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

Agenda

1. General information
2. Overview of the course
3. Data visualization: Why and How
4. Introduction to the Law of Large Numbers
5. Understand the Law of Large Numbers with visualization

General Information

General Information

Dr YUAN Yancheng Office: **TU 821**
Email: **yancheng.yuan@polyu.edu.hk**

Lecture time: **Wednesday 11:30 -- 13:20.**
Tutorial time: **Monday 9:30 -- 10:20.**

Tutors:

Mr. Luo Yuanhang Email: chattelion.luo@connect.polyu.hk
Mr. Wang Runshu Email: 23036798r@connect.polyu.hk

Consultation Hours: TBD.

Assessment Methods

1. **Continuous Assessment** (**40%**)

   (i)Two assignments (10% each).
   (ii) Midterm (20%). **Tentative schedule: Week 8.**

2. **Examination (40%)**

3. **Project (20%)**

   Information and requirements of projects will be announced in Week 4.

Policy for late submission



All the deadlines for submissions are **strict**.

Tentative deadlines

Assig 1: **Release date: Sun of Week 4.** **Due date: Sun of Week 7**.

Assig 2: Release date: Sun of Week 6. **Due date: Sun of Week 10**.

Project: **Information Release date**: Lecture of Week 3.

Due date: **Sun of Week 13.**

Overview of the Course

Part I Basic visualization with applications in probability



1. **Visualization tools in Python**

2. **Understanding the Law of Large Numbers with visualization**

3. **Understanding the Central Limit Theorem with visualization**

4. **Monte Carlo methods with visualization**

Part II Time series data

1. **Time series data and its visualization**

2. **Time series data analysis and visualization with applications in financial data**

Part III Data Analytics: Regression analysis



1. **Linear regression with applications**

2. **Logistic regression with applications**

3. **Data preprocessing with visualization**

Part IV Graph data

1. **Graph data and its visualization.**

2. **Visualization and analysis of Erdős–Rényi model and social networks.**

Part V Dimension reduction for high-dimensional data



1. **High dimensional data.**

2. **Dimension reduction and visualization.**

Let’s Start!

“A Picture Is Worth a Thousand Words!”

Data visualization: Why and How

An Example

Year	Marriages	Divorces	Population	Marriages_per_1000	Divorces_per_1000
1910	948000	83000	92407000	10.3	0.9
1911	955000	89000	93863000	10.2	0.9
1912	1005000	94000	95335000	10.5	1
1913	1021000	91000	97225000	10.5	0.9
1914	1025000	101000	99111000	10.3	1
1915	1008000	104000	100546000	10	1
1916	1076000	114000	101961000	10.6	1.1
1917	1144000	122000	103414000	11.1	1.2
1918	1000000	116000	104550000	9.6	1.1
1919	1150000	142000	105063000	10.9	1.4
1920	1274000	171000	106461000	12	1.6
1921	1164000	160000	108538000	10.7	1.5
1922	1134000	149000	110049000	10.3	1.4
1923	1230000	165000	111947000	11	1.5
1924	1185000	171000	114109000	10.4	1.5
1925	1188000	175000	115829000	10.3	1.5
1926	1203000	185000	117397000	10.2	1.6
1927	1201000	196000	119035000	10.1	1.6
1928	1182000	200000	120509000	9.8	1.7
1929	1233000	206000	121878000	10.1	1.7
1930	1127000	196000	123188000	9.1	1.6
1931	1061000	188000	124149000	8.5	1.5
1932	982000	164000	124949000	7.9	1.3
1933	1098000	165000	125690000	8.7	1.3
1934	1302000	204000	126485000	10.3	1.6
1935	1327000	218000	127362000	10.4	1.7
1936	1369000	236000	128181000	10.7	1.8
1937	1451000	249000	128961000	11.3	1.9
1938	1331000	244000	129969000	10.2	1.9
1939	1404000	251000	131028000	10.7	1.9
1940	1596000	264000	132122000	12.1	2
1941	1696000	293000	133402000	12.7	2.2
1942	1772000	321000	134860000	13.1	2.4
1943	1577000	359000	136739000	11.5	2.6
1944	1452000	400000	138397000	10.5	2.9
1945	1613000	485000	139928000	11.5	3.5
1946	2291000	610000	141389000	16.2	4.3

Source from website.

An example



Source from website.

