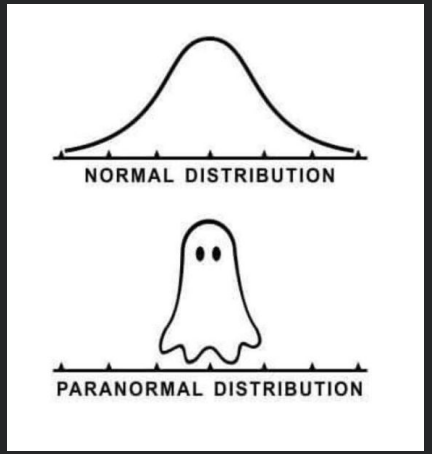
Statistical Distribution I

Data Science Immersive



FLATIRON SCHOOL

Agenda

- Introduce statistical measurements and terminology
- Introduce concept and application of statistical distributions
- Understand the use case of different types of distributions

After today, you will be able to...

- explain central tendency and measurement of dispersion
- Identify different properties of statistical distribution and their application to real world problem
- Understand Stem and Leaf plots, and other visual tools for representing statistical properties
- Understand that different types of statistical distributions

Terminology

- Random Variable

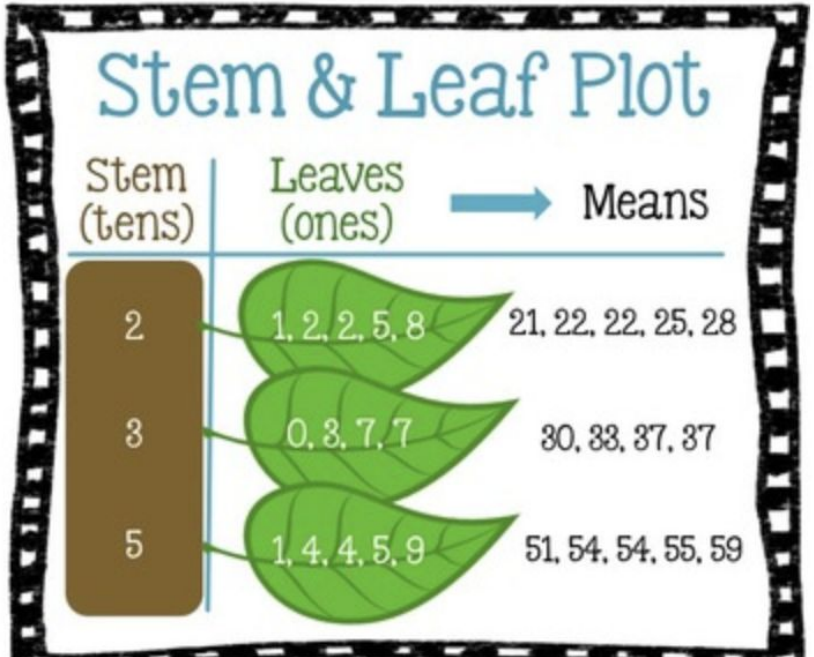
  Is a variable whose value results from the process of a random experiment
- Discrete variable

  A discrete variable can only take on count values
- Continuous variable

  Can take on any possible values

Stem-And-Leaf Plot

- Stem and Leaf plot is a good way for us to get information on the distribution of our data, because it gives us individual information on the data instead of a general distribution.



Stem-And-Leaf Plot

- Stem:
  - The stem is the first digit or digits of the data
- Leaves:
  - The leaves are the last digits of the the data

Measure of Central Tendency

- Mean: arithmetic mean of numbers
- Median: number on the 50th percentile
- Mode: most commonly occurring number
- Any other robust measurement for central tendency?

