2
3  def main():
4      x = random.random()
5      y = random.random()
6      z = x + y
7      print(z)
8
9  if __name__ == '__main__':
10     main()
```

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 $X_1, X_2, \ldots, X_n$ with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$.

As $n \to \infty$

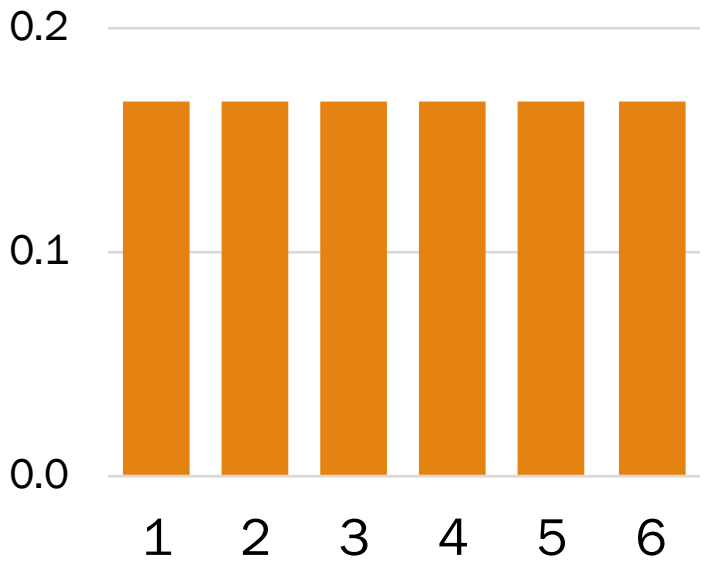
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

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

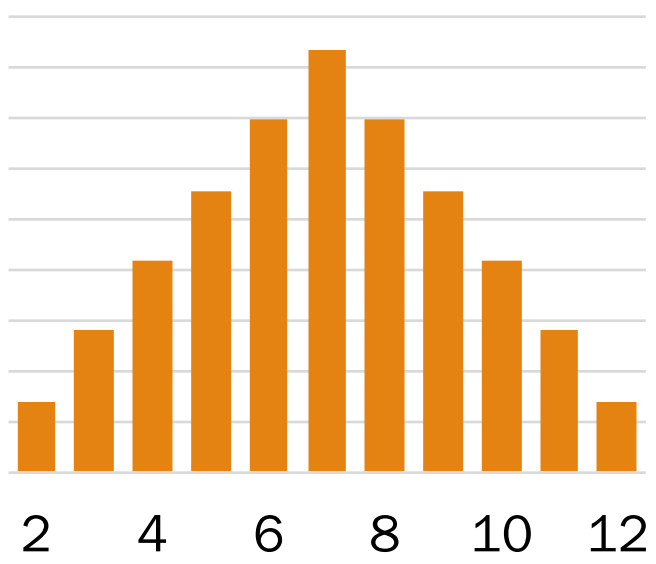
True happiness

Sum of dice rolls

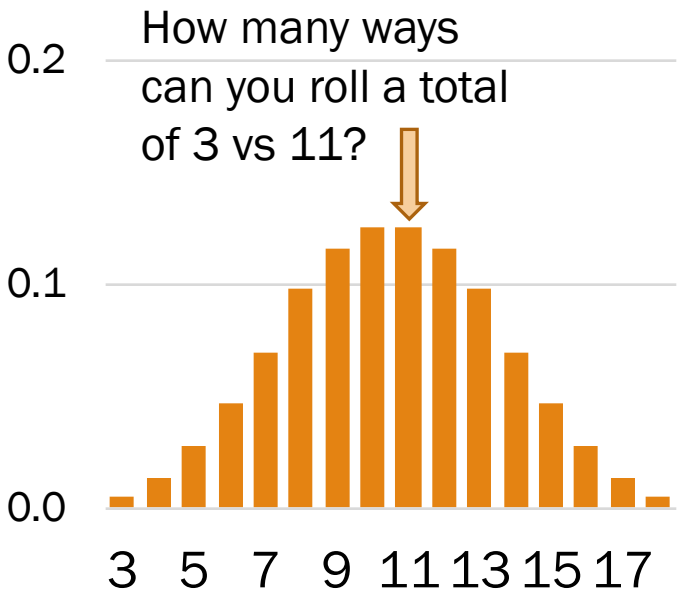
Roll n independent dice. Let X_i be the outcome of roll i. X_i are i.i.d.