Four Keys in Data Visualization

1. Stop worrying about about tools for visualization

- **A good visualization can be achieved even with Excel.**
- **Story will dictate tools.**

Four Keys in Data Visualization

2. Clearly define the story

- **Before you look at the data.**
- **Blindly data mining = bad time.**
- **What should the visualization communicate?**

Four Keys in Data Visualization



3. Pick the right chart.

- **What chart tells your story better?**
- **Without adding too much additional information.**

Four Keys in Data Visualization

4. Assess your visualization

- **Can you immediate see the story?**
- **Are there distractions?**

Story-ink ratio

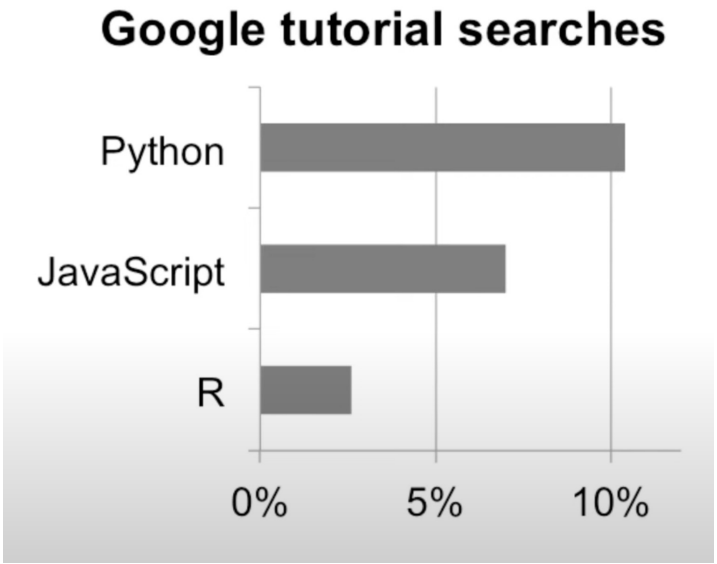
$$Story - ink\ ratio = \frac{Stroy\ ink}{Total\ ink\ used\ to\ print\ the\ graphic}$$

- **We should maximize the story-ink ratio when we do the visualization.**

Choose the Right Chart

Bar Chart

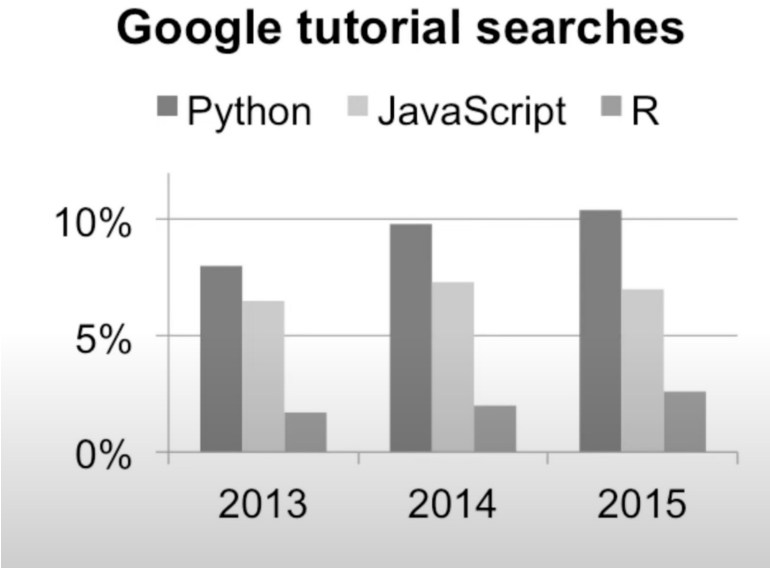
- **Most versatile chart type.**
- **Good for comparisons among categories.**



Source from website.

