# 2 Mathematical Foundation of Big Data

**Syllabus**

At the end of this unit, you should be able to understand and comprehend the following syllabus topics :

- Probability
  - Random Variables and Joint Probability
  - Conditional Probability
  - Tail bounds
  - Pair-wise independence and universal hashing
  - Approximate counting
  - Approximate median
- Concept of Markov chains
  - Markov chains and random walks
- Data Streaming Models and Statistical Methods
  - Flajolet Martin algorithm
  - Distance Sampling and Random Projections
  - Bloom filters
  - Mode
  - Variance
  - Standard deviation
  - Correlation analysis
  - Analysis of Variance

## 2.1 Need of Statistics in Data Science and Big Data Analytics

**Note :** Statistics is a vast domain in itself. This section is to just brush up some elementary concepts useful in machine learning and is not intended to give you a deep perspective on those topics. It is assumed that you either know these topics already or would independently study if you want to know details about them. Your understanding of the topics covered here in is crucial to understanding the concepts laid out in several chapters in this book. Please be thorough with these concepts.

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.

***Definition :*** *The statistical methods use the principles of statistics to carry out data analytics.*

In Big Data Analytics, it is primarily used for

1. Model Planning and Building
2. Model Evaluation and
3. Model Deployment

Let's learn about some of the statistical methods.

### 2.1.1 Sampling Distributions

- I will tell you a joke – a father sends his son to buy a matchbox and asks him to verify that it has good matchsticks. The son buys a matchbox, and as his father wanted, lights up all the matchsticks and verifies that they are good! He returns home and hands over the matchbox with all lit up matchsticks to his father. What happened next, I will leave that to your imagination. But what was the problem with the son's approach of verifying that the matchsticks are good? I hope you got it that he could have just randomly picked a few matchsticks and tested them instead of lighting up the whole matchbox, isn't it?
- Suppose that you are building a product. Do you have infinite resources to experiment on the entire population to get feedback about your product? Can you infinitely reach out to every person on earth and carry on certain investigation or draw conclusions such as whether a vaccine is effective against a virus or not? That's precisely where sampling comes into the picture.

***Definition :*** *Sampling is the act, process, or technique of selecting a representative part of a population for the purpose of determining parameters or characteristics of the whole population.*

So, what is a population and how is it different from a sample?

- Population refers to every member of a group/set of data e.g. all the citizens of India, all the residents of a city, all the students of a school, etc. The process of conducting a survey to collect data for the entire population is called a census.
- Sample refers to a subset of members drawn from the population that time and resources allow one to measure. The process of conducting a survey to collect data for a sample is called a sample survey.

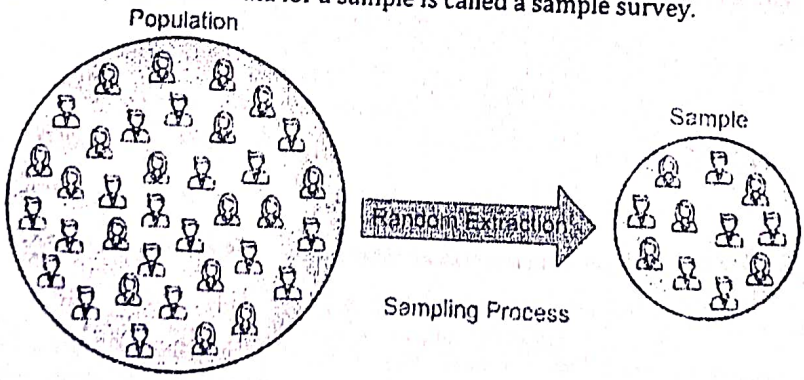


**Fig. 2.1.1**

- The sample needs to be unbiased and representative of the population, and needs to be extracted randomly out of the relevant population.
- A parameter is a number/quantity that describes (a characteristic of) the entire population, whereas a statistic is a number/quantity that describes a sample. The number used to describe characteristic of a population/sample could be a measure of central tendency (mean, median or mode) or a measure of dispersion (variance, standard deviation, range or inter-quartile range).
- Table 2.1.1 shows the notations and formulae of few important and widely used measures in context of population and sample. In the table below, $x$ denotes individual value in the population or sample set.

Let's learn about some of the statistical methods.

## 2.1.1 Sampling Distributions

- I will tell you a joke – a father sends his son to buy a matchbox and asks him to verify that it has good matchsticks. The son buys a matchbox, and as his father wanted, lights up all the matchsticks and verifies that they are good! He returns home and hands over the matchbox with all lit up matchsticks to his father. What happened next, I will leave that to your imagination. But what was the problem with the son's approach of verifying that the matchsticks are good? I hope you got it that he could have just randomly picked a few matchsticks and tested them instead of lighting up the whole matchbox, isn't it?
- Suppose that you are building a product. Do you have infinite resources to experiment on the entire population to get feedback about your product? Can you infinitely reach out to every person on earth and carry on certain investigation or draw conclusions such as whether a vaccine is effective against a virus or not? That's precisely where sampling comes into the picture.

✍ ***Definition : Sampling is the act, process, or technique of selecting a representative part of a population for the purpose of determining parameters or characteristics of the whole population.***

So, what is a population and how is it different from a sample?

- Population refers to every member of a group/set of data e.g. all the citizens of India, all the residents of a city, all the students of a school, etc. The process of conducting a survey to collect data for the entire population is called a census.
- Sample refers to a subset of members drawn from the population that time and resources allow one to measure. The process of conducting a survey to collect data for a sample is called a sample survey.

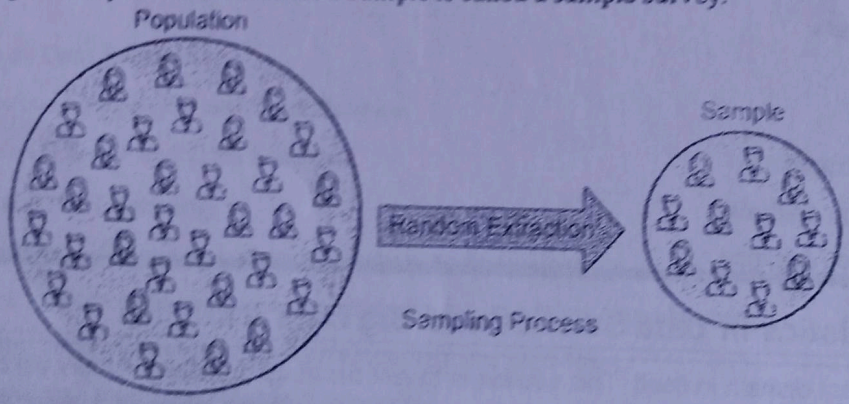


**Fig. 2.1.1**

- The sample needs to be unbiased and representative of the population, and needs to be extracted randomly out of the relevant population.
- A parameter is a number/quantity that describes (a characteristic of) the entire population, whereas a statistic is a number/quantity that describes a sample. The number used to describe characteristic of a population/sample could be a measure of central tendency (mean, median or mode) or a measure of dispersion (variance, standard deviation, range or inter-quartile range).
- Table 2.1.1 shows the notations and formulae of few important and widely used measures in context of population and sample. In the table below, $x$ denotes individual value in the population or sample set.

Table 2.1.1

| | Population (Parameters) | Sample (Statistics) |
|---|---|---|
| Size | N | n |
| Mean | $\mu = \frac{1}{N}\Sigma x$ | $\bar{x} = \frac{1}{n}\Sigma x$ |
| Variance | $\sigma^2 = \frac{1}{N}\Sigma (x - \mu)^2$ | $s^2 = \frac{1}{(n-1)}\Sigma (x - \bar{x})^2$ |
| Standard deviation | $\sigma = \sqrt{\sigma^2}$ | $s = \sqrt{s^2}$ |

## 2.1.2 General Statistics

Let's revise some common statistical concepts.

### 2.1.2(A) Mean

Assume a simple dataset.

$$X = [1\ 2\ 4\ 6\ 12\ 15\ 25\ 45\ 68\ 67\ 65\ 98]$$

X refers to the entire dataset and elements within the dataset are denoted as subscript. For example, $X_3$ refers to the 3rd element in the sample whose value is 4.

Mean of the sample is calculated as

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

$\bar{X}$, pronounced as (X-bar), denotes the mean of the sample. Mean is the average of the sample values in a given dataset.

So, for the given dataset, mean is calculated as

$$\bar{X} = \frac{1 + 2 + 4 + 6 + 12 + 15 + 25 + 45 + 68 + 67 + 65 + 98}{12}$$

$$= \frac{408}{12} = 34$$

### 2.1.2(B) Median

The median is the value separating the higher half from the lower half of a data sample, a population, or a probability distribution. For a data set, it may be thought of as "the middle" value. The basic feature of the median in describing data compared to the mean (often simply described as the "average") is that it is not skewed by a small proportion of extremely large or small values, and therefore provides a better representation of a "typical" value. Median income, for example, may be a better way to suggest what a "typical" income is, because income distribution can be very skewed.

To find the median value in a data set,

1. Sort the dataset
2. Pick the middle element of the data set
   (a) If the count of elements in the data set is odd, then it is just the middle element. For example, if there are 7 elements in the data set, the middle element is 4th one.

(b) If the count of elements in the data set if even, then take the average of the two middle elements. For example, if there are 6 elements in the data set, take the average of 3rd and the 4th element.

Let's see a quick example.

Assume, that you have a dataset such as the following.

$$X = [12, 3, 4, 5, 9, 80, 74]$$

First, sort the elements.

$$X = [3, 4, 5, 9, 12, 74, 80]$$

Since, the data set has 7 elements, just pick the middle element, which is the 4th one. Hence, the median value of the data set X = [3, 4, 5, 9, 12, 74, 80] is 9.

Now, assume that the data set is X = [12, 3, 4, 5, 9, 80]

First, sort the elements.

$$X = [3, 4, 5, 9, 12, 80]$$

Since, the data set has 6 elements, pick the 3rd element and the 4th element and average them out. So, the median value would be $\frac{5+9}{2} = 7$. Hence, the median of the data set X = [3, 4, 5, 9, 12, 80] is 7.

### 2.1.2(C) Mode

The mode for a set of data is the value that occurs most frequently in the set. It is possible for the greatest frequency to correspond to several different values, which results in more than one mode. Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal. In general, a data set with two or more modes is multimodal. At the other extreme if each data value occurs only once, then there is no mode for that data set.

Let's take an example.

Assume that you have the following data set.

$$X = [3, 5, 7, 3, 12, 5]$$

Number 3 and 5 are repeated twice in the data set. Hence, there are two elements that are most repeated in the data set. Hence, it a bimodal data set and the modes of the data set X = [3, 5, 7, 3, 12, 5] are 3 and 5.

Assume that you have the following data set.

$$X = [3, 5, 7, 3, 12, 3, 5, 3]$$

Here, the number 3 is repeated 4 times and is the most frequently appearing value in the data set. Hence, the mode of the data set X = [3, 5, 7, 3, 12, 3, 5, 3] is 3.

### 2.1.2(D) Mid-range

The mid-range is the average of the largest and smallest values in the data set.

For example, for the data set X = [3, 5, 7, 3, 12, 5], mid-range can be calculated as $\frac{3+12}{2} = 7.5$

### 2.1.2(E) Range

The range of a data set is the difference between the largest and smallest values in the data set.

For example, for the data set X = [3, 5, 7, 3, 12, 5], range can be calculated as 12 - 3 = 9.

### 2.1.3 Standard Deviation

- As you understand, you may not have all the data that you want every time. Most of the times, you draw samples from a large population in the hope that it represents the entire population to a great extent. By using a sample of the population, you can work out what is most likely to be the measurement if you used the entire population.
- Assume a simple dataset.

  $X = [1\ 2\ 4\ 6\ 12\ 15\ 25\ 45\ 68\ 67\ 65\ 98]$

  X refers to the entire dataset and elements within the dataset are denoted as subscript. For example, $X_3$ refers to the 3rd element in the sample whose value is.
- Mean of the sample is calculated as

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

$\bar{X}$, pronounced as (X-bar), denotes the mean of the sample. Mean is the average of the sample values.

