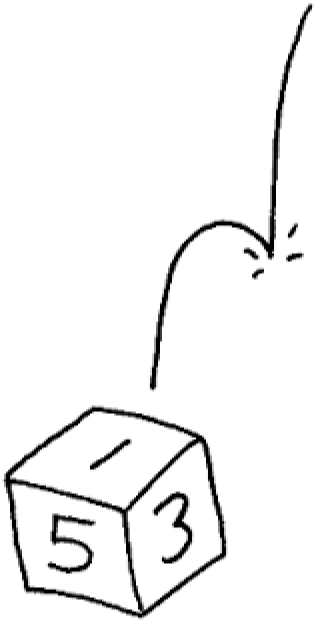
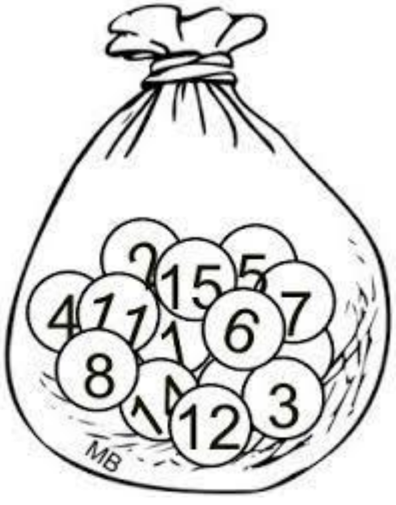
Lecture 11

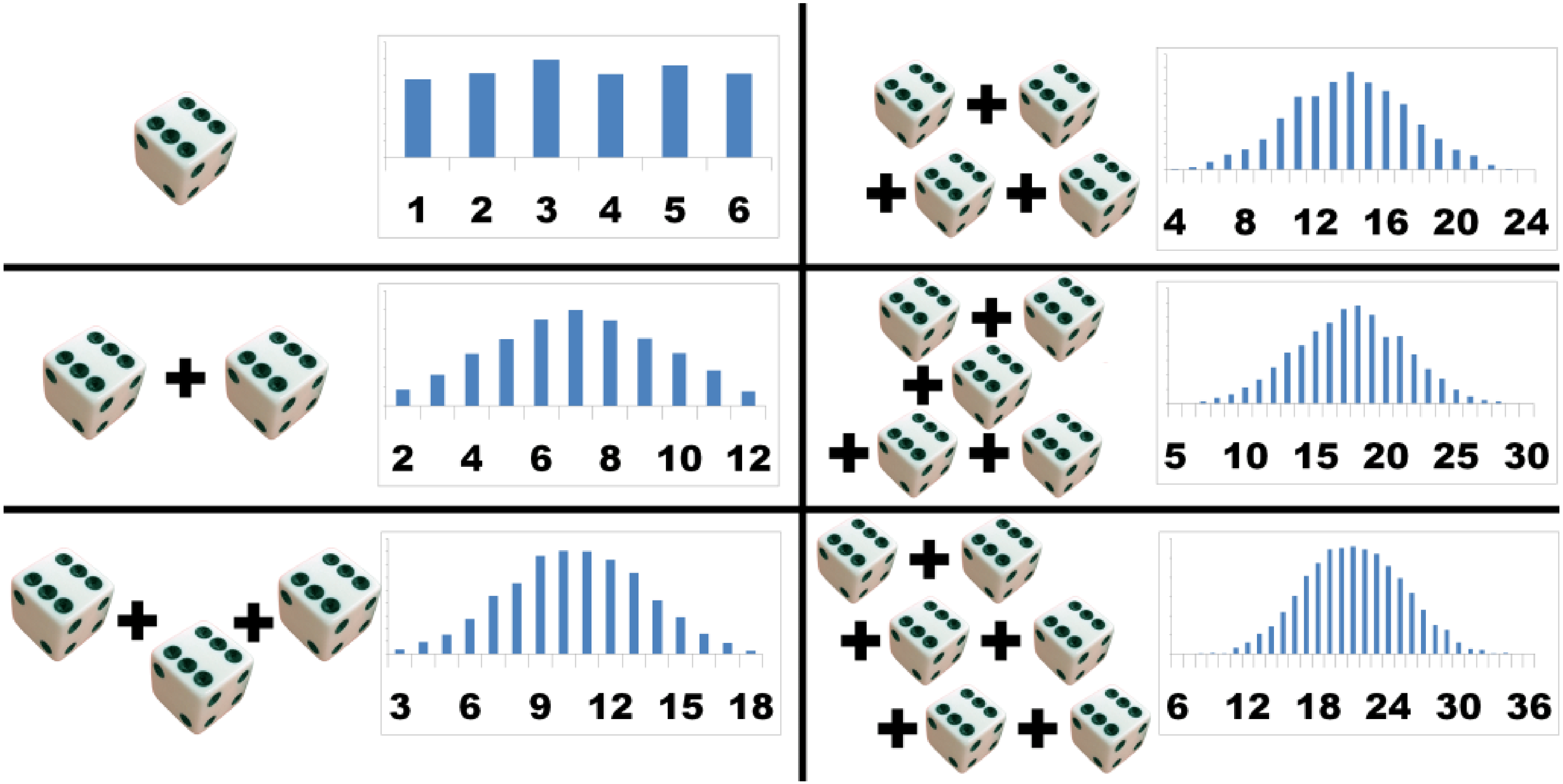
- Central limit theorem
  (中心极限定理，CLT)





ONE BIG DAY

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



Explain why Normal distribution is important

Important Inequalities

Chebyshev's inequality. If X is a R.V. with finite mean μ and variance σ^2, then for any value $k > 0$,

$$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2} \quad \text{or} \quad P\{|X-\mu| < k\} \geq 1 - \frac{\sigma^2}{k^2}$$

Can be derived from the Markov's inequality. P395 of textbook

Estimate (derive bounds on) the probability without knowing exact PDF.

