18: Central Limit Theorem

Lisa Yan and Jerry Cain

October 23, 2020

Quick slide reference

i.i.d. random variables

Another big day

Up until this point, we've mostly covered traditional probability topics:

- Equally likely outcomes
- Conditional probability, independence, random variables
- Joint probability distributions, conditional expectation

We have done some awesome applications:

- Federalist Papers: Authorship identification
- WebMD: General Inference

Today

- Our last big topic in **traditional probability** before we move onto modern-day statistical analysis!

Independence of multiple random variables

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 $X_1, X_2, \ldots, X_n$ 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.!

Central Limit Theorem

(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 $n \to \infty$

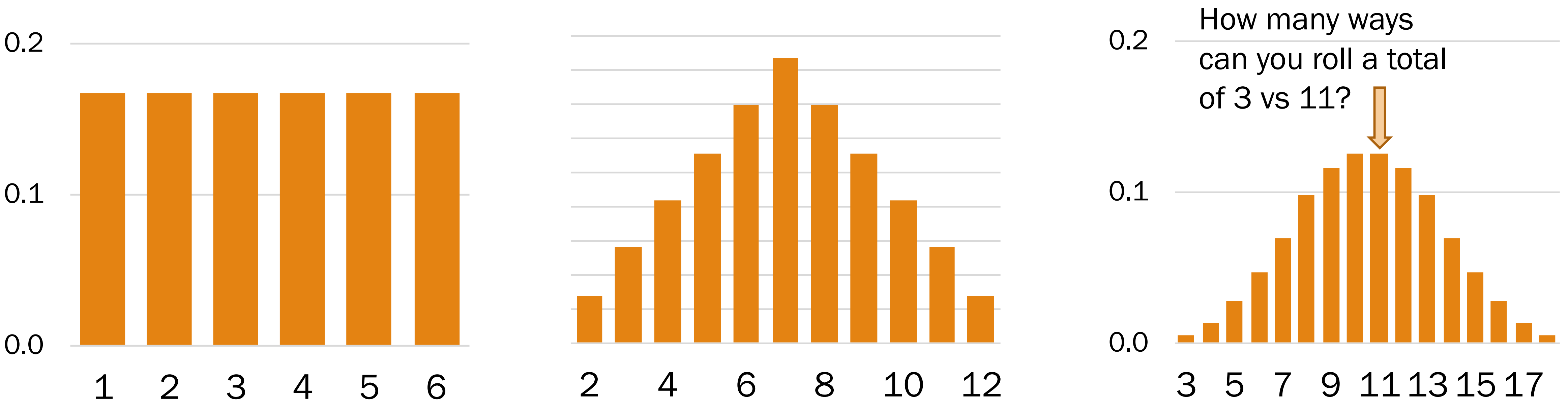
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

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

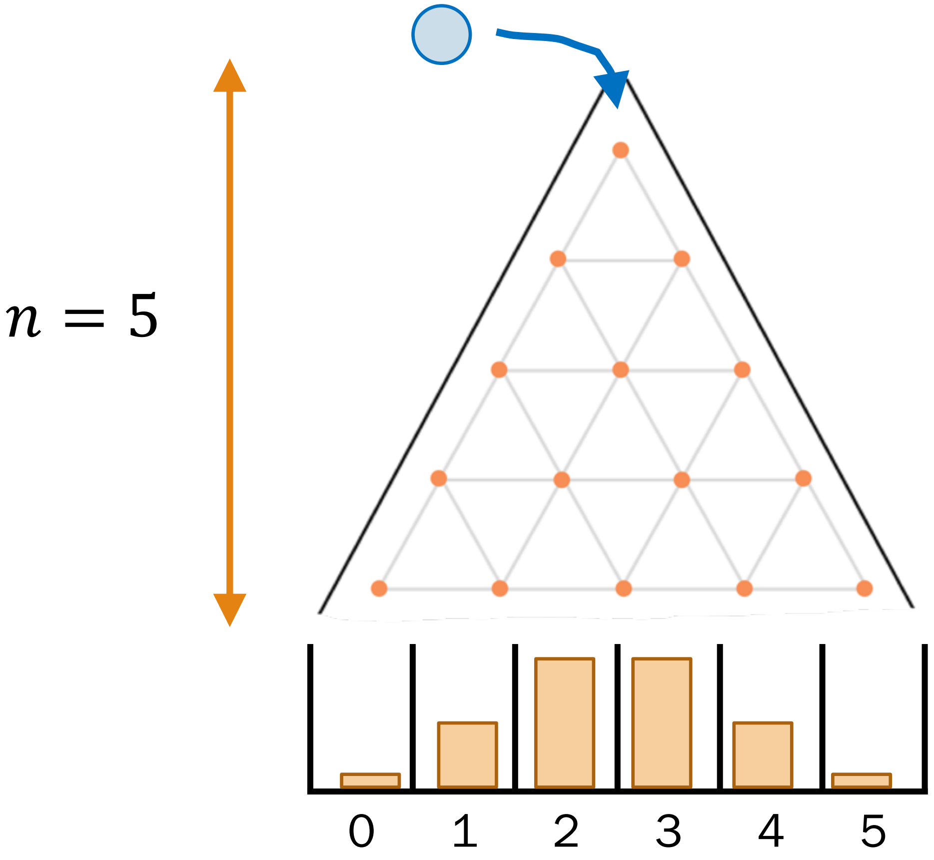
$\sum_{i=1}^{2} X_i$ Sum of 2 dice rolls

