THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

AMA 2233

Data Analytics and Visualization

Yancheng Yuan

Department of Data Science and Artificial Intelligence

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

Recap of last lecture

The sample mean and (weak) Law of Large Numbers



Sample mean: For independent and identically distributed (i.i.d.) random variables $X_1, X_2, \dots, X_n$, the sample mean is defined as

$$\overline{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}.$$

Question: Does the sample mean a good approximation to the expectation?

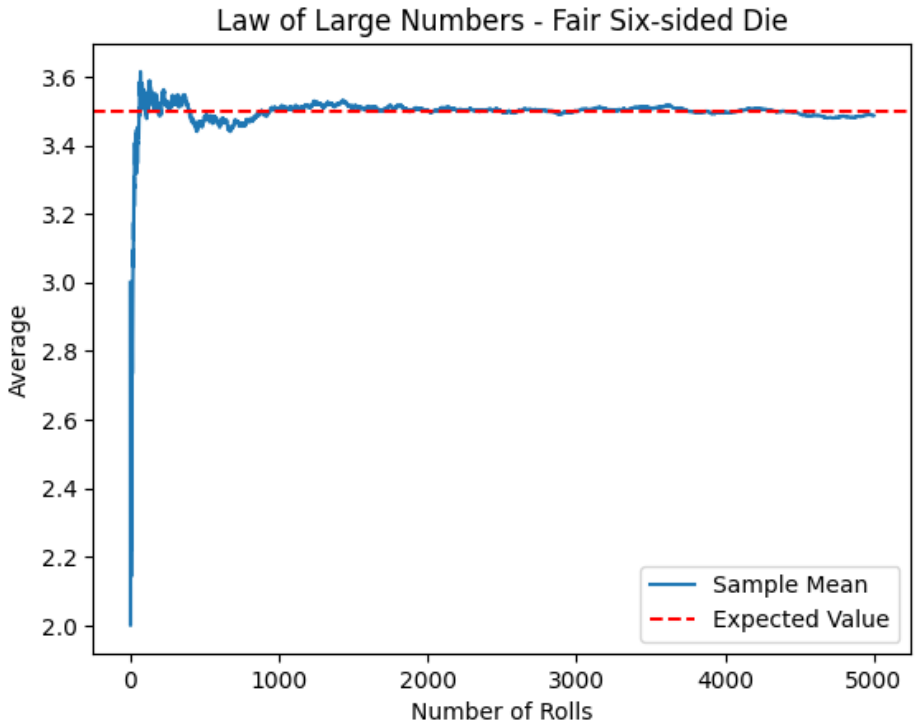
Weak Law of Large Numbers: The sample mean converges in probability to the expected value of a random variable:

$$\overline{X}_n \xrightarrow{P} E[X] =: \mu \text{ , as } n \to \infty.$$

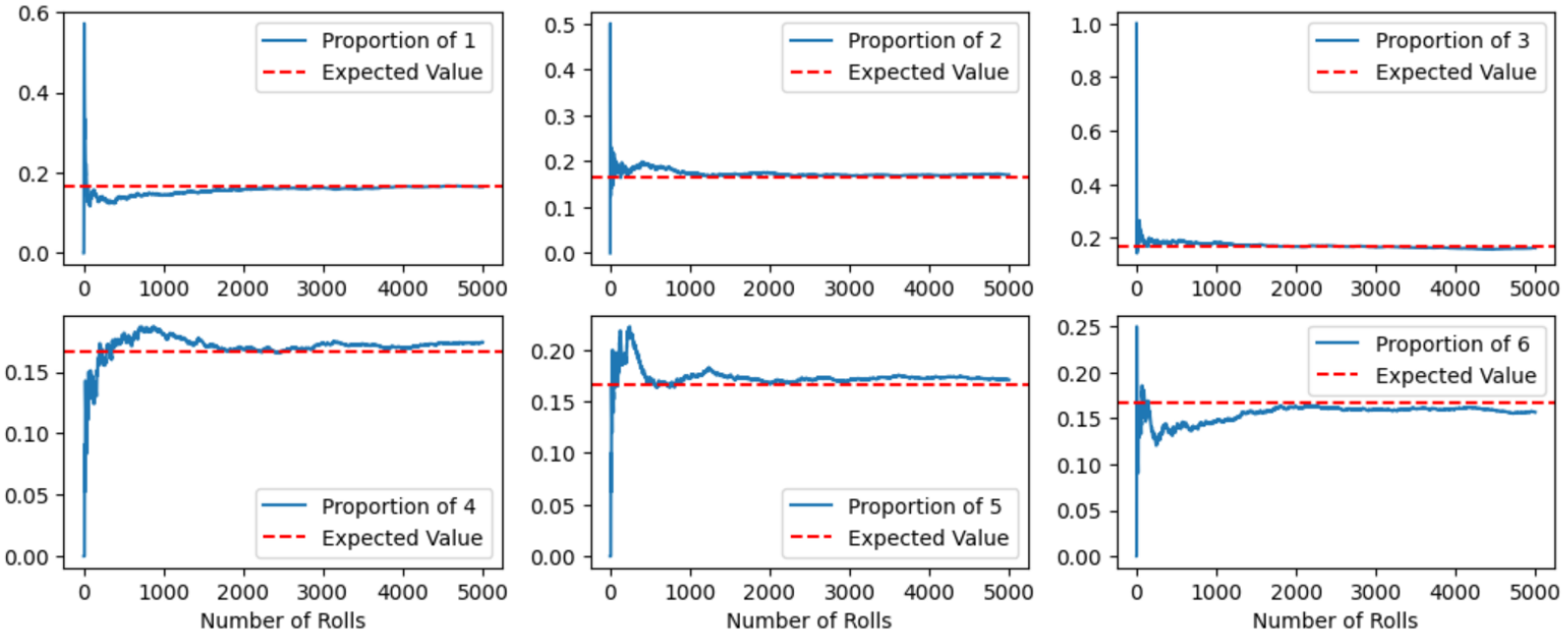
In other words, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{X_n} - \mu| < \varepsilon) = 1.$$

Visualization of sample mean



Fairness of the dice



Agenda

1. A quantitative bound on the approximation of the expectation by sample mean.
2. What will the sum of i.i.d. random variables converge to?
3. An introduction to the Central Limit Theorem
4. Understanding CLT from the visualization
5. An application of CLT.