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



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

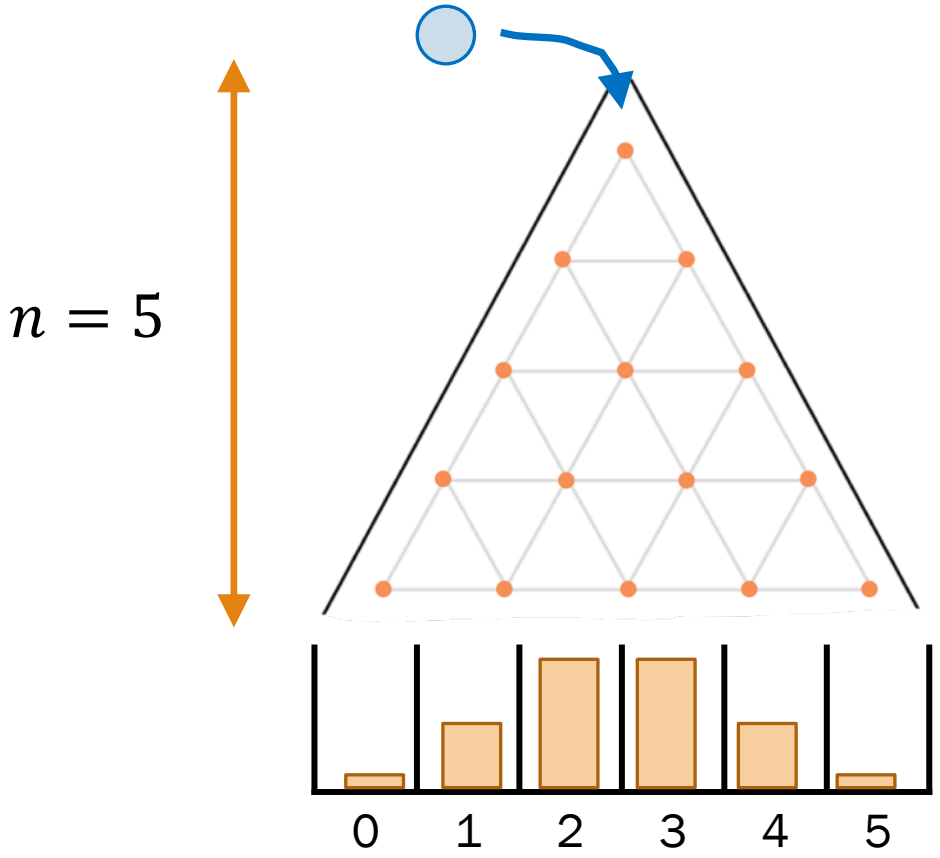


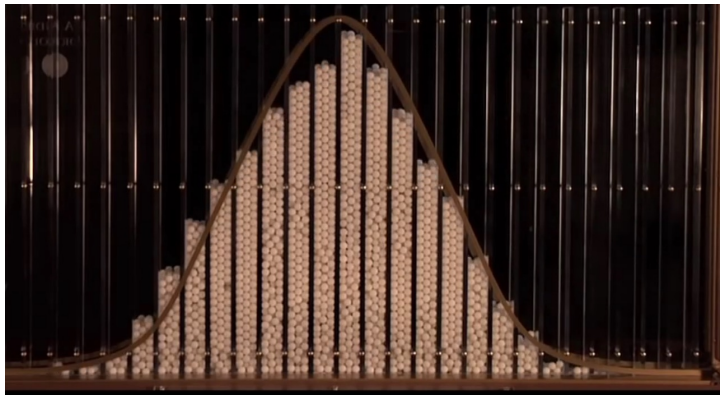
$\sum_{i=1}^{3} X_i$ Sum of 3 dice rolls

CLT explains a lot

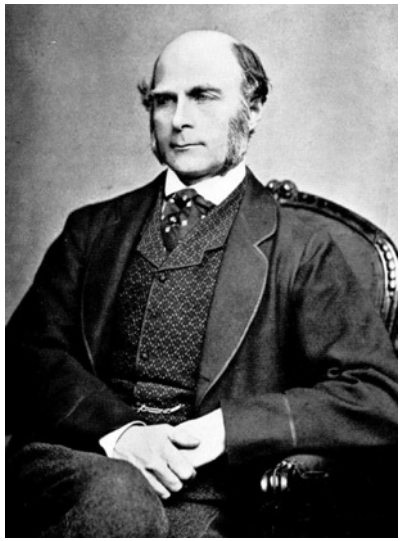
$$\text{As } n \to \infty \qquad \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

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





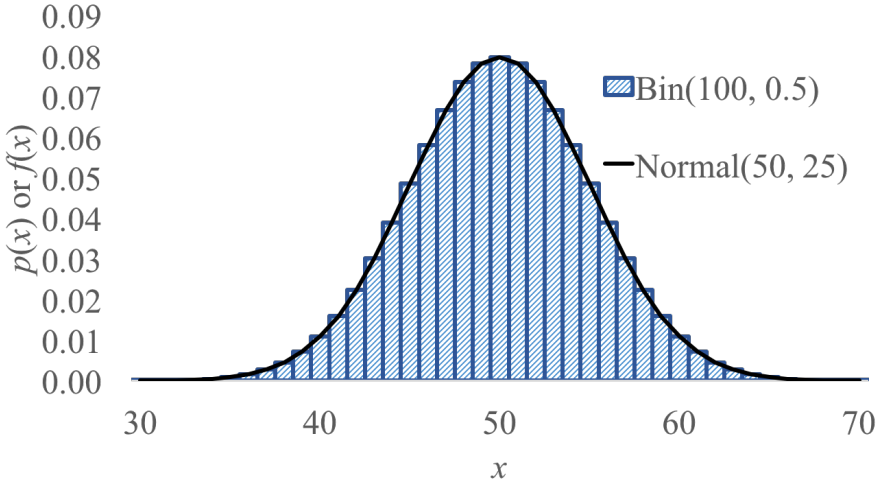
