Week Five: Random Variables and Distributions

CS 217

Random Variables

- **Random Variable**: A variable whose possible values are outcomes of a ‘random’ phenomenon
- Throwing a dice or flipping a coin is inherently random but the probable outcomes of each result are not random
- A **probability distribution** is a mathematical distribution that provides the probabilities of occurrence of different outcomes of an experiment

Random Variables

There are two types of random variables:

- Discrete - obtained by counting
  - A discrete variable has a finite amount of possible values, i.e. Heads or Tails for the flip of a coin or 1-6 for the roll of a die
- Continuous - obtained by measuring
  - A continuous variable has an **infinite** amount of possible values, i.e. if I were to measure the height of every student in the class I could round to the nearest inch, or be precise to thirty decimal places
  - You can be 5 feet 8 inches, or 5 feet 8.3 inches, or 5 feet 8.27 inches, or 5.8272 inches, etc...

Random Variables

- **Probability Distribution** - mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment

Random Variables

- Flipping a single coin is an example of a **Bernoulli distribution**, where there is a probability of an event occurring in a single trial
- Flipping multiple coins is an example of a **Binomial distribution**, where there is a probability of a number of ‘successes’ occurring in multiple **independent** experiments

Binomial Distribution

- The binomial distribution has two inputs:
  - n: the number of trials
  - p: the probability of success for a given trial
- If we flip three coins, what are n and p?

Binomial Distribution

- The binomial distribution has two inputs:
  - *n*: the number of trials
  - *p*: the probability of success for a given trial
- If we flip three coins, what are n and p?
  - N = 3
  - P = 0.5

Binomial Distribution

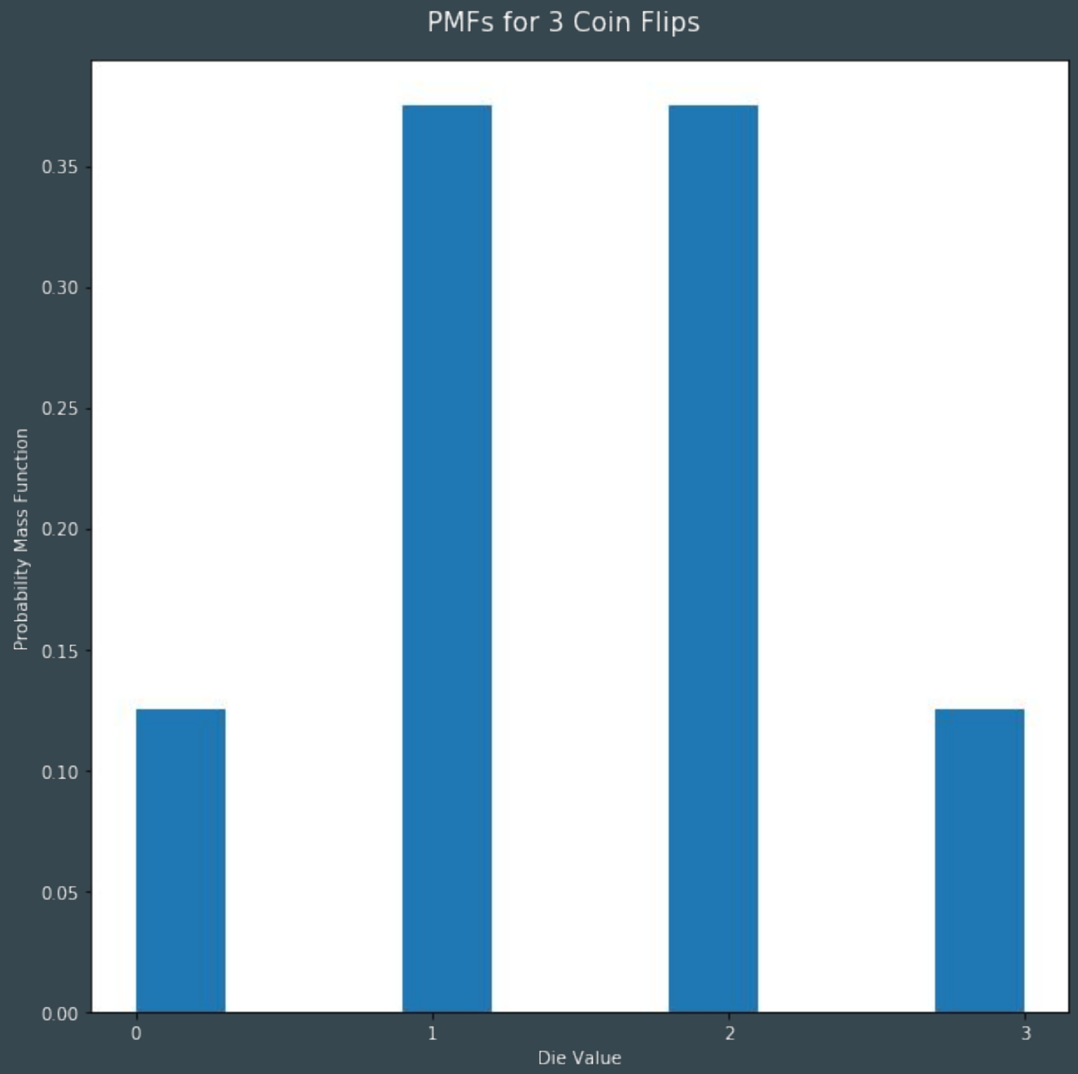
- To the right are the **eight** possible events that occur if we flip a coin three times
- 1 of the events has three heads
- 3 of the events have two heads
- 3 of the events have one head
- 1 of the events have zero heads

