Statistics in Data Science

overview

1. **Introduction to Statistics in Data Science**
2. **Distributions**
3. **Distribution Estimators: MoM, MLE, KDE**
4. **Point Estimates**
5. **Statistical Hypothesis Testing**
6. **Correlation**
7. **Practical Examples**

2. Distributions

- Define probability distribution.

  Probability measures how likely it is for an event to occur on a scale from 0 (the event never occurs) to 1 (the event always occurs).

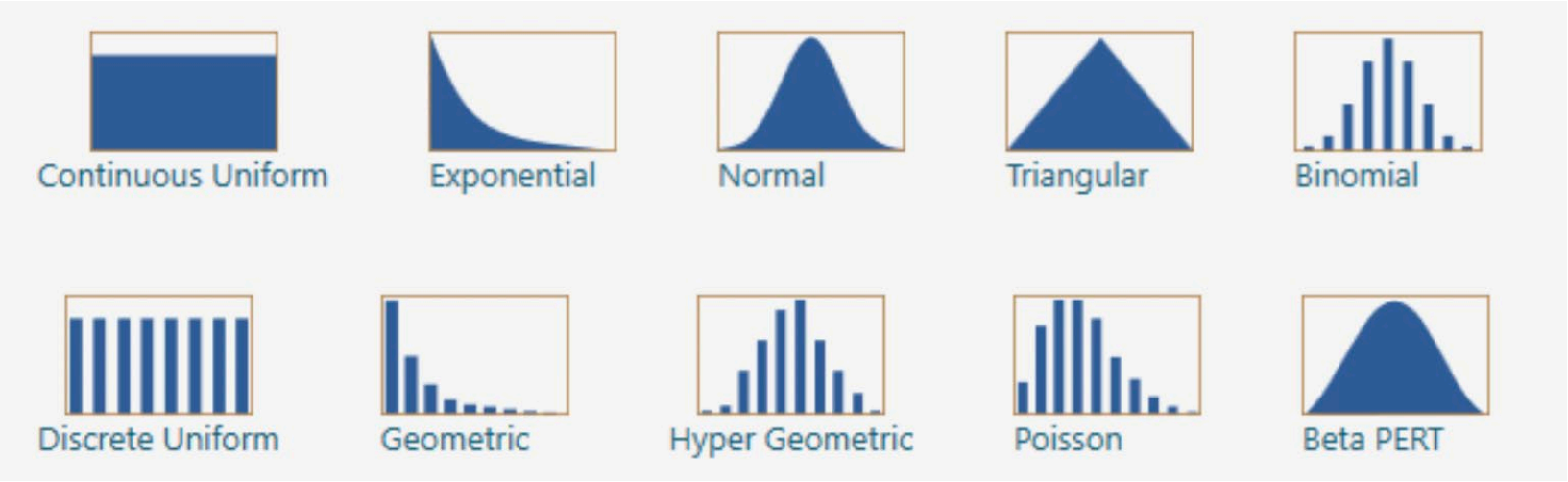
- Discrete and continuous distributions.
- common probability distributions:

  Uniform distribution; Normal distribution; Exponential distribution; Poisson distribution

- Include probability density functions (PDFs) and cumulative distribution functions (CDFs).

- A **discrete random variable** is one whose set of assumed values is countable (arises from counting).
  - values are drawn from a finite set of states.
  - Simply, this means that if I pick any two consecutive outcomes. I can't get an outcome that's in between.
  - In mathematics, we would say that the list of outcomes is countable.

- A **continuous random variable** is one whose set of assumed values is uncountable (arises from measurement.).
  - Values are drawn from a range of real-valued numerical values.