A quantitative bound on the approximation

A quantitative result on the approximation

Chebyshev's inequality: Let X be a random variable with finite expected value μ and non-zero variance σ^2. Then, for any $t > 0$, we have

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Example (Fair Coin Toss): We flip a fair coin n = 1000 times and denote S_n as the number of Heads in the n experiments. Denote $\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}$ as the proportion for Heads. We know that

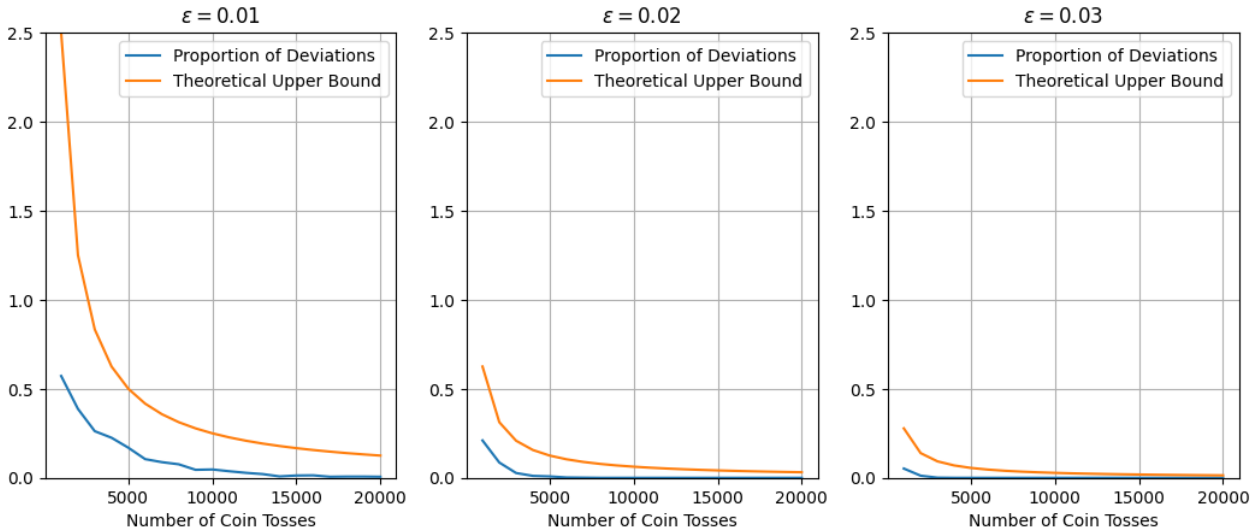
$$E[\overline{X}_n] = 0.5, \ Var[\overline{X}_n] = \sigma^2 = 1/4000.$$

Then, we know

$$P(|\overline{X}_{1000} - 0.5| \geq 0.1) \leq \frac{\sigma^2}{0.01} = 0.025.$$

Visualization

Experiment setting: Consider tossing a fair coin n times. The number of heads follows a binomial distribution. We select some numbers of coin tosses n. For each n, we perform 1000 independent random experiments.



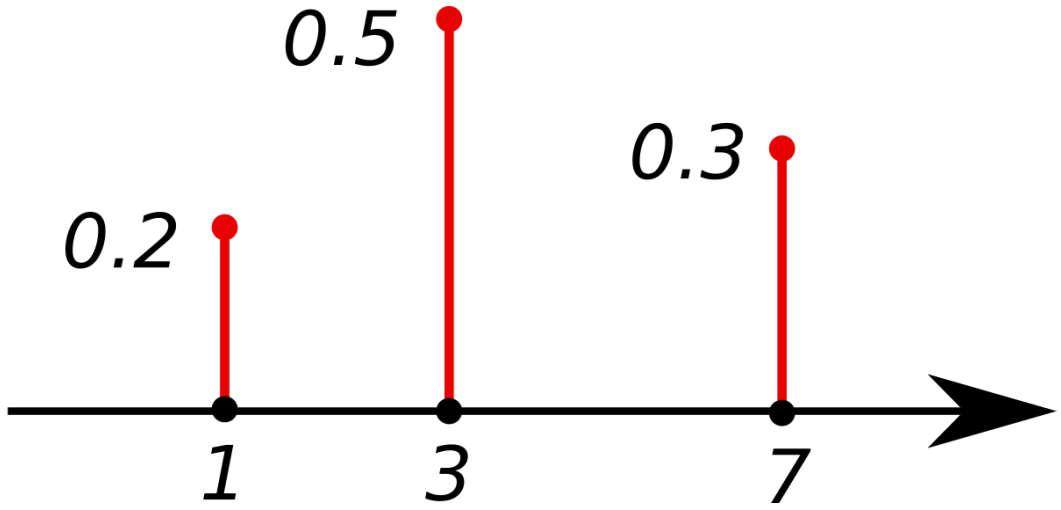
Question: What can we say about the random variable

$$S_n - n\mu \, ?$$

We need the Central Limit Theorem.

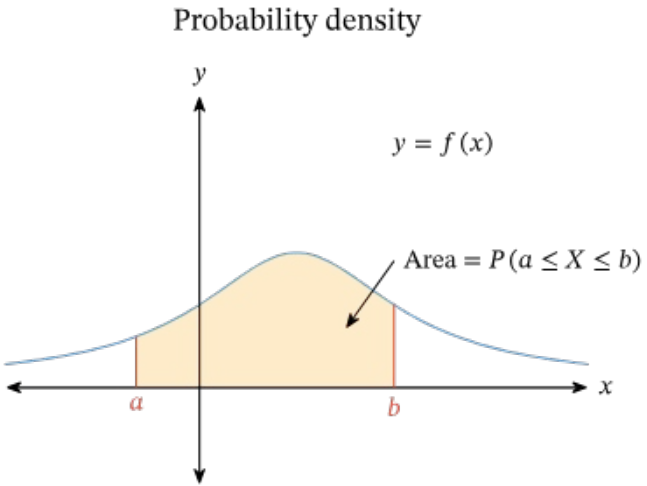
The sum of random variables

Probability mass function of a discrete random variable



Probability mass function: The mass function of a discrete random variable X is the function $f: R \to [0, 1]$ given by $f(x) = P(X = x)$.

Probability density function of a continuous random variable



Probability density function: An integrable function $f: R \to [0, \infty)$ is the probability density function of the continuous random variable X if its distribution function $F(x) = P(X \leq x)$ can be written as

$$F(x) = \int_{-\infty}^{x} f(u)\, du.$$

But, how to visualize a density function of $S_n - n\mu$?

Step 1: Fix a large integer n

Step 2: Randomly sample $X_1, X_2, \ldots, X_n$ and compute $S_n - n\mu$

Step 3: Repeat Step 2 for N times, and visualize the proportion result in the histogram.