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

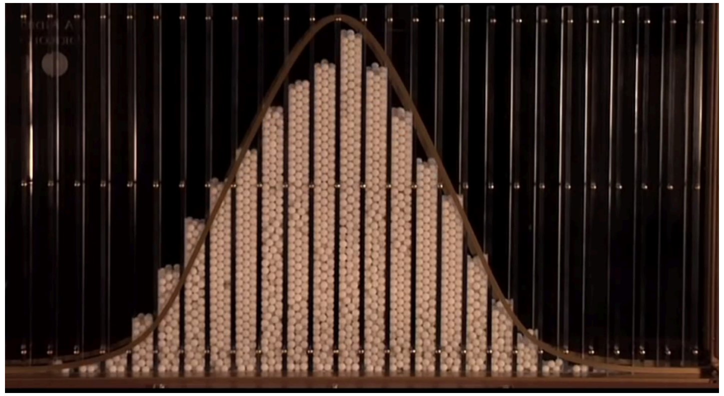
CLT explains a lot

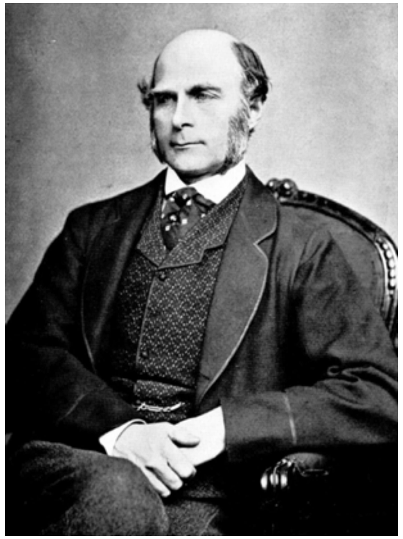
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad \text{As } n \to \infty$$

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



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

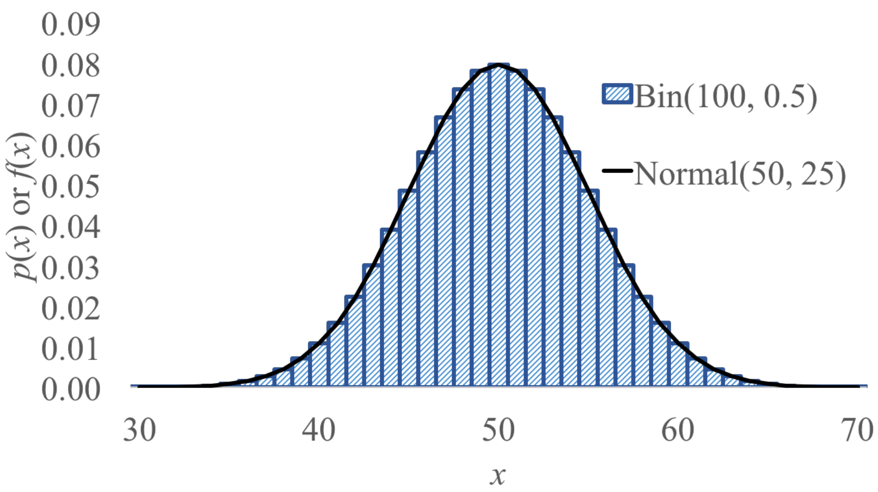




CLT explains a lot

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$ 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 $i = 1, \dots, n$, where X_i are i.i.d.
$E[X_i] = p$, $\text{Var}(X_i) = p(1-p)$