- **Continuous Probability Distributions**
- A continuous probability distribution summarizes the probability for a continuous random variable.
- The probability distribution function, defines the probability distribution for a continuous random variable.
- Continuous probability distribution also has a cumulative distribution function, or CDF, that defines the probability of a value less than or equal to a specific numerical value from the domain.
- Distributions include:

  Normal or Gaussian distribution; Exponential distribution; Pareto distribution
- Examples:

  The probabilities of the heights of humans; The probabilities of income levels

Continous Probabilty Terminology:

- **PDF: Probability Density Function,** returns the probability of a given continuous outcome.

$$\int_a^b f(x)dx = P(a \leq X \leq b)$$

- **CDF: Cumulative Distribution Function,** returns the probability of a value less than or equal to a given outcome.

$$f(x) = P(X \leq x)$$

- **PPF: Percent-Point Function,** returns a discrete value that is less than or equal to the given probability.
  - Inverse of CDF

- **Discrete Probability Distributions**
- A discrete probability distribution summarizes the probabilities for a discrete random variable.
- The probability mass function, or PMF, defines the probability distribution for a discrete random variable.
  - It is a function that assigns a probability for specific discrete values.
- A discrete probability distribution has a *cumulative distribution function, or CDF.*
  - This is a function that assigns a probability that a discrete random variable will have a value of less than or equal to a specific discrete value.
- Distribution include:

  Bernoulli and binomial distributions; Poisson distribution.
- Examples:

  The probabilities of dice rolls form a discrete uniform distribution.

  The probabilities of coin flips.

Discrete Probabilty Terminology:

- **PMF: Probability Mass Function**, returns the probability of a given outcome.

$$f(x) = P(X = x)$$

- **CDF: Cumulative Distribution Function**, returns the probability of a value less than or equal to a given outcome.

$$f(x) = P(X \leq x)$$

- **PPF: Percent-Point Function**, returns a discrete value that is less than or equal to the given probability.
  - Inverse of CDF

Uniform distribution

- Continuous & Discrete
- Each value within a certain range is equally likely to occur, and values outside of the range never occur.
- Example: a die roll has six possible outcomes: 1,2,3,4,5, or 6. There is a 1/6 probability for each number being rolled.

Function:

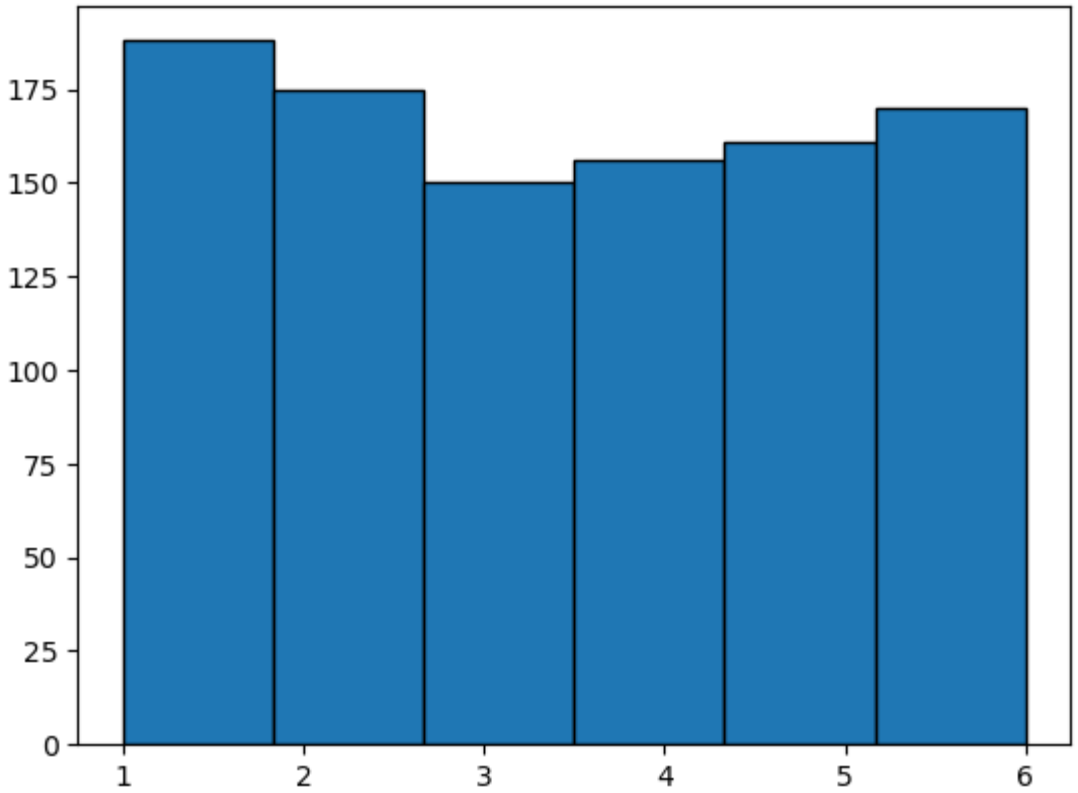
$$f(x;,a,b) = \frac{1}{(b-a)}, \ for\ a \leq x \leq b$$

