Lecture 1
Data Analysis Algorithm I: Statistics

Yue Ma, July 2024

Contents

Lecture 0

Lecture 1

The Very Basic Example

The easiest case: Flip a coin

Case	Probability
Head	0.5
Tail	0.5

Question: What if you repeat for multiple times?

More Options

Roll a dice

Case	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

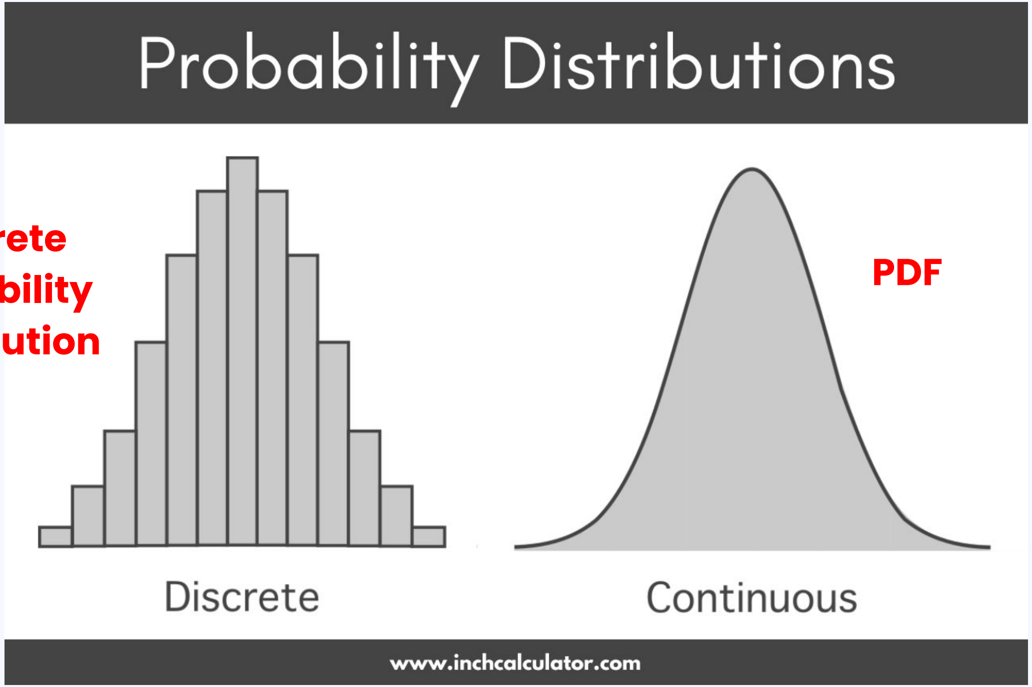
Non-uniform Distribution

Roll a dice (not uniform)

Case	Probability
1	1/12
2	1/12
3	1/3
4	1/6
5	1/6
6	1/6

PDF (Probability Density Function)