HHH	HHT
HTH	THH
THT	HTT
TTH	TTT

Binomial Distribution

- To the right are the **eight** possible events that occur if we flip a coin three times
- ⅛ of the time you will get three heads
- ⅜ of the time you will get two heads
- ⅜ of the time you will get one head
- ⅛ of the time you will get zero heads
- The probability that a **discrete random variable** is equal to a given value is called a **probability mass function**

HHH	HHT
HTH	THH
THT	HTT
TTH	TTT



PMFs for 3 Coin Flips
Probability Mass Function
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
1
2
3
Die Value

Probability Mass Function

- The probability mass function can be found mathematically with the equation to the right
- For example, if we want to find the PMF of getting two heads in three trials, we could use the equation to the right
- 3 * 0.5 * 0.5 * 0.5 = ⅜!

$$\binom{n}{k} p^k (1-p)^{n-k}$$

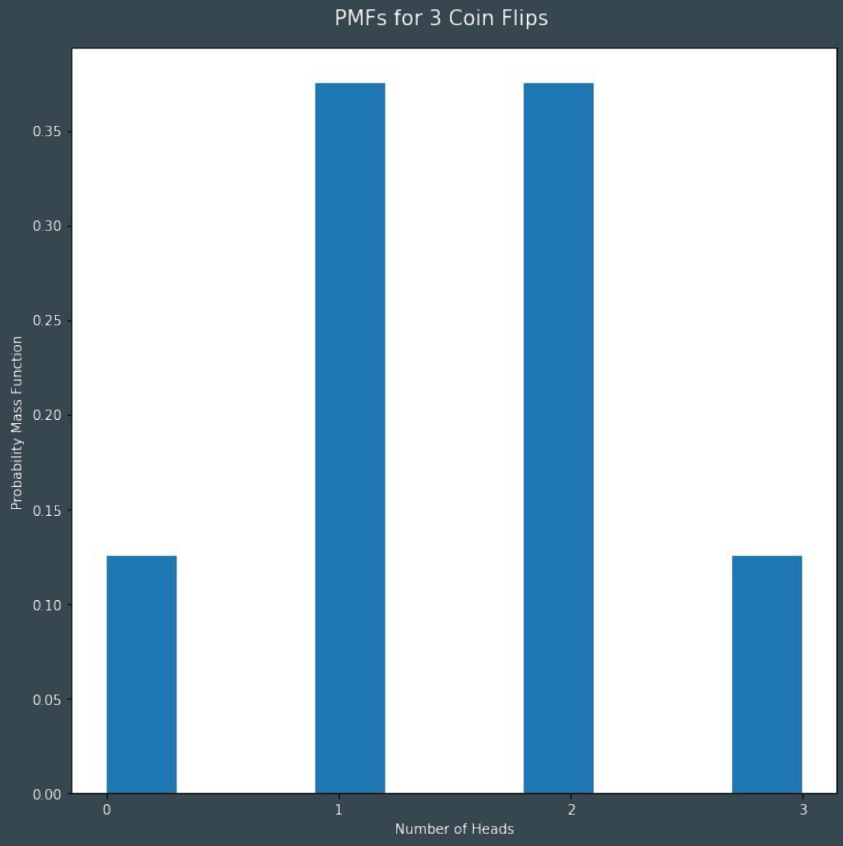
$$\binom{3}{2} 0.5^2 (0.5)^1$$

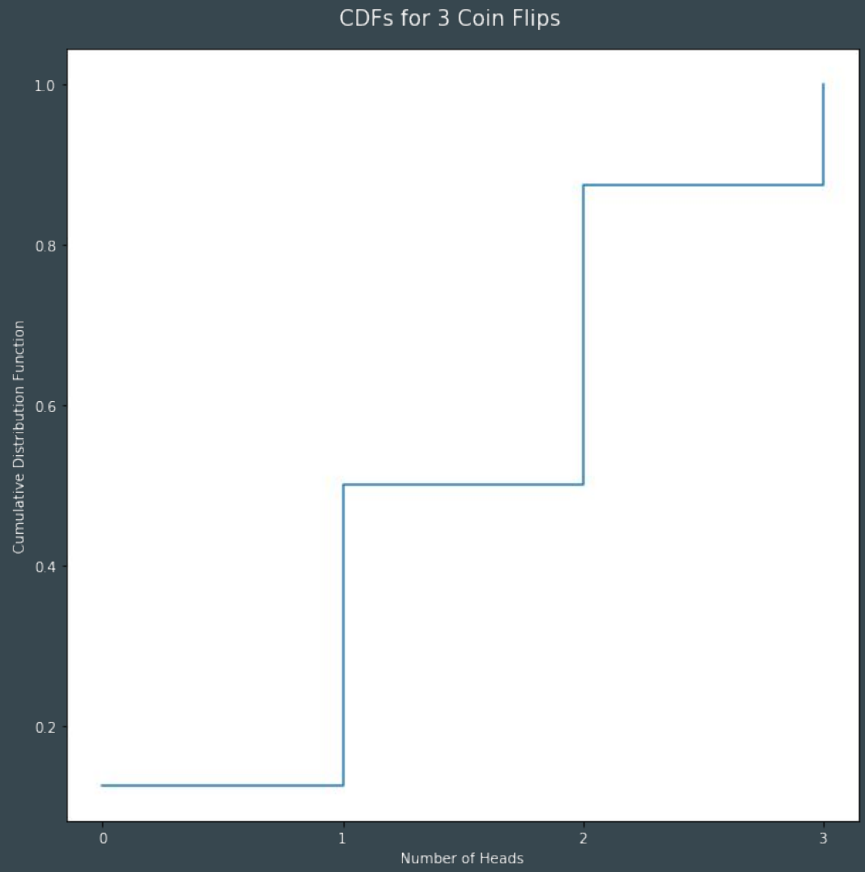
Cumulative Distribution Function

- The **cumulative distribution function** for a discrete variable is the probability that the distribution will have a value **less than or equal to** a certain value
- For a discrete distribution it can be obtained by adding up all of the **probability mass functions** up to and including that number

Cumulative Distribution Function

Number of Heads	PMF	CDF
0	1/8	1/8
1	3/8	4/8
2	3/8	7/8
3	1/8	8/8


PMFs for 3 Coin Flips
Probability Mass Function
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
1
2
3
Number of Heads


CDFs for 3 Coin Flips
Cumulative Distribution Function
0.2
0.4
0.6
0.8
1.0
0
1
2
3
Number of Heads