Bar Chart

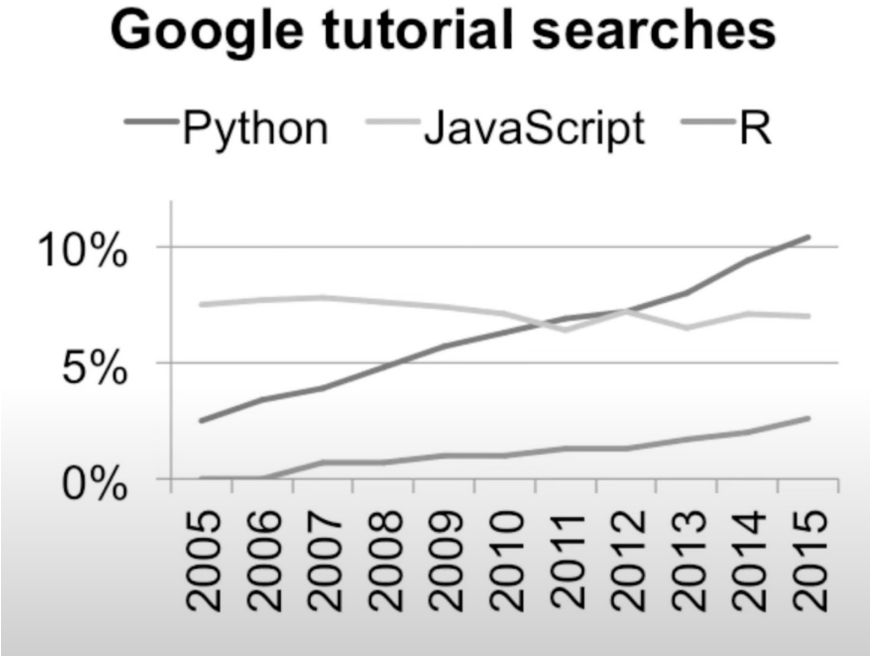
- **Most versatile chart type.**
- **Good for comparisons among categories.**
- **Comparison over short period of time.**



Source from website.

Line Chart

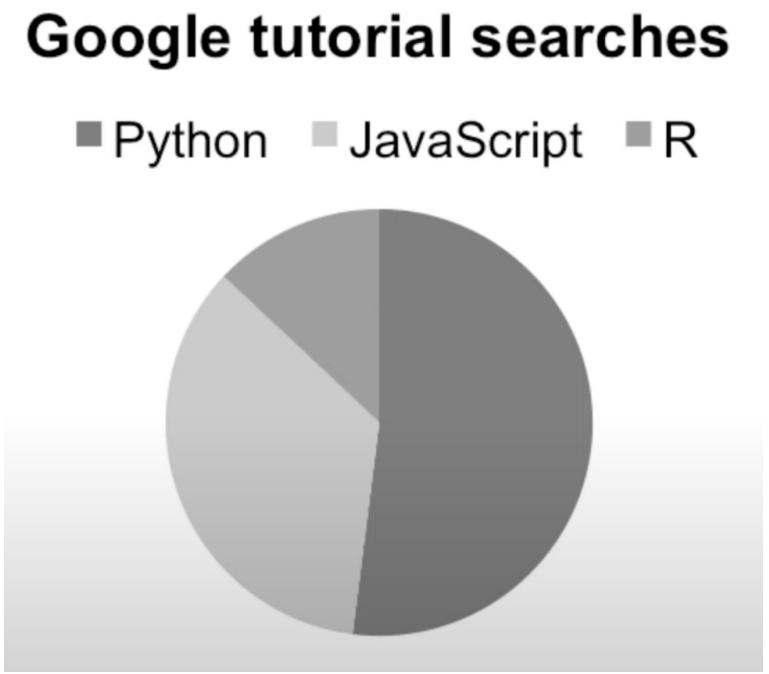
- **For a sequence of time over time.**
- **Good for comparing trends.**



Source from website.

Pie Chart

- **Show proportions.**
- **Best for few categories.**
- **Parts must sum to a meaningful whole.**



Source from website.

Pie Chart

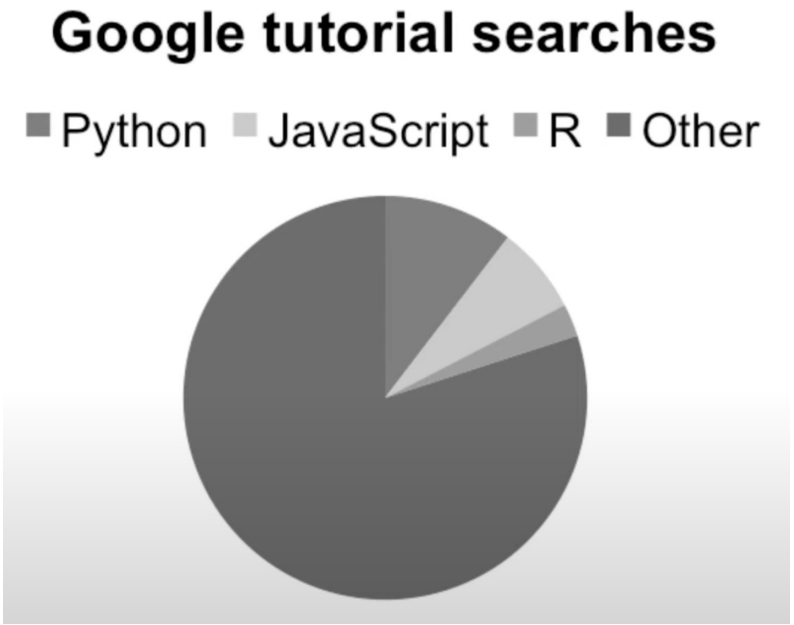
- **Show proportions.**
- **Best for few categories.**
- **Parts must sum to a meaningful whole.**