So, for the given dataset, mean is calculated as

$$\bar{X} = \frac{1+2+4+6+12+15+25+45+68+67+65+98}{12}$$

$$= \frac{408}{12} = 34$$

- Unfortunately, the mean doesn't tell us a lot about the data except for a sort of middle point. For example, these two data sets have exactly the same mean (10) but are obviously quite different.

  $D = [0\ 8\ 12\ 20]$ and $E = [8\ 9\ 11\ 12]$
- So what is different about these two sets? It is the spread of the data that is different. The Standard Deviation (SD) of a data set is a measure of how spread out the data is. How do you calculate it? The English definition of the SD is "The average distance from the mean of the data set to a point". The way to calculate it is to compute the squares of the distance from each data point to the mean of the set, add them all up, divide by n - 1, and then take the positive square root.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

- Let's calculate the standard deviation of the samples D and E.

| X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 0 | - 10 | 100 |
| 8 | - 2 | 4 |
| 12 | 2 | 4 |
| 20 | 10 | 100 |
| | Total | 208 |

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{208}{3}} = 8.32$$

| X | $X - \bar{X}$ | $(X - \bar{X})^2$ |
|---|---|---|
| 8 | - 2 | 4 |
| 9 | - 1 | 1 |
| 11 | 1 | 1 |
| 12 | 2 | 4 |
| | Total | 10 |

$$s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{10}{3}} = 1.82$$

- So, what does SD calculation tell you about the samples? The first set has a much larger standard deviation due to the fact that the data is much more spread out from the mean whereas the data is closer to the mean in the second set. If the set would have been [10 10 10 10], then the SD would have been 0 as none of the datapoints deviate from the mean.

### 2.1.4 Variance

- Variance is another measure of the spread of data in a data set. It is calculated as $s^2$ where s is the standard deviation. So, if standard deviation is 1.82, then variance is $1.82^2 = 3.31$.

Mathematically, it can be written as

$$\text{var}(X) = s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

### 2.1.5 Covariance

- Standard deviation and variance work on 1-dimensional data. However, many data sets have more than one dimension, and the aim of the statistical analysis of these data sets is usually to see if there is any relationship between the dimensions. For example, you might have a data set having both, the height of all the students in a class, and the marks they received for a test. You could then perform statistical analysis to see if the height of a student has any effect on her marks.
- Standard deviation and variance only operate on 1-dimension. So, you can only calculate the standard deviation for each dimension of the data set independently of the other dimensions in the dataset. However, it is useful to have a similar measure to find out how much each of the dimensions vary from the mean with respect to each other.
- Covariance is such a measure. Covariance is always measured between 2 dimensions. If you calculate the covariance between one dimension and itself, you get the variance for that dimension.
- So, if you have a 3-dimensional data set (x, y, z), then you could measure the covariance between x and y dimensions, x and z dimensions, and y and z dimensions. Measuring the covariance between x and x, y and y, or z and z would give you the variance of the x, y and z dimensions respectively.
- The mathematical formula for covariance is very similar to the mathematical formula for variance. The formula for variance could also be written like

$$\text{var}(X) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

Covariance is mathematically computed as

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Note here that cov (X, Y) = cov (Y, X) as the calculation would remain unchanged.

- Let's take a sample data set having 2-dimesions, number of hours of study and marks obtained. Does number of hours affect marks obtained?

| Hours (H) | Marks (M) |
|---|---|
| 9 | 39 |
| 15 | 56 |
| 25 | 93 |
| 14 | 61 |
| 10 | 50 |
| 18 | 75 |
| 0 | 32 |
| 16 | 85 |
| 5 | 42 |
| 19 | 70 |
| 16 | 66 |
| 20 | 80 |

- You could calculate the covariance between hours of study and marks obtained to establish if there is any relationship between them.

| Hours (H) | Marks (M) | $H_i - \bar{H}$ | $M_i - \bar{M}$ | $(H_i - \bar{H})(M_i - \bar{M})$ |
|---|---|---|---|---|
| 9 | 39 | - 4.92 | - 23.42 | 115.23 |
| 15 | 56 | 1.08 | - 6.42 | - 6.93 |
| 25 | 93 | 11.08 | 30.58 | 338.83 |
| 14 | 61 | 0.08 | - 1.42 | - 0.11 |
| 10 | 50 | - 3.92 | - 12.42 | 48.69 |
| 18 | 75 | 4.08 | 12.58 | 51.33 |
| 0 | 32 | - 13.92 | - 30.42 | 423.45 |
| 16 | 85 | 2.08 | 22.58 | 46.97 |
| 5 | 42 | - 8.92 | - 20.42 | 182.15 |
| 19 | 70 | 5.08 | 7.58 | 38.51 |

| Hours (H) | Marks (M) | $H_i - \bar{H}$ | $M_i - \bar{M}$ | $(H_i - \bar{H})(M_i - \bar{M})$ |
|---|---|---|---|---|
| 16 | 66 | 2.08 | 3.58 | 7.45 |
| 20 | 80 | 6.08 | 17.58 | 106.89 |
| | | | | 1352.42 |

$$\text{cov}(H, M) = \frac{1352.42}{11} = 122.94$$

So what does it tell you?

The exact value is not as important as it's sign (positive or negative).

- If the value is positive, as it is here, then it indicates that both dimensions increase together, meaning that, in general, as the number of hours of study increased, so did the marks obtained.
- If the value is negative, then as one of the dimensions increases, the other decreases. If you had ended up with a negative covariance here, then that would have meant that as the number of hours of study increased, the marks obtained decreased.
- In the last case, if the covariance is zero, it indicates that the two dimensions are independent of each other.

### 2.1.6 Mean Absolute Deviation

The average absolute deviation (AAD) of a data set is the average of the absolute deviations from a central point. In the general form, the central point can be a mean, median, mode, or the result of any other measure of central tendency or any reference value related to the given data set. AAD includes the mean absolute deviation and the median absolute deviation (both abbreviated as MAD).

Using the following steps, you can calculate the mean absolute deviation (MAD) of a data set.

1. Calculate the desired central tendency point (mean, median, or mode) of the data set.
2. Use the formula $\frac{1}{n}\sum_{i=1}^{n} |x_i - m(X)|$ where m (X) is the central point that you calculated in step 1. Note here that || means that you need to take the absolute value of the difference (without sign) for this calculation.

Let's take an example.

Suppose that the data set is X = [2, 2, 3, 4, 14]. Based on the various measures of central tendency, you can calculate mean absolute deviation as following.

| Measure of Central Tendency m (X) | Mean Absolute Deviation |
|---|---|
| Mean = 5 | $\frac{\lvert 2-5\rvert + \lvert 2-5\rvert + \lvert 3-5\rvert + \lvert 4-5\rvert + \lvert 14-5\rvert}{5} = \frac{18}{5} = 3.6$ |
| Median = 3 | $\frac{\lvert 2-3\rvert + \lvert 2-3\rvert + \lvert 3-3\rvert + \lvert 4-3\rvert + \lvert 14-3\rvert}{5} = \frac{14}{5} = 2.8$ |
| Mode = 2 | $\frac{\lvert 2-2\rvert + \lvert 2-2\rvert + \lvert 3-2\rvert + \lvert 4-2\rvert + \lvert 14-2\rvert}{5} = \frac{15}{5} = 3$ |

### 2.1.7 ANOVA (Analysis of Variance)

**Q. Explain the following : ANOVA** SPPU - Aug. 18, 4 Marks , Oct. 19, 5 Marks

- When there are more than two groups for analysis, you use ANOVA.
- For example, consider a simple case where you want to determine if there is any difference in IQ levels of children of 20 different schools in a city. Each school has around 500 children from standard 1 to standard 10th. Is there a school which is better than others?
- You would likely need to carry out hypothesis testing for each school with the other 19 schools. Additionally, you would need to pick random samples (say 10 or so) from 500 children in each school for analysis. This is problematic for the following reasons –
  1. You need to carry out multiple tests for each pair of schools and it might be difficult to correlate the results to draw a clear conclusion.
  2. You are picking multiple samples for analysis and hence there are increased chances of making Type I or Type II errors in hypothesis testing.
- ANOVA (Analysis of Variance) helps to address these issues.

✍ ***Definition :*** *ANOVA is a hypothesis testing technique that tests various groups to find out if there is a difference between them.*

- The null hypothesis of ANOVA is that all the population means are equal. The alternative hypothesis is that at least one pair of the population means is not equal. The hypotheses are represented as following.

  $H_0 : \mu_1 = \mu_2 = \ldots = \mu_n$

  $H_A : \mu_i \neq \mu_j$ (for at least one pair of $i$ and $j$)
- ANOVA requires calculating the variance of means between-groups and within-groups. Based on the calculated means, the ratio of between-group variance and within-group variance is calculated. The larger is the ratio, the higher is the probability of populations being different (alternative hypothesis is true).
- ANOVA calculations could be tedious and are carried out using software tools or programs.

## 2.2 Concepts of Probability

- Probability theory isn't new for you. You would have already studied it in your school days.

  As you know,

✍ ***Definition :*** *Probability is the branch of mathematics concerning numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true.*

- The probability of an event is a number between 0 and 1. 0 indicates impossibility of the event and 1 indicates certainty (or true that event would or has occurred). The higher the probability of an event, the more likely it is that the event will occur. A simple example is the tossing of a fair (unbiased) coin. Since the coin is fair, the two outcomes ("heads" and "tails") are both equally probable.
- Hence, the probability of "heads" equals the probability of "tails" and since no other outcomes are possible, the probability of either "heads" or "tails" is $\frac{1}{2}$ (which could also be written as 0.5 or 50%).

- Let's understand a few more terms that are used with probability.

*Definition : A random experiment is a mechanism that produces a definite outcome that cannot be predicted with certainty.*

*Definition : The sample space associated with a random experiment is the set of all possible outcomes.*

*Definition : An event is a subset of the sample space. An event E is said to occur on a particular trial of the experiment if the outcome observed is an element of the set E.*

- So, what is the sample space of tossing a fair coin? The outcomes could be labelled h for heads and t for tails. Then the sample space is the set S = {h, t}. When you toss a coin, the event could result in either a heads or a tails. So, if you are interested in heads event, then E = {h}. To find the probability, you divide the count of events by the total number elements in the sample space. Hence, in this case, $P(h) = \frac{\{h\}}{\{h, t\}} = \frac{1}{2} = 50\ \%$
- You can write the expression such as p (A) that denotes the probability that the event A is true. For example, A might be the logical expression "it will rain tomorrow". The probability for any event ranges between 0 and 1.
- Hence, $0 \leq p(A) \leq 1$, where p (A) = 0 means the event definitely will not happen, and p (A) = 1 means that the event definitely will happen.
- You write $p(\bar{A})$ to denote the probability of the event not A. This could be written as p, $p(\bar{A}) = 1 - p(A)$. You also commonly write A = 1 to mean the event A is true, and A = 0 to mean the event A is false.

### 2.2.1 Fundamental Rules of Probability

Let's understand some of the common rules around probability.

#### 1. Probability of a Union of Two Events

- Given two events, A and B, the probability of A or B could be defined as follows.

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

- If the events are mutually exclusive, meaning that only one of them is possible to happen, then you could simply write the following.

$p(A \cup B) = p(A) + p(B)$ as $p(A \cap B) = 0$ since it is not possible that both the events can happen at the same time.

- For example, what is the probability of getting an even number when you roll a dice?