Galton Board, by Sir Francis Galton (1822-1911)



CLT explains a lot

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

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



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.

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

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

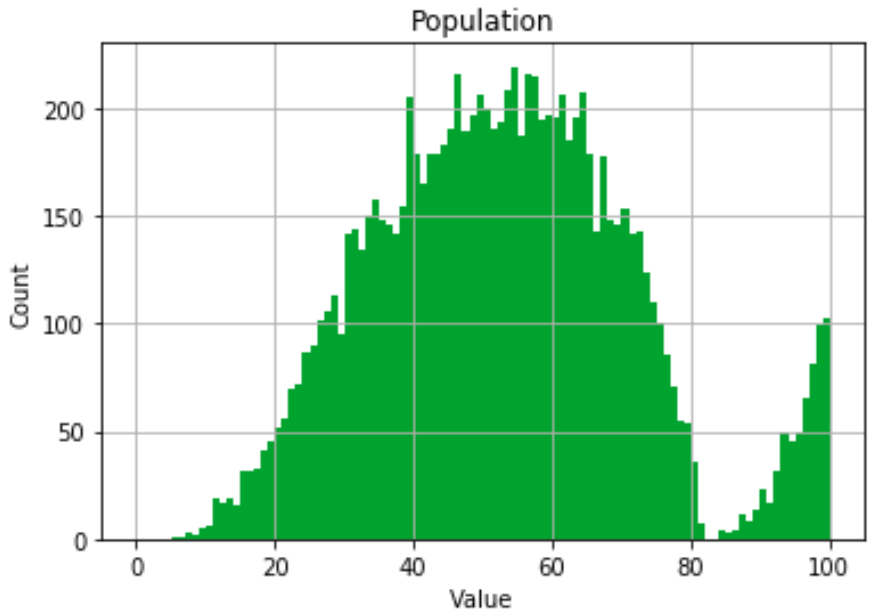
$$X \sim \mathcal{N}(n\mu, n\sigma^2) \qquad (\text{CLT, as } n \to \infty)$$

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

