Statistics for Data Science-1

Week 1 Graded Assignment Solution

The education minister wants to know the status of campus placements of B.Tech students in different engineering institutes of India. An analyst did a survey on the randomly selected four IITs of India and analysed the status of campus placements. Based on the information given, answer the questions (1), (2) and (3).

1. Identify the sample and population.

   (a) The sample consists of all the engineering institutes of India and the population consists of randomly selected four IITs of India.

   (b) The sample consists of all the IITs of India and the population consists of all the engineering institutes of India.

   (c) The sample consists of all IITs of India and the population consists of randomly selected four IITs of India.

   (d) The sample consists of four randomly selected IITs of India and the population consists of all the engineering institutes of India.

   Answer: d

   Solution:

   By definition, population is the entire collection of elements we are interested in. Here, the purpose of the survey is to know the status of campus placements of B.Tech students in different engineering institutes of India. Hence, the population will be all the engineering institutes of India.

   Also, sample is a subset of the population which is being studied. Since, the four IITs of India is selected to know the status of campus placement. Therefore, sample is four randomly selected IITs of India.

   Thus, the sample consists of four randomly selected IITs of India and the population consists of all the engineering institutes of India.

   Hence, option (d) is correct.

2. The report given by an analyst to the education minister about the status of campus placements states that "The campus placement of B.Tech students is 95% in the different engineering institutes of India". The given statement of analyst is based on which kind of statistical analysis?

   (a) Descriptive Statistics

   (b) Inferential Statistics

   Answer: b

   Solution:

Making conclusions from the sample data comes under inferential statistics. Here, analyst makes the conclusion in the report that "The campus placement of B.Tech students is 95% in the different engineering institutes of India" based on the information of four randomly selected IITs of India. Therefore, the study is inferential statistics.
Hence, option (b) is correct.

3. Is the conclusion of this study made by analyst on the basis of chosen sample reliable?

   (a) Yes

   (b) No

   Answer: b
   The objective of the survey is to know the status of campus placements of B.Tech students in different engineering institutes of India, but the institutes are selected only from IITs and not from different engineering institutes of India. Therefore, this sample is not a good representative of the population, as the status of the campus placement of B.Tech students could vary in various engineering institutes of India.
   Hence, option (b) is correct.

The data of five different types of fertilizers used by farmers of a village is tabulated in Table 1.1.G. Based on the information given, answer the questions (4), (5), (6), (7) and (8).

Fertilizers	Types of Fertilizers	Area of fields (In acres)	Types of Crops	Amount of fertilizers (In Kg)
Nitrogen	Inorganic	1	Rice	200
Phosphorus	Inorganic	2	Wheat	400
Manure	Organic	1.5	Potato	300
Compost	Organic	1.3	Rice	260
Potassium	Inorganic	1.6	Pulse	320

Table 1.1.G

4. Which of the following statements is/are true?

   (a) Inorganic is a case and Types of Fertilizers is a variable.

   (b) Rice is a case.

   (c) Manure is a case.

   (d) Amount of fertilizers is a variable.

(e) Nitrogen is a variable.

Answer: c, d

Solution:

Here, the specification data of the five different types of fertilizers used by farmers of a village are collected. So each specification (columns of the table) i.e. Fertilizers, Types of Fertilizers, Area of fields(In acres), Types of Crops and Amount of fertilizers(In Kg) is a variable.

Observation is an individual data point for which the entire data is being collected. So, here each value corresponding to which each of the specification noted is a case.

Thus, it is clear that Manure is a case.

Hence, options (c) and (d) are correct.

5. What is the scale of measurement of "Types of Crops"?

   (a) Ordinal Scale

   (b) Nominal Scale

   (c) Interval Scale

   (d) Ratio Scale

Answer: b

Solution:

Types of Crops is a categorical variable and it has four different categories of crops, i.e., Rice, Wheat, Potato and Pulse which are just labels.

Since, there is no particular order among the types of crops. Therefore, it has nominal scale of measurement.

Hence, option (b) is correct.

6. What kind of variable is "Area of fields"?(More than one option can be correct)

   (a) Categorical

   (b) Numerical

   (c) Discrete

   (d) Continuous

Answer: b, d

Solution:

Since Area of fields has numeric properties and can have arithmetic operations performed on it, it follows that Area of fields is a numerical variable. Moreover, it can take any value greater than 0 and have to measure in acres. Therefore, Area of fields is a continuous numerical variable.

