Introduction to Statistical Methods

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Course No: DSECL ZC413 / AIML ZC418
Course Title: ISM
WEBINAR 1 : 06.06.2023

Topics - Webinar

- Descriptive Statistics
  - Measures of Central Tendency
  - Measures of Variability
- Probability
  - Probability – Introduction and Basics
  - Conditional probability

Measures of Central Tendency

A psychologist wrote a computer program to simulate the way a person responds to a standard IQ test. To test the program, he gave the computer 15 different forms of a popular IQ test and computed its IQ from each form

IQ Values:

134 136 137 138 138 143 144 144 145 146 146 146
147 148 153

Find the following Statistical measures:

i. Mean, median, and mode
ii. Rang, Variance and standard deviation
iii. Five point summary of the data.
iv. The interquartile range.
v. Identify potential outliers, if any.
vi. Construct and interpret a boxplot

Measures of Central Tendency

2 3 4 5 Average.

Arranging the data in ascending order:

M

134, 136, 137, 138, 138, 143, 144, 144, 145, 146, 146, 146, 147, 148, 153

N=15 N=

S.No	Formula	Solution
Mean	$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$	μ = 143
Median	$p = \frac{n+1}{2}^{th} = \frac{15+1}{2} = 8^{th}$ item	Median = 144
Mode	The mode is the value or values that occur most frequently in the data set. A data set can have more than one mode, and it can also have no mode	Mode = 146

x	f
134	1
136	1
⋮	
144	2
$\boxed{146 \quad 3}$	

Mode

144 3
146 3
148 3

147 3 / Unimodal

$\boxed{146, 147}$ Bimodal.

Multimoda

Mode, Bimodal, and Multimodal

A given set of data may have one or more than one Mode. A set of numbers with one Mode is unimodal, a set of numbers having two Modes is bimodal, a set of numbers having three Modes is trimodal, and any set of numbers having four or more than four Modes is known as multimodal.

A set of values having a single unique mode is said to be unimodal, one with two modes is called bimodal, and one with three modes is called trimodal.

Bimodal Mode – A set of data including two modes is identified as a bimodal model. This indicates that there are two data values that possess the highest frequencies. For example, the mode of data set B = { 8, 12, 12, 14, 15, 19, 17, 19} is 12 and 19 as both 12 and 19 are repeated twice in the given set.

No mode:
If no number in a set of numbers occurs more than once, that set has no mode: 3, 6, 9, 16, 27, 37, 48.

Population $\mu =$
Sample $\bar{x}$

Measures of Central Tendency

Range	$\text{Range} = x_n - x_1$	Minimum = 134 Maximum =153 Range R = 19
Variance	For a Population $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$ For a Sample $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	Variance = 26 x $(x-\mu)^2$ Σx $\mu =$
Standard deviation	For a Population $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$ For a Sample $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	Standard Deviation = 5.09901951

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{\sum_{i=1}^{N} (X^2 - 2\mu X + \mu^2)}{N}$$

$$= \frac{\sum X^2}{N} - \frac{2\mu \boxed{\sum X}}{N} + \frac{N\mu^2}{N}$$

$$= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2$$

$$\boxed{\sigma^2 = \frac{\sum X^2}{N} - \mu^2}$$

$\sum_{i=1}^{N} \mu^2$ cons

Σx_i^2

$= \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$

x | x^2

Σx Σx^2

Measures of Central Tendency

Min

134	136	137	138	138	143	144	144	145	146	146	146	147
148	153		Q1			M					Q3	
	Max.											

- **Minimum**.
- **Q1** (the first quartile, or the 25% mark). 1/4 Nth
- **Median**.
- **Q3** (the third quartile, or the 75% mark).
- **Maximum**.