Proof: $P\{|X-\mu| \geq k\} = \int_{|x-\mu|\geq k} f(x)dx$

$$\leq \int_{|x-\mu|\geq k} \frac{(x-\mu)^2}{k^2} f(x)dx$$

$$\leq \frac{1}{k^2}\int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx = \frac{D(X)}{k^2} = \frac{\sigma^2}{k^2}$$

$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2}$ or $P\{|X-\mu| < k\} \geq 1-\frac{\sigma^2}{k^2}$

Ex. Given $k = 3\sigma$ or 4σ, estimate the probability of $P\{|X-\mu| < k\}$ with Chebyshev's inequality.

Ex. Given $E(X) = -2$, $D(X) = 1$, $E(Y) = 2$, $D(Y) = 4$, and X, Y are independent, find the probability of $P\{|X+Y| \geq 5\}$.

$$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2} \quad \text{or} \quad P\{|X-\mu| < k\} \geq 1-\frac{\sigma^2}{k^2}$$

Ex. Given $k = 3\sigma$ or 4σ, estimate the probability of $P\{|X - \mu| < k\}$ with Chebyshev's inequality.

Sol.

$$P\{|X-\mu| < 3\sigma\} \geq 1-\frac{\sigma^2}{k^2} = 1-\frac{1}{9} = 0.8889$$

$$P\{|X-\mu| < 4\sigma\} \geq 1-\frac{\sigma^2}{k^2} = 1-\frac{1}{16} = 0.9375$$

Ex. Given $E(X) = -2$, $D(X) = 1$, $E(Y) = 2$, $D(Y) = 4$, and X, Y are independent, find the probability of $P\{|X + Y| \geq 5\}$.

Sol. $P\{|X+Y-E(X+Y)| \geq 5\} \leq D(X+Y)/5^2 = 1/5$

$$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2} \quad \text{or} \quad P\{|X-\mu| < k\} \geq 1-\frac{\sigma^2}{k^2}$$

Understand Chebyshev's Inequality

$$P\{|X-\mu| \geq n\sigma\} \leq \frac{1}{n^2} \text{ or } P\{|X-\mu| < n\sigma\} \geq 1-\frac{1}{n^2}$$

Intuitive explanation:
Given any practical dataset, the probability of deviating $n\sigma$ away from the mean is lower bounded by $1-1/n^2$.

- More than 3/4 of the data are within 2σ range.
- More than 8/9 of the data are within 3σ range.
- More than 24/25 of the data are within 4σ range.

Chebyshev's inequality **bounds** the data with simple statistics, e.g., mean and variance.

An example of the exchange rate of euro in a week.
The red line indicates an average over the last 20 days.
The green, blue and yellow lines mark the 2σ, 3σ and 5σ ranges, respectively.

知乎：
[切比雪夫不等式到底是个什么概念?]

Ex. (**Determining the Required Number of Observations**) Suppose that a random sample is to be taken from a distribution with unknown mean μ, and the standard deviation σ is 2 units or less. We shall determine how large the sample size n must be in order to make the probability of $|\overline{X_n} - \mu|$ less than 1 unit to be at least 0.99.

$(\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i)$

Chebyshev's inequality $P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2}$ or $P\{|X-\mu| < k\} \geq 1 - \frac{\sigma^2}{k^2}$

Since $\sigma^2 \leq 4$, it follows from the Chebyshev's inequality that for every sample size n

$$P\{|\overline{X_n} - \mu| < 1\ \mathrm{unit}\} \geq 1 - \frac{\sigma^2\ \mathrm{unit}^2}{n\ (1\ \mathrm{unit})^2}$$

Since the probability must satisfies $P\{|\overline{X_n} - \mu| < 1\} \geq 0.99$, it follows that $\frac{\sigma^2}{n} \leq \frac{4}{n} \leq 0.01 \implies n \geq 400.$

Weak law of large number

Given $X_1, \dots, X_n$ are i.i.d. R.V.s with finite mean $E[X_i] = \mu$.

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right\} \Rightarrow 0$$

as $n \to \infty$

Simple proof with Chebyshev's inequality:

$$E\left[\frac{X_1+\cdots+X_n}{n}\right] = \mu \quad \text{and} \quad D\left(\frac{X_1+\cdots+X_n}{n}\right) = \frac{\sigma^2}{n}$$

From Chebyshev's inequality:

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

It follows

$$P\left\{\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{n\varepsilon^2} \Rightarrow 0 \quad (\text{as } n \to \infty)$$

Laws of Large Number vs. Central Limit Theorem

Laws of large number

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \Rightarrow E(X), \quad n \to +\infty$$

…… converges to the expected average.

Central Limit Theorem

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \qquad n \to +\infty$$

…… follows a normal distribution.

Central Limit Theorem (中心极限定理)

Given $X_1, \dots, X_n$ are **i.i.d.** R.V.s with mean μ and variance $\sigma^2 > 0$.