Hence, options (b) and (d) are correct.

7. What is the scale of measurement of "Amount of Fertilizers"?

   (a) Ordinal Scale

   (b) Nominal Scale

   (c) Interval Scale

   (d) Ratio Scale

   Answer: d
   Solution:
   Amount of Fertilizers can have a meaningful interval. It also has an absolute zero. Hence, it comes under the ratio scale of measurement.
   Hence, option (d) is correct.

8. Is the data given in Table 1.1.G structured or unstructured?

   (a) The data is structured

   (b) The data is unstructured

   Answer: a
   Solution:
   Since the data of the five types of fertilizers used by farmers of a village can be organized in a well-defined tabular form. Therefore, it comes under the structured data.
   Hence, option (a) is correct.

9. The data of Netflix subscribers at the end of year 2020 across different Asian countries is recorded. Based on this, choose the correct option:

   (a) It is time series data

   (b) It is cross-sectional data

   Answer: b
   Solution:
   Since the data of Netflix subscribers are recorded across different Asian countries at a same time (i.e., at the end of year 2020), not at different time-intervals. Therefore, the data collected is cross-sectional data.
   Hence, option (b) is correct.

10. Choose the incorrect statement(s):

    (a) Stock price of Titan is numeric and continuous variable.

    (b) Number of assignments submitted by a student has an interval scale of measurement.

    (c) Soccer positions (i.e. Defender, Midfielder, Forward) has an ordinal scale of measurement.

(d) The education level of a person has a nominal scale of measurement.

Answer: b, c, d

Solution:

Since stock price of Titan has numeric properties and can have arithmetic operations performed on it, it follows that Stock Price of Titan is a numerical variable. Moreover, it can take any value (any non-negative value). Therefore, Stock price of Titan is a numeric and continuous variable. Therefore, option(a) is correct.

Number of assignments submitted by a student can have a meaningful interval. It also has an absolute zero. Hence, it comes under the ratio scale of measurement. Therefore, option(b) is incorrect.

Soccer positions (i.e. Defender, Midfielder, Forward) are just labels. There is no particular order among positions. Thus, it has nominal scale of measurement. Therefore, option(c) is incorrect.

There is a particular order in the education level of a person like 12^{th}, Undergraduate, graduate etc. Thus, it has an ordinal scale of measurement. Therefore, option(d) is incorrect.

Hence, options(b), (c) and (d) are correct.

Statistics for Data Science-1

Week-2 Graded Assignment Solution

1. Which of the following statements is/are incorrect?

   (a) To represent the share of a particular category, bar chart is the most appropriate graphical representation.

   (b) The multiplication of the total number of observations and relative frequency of a particular observation should be equal to the frequency of that observation.

   (c) Mean can be defined for a categorical variable.

   (d) Mode of a categorical variable is the widest slice in a pie chart.

   Answer: a, c

   Solution:

   To show the share of a particular category, pie chart is a most appropriate graphical representation. Thus, the statement of option (a) is incorrect.

   Suppose we have n observations and their corresponding frequencies are $f_1, f_2, \ldots f_n$ respectively.

   By the definition, Relative frequency for i^{th} observation can be defined as $R_{f_i} = \frac{f_i}{N}$; $i = 1, 2..., n$

   Thus, $f_i = R_{f_i} \times N$ which implies that the multiplication of the total number of observations and relative frequency of a particular(i^{th}) observation is equal to the frequency of that observation. Thus, the statement of option (b) is correct.

   Since we cannot perform any meaningful mathematical operations on categorical data. And, it is required to perform mathematical operation while computing mean of a dataset which is not possible in the case of categorical data. Thus, the statement of option(c) is incorrect.

   In a pie chart, the widest pie/slice will always have the highest frequency. Thus, mode will be the widest slice in a pie chart. Thus, the statement of option(d) is correct.

   Therefore, options (a) and (c) are correct.

Figure 2.1.G shows the pie chart representation of the weightage distribution of 5 different subjects in an exam. Based on this information, answer questions (2) and (3).