Source from website.

Pie Chart

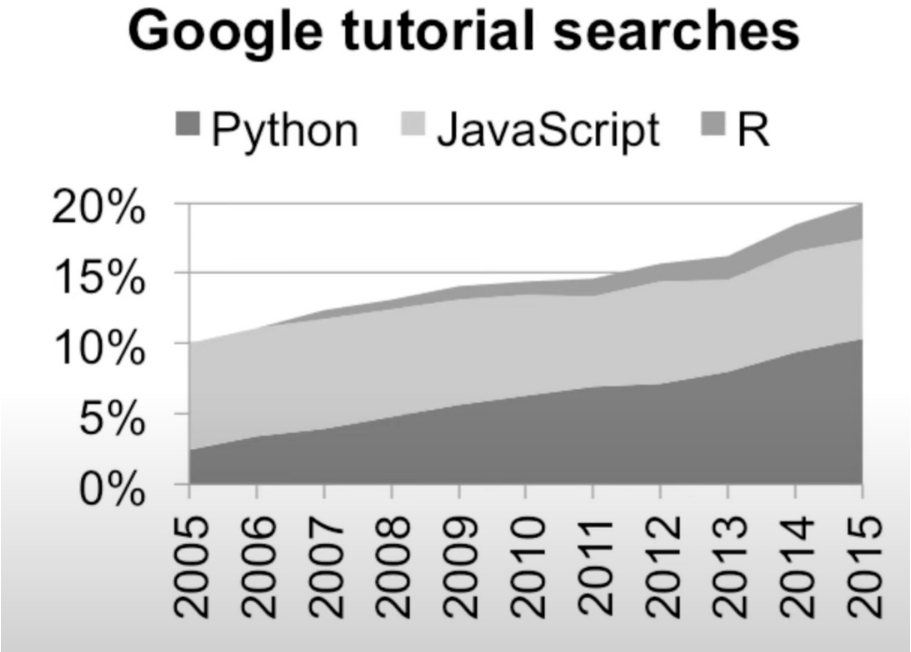
- **Show proportions.**
- **Best for few categories.**
- **Parts must sum to a meaningful whole.**
- **Show or highlight simple proportion for one category.**



Source from website.

Stacked Area Chart

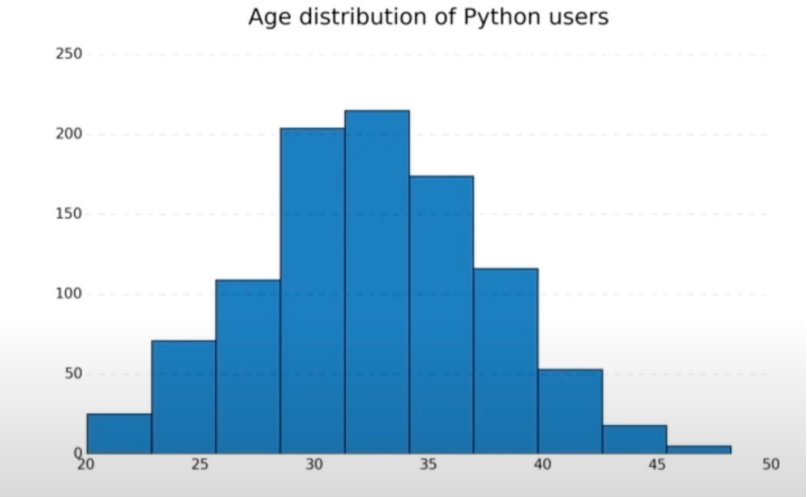
- **Best for visualizing cumulative proportions.**
- **Poor at showing specific values.**



Source from website.