Even numbers that possible to get on a roll of dice are {2, 4, 6} = 3

Total possible outcomes are {1, 2, 3, 4, 5, 6} = 6

So,

$$p(2, 4, 6) = \frac{3}{6} = 50\ \%$$

You could have also approached this problem as a union of mutually exclusive events.

$$p(2, 4, 6) = p(2) + p(4) + p(6)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = 50\ \%$$

#### 2. Joint Probabilities

- The probability of the joint event A and B can be defined as following.

$$p(A, B) = p(A \cap B) = p(A \mid B) \times p(B)$$

- This is sometimes called the product rule. You will learn about in detail later on.
- $p(A \mid B)$ is read as probability of A given B is true. $p(A \mid B)$ is called as conditional probability.

### 3. Conditional Probability

- The conditional probability of event A, given that event B is true, is defined as following :

$$p(A \mid B) = \frac{p(A, B)}{p(B)} \quad \text{where } p(B) > 0$$

## 2.3 Tail Bounds

- For a random variable X, the tails of X are the parts of the probability distribution that are "far" from its mean. Following are a few examples of distributions with tails highlighted.

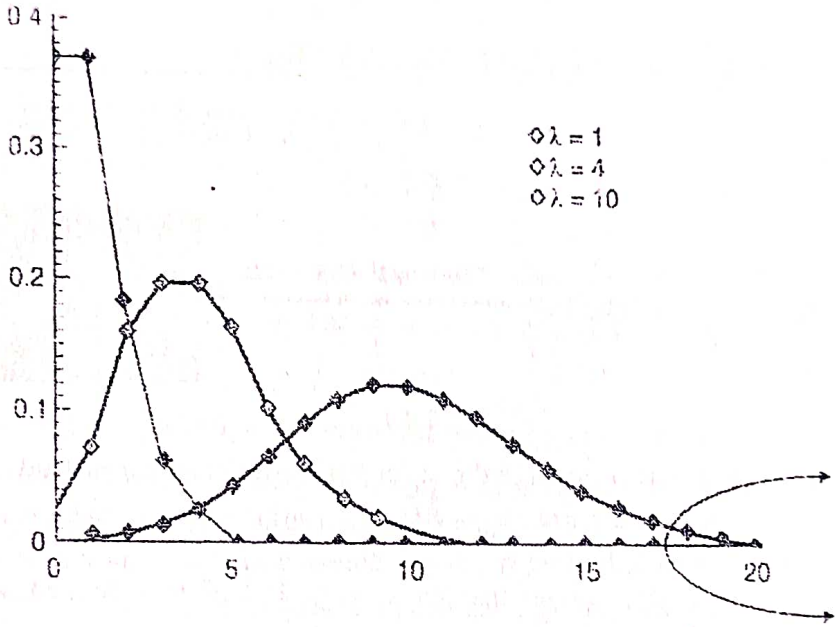


Fig. 2.3.1

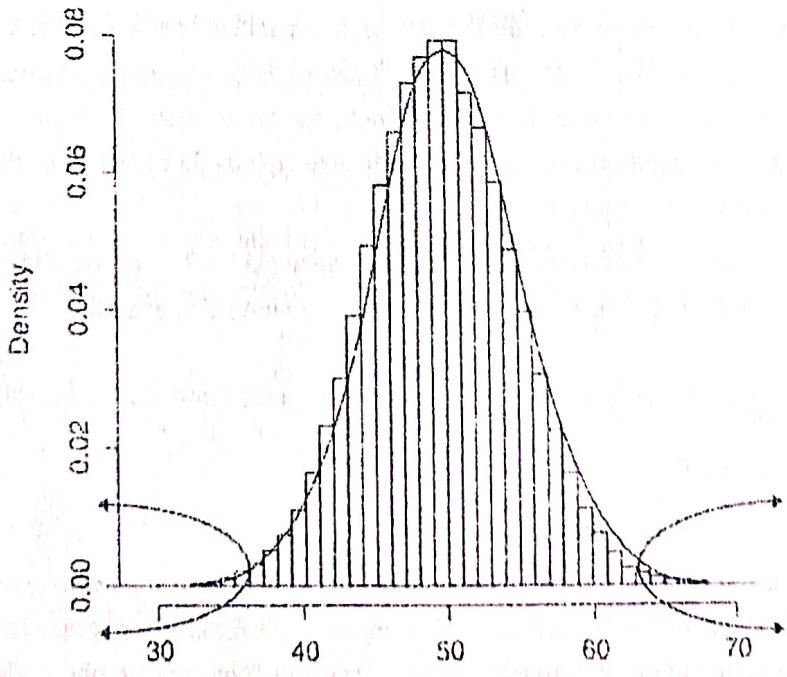


Fig. 2.3.2

- Basically, tail bounds give you an idea about how soon the distribution tends to 0. A distribution at times could be heavy tailed. A heavy tailed distribution has more weight in the tails.

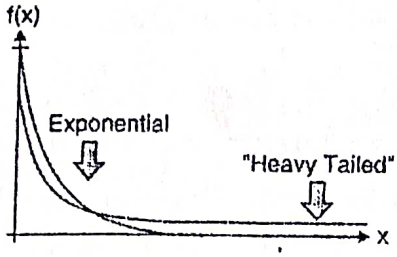


Fig. 2.3.3

- A heavy tailed distribution has a tail that is heavier than an exponential distribution. In other words, a distribution that is heavy tailed goes to zero slower than one with exponential tails which means that there will be more bulk under the curve of the probability distribution. Heavy tailed distributions tend to have many outliers with very high values. The heavier the tail, the larger the probability that you will get one or more disproportionate values in a sample.

## 2.4 Pair-Wise Independence and Universal Hashing

***Definition :*** *Hashing, in general, is the process of taking any length of input information and finding a unique fixed length representation of that input information.*

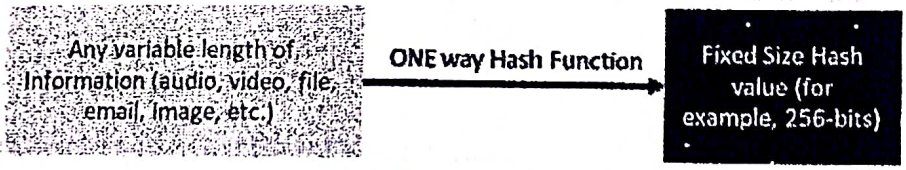


Fig. 2.4.1

- Hashing is the process of finding a unique message digest (or hash value) that corresponds to the input information. The length of input could be just one character or a huge video file. Hashing always produces a fixed size representation of the information. Hashing, in general, is used in several domains such as information security, cryptocurrency, high-performance programming, and for creating quick lookup tables.
- Hashing can be thought of as a way to rename an address space. For instance, a router at the internet backbone may wish to have a searchable database of destination IP addresses of packets that are flying by. An IP address could be up to 128 bits, so the number of possible IP addresses is $2^{128}$ which is too large to let you have a table indexed by IP addresses. Hashing allows you to rename each IP address by fewer bits. Furthermore, this renaming is done probabilistically, and the renaming scheme is decided in advance before you have seen the actual addresses. In other words, the scheme is ignorant to the actual addresses.
- Formally, you want to store a subset S of a large universe U (where $|U| = 2^{128}$ in the above example) and $|S| = m$ is a relatively small subset. For each $x \in U$, you want to support the following 3 operations.
  - Insert (x). Insert x into S.
  - delete (x). Delete x from S.
  - query (x). Check whether $x \in S$.
- A hash table can support all these 3 operations.
- For real-life applications, hash functions needs to have randomness so that it is hard to guess the hash value. One such property that is basic but is required for ensuring randomness of hash functions is pairwise independence. The goal of hash functions is to map elements from a large domain to a small one. Typically, to obtain the required guarantees, you would need not just one function, but a family of hash functions, where you would use randomness to sample a hash function from this family.

Let $H = (h : U \rightarrow R)$ be a family of functions, mapping elements from a (large) universe U to a (small) range R.

- A family $H = (h : U \rightarrow R)$ is said to be pairwise independent, if for any two distinct elements $x_1 \neq x_2 \in U$, and any two (possibly equal) values $y_1, y_2 \in R$,

$$\Pr_{h \in H}[h(x_1) = y_1 \text{ and } h(x_2) = y_2] = \frac{1}{|R|^2}$$

- Generally speaking, you encounter a trade-off. The more random H is, the greater the number of random bits needed to generate a function h from this class, and the higher the cost of computing h.

## 2.5 Approximate Counting

- Approximate Counting is a technique that allow you to count a large number of events using a very small amount of computer memory. It was invented by Robert Morris in 1977. That time, Morris was working on a programming situation that required using a large number of counters to keep track of the number of occurrences of many different events. That time computer memory was extremely limited, and it was not straightforward to store large numbers in say 8-bit registers.
- An n-bit register can ordinarily only be used to count up to $2^n - 1$. The counters were 8-bit bytes and because of the limited amount of storage available on the machine being used, it was not possible to use 16-bit counters. Using an intermediate size counter on a byte-oriented machine would have considerably increased both the complexity and running time of the program. The resulting limitation of the maximum count to 255 led to inaccuracies in the results, since the most common events were recorded as having occurred 255 times when in fact some of them were much more.
- Morris' approximate counting technique uses probabilistic techniques to increment the counter. Although it does not guarantee the exactness of count, it does provide a fairly good estimate of the true value while inducing a minimal and yet fairly constant relative error. Let's understand how it works.
- The most obvious way to count more than 255 events in an 8-bit register is to count only every other event. This can be done with a modest amount of error by simply flipping a coin at every event to decide whether or not to make the count. Not only is the expected error small, but it can be precisely described.
- In particular, if the number of events that have occurred is n, then the expected value for the value v in the counter is $\frac{n}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{v}{2}}$. For example, if 400 events have occurred, the value von the counter would be 200 and standard deviation would be $\sigma = 10$. So, you can expect that 95 percent of the time, the number of events estimated from 2v is within 40 of the actual count, an error of just 10 percent! That is precisely what is approximate counting
- Morris generalised the approximate counting function using log to minimise the error as following :

$$v(n) = \left( \frac{\log\left(1 + \frac{n}{a}\right)}{\log\left(1 + \frac{1}{a}\right)} \right)$$

- Where the parameter a controls both the maximum count that can be held in the register and the expected error. The constant $\log\left(1 + \frac{1}{a}\right)$ in the denominator serves only to force $n = 1$ to correspond to $v = 1$ so that the random procedures have no effect on the first count and counts of 0 and 1 are represented exactly.

## 2.6 Approximate Median

- As you understand, calculating median involves sorting the dataset and then picking the middle element. Often times, the dataset is too large to sort and then find median using the conventional method. Hence, you could use techniques to approximate the median value. Median of Medians is one such median approximation technique. Let's learn about it.
- The median of medians is an approximate median selection algorithm. It is frequently used to supply a good pivot (central point) for an exact selection algorithm. The median of medians algorithm chooses its pivot in the following clever way.
  1. Divide the list into sublists of length five. (Note that the last sublist may have length less than five)
  2. Sort each sublist and determine its median directly.
  3. Use the median of medians algorithm to recursively determine the median of the set of all medians from the previous step. (This step is what gives the algorithm its name)
  4. Use the median of the medians from step 3 as the pivot.

  Let's see an example.
- Suppose that you have the following list of unsorted elements for which you are interested to find the approximate median.

  [2, 3, 5, 4, 1, 12, 11, 13, 16, 7, 8, 6, 10, 9, 17, 15, 19, 20, 18, 23, 21, 22, 25, 24, 14]
- Divide the entire list into sublists of 5 elements each.
  - [2, 3, 5, 4, 1]
  - [12, 11, 13, 16, 7]
  - [8, 6, 10, 9, 17]
  - [15, 19, 20, 18, 23]
  - [21, 22, 25, 24, 14]
- Sort each of the sublists and find respective medians.

| Original Sublist | Sorted Sublist | Median from Sorted Sublist |
|---|---|---|
| [2, 3, 5, 4, 1] | [1, 2, 3, 4, 5] | 3 |
| [12, 11, 13, 16, 7] | [7, 11, 12, 13, 16] | 12 |
| [8, 6, 10, 9, 17] | [6, 8, 9, 10, 17] | 9 |
| [15, 19, 20, 18, 23] | [15, 18, 19, 20, 23] | 19 |
| [21, 22, 25, 24, 14] | [14, 21, 22, 24, 25] | 22 |

- Now, sort the sublist medians and then find median of the sorted sublist medians which is [3, 9, 12, 19, 22] = 12.
- So, the approximate median is 12.
- By the way, if you were to sort the entire list and find the true median it would come out as 13. So, an approximate median of 12 is pretty close, isn't it?
- Note here that if there were more elements in the list, say 100 elements, you would have first made initial 20 sublists of 5 elements each, and then you would have got 20 sublist medians from them. You would then recursively create sublist of the 20 sublist medians that you got. So, if you again make a sublist of 5 elements each form sublist medians, you would have got 4 sublists. You would again find median for each of these 4 sublists. Finally, out of 4 medians that you would have got, you would find the approximate median.

