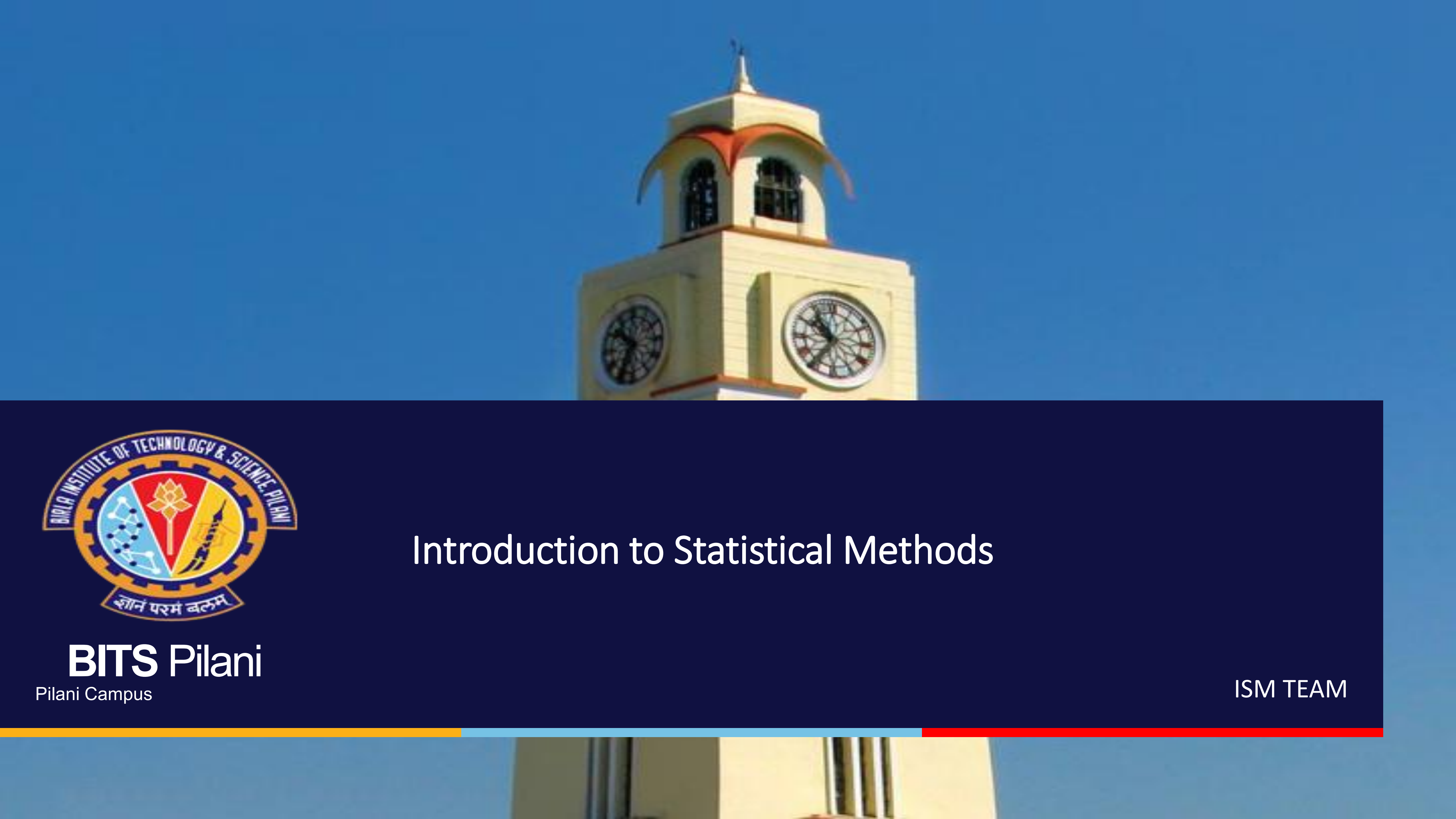


BITS Pilani
Pilani Campus

Introduction to Statistical Methods

ISM TEAM

BITS Pilani

Pilani Campus

Course No: AIML ZC418

Webinar 1 : 11.11.2025

Team ISM

Topics - Webinar

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

Measures of Central Tendency

1) 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. Range, Variance and standard deviation

iii. The interquartile range.

v. Identify potential outliers, if any.

vi. Construct and interpret a boxplot

Measures of Central Tendency

Arranging the data in ascending order:

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

N=15

Measure	Formula	Solution
Mean	$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$	μ = 143
Median	$p = \frac{n+1}{2}$	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

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.

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:

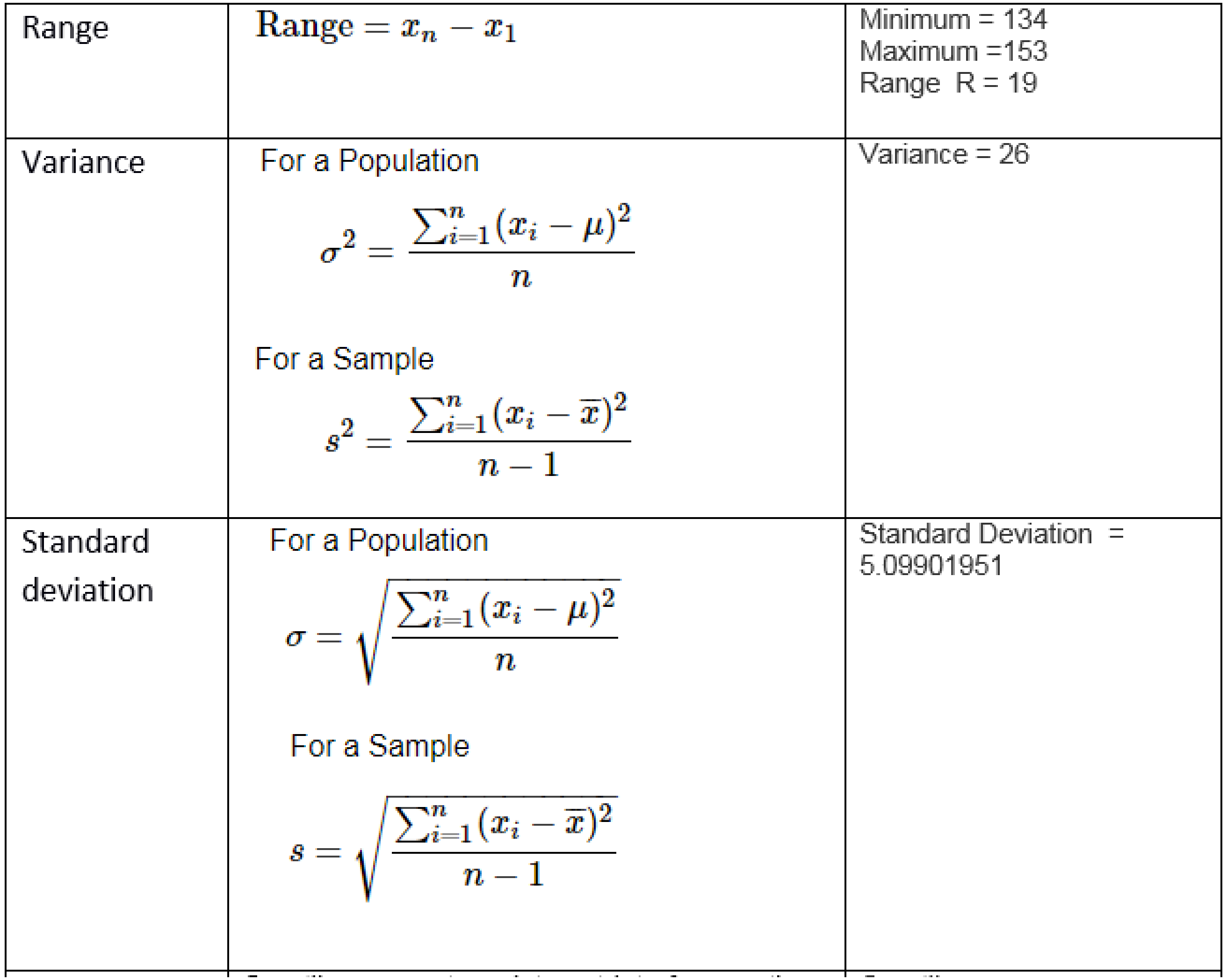
If no number in a set of numbers occurs more than once, that set has no mode: 3, 6, 9, 16, 27, 37, 48.

A unimodal mode is a set of data with only one mode.

A bimodal mode is a set of data that has two modes.

A trimodal mode is a set of data that has three modes.