Histograms

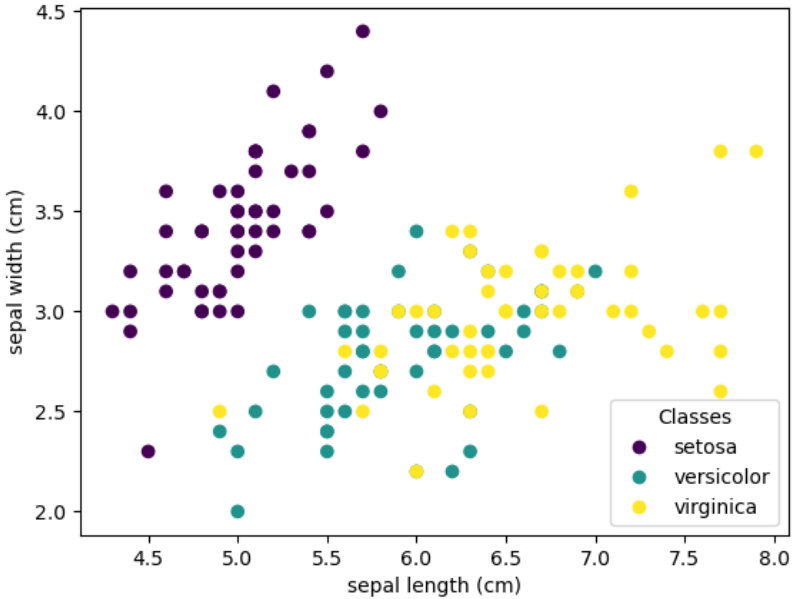
- **Show distributions and spread of data.**



Source from website.

Scatter Plots

- **Powerful chart.**
- **Show correlation.**
- **Can scale dot size to show the third variable.**



Source from website.

Basic probability and the Law of Large Numbers

Start with the coin toss

Coin toss: How to determine whether a coin is fair or not via coin toss?

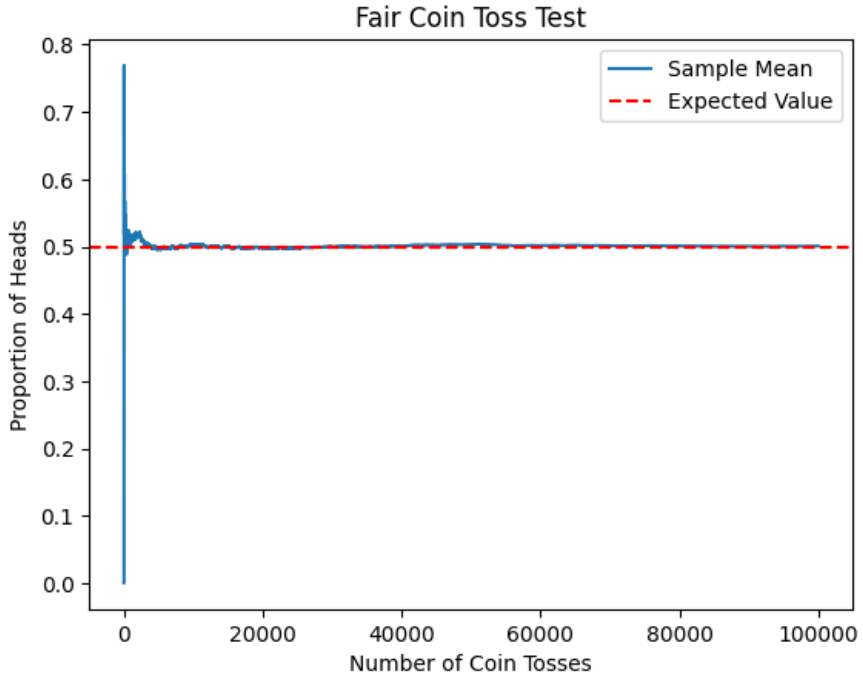
Random Experiments



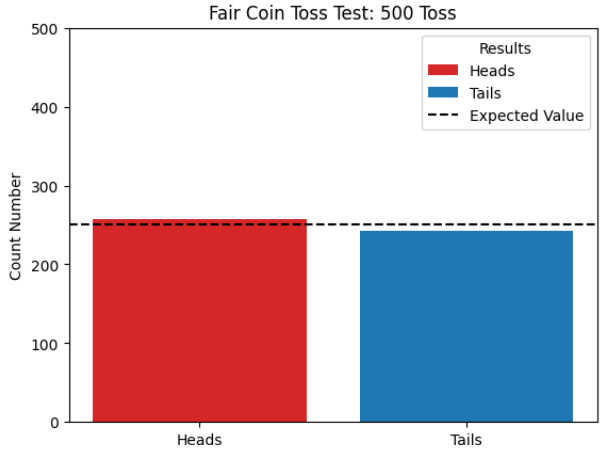
Flip the coin multiple times and count the number of heads and tails.