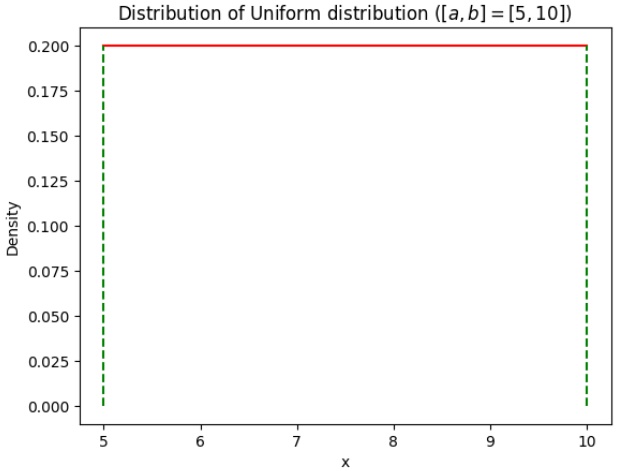
Visualization of $S_n - n\mu$ for uniform distribution

Uniform distribution on [a, b]: The density function for a uniformly distributed random variable X is

$$f(x) = \frac{1}{b - a}$$

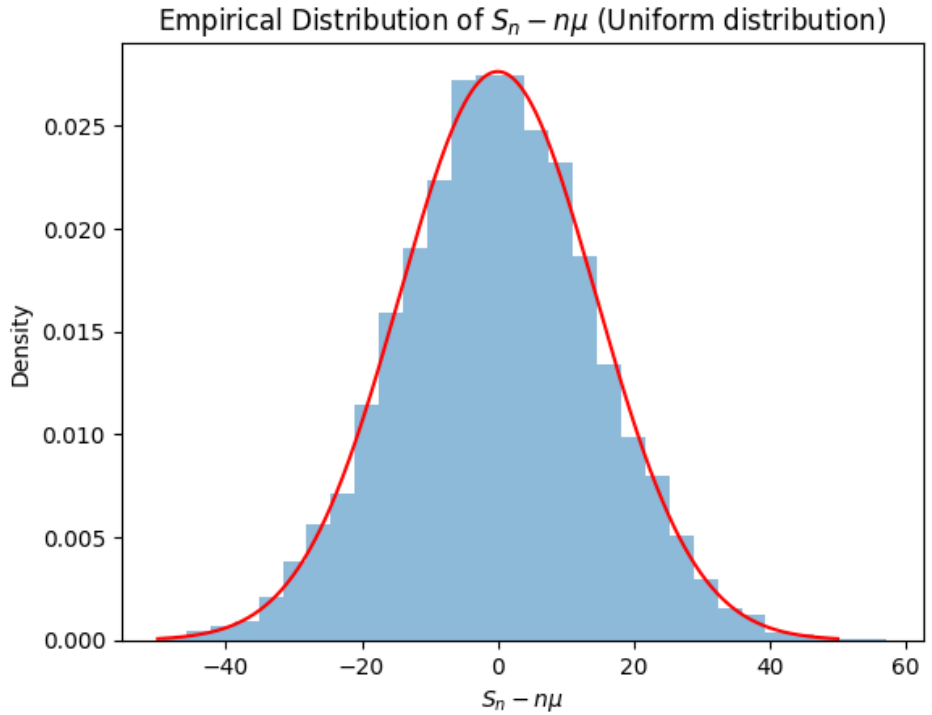
Then

$$\mu = \frac{a + b}{2}, \qquad \sigma^2 = \frac{(b - a)^2}{12}$$



Visualization of $S_n - n\mu$ for uniform distribution

Visualization setting: n = 100, number of random experiments N = 10000.



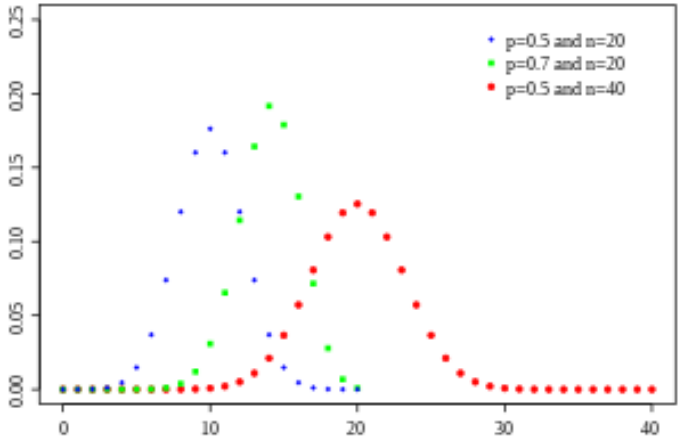
Visualization of $S_n - n\mu$ for binomial distribution

Binomial distribution: The probability mass function for a binomial distribution is

$$\mathrm{P}(x = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

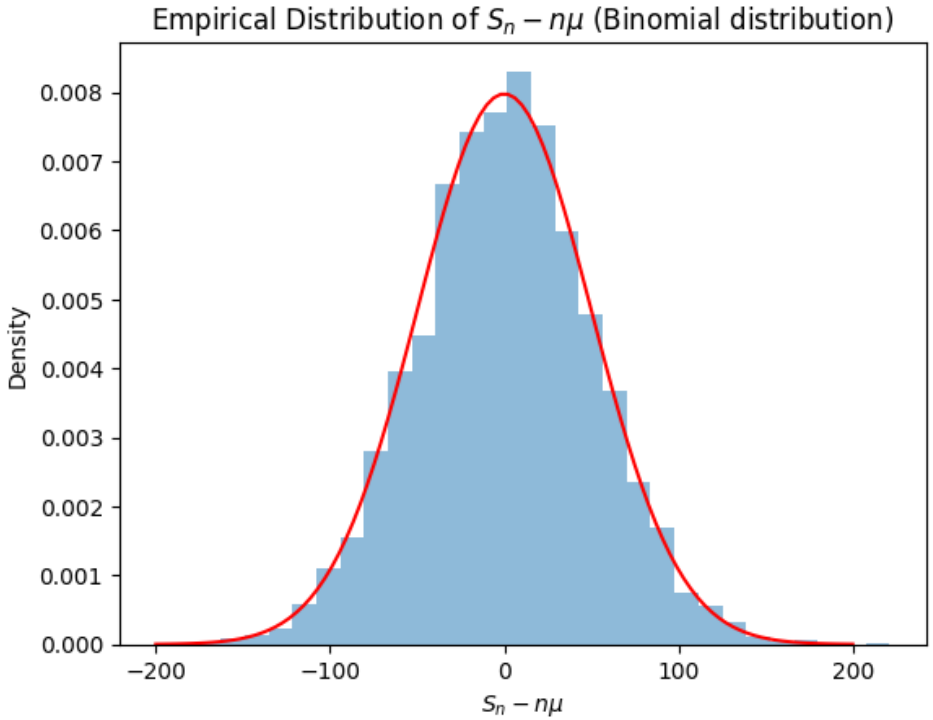
Then

$$\mu = np, \qquad \sigma^2 = np(1-p)$$



Visualization of $S_n - n\mu$ for binomial distribution

Visualization setting: p = 0.5, n = 100, sample size = 100, number of random experiments N = 10000.