Discrete Probability Distribution

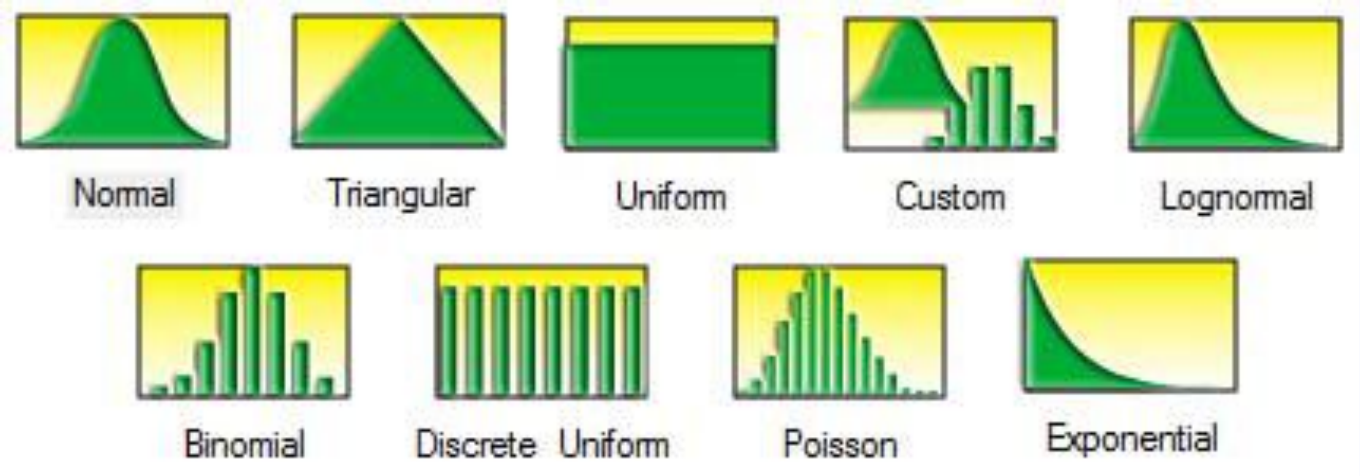


PDF

From Discrete to **Continuous**

(or the opposite)

Examples of PDF



Probability Distributions have various forms

Example: Gaussian Distribution

Normal Distribution Formula

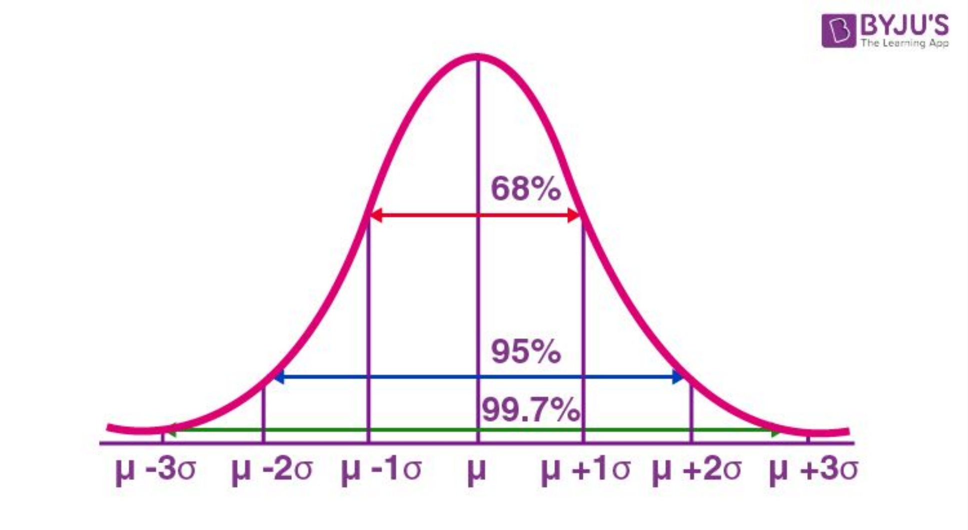
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ mean of x

$\sigma =$ standard deviation of x

$\pi \approx$ 3.14159 ...

$e \approx$ 2.71828 ...



Hypothesis Test: p-value

Easiest case for null hypothesis test: Find the p-value

When you have only 1 hypothesis

the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct

"Extreme" doesn't have a unique definition. There are lots of choices

Hypothesis Test: p-value

Let's go back to the coin flipping example

Case	**Probability**
Head	0.5
Tail	0.5

Data: Number of Heads after repeating for 100 times

Hypothesis: The coin is uniform (probability is 0.5-0.5)

Hypothesis Test: p-value

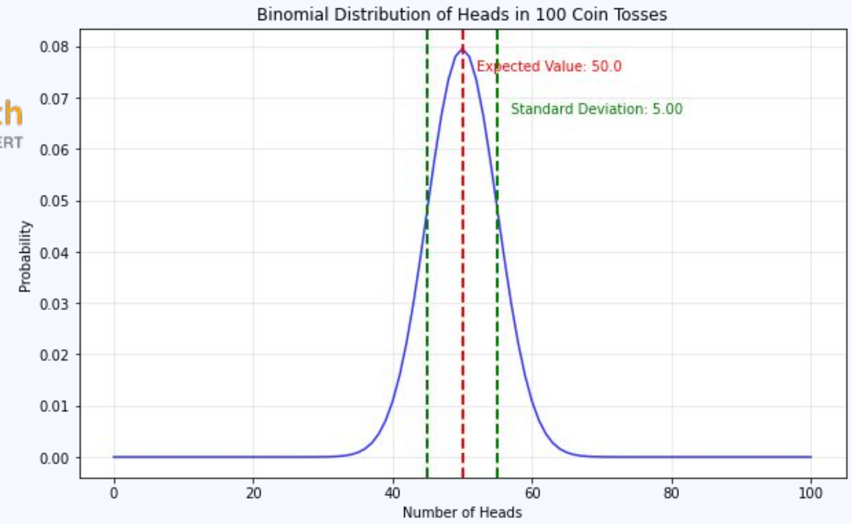
With the binomial test formula:

Binomial Distribution Formula

cuemath
THE MATH EXPERT

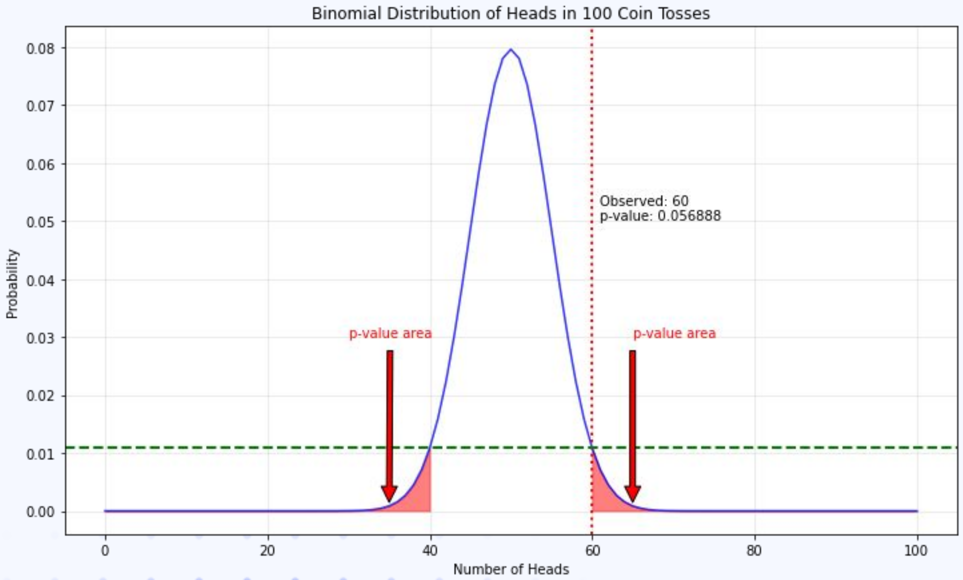
$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

where
n = the number of trials (or the number being sampled)
x = the number of successes desired
p = probability of getting a success in one trial
q = 1 - p = the probability of getting a failure in one trial



Hypothesis Test: p-value

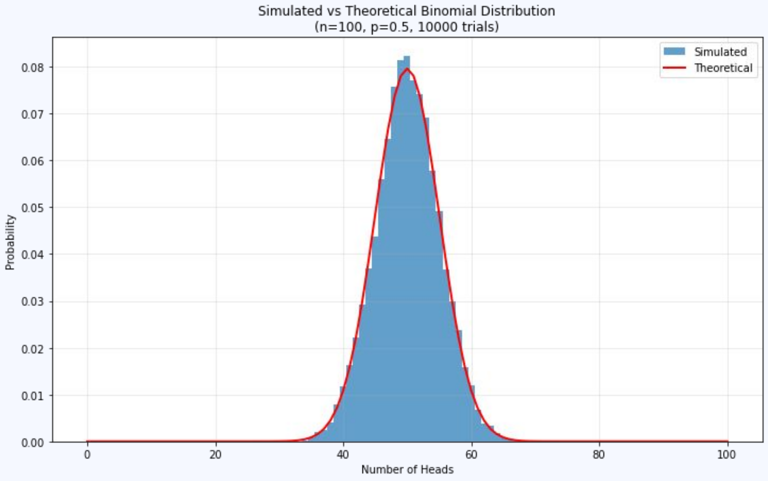
With the binomial test formula:



Hypothesis Test: p-value

Or, we use Monte Carlo to get PDF

- For each experiment: we repeat for 100 times, record the number of heads (generate 100 random numbers)
- We do 10000 experiments
- Draw a histogram of the results
- Then we get a numerical PDF



Homework

Try to play with and read the code, then do the following tasks:

- Try to use a **non-uniform coin** (set the value to < or >0.5), and check the results **(attach some plots).**
- For the dice rolling case, which data can we choose to judge the uniformity? Explain your idea.
- Explain what is histogram in your own words, and try to adjust the **binning of the histogram** for the numerical simulation **(attach some plots)**
- For the numerical method, how many “experiment” do we need? Can you come up with a way to judge if it’s enough?