Measures of Variability



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 $s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$	Variance = 26
Standard deviation	For a Population $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$ For a Sample $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$	Standard Deviation = 5.09901951

$$
\begin{aligned}
\sigma^2 &= \frac{\sum(X-\mu)^2}{N} \\
&= \frac{\sum(X^2-2\mu X+\mu^2)}{N} \\
&= \frac{\sum X^2}{N} - \frac{2\mu\sum X}{N} + \frac{N\mu^2}{N} \\
&= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2 \\
&= \frac{\sum X^2}{N} - \mu^2
\end{aligned}
$$

Measures of Central Tendency

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

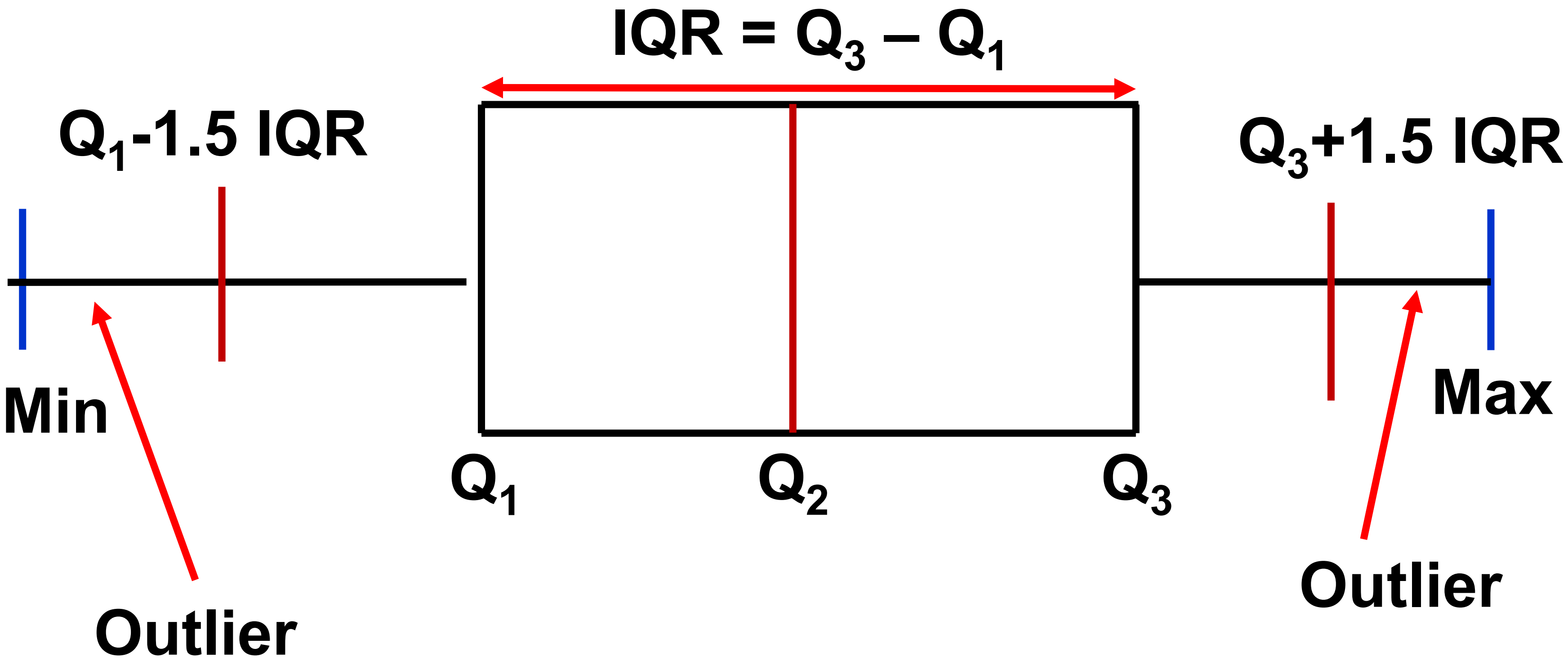
148 153

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

Measures of Central Tendency

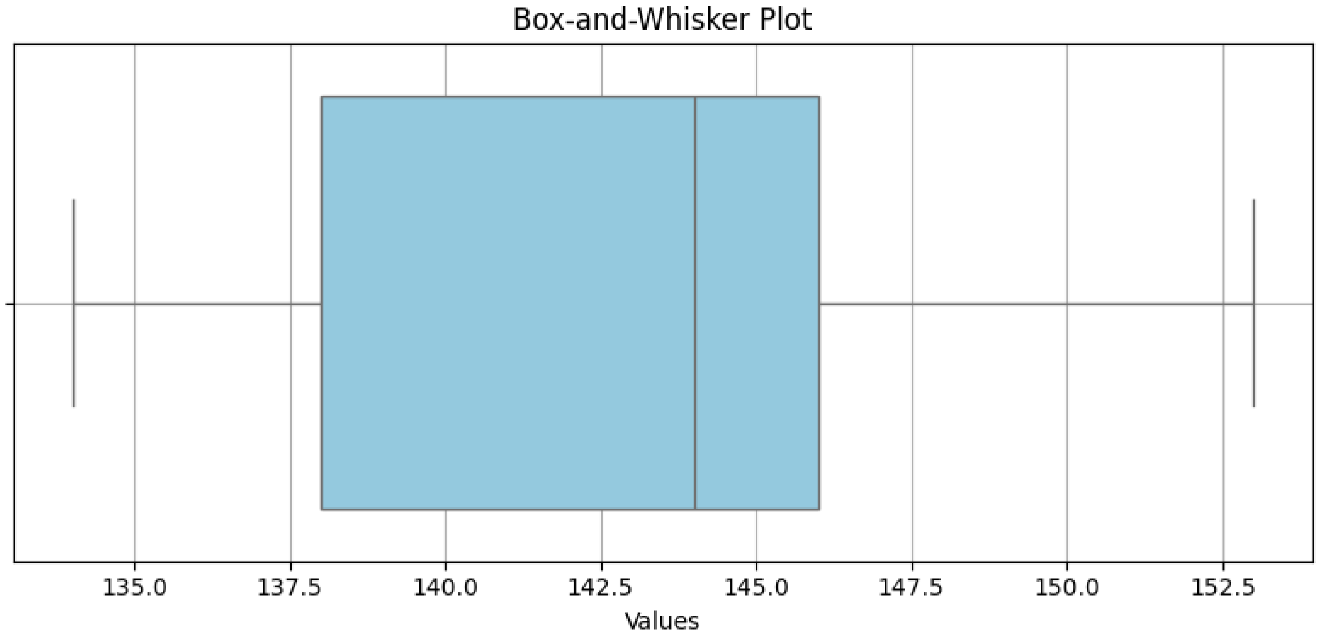
Quartiles	Quartiles separate a data set into four sections. The median is the second quartile Q_2. It divides the ordered data set into higher and lower halves. The first quartile, Q_1, is the median of the lower half not including Q_2. The third quartile, Q_3, is the median of the higher half not including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

Graphical Representation:Box plot

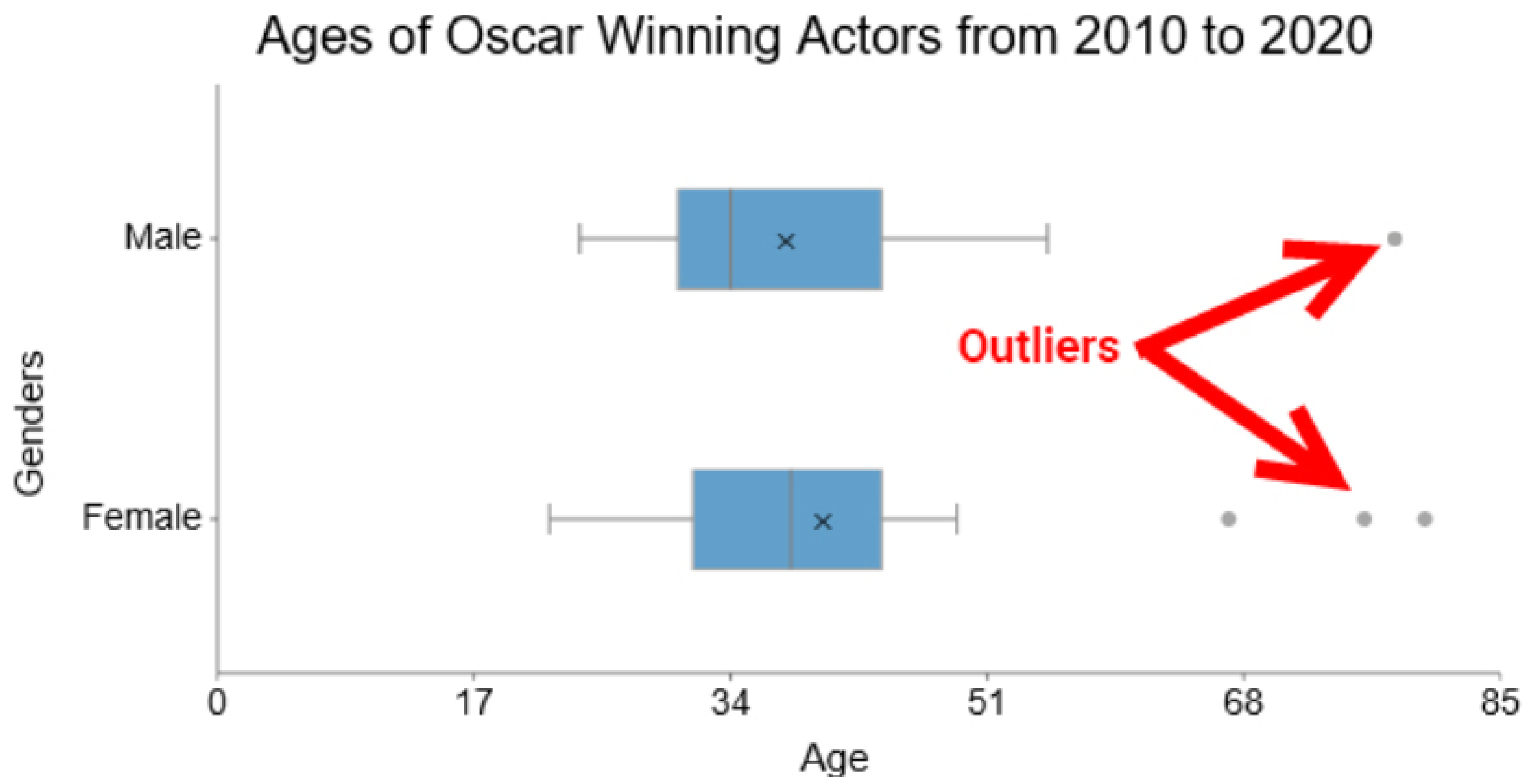


Box-plot for the given data:

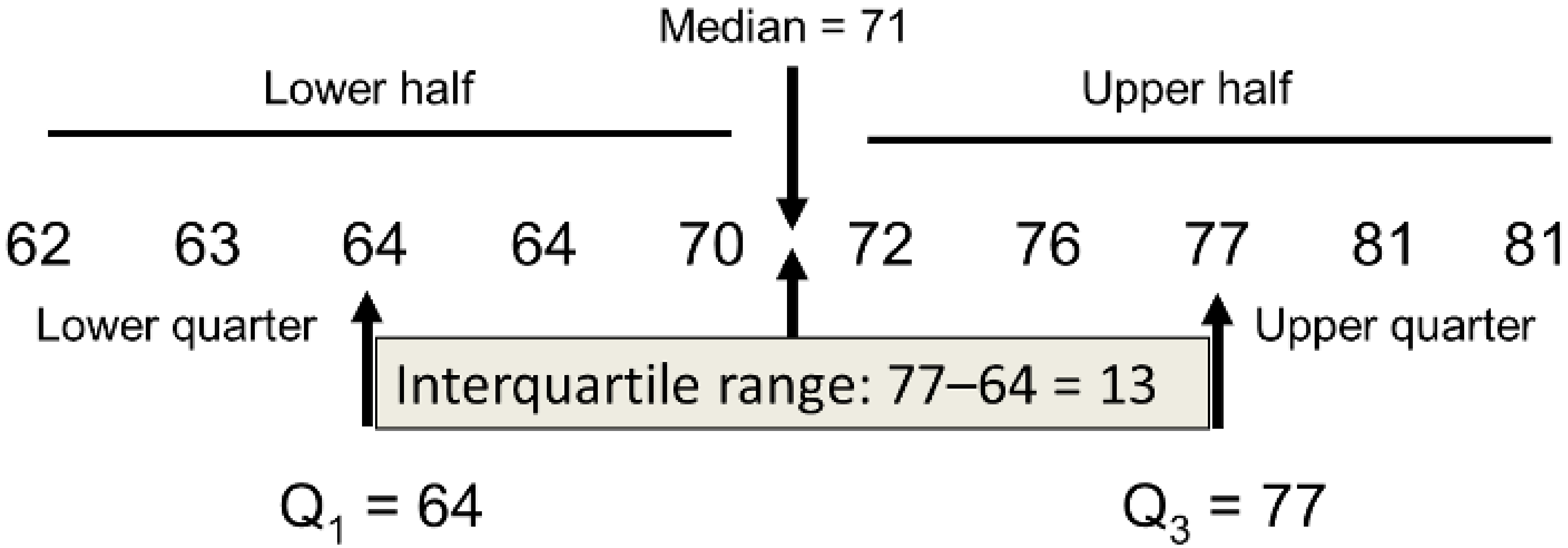
Population size: 15
Median: 144
Minimum: 134
Maximum: 153
First quartile: 138
Third quartile: 146
Interquartile Range: 8
Outliers: none



Real-time Example:



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



Example:

Data: 5, 7, 8, 9, 10, 12, 13, 15, 18, 30

Five- point summary:

- **Q1** = 8
- **Q2 (Median)** = 11
- **Q3** = 15
- **IQR** = 15 - 8 = 7
- **Lower bound** = 8 - (1.5 × 7) = -2.5
- **Upper bound** = 15 + (1.5 × 7) = 25.5

✅ Any value **> 25.5** or **< -2.5** is an **outlier** → here, **30** is an outlier.