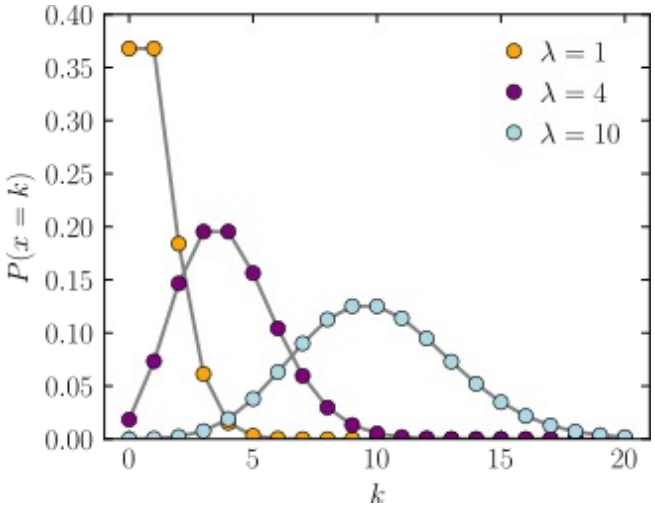
Visualization of $S_n - n\mu$ for Poisson distribution

Poisson distribution: A random variable X is said to have a Poisson distribution, with parameter $\lambda > 0$, if it has a probability mass function given by

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

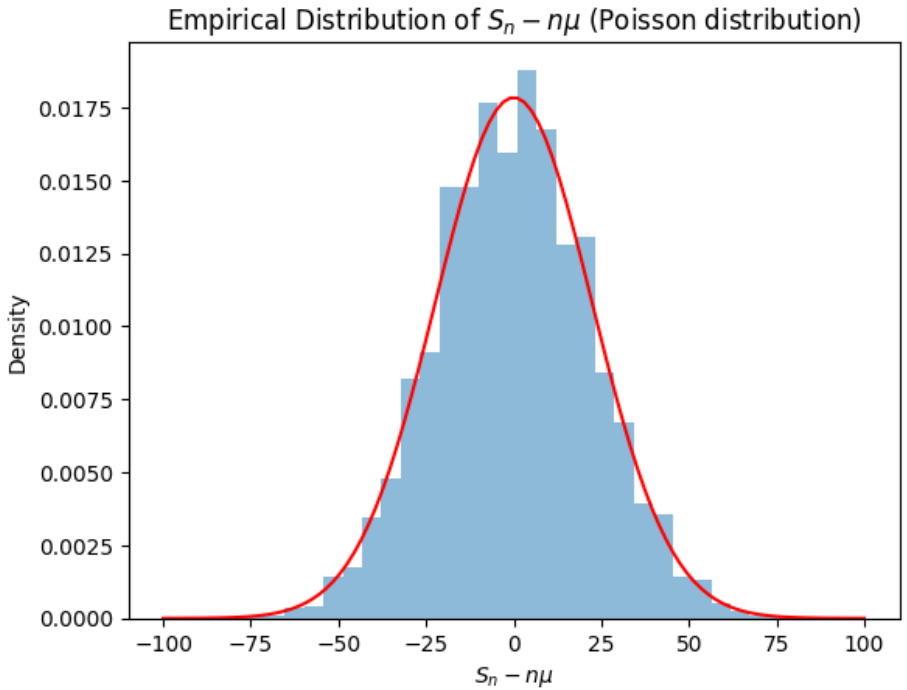
Then,

$$\mu = \sigma^2 = \lambda.$$



Visualization of $S_n - n\mu$ for Poisson distribution

Visualization setting: $\lambda = 5$, sample size = 100, number of random experiments N = 10000.



Some observations

1. $\sigma(S_n - n\mu)$ grows in the scale of $\sqrt{n}$.

2. The density function of $S_n - n\mu$ looks like a bell curve.



The Central Limit Theorem

Central Limit Theorem

Central Limit Theorem: Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with finite mean μ and finite non-zero variance σ^2, and let

$$\boldsymbol{S_n = X_1 + X_2 + \cdots + X_n}.$$

Then

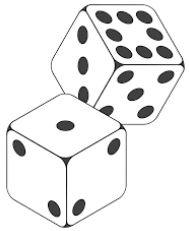
$$\frac{\boldsymbol{S_n - n\mu}}{\sqrt{\boldsymbol{n\sigma^2}}} \xrightarrow{\boldsymbol{D}} \boldsymbol{N(0,1)} \quad \text{as } n \to \infty.$$

Remark

1. **No requirements on the distribution of X.**
2. **The variance σ^2 should be finite and non-zero.**
3. **The sample size n should be large enough.**

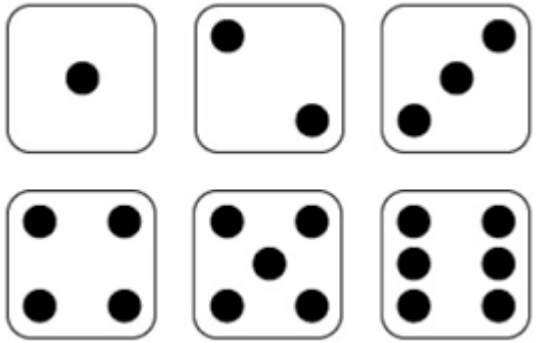
Distribution doesn’t matter.