## 2.7 Random Variables

- Tell me a number. You might have said a number without caring whether it is large or small or negative or positive or decimal, right? If I ask the same question to 100 people, how likely am I to get the common answer? Probably not very likely because the range of numbers is nearly infinite. So, that's precisely what a random variable is. Here the number that I get from various people is random, picked up from the entire range of numbers. You won't say 7 if I ask you to tell me a number, at least if you are sane!

✍ ***Definition :*** *A random variable (also called as stochastic variable) is a type of variable in statistics whose possible values depend on the outcomes of a certain random phenomenon.*

- Since a random variable can take on different values, it is commonly labelled with a letter (e.g., variable "X"). Each variable possesses a specific probability distribution function (a mathematical function that represents the probabilities of occurrence of all possible outcomes). So, for example, if I ask you to give me a random even number between 1 and 10, you could only possibly give me {2, 4, 6, 8}. Hence, the probability of getting one of these numbers in a random experiment would be $\frac{4}{10} = 40\ \%$.
- Random variables are classified into discrete and continuous variables. The main difference between the two categories is the type of possible values that each variable can take. In addition, the type of (random) variable implies the particular method of finding a probability distribution function.

### 2.7.1 Discrete Random Variables

- A discrete random variable is a (random) variable whose values take only a finite number of values. The best example of a discrete variable is a dice. Throwing a dice is a purely random event. At the same time, the dice can take only a finite number of outcomes {1, 2, 3, 4, 5, 6}.
- Each outcome of a discrete random variable contains a certain probability. For example, the probability of each dice outcome is $\frac{1}{6}$ because the outcomes are of equal probabilities. Note that the total probability outcome of a discrete variable is equal to 1.
- So, if you were to plot the probability distribution for a discrete random variable, in this case for dice thrown, it could look like the Fig. 2.7.1.

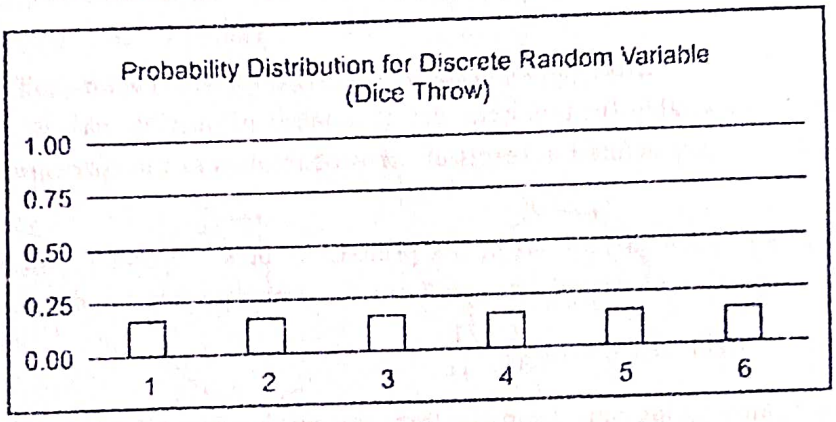


**Fig. 2.7.1**

- Also, the example I provided earlier, give me a random even number between 1 and 10, falls under discrete random variable category.
- In case of discrete random variables, you denote the probability of the event that X = x by p (X = x), or just p (x) for short. Here p (x) is called a probability mass function or pmf. This satisfies the properties $0 \leq p\ (x) \leq 1$ and $\sum_{x \in X} p\ (x) = 1$. So, X is the event where $x$ is one of the discrete random values and as you know the total probability outcome of a discrete variable is equal to 1.

### 2.7.2 Continuous Random Variables

- Unlike discrete variables, continuous random variables can take on an infinite number of possible values. One of the examples of a continuous variable is the returns of stocks that you own. The returns can take an infinite number of possible values (as percentages).
- Also, the example I provided earlier, tell me a number, falls under continuous random variable category where you could have answered any number.
- Due to the above reason, the probability of a certain outcome for the continuous random variable may look like to be zero. However, there is always a non-negative probability that a certain outcome will lie within the interval between two values.
- In case of continuous random variable, the probability density function or pdf is defined as,

$$f(x) = \frac{d}{dx} F(x)$$

- Given a pdf, you can compute the probability of a continuous variable being in a finite interval as following.

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

Where a and b are desired intervals where a possible value of X could be found.

### 2.7.3 Multiple Random Variables

- So far you learnt about one random variable in a sample space, but in most of the practical purposes there are two or more random variables on the same sample space. For these cases a random vector must be defined to contain the multiple random variables that you are interested in.
- An n - dimensional random vector is a function from a sample space S into $R^n$. In the bivariate (2 variables) case, n = 2.
- For example, consider the experiment of tossing two fair dice. The sample space for this experiment has 36 equally likely points. Let X = sum of the two dice and Y = | difference of two dice | (just considering the actual difference without sign).
- In this way, you have defined then bivariate random vector (X, Y). This random vector (X, Y) is called a discrete random vector because it has only a countable (in this case, finite) number of possible values. The probabilities of events defined in terms of X and Y are just defined in terms of the probabilities of the corresponding events in the sample space S.
- For example, suppose that you want to find out of the probability of X = 5 and Y = 3. So, you can get X = 5 and Y = 3 only when the first dice gets 4 and second dice gets 1 or when first dice gets 1 and second dice gets 4.

$$P(X = 5, Y = 3) = p(\{4, 1\}, \{1, 4\}) = \frac{2}{36} = \frac{1}{18}$$

- So, out of 36 total possible combinations, only 2 combinations can give X = 5 and Y = 3.

### 2.7.4 Markov Models

Markov Models is one of the most commonly used techniques in machine learning. It is commonly used for speech analysis, facial recognition, speech tagging, and gesture recognition. Let's dive deeper into understanding it.

**Markov Process**

***Definition : A Markov process is a chain of events that is memoryless.***

- It means that what is going to happen next depends only on the current state and not states previous to the current state.

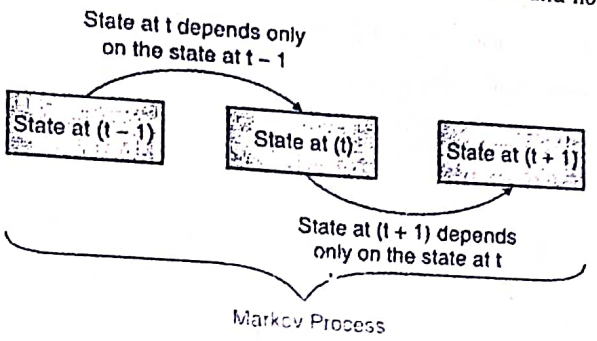


**Fig. 2.7.2 : Markov Process**

- For example, you decide your next move on a chess board based on the current state of the board and not based on the previous states of the board. Another example could be what you eat in your current meal depends on what you ate in your previous meal. So, if your previous meal was heavy, probably you would like to eat something light as the next meal.

## Markov Assumptions

Markov processes (and models based on these processes) make the following assumptions.

1. The number of possible outcomes or states is finite.
2. The outcome at any stage depends only on the outcome of the previous stage.
3. The probabilities of transitioning from one state to another state are constant over time.

## Markov Chain

**Definition :** *A Markov chain is a mathematical model that represents the state transitioning probabilities of a Markov process.*

- As you understand, the next state of a Markov process depends only upon the current state. Markov chain assigns state transitioning probabilities for a Markov process.
- For example, the Fig. 2.7.3 illustrates a simple Markov chain for weather.

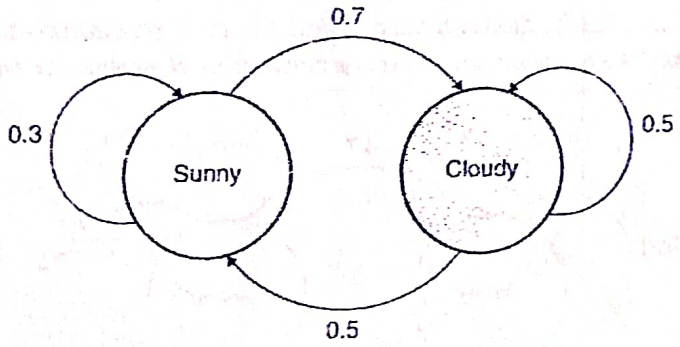


**Fig. 2.7.3 : Markov chain for weather**

- Assume that weather could only have two states – sunny or cloudy. Markov chain represents the state transitioning probabilities between the states.
- So, for the given Markov chain,

Probability of sunny remaining sunny = P (Sunny | Sunny) = 0.3

Probability of sunny becoming cloudy = P (Cloudy | Sunny) = 0.7

Probability of cloudy remaining cloudy = P (Cloudy | Cloudy) = 0.5

Probability of cloudy becoming sunny = P (Sunny | Cloudy) = 0.5

- As you know, note that the sum of probabilities of all outgoing arrows from a state should be 1. It cannot be higher than 1 (sure to happen).
- You can represent the four state transitioning probabilities as a matrix.

| | To Sunny | To Cloudy | |
|---|---|---|---|
| From Sunny | 0.3 | 0.7 | = 1 |
| From Cloudy | 0.5 | 0.5 | = 1 |

State Transitioning probabilities = $\begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix}$

### 2.7.5 Random Walk

A random walk is a mathematical object, known as a stochastic or random process, that describes a path that consists of a succession of random steps on some mathematical space such as the integers. For example, random walk on the integer number line, which starts at 0 and at each step moves +1 or −1 with equal probability. Markov chains helps you to carry out random walks on the established set of states and transition from one state to another.

#### Initial State

- Initial state or the starting state just represents the beginning of a Markov process. For example, the way you arrange your chess board at first remains the same and then once the game proceeds, you can have various states of the board.
- Extending the previous example from Markov chain, assume that the current day is sunny. So, what is the probability of the next day to be sunny or cloudy?
- Let's solve it out.

#### Practice Questions

**Ex. 2.7.1 :** For the given Markov chain, predict the respective probabilities of weather for next 3 days assuming that the current day is Sunny.

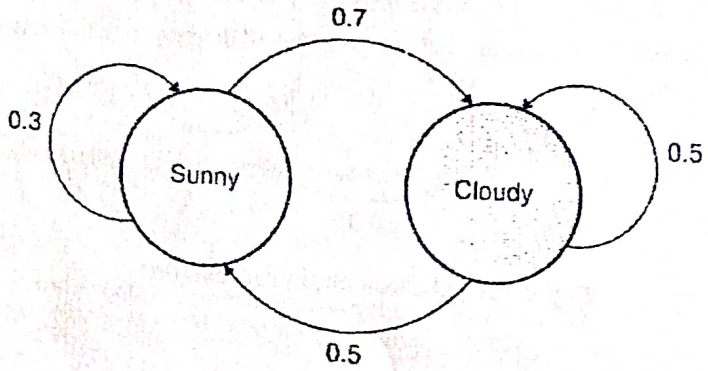


**Fig. P. 2.7.1**

**Soln. :**

- For the given Markov chain,

Probability of sunny remaining sunny = P (Sunny | Sunny) = 0.3

Probability of sunny becoming cloudy = P (Cloudy | Sunny) = 0.7

Probability of cloudy remaining cloudy = P (Cloudy | Cloudy) = 0.5

Probability of cloudy becoming sunny = P (Sunny | Cloudy) = 0.5

$$\text{State Transitioning Probabilities} = \begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix}$$

- Given that the current day is Sunny and not cloudy. So, it can be written as [1 0], where 1 is the probability of it being a sunny day (given in the question) and 0 is the probability of it being a cloudy day.

