Advanced analytics complete notes.pdf

Page no. 26

**Outliers :**

Outlier is an observation which deviates markedly from other members of same sample

Reasons of outliers?

▪ Human errors, e.g. data entry errors

▪ Instrument errors, e.g. measurement errors

▪ Data processing errors, e.g. data manipulation

▪ Sampling errors, e.g. extracting data from wrong sources

▪ Not an error, the value is extreme, just a ‘novelty’ in the data

A dilemma

▪ Outliers can be genuine values

▪ The trade-off is between the loss of accuracy if we throw away “good” observations, and the bias of our estimates if we keep “bad” ones

How to Identify them:

**Visualization techniques**: Visualization techniques such as box plots and scatter plots can also be

used to identify outliers by visually identifying any points that fall outside of the general pattern of the

data.

**Z-score method**: This method calculates the standard deviation and mean of the data, and any

observation that falls more than 3 standard deviations away from the mean is considered an outlier.

X bar - 3 \*sigma X bar + 3 \* sigma

|--------------------------- X bar (mean) ---------------------------|

**Interquartile range method**: This method calculates the interquartile range (IQR) of the data, and

any observation that falls outside of the lower and upper limits of the box plot, which are defined as

Q1–1.5 \* IQR and Q3 + 1.5 \* IQR, respectively, is considered an outlier.

**Clustering methods**: Clustering methods such as DBScan and KMeans can also be used to identify

outliers by grouping similar data points together and identifying any data points that do not belong to

any cluster.

How to handle them:

**Deleting the outlier observations**: This is a simple method, but it can lead to a loss of information if

the outliers are actually meaningful observations.

**Trimming the data**: This method involves removing a certain percentage of the largest and smallest

observations.

**Winsorizing**: This method replaces the outliers with a value that is closer to the center of the data.

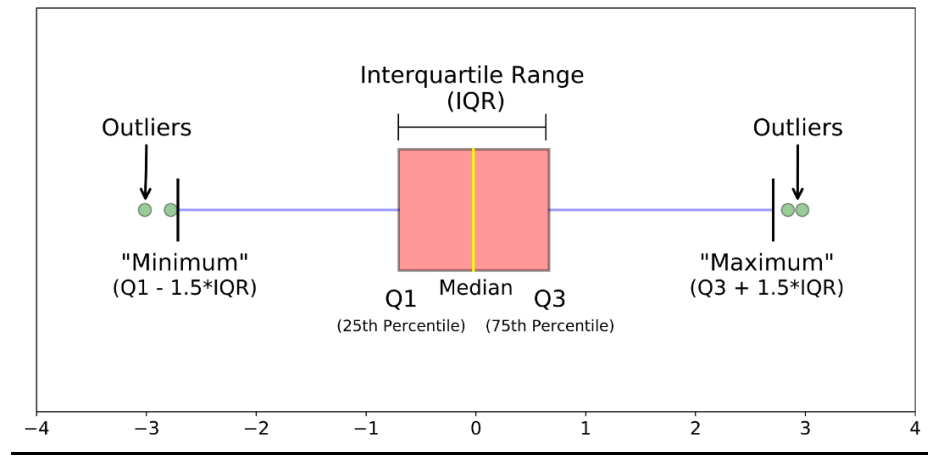
Either mean , median , mode

**Log transformation**: This method can be used when the data is positively skewed and the outliers are

on the high end of the distribution. Instead of x replace it with log(x), each value is replaced.

**Z-score standardization**: This method replaces each observation with its z-score, which is the

number of standard deviations away from the mean.

**Cap and floor**: This method replaces the outlier with a maximum and minimum value respectively

Page no. 5

**Factorial :**

We define, 0! = 1

For any positive integer n,

n! = n\*(n − 1)\*(n − 2)\*….\*1

e.g.

1! = 1

2! = 2\*1

3! = 3\*2\*1

4! = 4\*3\*2\*1 and so on.

