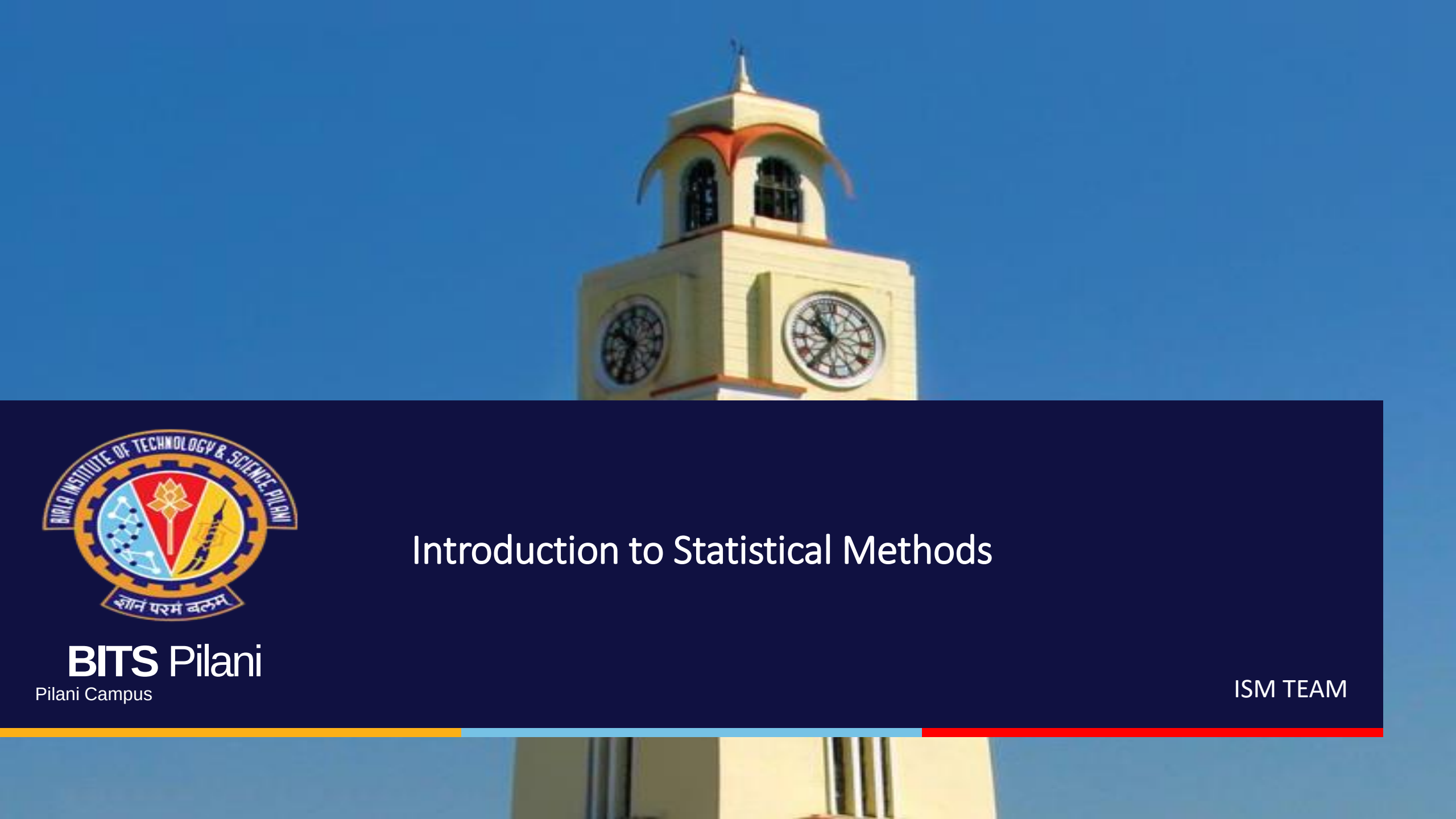


Introduction to Statistical Methods

BITS Pilani

Pilani Campus

ISM TEAM

innovate achieve lead

BITS Pilani

Pilani Campus

Course No: AIMLCZC418

Webinar 1 : 12.12.2024

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. Rang, 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

S.No	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 $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	Variance = 26
Standard deviation	For a Population $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$ For a Sample $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	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