Measures of Central Tendency

Quartiles	Quartiles separate a data set into four sections. The median is the second quartile Q_2. It divides the ordered data set into higher and lower halves. The first quartile, Q_1, is the median of the lower half not including Q_2. The third quartile, Q_3, is the median of the higher half not including Q_2.	Quartiles: Q_1 --> 138 Q_2 – Mead --> 144 Q_3 --> 146
Interquartile range	IQR=Q3−Q1	Interquartile Range IQR =8
Potential outliers, if any.	$\text{Upper Fence} = \overset{146}{Q_3} + \boxed{1.5 \times \overset{8}{IQR}} = 158$ $\text{Lower Fence} = Q_1 - \boxed{1.5 \times IQR} = 126$	none

Major outliers : $Q_3 + 3(IQR)$

Measures of Central Tendency – Box plot

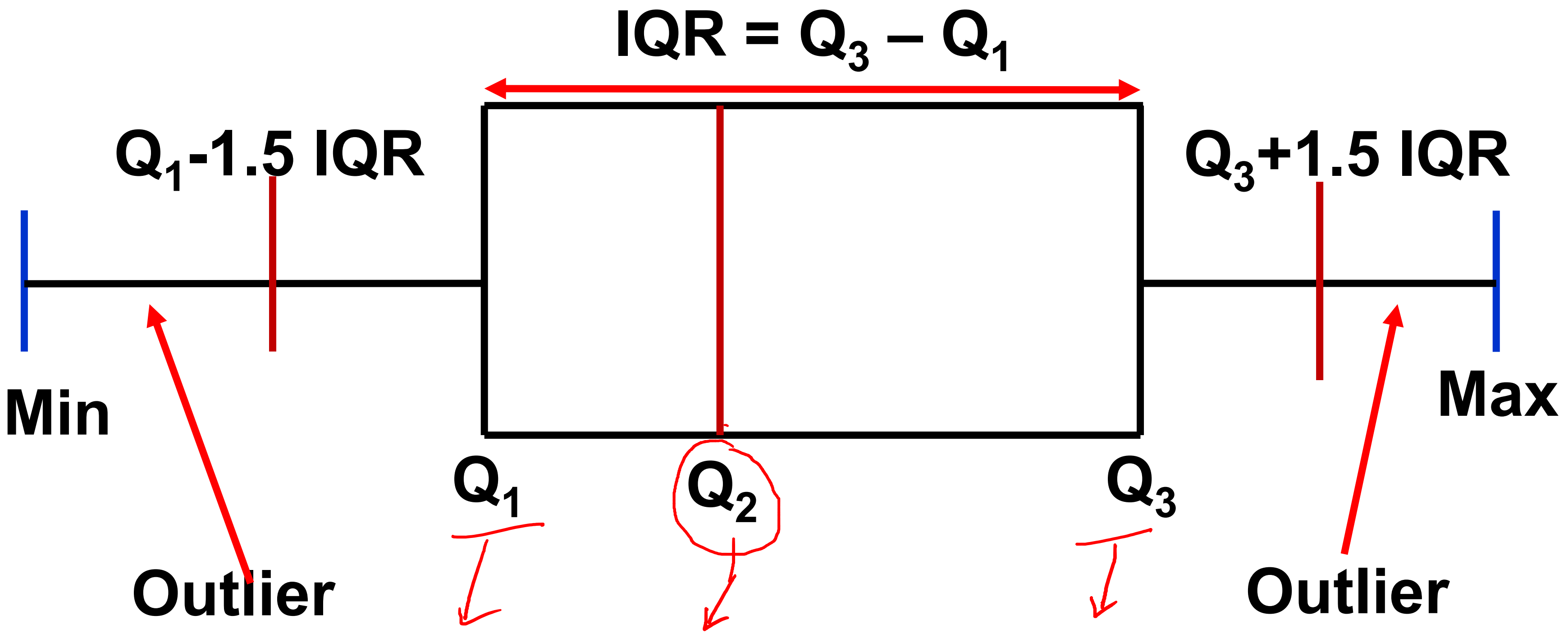
Population size: 15
Median: 144
Minimum: 134
Maximum: 153
First quartile: 138
Third quartile: 146
Interquartile Range: 8
Outliers: none

(126, 158)

M = middle $\frac{N+1}{2}$th.

$Q_1 = \frac{1}{4} (N+1)$

$Q_3 = \frac{3}{4}$



Average

Q1 , Median (Q2) and Q3

Example of Quartiles (n is odd)

Suppose the distribution of math scores in a class of 19 students in ascending order is:

- 59, 60, 65, 65, 68, 69, 70, 72, 75, **75**, 76, 77, 81, 82, 84, 87, 90, 95, 98

First, mark down the median, Q2, which in this case is the 10th value: 75.