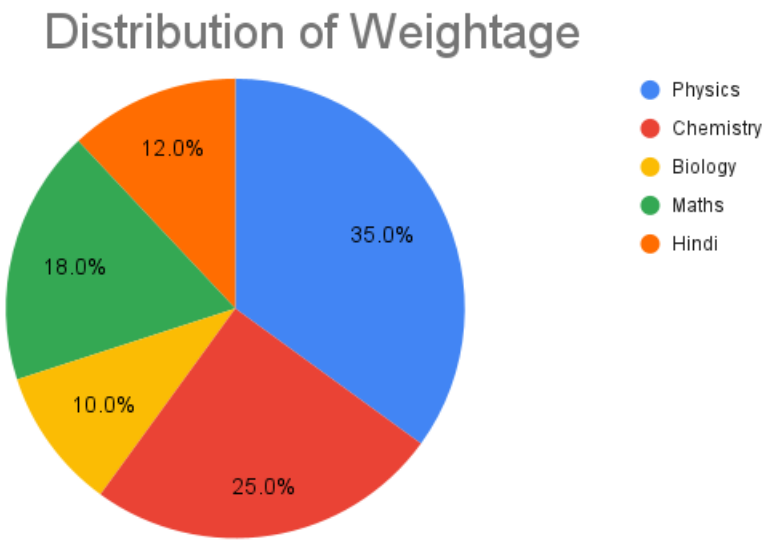


Figure 2.1.G: Weightage distribution of 5 different subjects

2. If the exam is for a total of 500 marks, then what is the aggregate distribution of marks in Physics, Maths and Biology?
   Answer: 315
   Solution:
   Since, the exam is for a total of 500 marks and Weightage of Physics is 35%.
   Therefore, marks in Physics $= 500 \times \frac{35}{100} = 175$.
   Similarly, as weightage of Maths is 18%.
   Therefore, marks in Maths $= 500 \times \frac{18}{100} = 90$.
   And, weightage of Biology is 10%.
   Therefore, marks in Biology $= 500 \times \frac{10}{100} = 50$.
   Hence, aggregate distribution of marks in Physics, Maths and Biology is 175+90+50 = 315.

3. Choose the correct statement(s):

   (a) The pie chart is misleading because it does not obey the area principle.

(b) The pie chart has round off errors.

(c) The pie chart is not a misleading graph.

(d) The slices of pie chart adds up to 100%.

Answer: c, d

Solution:

From the figure 2.1.G., it is clear that pie chart obeys the area principle as area occupied by a part of the chart is correspond to the amount of the data it represents.
Also, the slices of pie chart adds up to 100% as $12\%+35\%+25\%+10\%+18\% = 100\%$.
Thus, option(c) and (d) are correct.

Table 2.1.G represents the distribution of 200 cricket players trained by different cricket academies in Chennai.

Academy	Number of Players
A	a
B	b
C	50
D	d
E	75

Table 2.1.G

If each academy has trained at least one player, then based on the given information, answer questions (4), (5), (6) and (7).

4. What is the combined relative frequency of the academy A, B and D? (Enter the answer correct to 3 decimal places)

Answer: 0.375, Range: 0.370,0.380

Solution:

It is given that total number of cricket players is 200, i.e., $N = 200$.

Relative frequencies corresponding to academy C and academy D will be $\frac{50}{200} = 0.25$ and $\frac{75}{200} = 0.375$.

Let relative frequencies corresponding to academy A, B and D are R_{f_A}, R_{f_B} and R_{f_D} respectively.

Since, sum of all relative frequencies is equal to 1.

Therefore,

$$R_{f_A} + R_{f_B} + 0.25 + R_{f_D} + 0.375 = 1$$

$$R_{f_A} + R_{f_B} + R_{f_D} = 1 - (0.375 + 0.25) = 0.375$$

Hence, combined relative frequency of the academy A, B and D is 0.375

5. Median of the given data is:

   (a) Academy C

   (b) Academy E

   (c) Academy D

   (d) Median is not defined for the given data

   (e) Insufficient data

   Answer: d

   Solution:
   The given dataset has nominal scale of measurement and can not be ordered or ranked in an order. Hence, we can not compute median for it (as the first step in computation of median, i.e. arrange the dataset in ascending order, can not be performed).
   Hence, option (d) is correct.