Example

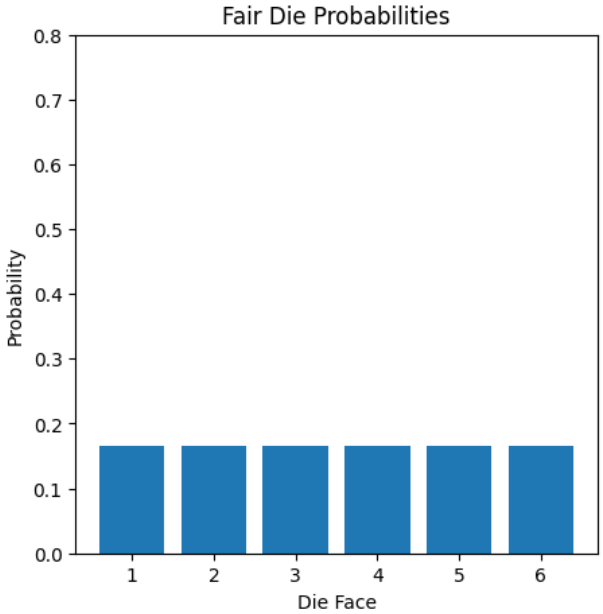


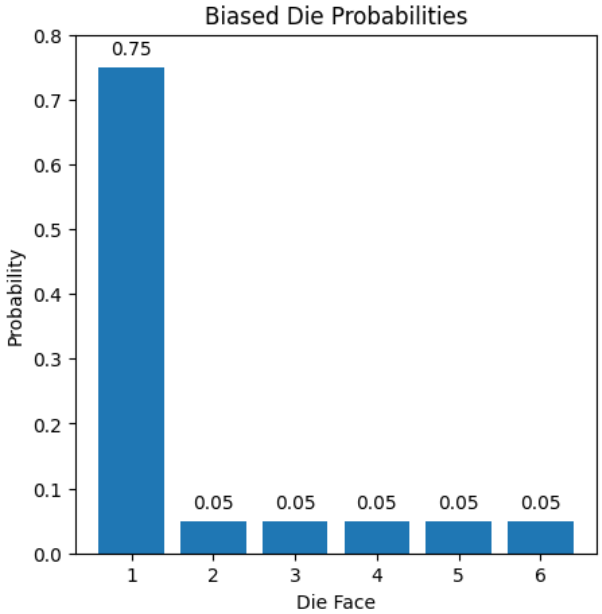
Dice Rolling: Rolling a fair and an unfair dice.

Possible outcomes:

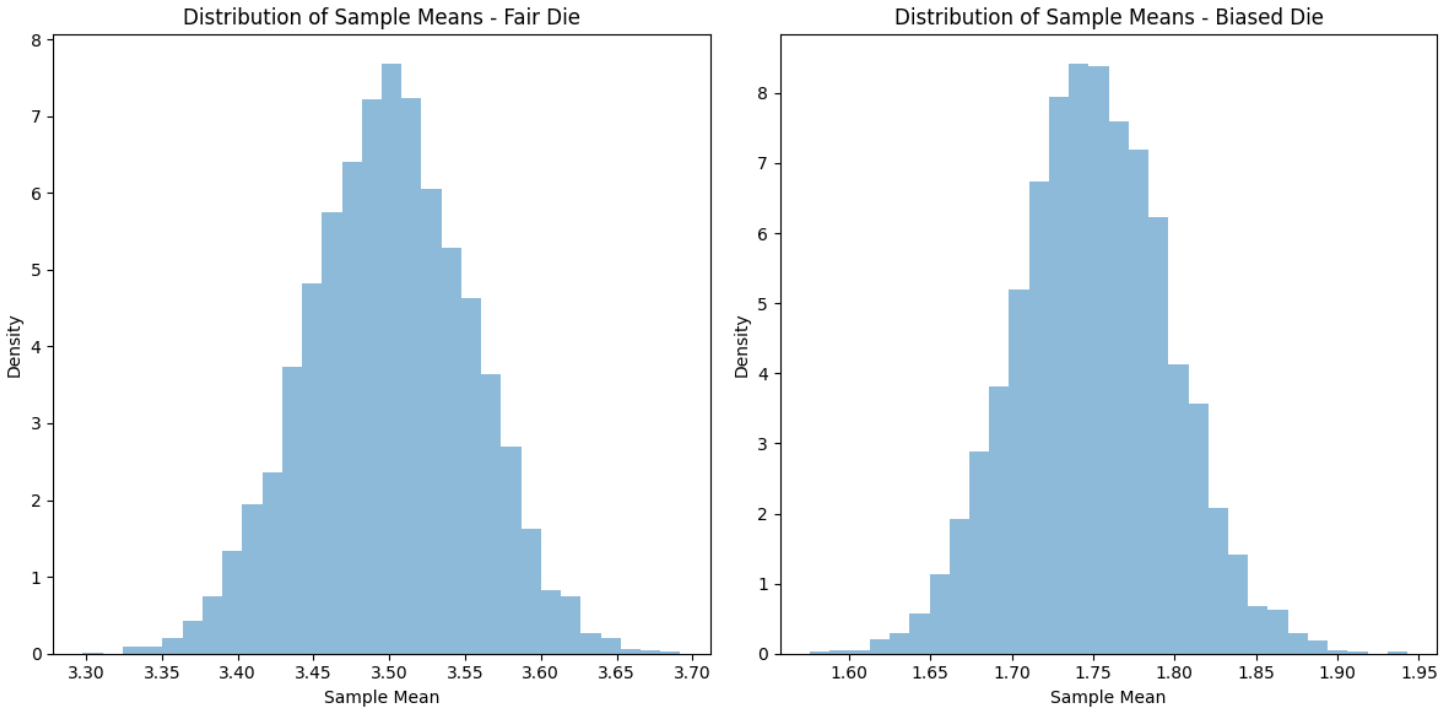


Two scenario



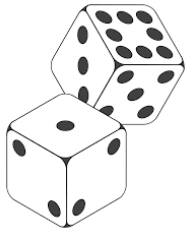


Results

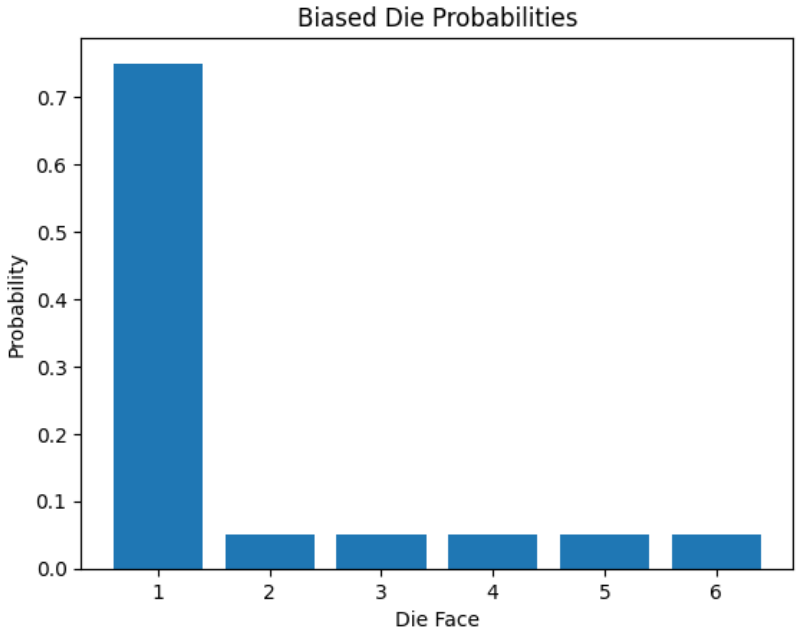

Distribution of Sample Means - Fair Die
Density
Sample Mean
0
1
2
3
4
5
6
7
8
3.30
3.35
3.40
3.45
3.50
3.55
3.60
3.65
3.70
Distribution of Sample Means - Biased Die
Density
Sample Mean
0
1
2
3
4
5
6
7
8
1.60
1.65
1.70
1.75
1.80
1.85
1.90
1.95