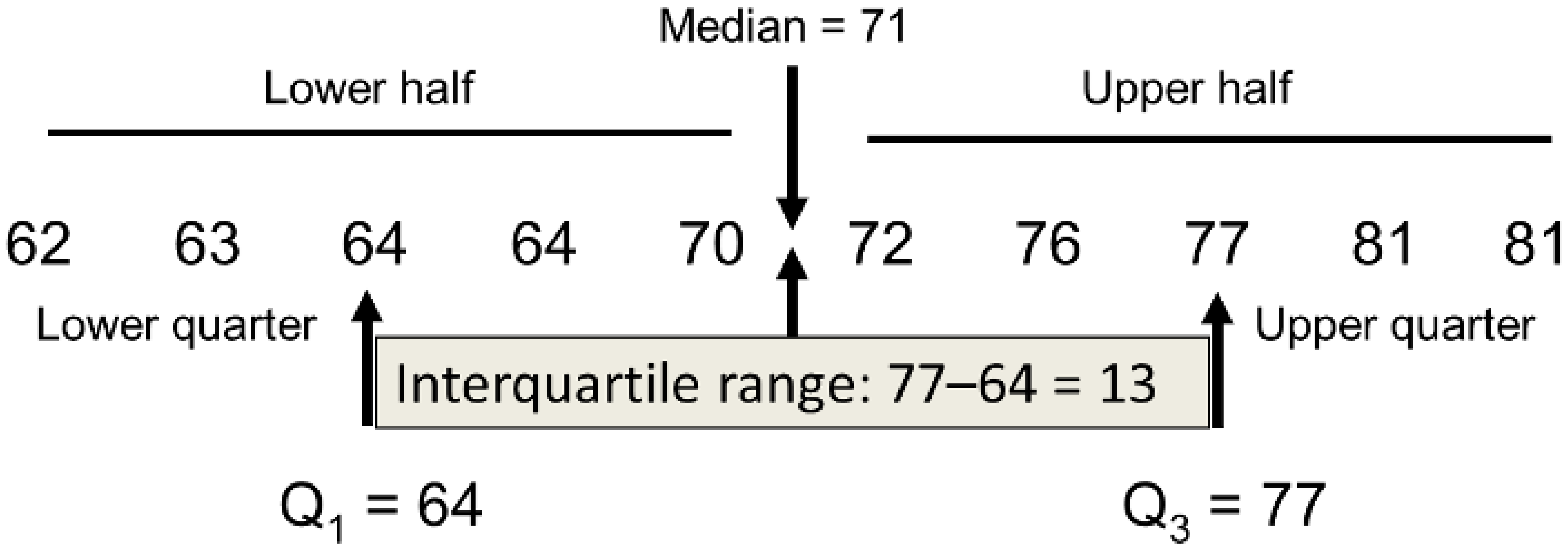
Q1 is the central point between the smallest score and the median. In this case, Q1 falls between the first and fifth score: 68. (Note that the median can also be included when calculating Q1 or Q3 for an odd set of values. If we were to include the median on either side of the middle point, then Q1 will be the middle value between the first and 10th score, which is the average of the fifth and sixth score—(fifth + sixth)/2 = (68 + 69)/2 = 68.5).

Q3 is the middle value between Q2 and the highest score: 84. (Or if you include the median, Q3 = (82 + 84)/2 = 83).

Now that we have our quartiles, let's interpret their numbers. A score of 68 (Q1) represents the first quartile and is the 25th percentile. 68 is the median of the lower half of the score set in the available data—that is, the median of the scores from 59 to 75.

Q1 tells us that 25% of the scores are less than 68 and 75% of the class scores are greater. Q2 (the median) is the 50th percentile and shows that 50% of the scores are less than 75, and 50% of the scores are above 75. Finally, Q3, the 75th percentile, reveals that 25% of the scores are greater and 75% are less than 84.

