IT3011 - Theory & Practices in Statistical Modelling

Prerequisites for Statistical Modelling

H.M. Samadhi Chathuranga Rathnayake

Set Theory

Introduction to Sets

- Simply put, a set in mathematics is a grouping of different items.
- Any group of objects, such as a collection of numbers, days of the week, different kinds of cars, etc., can be included in a set.
- Every item in the set is called an element of the set.
- Curly brackets are used while writing a set.
- Example-

$$A = \{1,2,3,4,5\}$$

- In mathematics, a set is a well-defined collection of objects.
- Sets are named and represented using a capital letter.

Elements of a Set

- Either elements or members of a set are terms used to describe the items that make up a set.
- A set's components are denoted by curly brackets enclosing them and commas separating them.
- To denote that an element is contained in a set, the symbol '∈' is used.
- In the above example, $2 \in A$.
- If an element is not a member of a set, then it is denoted using the symbol '∉'.
- Here, $7 \notin A$.

Cardinal Number

- The cardinal number, cardinality, or order of a set denotes the total number of elements in the set.
- For the above example, n(A) = 5.
- Sets are defined as a collection of unique elements.
- One requirement for defining a set is that all of its components must be connected to one another and possess a similar characteristic.
- We can state that all the members of a set are the months of the year, for instance, if the elements of a set are the names of the months in a year.

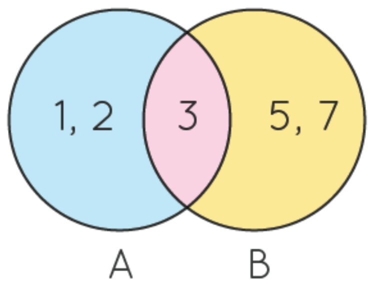
Representation of Sets

- There are different set notations used for the representation of sets.
- They differ in the way in which the elements are listed.

  - Semantic form
    - The semantic notation describes a statement to show what are the elements of a set. For example, Set A is the list of the first five odd numbers.
  - Roster form
    - The most common form used to represent sets is the roster notation in which the elements of the sets are enclosed in curly brackets separated by commas. For example, Set B = {2,4,6,8,10}
  - Set builder form
    - The set builder notation has a certain rule or a statement that specifically describes the common feature of all the elements of a set. For example, A = { k | k is an even number, $k \leq 20$}.

Visual Representation of Sets

- Each set is represented as a circle in a Venn Diagram, which is a visual representation of sets.
- Inside the circles are the components of a set.
- The universal set, which is represented by a rectangle sometimes encloses the circles.
- The relationship between the two sets is seen in the Venn diagram.



Set A = {1, 2, 3}

Set B = {3, 5, 7}

Elements of set A are 1, 2, 3

Element of set B are 3, 5, 7

Common element of set A and B is 3.

Sets Symbols

Symbols	Meaning
U	Universal set
n(X)	Cardinal number of set X
$b \in A$	'b' is an element of set A
$a \notin B$	'a' is not an element of set B
{}	Denotes a set
$\emptyset$	Null or empty set
$A \cup B$	Set A union set B
$A \cap B$	Set A intersection set B
$A \subseteq B$	Set A is a subset of set B
$B \supseteq A$	Set B is the superset of set A

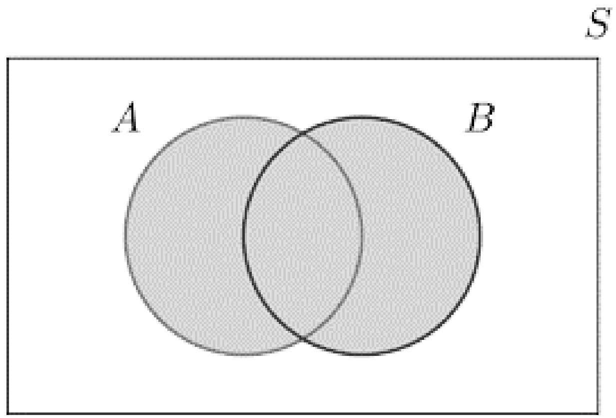
Set Operations

- Set operations are the operations that are applied on two or more sets to develop a relationship between them. There are four main kinds of set operations which are as follows.