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

where $n \to \infty$.

Alternatively,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1)$$

or

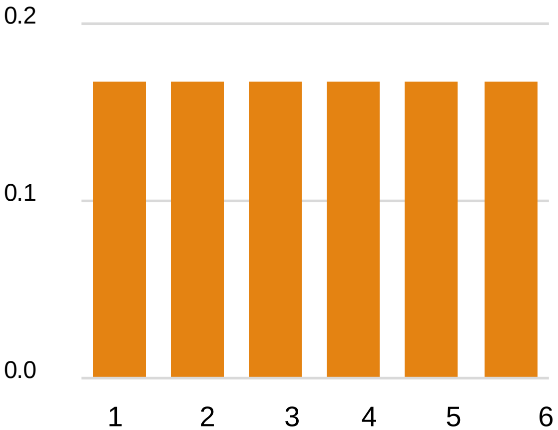
$$P\left\{\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq a\right\} \Rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

where $n \to \infty$.

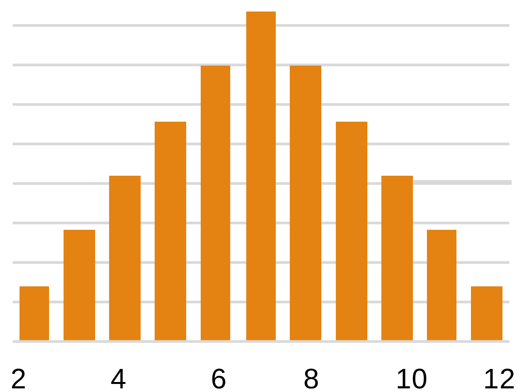
Proof of CLT is not simple.

Example of CLT: Sum of dice rolls

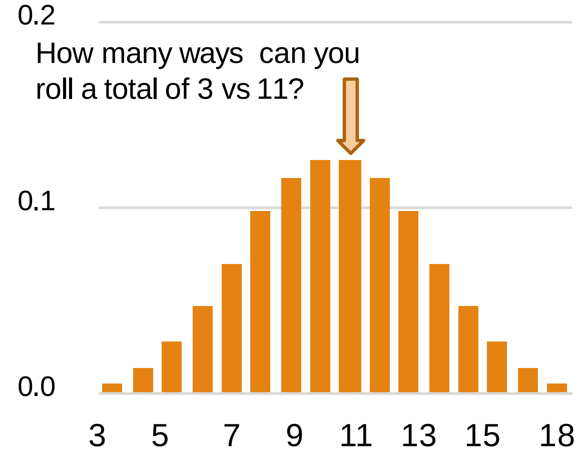
Roll n independent dice. Let X_i be the outcome of roll i.