Q1 , Median (Q2) and Q3 (n is even)



Consider the following statistical summary of a dataset. Write at least three useful observations as a part of data pre – processing. **[Midsem Sep 2023]**

	Nr	Cells	QValue	Fat	Protein
count	969.000000	969.000000	969.000000	969.000000	969.000000
mean	7.074303	358.284830	90.016584	3.620279	3.300196
std	4.759793	344.324223	4.998924	0.349956	0.136071
min	1.000000	0.000000	62.650000	2.240000	2.750000
25%	3.000000	157.000000	87.570000	3.410000	3.220000
50%	6.000000	283.000000	90.750000	3.610000	3.310000
75%	10.000000	476.000000	93.400000	3.820000	3.380000
max	20.000000	5226.000000	100.000000	5.420000	3.840000

Q1

Q3

I. Direct Observation

- Mea

Range

- Coeff. Var = Sd/Mean → Protein
- Range
- $Q_3 - Q_1$ → Protein

Contin….

Observation 1: The variation in Protein is found very low when compared with remaining using coefficient of variation = sd/mean

Observation 2: The range in Protein is found very low when compared with remaining using range = max-Min

Observation 3: The middle range in Protein is found very low when compared with remaining using quartile range = Q3(75%)-Q1(25%)

Etc…………

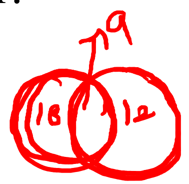
Basic Probability

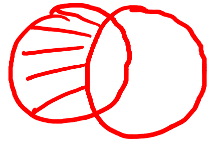
In a sample of 100 nails, 18 of them have ~~only~~ defective heads, 12 contain defective tail ends and 9 contain both defectives.

What is the probability that a nail selected at random is either a defective head or a defective tail?

A B $A \cap B$

$A \cup B$





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

$$= \frac{18}{100} + \frac{12}{100} - \frac{9}{100}$$

$$= 0.21$$

Basic Probability

$P(A-B) \neq P(A) - P(B)$ Not

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$ b). $P(A^c \cap B)$ c). $P(A \cap B^c)$ $\rightarrow P(A) - P(A \cap B)$

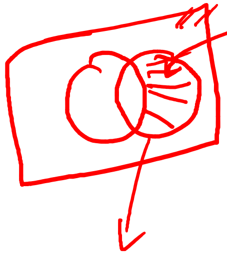
d). $P(A^c \cap B^c)$ e). $P(A^c \cup B^c)$ f). $P((A \cup B)^c)$

$P(A \cup B)^c$

$= 1 - P(A \cup B)$

$P(A^c \cap B) = P(B - (A \cap B)) = P(B) - P(A \cap B)$

$P(A^c)P(B)$



Basic Probability

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

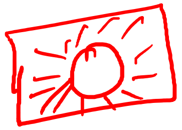
a). $P(A \cup B)$ b). $P(A^c \cap B)$ c). $P(A \cap B^c)$

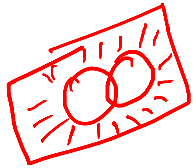
d). $P(A^c \cap B^c)$ e). $P(A^c \cup B^c)$ f). $P((A \cup B)^c)$

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$ b). $P(A^C \cap B)$ c). $P(A \cap B^C)$

d). $P(A^C \cap B^C)$ e). $P(A^C \cup B^C)$ f). $P((A \cup B)^C)$





Solution

a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/3 - 1/5 = 19/30 = 0.6333$

b) $P(A^C \cap B) = P(B - A) = P(B - (A \cap B)) = P(B) - P(A \cap B) = 1/3 - 1/5 = 2/15 = 0.1333$

c) $P(A \cap B^C) = P(A - B) = P(A - (A \cap B)) = P(A) - P(A \cap B) = 1/2 - 1/5 = 3/10 = 0.3$

d) $P(A^C \cap B^C) = P((A \cup B)^C) = 1 - P(A \cup B) = 1 - 19/30 = 11/30 = 0.3667$

e) $P(A^C \cup B^C) = P((A \cap B)^C) = 1 - P(A \cap B) = 1 - 1/5 = 4/5 = 0.8$

f) $P((A \cup B)^C) = 1 - P(A \cup B) = 1 - 19/30 = 11/30 = 0.3667$

Basic Probability

Consider the following data from the industry. Let D be the set of software developers in a company . Some subsets of D and some of probability that a random employee will belong to the set are represented below. Find the missing term in the table with justification.

Try to identify inconsistencies also (if exists and validate them). **[Midsem Jan 2023]**

Set in words	**Symbol**	**Probability**
All software developers	D	?
All software testing engineers	T	?
All software developers ~~and~~ / or software testing engineers	$D \cup T$	0.25
All software developers who are not software testing engineers	$D - T$	0.05
All software testing engineers who are not software developers	$T - D$	0.10
All software developers and software testing engineers	$T \cap D$	?