  SIQR--semi interquartile range

Measure of Dispersion

- Absolute Deviation

  The simplest form of dispersion, calculated by taking the difference between a number and the average

  Eg. in a list [2,10,20,30], the absolute deviation of 2 is |[(2+10+20+30) / 4] - 2| = 13.5

Measure of Dispersion

- Variance

  The variance is calculated by taking the squared difference from the mean and add them all up

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S^2 = \frac{\sum (X - \overline{X})^2}{N-1}$$

  E.g σ^2 of [2,10,20,30] is 147.67

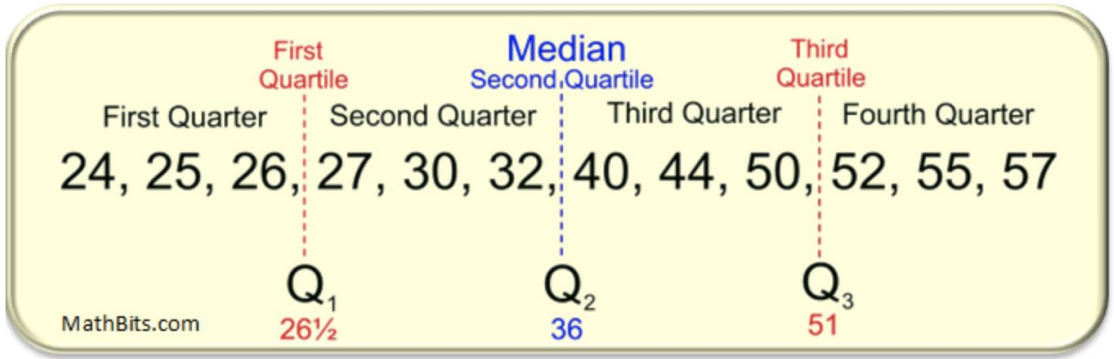
Measure of Dispersion

- Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Quartiles

- The **quartiles** of a data set divides the data into four equal parts, with one-fourth of the data values in each part. The second quartile position is the median of the data set, which divides the data set in half as shown for a simple dataset below:



Quartiles

- IQR--Interquartile Range
  - The interquartile range (IQR) is a measure of where the “middle fifty” is in a data set. It is the difference between the upper quartile and the lower quartile → Q3 - Q1
- SIQR--Semi Interquartile Range
  - SIQR is one-half of IQR → (Q3 - Q1) / 2
- Why are these measurements of central tendency **better** than mean, mode or median?

Probability Mass Function

- Probability Mass Function is a function that maps the frequency of a **discrete** set of data to distribution

$$f_X(x) = \Pr(X = x) = P(\{s \in S : X(s) = x\})$$

- The values of pmf must integrate to 1
- f(x) can only take on values [0,1]

Cumulative Mass Function

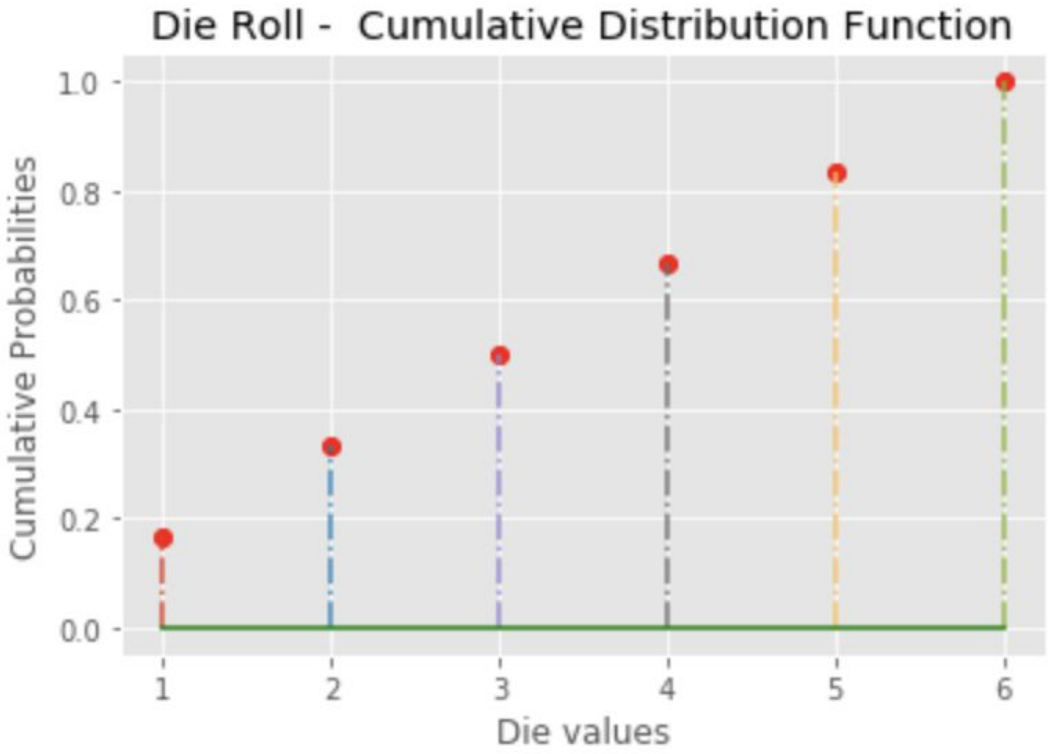
- Cumulative frequency sums the frequencies at and below a particular value