  As per Markov process, the next state depends only upon the current state $P(\text{Weather}_t \mid \text{Weather}_{t-1})$.

- Next state probability could be calculated by multiplying state transition matrix with the current state.

$$P(W_1 \mid W_0) = [1 \quad 0] \times \begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix} = [0.3 \quad 0.7]$$

- So, given that today is sunny, there is 0.3 probability that the next day would be sunny and 0.7 probability that the next day would be cloudy. This is also evident from the given Markov chain.

Let's calculate the probabilities for the next two days.

$$P(W_2 \mid W_1) = [0.3 \quad 0.7] \times \begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix} = [0.44 \quad 0.56]$$

$$P(W_3 \mid W_2) = [0.44 \quad 0.56] \times \begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix} = [0.41 \quad 0.59]$$

- Hence, the weather probabilities for next 3 days are as following.

| | Day 1 | Day 2 | Day 3 |
|---|---|---|---|
| Sunny | 0.3 | 0.44 | 0.41 |
| Cloudy | 0.7 | 0.56 | 0.59 |

---

**Ex. 2.7.2 :** Consider Markov chain model for 'Rain' and 'Dry' is shown in following Fig. P. 2.7.2.

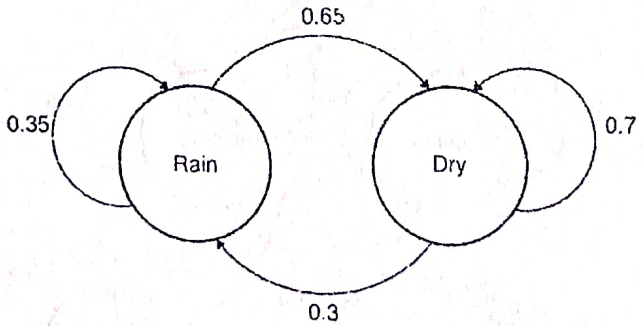


**Fig. P. 2.7.2**

Initial Probabilities are P('Rain')=0.4 and P('Dry') = 0.6

Calculate the probability for a sequence of states {'Dry', 'Rain', 'Rain', 'Dry'}

**Soln. :** For the given Markov chain,

$$P(\text{Rain} \mid \text{Rain}) = 0.35$$

$$P(\text{Dry} \mid \text{Rain}) = 0.65$$

$$P(\text{Rain} \mid \text{Dry}) = 0.3$$

$$P(\text{Dry} \mid \text{Dry}) = 0.7$$

$$\text{State Transitioning Probabilities} = \begin{bmatrix} 0.35 & 0.65 \\ 0.3 & 0.7 \end{bmatrix}$$

$$\text{Initial State} = [0.4 \quad 0.6]$$

$$P(W_1 \mid W_0) = [0.4 \quad 0.5] \times \begin{bmatrix} 0.35 & 0.65 \\ 0.3 & 0.7 \end{bmatrix} = [0.32 \quad 0.68]$$

So, for state 1, $P_1$ (Rain) = 0.32 and $P_1$ (Dry) = 0.68

$$P(W_2 \mid W_1) = [0.32 \quad 0.68] \times \begin{bmatrix} 0.35 & 0.65 \\ 0.3 & 0.7 \end{bmatrix} = [0.316 \quad 0.684]$$

So, for state 2, $P_2$ (Rain) = 0.316 and $P_2$ (Dry) = 0.684

$$P(W_3 \mid W_2) = [0.316 \quad 0.684] \times \begin{bmatrix} 0.35 & 0.65 \\ 0.3 & 0.7 \end{bmatrix} = [0.3158 \quad 0.6842]$$

So, for state 3, $P_3$ (Rain) = 0.3158 and $P_3$ (Dry) = 0.6842

$$P(W_4 \mid W_3) = [0.3158 \quad 0.6842] \times \begin{bmatrix} 0.35 & 0.65 \\ 0.3 & 0.7 \end{bmatrix} = [0.316 \quad 0.684]$$

So, for state 4, $P_4$ (Rain) = 0.316 and $P_4$ (Dry) = 0.684

You need probability for state transition as {'Dry', 'Rain', 'Rain', 'Dry'}. Multiply the respective probability of the 4 states.

$$P(\text{'Dry', 'Rain', 'Rain', 'Dry'}) = P_1(\text{Dry}) \times P_2(\text{Rain}) \times P_3(\text{Rain}) \times P_4(\text{Dry})$$

$$= 0.68 \times 0.316 \times 0.3158 \times 0.684 = 0.05$$

### 2.7.5(A) Steady State

- Markov chains are also useful for determining the chances of reaching a steady state where the probabilities for reaching a particular state is more or less the same over a longer period of time. It helps you to determine and answer questions like what percentage of times a particular state would be reached.
- Let's continue the previous example.

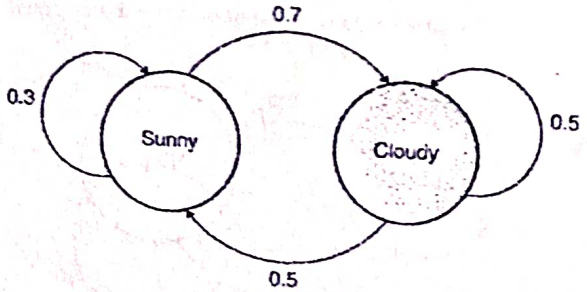


Fig. 2.7.4

- Earlier you found out that for the given Markov chain, the respective probabilities for various states are as following.

| | Day 1 | Day 2 | Day 3 |
|---|---|---|---|
| Sunny | 0.3 | 0.44 | 0.41 |
| Cloudy | 0.7 | 0.56 | 0.59 |

- If you keep on doing this exercise, you will find that the probabilities would more or less remain steady for days after days. That is when you say that you have reached the steady state (or say known routine or schedule). Let's call those probabilities be $\pi_1$ probability of being sunny in the steady state) and $\pi_2$ (probability of being cloudy in the steady state).

So,

| | Day 1 | Day 2 | Day 3 | Day... | Day n |
|---|---|---|---|---|---|
| Sunny | 0.3 | 0.44 | 0.41 | ... | $\pi_1$ |
| Cloudy | 0.7 | 0.56 | 0.59 | ... | $\pi_2$ |

- Assume that you want to calculate the steady state probability that the next day would be sunny.

You can write it as

$$\pi_1 = \pi_1 \times 0.3 + \pi_2 \times 0.5$$ [as 0.3 is the probability of a sunny day remaining sunny and 0.5 is the probability of a cloudy day becoming sunny the next day].

Also, $$\pi_1 + \pi_2 = 1$$ [as sum of probabilities is 1]

$$\pi_2 = 1 - \pi_1$$

Put the value of $\pi_2$ in the first equation you get,

$$\pi_1 = \pi_1 \times 0.3 + (1 - \pi_1) \times 0.5$$

$$\pi_1 = \frac{5}{12} = 0.42$$

$$\text{Hence, } \pi_2 = \frac{7}{12} = 0.58$$

- So, in the steady state there is 0.42 (or 42%) chance that the day would be sunny and 0.58 (58%) chance that the day would be cloudy. You could also say that 42% of all days would be sunny and 58% of all days would be cloudy. Very powerful derivation, isn't it?

## Practice Questions

**Ex. 2.7.3 :** A restaurant provides special discount on one of the items a day on pizza, burger, and hotdog as per the given Markov chain. Find out which items would be on discount for next 3 days if today the discount is on pizza.

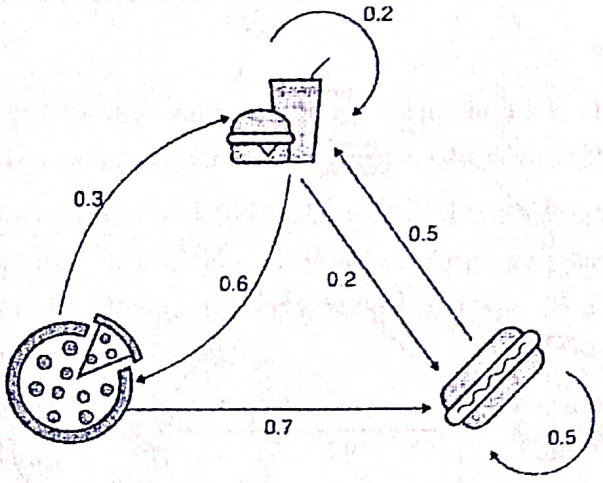


Fig. P. 2.7.3

- If you keep on doing this exercise, you will find that the probabilities would more or less remain steady for days after days. That is when you say that you have reached the steady state (or say known routine or schedule). Let's call those probabilities be $\pi_1$ probability of being sunny in the steady state) and $\pi_2$ (probability of being cloudy in the steady state).

So,

| | Day 1 | Day 2 | Day 3 | Day ... | Day n |
|---|---|---|---|---|---|
| Sunny | 0.3 | 0.44 | 0.41 | ... | $\pi_1$ |
| Cloudy | 0.7 | 0.56 | 0.59 | ... | $\pi_2$ |

- Assume that you want to calculate the steady state probability that the next day would be sunny.

You can write it as

$$\pi_1 = \pi_1 \times 0.3 + \pi_2 \times 0.5$$

[as 0.3 is the probability of a sunny day remaining sunny and 0.5 is the probability of a cloudy day becoming sunny the next day].

Also, $\pi_1 + \pi_2 = 1$ [as sum of probabilities is 1]

$$\pi_2 = 1 - \pi_1$$

Put the value of $\pi_2$ in the first equation you get,

$$\pi_1 = \pi_1 \times 0.3 + (1 - \pi_1) \times 0.5$$

$$\pi_1 = \frac{5}{12} = 0.42$$

$$\text{Hence, } \pi_2 = \frac{7}{12} = 0.58$$

- So, in the steady state there is 0.42 (or 42%) chance that the day would be sunny and 0.58 (58%) chance that the day would be cloudy. You could also say that 42% of all days would be sunny and 58% of all days would be cloudy. Very powerful derivation, isn't it?

## Practice Questions

**Ex. 2.7.3 :** A restaurant provides special discount on one of the items a day on pizza, burger, and hotdog as per the given Markov chain. Find out which items would be on discount for next 3 days if today the discount is on pizza.

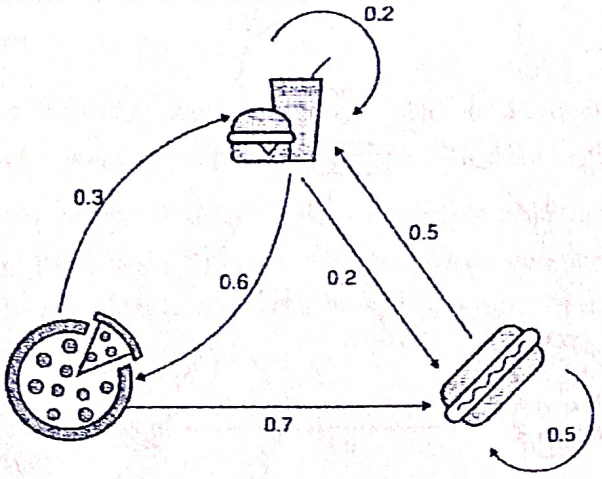


Fig. P. 2.7.3

**Soln. :**

First, arrange the state transition probabilities as a matrix.

| States | Burger | Pizza | Hotdog |
|---|---|---|---|
| Burger | 0.2 | 0.6 | 0.2 |
| Pizza | 0.3 | 0 | 0.7 |
| Hotdog | 0.5 | 0 | 0.5 |

$$\text{State Transitioning Probabilities} = \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Today, the discount is on Pizza.