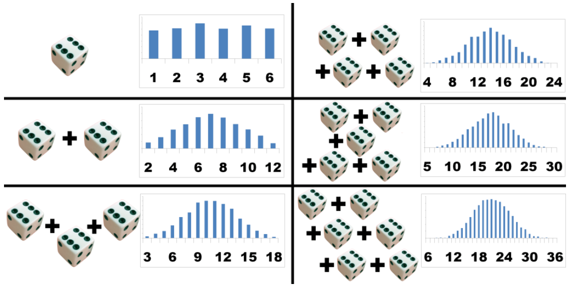
$\sum_{i=1}^{1} X_i$ sum of 1 die



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

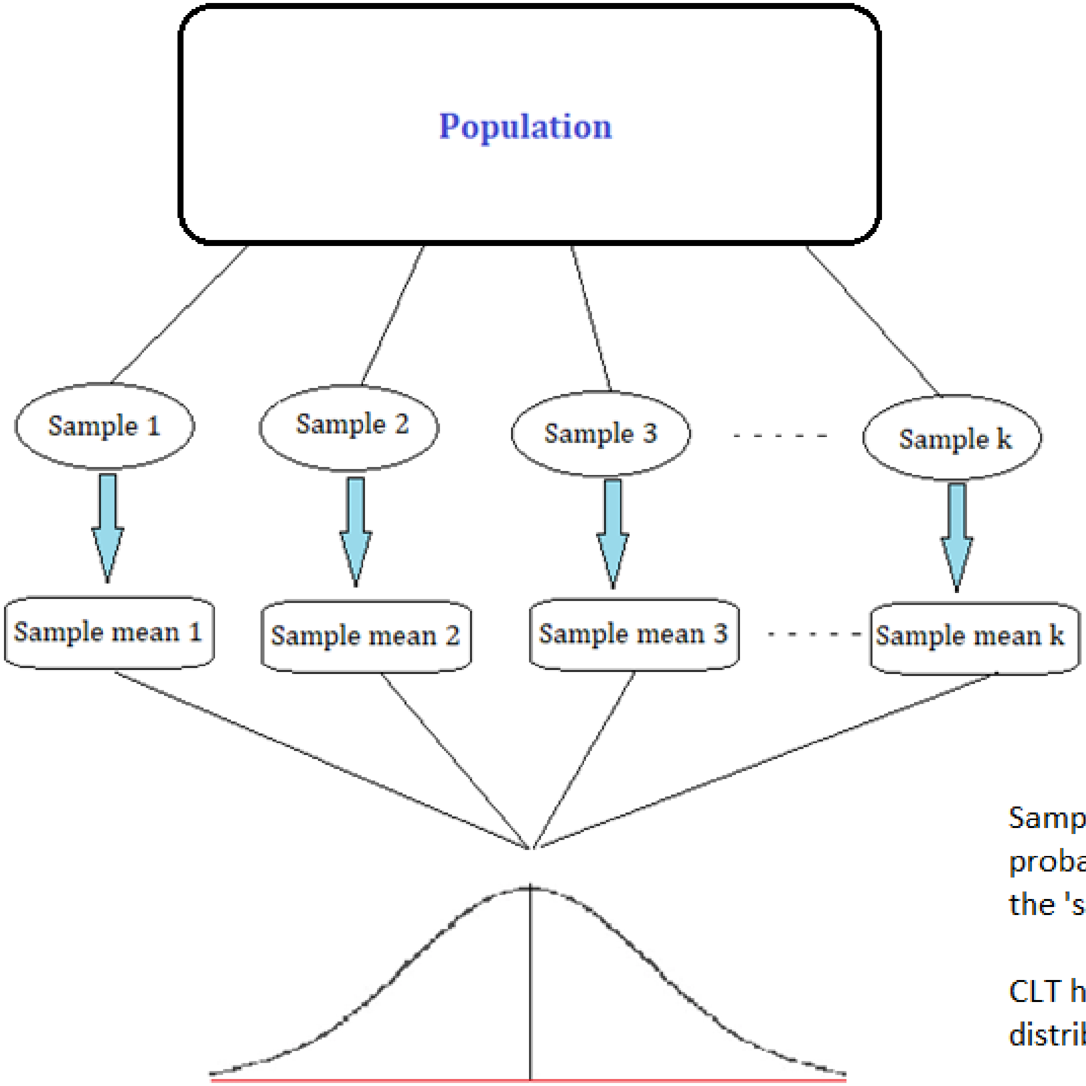


$\sum_{i=1}^{3} X_i$ sum of 3 dice



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

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



General CLT statement

Population
(with unknown distribution)
↓
Samples
(with arbitrary sizes, n)
↓
Sample means/sums
(computed from samples)
↓
Normally distributed

Sample mean is a random variable itself and thus has a probability distribution like any other random variable called the 'sampling distribution'!

CLT helps us to figure out the parameters of the sampling distribution

Figure from Exp. 4

Important variants of CLT

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \frac{\bar{X} - \mu}{\sqrt{n}\sigma/n} \sim \mathcal{N}(0,1)$$

CLT works for
sum or average of R.V.s

$$\Downarrow$$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2), \quad \bar{X} - \mu \sim \mathcal{N}\left(0, \frac{1}{n}\sigma^2\right), \quad \bar{X} \sim \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$$

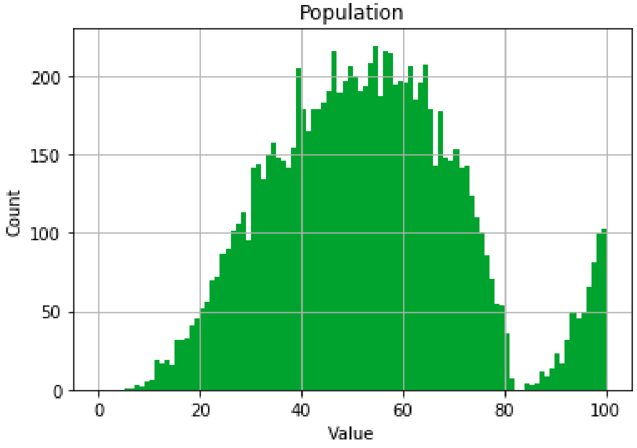
Different form of
normal variables by CLT

CLT explains a lot

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

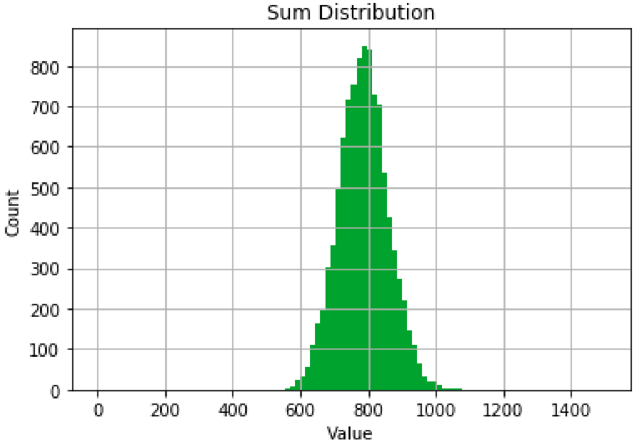
as $n \to \infty$

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

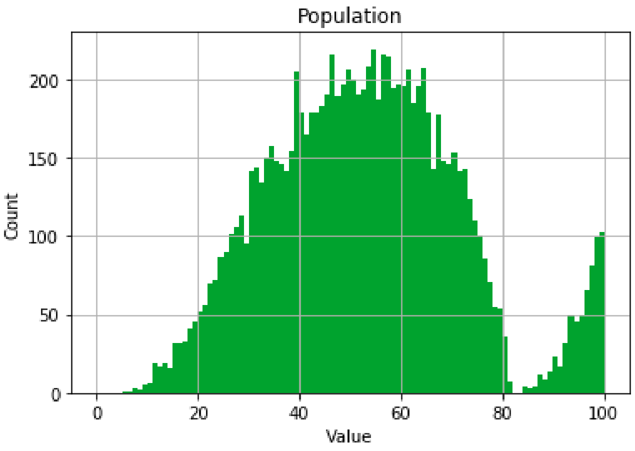
Not normally distributed, even with unknown pdf.