$D - (D \cap T)$

$D \cup T^c$

$P(D - T) =$

$P(D - (D \cap T)) = \cancel{P(D} \; 0.05$

$P(D) - P(D \cap T) = 0.05$

$P(T) - P(D \cap T) = 0.10$

$[P(D) + P(T) - P(D \cap T)] - P(D \cap T) = 0.15$

$0.25 - P(D \cap T) = 0.15$

$P(D \cap T) = 0.1$

$P(D) = 0.15$

$P(T) = 0.2$

Basic Probability

A political leader has submitted his nomination to compete in two different electoral constituencies namely A1 and A2. The probability of wining in constituency A1 and A2 is 0.80 and 0.65 respectively. The probability of losing at least one of the constituencies is 0.35. What will be the probability that he will win in one of the constituencies?

A1 = A, A2 = B

$P(A) = 0.8$

$P(B) = 0.65$

A^c, B^c

$P(A^c \cup B^c) = 0.35$

$P((A \cap B)^c) = 0.35$

$\boxed{P(A \cap B) = 0.65}$

$P(A^c \cup B^c) = (A \cap B)^c$

$A^c \cup B^c = (A \cap B)^c$

$A \cup B$

$P(A^c) = 1 - P(A)$

Required $= P(A \cup B) - P(A \cap B)$

$= 0.8 - 0.65$

$= 0.15$

Basic Probability

A political leader has submitted his nomination to compete in two different electoral constituencies namely A1 and A2. The probability of wining in constituency A1 and A2 is 0.80 and 0.65 respectively. The probability of losing at least one of the constituencies is 0.35. What will be the probability that he will win in one of the constituencies?

0.65 0

Assume that A, B be the events defined as follows: **A :** "Winning in constituency A1" **B :** "Winning in constituency A2" **Given:** **P(A) = 0.80, P(B) = 0.65** and $P(\bar{A} \cup \bar{B}) = 0.35$ **Now,** $\because P(\bar{A} \cup \bar{B}) = 0.35$ $\therefore P(\overline{A \cap B}) = 0.35$ $\Rightarrow 1 - P(A \cap B) = 0.35$ $\Rightarrow \boxed{P(A \cap B) = 0.65}$	**Then,** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ **= 0.80 + 0.65 – 0.65** $\therefore P(A \cup B) = 0.80$ Then, $P(\textit{He will win in one of the constituencies}) = P(A \cup B) - P(A \cap B)$ $= 0.80 - 0.65$ $\therefore P(\textit{He will win in one of the constituencies}) = 0.15$ $P(\textit{He will win in constituency A1 ONLY}) = P(A) - P(A \cap B)$ **= 0.80 – 0.65** **= 0.15** $P(\textit{He will win in constituency A2 ONLY}) = P(B) - P(A \cap B)$ **= 0.65 – 0.65** **= 0**

Independent events

Comment on the statement:

"If two events are mutually exclusive, then they are independent also and vice versa"

$P(A \cap B) = 0$

$P(A) \times P(B) = 0$

either $P(A) = 0$ or $P(B) = 0$

Two independent events cannot be mutually exclusive events - unless one or both events have a probability of zero (meaning one of the events is impossible).

Independent events

Let A and B be the two possible outcomes of an experiment and suppose $P(A) = 0.4$, $P(B) = p$ and $P(A \cup B) = 0.7$

(i) For what choice of ‘p’ are A and B mutually exclusive?

(ii) For what choice of ‘p’ are A and B independent?

Independent events

No common element

Let A and B be the two possible outcomes of an experiment and suppose P(A) = 0.4, P(B) = p and P(AUB) = 0.7

(i) For what choice of 'p' are A and B mutually exclusive?