Sample size n matters

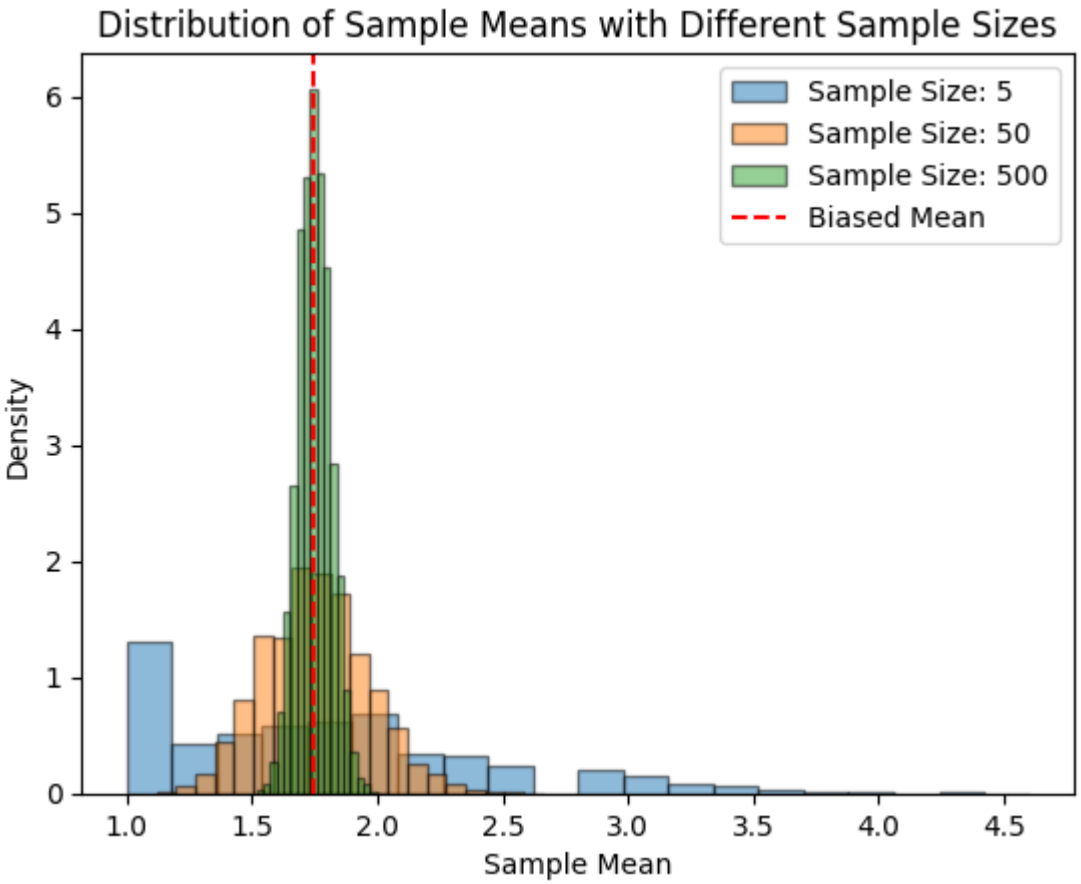
Example



Dice Rolling: Rolling a biased dice.



Results


Distribution of Sample Means with Different Sample Sizes
Sample Size: 5
Sample Size: 50
Sample Size: 500
Biased Mean
Density
Sample Mean
0
1
2
3
4
5
6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

An application

Example

Optimize the weekly restocking of seltzer water

sales_dataset.csv

sales_dataset.csv > data

1	sales
2	59
3	17
4	3
5	14
6	77
7	7
8	88
9	91
10	13
11	15
12	48
13	33
14	5
15	6
16	5
17	47

A glimpse into a small portion of the sales dataset.

Investigate every store carefully?

😅 Time consuming!