Binomial Distribution

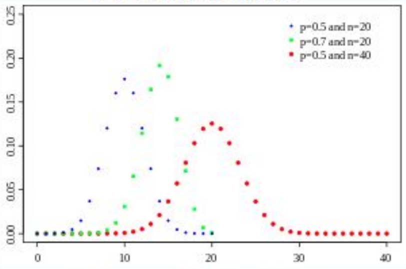
- The **mean** of a binomial distribution is n * p, which in this case is 1.5
    - If we flip three coins, we'll get 1.5 heads on average
- The **variance** of a binomial distribution is n * p * (1 - p), which in this case is 0.75

Binomial Distribution

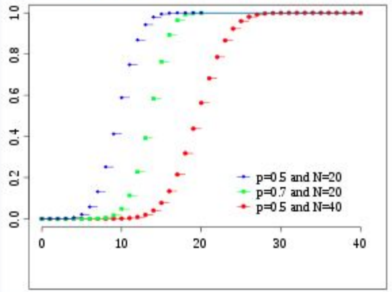
- For a given discrete distribution, these are some of the important metrics:
  - Inputs
  - Mean
  - Variance
  - PMF Formula
  - CDF Formula

Binomial distribution

Probability mass function



Cumulative distribution function



Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \ldots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial
Support	$k \in \{0, 1, \ldots, n\}$ – number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$