  E.g. 5 students rating the instructor on a scale of 0 to 2

Rating	Frequency	Cumulative Frequency	Cumulative frequency probability
0	1	1	0.2
1	1	3	0.2
2	3	5	0.6

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

Cumulative Density Function

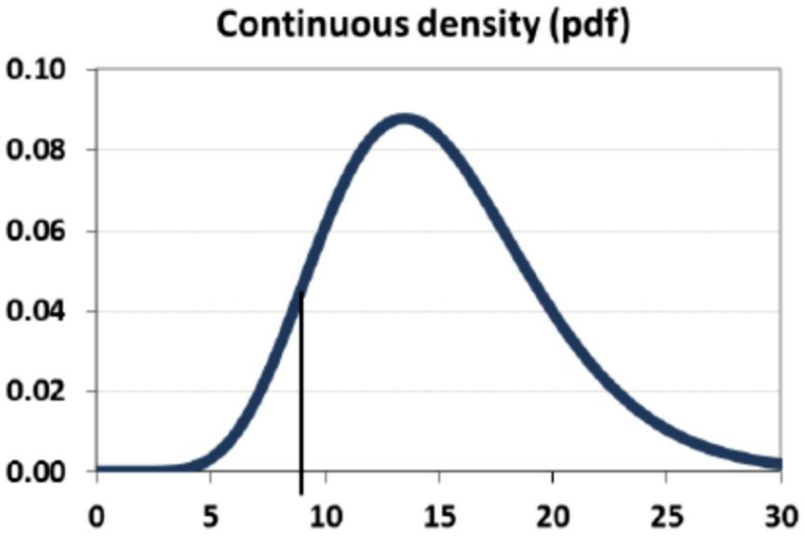


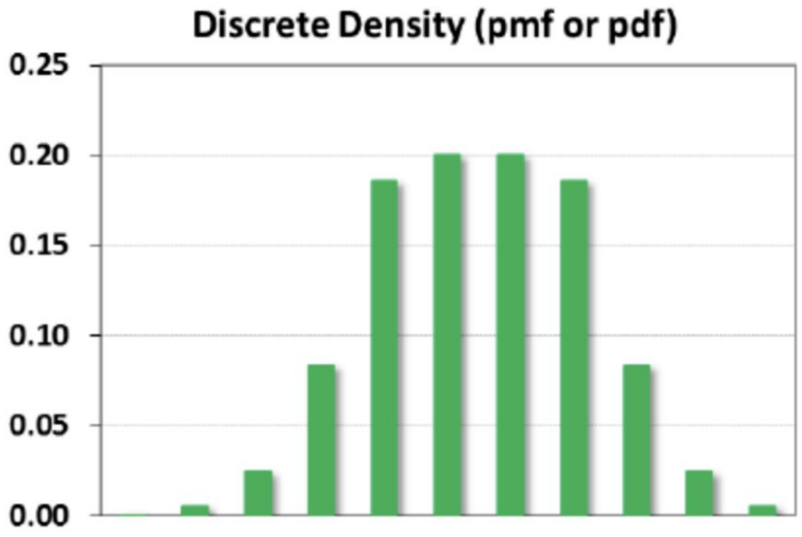
Probability Density Function

- A Probability Density Function (PDF) is analogous to PMFs as we just discussed, but for continuous rather than discrete variables.
- The probability of any given point in a PDF is 0!
-