CLT explains a lot

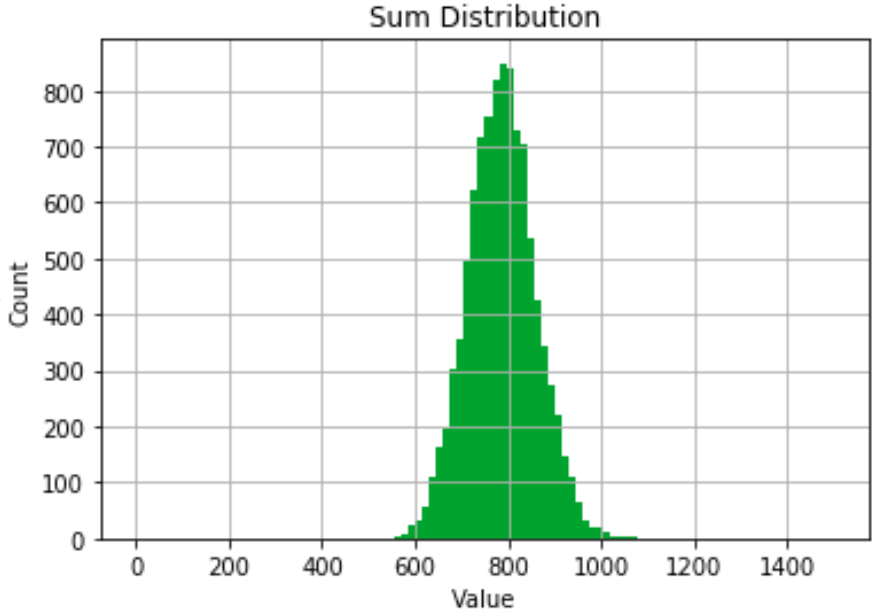
$$\text{As } n \to \infty \qquad \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

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



Distribution of X_i

Sample of size 15, **sum** values

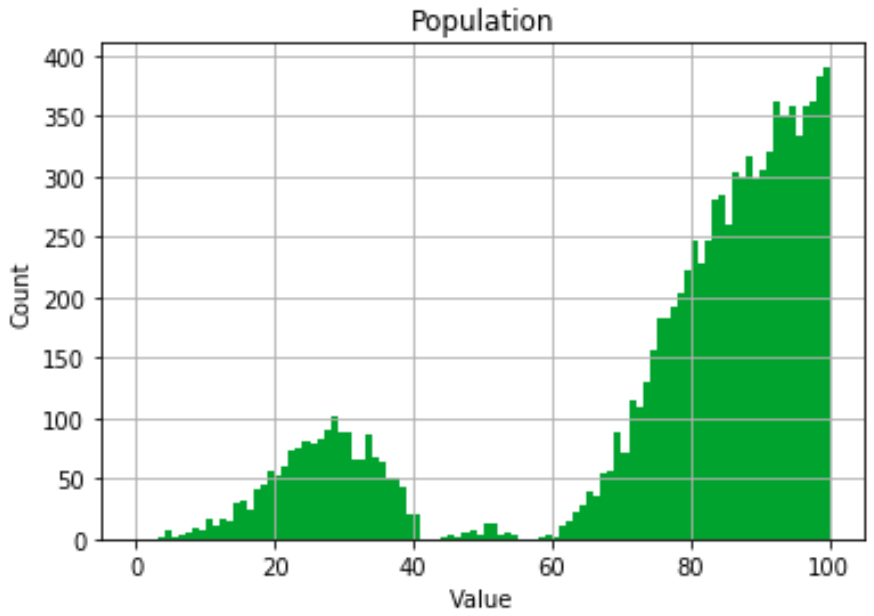


Distribution of $\sum_{i=1}^{15} X_i$

CLT explains a lot

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

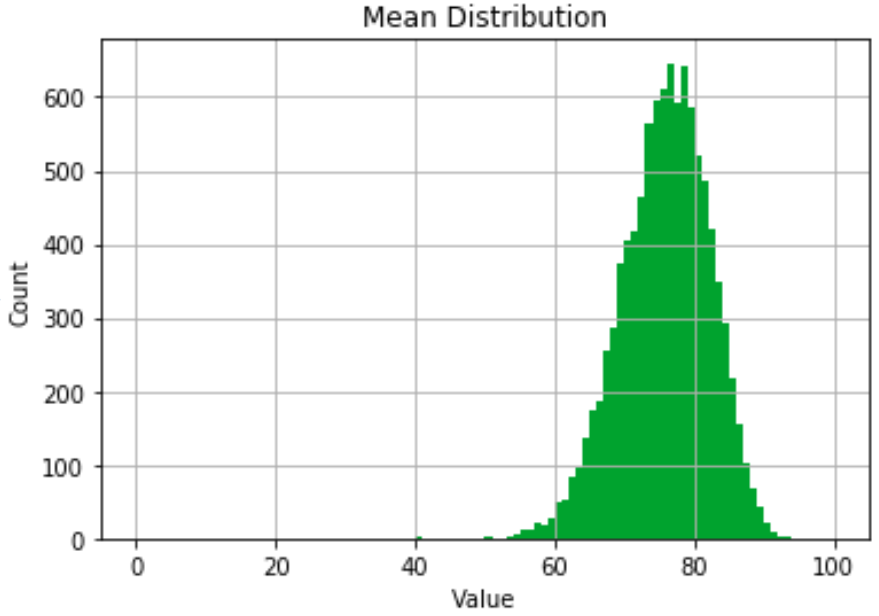
The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.