  - Union of sets
  - Intersection of sets
  - Complement of a set
  - Difference between sets / Relative Complement

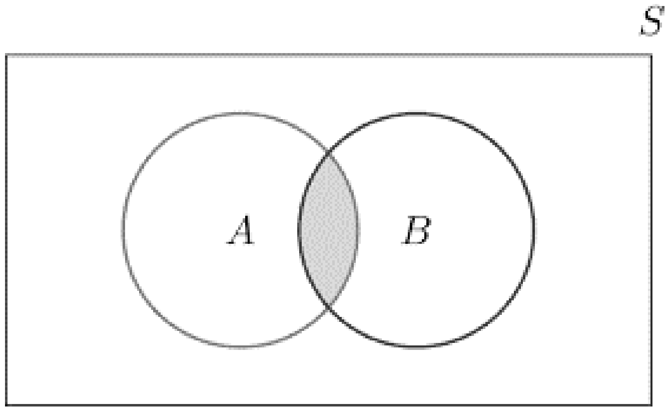
Union

- For two given sets A and B, A∪B (read as A union B) is the set of distinct elements that belong to set A and set B or both.



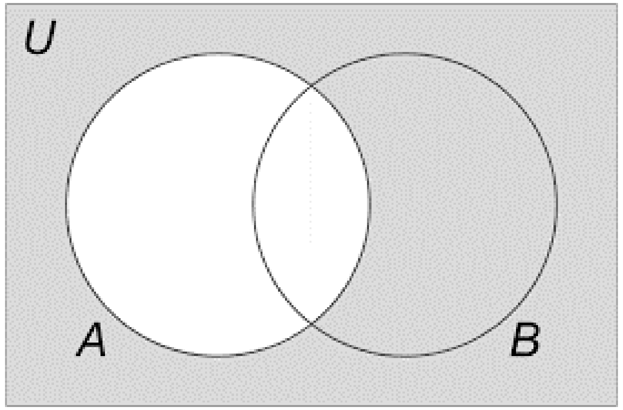
Intersection

- For two given sets A and B, A∩B (read as A intersection B) is the set of common elements that belong to set A and B.



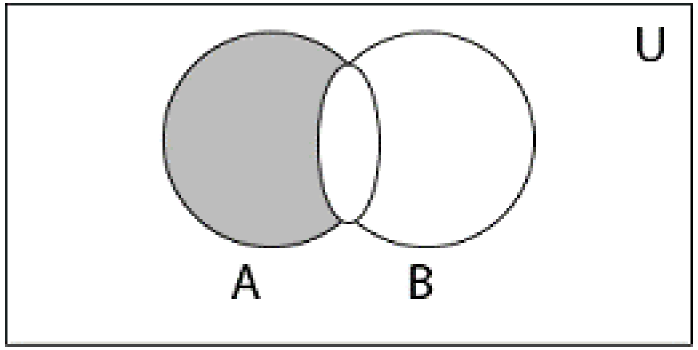
Complement

- The complement of a set A denoted as A′ or A^c (read as A complement) is defined as the set of all the elements in the given universal set(U) that are not present in set A.

