Unit - II

DATE: 13/10/23
PAGE:

Data Exploration:

It is the first in data analysis involving the use of data visualization tools and statistical techniques to uncover data set characteristics and initial patterns.

Why is data exploration important?

- Humans are visual learners, able to process visual data much more easily than numerical data. visually
- We can represent data with the help of data exploration.

Data profiling is a process of reviewing and cleaning data to better understand how its structured and maintain data quality standards beneath an organisation.

Fundamental of Statistics for data Analytics.
Statistics is a branch of mathematics, it deals with the analysis, interpretation, presentation of organisation of data.

Types of statistics: Statistics is divided into two category -

① Descriptive Statistics - provide information on
② ~~Infer~~ the distribution, central tendency and variability of data.
It includes mean, median, mode, variance, standard deviation, range.

② Inferential statistics: It is used to make prediction or draw conclusions about a population based on sample data. It includes hypothesis testing, parameter estimation and relationship b/w two variables.

Benefit of statistics for data analytics

① Summarise and describe data
② Identify pattern trend & relationship b/w data.
③ Make future prediction and find conclusion.
④ Hypothesis test.

Fundamental terms

i) Probability -
ii) Population & sample -
entire group data ex, it (2 out of 10)
iii) Distribution of data
how data is spread out

Normal Uniform Skewed

iv) Measure of central tendency
It describe the central value of data set by using mean, median and mode.

v) Variability It refers how data is spread out. find variability by using std deviation

vi) Central limit theorem.
conditional probability (event already occured)
vii) Co-variance & co-relation
viii) Hypothesis testing
ix viii) Student test, T-test and F
x) Type 1 error & Type 2 error

① find mean, median and mode

Number	7	8	9	10	11	12	13
frequency	3	8	12	15	14	17	9

Mean = 10.48 12.51

$\frac{n+1}{2} = \frac{8}{2}$ Median = 10

Mode = 12

CF (Cumulative freq) = 78 (even)

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th}\text{term} + \left(\frac{n}{2}+1\right)^{th}\text{term}}{2}$$

$$= \frac{11+11}{2} = \frac{22}{2} = 11$$

$\frac{7}{2}$

wt	46	48	50	52	53	54	55
no of people	7	5	8	12	10	2	1

Median

CF = 45 (odd)

Median = $\left(\frac{n+1}{2}\right)^{th}$ term

= 23 th term

Median = $l + \left(\frac{\frac{N}{2} - cf}{f}\right) h$ (l = previous class, h = missing class)

Median = 52

Mean = $\frac{(46\times7 + \dots)}{45}$ = 50.64

Mode = 52

Variance {3, 5, 8, 1} 24.625

Mean = $\frac{17}{4}$ = 4.25

Variance = [Each value − (mean)]²

= $\frac{(3-4.25)^2 + (5-4.25)^2 + (8-4.25)^2 + (1-4.25)^2}{4}$

= $\frac{1.5625 + 0.5625 + 14.0625 + 10.5625}{4}$

= 6.6875

DATE :
PAGE :

σ^2 = Variance = $\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$ (Sample)

σ^2 = Variance = $\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$ (Population)

σ (Standard deviation) = 2.586

Co-variance =

$cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ (Sample)

$cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$ (Population)

Corelation = $\frac{cov(x,y)}{\sigma_x \sigma_y}$

Cor $(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$

find mean, variance, covariance and correlation

Scores	freq	Midpoint (x)
40-45	3	42.5
34-39	1	36.5
28-33	4	30.5
22-27	5	24.5
16-21	2	18.5
10-15	5	12.5

$N = 20$

Mean $= \frac{\sum fx}{N} = 3.66$

$= \frac{508}{20} = 25.4$

$= 3.67$

Median $=$ $cf = 20$

$\frac{(n+1)^{th} \text{ term}}{2}$

$\left(\frac{n}{2}\right)^{th} \text{ term} + \left(\frac{n}{2}+1\right)^{th} \text{ term}$ / 2

$10^{th} \text{ term} + 11^{th} \text{ term}$ / 2

$= \frac{24.5 + 24.5}{2}$

$= \frac{490}{2} = 24.5$

Mode = highest freq $= 24.5, 12.5$

$\bar{x} = 7$

$\bar{y} = 27.4$

DATE :
PAGE :

Variance $= \frac{[\text{Each value} - \text{mean}]^2}{n-1}$ (sample data)

$= \frac{1}{4}\left[3(292.41) + (123.21) + (26.01) + (0.81) + (47.61) + 5(166.41)\right]$

$= 107.15$

877.23 + 123.21 + 104.04 + 4.05 + 95.22 + 832.05

X	Y	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})(y-\bar{y})$
2	38	-5	10.6	-53
4	36	-3	8.6	-25.8
8	24	1	-3.4	-3.4
9	17	2	-10.4	-20.8
12	22	5	-5.4	-27

Sum =

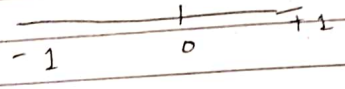
$\text{Cov}(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$ ← mean