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

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

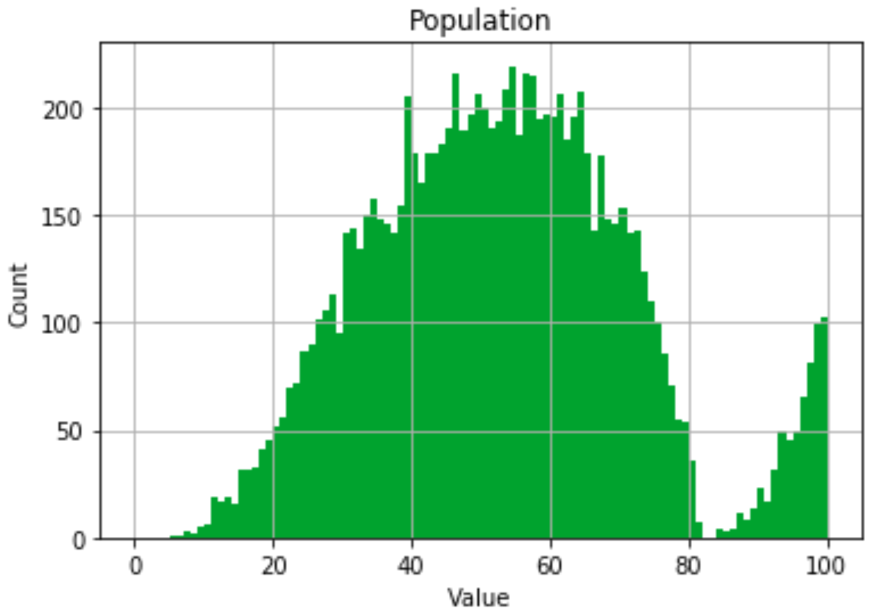
$$X \sim \mathcal{N}(np, np(1-p))$$ (substitute mean, variance of Bernoulli)