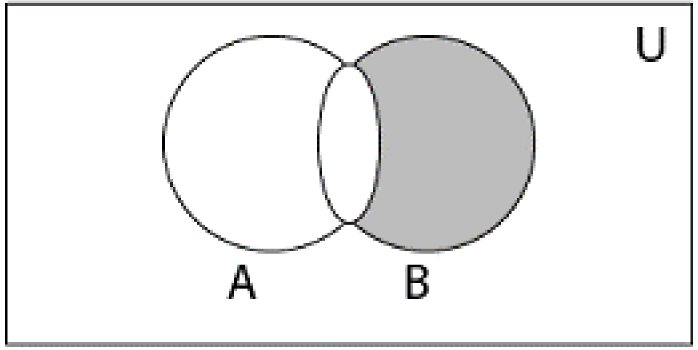


Set Difference

- The set operation difference between sets implies subtracting the elements from a set which is similar to the concept of the difference between numbers.
- The difference between sets A and set B denoted as A − B lists all the elements that are in set A but not in set B.



A - B



B - A

Properties of Sets

Property	Example
Commutative Property	$A \cup B = B \cup A$ $A \cap B = B \cap A$
Associative Property	$(A \cap B) \cap C = A \cap (B \cap C)$ $(A \cup B) \cup C = A \cup (B \cup C)$
Distributive Property	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Identity Property	$A \cup \emptyset = A$ $A \cap U = A$
Complement Property	$A \cup A' = U$
Idempotent Property	$A \cap A = A$ $A \cup A = A$

Properties of Sets

- De-Morgan's Law - The De Morgan's law states that for any two sets A and B, we have $(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$
- Further the following properties can also be seen.

$$A \cap \emptyset = \emptyset$$

$$A \cup \emptyset = A$$

$$A \cap B \subseteq A$$

$$A \subseteq A \cup B$$

Set Operations

Example 1: In a school, every student plays either football or soccer or both. It was found that 200 students played football, 150 students played soccer and 100 students played both. Find how many students were there in the school using the set operation formula.

Solution: Let us represent the number of students who played football as $n(F)$ and the number of students who played soccer as $n(S)$. We have $n(F) = 200$, $n(S) = 150$ and $n(F \cap S) = 100$. We know that,

$n(F \cup S) = n(F) + n(S) - n(F \cap S)$

Therefore, $n(F \cup S) = (200+150) - 100$

$n(F \cup S) = 350 - 100 = 250$

Answer: Hence the total number of students in the school is 250.

Example 2: If A = {a, b, c, d, e}, B = {a, e, i, o, u}, U = {a, b, c, d, e, f, g, h, i, j, k, l, o, u}. Perform the following operations on sets and find the solutions.

a) $A \cup B$

b) $A \cap B$

c) A'

d) $A - B$

Solution: a) $A \cup B$ = {a, b, c, d, e, i, o, u}

b) $A \cap B$ = {a, e}

c) A' = {f, g, h, i, j, k, l, o, u}

d) $A - B$ = {b, c, d}

Permutations & Combinations

Permutations

- All possible combinations by considering the order.
- There are basically two types of permutation:
  - **Repetition is Allowed**
  - **No Repetition**

Permutations with Repetition

- When a thing has n different types, we have n choices each time!

- For example: choosing 3 of those things, the permutations are:

$$n \times n \times n$$

- choosing r of something that has n different types, the permutations are:

$$n \times n \times \ldots (r\ times) = n^r$$

Permutations with Repetition

- Example: Assume there are 10 numbers to choose from (0,1,2,3,4,5,6,7,8,9) and we choose 3 of them:

$$10 \times 10 \times 10 = 10^3 = 1{,}000\, permutations$$

Permutations without Repetition

- When n is the number of things to choose from, and we choose r of them, no repetitions, order matters.

$$n_{P_r} = \frac{n!}{(n-r)!}$$

Permutations without Repetition

- Example: How many ways can first and second place be awarded to 10 people?

$$\frac{10!}{(10-2)!} = 90$$

Combinations

- All possible combinations without considering the order.
- There are basically two types of combinations:
  - **Repetition is Allowed**
  - **No Repetition**

Combinations without Repetition

- When n is the number of things to choose from, and we choose r of them, no repetitions, order doesn't matter.

