Linear Algebra & Statistical Modeling for Data Science

UNIT II

Random Variables and Probability Functions,notion of probability, distributions,Mean, variance, Covariance, Hypotheses Testing, Type 1 and Type 2 errors. Testing for parameters of a normal distribution and for percentages based on a single sample and based on two samples. Introduction to the chisquared test. The concept of p-value. Mean-square estimation and Kalman filtering.

06 - Hrs

Recap

- What is Linear Equations
- Solving Linear Eqns.
- What is Hyperplane in Linear Algebra
- What is EigenValue
- EigenVector
- Properties of EigenValue
- Intro. Statistical Modeling

Introduction – Statistics Model

- The **health and welfare council** wants to survey your city to understand the lifestyle of the residents and if need be, improve the facilities provided based on the conclusions drawn from the data collected.
- **Some of the data being collected for this analysis is listed below:**

  1. Personal information such as name, age, income of household and educational qualifications.

  2. Quality of water supply across the city.

  3. Medical facilities such as availability of hospitals, doctors etc.
- The collection, analysis, interpretation, presentation and organization of data is termed as **statistics.**

Introduction – Statistics Model

- Statistics is used to study a population/process (data).
- In the example cited earlier, the data is personal information of residents, water being supplied and information about the medical facilities.
- We can leverage the analysis of this data to answer a few questions which could be but not limited to,

  1. What is the average income of households in the city?
  2. Which is the most common disease in a given area?
  3. How much water does a household use in a month?
  4. How good is the water being supplied?
  5. Are there enough medical facilities being provided?

- When all data required for observation/analysis is collected and studied, the data is referred to as the population.
- On the other hand, when limited data is being collected/analyzed, this data is referred to as sample and is used as an indicative of the entire population.

Types of Statistics

- **Example:**
- **Population:** Data is being collected from all the residents of the city about their age, educational qualification and medical history.
- **Sample:** On the other hand to analyze the water supply, a few litres of water are being collected from different water supply lines that run in the city.
- Statistics can be broadly classified into descriptive and inferential statistics.

1. Descriptive statistics:

- Summarization of data to describe the main features of the sample.
- **For example,** in the survey cited earlier, the knowledge of descriptive statistics can be leveraged to answer some questions like:

1. What is the average income of households in the city?

2. Which is the most common disease in a given area?

Types of Statistics

2. Inferential statistics:

- When working with samples of the population the techniques and processes we use to draw conclusions come under inferential statistics
- For example, in the survey cited earlier, inferential statistics can be used to answer the following questions:

1. How good is the water being supplied?

2. Are there enough medical facilities being provided?

Statistical Modeling

- Statistical modeling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.
- When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically.
- Rather than sifting through the raw data, this practice allows them to identify relationships between variables, make predictions about future sets of data, and visualize that data so that non-analysts and stakeholders can consume and leverage it.
- When you analyze data, you are looking for patterns,"Say Hello". You are using a sample to make an inference about the whole.

Reasons to Learn Statistical Modeling

- While **data scientists** are most often tasked with building models and writing algorithms, analysts also interact with statistical models in their work on occasion.
- For this reason, **analysts** who are looking to excel should aim to obtain a solid understanding of what makes these models successful.
- As machine learning and artificial intelligence become more commonplace, more and more companies and organizations are leveraging **statistical modeling** in order to make predictions about the future data.
- Below are some of the benefits that come from having a thorough understanding of statistical modeling.

1. You will be better equipped to choose the right model for your needs.
2. You will be better able to prepare your data for analysis.
3. You will become a better communicator.

Techniques in Statistical Modeling

- There are several statistical modeling techniques used during data exploration/analysis/ prediction etc,

  1. Linear Regression

  2. Classification

  3. Tree-Based Methods

  4. Supervised Learning

  5. Unsupervised Learning

  6. Neural Networks

What is Probability?

- Probability is a mathematical subject which enables us in determining or predicting how likely it is that an event will happen.
- The probability of occurrence is assigned a value from 0 to 1.
- When the value assigned is 1, it implies that the event will happen with all certainty.
- On the other hand when it is 0, it implies that the event is not likely to take place.
- Thus, we can be more certain of an event's occurrence when its probability is higher.

Chance is a possibility of something happening.