CLT explains a lot

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

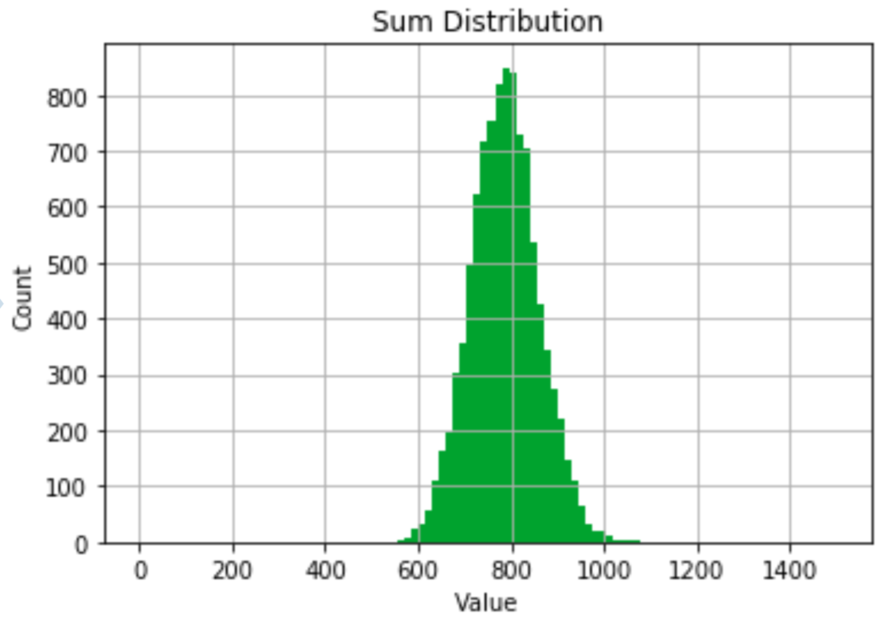
As $n \to \infty$

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



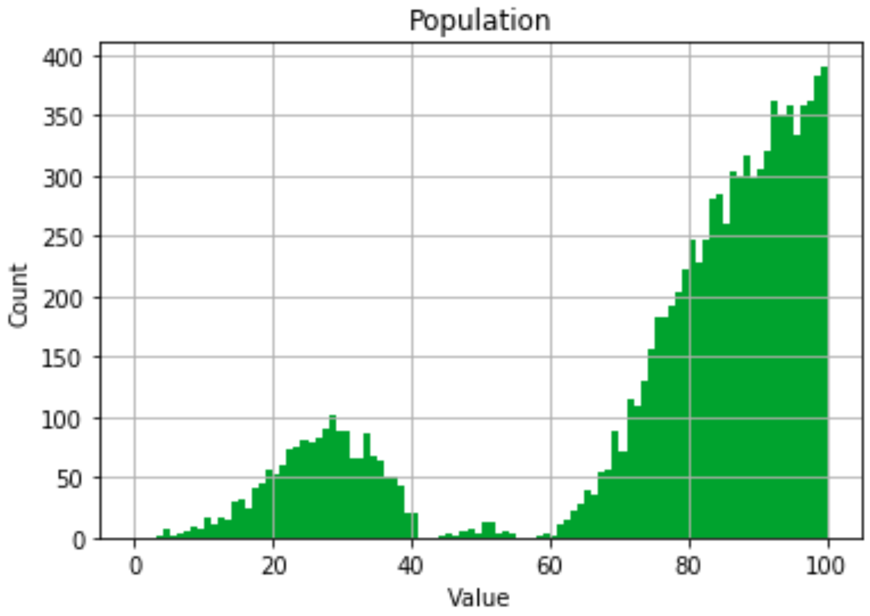
Distribution of $\sum_{i=1}^{15} X_i$

CLT explains a lot

As $n \to \infty$

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

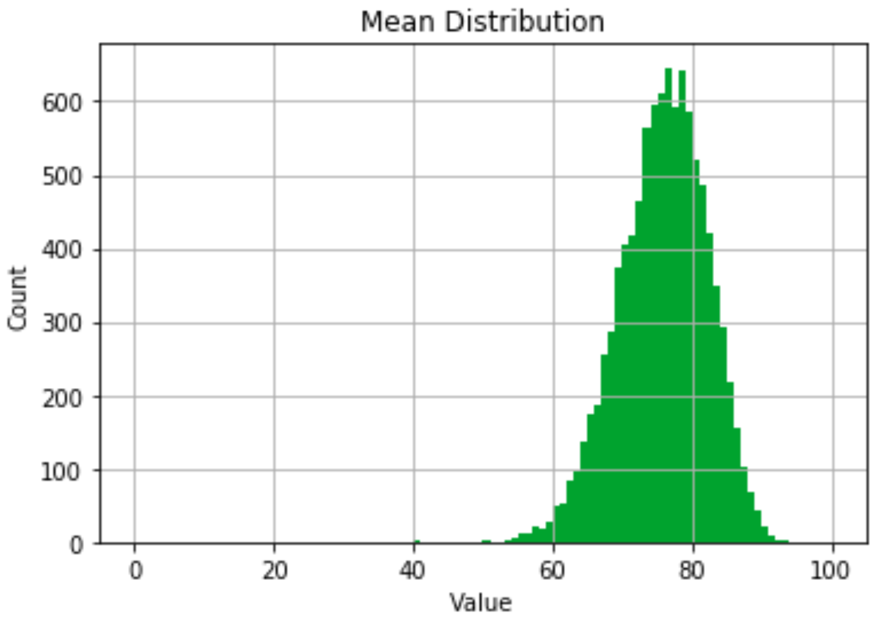
The sum of n **i.i.d.** random variables is normally distributed with mean $n\mu$ and variance $n\sigma^2$.