Hence, Initial state = [0 1 0]

Let's calculate the probabilities for next 3 days.

$$P(D_1 \mid D_0) = [0 \quad 1 \quad 0] \times \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix} = [0.3 \quad 0 \quad 0.7]$$

$$P(D_2 \mid D_1) = [0.3 \quad 0 \quad 0.7] \times \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix} = [0.41 \quad 0.18 \quad 0.41]$$

$$P(D_3 \mid D_2) = [0.41 \quad 0.18 \quad 0.41] \times \begin{bmatrix} 0.2 & 0.6 & 0.2 \\ 0.3 & 0 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix} = [0.34 \quad 0.25 \quad 0.41]$$

Hence, the probabilities for discounts on respective items for next 3 days are as following.

| | Day 1 | Day 2 | Day 3 |
|---|---|---|---|
| Burger | 0.3 | 0.41 | 0.34 |
| Pizza | 0 | 0.18 | 0.25 |
| Hotdog | 0.7 | 0.41 | 0.41 |

**Ex. 2.7.4 :** A restaurant provides special discount on one of the items a day on pizza, burger, and hotdog as per the given Markov chain. Find out the percentage of days on which the respective items are on discount.

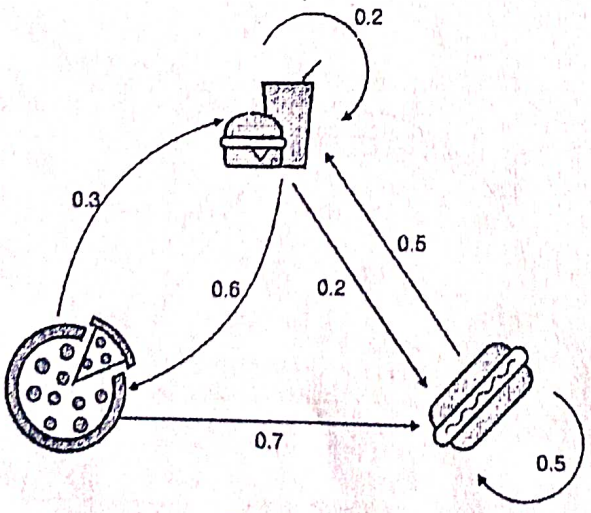


Fig. P. 2.7.4

**Soln. :**

Assume that

$\pi_B$ = Steady state probability for discount on burger

$\pi_P$ = Steady state probability for discount on pizza

$\pi_H$ = Steady state probability for discount on hotdog

- You can write steady state equations as following.

$$\pi_B = 0.2\pi_B + 0.3\pi_P + 0.5\pi_H$$

$$\pi_P = 0.6\pi_B$$

$$\pi_H = 0.2\pi_B + 0.7\pi_P + 0.5\pi_H$$

$$\pi_B + \pi_P + \pi_H = 1$$

- Solve these linear equations.

$$\pi_B = 0.2\pi_B + 0.3\,(0.6\pi_B) + 0.5\,(1 - \pi_B - \pi_P)$$

$$= 0.2\pi_B + 0.18\pi_B + 0.5 - 0.5\pi_B - 0.5 \times 0.6\pi_B$$

$$\pi_B = 0.5 - 0.42\pi_B$$

$$\pi_B = 0.35$$

$$\pi_P = 0.6\pi_B = 0.6 \times 0.35 = 0.21$$

$$\pi_H = 1 - \pi_B - \pi_P = 1 - 0.35 - 0.21 = 0.44$$

Hence,

- The steady state probability for discount on burger = 0.35 = 35% of the days the restaurant is likely to provide discount on burger
- The steady state probability for discount on pizza = 0.21 = 21% of the days the restaurant is likely to provide discount on pizza
- The steady state probability for discount on hotdog = 0.44 = 44% of the days the restaurant is likely to provide discount on hotdog

## 2.7.5(B) Hidden Markov Model

- A Markov chain is useful when you need to compute the probability for a sequence of observable events (states). In many cases, however, the states that you may be interested in could be hidden i.e. they are not directly observable.
- For example, you often don't know the reason behind someone being in a bad mood. There could be various reasons about that person being in a bad mood. If that person is your best friend, you may ask directly. But, if the person is a stranger or is someone who could get offended if you ask directly, you typically try to guess the probable reasons. So, these states are hidden and not directly observable.
- Let's understand it better with the following Fig. 2.7.5.

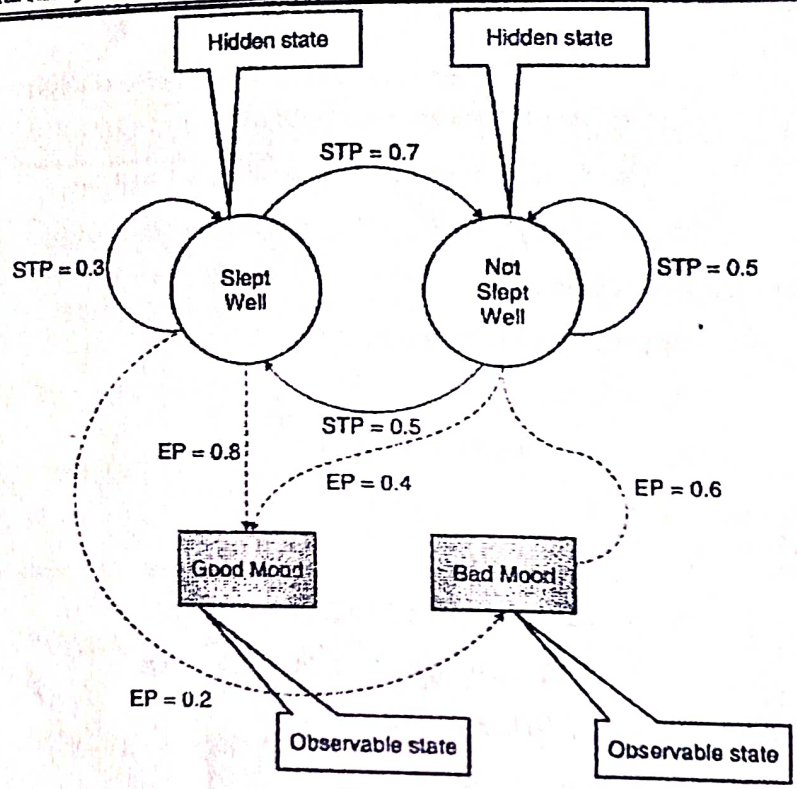


Fig. 2.7.5

- So, you can observe someone to be in either good or bad mood. Pretty straight forward, isn't it? But what is not directly observable (or known) is the reason behind that mood. Hence, the reasons are not observable, but those reasons do directly affect the observed state. These reasons or states are "hidden" from you as they are not directly observable.
- From your previous learning about Markov process, you have state transition probabilities (STP) that govern the transition between the "hidden" states. Based on the hidden states, the probabilities of the observed states are called emission probabilities (EP). Emission probabilities are outcomes of hidden states and are governed by state transition probabilities. So, let's write these down in the matrix format.
- State Transition Probabilities are given as following.

| | Slept Well | Not Slept Well |
|---|---|---|
| Slept Well | 0.3 | 0.7 |
| Not Slept Well | 0.5 | 0.5 |

$$\text{State Transition Probabilities} = \begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix}$$

- Emission Probabilities are given as following.

| | Good Mood | Bad Mood |
|---|---|---|
| Slept Well | 0.8 | 0.2 |
| Not Slept Well | 0.4 | 0.6 |

$$\text{Emission Probabilities} = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}$$

- A hidden Markov model (HMM) allows you to account for both – observable states as well as hidden states. The goal is to learn about hidden states by observing observable states.

## Practice Questions

**Ex. 2.7.5 :** Consider the following Hidden Markov Model where a professor picks a shirt colour based on his mood. Given that he picked green, blue, and red shirt for three days, predict his mood on those 3 days respectively.

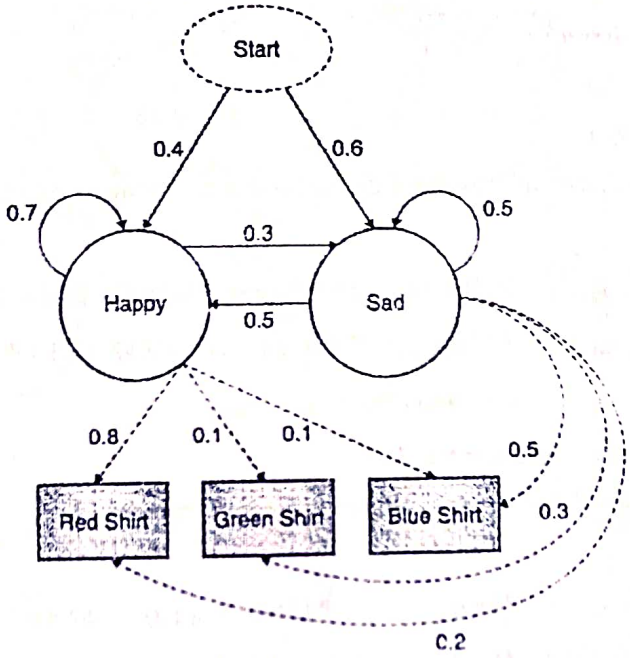


**Fig. P. 2.7.5**

**Soln. :**

- Happy and Sad are hidden states whereas Red Shirt, Green Shirt, and Blue shirt are observed states (as you can see the professor wearing those coloured shirts).
- For the given Markov chain,

$$P(\text{Happy} \mid \text{Happy}) = 0.7$$

$$P(\text{Sad} \mid \text{Happy}) = 0.3$$

$$P(\text{Happy} \mid \text{Sad}) = 0.5$$

$$P(\text{Sad} \mid \text{Sad}) = 0.5$$

$$\text{State Transitioning Probabilities} = \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\text{Initial State} = [0.4 \quad 0.6]$$

$$P(\text{Mood}_{\text{Day 1}} \mid \text{Mood}_{\text{Day 0}}) = [0.4 \quad 0.6] \times \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{bmatrix} = [0.58 \quad 0.42]$$

$$P(\text{Mood}_{\text{Day 3}} \mid \text{Mood}_{\text{Day 1}}) = [0.58 \quad 0.42] \times \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{bmatrix} = [0.62 \quad 0.38]$$

$$P(\text{Mood}_{\text{Day 3}} \mid \text{Mood}_{\text{Day 2}}) = [0.62 \quad 0.38] \times \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{bmatrix} = [0.62 \quad 0.38]$$

- So, based on the state transition probabilities, the mood on the respective 3 days are as following.

| | Day 1 | Day 2 | Day 3 |
|---|---|---|---|
| Happy | 058 | 0.62 | 0.62 |
| Sad | 0.42 | 0.38 | 0.38 |

- Now, as you understand, these states are not directly observable as these are hidden. Based on the emitted states and their emission probabilities, you need to find out the probabilities of these moods on the respective three days.

It is given that

1. On Day 1, shirt colour is Green
2. On Day 2, shirt colour is Blue
3. On Day 3, shirt colour is Red

- Probability of wearing a green colour shirt on day 1 depends upon the mood on day 1.

Hence,

$$P(\text{Green} \mid \text{Mood}_{\text{Day 1 = Happy}}) = P(\text{Happy}_{\text{Day 1}}) \times P(\text{Green} \mid \text{Happy}) = 0.58 \times 0.1 = 0.058$$

$$P(\text{Green} \mid \text{Mood}_{\text{Day 1 = Sad}}) = P(\text{Sad}_{\text{Day 1}}) \times P(\text{Green} \mid \text{Sad}) = 0.42 \times 0.3 = 0.13$$

$$P(\text{Green} \mid \text{Mood}_{\text{Day 1 = Sad}}) > P(\text{Green} \mid \text{Mood}_{\text{Day 1 = Happy}})$$

Hence, on Day 1, it is likely that the professor was Sad.