Parameters

- a is the minimum value
- b is the maximum value

```
import numpy as np
import matplotlib.pyplot as plt

rolls = np.random.randint(1, 7, 1000)
plt.hist(rolls, bins=6, edgecolor='black')
plt.show()
```



Normal distribution

- Continuous
- A normal distribution is defined by its center (mean) and spread (standard deviation.).
- Many common statistical tests assume distributions are normal.

Function

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Parameters

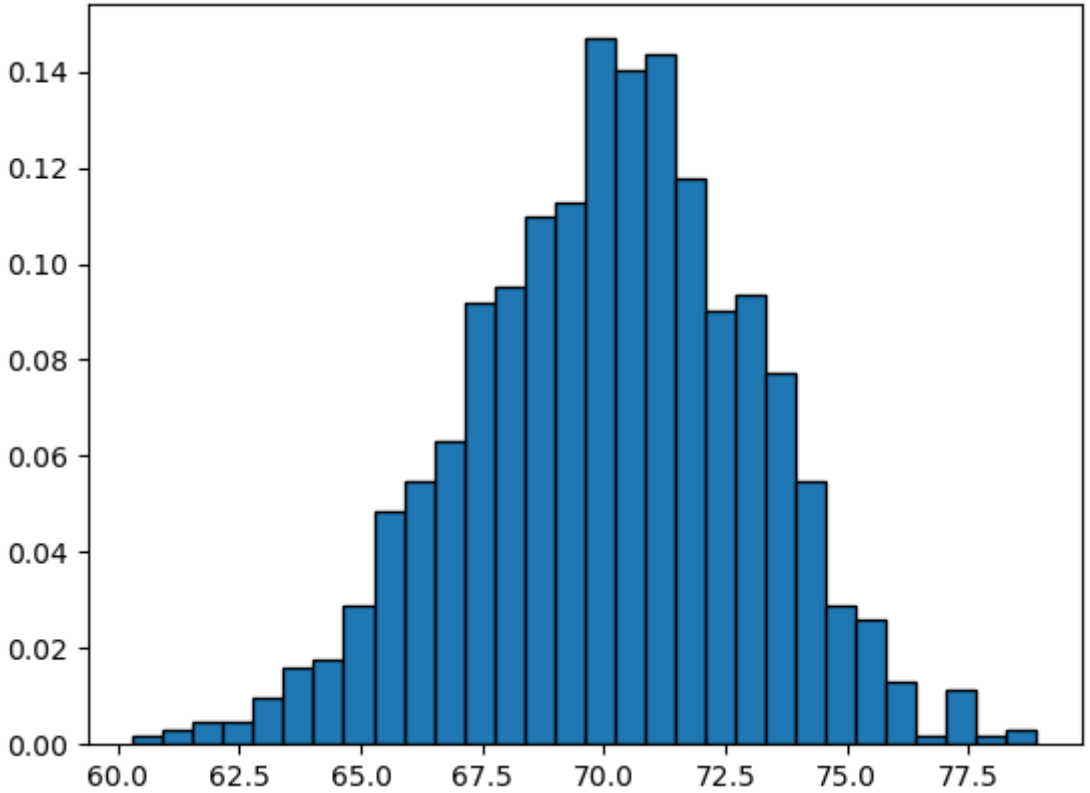
- the mean, μ - the point where the centre of the distribution is, and
- the standard deviation, σ, - how spread out the distribution is.

- **Example**: Heights of adult humans

  The height of adult men is normally distributed with a mean of 70 inches and a standard deviation of 3 inches.

```
from scipy.stats import norm
import numpy as np

data = np.random.normal(70, 3, 1000)
plt.hist(data, bins=30, density=True, edgecolor='black')
plt.show()
```



Binomial distribution

- The Binomial Distribution is a discrete probability distribution that models the number of successes in a fixed number of independent and identically distributed Bernoulli trials. Each Bernoulli trial has only two possible outcomes: success (usually denoted as "1") and failure (usually denoted as "0"). The Binomial Distribution is characterized by two parameters:

1. n (Number of Trials): It represents the total number of trials or experiments, each with a binary outcome (success or failure).