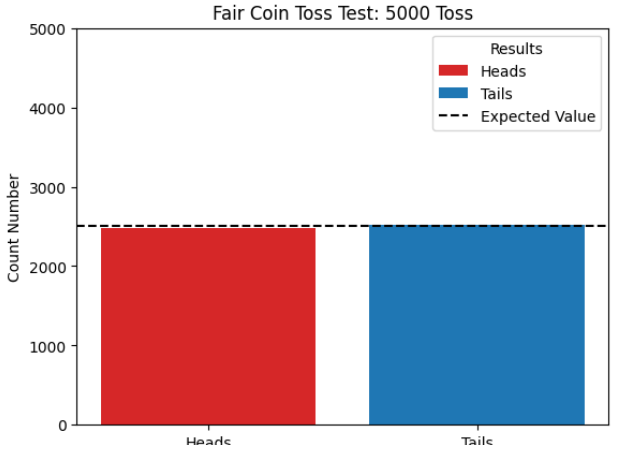
Experiment results visualization

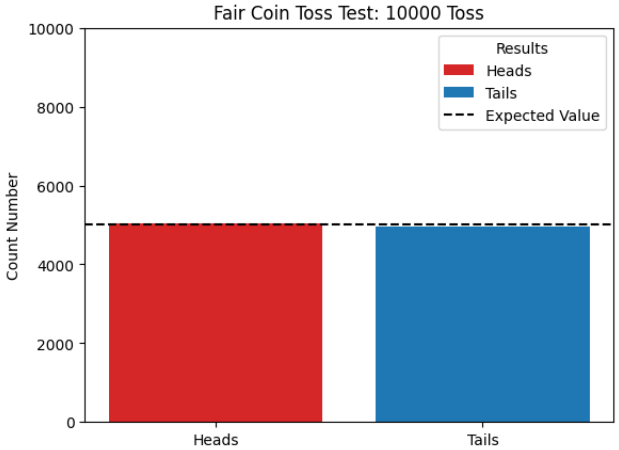
Visualization of Proportion of Heads for up to 100,000 experiments

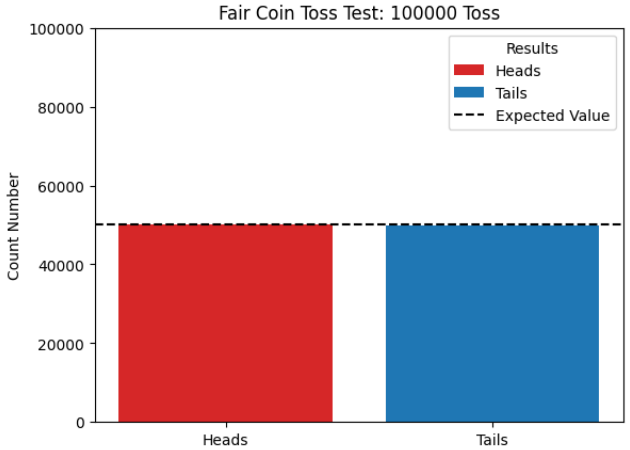


Experiment results visualization









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?