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

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

CLT example

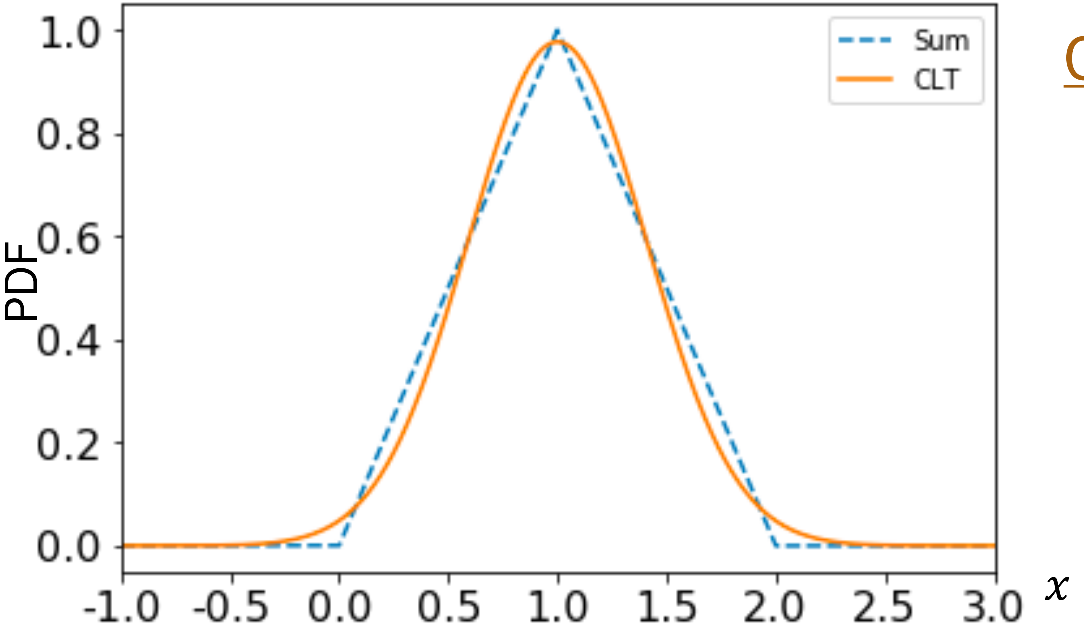
Sum of n independent Uniform RVs

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

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

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

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

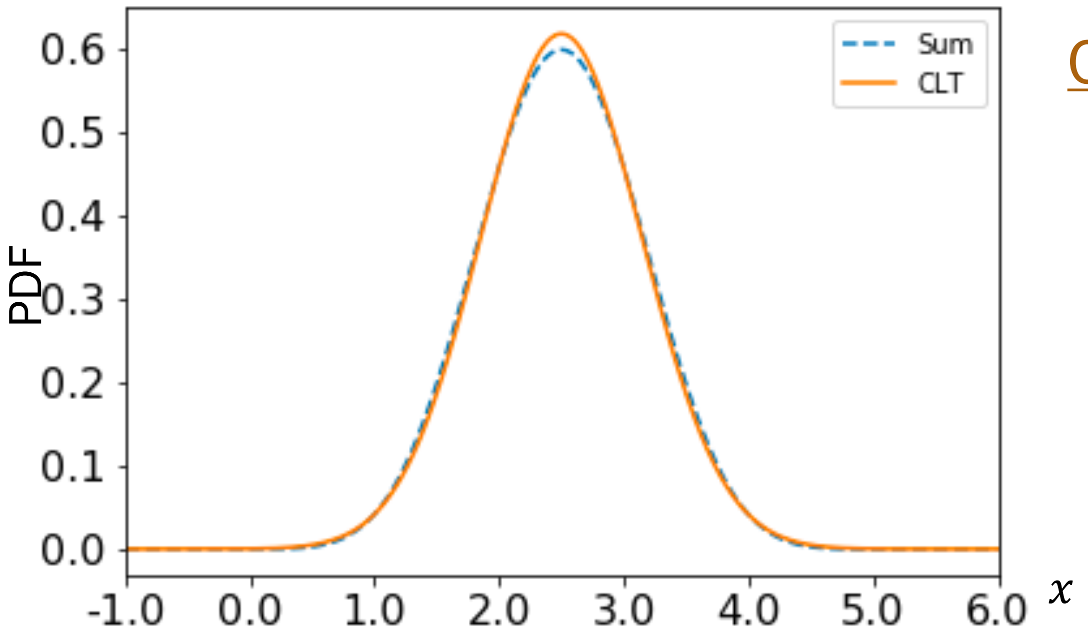
Sum of n independent Uniform RVs

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

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

CLT approximation

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

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

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

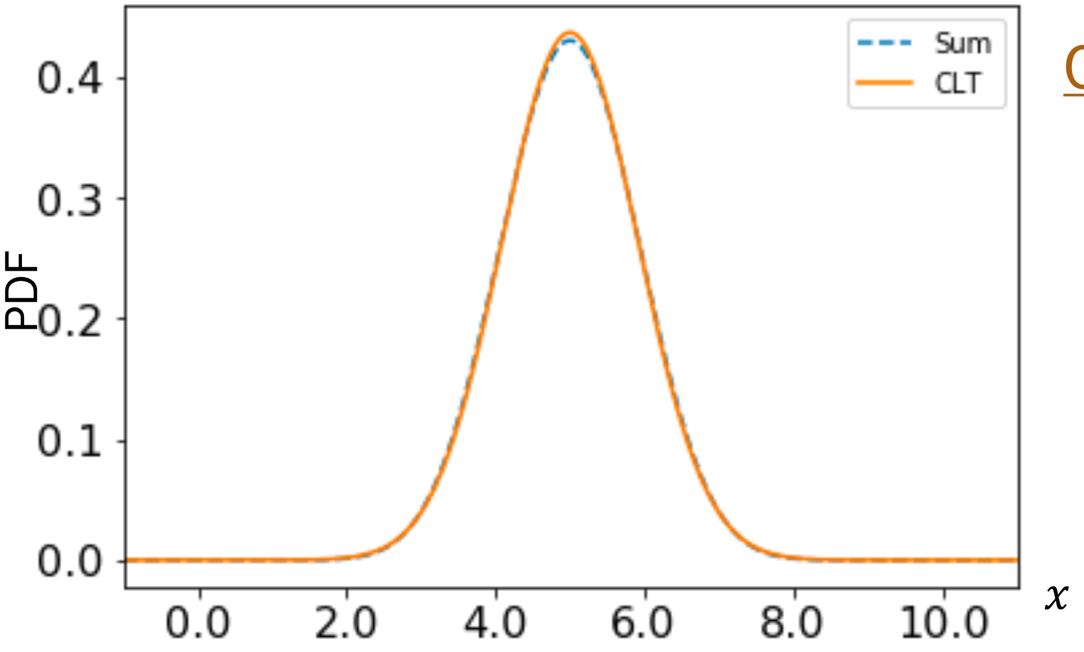
Sum of n independent Uniform RVs

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

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

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

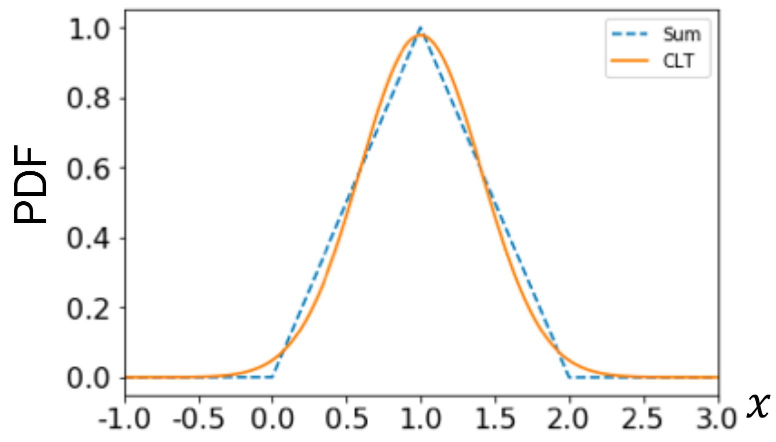
$$= \Phi\left(\frac{10/3 - 5}{\sqrt{5/6}}\right) \approx 0.0339$$