2)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
```

Contd….

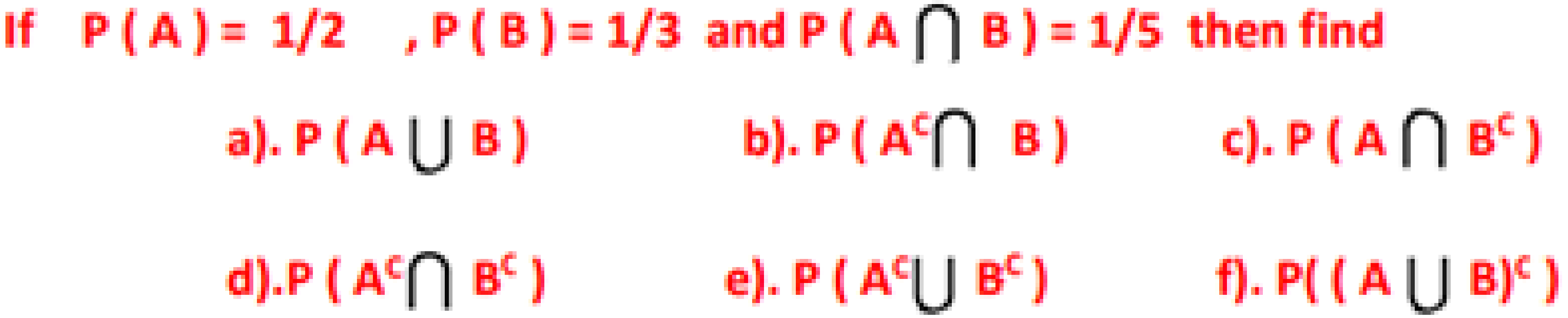
Observation 1: The variation in Protein is found very low when compared with remaining using coefficient of variation = sd/mean

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…………

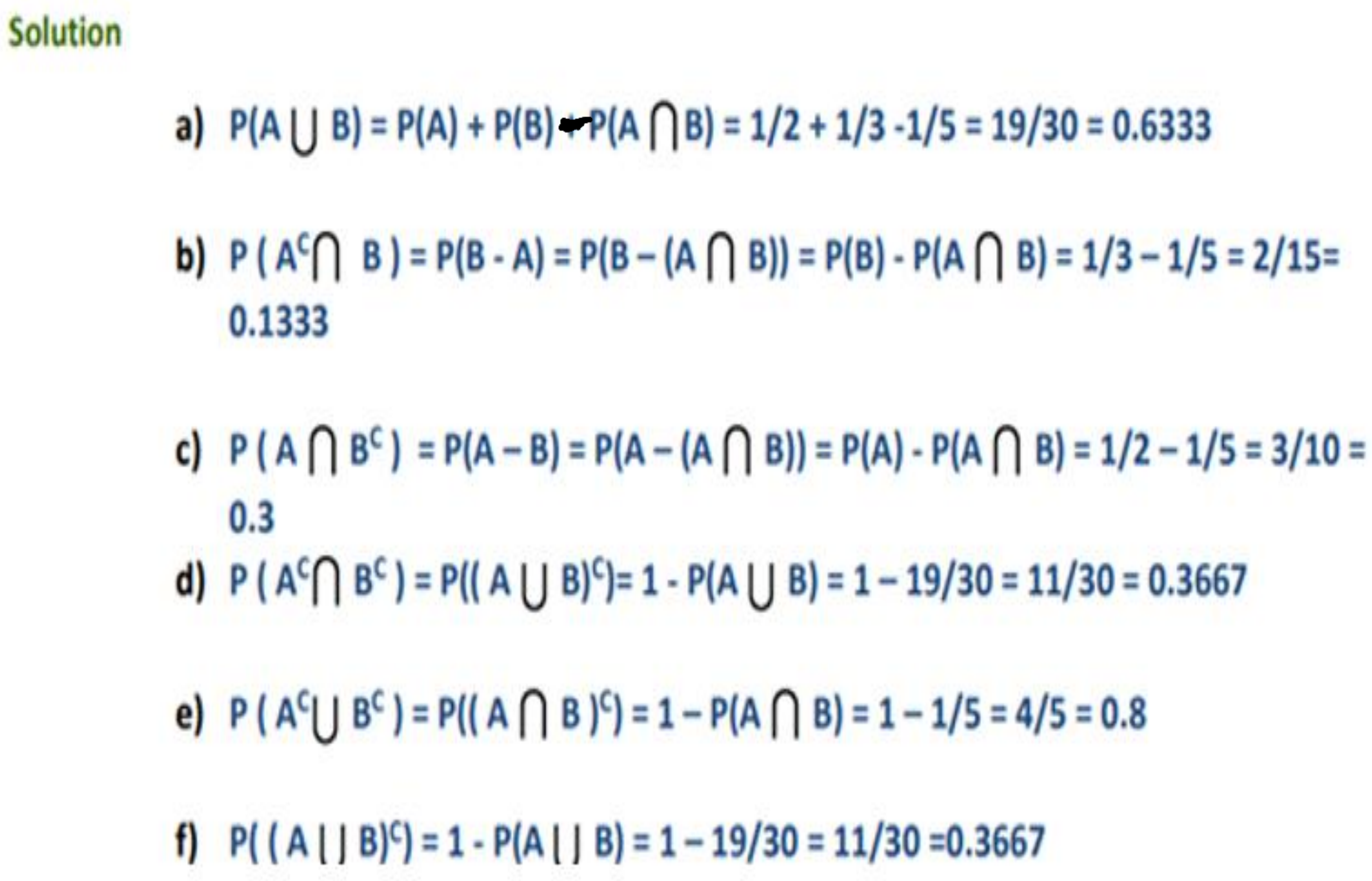
Basic Probability



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

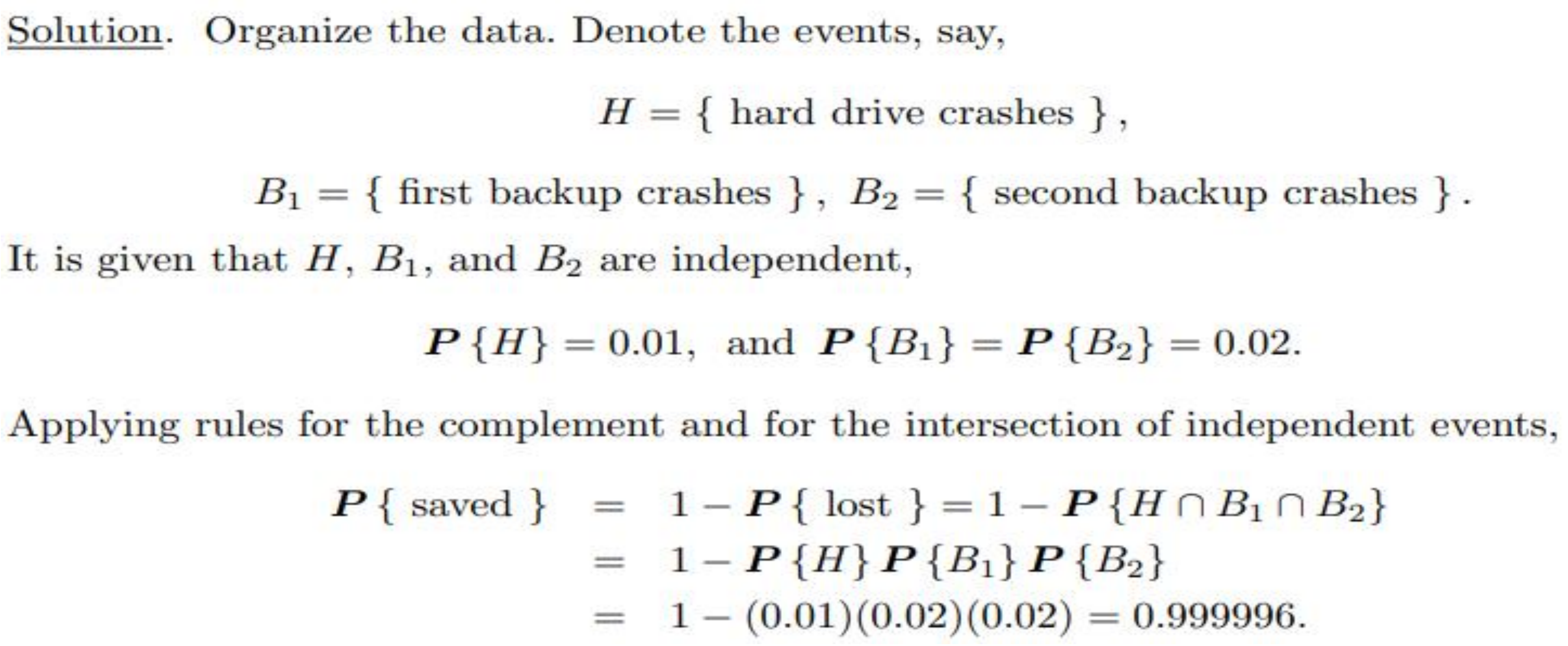
d) $P(A^c \cap B^C) = P((A \cup B)^C) = 1 - P(A \cup B) = 1 - 19/30 = 11/30 = 0.3667$

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

4) There is a 1% probability for a hard drive to crash. Therefore, it has two backups, each having a 2 % probability to crash, and all three components are independent of each other. The stored information is lost only in an unfortunate situation when all three devices crash. What is the probability that the information is saved?



Solution. Organize the data. Denote the events, say,

$$H = \{ \text{ hard drive crashes } \},$$

$$B_1 = \{ \text{ first backup crashes } \},\ B_2 = \{ \text{ second backup crashes } \}.$$

It is given that H, B_1, and B_2 are independent,

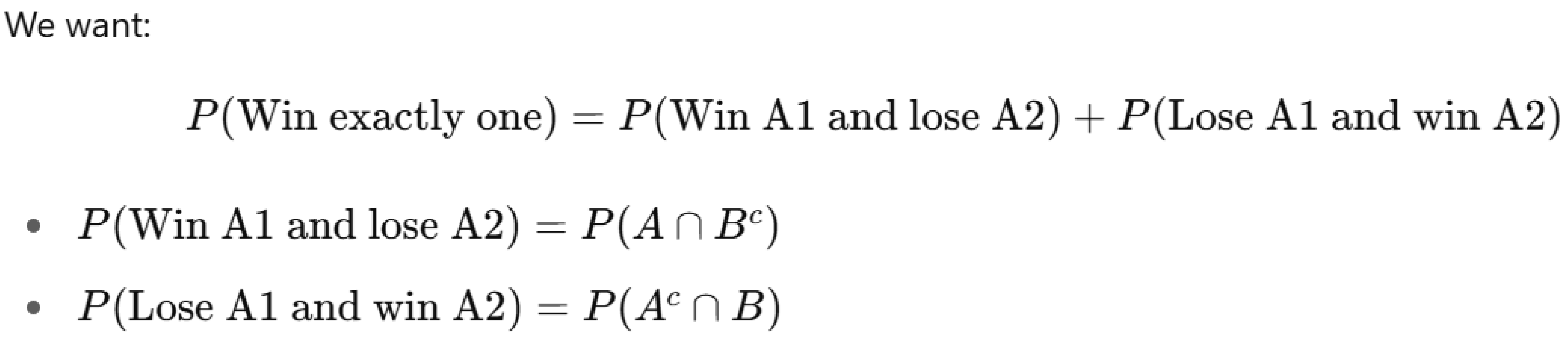
$$\boldsymbol{P}\{H\} = 0.01, \quad \text{and} \quad \boldsymbol{P}\{B_1\} = \boldsymbol{P}\{B_2\} = 0.02.$$

Applying rules for the complement and for the intersection of independent events,

$$\begin{aligned} \boldsymbol{P}\{ \text{ saved } \} &= 1 - \boldsymbol{P}\{ \text{ lost } \} = 1 - \boldsymbol{P}\{H \cap B_1 \cap B_2\} \\ &= 1 - \boldsymbol{P}\{H\}\boldsymbol{P}\{B_1\}\boldsymbol{P}\{B_2\} \\ &= 1 - (0.01)(0.02)(0.02) = 0.999996. \end{aligned}$$

Basic Probability

5) 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 exactly one of the constituencies?



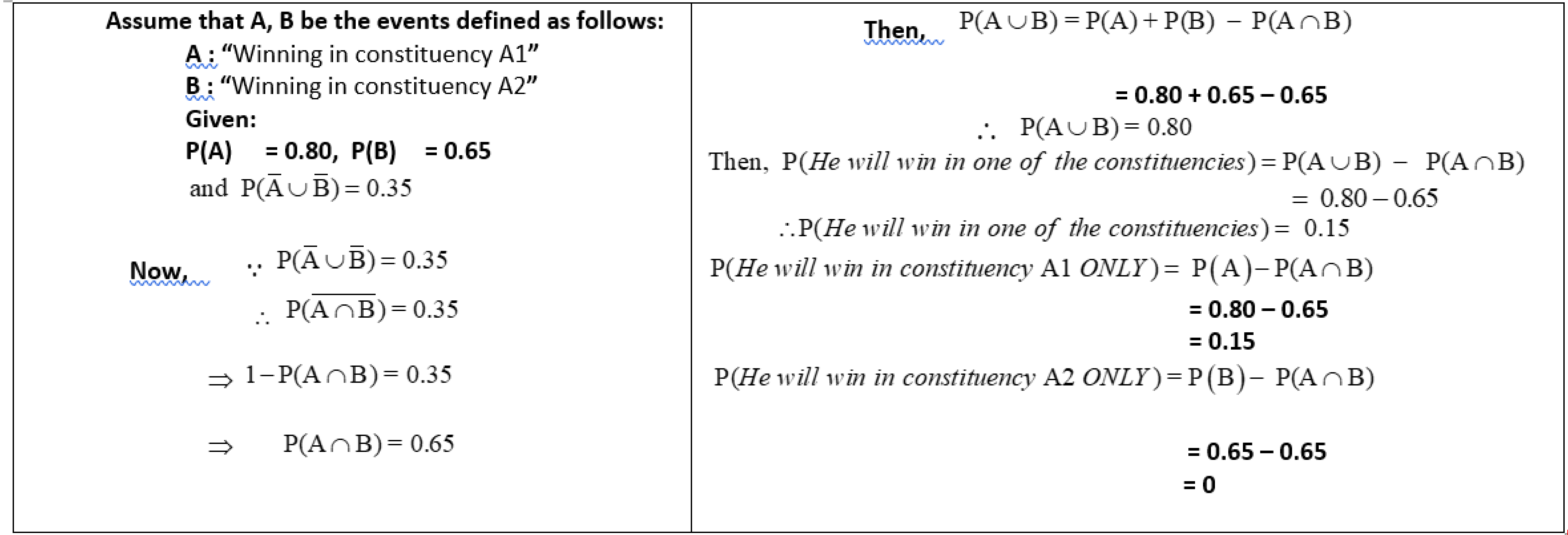
We want:

$$P(\text{Win exactly one}) = P(\text{Win A1 and lose A2}) + P(\text{Lose A1 and win A2})$$

- $P(\text{Win A1 and lose A2}) = P(A \cap B^c)$
- $P(\text{Lose A1 and win A2}) = P(A^c \cap B)$

Solution:

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?



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, $\because P(\bar{A}\cup\bar{B}) = 0.35$