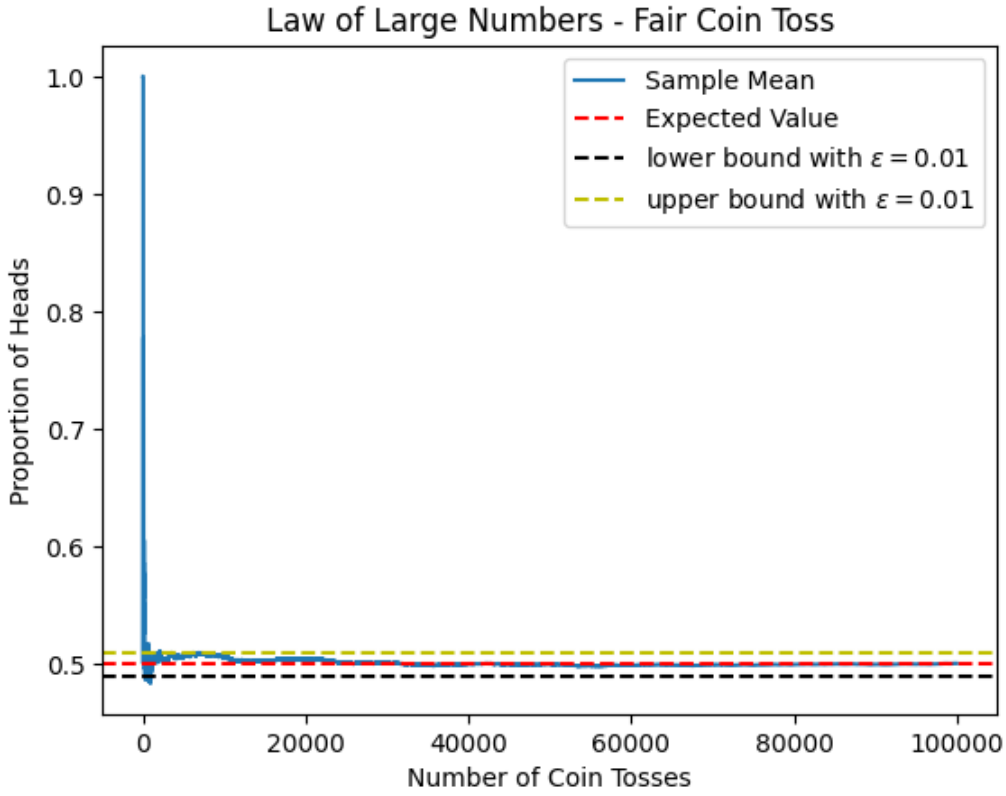
Weak Law of Large Numbers: The sample mean converges in probability to the expected value of a random variable:

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

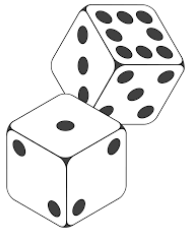
In other words, for any $\varepsilon > 0$,

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

A visualization

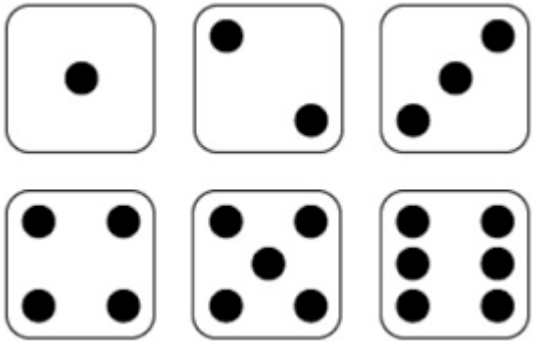


Another example



Dice Rolling: Determine the expected value of the results.

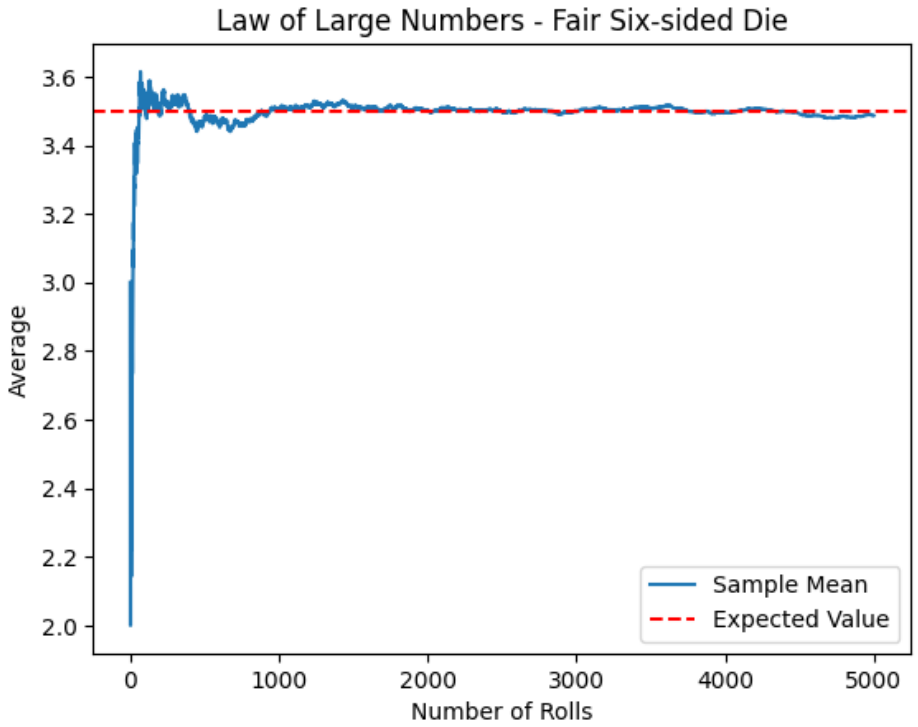
Possible outcomes:



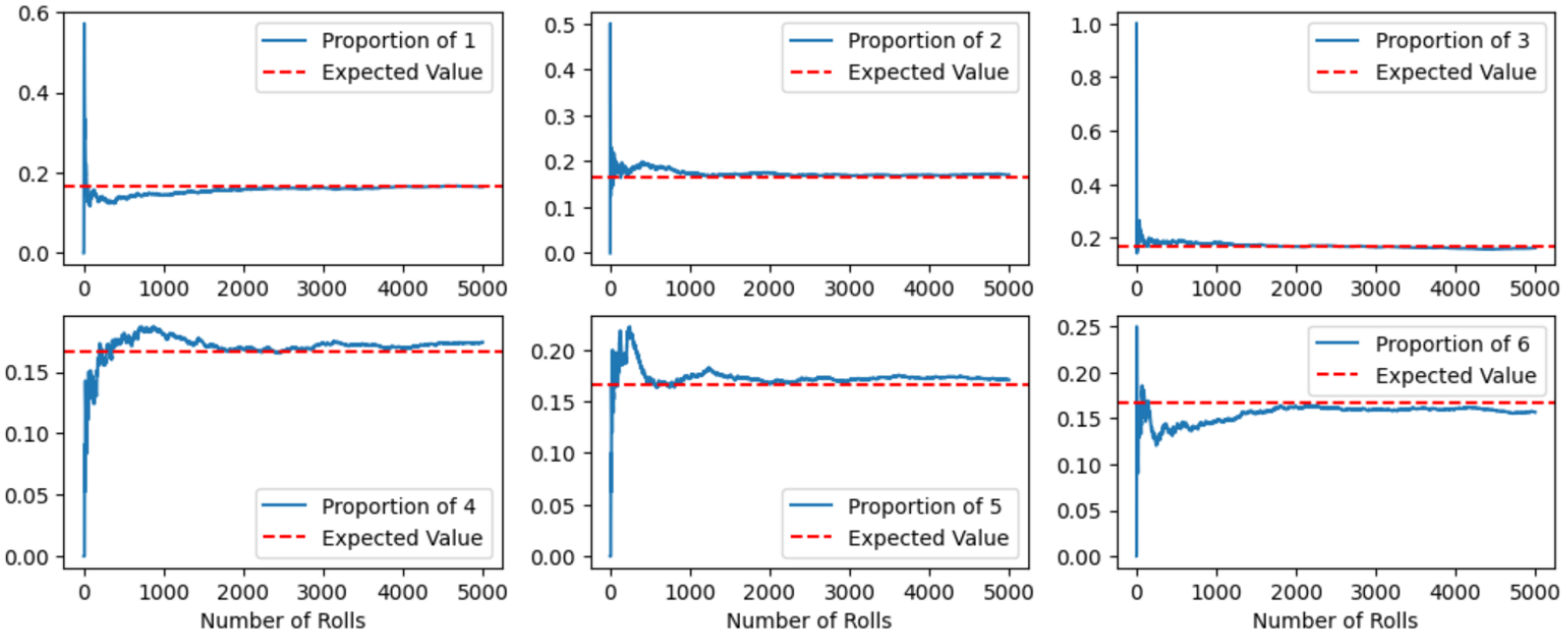
Expected value: $\frac{1}{6} * (1 + 2 + 3 + 4 + 5 + 6) = 3.5.$

Variance: $\mathbf{E}\left[(\mathbf{X} - \boldsymbol{\mu})^2\right] = \frac{\mathbf{35}}{\mathbf{12}}.$

Visualization of sample mean



Fairness of the dice



Until now

(a) We know that the sample mean $\overline{X}_n$ converges to the expected value of a random variable in probability.

(b) Have a visualization understanding of the weak Law of Large Numbers on two examples.

Summary

1. Introduction to some basic visualization ideas.

2. Introduction to the Law of Large Numbers

3. Understand the Law of Large Numbers with visualization

More on Law of Large Numbers: https://en.wikipedia.org/wiki/Law_of_large_numbers


Thank You!