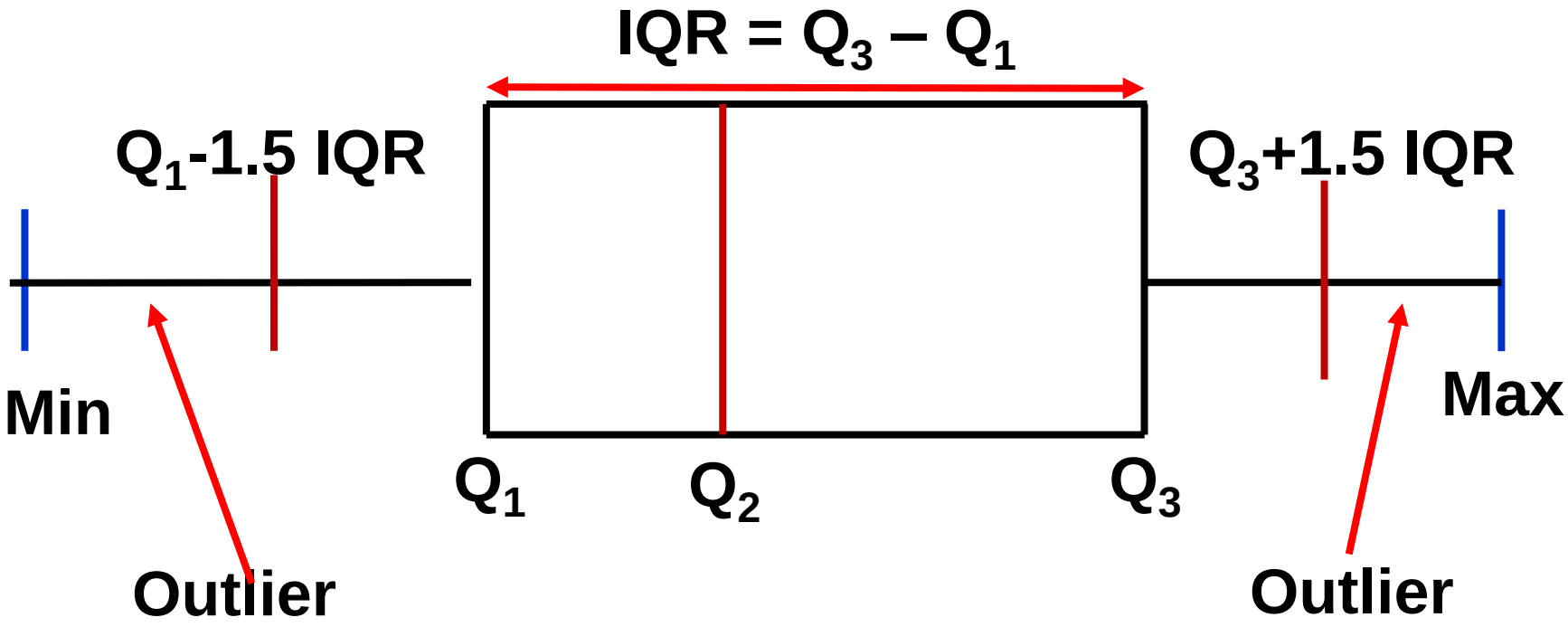
- **Minimum**.
- **Q1** (the first quartile, or the 25% mark).
- **Median**.
- **Q3** (the third quartile, or the 75% mark).
- **Maximum**.

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

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



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 10^{th} value: 75.