Distribution of X_i

Sample of size 15, **average** values

(sample mean)



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

Proof of CLT

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

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

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)

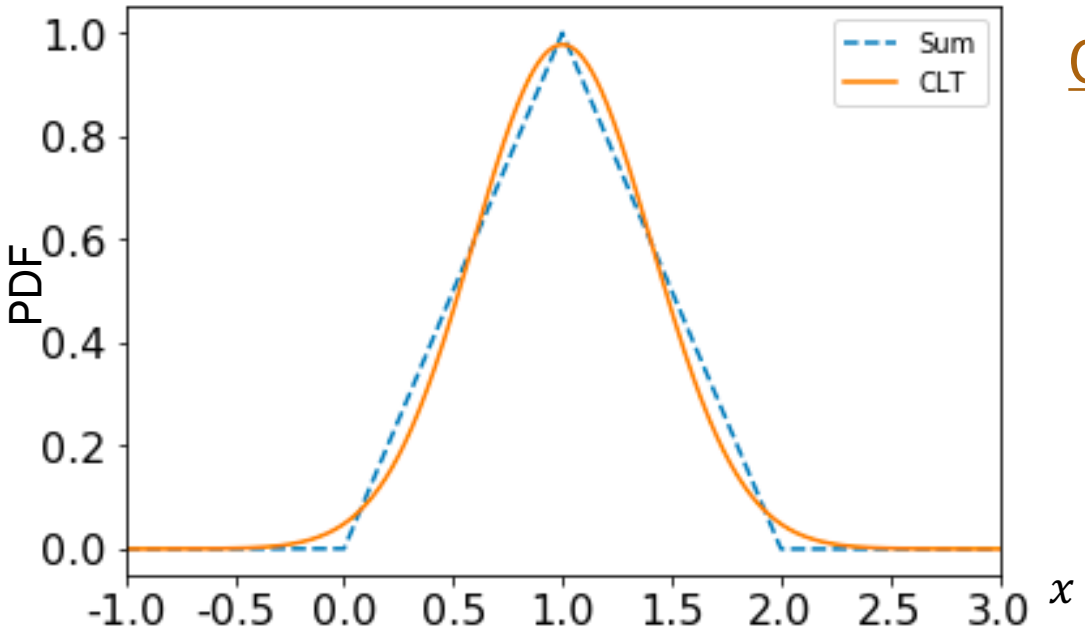
CLT example

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

CLT approximation

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

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

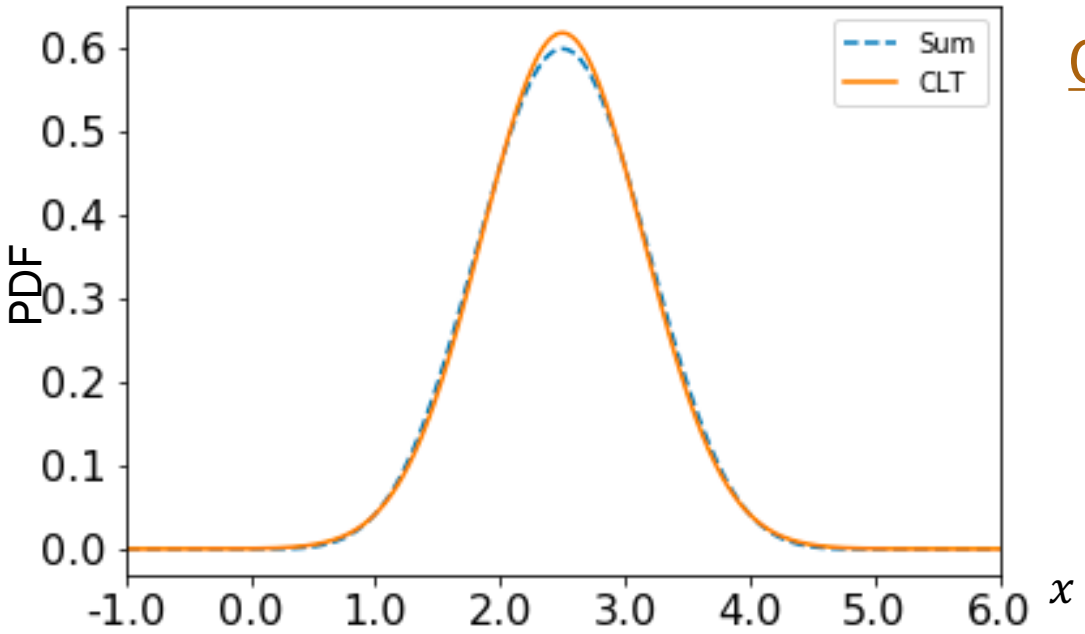
$$= \Phi\left(\frac{2/3 - 1}{\sqrt{1/6}}\right) \approx 0.2071$$

Sum of n independent Uniform RVs

Let $X = \sum_{i=1}^{n} X_i$ be sum of 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

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