1. Is it possible that we observe a black colour sun?

Answer: It is not possible to observe a black colour sun. The chance of this happening is zero.

2. Is it possible that if today is a Monday, tomorrow is a Tuesday?

Answer: Definitely, it is possible. There is a 100 % chance that tomorrow is a Tuesday if it is a Monday today.

3. Is it possible that we have a pizza for dinner?

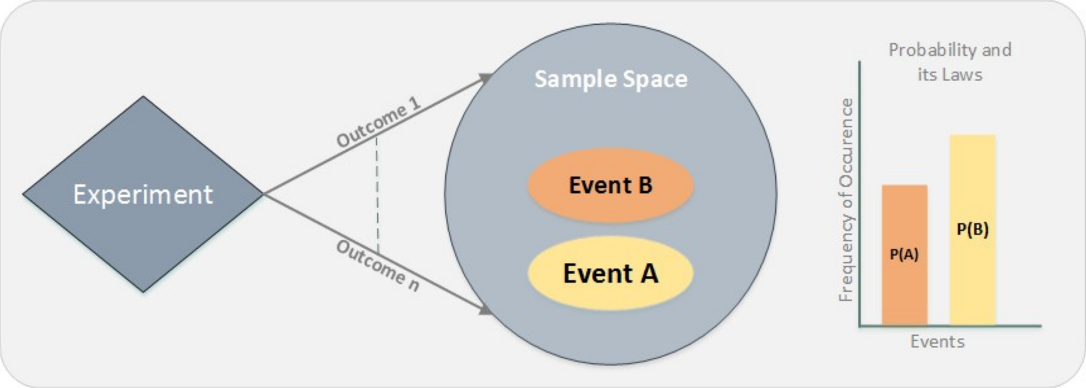
Answer: It may or may not be possible, there is some chance. We may not be absolutely sure about having a pizza for dinner.

- **Observation:**

Example 1 and example 2 had a clear possibility of occurrence. However, example 3 had an unclear possibility of occurrence.

Probabilistic Model

- The probabilistic model is a generic structure which describes the random outcomes of an activity.
- This model helps assign the likelihood of occurrence to a collection (set) of the random outcomes.



The process of observation of an activity is termed as an **experiment.**
The results of an observation are termed as **outcomes** of the Experiment.
The events for which we cannot calculate the outcomes, those experiments are called **random experiments.**

Introduction to Probability Distributions

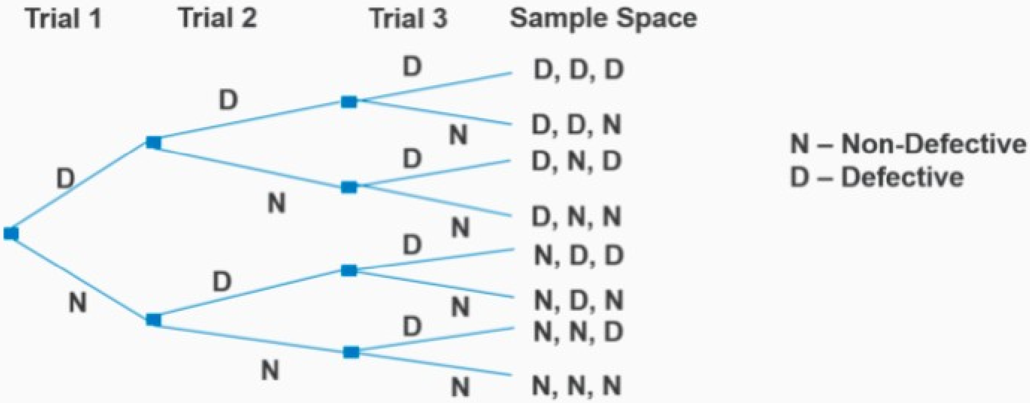
- **Consider a Scenario: Random Variable**
- Sashi wants to sell home-made ice-creams and waffle-cones in a mobile ice-cream van.
- She wants to sell 5 different flavors of ice-cream namely, Vanilla, Chocolate, Strawberry, Black-Currant and Pistachio.