CLT explains a lot

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

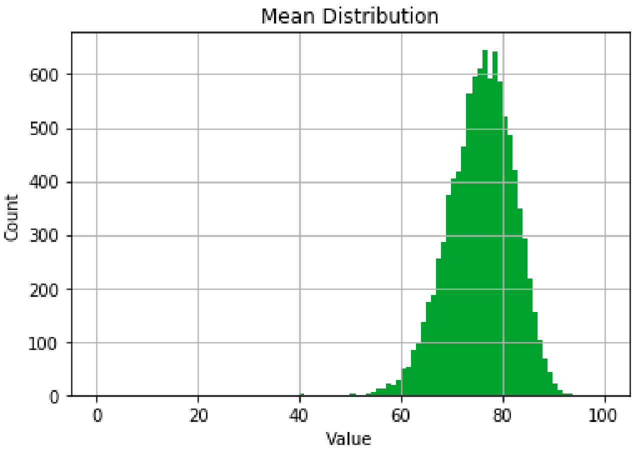
as $n \to \infty$

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



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

Not normally distributed, even with unknown pdf.

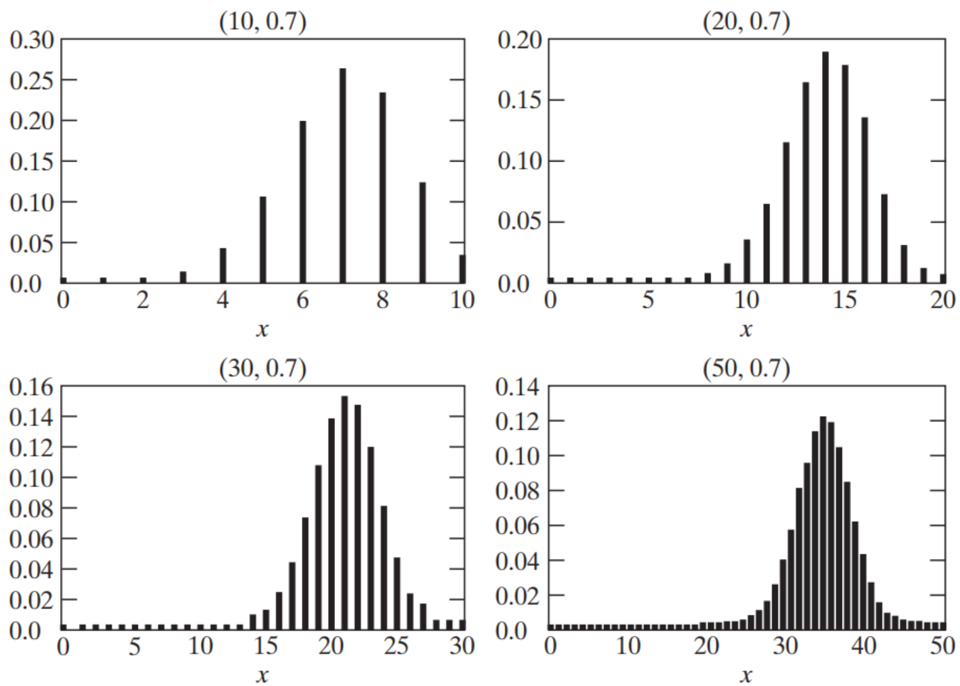
Normal Approximation to $b(n,p)$

The DeMoivre-Laplace limit theorem (拉普拉斯中心极限定理)

If S_n denotes the number of successes that occur when n independent trials, each resulting in a success with probability p, are performed, then, for any $a < b$,

$$P\left\{a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right\} \to \Phi(b) - \Phi(a), \quad \text{as } n \to \infty.$$

(Special case: $\sum_{i=1}^{n} X_i$ follows **Binomial distribution**)



$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim \mathcal{N}(0,1)$$

The probability mass function of a binomial (n, p) random variable becomes more and more "normal" as n becomes larger and larger.

Website testing

- 100 people are given a new website design with A/B test.
- $X = \#$ people whose time on the website increases.
- The design actually has no effect, so P(time on site increases) = 0.5.
- CEO will endorse the new design if $X \geq 65$.

What is P(CEO endorses change)? *Give a numerical approximation*.

Approach 1: Binomial

Define

$X \sim b(n = 100, p = 0.5)$

$P(X \geq 65)$

$= \sum_{i=65}^{100} C_{100}^{i} 0.5^{i} 0.5^{100-i}$

≈ 0.0018

Approach 2: Approximate with Normal

Define & Approx.