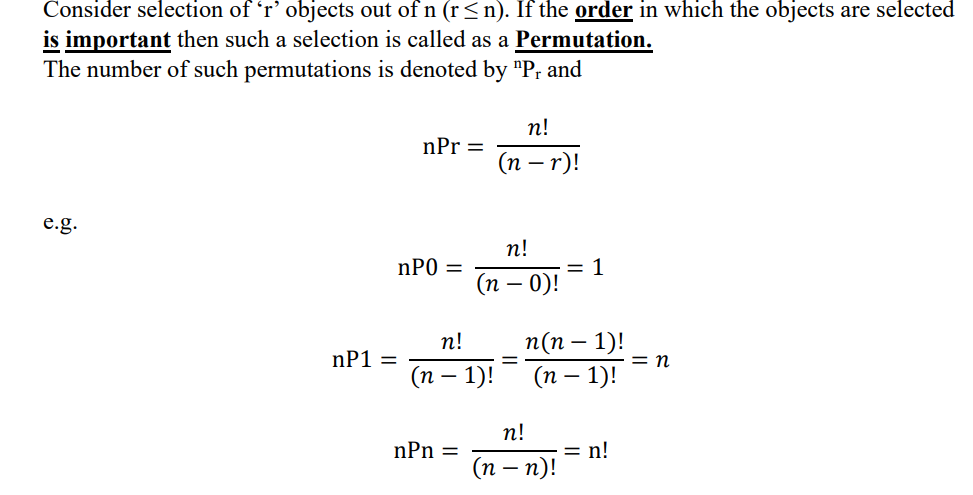
Consider 6! = 6\*5\*4\*3\*2\*1, which we can write as

6! = 6\*(5\*4\*3\*2\*1) = 6\*5! i.e. n! = n×[(n − 1)!]

6! = 6\*5\*(4\*3\*2\*1) = 6\*5\*4!

n! = n×(n − 1)×[(n − 2)!] and so on.

**Permutation :**

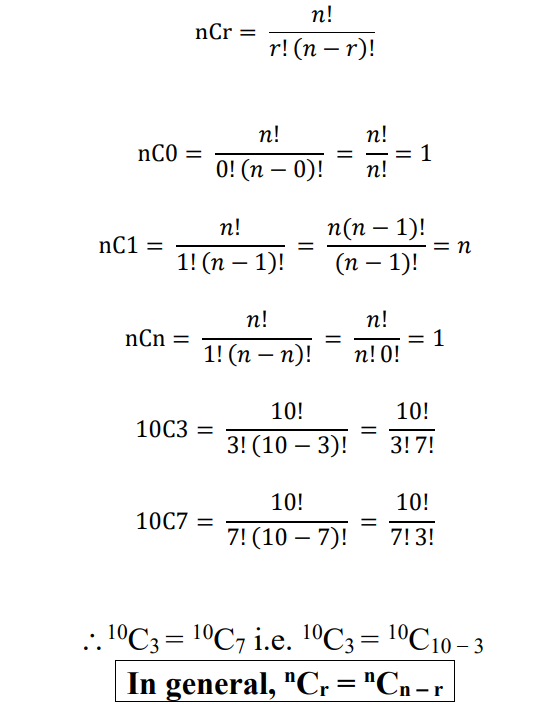


**Combination:**

Consider selection of ‘r’ objects out of n (r ≤ n). If the order in which the objects are selected

is not important then such a selection is called as a Combination.