$= \frac{130}{5} = 26$

Correlation means relation b/w two variables
It describe two things by two ways -
i) Direction of relation (↑↓ ↑↑ ↓↑ ↓↓)
ii) Degree of relation (means strong or weak relatn)

(↑↓)
+ −
negative correlation

correlation coefficient (r)
correlation values lies b/w −1 to 1. except 0



Type of correlation
There are two type of correlation
(+ve correlation) * Age premium
↑↑ ↓↓
+ +

(−ve correlation) * price & demand (onion)
↑↓ ↓↑
+ −

*

Subtype of correlation coefficient (r)
perfect +ve correlation value either −1 or 1
No correlation (If the value if r is 0)
strong +ve correlation (lie b/w 0.8 to 1)

Moderate +ve correlation (lie b/w 0.5 to 0.8)
low +ve correlation (r lie b/w 0.0 to 0.5)
just above 0
− ve
Strong −ve correlation (−0.8 to −1)

Imp.

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \times \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

ρ(x,y)

Karl Pearson Correlation coefficient formula

X	Y
9	15
0	16
7	14
6	13
5	11
4	12
3	10
2	8
1	9

r = ?

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \times \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

= 0.85
= 0.9

$$corr(x,y) = \frac{cov(x,y)}{\sqrt{var(x)}\sqrt{var(y)}} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

DATE: 28/11/23
PAGE:

Basic Probability

Probability is the chance that uncertain event will occur.

0 to 1
- 0 → impossible event
- 1 → confirm 100% occurance of event. (Sure to occur) (certain event)

There are three approaches to accessing the probability:

1. Priori - based on prior knowledge or process
2. Imperical probability - based on experiment
3. Subjective probability - based on combination of ~~subjective~~ an individual's past experience (personal opinion & analysis of particular situation)

① $P = \frac{X}{T}$

X = No. of ways in which event occurs.
Total no. of event (Possible outcomes)

P = No. of ways in which the event occurs
Total no. of possible outcomes.

$P = \frac{X}{T}$

Q1 what is

DATE:
PAGE:

Q. when Randomly selecting a day from the year 2015 what is the probability the day is in January.

$P = \frac{31}{365}$

Q1

	opted ML	Not opted ML	Total
Male	84	145	209
Female	76	134	210
Total	160	279	439

find the probability of selecting a male opting ML from the population described in the following table.

$P = \frac{84}{439}$

Q. Events - There are three type of an event.
1. Simple event (Single characteristics)
2. Joint event (more than two characteristic)

Complement of event A - All events that are not part of event A.

Sample Space - collection of all possible events.

All six faces of dice is called sample Spa

	Jan	Not Jan	
Wed	4	48	52
Not wed	27	286	313
Total	31	334	365

$P(\text{wed}) = \frac{52}{365}$

$P(\text{Jan}) = \frac{31}{365}$

$P(\text{Jan and wed}) = \frac{4}{365}$

$P(\text{not Jan and not wed}) = \frac{281}{365}$

Mutually exclusive events - Event that cannot occur simultaneously.

Ex: Randomly choosing a day from a year, 2015.

Event A = day in jan
Event B = day in feb
Event A and B are mutually exclusive event.

Collectively exhaustive event:
One of the event must occur.
The set of events covers the entire sample space.
Randomly chose a day from year 2015.

DATE: 28/11/23
PAGE:

Event A = weekday
Event B = weekend.
Event C = January.
Event D = Spring

Event C^A, B, D are collectively exhaustive but not mutually exclusive.

Event A and B are collectively exhaustive and also mutually exclusive.

Weekday cannot be Jan or Spring.
collectively exhaustive.

Joint and Marginal probability.

$P(A \text{ and } B) = \frac{\text{No. of outcome satisfying A and B}}{\text{Total no of outcomes.}}$

Addition Rule

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ — No. of outcomes of A or B / Total no of outcomes

If A and B are mutually exclusive then
$P(A \text{ and } B) = 0$
$P(A \text{ or } B) = P(A) + P(B)$

P(Jan or Wed) = P(Jan) + P(wed) − P(A and B)

$= \frac{31}{365} + \frac{52}{365} - \frac{4}{365}$

$= \frac{83 - 4}{365}$

Conditional Probability

$P(A/B) = \frac{P(A \text{ and } B)}{P(B)}$

$P(B/A) = \frac{P(A \text{ and } B)}{P(A)}$

Q. In a used carlot 70% have AC and 40% have GPS. 20% have both. What is the Probability that car has a GPS Given that it has AC.

$P(GPS / AC) = \frac{20}{70} = 0.2$

Independance

If two events are independent then P(A) is not affected by P(B).

$P(A/B) = P(A)$

Multiplication rule of two event A and B

$P(A \text{ and } B) = P(A/B) \cdot P(B)$

If A and B are independent

$P(A \text{ and } B) = P(A) \cdot P(B)$