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

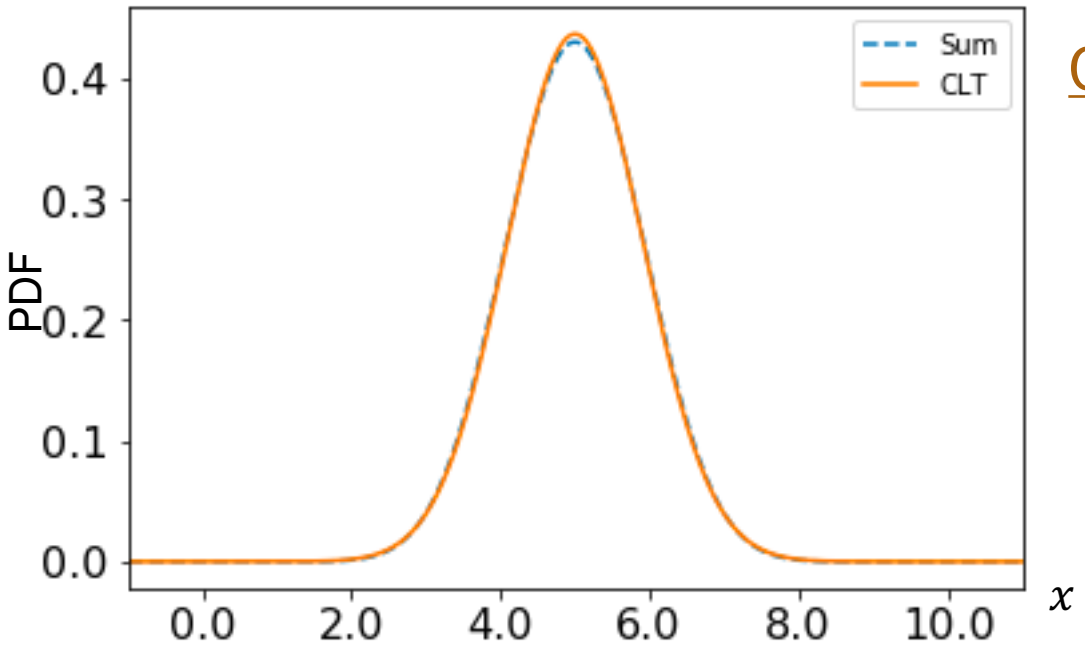
$$= \Phi\left(\frac{5/3 - 5/2}{\sqrt{5/12}}\right) \approx 0.0984$$

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

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

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

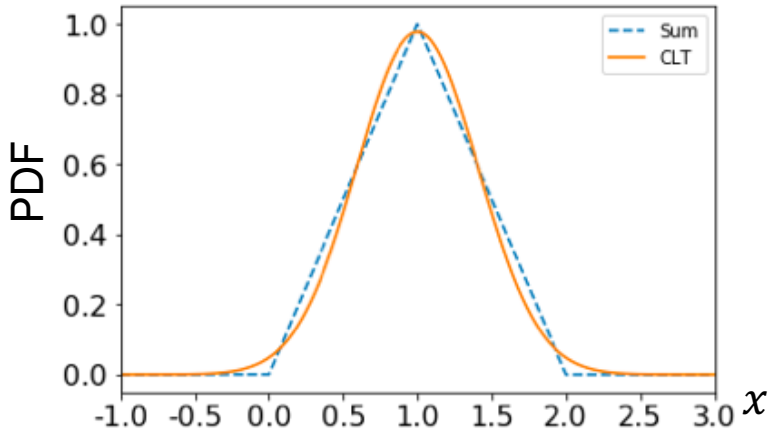
$$= \Phi\left(\frac{10/3 - 5}{\sqrt{5/6}}\right) \approx 0.0339$$

Sum of n independent Uniform RVs

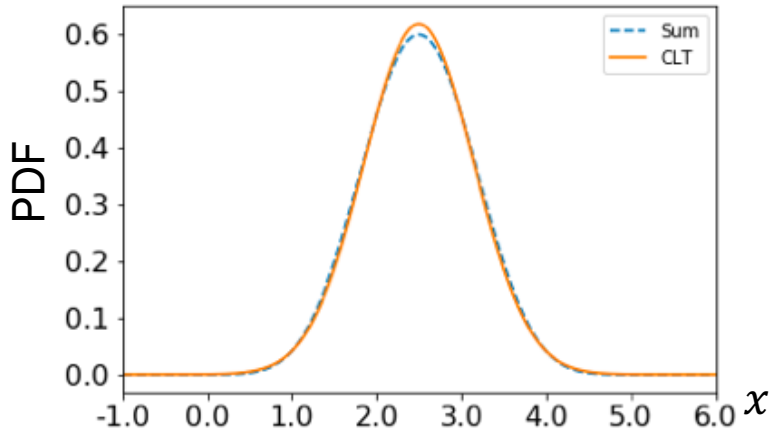
Let $X = \sum_{i=1}^{n} X_i$ be sum of i.i.d. RVs, where $X_i \sim \text{Uni}(0,1)$. $\mu = E[X_i] = 1/2$, $\sigma^2 = \text{Var}(X_i) = 1/12$

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

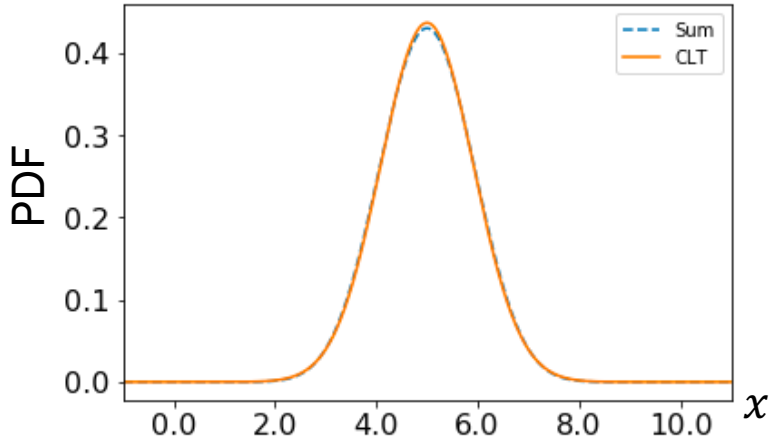
$n = 2$:



$n = 5$:



$n = 10$:



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.

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?

The Central Limit Theorem

- Consider I.I.D. random variables X_1, X_2, ...
  - X_i have same distribution with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - Let: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Central Limit Theorem:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \text{as} \quad n \to \infty$$

Demo

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

Demo Time!

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

By the Central Limit Theorem, the mean of IID variables are distributed normally.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

What about other functions?

Sum of iid? Normal

Average of iid? Normal

Max of iid?

What about other functions?