$\therefore P(\overline{A\cap B}) = 0.35$

$\Rightarrow 1-P(A\cap B) = 0.35$

$\Rightarrow P(A\cap B) = 0.65$

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, $P(\textit{He will win in one of the constituencies}) = P(A\cup B) - P(A\cap B)$

$= 0.80 - 0.65$

$\therefore P(\textit{He will win in one of the constituencies}) = 0.15$

$P(\textit{He will win in constituency A1 ONLY}) = P(A) - P(A\cap B)$

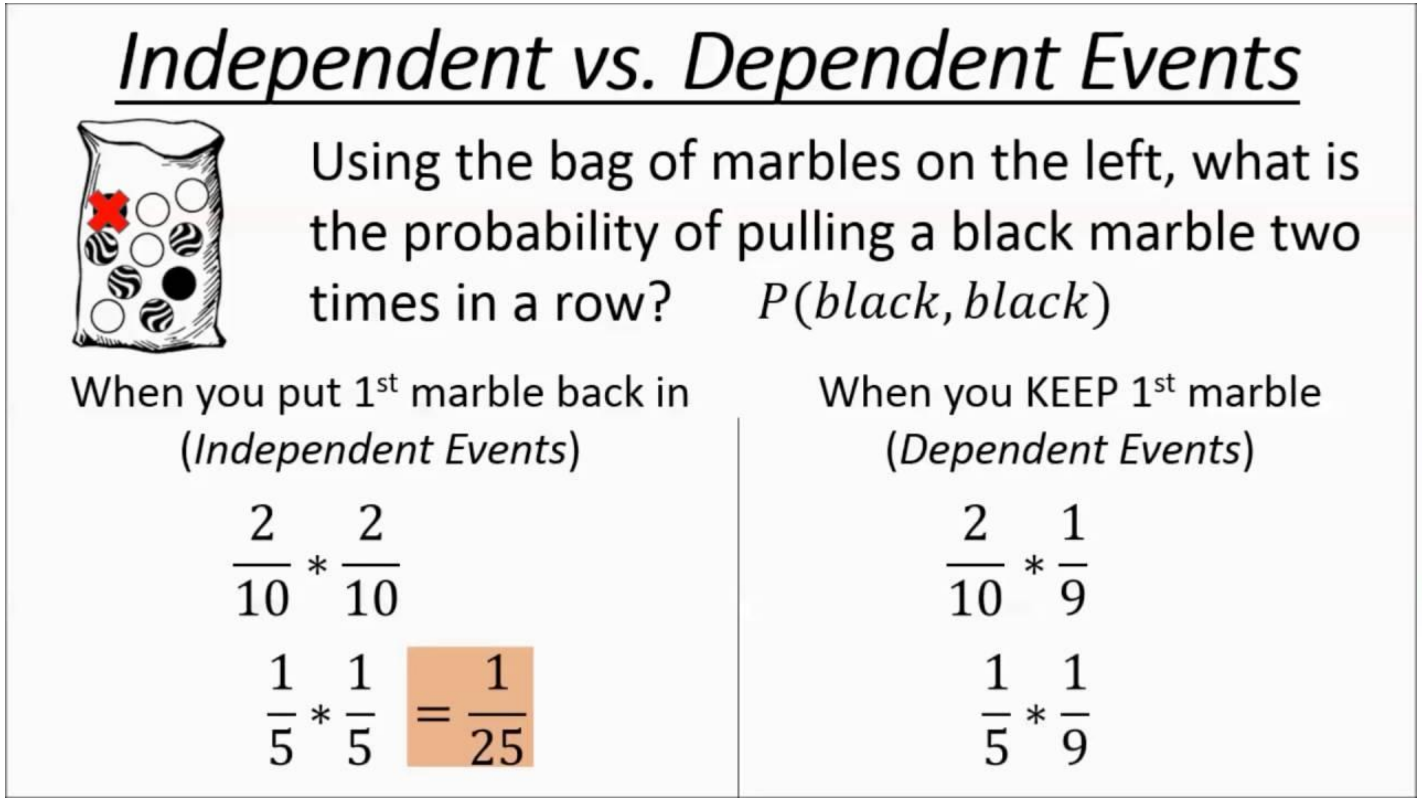
= 0.80 – 0.65

= 0.15

$P(\textit{He will win in constituency A2 ONLY}) = P(B) - P(A\cap B)$

= 0.65 – 0.65

= 0



Independent vs. Dependent Events

Using the bag of marbles on the left, what is the probability of pulling a black marble two times in a row? $P(black, black)$

When you put 1st marble back in (*Independent Events*)

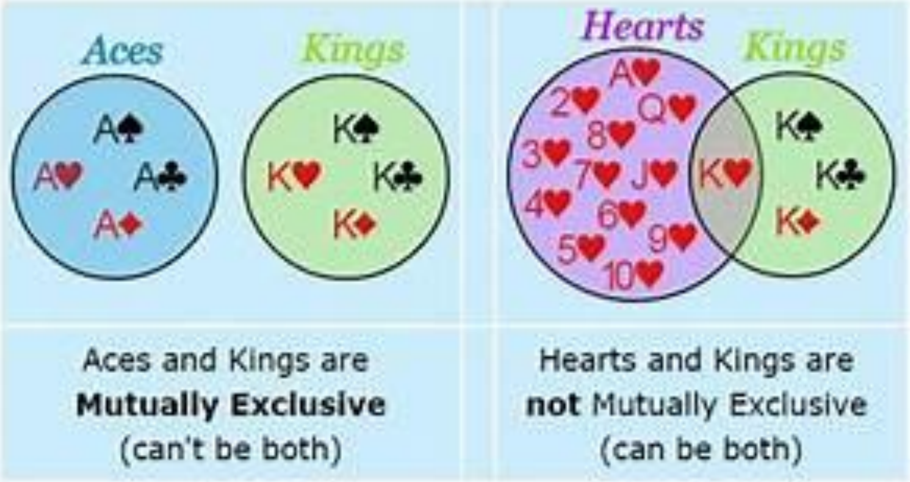
$$\frac{2}{10} * \frac{2}{10}$$

$$\frac{1}{5} * \frac{1}{5} = \frac{1}{25}$$

When you KEEP 1st marble (*Dependent Events*)

$$\frac{2}{10} * \frac{1}{9}$$

$$\frac{1}{5} * \frac{1}{9}$$



Aces and Kings are **Mutually Exclusive** (can't be both)

Hearts and Kings are **not** Mutually Exclusive (can be both)

Independent events

6) Comment on the statement:

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

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).

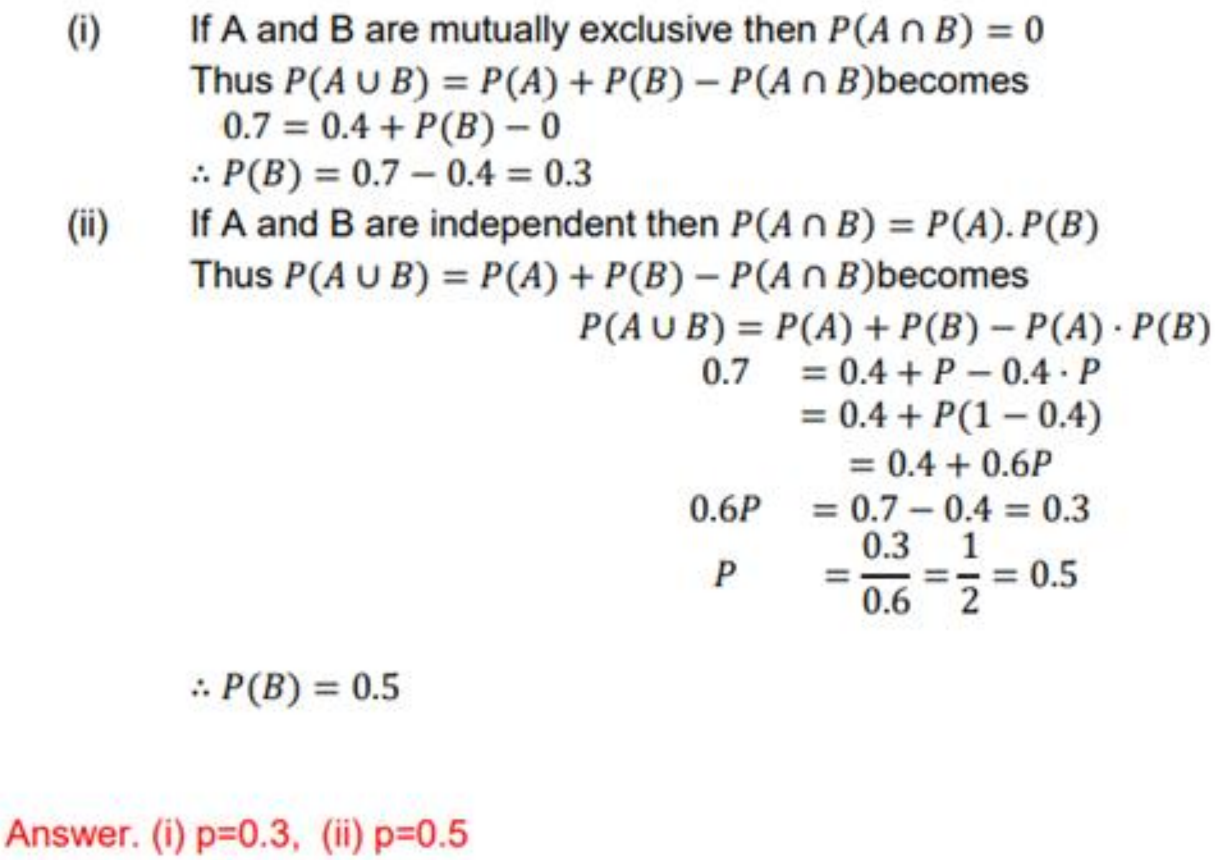
Independent events

7) 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?

Solution:



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

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

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A) \cdot P(B) \\ 0.7 &= 0.4 + P - 0.4 \cdot P \\ &= 0.4 + P(1 - 0.4) \\ &= 0.4 + 0.6P \\ 0.6P &= 0.7 - 0.4 = 0.3 \\ P &= \frac{0.3}{0.6} = \frac{1}{2} = 0.5 \end{aligned}$$

$$\therefore P(B) = 0.5$$

Answer. (i) p=0.3, (ii) p=0.5

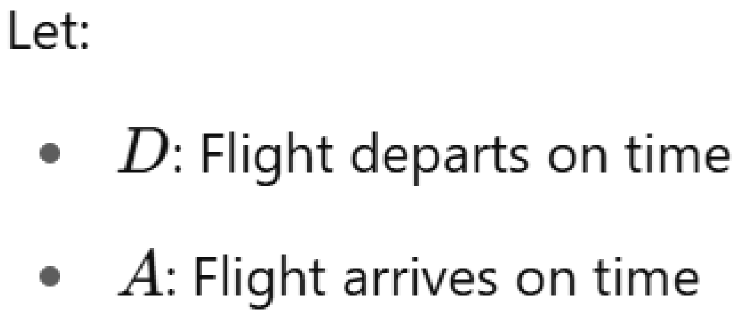
Conditional Probability

8. Ninety percent of flights depart on time. Eighty percent of flights arrive on time. Seventy-five percent of flights depart on time and arrive on time.

a) You are meeting a flight that departed on time. What is the probability that it will arrive on time?