2. p (Probability of Success): It represents the probability of success in a single trial. It is the same for each trial and remains constant throughout the experiments.

Function

$$f(x; p, n) = \binom{n}{x} (p)^x (1-p)^{(n-x)} \qquad \text{for } x = 0, 1, 2, \cdots, n$$

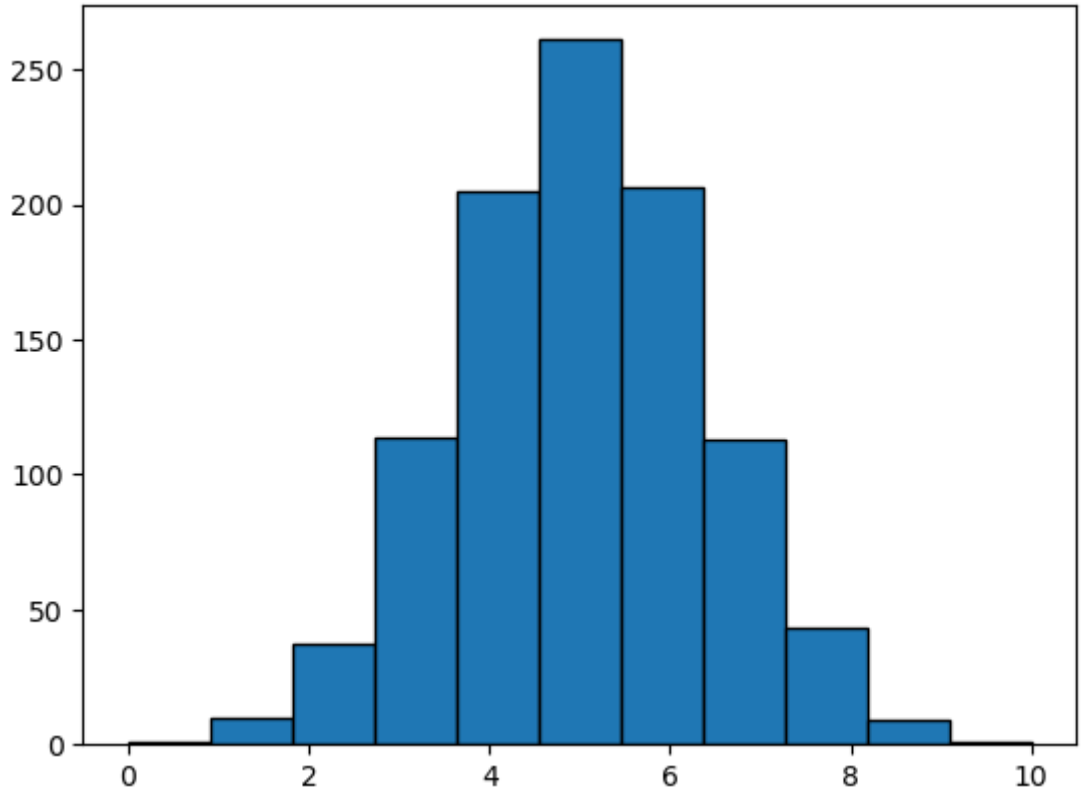
Parameters

- p - probability of success of a single trail
- n - nth trial

- **Example**: Flipping a coin 10 times and counting how many heads occur, or the number of successful free throws in basketball.

```
from scipy.stats import binom

data = binom.rvs(n=10, p=0.5, size=1000)
plt.hist(data, bins=11, edgecolor='black')
plt.show()
```



The Geometric and Exponential Distributions

- The geometric and exponential distributions model the time it takes for an event to occur.
- The ***geometric distribution*** is discrete; Models the number of trials it takes to achieve a success in repeated experiments with a given probability of success.