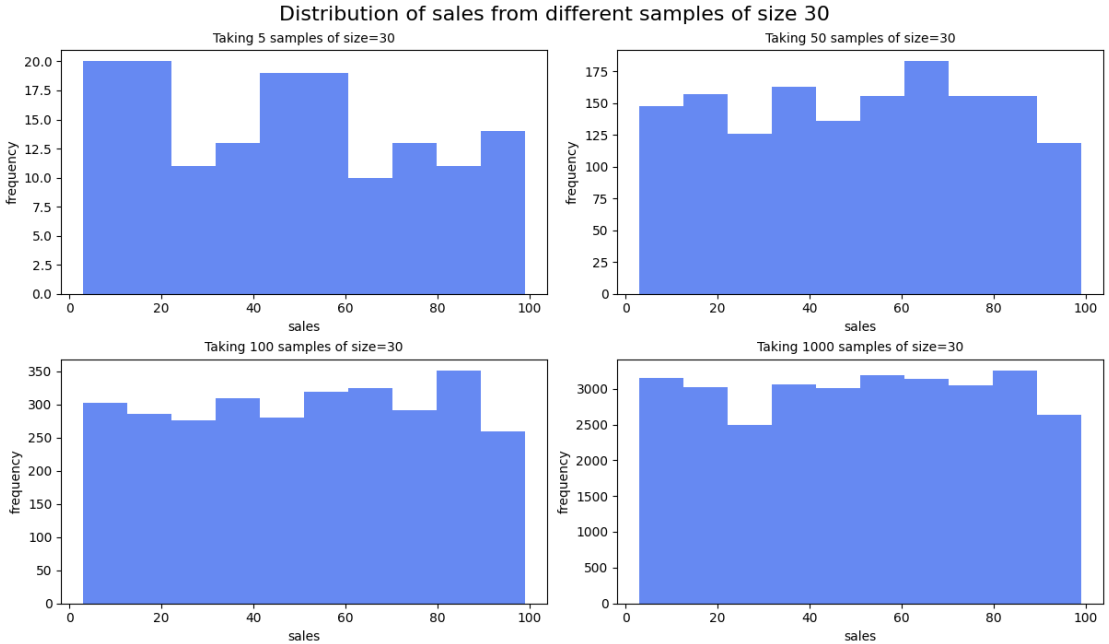
Central Limit Theorem 😁

#restocking = Mean * #stores



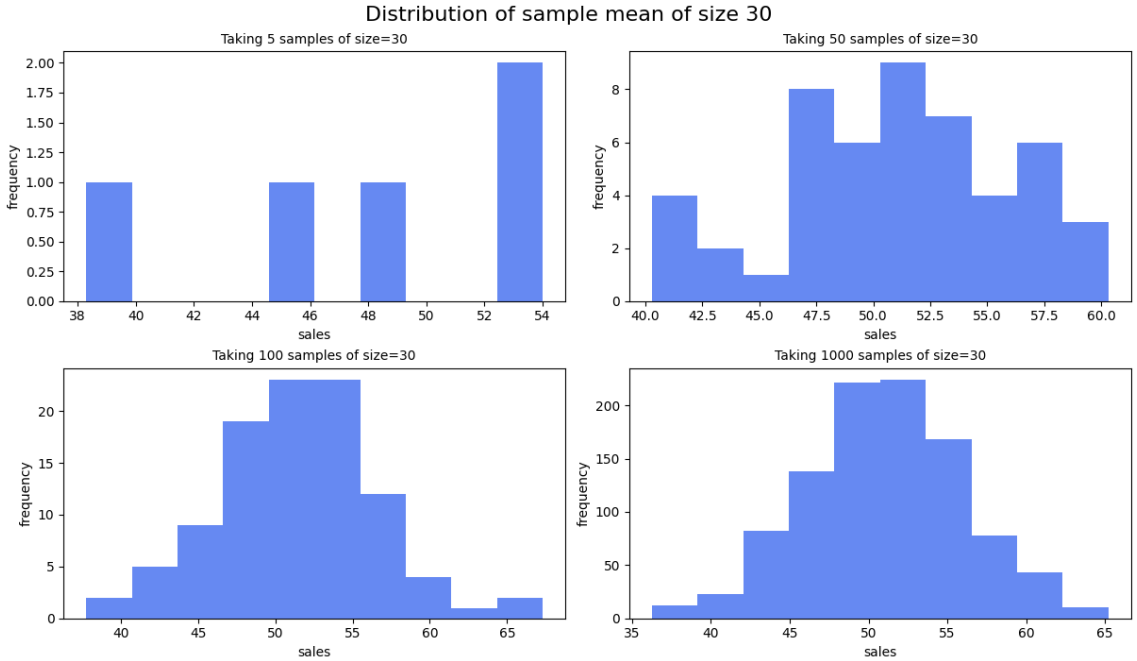
Example

Randomly choose 30 stores and repeat the experiments ‘n’ times;
plot the frequency of sales.