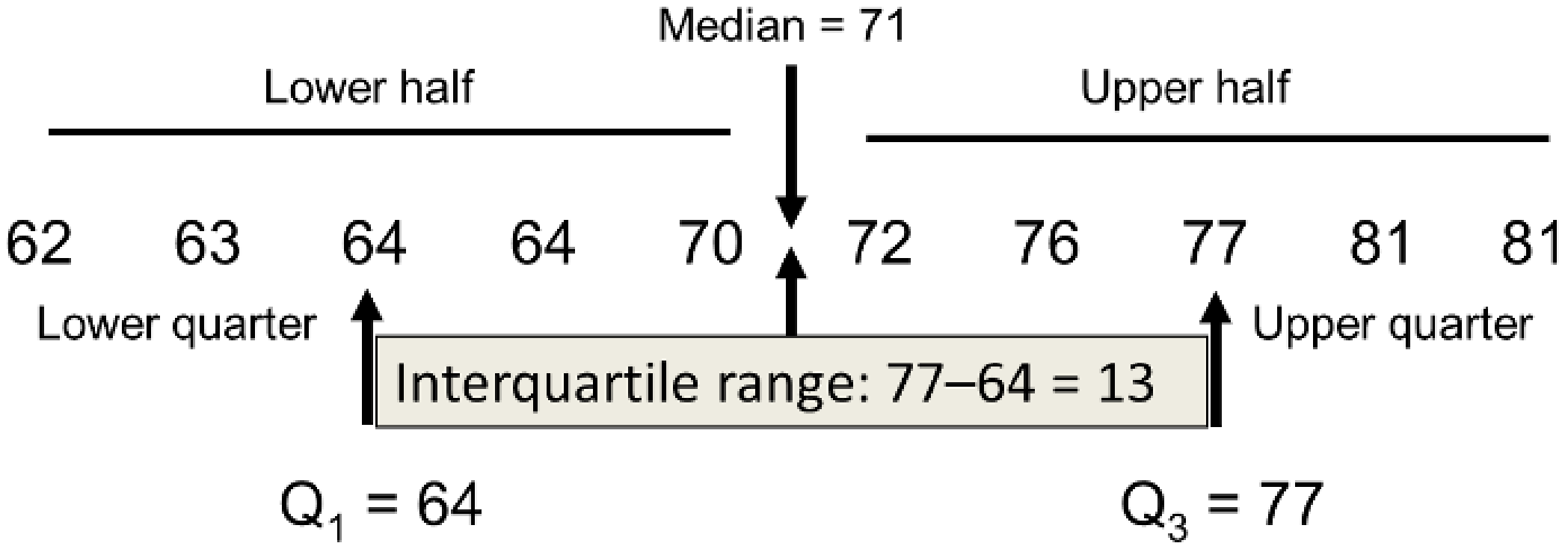
Q1 is the central point between the smallest score and the median. In this case, Q1 falls between the first and fifth score: 68. (Note that the median can also be included when calculating Q1 or Q3 for an odd set of values. If we were to include the median on either side of the middle point, then Q1 will be the middle value between the first and 10^{th} 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 25^{th} 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 50^{th} percentile and shows that 50% of the scores are less than 75, and 50% of the scores are above 75. Finally, Q3, the 75^{th} percentile, reveals that 25% of the scores are greater and 75% are less than 84.

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



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 re maining 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) 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,

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

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

4) 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) $P((A \mid B)^c) = 1 - P(A \mid B) = 1 - 19/30 = 11/30 = 0.3667$

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 one of the constituencies?

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*)

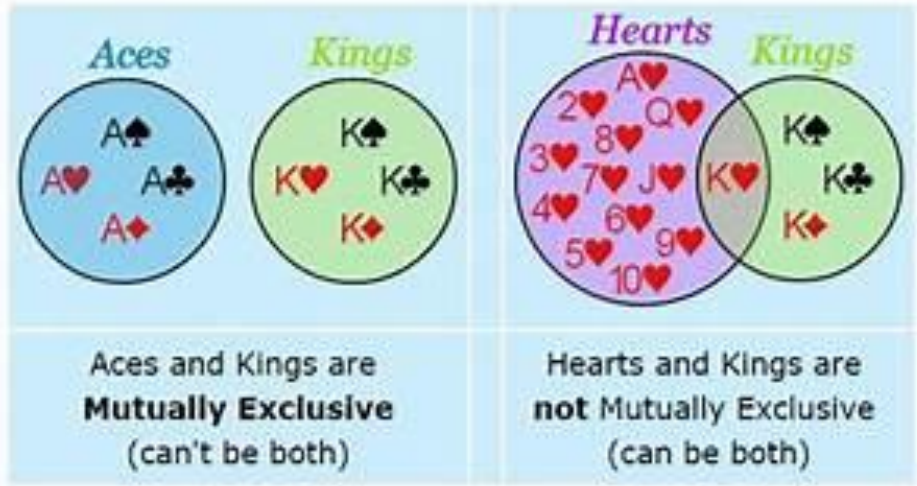
$$\frac{2}{10} * \frac{2}{10}$$

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

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

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

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



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

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

$\therefore P(B) = 0.5$

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

8. Given $P(A) = \frac{1}{4}$, $P(B) = \frac{1}{3}$, $P(A \cup B) = \frac{1}{2}$

Evaluate $P(A|B)$, $P(B|A)$, $P(A \cap \bar{B})$, $P(A|\bar{B})$

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

$P(A \cap B) = \frac{1}{4} + \frac{1}{3} - \frac{1}{2} = \frac{1}{12}$.

$P(A|B) = P(A \cap B) \mid P(B) = 1/4$.