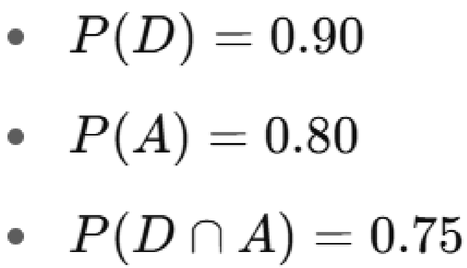
b) You have met a flight, and it arrived on time. What is the probability that it departed on time?

c) Are the events, departing on time and arriving on time, independent?



Let:

- D: Flight departs on time
- A: Flight arrives on time



- $P(D) = 0.90$
- $P(A) = 0.80$
- $P(D \cap A) = 0.75$

Solution:

a) You are meeting a flight that departed on time. What is the probability that it will arrive on time?

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

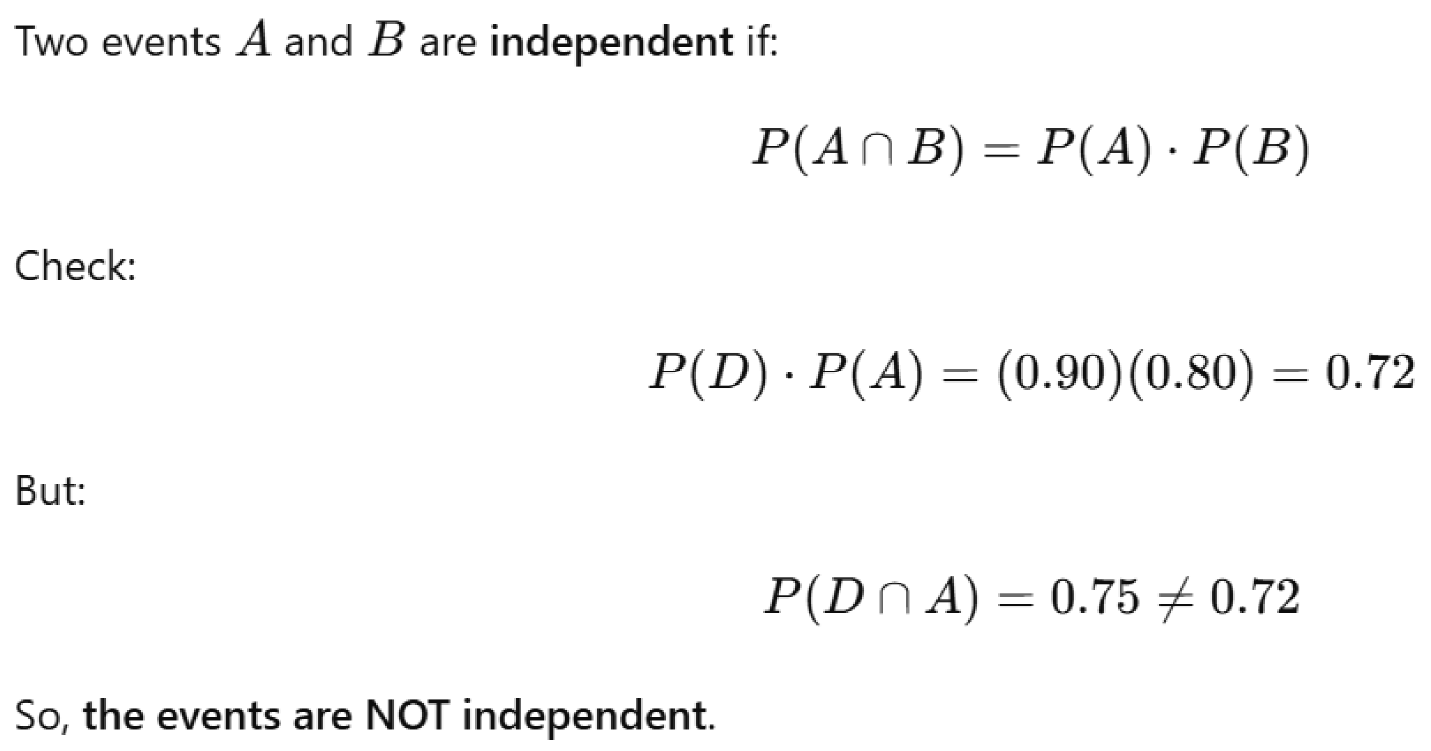
$$P(A \mid D) = \frac{0.75}{0.90} = \frac{75}{90} = \frac{5}{6} \approx 0.833$$

b) You have met a flight, and it arrived on time. What is the probability that it departed on time?

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

$$P(D \mid A) = \frac{0.75}{0.80} = \frac{75}{80} = \frac{15}{16} = 0.9375$$

c) Are the events, departing on time and arriving on time, independent?



Two events A and B are **independent** if:

$$P(A \cap B) = P(A) \cdot P(B)$$

Check:

$$P(D) \cdot P(A) = (0.90)(0.80) = 0.72$$

But:

$$P(D \cap A) = 0.75 \neq 0.72$$

So, **the events are NOT independent**.

Conditional Probability

9) 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

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.

Conditional Probability

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

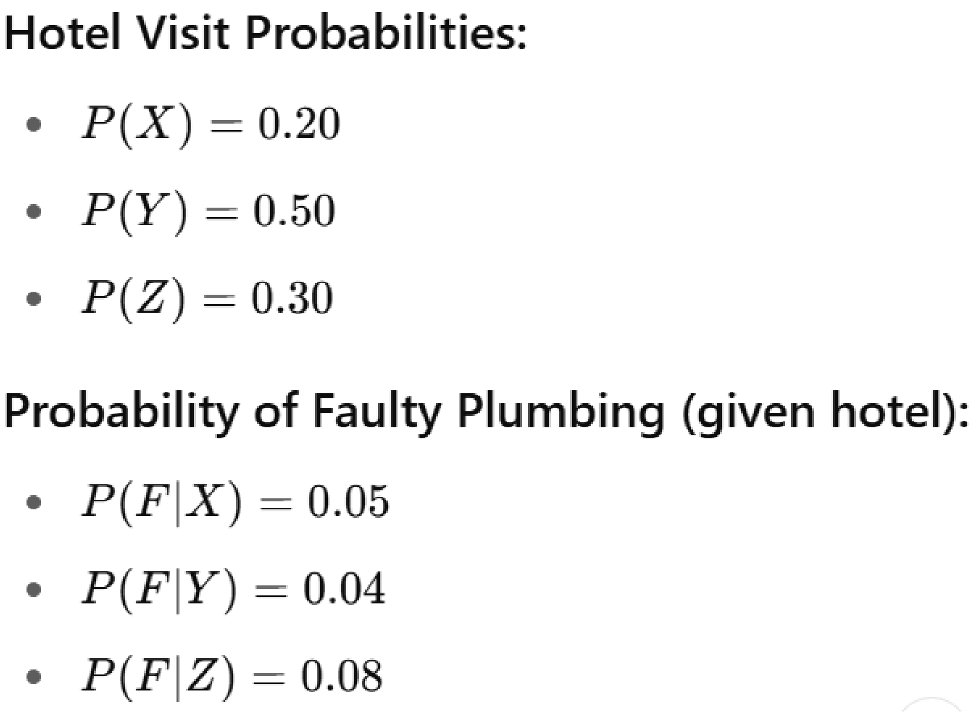
$$P(F' \cap A) = P(A) - P(A \cap F)$$
$$= 0.45 - 0.05$$
$$= 0.40$$

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

Total Probability

10) A businessman goes to hotels X, Y, Z -20%, 50%, 30% of the time, respectively. It is known that 5%, 4%, 8% of the rooms in X, Y, Z hotels have faulty plumbing. Determine the probability that the businessman goes to hotel with faulty plumbing.

Given:



Hotel Visit Probabilities:

- $P(X) = 0.20$
- $P(Y) = 0.50$
- $P(Z) = 0.30$

Probability of Faulty Plumbing (given hotel):

- $P(F|X) = 0.05$
- $P(F|Y) = 0.04$
- $P(F|Z) = 0.08$

Find the **total probability** that the businessman goes to a hotel with faulty plumbing:

$$P(F) = P(X) \cdot P(F|X) + P(Y) \cdot P(F|Y) + P(Z) \cdot P(F|Z)$$

$$P(F) = (0.20 \times 0.05) + (0.50 \times 0.04) + (0.30 \times 0.08)$$

$$= 0.010 + 0.020 + 0.024 = \boxed{0.054}$$

Total Probability

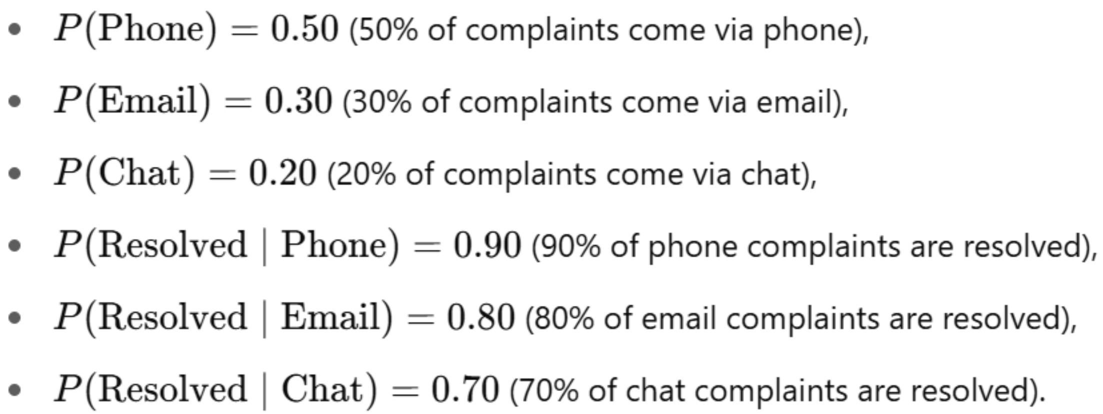
11) A company receives 50% of customer complaints via phone, 30% via email, and 20% via chat.

Resolution rates:

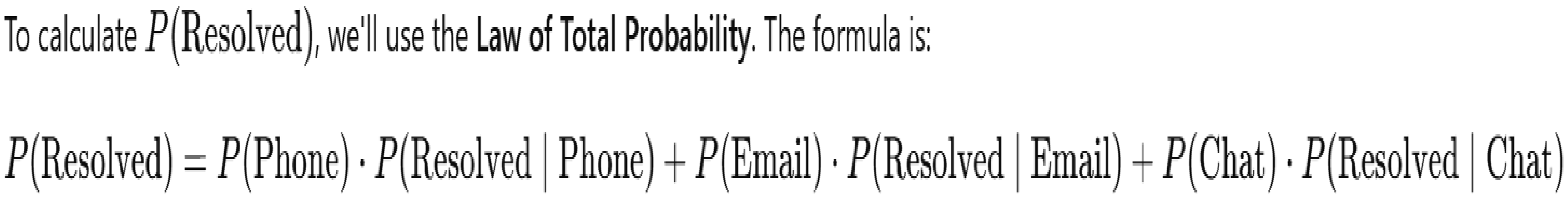
- Phone: 90%
- Email: 80%
- Chat: 70%

What is the probability that a randomly selected complaint is resolved?

Given:

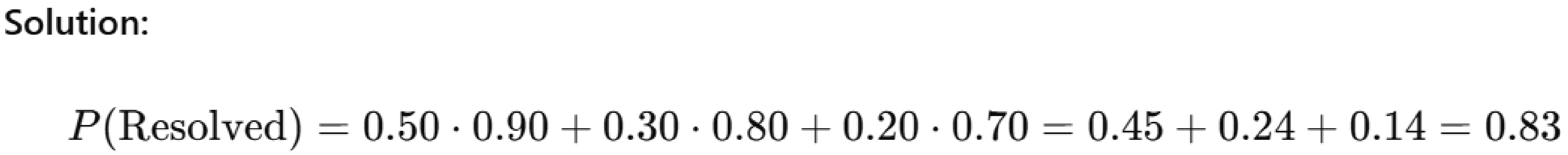


- $P(\text{Phone}) = 0.50$ (50% of complaints come via phone),
- $P(\text{Email}) = 0.30$ (30% of complaints come via email),
- $P(\text{Chat}) = 0.20$ (20% of complaints come via chat),
- $P(\text{Resolved} \mid \text{Phone}) = 0.90$ (90% of phone complaints are resolved),
- $P(\text{Resolved} \mid \text{Email}) = 0.80$ (80% of email complaints are resolved),
- $P(\text{Resolved} \mid \text{Chat}) = 0.70$ (70% of chat complaints are resolved).



To calculate $P(\text{Resolved})$, we'll use the **Law of Total Probability**. The formula is:

$$P(\text{Resolved}) = P(\text{Phone}) \cdot P(\text{Resolved} \mid \text{Phone}) + P(\text{Email}) \cdot P(\text{Resolved} \mid \text{Email}) + P(\text{Chat}) \cdot P(\text{Resolved} \mid \text{Chat})$$



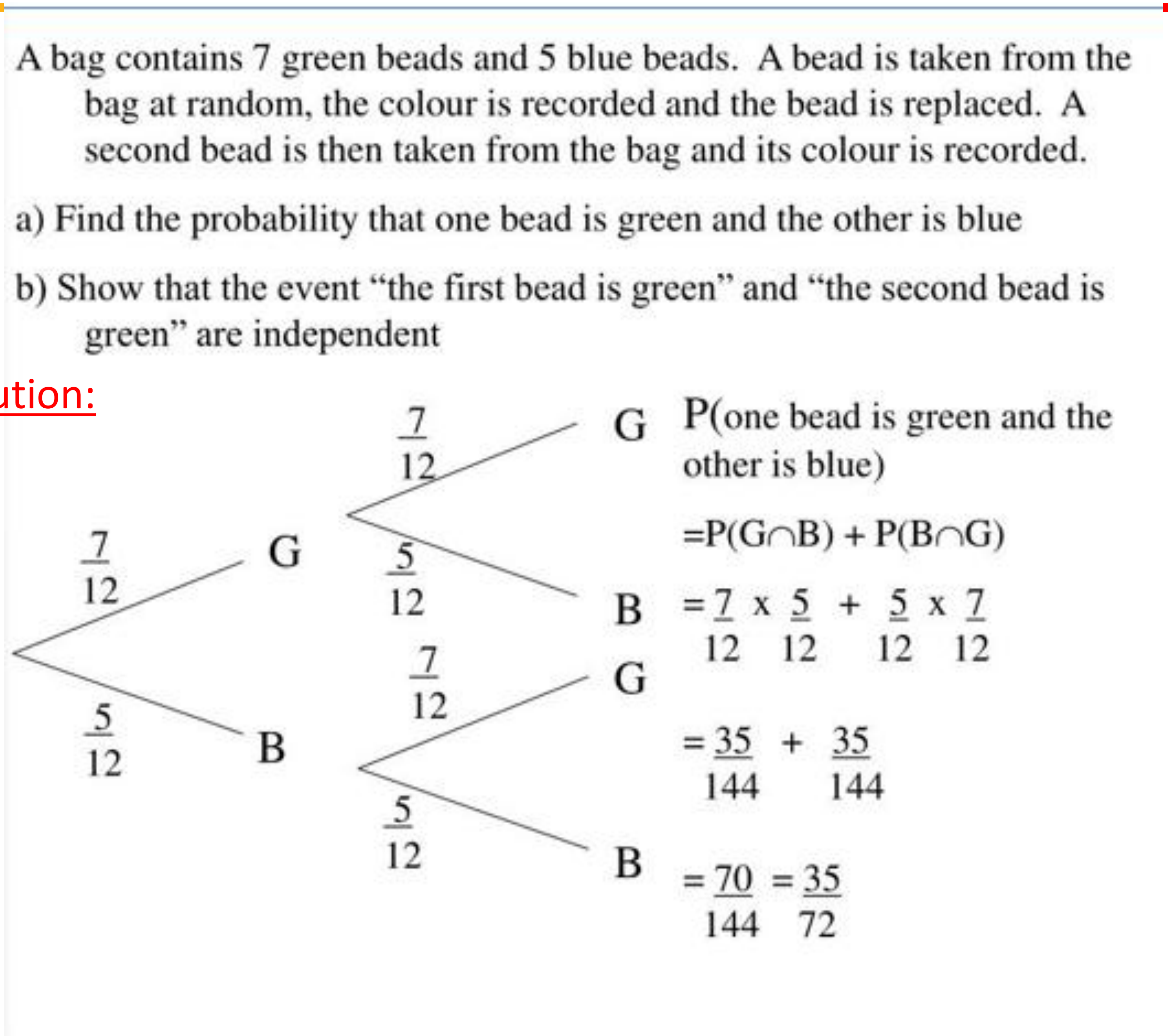
Solution:

$$P(\text{Resolved}) = 0.50 \cdot 0.90 + 0.30 \cdot 0.80 + 0.20 \cdot 0.70 = 0.45 + 0.24 + 0.14 = 0.83$$

Conclusion:

So, there is an **83%** chance that a randomly selected customer complaint will be resolved, based on the given resolution rates for each channel.

Problem:



12. A bag contains 7 green beads and 5 blue beads. A bead is taken from the bag at random, the colour is recorded and the bead is replaced. A second bead is then taken from the bag and its colour is recorded.

a) Find the probability that one bead is green and the other is blue

b) Show that the event "the first bead is green" and "the second bead is green" are independent

Solution:

$\frac{7}{12}$
G
$\frac{5}{12}$
B
$\frac{7}{12}$
G
$\frac{5}{12}$
B
$\frac{7}{12}$
G
$\frac{5}{12}$
B

P(one bead is green and the other is blue)