$Y \sim \mathcal{N}(\mu = np = 50, \sigma^2 = np(1-p) = 5)$

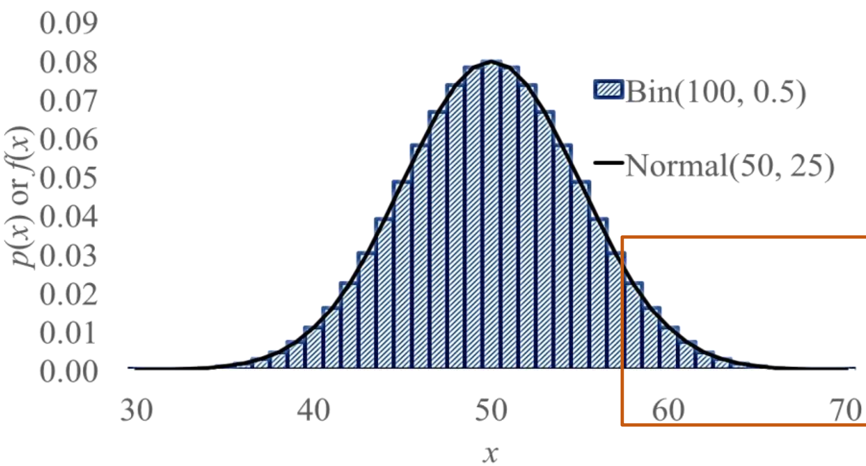
$P(X \geq 65) \approx P(Y \geq 65) = 1 - F_Y(65)$

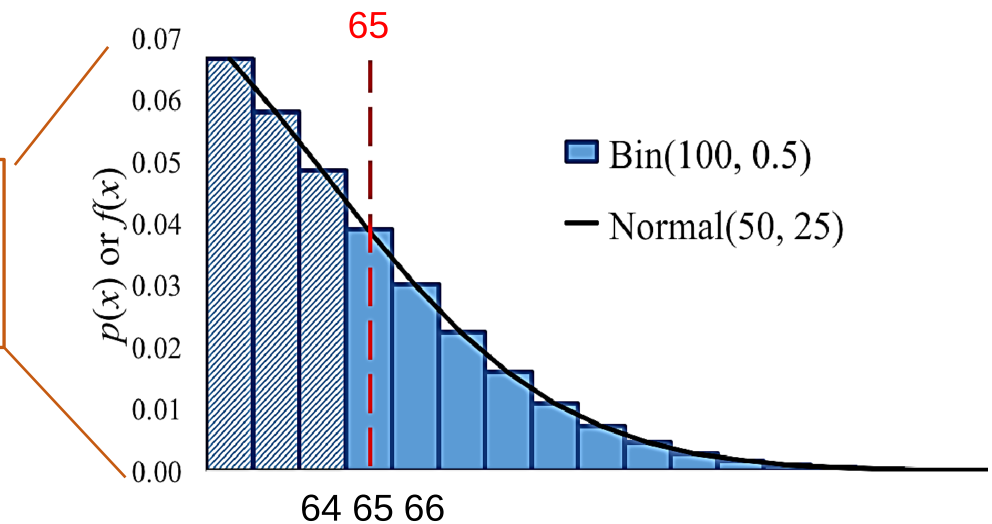
$= 1 - \Phi\left(\frac{65-50}{5}\right) = 1 - 0.9987 = 0.0013$??!

Something is wrong!

Website testing (with continuity correction)

In our website testing, $Y \sim \mathcal{N}(50, 25)$ approximates $X \sim b(100,0.5)$.





Approach 2: Approximate with Normal

You must perform a continuity correction when approximating a Binomial R.V. with a Normal R.V.

Define

$P(X \geq 65)$ Binomial

$\approx P(Y \geq 64.5) = 1 - F_Y(64.5)$ Normal

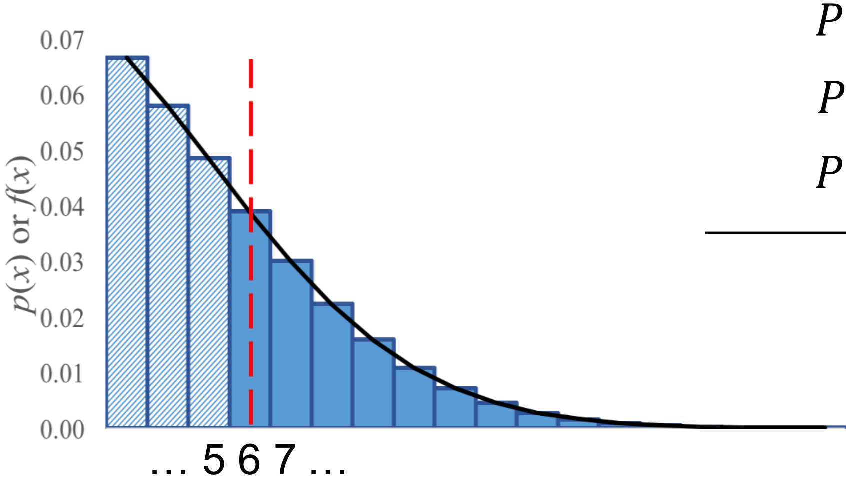
$= 1 - \Phi\left(\frac{64.5-50}{5}\right) = 0.0019$

Approach 2 is better than Approach 1.
Easier to compute!

Continuity correction

If $Y \sim \mathcal{N}\big(np, np(1-p)\big)$ approximates $X \sim b(n,p)$, how do we approximate the following probabilities?