6. Mode of the given data is:

   (a) Academy C

   (b) Academy E

   (c) Academy D

   (d) Mode is not defined for the given data

   (e) Insufficient data

   Answer: b

   Solution:
   There are a total of 200 cricket players, i.e. total frequency = 200.
   This implies that $a + b + d = 75$.
   Also, we know that each academy has trained at least one player, i.e. $a > 1, b > 1$ and $d > 1$.
   Therefore, the value of a, b and d will always be less than 75, which implies that Academy E will have the highest frequency for the given dataset.
   Hence, option (b) is correct.

7. Which of the following graphical representations is appropriate for the number of players in each academy for the given data in Table 2.1.G?

   (a) Bar chart

   (b) Pie chart

   (c) Pareto chart

   (d) Both bar chart and pareto chart

Answer: d

Solution:

Since, we are interested in the count/number of players, a bar chart would be a appropriate representation for the given dataset.

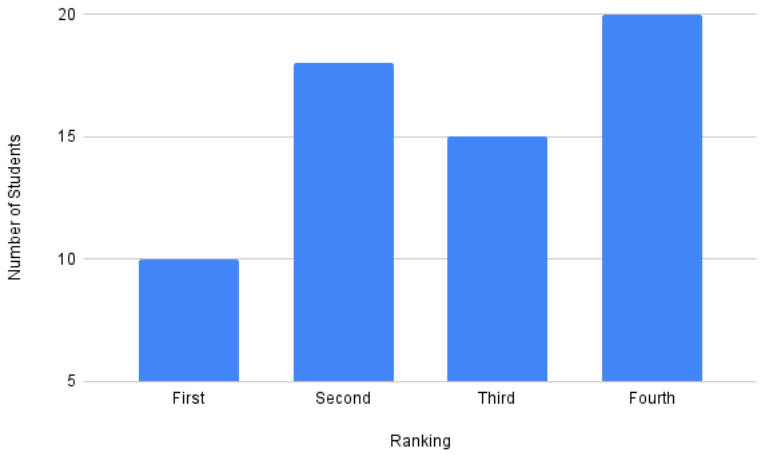
Also, the given data is nominal, which implies that we can arrange the bars in a specific order while plotting and it would be still appropriate for the representation of number of players.

But, a pie chart is used when we are interested in representation of proportion or percentage of players in each of the academy.

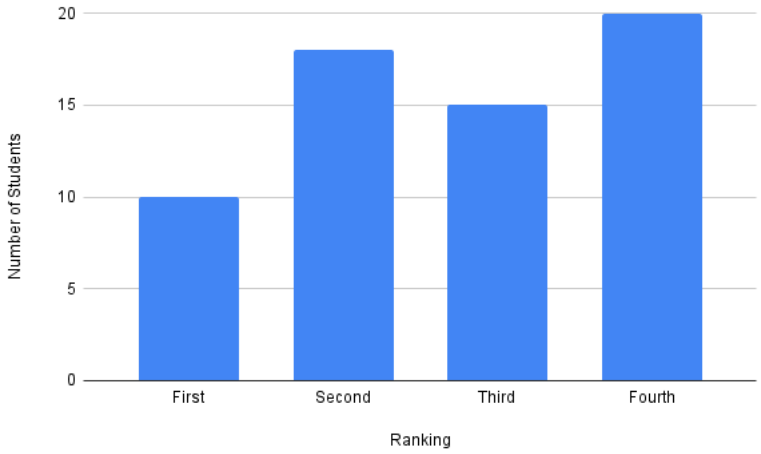
Hence, option (d) is correct.

8. The data of number of students sharing the same rank is collected. Which of the following is/are suitable to represent the collected data?