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{1}{3}.$$

$$P(A \cap \bar{B}) = P(A) - P(A \cap B) = \frac{1}{6}.$$

$$P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{1}{4}.$$

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.

Solution :- A: Event of faulty plumbing

$B_1 = X \quad B_2 = Y \quad B_3 = Z$

By theorem on total Probability

$$P[\text{Faulty Plumbing}] = P(A) = \sum_{i=1}^{3} P(B_i)\, P(A|B_i)$$

$$= P(X)\,P(A|X) + P(Y)\,P(A|Y) + P(Z)\,P(A|Z)$$

It is Known that

$$P(B_1) = P(X) = \frac{20}{100} = 0.2, \quad P(B_3) = P(Z) = 0.3$$

$$P(B_2) = P(Y) = \frac{50}{100} = 0.5, \quad P(A|X) = \frac{5}{100} = 0.05$$

$$P(A|Y) = \frac{4}{100} = 0.04$$

$$P(A|Z) = \frac{8}{100} = 0.08$$

$$\therefore P(A) = (0.2)(0.05) + (0.5)(0.04) + (0.3)(0.08)$$

$$= 0.054$$

Total Probability

11) Three machines A, B , C produce 50%, 30%, and 20% of the items in a fctory. The percentage of defective outputs of these machines are 3, 4 and 5 respectively. If an item is selected at random, what is the probability that it is defective? If a selected item is defective, what is the probability that it is from machine A?

P(A), P(B), P(C) → Probability of Choosing an item Produced by maching A, B, C.

P(E) → Probability of Choosing a defective item from the whole output.

P(E|A), P(E|B) P(E|C) → Probability of Choosing a defective items from A, B, C

Given $P(A) = 0.5$, $P(B) = 0.3$, $P(C) = 0.2$

$P(E|A) = 0.03$ $P(E|B) = 0.04$, $P(E|C) = 0.05$

$P(E) = P(A)\,P(E|A) + P(B)\,P(E|B) + P(C)\,P(E|C) = 0.037.$

$$P(A|E) = \frac{P(A)\,P(E|A)}{P(E)} = 0.4054$$

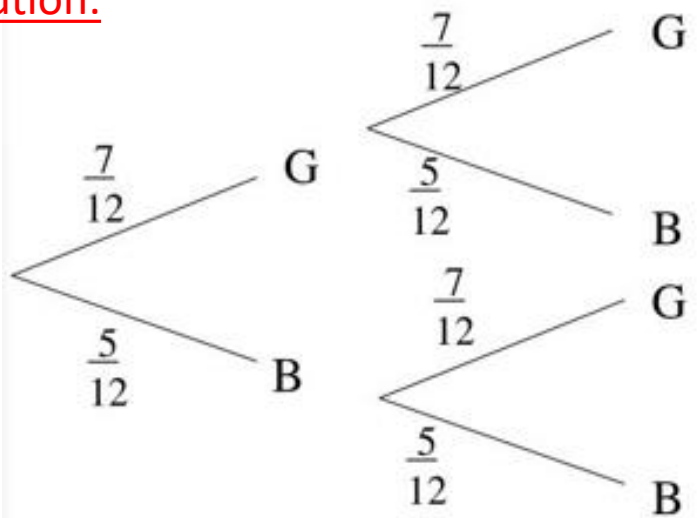
Problem:

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:



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

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

$= \frac{7}{12} \times \frac{5}{12} + \frac{5}{12} \times \frac{7}{12}$

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

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

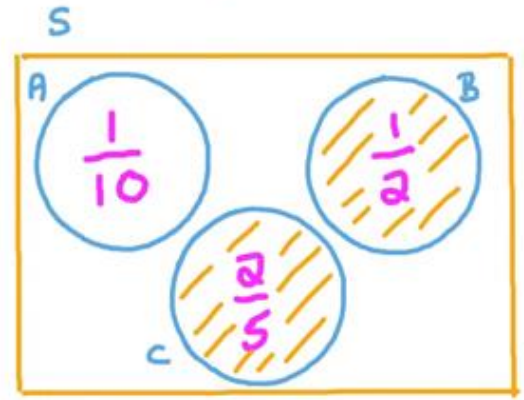
Problem:

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. $P(X \cup Y) = P(X) + P(Y)$



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

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

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

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. A manufacturing company produces certain types of output by 4 machines i.e A, B, C and D. Machine A produces 30%, Machine B produces 15 % 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 by which Machine? If so, find it.

Practice Problems

3. 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 3% and 5% 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? Also,based on this information write your observations.

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?

THANK YOU!