$$n_{C_r}(Binomial\ Coefficient) = \frac{n!}{(n-r)!\,r!}$$

Combinations without Repetition

- Example : Choosing 3 balls out of 16, or choosing 13 balls out of 16, have the same number of combinations:

$$\frac{16!}{(16-3)!\,3!} = 560$$

Combinations with Repetition

- When n is the number of things to choose from, and we choose r of them repetition allowed, order doesn't matter.

$$r+n-1_{C_r} = \frac{(r+n-1)!}{(n-1)!\,r!}$$

Combinations with Repetition

- Example: Let us say there are five flavors of ice-cream: banana, chocolate, lemon, strawberry and vanilla. We can have three scoops. How many variations will there be?

$$\frac{(3+5-1)!}{(5-1)!\,3!} = 35$$

- There are 35 ways of having 3 scoops from five flavors of ice-cream.

Summary

	Repeats allowed	No Repeats
Permutations (order matters):	n^r	$\frac{n!}{(n-r)!}$
Combinations (order doesn't matter):	$\frac{(r+n-1)!}{r!(n-1)!}$	$\frac{n!}{r!(n-r)!}$

Probability Theory & Random Variables

Random Experiments

The term experiment refers to the process of obtaining results of some phenomenon. The performance of an experiment is called as **a trial** and the observed result of a trial of an experiment is called as **an outcome**.

Ex:-

Experiment – Tossing a coin

Trial – Tossing a coin for a single time

Outcome – Obtaining Head (H)

If any outcome of an experiment is likely to occur as any other outcome, then they are called **equally likely outcomes**.

Ex:- When tossing a coin, possible outcomes are; obtaining head and obtaining tail. As this is a fair coin, two outcomes equally likely.

Sample Space

The set of all possible outcomes, of an experiment is called the sample space. This is denoted by “S”. A sample space of an experiment can be discrete or continuous.

When the sample space is finite of countably infinite, then it is called a discrete sample space.

Ex:- An experiment for tossing a single coin

$$S = \{H, T\}$$

Sample Space

Ex:- An experiment consists of tossing two fair coins simultaneously

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Sample Space

Ex:- An experiment for rolling a fair die

$$S = \{1,2,3,4,5,6\}$$

Sample Space

Ex:- There are 5 red, 7 blue and 10 green balls in a bag. Three balls are taken randomly.

Sample Space (S):

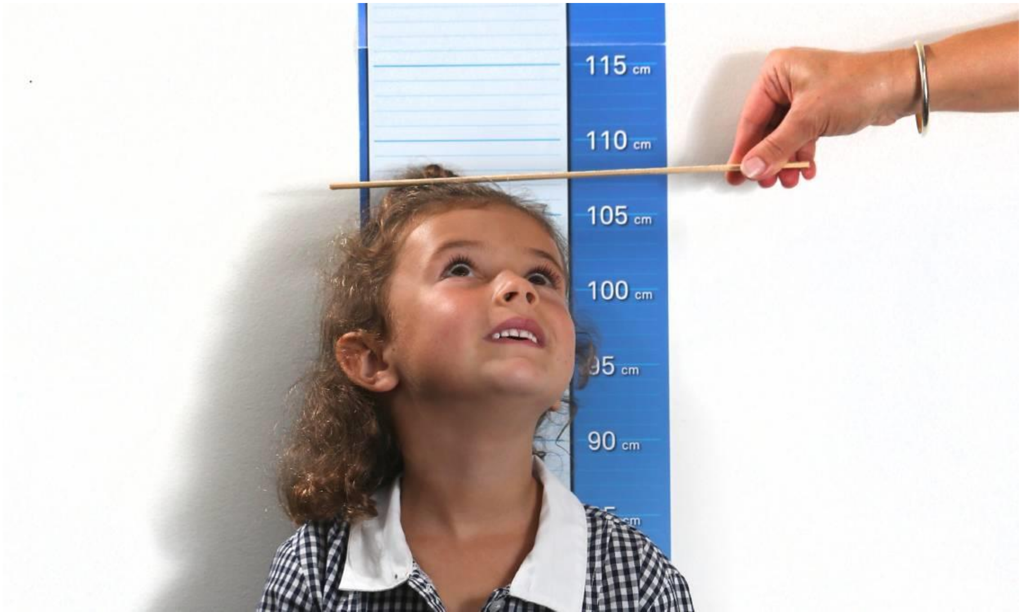
RRR	BBB	GGG
RRG	RRB	BBR
BBG	GGR	GGB
RBB	RGG	BRR
BGG	GRR	GBB
RBR	RGR	BRB
BGB	GRG	GBG
RBG	RGB	BRG
BGR	GRB	GBR