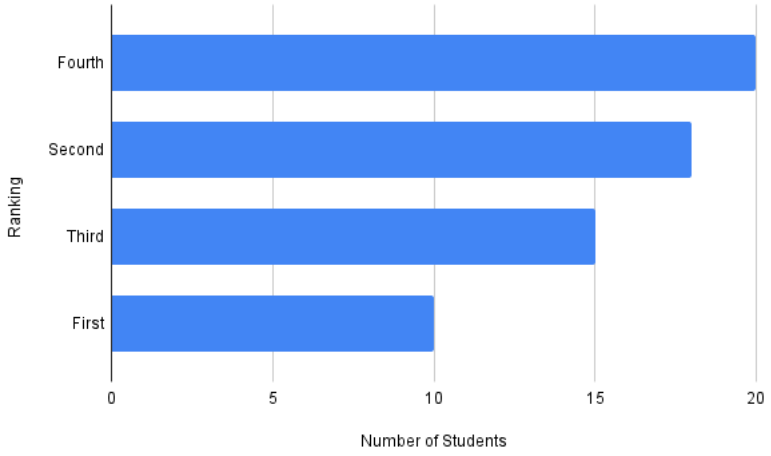
   a.



   b.



c.



Answer: b

Solution:

In option (a), there's a missing baseline resulting in a misleading plot.
In option (c), the order of categories is not retained. Therefore, it's not a suitable representation for the collected data because, in case of ordinal data we preserve the order of categories while plotting.
The option (b) is a good representation of the collected dataset as it is neither misleading nor order of categories is violated.
Hence, option (b) is correct.

9. Choose the correct statement about categorical data:

   (a) Categorical data have measurement units.

   (b) Categorical data can take numerical values, but no meaningful mathematical operations can be performed on it.

   (c) Categorical data is quantitative in nature.

   (d) All of the above

Answer: b

Solution:

Categorical data are also called qualitative variables and it identifies the group membership. Also, we cannot perform any meaningful mathematical operations on it.
Suppose, we have a categorical variable "Gender" with two categories "F" and "M" and we have coded "F" as 1 and "M" as 0. Here, categorical data have taken numerical values, but we cannot perform any meaningful mathematical operation on it.
Also, categorical data does not have any measurement units as it represents only categories or labels.

Hence, from above explanation it is clear that option (b) is correct.

The distribution of grades in a Statistics class consisting of 80 students is shown by a pie chart in Figure 2.2.G. Based on the information given, answer the questions (10) and (11)

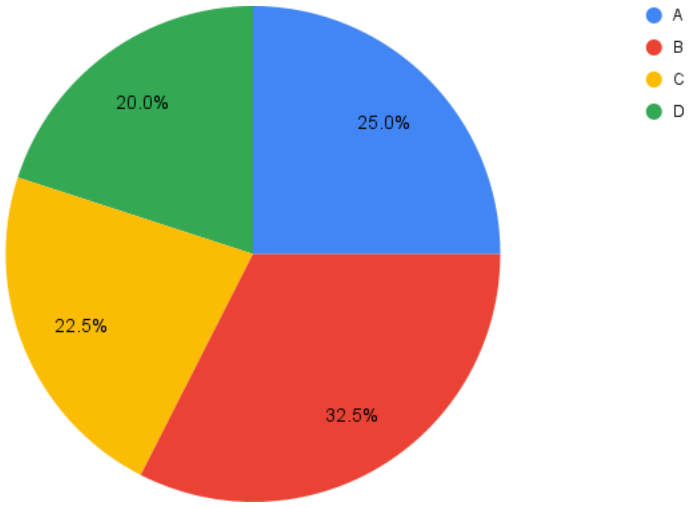


Figure 2.2.G: Distribution of grades in a Statistics class

10. How many students have secured B grade?
    Answer: 26

    Solution:
    Total number of students in the statistics class is 80 and distribution for grade B is 32.5%.
    So, number of students secured B grade is $80 \times \frac{32.5}{100} = \frac{2600}{100} = 26$.

11. What is the ratio of the students secured C grade to the students secured A grade?
    Answer: 0.9

    Solution:
    Total number of students in the statistics class is 80.
    Number of students secured C grade is $80 \times \frac{22.5}{100} = 18$
    Number of students secured A grade is $80 \times \frac{25}{100} = 20$
    Thus, required ratio $\frac{18}{20} = 0.9$

Statistics for Data Science-1

Week-3 Graded Assignment

1. The numbers a, b, c, d have frequencies $(x+6)$, $(x+2)$, $(x-3)$ and x respectively. If their mean is m, find the value of x. (Enter the value as next highest integer)