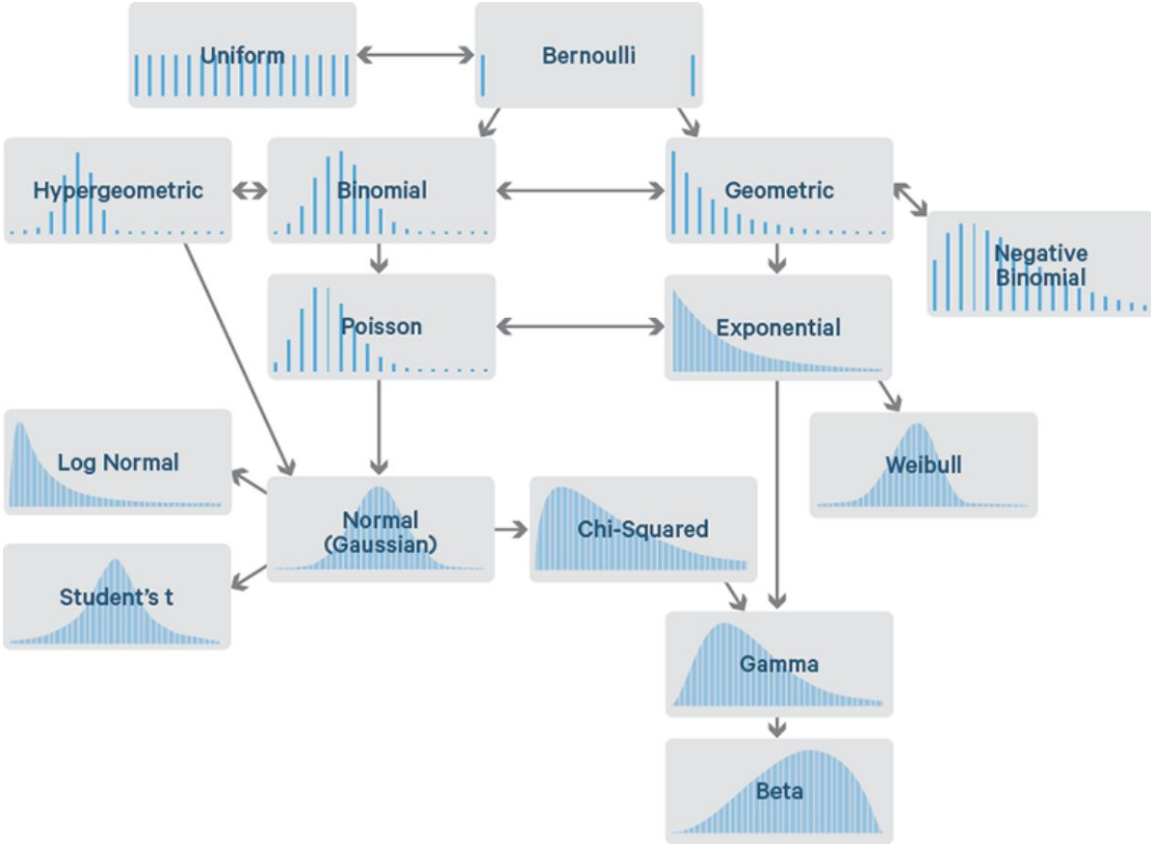
Introduction to Statistical Distribution

- Distribution is a function that represents the occurrence of an outcome in the experiment





Types of Different Distributions



Next Steps

- In module 2, we will discuss different types of statistical distribution and their use cases and applications, which include:
  - The normal/gaussian distribution
  - The standard normal distribution
  - The negative binomial distribution
  - Poisson Distribution