(ii) For what choice of 'p' are A and B independent?

element

Prob

P(A)

(i) If A and B are mutually exclusive then $P(A \cap B) = 0$

Thus $P(A \cup B) = P(A) + P(B) - \boxed{P(A \cap B)}$ becomes

$0.7 = 0.4 + P(B) - 0$

$\therefore P(B) = 0.7 - 0.4 = 0.3$

(ii) If A and B are independent then $P(A \cap B) = P(A).P(B)$

Thus $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$$

$$0.7 = 0.4 + P - 0.4 \cdot P$$

$$= 0.4 + P(1 - 0.4)$$

$$= 0.4 + 0.6P$$

$$0.6P = 0.7 - 0.4 = 0.3$$

$$P = \frac{0.3}{0.6} = \frac{1}{2} = 0.5$$

$\therefore P(B) = 0.5$

Answer. (i) $\boxed{p=0.3}$, (ii) $\boxed{p=0.5}$

Dependent events – Conditional Probability

A, B

$P(B) > 0$

$P(A/B)$

$P(\text{Pass}/\text{Team}) = 1.00$

$P(\text{Pass}) = 0.75$

If P (A) =1/2, P (B) =1/3, P (A/B) = 1/6 find i). P(B/A) and ii). P(AUB / A)

$$P(A/B) = \frac{P(A\cap B)}{P(B)} \Rightarrow P(A\cap B) = P(A/B)\,P(B)$$

$$P(B\cap A) = \frac{1}{6} \times \frac{1}{3}$$

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

$$P(A\cup B/A) = \frac{P[(A\cup B)\cap A]}{P(A)} = \frac{P(A)}{P(A)} = 1$$

A & B depend.

- $P(A/B) = \dfrac{P(A\cap B)}{P(B)}$
- A & B independent

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

- $\boxed{P(A)}$

$P($

$P[\text{Pass}/\text{highest}] =$

$P(\text{pass})$

Conditional Probability

Consider the following data. **[Midsem Jan 2023]**

A: the event that a patient selected is with high B.P

B: the event that a patient selected is diabetic

C: the event that a patient selected is with cancer

Events	A	B	C	A and B	B and C	A and C	(A and B and C)
Probability	0.35	0.40	0.45	0.22	0.25	0.22	0.15

i) If a patient is selected at random, what is the probability that he suffering from at least one of these health issues?

$P(A \cup B \cup C) =$

i) What is the probability that the patient selected is with B.P given that he is with either of the remaining two diseases?

$P(A / B \cup C)$

iii) What is the probability that the patient is diabetic given that he is suffering from at least one of the diseases mentioned?

$P(B / A \cup B \cup C)$

$P(A/B)$ → given already occured

Conditional Probability

F

A

In an online shopping survey, 30% of persons made shopping in Flipkart, 45% of persons made shopping in Amazon and 5% made purchases in both. If a person is selected at random, find

i) the probability that he makes shopping in at least one of two companies $P(F \cup A)$

ii).the probability that he makes shopping in Amazon given that he already made shopping in Flipkart. $P(A/F)$

iii).the probability that the person will not make shopping in Flipkart given that he already made purchase in Amazon.

$P[F^c/A]$

Conditional Probability

In an online shopping survey, 30% of persons made shopping in Flipkart, 45% of persons made shopping in Amazon and 5% made purchases in both. If a person is selected at random, find

i) the probability that he makes shopping in at least one of two companies

3rd lev. 0.25

ii).the probability that he makes shopping in Amazon given that he already made shopping in Flipkart.

iii).the probability that the person will not make shopping in Flipkart given that he already made purchase in Amazon.