Geometric distribution - Discrete

Function

$$f(x) = p^x(1-p)^{1-x}$$

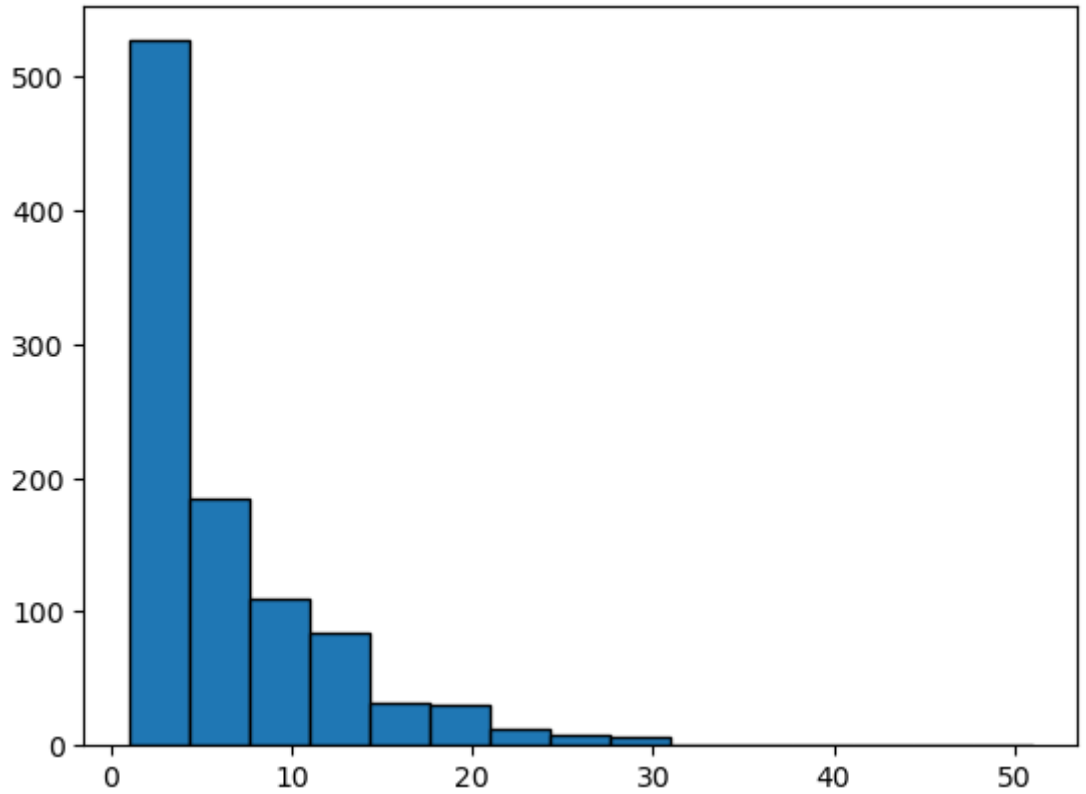
Parameters

- x represents the outcome and takes the value 1 or 0. So we could say that heads = 1 and tails = 0.
- p is a parameter that represents the probability of the outcome being 1.

- **Example**: Tossing a coin until you get heads or rolling a die until you get a six.

```
from scipy.stats import geom

data = geom.rvs(p=1/6, size=1000)
plt.hist(data, bins=15, edgecolor='black')
plt.show()
```



- The ***exponential distribution*** is a continuous analog of the geometric distribution; Models the amount of time you have to wait before an event occurs given a certain occurrence rate.

Exponential - Continuous

Function

$$f(x; \mu, \beta) = \frac{1}{\beta} e^{-(x-\mu)/\beta} \qquad x \geq \mu; \beta > 0$$