Sum of iid? Normal

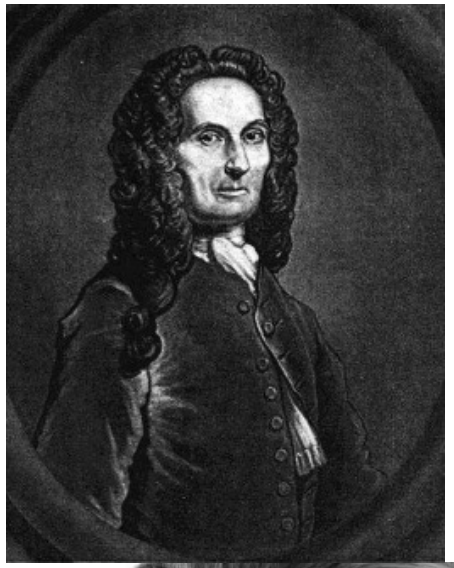
Average of iid? Normal

Max of iid? Gumbel

See Fisher Trippett Gnedenko Theorem

Once Upon a Time…

Abraham De Moivre



THE

DOCTRINE

OF

CHANCES:

OR,

A Method of Calculating the Probability of Events in Play.

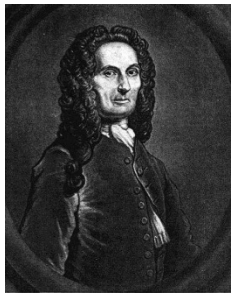
By *A. De Moivre.* F. R. S.

LONDON:

Printed by *W. Pearson,* for the Author. MDCCXVIII.

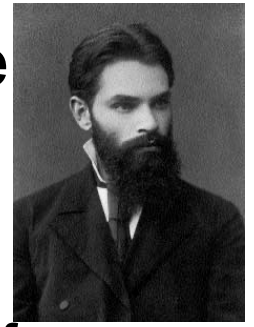
1733

Once Upon a Time…



- History of the Central Limit Theorem
  - 1733: CLT for X ~ Ber(1/2) postulated by Abraham de Moivre

  - 1823: Pierre-Simon Laplace extends de Moivre's work to approximating Bin(n, p) with Normal



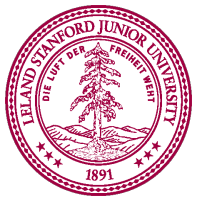
  - 1901: Aleksandr Lyapunov provides precise definition and rigorous proof of CLT
  - 2016: Beyonce releases Lemonade
    - It was her 6th album, bringing her total number of songs to 214
    - Mean quality of subsamples of songs is Normally distributed (thanks to the Central Limit Theorem)

Estimating Clock Running Time

- Have new algorithm to test for running time
  - Mean (clock) running time: μ = *t* sec.
  - Variance of running time: σ^2 = 4 sec^2.
  - Run algorithm repeatedly (I.I.D. trials), measure time
    - How many trials s.t. estimated time = $t \pm 0.5$ with 95% certainty?
    - X_i = running time of *i*-th run (for $1 \leq i \leq n$), $\bar{X}$ is the mean

---

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \sim N(t, \frac{4}{n})$$



$$0.95 = P(-0.5 < \bar{X} - t < 0.5) \qquad \bar{X} - t \sim N(0, \frac{4}{n})$$

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

$$= \Phi\Big(\frac{0.5 - 0}{\sqrt{4/n}}\Big) - \Phi\Big(\frac{-0.5 - 0}{\sqrt{4/n}}\Big)$$

$$= 2\phi(\frac{\sqrt{n}}{4}) - 1$$

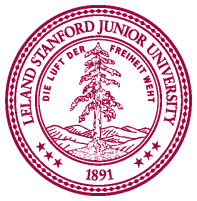
$$0.95 = 2\phi(\frac{\sqrt{n}}{4}) - 1$$

$$0.975 = \phi(\frac{\sqrt{n}}{4})$$

$$\phi^{-1}(0.975) = \frac{\sqrt{n}}{4}$$

$$1.96 = \frac{\sqrt{n}}{4}$$

$$n = 61.4$$



William Sealy Gosset
(aka Student)

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 ($X_1, X_2, \ldots, X_{10}$)
  - 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: $X = \sum_{i}^{n} X_i \to N(n\mu, n\sigma^2) \qquad \text{As } n \to \infty$
  - Determine P(X ≤ 25 or X ≥ 45) using CLT:

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

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

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

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

That’s all folks!