Discrete (e.g., Binomial) probability	Continuous (Normal) probability
$P(X = 6)$	$P(5.5 \leq Y \leq 6.5)$
$P(X \geq 6)$	$P(Y \geq 5.5)$
$P(X > 6)$	$P(Y \geq 6.5)$
$P(X < 6)$	$P(Y \leq 5.5)$
$P(X \leq 6)$	$P(Y \leq 6.5)$



Example of CLT: Dice game

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

as $n \to \infty$

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 probability (according to the CLT)…

1. Define R.V.s and state goal.

$$E[X_i] = 3.5$$
$$D[X_i] = 35/12$$

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

Approximate:

$$X \approx Y \sim \mathcal{N}\left(10 \cdot 3.5, 10 \cdot \frac{35}{12}\right)$$

2. Solve.

$P(Y \leq 25.5) + P(Y \geq 44.5)$ or $1 - P(25.5 \leq Y \leq 44.5)$

continuity correction

Example of CLT: Dice game

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

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 probability (according to the CLT)…

1. Define R.V.s and state goal.

$$E[X_i] = 3.5$$
$$D[X_i] = 35/12$$

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

Approximate:

$$X \approx Y \sim \mathcal{N}\left(10 \cdot 3.5, 10 \cdot \frac{35}{12}\right)$$

2. Solve.

$$P(Y \leq 25.5) + P(Y \geq 44.5) = \Phi\left(\frac{25.5-35}{\sqrt{10(35/12)}}\right) + \left(1 - \Phi\left(\frac{44.5-35}{\sqrt{10(35/12)}}\right)\right)$$
$$\approx \Phi(-1.76) + \big(1 - \Phi(1.76)\big) \approx (1 - 0.9608) + (1 - 0.9608) = 0.0784$$

Ex. 4g (P208) Let X be the number of times that a fair coin that is flipped 40 times lands on heads. Find the probability that $X = 20$. Use the normal approximation and then compare it with the exact solution.

Ex. 4g (P208) Let X be the number of times that a fair coin that is flipped 40 times lands on heads. Find the probability that $X = 20$. Use the normal approximation and then compare it with the exact solution.

Sol.

$$Y \sim \mathcal{N}(40 \cdot 0.5, 40 \cdot 0.5 \cdot 0.5)$$

$$P\{X = 20\} \approx P\{19.5 < Y < 20.5\} = P\left\{\frac{19.5-20}{\sqrt{10}} < \frac{Y-20}{\sqrt{10}} < \frac{20.5-20}{\sqrt{10}}\right\}$$
$$\approx P\left\{-0.16 < \frac{Y-20}{\sqrt{10}} < 0.16\right\}$$
$$= \Phi(0.16) - \Phi(-0.16) = 0.1272$$

Exact result: $P\{X = 20\} = C_{40}^{20} 0.5^{40} \approx 0.1254$

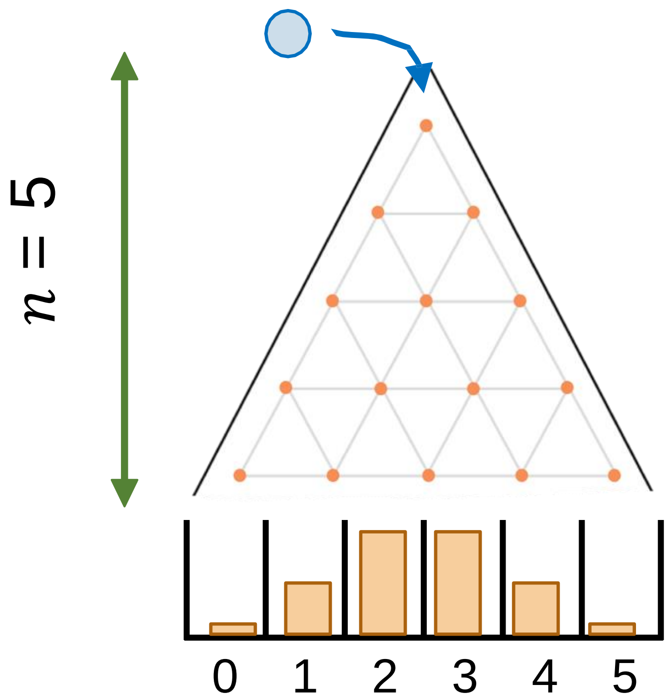
Normal approximation is accurate **when $np(1-p)$ is large**.

CLT explains a lot

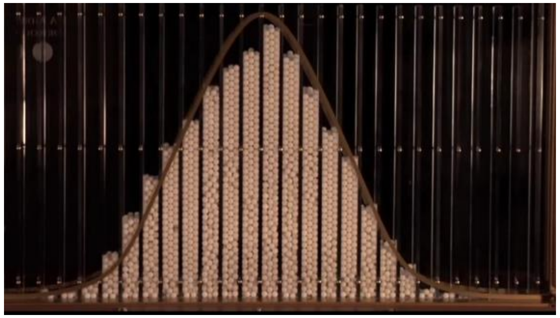
$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

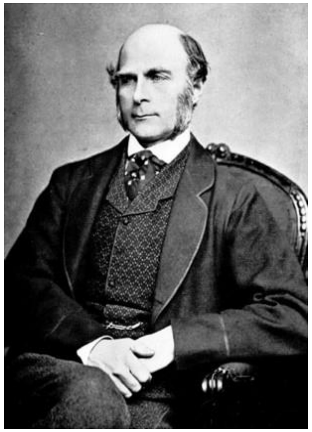
as $n \to \infty$

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)





Example of CLT: Crashing website

- Let $X = \#$ visitors per minute to a website, where $X \sim \pi(100)$.
- 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?