Parameters

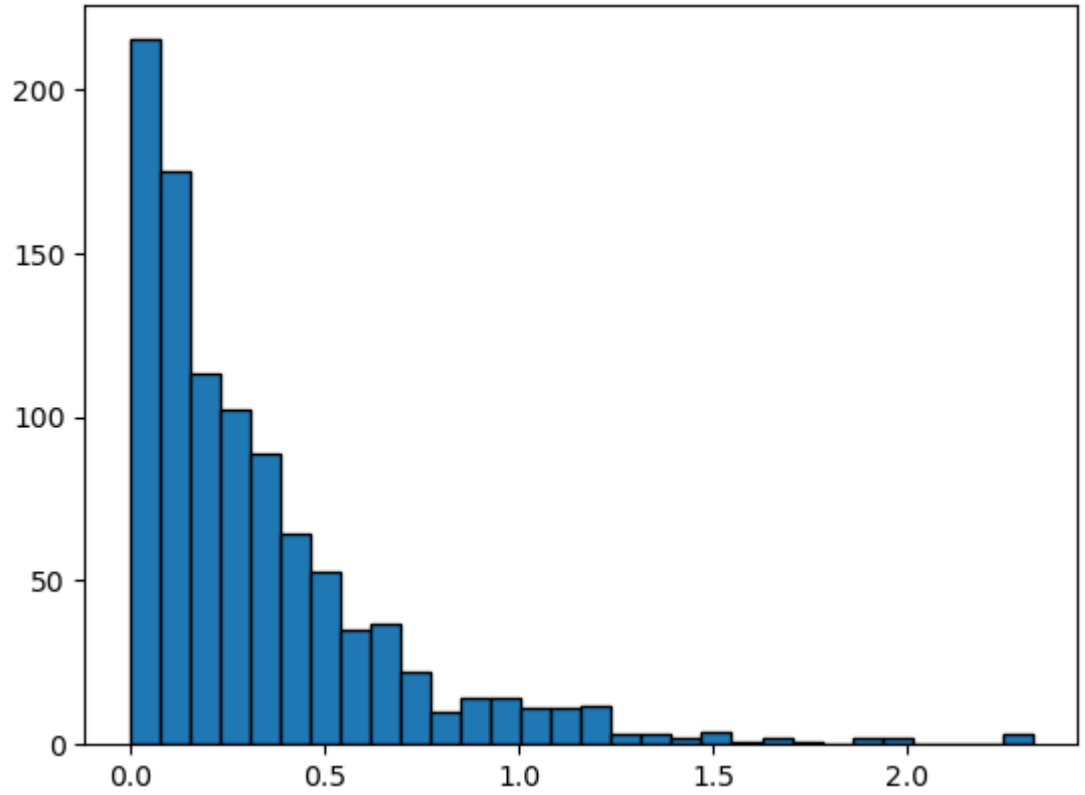
- μ is the location parameter and
- β is the scale parameter (the scale parameter is often referred to as λ which equals 1/β).
- The case where μ = 0 and β = 1 is called the standard exponential distribution.

- **Scenario**: A website gets an average of 3 hits per minute.

What is the probability that the next hit will come within the next 10 seconds?

```
from scipy.stats import expon

data = expon.rvs(scale=1/3, size=1000)
plt.hist(data, bins=30, edgecolor='black')
plt.show()
```



Poisson distribution

- Discrete
- The Poisson distribution models the probability of seeing a certain number of successes within a time interval
- where the time it takes for the next success is modeled by an exponential distribution.
- The Poisson distribution can be used to model traffic, such as the number of arrivals a hospital can expect in a hour's time or the number of emails you'd expect to receive in a week.

Function

$$f(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

Parameters

- $X = \{0, 1, 2, \ldots\}$
- $\lambda > 0$, where λ is both the mean and the variance of X.