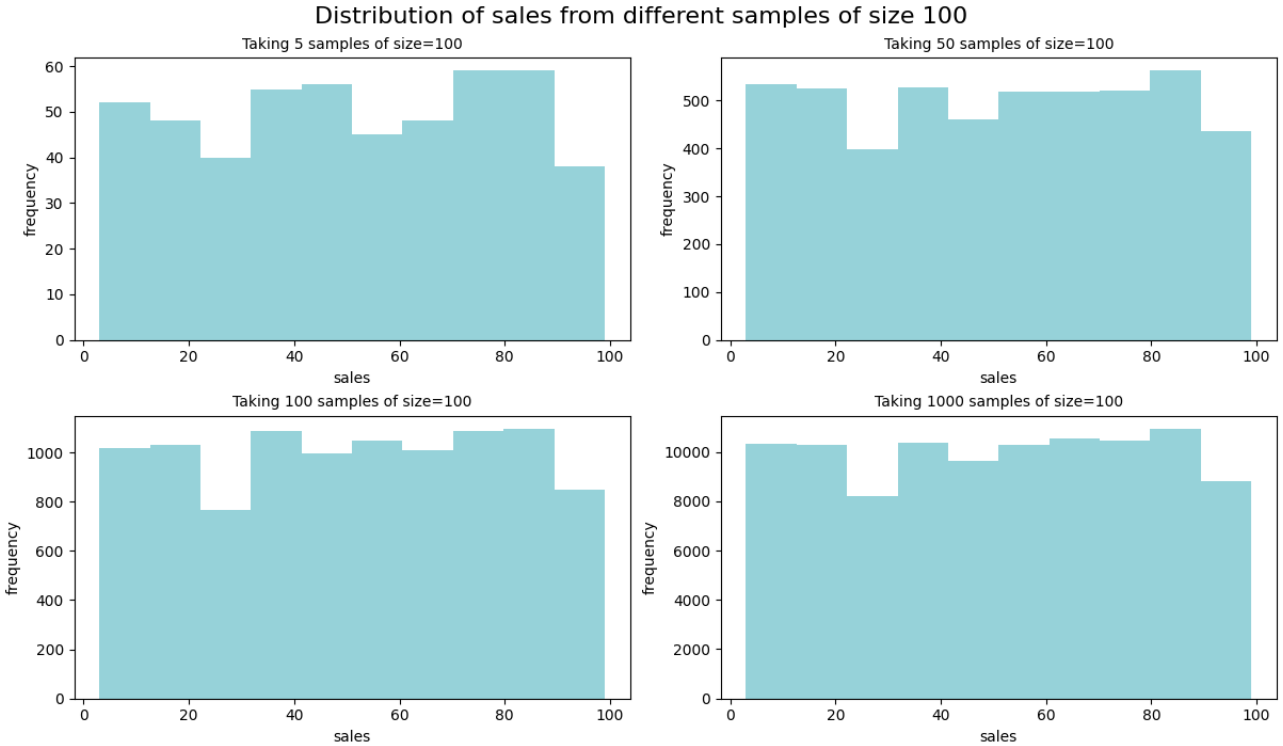
Example

For each experiment, calculate the sample mean and plot the distribution

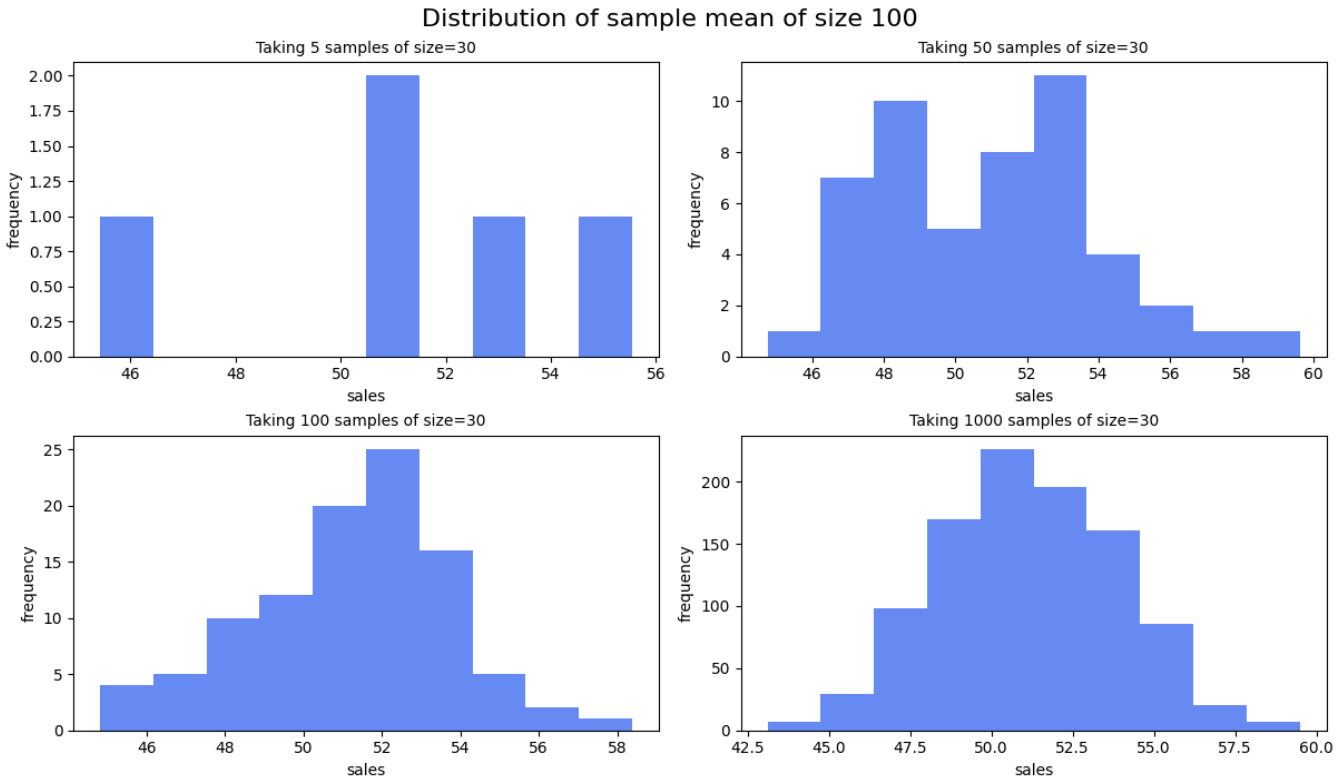


Example

Randomly choose 100 stores instead!

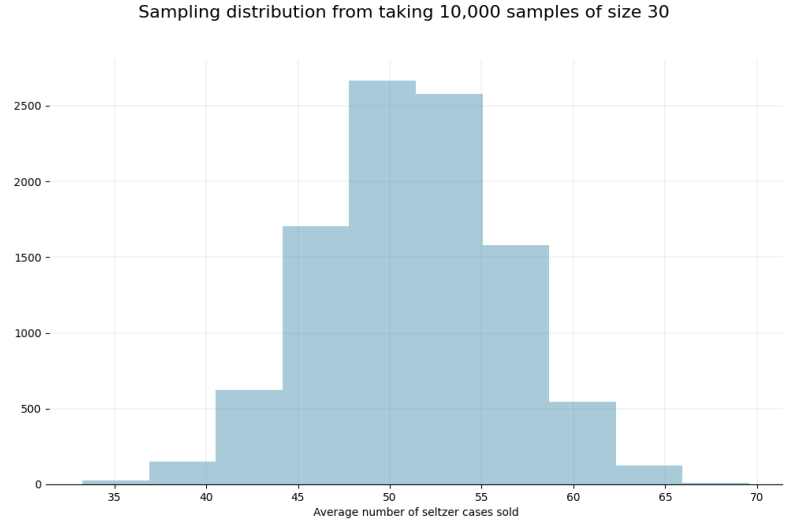


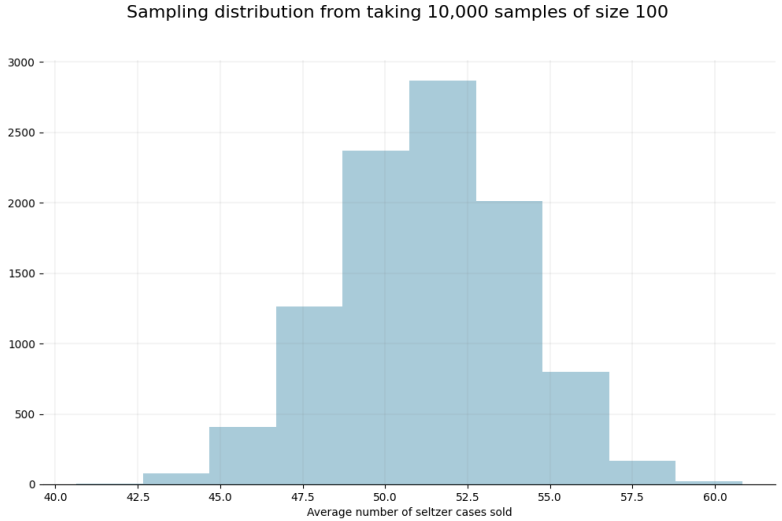
Example


Distribution of sample mean of size 100
Taking 5 samples of size=30
Taking 50 samples of size=30
Taking 100 samples of size=30
Taking 1000 samples of size=30
frequency
sales

Example

How about more samples?





Results

Choose 51 as the sample mean!

It is perfectly aligned with the actual mean value of the dataset.



```
Summary statistics for the population (sales dataset)
count    4200.000000
mean       51.235952
std        27.631507
min         3.000000
25%        27.000000
50%        52.000000
75%        75.000000
max        99.000000
Name: sales, dtype: float64
```

Summary

1. An introduction to the Central Limit Theorem

2. Understanding CLT from the visualization

Further reading:
https://en.wikipedia.org/wiki/Central_limit_theorem


Thank You!