Sum of n independent Uniform RVs

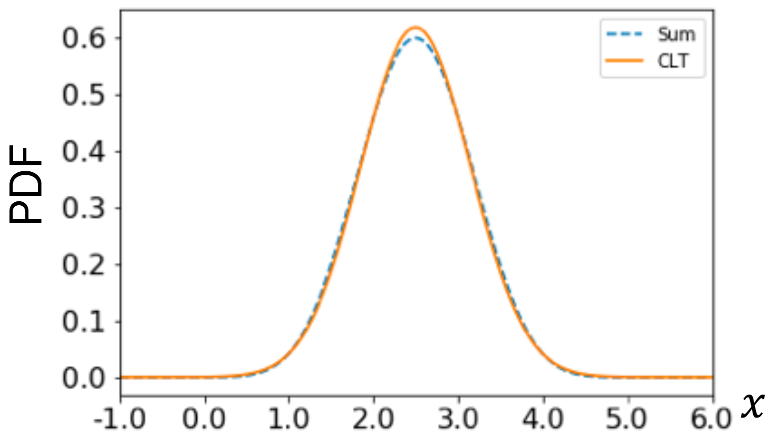
Let $X = \sum_{i=1}^{n} X_i$ be sum of i.i.d. RVs, where $X_i \sim \text{Uni}(0,1)$. $\mu = E[X_i] = 1/2$, $\sigma^2 = \text{Var}(X_i) = 1/12$

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

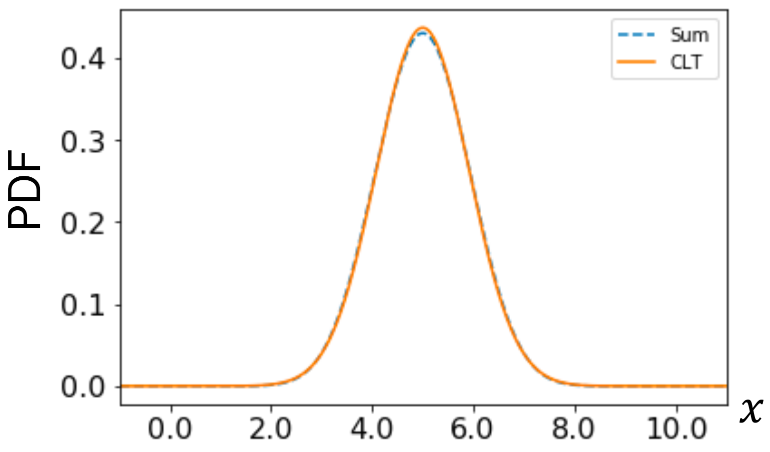
$n = 2$:



$n = 5$:



$n = 10$:



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.

Sum/average/max of i.i.d. random variables

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 $n \to \infty$:

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$	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 $X_1, X_2, \ldots, X_n$ be i.i.d., 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 i.i.d. RVs
?	Average of i.i.d. RVs (sample mean)
?	Max of i.i.d. RVs

Distribution of sample mean

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

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

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

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

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

Distribution of sample mean

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

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

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

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

$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ (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 μ and variance σ^2/n.

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

What about other functions?

Let $X_1, X_2, \ldots, X_n$ be i.i.d., where $E[X_i] = \mu$, 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 i.i.d. RVs
$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$	Average of i.i.d. RVs (sample mean)
Gumbel	Max of i.i.d. RVs

(see Fisher-Tippett Gnedenko Theorem)

(live)

18: Central Limit Theorem

Lisa Yan and Jerry Cain

October 23, 2020

Think

Slide 36 has a question to go over by yourself.

Post any clarifications here!

https://us.edstem.org/courses/2678/discussion/153773

Think by yourself: 2 min

(by yourself)

Quick check

What dimensions are the following RVs?
(Let X_i be i.i.d. with mean μ)

1. X_1
2. $(X_1, X_2, \dots, X_n)$
3. $\sum_{i=1}^{n} X_i$
4. $\frac{1}{n}\sum_{i=1}^{n} X_i$
5. $\frac{1}{n}\sum_{i=1}^{n} \mu$

A. 1-D random variable
B. n -D random variable (a vector)
C. not a random variable

(by yourself)

Quick check

What dimensions are the following RVs?
(Let X_i be i.i.d. with mean μ)

1. X_1
2. $(X_1, X_2, \dots, X_n)$ (aka a **sample**)
3. $\sum_{i=1}^{n} X_i$
4. $\frac{1}{n}\sum_{i=1}^{n} X_i$ (aka the **sample mean**)
5. $\frac{1}{n}\sum_{i=1}^{n} \mu$

A. 1-D random variable
B. n -D random variable (a vector)
C. not a random variable

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 $X \geq 45$.