Sample Space

If an experiment involves outcomes, assuming any value in a given interval of real numbers, it is then called a continuous sample space.

Ex:- An experiment for finding the height of teenagers in Sri Lanka (In cm). Assume that the minimum height obtained is 100 cm and the maximum height obtained is 190 cm in the study.

$$S = \{100 \leq x \leq 190\}$$



Events

An event is a subset of the sample space (S). Consider the tossing two coins. The subset,

$$A = \{(H, H), (H, T), (T, H)\}$$

shows the event of obtaining at least one head.

Probability

Chance of happening any event in an experiment.

Ex:-

In the previous example,

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{4} = 0.75$$

Probability

Probability concept is defined in following three ways

- Classical approach
    - If an event **A** occurs in **m** times from an experiment which has **n** equally likely outcomes, then the probability of the occurrence of event **A** is **m/n**.
    - Consider the previous example.
- Relative frequency approach
    - Assume that an event **A** has occurred $\mathbf{n_A}$ times from an experiment which is repeated **n** time under the same condition.
    - Then the probability of the occurrence of event **A** is approximated by $\mathbf{n_A/n}$

$$\boldsymbol{P(A) = \lim_{n\to\infty} \frac{n_A}{n}}$$

- Axiomatic approach - Kolmogorovs Axioms

Probability - Axiomatic approach

- A probability P is a rule (function) which assigns a positive number to each event (in the event space), and which satisfies the following axioms:

$$P(A) \geq 0$$

$$P(S) = 1 \; where\; S\; is\; the\; sample\; space$$

If $A_1, A_2, \ldots, A_n$ is a sequence of mutually exclusive events, then,

$$P(\cup A_i) = Probability\; of\; all\; events = \sum P(A_i)$$

Properties of Probability

- $If\ A \subseteq B, then\ P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$

Random Variables

A random variable is a function that associated a unique numerical value with every outcome of an experiment. Consider a random experiment with the sample space S. A random variable X(r) is a function that assigns a real number to each outcome (r) on S.

Usually a single letter X is used for X(r) to represent a random variable.

Random variables can be discrete or continuous.

Random Variables

Ex:- An experiment for tossing a single coin

$$S = \{H, T\}$$

X random variable can be defined as,

$$X = Number\ of\ heads\ obtained\ (0, 1)$$

Random Variables

Ex:- An experiment consists of tossing two fair coins simultaneously

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

X random variable can be defined as,

$$X = Number\ of\ heads\ obtained\ (0, 1, 2)$$

Random Variables

Ex:- An experiment for rolling a fair die

$$S = \{1,2,3,4,5,6\}$$

X random variable can be defined as,

$$X = Number\ obtained\ on\ the\ top\ (1,2,3,4,5,6)$$

Random Variables

Ex:- Ex:- There are 5 red, 7 blue and 10 green balls in a bag. Three balls are taken randomly.

X random variable can be defined as,

$$X = Number\ green\ balls\ taken\ (0,1,2,3)$$

Random Variables

Ex:- An experiment for finding the height of teenagers in Sri Lanka (In cm). Assume that the minimum height obtained is 100 cm and the maximum height obtained is 190 cm in the study.

$$S = \{100 \leq x \leq 190\}$$

X random variable can be defined as,

$$X = A\ height\ of\ a\ teenager\ (Any\ value\ from\ 100\ to\ 190)$$

Probability Associated with Random Variables

Using the values of random variables, associated probabilities can be found.