Solution: Given $P(F) = 30\% = 0.30$

$P(A) = 45\% = 0.45$

$P(F \cap A) = 5\% = 0.05$

i) $P(F \cup A) = P(F) + P(A) - P(F \cap A)$

$= 0.30 + 0.45 - 0.05 = 0.7$

ii) $P(A \mid F) = \frac{P(A \cap F)}{P(F)}$

$= \frac{0.05}{0.30} = 0.167$

iii) $P(F' \mid A) = \frac{P(F' \cap A)}{P(A)}$

$P(F' \cap A) = P(A) - P(A \cap F)$

$= 0.45 - 0.05$

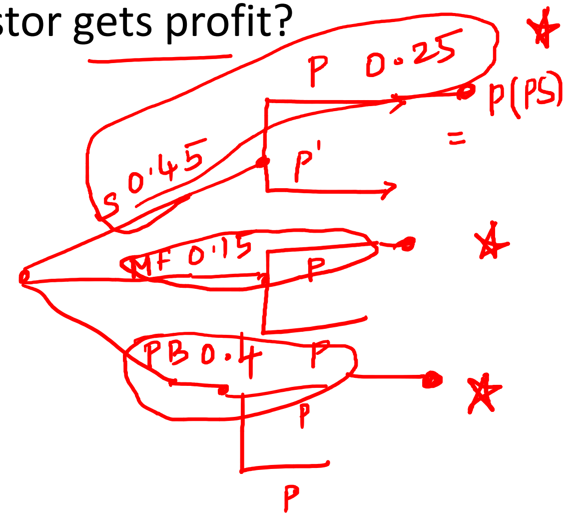
$= 0.40$

$P(F' \mid A) = \frac{0.40}{0.45} = 0.88$

Total Probability

$P(A \cup B) = P(A) + P(B)$

An investor made 45% of investment in stocks, 15% in mutual funds and the rest in his personal business. There are 25%, 10% and 20% of chances of obtaining profit in stocks, mutual funds and personal business respectively. What is the probability that the investor gets profit?



$P(\cancel{P}/S) = 0.25$

$P(P/MF) = 0.10$

$P(P/PB) = 0.2$

$P(S) = 0.45$

$P(MF) = 0.15$

$P(PB) = 0.4$

$P(\text{Profit})$

$= P(PS \text{ or } PMF \text{ or } PPB)$

$= P(PS)$

$P(PS) = P(S)\,P(P/S)$

An investor made 45% of investment in stocks, 15% in mutual funds and the rest in his personal business. There are 25%, 10% and 20% of chances of obtaining profit in stocks, mutual funds and personal business respectively.

a) What is the probability that the investor gets profit.

b) Given that the investor has got some profit, what are the probabilities that he got it from stocks, mutual funds and business.

Solution Q2:

Given $P(S) = 45\% = 0.45$, $P(M) = 15\% = 0.15$, $P(B) = 0.4$

$P(P/S) = 0.25$, $P(P/M) = 0.1$, $P(P/B) = 0.2$

a) $P(P) = P(S)P(P/s)+P(M)P(P/M)+P(B)P(P/B) = 0.2075$

b) $P(S/P) = P(S)P(P/s)/P(P) = 0.542169$

$P(M/P) = P(M)P(P/M)/P(P) =0.072289$

$P(B/P) = P(B)P(P/B)/P(P) = 0.385542$

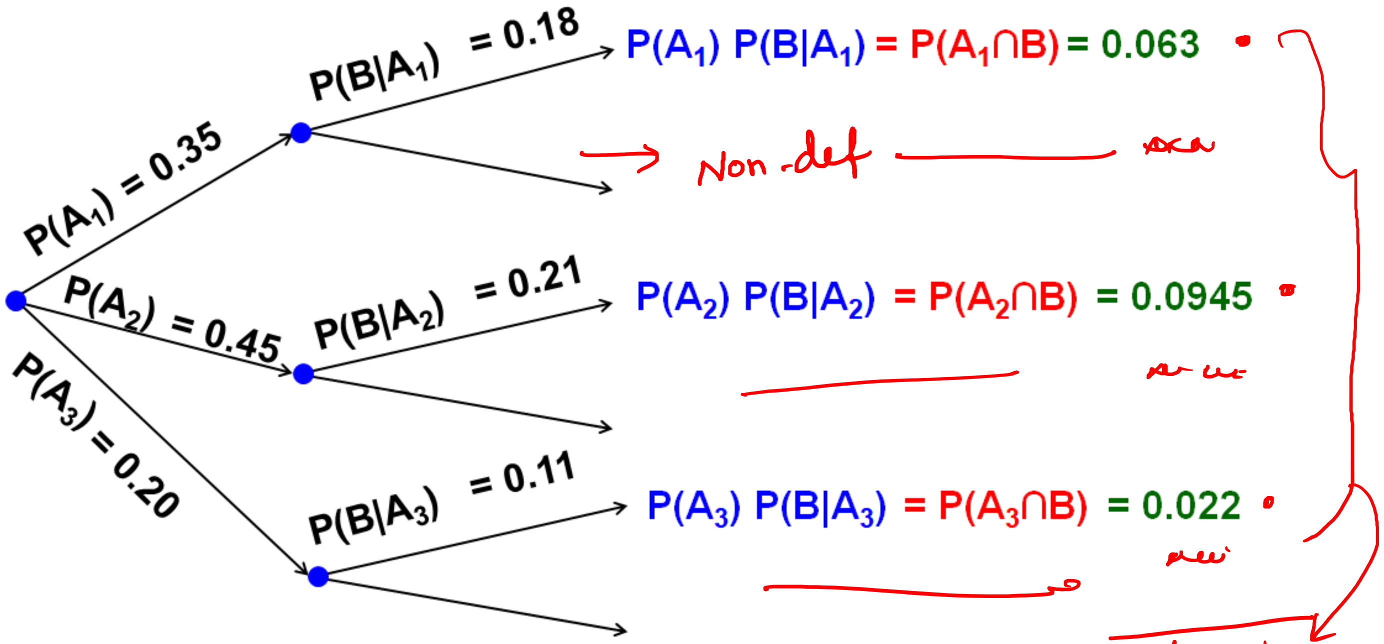
Total Probability

$P(X \cap Y) = P(X/Y)\,P(Y)$

$= P(Y/X)\,P(X)$

$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}$

Of 1000 car parts produced, it is known that 350 are produced in one plant, 450 parts in a second plant, and 200 parts in a third plant. Also, it is known that the probabilities are 0.18, 0.21, and 0.11 that the parts will be defective if they are produced in the first, second and third plants respectively. What is the probability that a randomly picked part from this batch is defective?



B

$P(B) = P(BA_1 \text{ or } BA_2 \text{ or } BA_3)$

$= P(BA_1) + \dots$

$= P(A_1) \times P(B/A_1)$

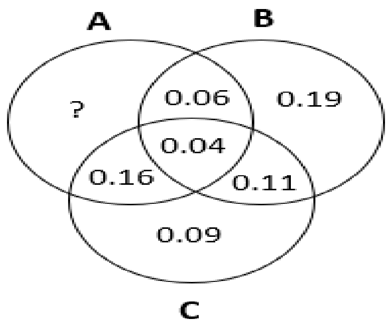
$=$

$P(B) = P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) + P(A_3)\,P(B|A_3) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_2 \cap B) = 0.1795$

An insurance company insured 1,000 taxi drivers, 2,000 car drivers and 3,000 truck drivers. The probability of their accident is 0.05, 0.15 and 0.1 respectively. One of the insured persons meets with an accident. What is the probability that he is a truck driver?

Exercise

1. If P (A) =1/2, P (B) =1/3, P (A/B) = 1/6 find i). P(B/A) ii). P(B/A') iii).P(AUB / A) iv).P(B/A) .
2. There are three events A, B and C. The probability of occurrence of at least one of them is 0.23. Using the probabilities given in the following Venn diagram, find the probability of the event A.



3. A manufacturing company produces certain types of output by 4 machines i.e A, B, C and D. Machine A produces 15%, Machine B produces 30 % and Machine C produces 30% of daily production. Based on experience it is observed that 1% of the output by Machine A is defective. Similarly, the defectives by other machines are 2% ,3% and 4% respectively. An item is drawn at random and found to be defective. Is it possible to find the defective item is produced bv which Machine? If so, find it.

Exercise

4. A manufacturer has three machine operators A, B and C. The first operator A produce 1% defective items, whereas the other two operators B and C produce 5% and 7% defective items respectively. A is on the job for 50% of the time, B is on the job for 30% of the time. A defective item is produced, what is the probability that it was produced by A, B, C? Based on this write your observations.

5. Consider the following data related to the employees, who are on travel. 40% check work email, 20% use cell phone to stay connected to work, 25% bring laptop with them, 23% check both work email and use cell phone to stay connected, and 50% neither check work email nor use a cell phone to stay connected nor bring a laptop. In addition, 88 out of every 100 who bring a laptop also check work email, and 70 out of every 100 who use a cell phone to stay connected also bring a laptop.

i) What is the probability that a randomly selected traveller who checks work email also uses a cell phone to stay connected?

ii) What is the probability that someone who brings a laptop on vacation also uses a cell phone to stay connected?

iii) If the randomly selected traveler checked work email and brought a laptop, what is the probability that he/she uses a cell phone to stay connected?

THANK YOU!