- Probability of wearing a blue colour shirt on day 2 depends upon the mood on day 2.

Hence,

$$P(\text{Blue} \mid \text{Mood}_{\text{Day 2 = Happy}}) = P(\text{Happy}_{\text{Day 2}}) \times P(\text{Blue} \mid \text{Happy}) = 0.62 \times 0.1 = 0.062$$

$$P(\text{Blue} \mid \text{Mood}_{\text{Day 2 = Sad}}) = P(\text{Sad}_{\text{Day 2}}) \times P(\text{Blue} \mid \text{Sad}) = 0.38 \times 0.5 = 0.19$$

$$P(\text{Blue} \mid \text{Mood}_{\text{Day 2 = Sad}}) > P(\text{Blue} \mid \text{Mood}_{\text{Day 2 = Happy}})$$

Hence, on Day 2, it is likely that the professor was Sad.

- Probability of wearing a red colour shirt on day 3 depends upon the mood on day 3.

Hence,

$$P(\text{Blue} \mid \text{Mood}_{\text{Day 3 = Happy}}) = P(\text{Happy}_{\text{Day 3}}) \times P(\text{Red} \mid \text{Happy}) = 0.62 \times 0.8 = 0.5$$

$$P(\text{Blue} \mid \text{Mood}_{\text{Day 3 = Sad}}) = P(\text{Sad}_{\text{Day 3}}) \times P(\text{Red} \mid \text{Sad}) = 0.38 \times 0.2 = 0.08$$

$$P(\text{Red} \mid \text{Mood}_{\text{Day 3 = Happy}}) > P(\text{Red} \mid \text{Mood}_{\text{Day 3 = Sad}})$$

Hence, on Day 3, it is likely that the professor was Happy.

- So, observing the colour of professor's shirt, you can guess his mood on the respective days as following.

| | Day 1 | Day 2 | Day 3 |
|---|---|---|---|
| Shirt Colour | Green | Blue | Red |
| Mood | Sad | Sad | Happy |

Powerful, isn't it ?

**Ex. 2.7.6 :** Consider two friends, Alice, and Bob, who live far apart from each other and who talk together daily over the telephone about what they did that day. Bob is only interested in three activities: walking in the park, shopping, and cleaning his apartment. The choice of what to do is determined exclusively by the weather on a given day. Alice has no definite information about the weather, but she knows general trends. Based on what Bob tells her he did each day; Alice tries

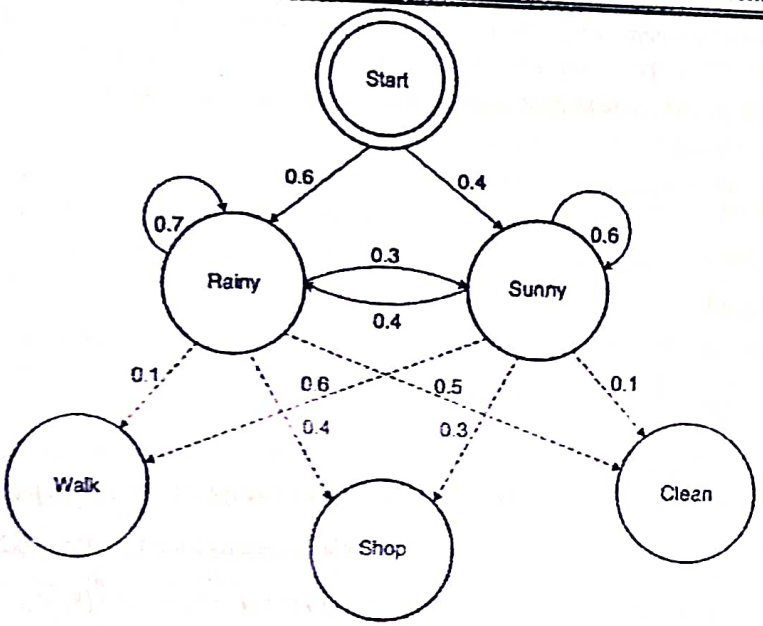


**Fig. P. 2.7.6**

Predict the weather on each of the 4 days, if Bob told Alice that he walked, shopped, shopped, and cleaned on those 4 days respectively.

**Soln. :**

- Rainy and Sunny are hidden states whereas walk, shop, and clean are observed states (as Alice knows this information directly from Bob over her telephonic conversation).
- For the given Markov chain,

$$P(\text{Rainy} \mid \text{Rainy}) = 0.7$$

$$P(\text{Sunny} \mid \text{Rainy}) = 0.3$$

$$P(\text{Rainy} \mid \text{Sunny}) = 0.4$$

$$P(\text{Sunny} \mid \text{Sunny}) = 0.6$$

$$\text{State Transitioning Probabilities} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

$$\text{Initial State} = [0.6 \quad 0.4]$$

$$P(\text{Weather}_{\text{Day 1}} \mid \text{Weather}_{\text{Day 0}}) = [0.6 \quad 0.4] \times \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = [0.58 \quad 0.42]$$

$$P(\text{Weather}_{\text{Day 2}} \mid \text{Weather}_{\text{Day 1}}) = [0.58 \quad 0.42] \times \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = [0.57 \quad 0.43]$$

$$P(\text{Weather}_{\text{Day 3}} \mid \text{Weather}_{\text{Day 2}}) = [0.57 \quad 0.43] \times \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = [0.57 \quad 0.43]$$

$$P(\text{Weather}_{\text{Day 4}} \mid \text{Weather}_{\text{Day 3}}) = [0.57 \quad 0.43] \times \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = [0.57 \quad 0.43]$$

- So, based on the state transition probabilities, the weather on the respective 4 days are as following.

| | Day 1 | Day 2 | Day 3 | Day 4 |
|---|---|---|---|---|
| Rainy | 058 | 0.57 | 0.57 | 0.57 |
| Sunny | 0.42 | 0.43 | 0.43 | 0.43 |

- Now, as you understand, these states are not directly observable as these are hidden. Based on the emitted states and their emission probabilities, you need to find out the probabilities of these weathers on the respective 4 days.

It is given that

1. On Day 1, Bob walked
2. On Day 2, Bob shopped
3. On Day 3, Bob shopped
4. On Day 4, Bob cleaned

- Probability of walking on day 1 depends upon the weather on day 1.

Hence,

$$P(\text{Walk} \mid \text{Weather}_{\text{Day 1 = Rainy}}) = P(\text{Rainy}_{\text{Day 1}}) \times P(\text{Walk} \mid \text{Rainy}) = 0.58 \times 0.1 = 0.058$$

$$P(\text{Walk} \mid \text{Weather}_{\text{Day 1 = Sunny}}) = P(\text{Sunny}_{\text{Day 1}}) \times P(\text{Walk} \mid \text{Sunny}) = 0.42 \times 0.6 = 0.25$$

$$P(\text{Walk} \mid \text{Weather}_{\text{Day 1 = Sunny}}) > P(\text{Walk} \mid \text{Weather}_{\text{Day 1 = Rainy}})$$

Hence, on Day 1, it is likely that the weather was Sunny.

- Probability of shopping on day 2 depends upon the weather on day 2.

Hence,

$$P(\text{Shop} \mid \text{Weather}_{\text{Day 2 = Rainy}}) = P(\text{Rainy}_{\text{Day 2}}) \times P(\text{Shop} \mid \text{Rainy}) = 0.57 \times 0.4 = 0.23$$

$$P(\text{Shop} \mid \text{Weather}_{\text{Day 2 = Sunny}}) = P(\text{Sunny}_{\text{Day 2}}) \times P(\text{Shop} \mid \text{Sunny}) = 0.43 \times 0.4 = 0.13$$

$$P(\text{Shop} \mid \text{Weather}_{\text{Day 2 = Rainy}}) > P(\text{Shop} \mid \text{Weather}_{\text{Day 2 = Sunny}})$$

Hence, on Day 2, it is likely that the weather was Rainy.

- Probability of shopping on day 3 depends upon the weather on day 3.

Hence,

$$P(\text{Shop} \mid \text{Weather}_{\text{Day 3 = Rainy}}) = P(\text{Rainy}_{\text{Day 3}}) \times P(\text{Shop} \mid \text{Rainy}) = 0.57 \times 0.4 = 0.23$$

$$P(\text{Shop} \mid \text{Weather}_{\text{Day 3 = Sunny}}) = P(\text{Sunny}_{\text{Day 3}}) \times P(\text{Shop} \mid \text{Sunny}) = 0.43 \times 0.3 = 0.13$$

$$P(\text{Shop} \mid \text{Weather}_{\text{Day 3 = Rainy}}) > P(\text{Shop} \mid \text{Weather}_{\text{Day 3=sunny}})$$

Hence, on Day 3, it is likely that the weather was Rainy.

- Probability of cleaning on day 4 depends upon the weather on day 4.

Hence,

$$P(\text{Clean} \mid \text{Weather}_{\text{Day 4 = Rainy}}) = P(\text{Rainy}_{\text{Day 4}}) \times P(\text{Clean} \mid \text{Rainy}) = 0.57 \times 0.5 = 0.29$$

$$P(\text{Clean} \mid \text{Weather}_{\text{Day 4 = Sunny}}) = P(\text{Sunny}_{\text{Day 4}}) \times P(\text{Clean} \mid \text{Sunny}) = 0.43 \times 0.1 = 0.043$$

$$P(\text{Clean} \mid \text{Weather}_{\text{Day 4 = Rainy}}) > P(\text{Clean} \mid \text{Weather}_{\text{Day 4=Sunny}})$$

Hence, on Day 4, it is likely that the weather was Rainy.

- So, talking with Bob and knowing his day's activity, Alice can guess the weather on the respective days as following.

| | Day 1 | Day 2 | Day 3 | Day 4 |
|---|---|---|---|---|
| Activity | Walk | Shop | Shop | Clean |
| Weather | Sunny | Rainy | Rainy | Rainy |

## 2.8 Flajolet-Martin (FM) Algorithm (LogLog Counting)

- You would have often come across a scenario where you needed to find the count of unique elements in a set. For example, in the set [1, 2, 3, 4, 2, 1, 4, ,4, 2, 1, 3] the unique elements are [1, 2, 3, 4] and the count of unique elements is 4. If the input dataset or stream is too large, you would need to store every value before you can confidently determine the exact count of unique elements, isn't it? The Flajolet-Martin (FM) algorithm precisely solves this problem where you don't need to store or know every value in the stream before approximating the count of unique elements. It is also known as LogLog counting.
- The FM algorithm approximates the number of unique objects in a stream or a database in one pass. If the stream contains $n$ elements with $m$ of them unique, this algorithm runs in O (n) time and needs O (log (m)) memory. So, as you see, it uses very less memory to approximate unique number of objects in a stream. Note here that the FM algorithm gives an approximation of the count and not the exact count. It may have approximation errors.
- The algorithm works as following.
  1. A hash converts every element received from the stream into a number.
  2. The algorithm converts the number into binary.
  3. The algorithm counts the number of trailing zeros in the binary number and tracks of the maximum number it sees as n.
  4. The algorithm estimates the number of distinct elements passed in the stream as $2^n$.
- Let's see a simple example.
- Suppose you have a data stream such as x = [1, 3, 5, 7, 5, 2, 7] and the hash function is h (x) = $(3x_i + 1)$ mod 5

The algorithm works as following.

| x | Hash Value h (x) | Binary of hash value | Number of trailing zeroes |
|---|---|---|---|
| 1 | (3 × 1 + 1) mod 5 = 4 | 100 | 2 |
| 3 | (3 × 3 + 1) mod 5 = 0 | 000 | 0 (as value is 0) |
| 5 | (3 × 5 + 1) mod 5 = 1 | 001 | 0 |
| 7 | (3 × 7 + 1) mod 5 = 2 | 010 | 1 |
| 5 | (3 × 5 + 1) mod 5 = 1 | 001 | 0 |
| 2 | (3 × 2 + 1) mod 5 = 2 | 010 | 1 |
| 7 | (3 × 7 + 1) mod 5 = 2 | 010 | 1 |

- So, the maximum number of trailing zeroes is 2. Hence, the approximate count of unique elements in the stream is equal to $2^2 = 4$.
- As you see, the exact count of unique elements is 5 but approximate count of unique elements is very close to the actual count.