To the demo!

Breakout Rooms

Check out the question on the next slide (Slide 36). Post any clarifications here!

https://us.edstem.org/courses/2678/discussion/153773

Breakout rooms: 3 min

🤔

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 $X \le 25$ 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: ?

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.

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

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

Dice game

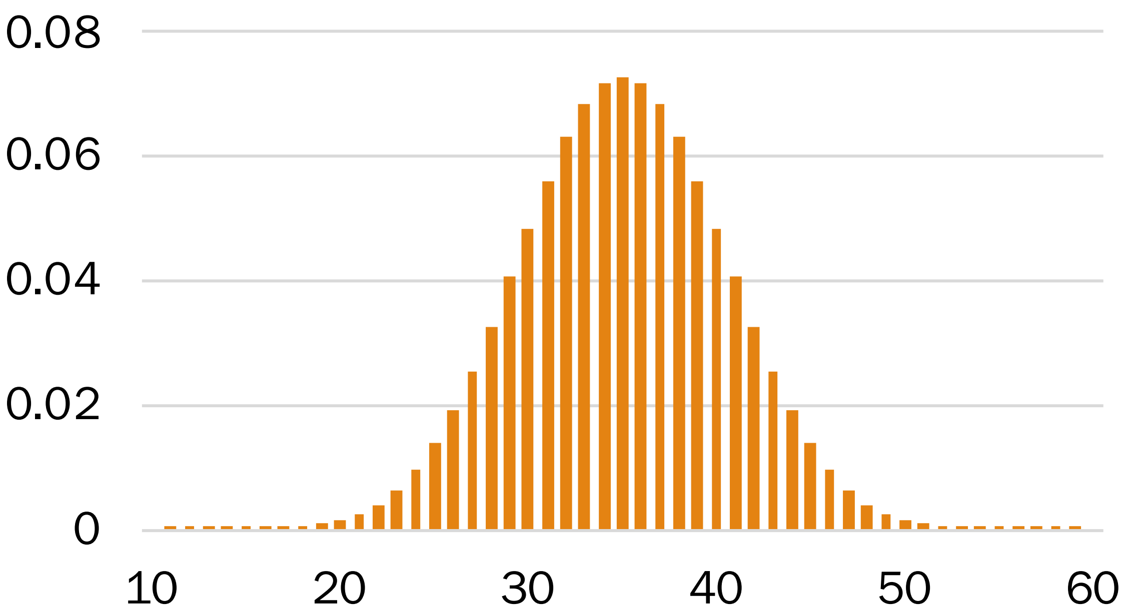
As $n \to \infty$: $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$

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)…

Check out the code!



(by CLT)

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

(exact, by computer)

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

(exact, by computer)

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

Summary: Working with the CLT

Let $X_1, X_2, \ldots, X_n$ i.i.d., 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 i.i.d. RVs

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

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

⚠️

If X_i is discrete:

Use the continuity correction on Y!

Interlude for jokes/announcements

Announcements

Quiz #2 Review session

When: Monday 10/26 7pm-9pm PT
Recorded: yes

Zoom link

Quiz #2 Info and practice: Exam page link
Covers PS3, PS4 (i.e., up to and including Lecture 15)

Think

Slide 43 has a question to go over by yourself.

Post any clarifications here!

https://us.edstem.org/courses/2678/discussion/153773

Think by yourself: 2 min

(by yourself)

Crashing website

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

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

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

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

Check out the code!

Solve

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

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 (i.i.d. 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 (i.i.d. 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 (i.i.d. 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.

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

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

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

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

$\Rightarrow$ $n \approx 62$

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 (i.i.d. 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, σ^2/n decreases
- The probability that our sample mean $\bar{X}$ is *close* to the true mean μ increases

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
(of the Galton Board)

Next time

Central Limit Theorem:

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

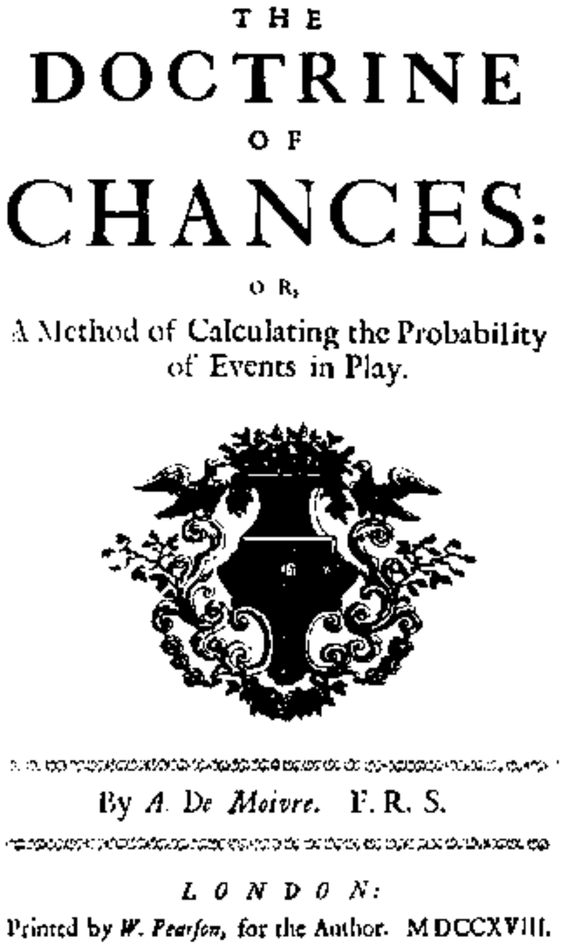
In real life:

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

...until next time!

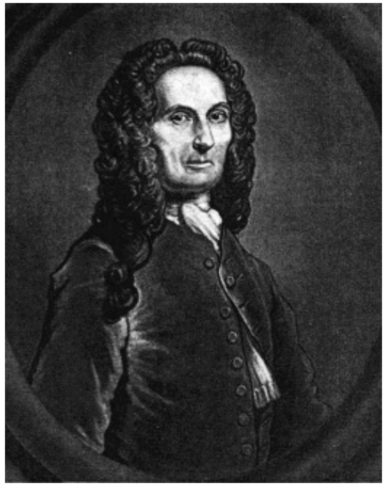
Extra: History of the CLT

Once upon a time…


THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability of Events in Play.

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

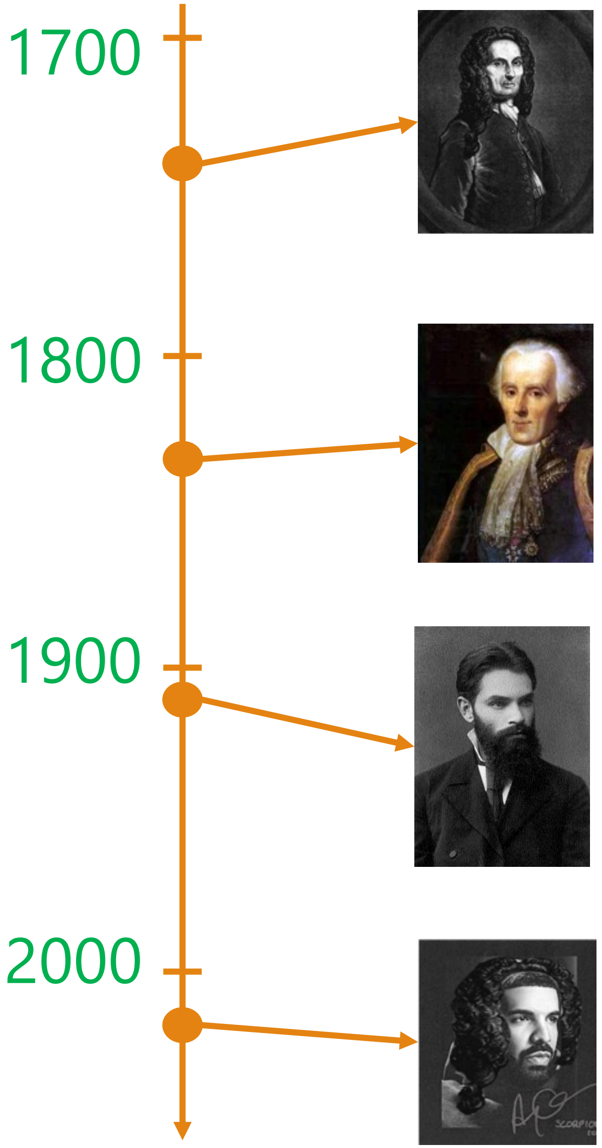
LONDON:
Printed by W. Pearson, for the Author. MDCCXVIII.



Abraham de Moivre
CLT for $X \sim \text{Ber}(1/2)$
1733

Aubrey Drake Graham
(Drake)

A short history of the CLT



1733: CLT for $X \sim \text{Ber}(1/2)$
postulated by Abraham de Moivre

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

1901: Alexandr Lyapunov provides precise definition and rigorous proof of CLT

2018: Drake releases *Scorpion*

- It was his 5th studio album, bringing his total # of songs to 190
- Mean quality of subsamples of songs is normally distributed (thanks to the Central Limit Theorem)