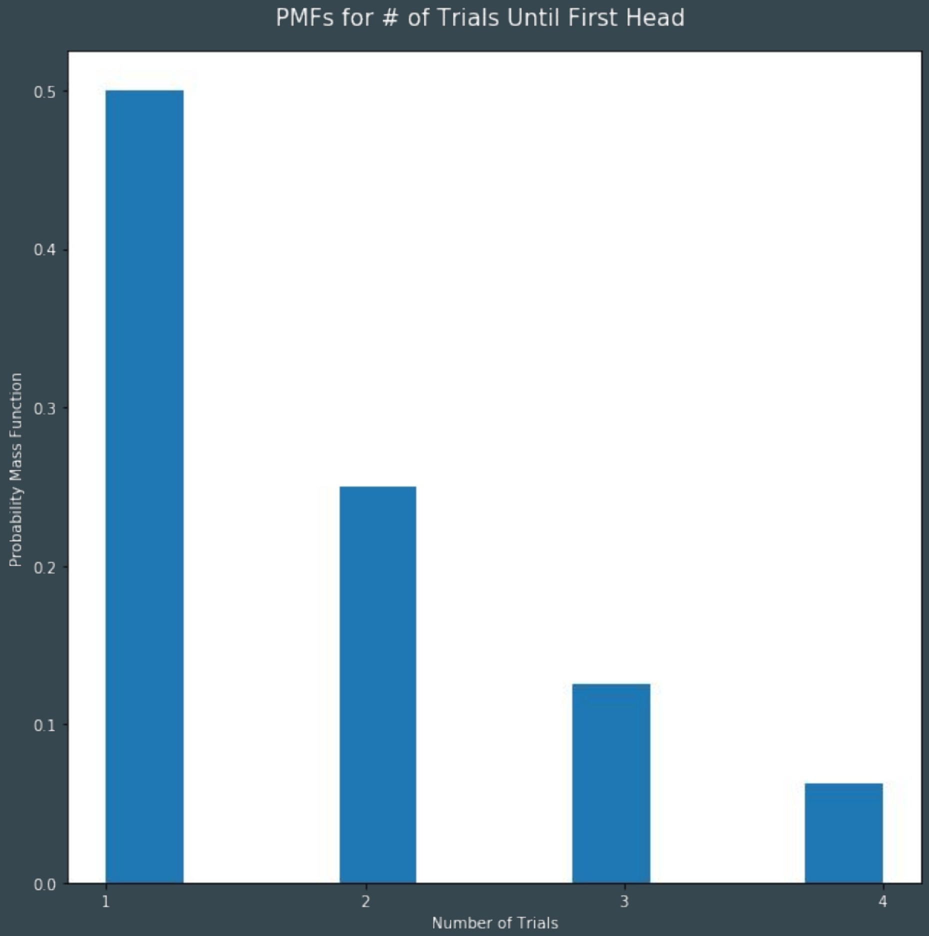
Bernoulli Distribution

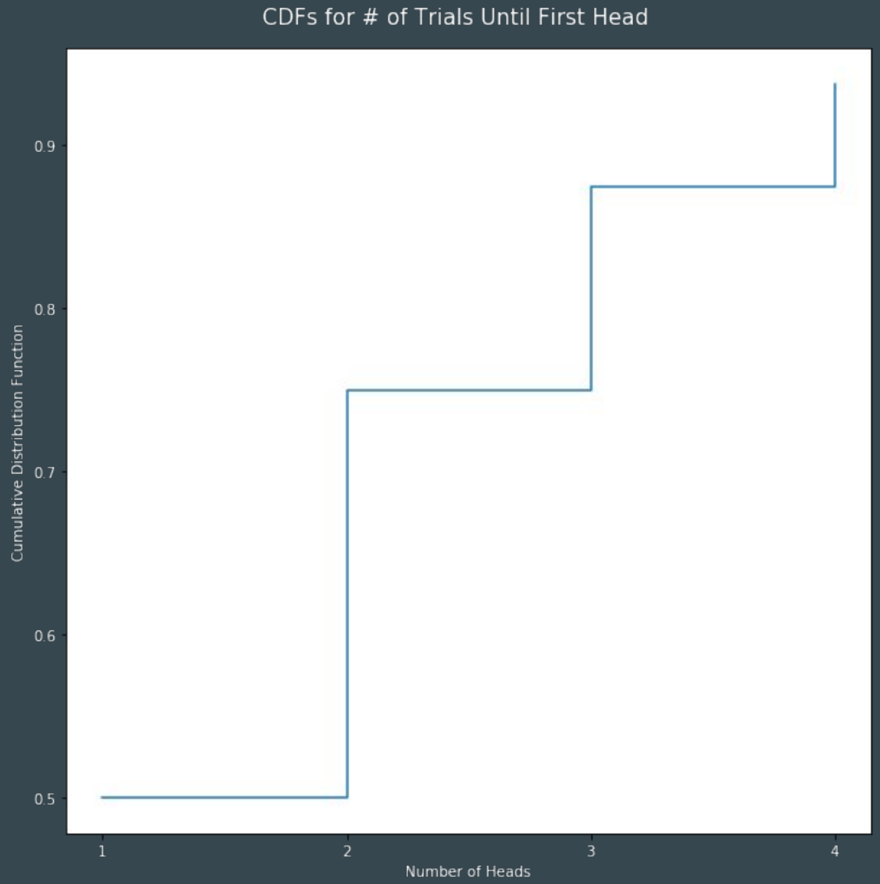
- On that same token, a Bernoulli distribution also has these same metrics
  - Inputs
  - Mean
  - Variance
  - PMF Formula
  - CDF Formula

Bernoulli	
Parameters	$0 \leq p \leq 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
pmf	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0, 1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Mode	$\begin{cases} 0 & \text{if } p < 1/2 \\ 0, 1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Variance	$p(1 - p) = pq$

Geometric Distribution

- What if we wanted to see the distribution of how many times we would need to flip a coin to get a head?
- If we flip a coin once, there are two possibilities: {**H**, T}. There is a 0.5 chance that it will take one flip to get the first head
- If we flip a coin twice, there are four possibilities: {HH, HT, **TH**, TT}. There is a 0.25 chance that it will take two flips to get the first head
- If we flip a coin three times, there are eight possibilities: {HHH, HHT, HTH, THH, HTT, THT, **TTH**, TTT}. There is a ⅛ chance that it will take eight flips to get the first head


PMFs for # of Trials Until First Head
Probability Mass Function
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
Number of Trials


CDFs for # of Trials Until First Head
Cumulative Distribution Function
0.5
0.6
0.7
0.8
0.9
1
2
3
4
Number of Heads