$$=P(G\cap B) + P(B\cap G)$$

$$= \frac{7}{12} \text{ x } \frac{5}{12} + \frac{5}{12} \text{ x } \frac{7}{12}$$

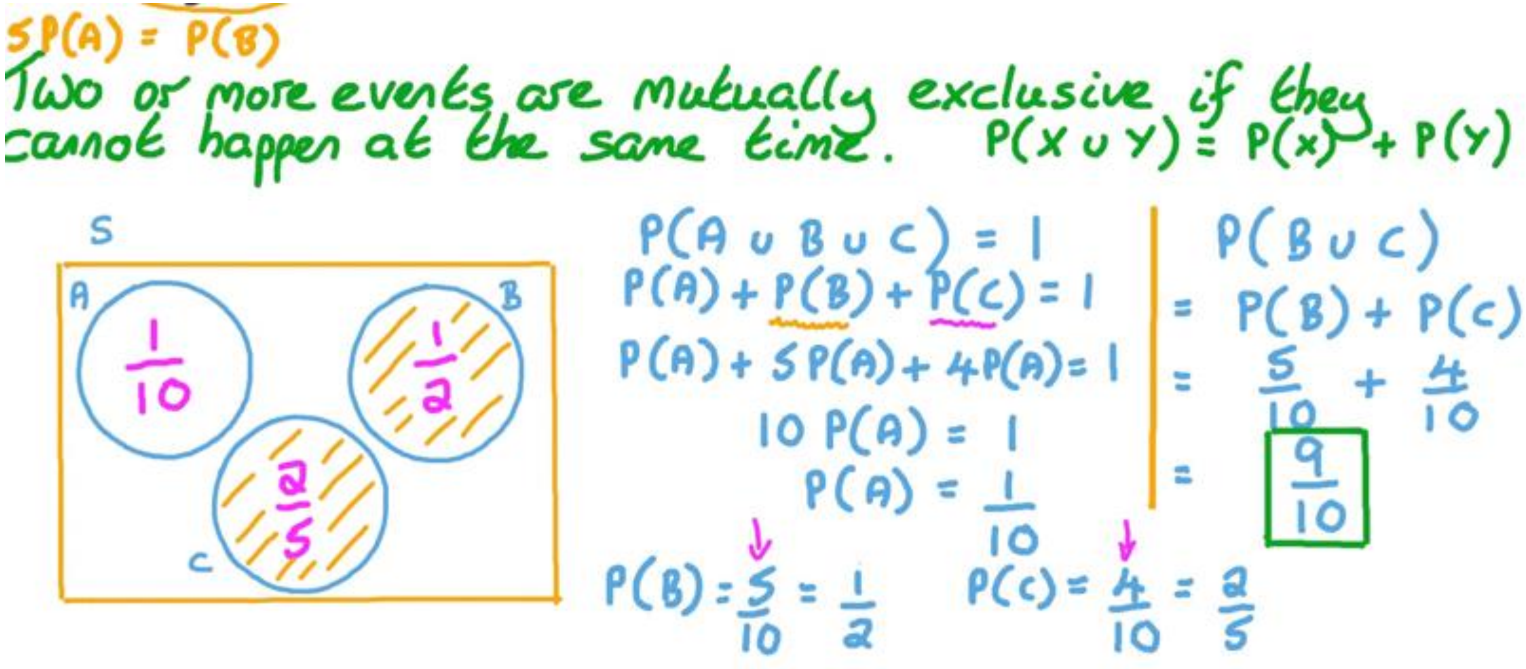
$$= \frac{35}{144} + \frac{35}{144}$$

$$= \frac{70}{144} = \frac{35}{72}$$

Problem:

13. Suppose A, B and C are three mutually exclusive events in a sample space. Given that S= AUBUC, P(A) = (1/5)P(B), and P(C)= 4P(A), find P(BUC).

Solution:



$5P(A) = P(B)$

Two or more events are mutually exclusive if they cannot happen at the same time. $P(X \cup Y) = P(x) + P(Y)$

$$P(A \cup B \cup C) = 1$$
$$P(A) + P(B) + P(C) = 1$$
$$P(A) + 5P(A) + 4P(A) = 1$$
$$10\,P(A) = 1$$
$$P(A) = \frac{1}{10}$$

$$P(B) = \frac{5}{10} = \frac{1}{2} \qquad P(C) = \frac{4}{10} = \frac{2}{5}$$

$$P(B \cup C) = P(B) + P(C) = \frac{5}{10} + \frac{4}{10} = \boxed{\frac{9}{10}}$$

S
A
B
C
1/10
1/2
2/5

Practice Problems

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

2. Three machines A, B , C produce 50%, 30%, and 20% of the items in a factory. The percentage of defective outputs of these machines are 3, 4 and 5 respectively.

(i) If an item is selected at random, what is the probability that it is defective? (Ans: 0.037)

(ii) If a selected item is defective, what is the probability that it is from machine A? (Ans: 0.4054)

Practice Problems

3. A weather app predicts rain with 90% accuracy when it actually rains. It wrongly predicts rain 10% of the time when it doesn't rain. Suppose it rains only 10% of days in a year. You check the app and it predicts rain.
What is the probability it will actually rain? (Ans: 0.5)

4. 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 traveller checked work email and brought a laptop, what is the probability that he/she uses a cell phone to stay connected?

TEXT BOOKS

T1 : Statistics for Data Scientists, An introduction to Probability, Statistics and Data Analysis, Maurits Kaptein et al, Springer 2022

T2 : Probability and Statistics for Engineering and Sciences, 8th Edition, Jay L Devore, Cengage Learning

T3 : Introduction to Time Series and Forecasting, Second Edition, Peter J Brockwell, Richard A Davis, Springer.

Questions please…

THANK YOU!