Example of CLT: Crashing website

- Let $X = \#$ visitors per minute to a website, where $X \sim \pi(100)$.
- The server crashes if there are ≥ 120 requests/minute.

What is P(server crashes in next minute)?

<u>Strategy</u>:
Poisson (exact)

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

<u>Strategy</u>:
CLT
(approx.)

State approx. goal:

$X = X_1 + \cdots X_i + \cdots + X_n$,
where $X_i = 0$ or 1
$X \sim \pi(100)$

$X \approx Y \sim \mathcal{N}(\mu, \sigma^2)$
$\mu = \sigma^2 = 100$

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

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

Ex. There are 3000 people of the same age participating in an insurance program. The mortality rate of these people in a year is 0.1%. The insurance premium is 10 yuan, and the family members can receive 2,000 yuan from the insurance company in the event of death.

Find the probability that

(1) the insurance company makes a profit of not less than 10,000 yuan in a year.

(2) the insurance company gains no profit from this program.

Hint: no continuity correction is needed since n is large.

Ex. There are 3000 people of the same age participating in an insurance program. The mortality rate of these people in a year is 0.1%. The insurance premium is 10 yuan, and the family members can receive 2,000 yuan from the insurance company in the event of death.

Find the probability that

(1) the insurance company makes a profit of not less than 10,000 yuan in a year.

Sol. Let X denotes the number of mortalities in one year.

$X \sim b(3000, 0.001)$, $\mu = 3$, $\sigma^2 \approx 2.997$.

The profit can be computed by $3000 \times 10 - 2000 \cdot X$.

Ex. There are 3000 people of the same age participating in an insurance program. The mortality rate of these people in a year is 0.1%. The insurance premium is 10 yuan, and the family members can receive 2,000 yuan from the insurance company in the event of death.

Find the probability that

(1) the insurance company makes a profit of not less than 10,000 yuan in a year.

$$\Pr\{\text{Profit} \geq 10000\} = P\{3000 \times 10 - 2000 \cdot X \geq 10000\} = P\{0 \leq X \leq 10\}$$

From CLT (no continuity correction is needed),

$$P\{0 \leq X \leq 10\} = P\left\{\frac{0-3}{1.7312} \leq \frac{X-3}{1.7312} \leq \frac{10-3}{1.7312}\right\}$$

$$\approx \Phi(4.043) - [1-\Phi(1.733)] = 0.96$$

Ex. There are 3000 people of the same age participating in an insurance program. The mortality rate of these people in a year is 0.1%. The insurance premium is 10 yuan, and the family members can receive 2,000 yuan from the insurance company in the event of death.

Find the probability that

(2) the insurance company gains no profit from this program.

$$P\{\text{Profit} < 0\} = P\{3000 \times 10 - 2000 \cdot X < 0\} = P\{X > 15\}$$

From CLT,

$$P\{X > 15\} = 1 - P\{0 \leq X \leq 15\} = 1 - P\left\{\frac{0-3}{1.7312} \leq \frac{x-3}{1.7312} \leq \frac{15-3}{1.7312}\right\}$$

$$\approx 1 - \left[\Phi\left(\frac{12}{1.7312}\right) - \Phi\left(\frac{-3}{1.7312}\right)\right] = 1 - 0.9582 = 0.0418$$

CLT for independent R.V.s

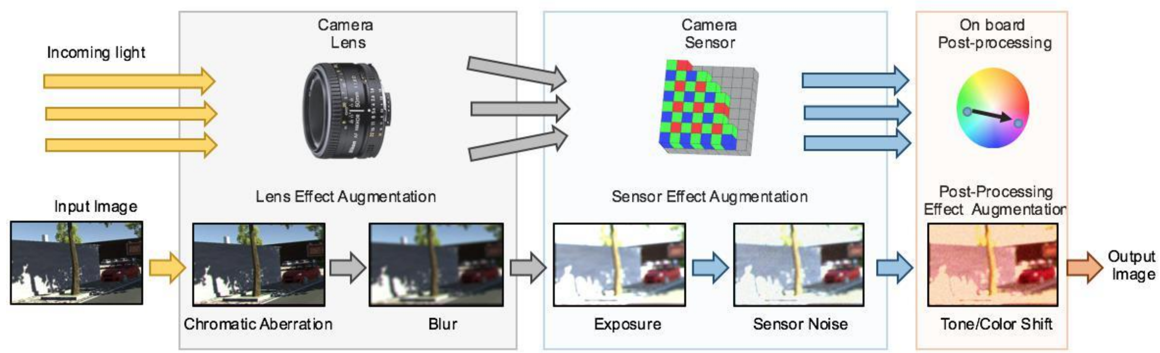
$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim \mathcal{N}(0,1)$$

Let $X_1, X_2, \ldots$ be a sequence of independent R.V.s having respective means and variances μ_i and σ_i^2. Under certain conditions (No X_i is dominating P405 of textbook).

$$P\left\{\frac{\sum_{i=1}^{n}(X_i-\mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \le a\right\} \to \Phi(a) \ \text{ as } \ n \to \infty$$

CLT for independent but not identically distributed variables.
Useful in practice.

Example:
Noise in digital image.



Quiz 4 (LAST ONE)

Date: 2-Dec-2024

Scope: statistics of R.V.s, C.L.T.

Open-book exam, with fill-in-the-blank and multiple-choice questions.

Time: ~50 mins