The number of such combinations is denoted by nCr and



**Probability :**

**Basic Terms:**

**Random Experiment:** Exact outcome cannot be predicted but set of all possible outcomes is known.

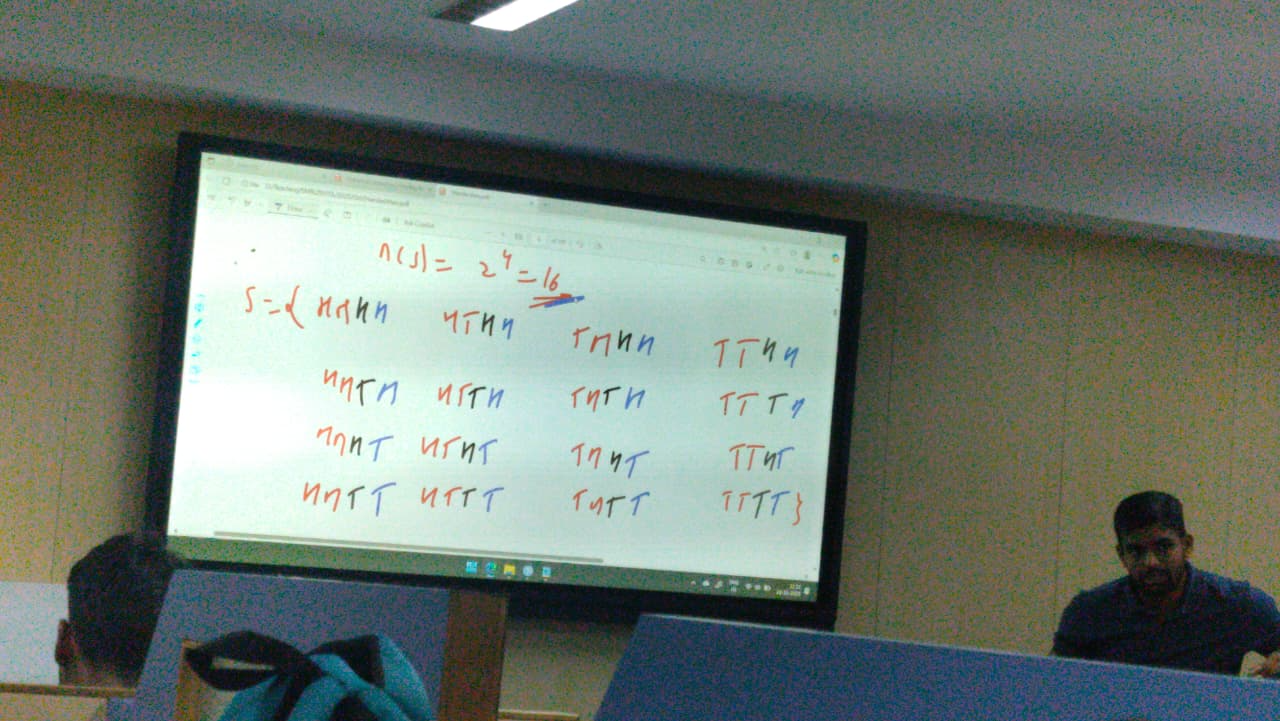
When tossed 1 coin , number of sample space , S

S = {H,T} ; n(S) = 2 ; 21 = 2

2 coins S = {HH,HT,TH,TT} ; n(S) = 4 ; 22 = 4

3 coins S = 23 = 8 , ; n(S) = 8

4 coins 🡪



1 Dice = 61 = 6 ; S = {1,2,3,4,5,6}

2 Dice = 62 = 36 ; S = { (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) }

**Event**:

Any subset of the sample space is called as an ‘Event’ and is denoted by any capital letter like

A, B, C or A1, A2, A3, ……

**Favourable cases**:

The cases which ensure the happening of an event A, are called as the cases favourable to the

event A. The number of cases favourable to event A is denoted by n(A).

Types of Events

**Elementary Event**: An event consisting of a single outcome is called an elementary event.

**Certain Event**: The sample space is called the certain event if all possible outcomes are

favourable outcomes. i.e. the event consists of the whole sample space. Here Probability is 1

**Impossible Event**: The empty set is called impossible event as no possible outcome is

Favorable

**Union of Two Events**

Let A and B be two events in the sample space S. The union of A and B is denoted by A∪B

and is the set of all possible outcomes that belong to at least one of A and B.

A∪B = combine

**Exhaustive Events (Exhaustive = Complete)**

Two events A and B in the sample space S are said to be exhaustive if A∪B = S

**Intersection of Two Events**

Let A and B be two events in the sample space S.

The intersection of A and B is the event consisting of outcomes that belong to both the events

A and B.

**Mutually Exclusive Events**Event A and B in the sample space S are said to be mutually exclusive if they have no

outcomes in common. (A intersection B = Phi ( Empty) ). In other words, the intersection of mutually exclusive events is empty. Mutually exclusive events are also called **disjoint events.**

**Complementary events.**If two events A and B are mutually exclusive and exhaustive.

Ex., Winning and losing is an only option, one team wins other loses.

**Equally Likely Cases:**

Cases are said to be equally likely if they all have the same chance of occurrence i.e. no case

is preferred to any other case.

**Probability Introduction :**

Chance is the occurrence of events in the absence of any obvious intention or cause. It is,

simply, the possibility of something happening. When the chance is defined in Mathematics,

it is called probability.

Probability is the extent to which an event is likely to occur, measured by the ratio of the

favourable cases to the whole number of cases possible.

Mathematically, the probability of an event occurring is equal to the ratio of a number of cases

favourable to a particular event to the number of all possible cases.

The theoretical probability of an event is denoted as P(E).

𝑃(𝐸) = Number of Outcomes Favourable to E / Number of all Possible Outcomes of the Experiment

**ODDS (Ratio of two complementary probabilities):**

Let n be number of distinct sample points in the sample space S. Out of n sample points, m

sample points are favourable for the occurrence of event A. Therefore remaining (n-m)

sample points are favourable for the occurrence of its complementary event A'.