Ex:- An experiment consists of tossing two fair coins simultaneously

$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

X random variable can be defined as,

$$X = Number\ of\ heads\ obtained\ (0,1,2)$$

$$P(x=1) = \frac{2}{4} = \frac{1}{2}$$

$$P(x \geq 1) = P(x=1) + P(x=2) = \frac{3}{4}$$

Probability Mass Function & Probability Density Function

Probability of a random variable with any value can be represented in a functional form.

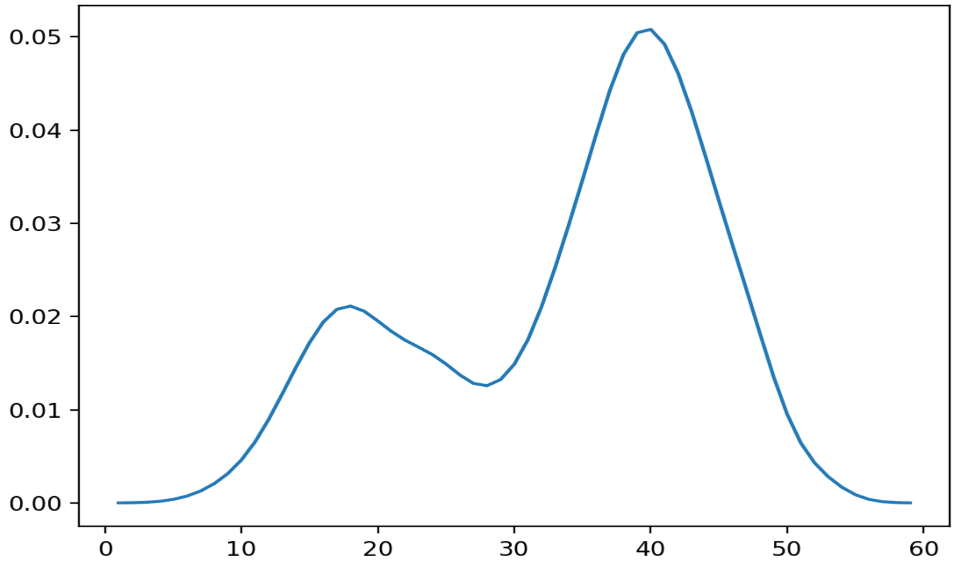
Discrete Random Variables
Probability Mass Function

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

Continuous Random Variables
Probability Density Function

Probability Distribution

Any function has a distribution or an arrangement with its inputs and outputs.



In a random variable, this arrangement is called as a **Probability Distribution**. With the nature of the random variable, the distribution can be a Discrete Probability Distribution or a Continuous Probability Distribution.

Probability Distribution

- When the random variable is a discrete random variable, the distribution is a discrete probability distribution. There are several discrete distributions with the nature of the random variable.
- When the random variable is a continuous random variable, the distribution is a continuous probability distribution. There are several continuous distributions with the nature of the random variable.

Moments of Discrete Probability Distribution

- Let X be a random variable. The mean of X, denoted by μ_x or E(X), is defined by,

$$E(X) = \sum_x P_X(x) \ .X$$

- Let X be a random variable and let μ_x be E(X). The variance of X, denoted by σ_x^2 or V(X) is defined by,

$$V(X) = \sum_x (x - \mu_X)^2 P_X(x)$$

Moments of Continuous Probability Distribution

- Let X be a random variable. The mean of X, denoted by μ_x or E(X), is defined by,

$$E(X) = \int_{-\infty}^{\infty} x.f_X(x)\,dx$$

- Let X be a random variable and let μ_x be E(X). The variance of X, denoted by σ_x^2 or V(X) is defined by,

$$V(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2.f_X(x)\,dx$$