## 2.9 Bloom Filters

- Often times, you are interested to figure out if an item is in a set or not. For example, when you create an email account, your chosen username is checked if it already exists or not. If it already exists, you are forced to choose a different username until you come up with a username that is not already taken.

- A Bloom filter is a simple, space-efficient probabilistic data structure based on hashing that represents a set in a way that allows membership queries to determine whether an element is a member of the set. False positives (wrongly returning that an element exists when actually it is not) are possible, but not false negatives (wrongly returning that the element does not exist when actually it does). In many applications, the space savings afforded by Bloom filters outweigh the drawbacks of a small probability for a false positive.
- Various extensions of Bloom filters can be used to handle alternative settings, such as when elements can be inserted and deleted from the set, and more complex queries, such as when each element has an associated function value that should be returned. Bloom filters have several applications. For example, Google Bigtable, Apache HBase and Apache Cassandra and PostgreSQL use Bloom filters to reduce the disk lookups for non-existent rows or columns. Avoiding costly disk lookups considerably increases the performance of a database query operation.
- Let's understand how Bloom filters work.
  1. An empty Bloom filter is a bit array of m bits, all set to 0 initially.
  2. You define k different hash functions each of which maps or hashes elements of a set into one of the m array positions.
  3. To add an element to the Bloom filter array, you feed it to each of the k hash functions to get k array positions. You then set the bits at all these k positions to 1.
  4. To query for an element (test whether it is in the set), you feed it to each of the k hash functions to get k array positions.
     (a) If any of the bits at these k positions is 0, then the element is definitely not in the set. If the element was to be in the set, then all the k positions in the Bloom filter array should have been 1.
     (b) If all the bits at k positions are 1, then either the element is in the set, or the bits have by chance been set to 1 during the insertion of other elements, resulting in a false positive. In a simple Bloom filter, there is no way to distinguish between the two cases, but more advanced techniques can address this problem.
- Let's see a simple example to understand this.
- Suppose the initial Bloom filter has an array of 10 positions all initialised to 0.

**Initial Bloom Filter**

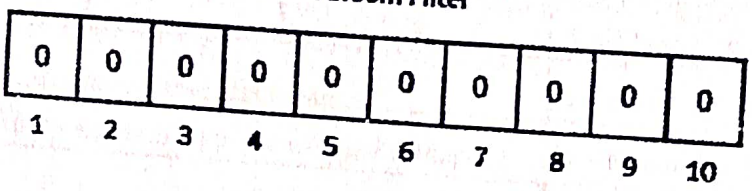


**Fig. 2.9.1**

- Suppose that you have two words {cat, rat}. Also, assume that you have three hash functions. For inserting the elements in the set into the Bloom filter, you pass each element through these three hash functions to get the respective bit positions in the Bloom filter.
  - $h_1$ (cat) = bit position 1
  - $h_2$ (cat) = bit position 9
  - $h_3$ (cat) = bit position 4
  - $h_1$ (rat) = bit position 2
  - $h_2$ (rat) = bit position 3
  - $h_3$ (rat) = bit position 6

- You then set these bit positions to 1.

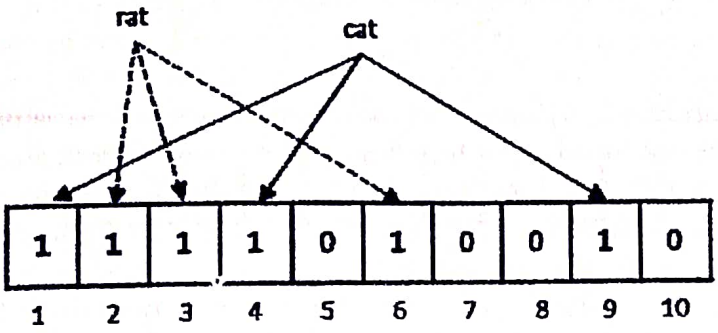


Fig. 2.9.2

- Now, suppose that you have to determine if the element *rat* is there in the set or not. You pass the element rat to the same three hash functions to get the bit positions that you should check to be 1.

- Passing through the hash functions, you would find the bit positions as following.
  - $h_1$ (rat) = bit position 2
  - $h_2$ (rat) = bit position 3
  - $h_3$ (rat) = bit position 6
- You would then check the bit positions 2, 3 and 6 in the Bloom filter to check if all the three bit positions are set to 1 or not.

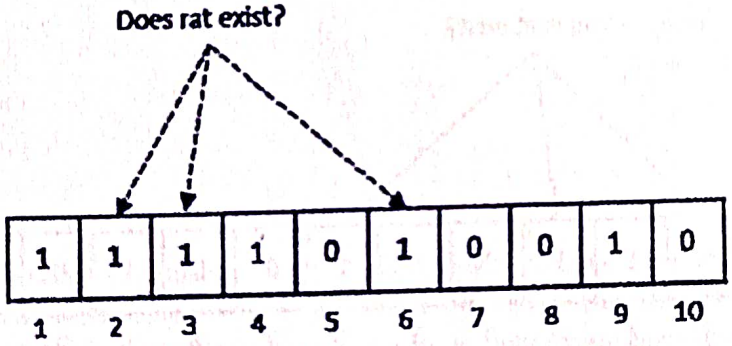


Fig. 2.9.3

- You find that the positions 2, 3, and 6 are all set to 1. Hence, on the basis of Bloom filter output, you can say that the element rat probably exists.
- Now, suppose that you want to determine if the element dog exists. You pass the element dog to the same three hash functions to get the bit positions that you should check to be 1.
- Passing through the hash functions, assume that you got the bit positions as following.
  - $h_1$ (dog) = bit position 2
  - $h_2$ (dog) = bit position 5
  - $h_3$ (dog) = bit position 10
- You would then check the bit positions 2, 5 and 10 in the Bloom filter to check if all the three bit positions are set to 1 or not.

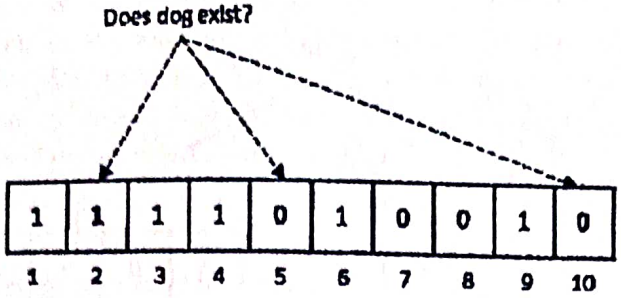


Fig. 2.9.4

- You find that only the bit position 2 is set to 1 and the bit positions 5 and 10 are set to 0. Hence, you conclude that the element dog is not in the set.
- Now, let's see a false positive scenario where an element say mat is checked for existence in the set. You pass the element mat to the same three hash functions to get the bit positions that you should check to be 1.

- Passing through the hash functions, assume that you got the bit positions as following.
  - $h_1$ (mat) = bit position 1
  - $h_2$ (mat) = bit position 3
  - $h_3$ (mat) = bit position 6
- You would then check the bit positions 1, 3 and 6 in the Bloom filter to check if all the three bit positions are set to 1 or not.

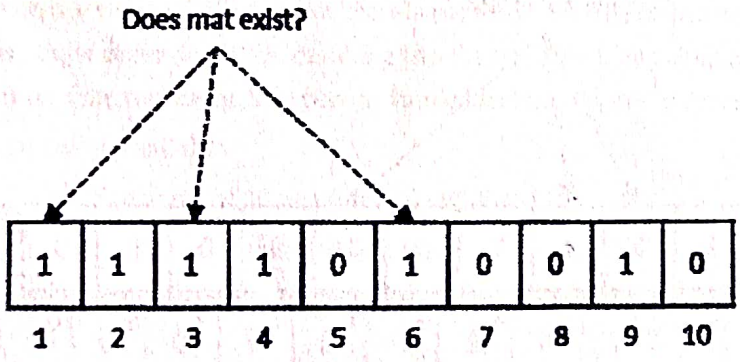


Fig. 2.9.5

- From the Bloom filter, you find that the bit positions 1, 3, and 6 are all set to 1. You wrongly conclude that the element mat exists in the set where you know clearly that those positions were set to 1 when you inserted the elements cat and rat.

## 2.10 Distance Sampling and Random Projections

- Before continuing here, please read about dimensionality reduction topic covered under Unit 4 - "Data Reduction".
- Dimensionality reduction techniques such as Principal Component Analysis (PCA) transform the data linearly into a lower-dimensional space but are expensive. The time taken to find the transformation (which is a matrix comprising the eigenvectors of the covariance matrix) is cubic in the number of dimensions. This makes it infeasible for datasets with a large number of attributes. A far simpler alternative is to use a random projection of the data into a subspace with a predetermined number of dimensions.

- In essence, Random Projection is an extremely simple linear dimensionality reduction method. Just like any other linear dimensionality reduction algorithm, Random Projection reduces the dimension of samples by applying a linear transformation to input data, so that each output feature is computed as a linear combination of the original features. However, the main difference between Random Projection and other approaches is that it generates the projection matrix from a random distribution. Therefore, as opposed to other methods where training data is required to select an appropriate projection matrix, Random Projection is a data-independent method. This means that no knowledge about the distribution of data is required to generate the projection matrix. Surprisingly, if an appropriate distribution is used to generate the entries of the projection matrix, the structure of data in the high-dimensional input feature space can be mostly preserved after the projection. Moreover, the projection matrix can be sparse and made of integers, allowing for computational savings and efficient implementation in database environments.
- In Random Projections, the original data $X \in \mathbb{R}^a$ is transformed to the lower dimensional $S \in \mathbb{R}^k$ with $k << a$ using $S = RX$ where the columns of R are realisations of independent and identically distributed (i.i.d.) zero-mean normal variables, scaled to have the unit length.
- Gaussian random matrix and Sparse random matrix are the two major types of random matrices used for random projections. The Gaussian Random Projection reduces the dimensionality by projecting the original input space on a randomly generated matrix where components are drawn from the $N\left(0, \frac{1}{n_{components}}\right)$ whereas the Sparse Random Projection reduces the dimensionality by projecting the original input space using a sparse random matrix. Sparse random matrices are an alternative to dense Gaussian random projection matrix that guarantees similar embedding quality while being much more memory efficient and allowing faster computation of the projected data. If you define $S = \frac{1}{density}$, then the elements of the random matrix are drawn as following.

$$\begin{cases} -\sqrt{\frac{S}{n_{components}}}\frac{1}{2s} \\ 0 \text{ with probability } 1 - \frac{1}{s} \\ +\sqrt{\frac{s}{n_{components}}}\frac{1}{2s} \end{cases}$$

where $n_{components}$ is the size of the projected subspace.

- Random Projection has become a widespread tool for dimensionality reduction, especially in large-scale applications where the volume of data or the dimensionality of samples is too big for alternative methods. For instance, Random Projection has been successfully used to accelerate tasks such as multivariate correlation analysis, high-dimensional data clustering, image search, and texture classification.

## Review Questions

Here are a few review questions to help you gauge your understanding of this chapter. Try to attempt these questions and ensure that you can recall the points mentioned in the chapter.

**Q. 1** With an example, explain mean of a dataset. **[4 Marks]**

**Q. 2** With an example, explain median of a dataset. **[4 Marks]**

**Q. 3** With an example, explain mode of a dataset. **[4 Marks]**

**Q. 4** With an example, explain mid-range of a dataset. **[4 Marks]**

**Q. 5** With an example, explain range of a dataset. **[4 Marks]**

**Q. 6** With an example, explain standard deviation of a dataset. **[6 Marks]**