Probabilistic Model

- Sashi plans to begin selling ice-creams in the month of March. She manufactures a batch of 1200 waffle-cones for the same.
- She then examines random waffle-cones and judges each cone as either **“defective”** or **“non-defective”.** She decides to examine 3 waffle-cones.
- The process of observation of activity is termed as an **experiment.**
- The results of the observation are termed as **outcomes of the experiment.**
- **Random experiments** are those experiments whose outcomes can't be predicted.
- Examining a random waffle-cone and judging each cone as either **defective** or **non-defective**, is the experiment in the above scenario.
- The outcome of the above experiment can be either **defective(D**) or **non-defective(N).**
- Since the exact outcome of the experiment (D or N) cannot be predicted, the above experiment is a **random experiment.**

Trial and Sample Space

- Individual repetitions of the same **random experiment** are termed as the **trial.** The set of all possible outcomes in a random experiment is called **Sample space.**
- Examining a waffle-cone is a trial within the experiment composed of examining 3 waffle cones.
- The sample space can be defined for the experiment using a tree diagram as shown below:



Event

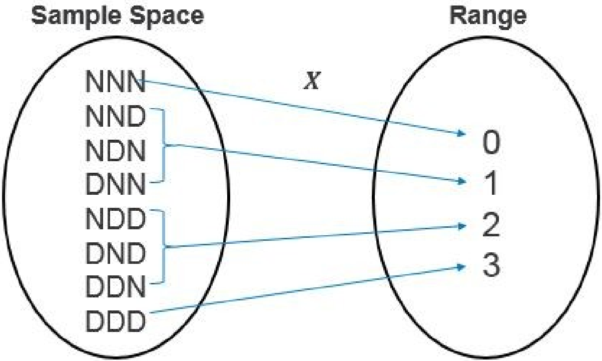
- One particular outcome or a set of some outcomes from the **entire sample space** is termed as an **event.**
- In the experiment, the event of interest can be composed of all the outcomes in which the total no. of defective cones is two.
- Then the event E is,

  E = {DDN, DND, NDD}
- Hence, there are 3 ways in which the above event can occur.
- In the experiment, the event of interest can be composed of all the outcomes in which the total no. of defective cones is one or more.
- Then the event E is,

  E = {DDD, DNN, NDN, NND, DDN, DND, NDD}
- Hence, there are 7 ways in which the above event can occur.

Random Variable

- A function **X** can be defined or explained on the **Sample Space** as a **relation** where each **independent sample space** outcome is mapped to a **numerical value**, based on the **event** of interest.
- **For example,** if we go through the total number of defective cones in 3 trials then, X can be depicted as follows:



X is called as Random Variable.

Random Variable

- A real-valued function, defined over a sample space is called a **random variable.**
- Only one real value is assigned by function to each individual outcome.
- **X or any other uppercase letter denotes a random variable.**
- We use a lowercase letter, for example, x to denote the **real value** that can be mapped with a **random variable map** and it's each outcome inside the sample space.
- X is not a variable like in Algebra. e.g. x+2 = 7, where x value is unknown.
- X is a function and can be depicted as follows:
- X = {x1, x2, x3, ...} or X = xk, where k = 1,2,3,...

- We can explain random variable X in the ice-cream scenario as follows:
- X = Total no. of defective cones in 3 trials = {0,1,2,3}
- So, few points to understand here are:
- We might know/be aware of the possible values of X i.e. 0,1,2,3 before we begin the experiment, we can't be sure what values will be taken at the experiment end.
- Also, X can assume different values each time the experiment is performed.
- Due to its trial to trial variability in value and non-predictable nature, **X is called a random variable.**

Probability Distribution

- Distributions we can understand how the random variable behaves. When the possibility of random variable values is associated with each of its probabilities, we get its Probability Distribution.
- The probability distribution is usually represented through either a table or a Graph**(usually a histogram).**
- Recall, for a Finite Sample Space S, then the **probability P(A),** is a real number assigned to the event A such that,

  0<=P(A)<= 1 and P(S)=1

Types of Probability Distributions

- Probability Distributions can be **Discrete or Continuous.**
- The associated probability distribution for a random variable with **discrete values** is called a **Discrete Probability Distribution**
- Discrete Probability Distributions are described by using the **Probability Mass Function (PMF).**
- The associated probability distribution for a random variable with **continuous (or approx. continuous)** values is called a **Continuous Probability Distribution**
- Continuous Probability Distributions are described by using the **Probability Density Function (PDF).**