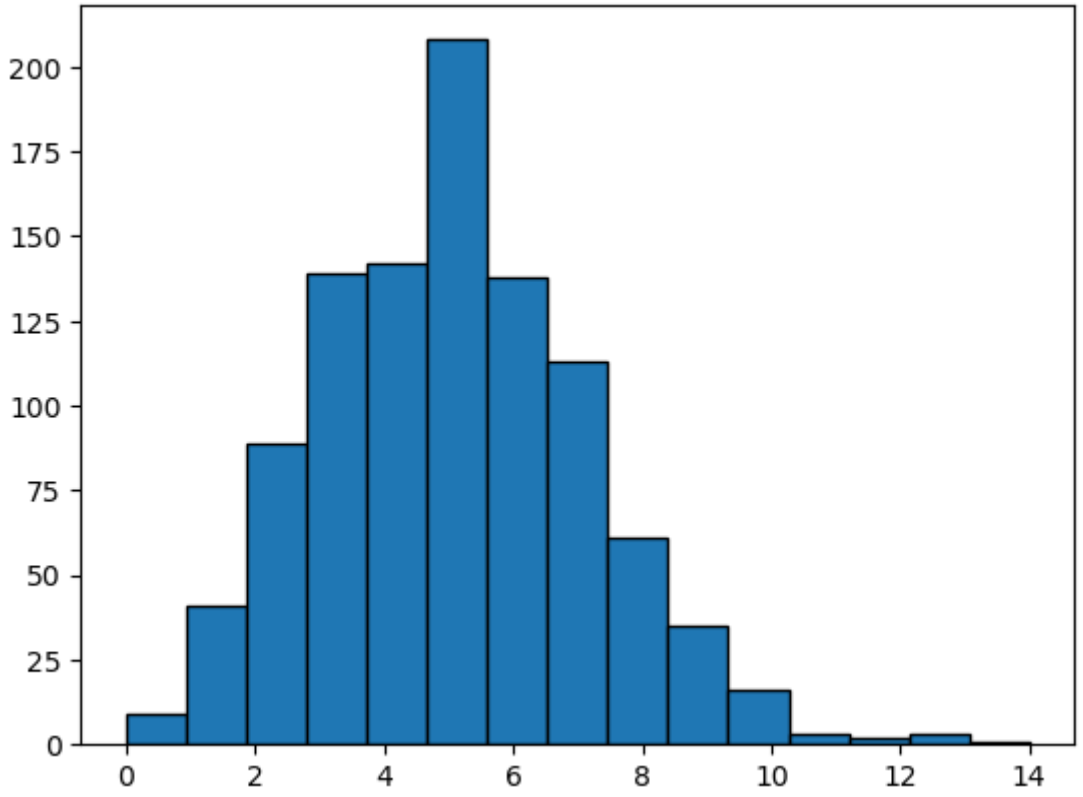
$$E(X) = \mathrm{Var}(X) = \lambda$$

- e = 2.71828

- **Example**: A call center receives an average of 5 calls per minute. What is the probability of receiving exactly 8 calls in a minute?

```
from scipy.stats import poisson
import matplotlib.pyplot as plt

data = poisson.rvs(mu=5, size=1000)
plt.hist(data, bins=15, edgecolor='black')
plt.show()
```



Quick check:

- **Scenario**: Counting the number of cars passing a checkpoint in an hour.

  Poisson distribution, **discrete**.

- **Scenario**: Measuring the time between phone calls at a call center.

  Exponential distribution, **continuous**.

- **Scenario**: Rolling a die multiple times until you roll a 6.

  Geometric distribution, **discrete**.

- **Scenario**: Measuring the weights of apples in a grocery store.

  Normal distribution, **continuous**.

Summary

Discrete Distributions

Discrete distributions involve countable outcomes (integers or specific events). The probability is assigned to each individual outcome.

- **Key Features**:
  - Probability is assigned to specific, distinct values.
  - Probabilities for all outcomes sum to 1.
- **Examples**:
  1. **Binomial Distribution**: Counts the number of successes in a fixed number of independent trials.
     - **Scenario**: Flipping a coin 10 times and counting the number of heads.
     - **Type**: Discrete (you count the exact number of successes).
  2. **Poisson Distribution**: Describes the number of events occurring in a fixed interval of time or space.
     - **Scenario**: The number of emails you receive per hour.
     - **Type**: Discrete (you count a specific number of events).
  3. **Geometric Distribution**: Counts the number of trials until the first success.
     - **Scenario**: Rolling a die until you get a 6.
     - **Type**: Discrete (you count the number of attempts).

Summary

Continuous Distributions

Continuous distributions involve outcomes that can take any value within a range. Probabilities are assigned to ranges of values rather than specific points.

- **Key Features**:
  - Probability is represented by the area under a curve.
  - Probabilities are assigned to ranges (intervals) of values.
  - The probability of any single, exact value is 0.
- **Examples**:
  1. **Normal Distribution (Gaussian)**: Describes a bell-shaped curve where most values cluster around the mean.
     - **Scenario**: Heights of adult men, with most around the average height.
     - **Type**: Continuous (height can take any value within a range).
  2. **Exponential Distribution**: Describes the time between events in a Poisson process.
     - **Scenario**: Time between buses arriving at a stop.
     - **Type**: Continuous (time can be any positive value).
  3. **Uniform Distribution**: Every outcome within a given range has an equal probability.
     - **Scenario**: Choosing a random number between 0 and 1.
     - **Type**: Continuous (any value between 0 and 1 is equally likely).

Check all the figures and examples of there distributions in notebook.