Geometric Distribution

- The geometric distribution only has one input, *p*, compared to the two inputs for the binomial distribution, *p* and *n*
- The mean of the geometric distribution is 1/p, which in our case is 1/0.5, or two. On average, it will take two coin flips to get our first head
- The variance of the geometric distribution is (1 - p) / p^2, which in our case is 0.5/0.25, or also 2
- The PMF for a given # of trials, k, is (1 - p) ^ (k-1) * p
- The CDF for a given # of trials, k, is 1 - (1 - p) ^ k

$$Variance : \frac{1-p}{p^2}$$

$$PMF : 1 - p^{k-1}p$$

$$CDF : 1 - (1-p)^k$$

Geometric Distribution

- How many people do you have to meet, on average, to find someone with the same birthday as you?
- What is the probability of the 100th person you meet being the first to share the same birthday as you?
- What is the probability that one of the first 100 people you meet with share the same birthday as you?

Geometric Distribution

- How many people do you have to meet, on average, to find someone with the same birthday as you?
  - *1/365*
- What is the probability of the 100th person you meet being the first to share the same birthday as you?
  - *(364/365) ^ 99 * (1/365) = 0.002*
- What is the probability that one of the first 100 people you meet with share the same birthday as you?
  - *1 - (364/365) & 100 = 0.24*

Uniform Distribution

- A **uniform distribution** is one where there is an equal opportunity of all outcomes occurring
- It can be either discrete or continuous, however we will think of just a discrete distribution for now
- A single dice roll is an example of this, as there is an equal chance of all outcomes occurring

Poisson Distribution

- A **poisson distribution** measures the probability of a given number of events happening in a fixed interval of time (as opposed to the **binomial distribution** which measures the probability of a given number of events happening in a fixed **number of trials**)
- With the poisson distribution, there is the assumption that the occurence of each event is independent from each other
- An example is the number of babies born in a hospital per hour, since the time one baby is born has nothing to do with when another baby is born
- A more flawed application is the number of trains that arrive at a platform in a given hour
    - ***Why is this flawed?***

Poisson Distribution

- A **poisson distribution** has **one input**: lambda, which is the expected number of occurrences in a given time
- Lambda is both the mean and variance of the poisson distribution
- Say, on average, 2 trains arrive every ten minutes at the 145th Street A stop. What is the probability that 0 trains will arrive?

$$PMF: \frac{\lambda^k e^{-\lambda}}{k!}$$

Poisson Distribution

- A **poisson distribution** has **one input**: lambda, which is the expected number of occurrences in a given time
- Lambda is both the mean and variance of the poisson distribution
- Say, on average, 2 trains arrive every ten minutes at the 145th Street A stop. What is the probability that 0 trains will arrive?
    - *The probability is around 13%*

$$PMF: \frac{\lambda^k e^{-\lambda}}{k!}$$

$$PMF: \frac{2^0 e^{-2}}{0!}$$

Poisson Distribution

- The **poisson distribution** can also be used as an approximation to the binomial distribution when there are a high number of trials ($n > 100$) and a low probability ($p < 0.05$)
- It is considered easier to work with than the binomial distribution because it only requires 1 input as compared to 2, and its CDF function is easier to calculate
- Of course we can easily use either function with Python

$$PMF : \frac{\lambda^k e^{-\lambda}}{k!}$$

$$PMF : \frac{2^0 e^{-2}}{0!}$$

Discrete Distributions

- Of course a discrete event can occur that doesn't follow a common distribution
- In that case we can use the traditional measures for mean, variance, PMF, and CDF

Discrete Distributions

- Say we roll two dice. Below is the sample space of all possible outcomes.

	1	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Discrete Distributions

- We can obtain our metrics via **counting**

Outcome	PMF	CDF
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36

Outcome	PMF	CDF
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	16/36

Discrete Distributions

- The mean is equal to the sum of each outcome multiplied by its respective PMF: (2 * 1/36) + (3 * 2/36) + (4 * 3/36) etc…

Outcome	PMF	CDF
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36

Outcome	PMF	CDF
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	16/36

Discrete Distributions

- The variance is equal to the sum of each outcome minus the mean squared multiplied by its respective PMF: ((2 - 7)^2) * 1/36) + ((3 - 7)^2) * 2/36)

Outcome	PMF	CDF
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36

Outcome	PMF	CDF
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	16/36