Solution:

$$\frac{a(x+6)+b(x+2)+c(x-3)+dx}{(x+6)+(x+2)+(x-3)+x} = m$$

$$\frac{ax+6a+bx+2b+cx-3c+dx}{4x+5} = m$$

$$ax+bx+cx+dx+6a+2b-3c = m(4x+5) = (4m)x+5m$$

$$(a+b+c+d-4m)x = 5m-6a-2b+3c$$

$$x = \frac{(5m-6a-2b+3c)}{(a+b+c+d-4m)}$$

Suppose, we substitute values of a, b, c, d and m as 2, 7, 9, 17 and 6.88 respectively, then

$$x = \frac{(5\times 6.88)-(6\times 2)-(2\times 7)+(3\times 9)}{(2+7+9+17-(4\times 6.88)} = 4.73$$

Hence, $x = 5$

The mean and sample standard deviation of the dataset consisting of N observations is m and s respectively. Later it is noted that one observation x is wrongly noted as p. Based on the given information, answer questions (2) and (3).

2. What is the mean of the original dataset? (Correct up to 2 decimal place accuracy)
Solution:
Let the sum of all the observations of noted dataset be T and for the original dataset be T'.

$$Mean = \frac{T}{N} = m$$

$$T = m \times N$$

Therefore, $T' = T - p + x$. Hence, Mean for original dataset= $\frac{T'}{N}$

Suppose, we substitute values of N, m, s, x and p as 8, 13, 8, 18 and 13 respectively.
Let the sum of all the observations of the noted dataset be T and for the original dataset be T'.

$$Mean = \frac{T}{8} = 13$$
$$T = 13 \times 8 = 104$$

Therefore, $T' = T - p + x = 104 - 13 + 18 = 109$.
Hence, Mean for original dataset= $\frac{T'}{N} = \frac{109}{8} = 13.625$

3. What is the sample variance of the original dataset? (Correct up to 2 decimal place accuracy)
Solution:

$$\text{Sample variance, } s^2 = \frac{\Sigma(x_i - \overline{x})^2}{N-1} = \frac{\Sigma(x_i^2 - 2x_i\overline{x} + \overline{x}^2)}{N-1} = \frac{\Sigma x_i^2 - 2\overline{x}\Sigma x_i + N\overline{x}^2}{N-1}$$
$$\Rightarrow s^2 = \frac{\Sigma x_i^2 - 2\overline{x}(N\overline{x}) + N\overline{x}^2}{N-1} = \frac{\Sigma x_i^2}{N-1} - \left(\frac{N\ \overline{x}^2}{N-1}\right)$$

Let Σx_i^2 be equals to A for noted dataset and for the original dataset be equals to B.
So, $B = A - p^2 + x^2$

where, $A = \left(s^2 + \frac{N\ m^2}{N-1}\right) \times (N-1)$

Also, Mean of original dataset= $\frac{T'}{N}$

Hence, sample variance for the original dataset = $\frac{B}{N-1} - \left(\frac{N \times \left(\frac{T'}{N}\right)^2}{N-1}\right)$

$= \frac{B}{N-1} - \frac{T'^{\,2}}{N(N-1)}$

Suppose, we substitute values of N, m, s, x and p as 8, 13, 8, 18 and 13 respectively.

Let Σx_i^2 be equals to A for noted dataset and for the original dataset be equals to B.
So, $A = \left(8^2 + \frac{8 \times 13^2}{7}\right) \times (8-1) = 1800$
Therefore, $B = 1800 - 13^2 + 18^2 = 1955$
Hence, sample variance for the original dataset = $\frac{1955}{8-1} - \frac{109^2}{8 \times 7} = 67.125$

4. Let the data $x_1, x_2, ..., x_n$ represent the retail prices in rupees of a certain commodity in n randomly selected shops in a particular city. What will be the sample variance in the retail prices, if c rupees is added to all the retail prices? (Correct up to 2 decimal place accuracy)

Solution:

Mean $= \dfrac{x_1 + x_2 + ... + x_n}{n}$

If c rupees is added to all the retail prices, then the new prices will be $y_i = x_i + c$; $i = 1, 2, ..., n$

Then, New variance = Old variance.

i.e,

$$\frac{\Sigma(y_i - \overline{y})^2}{n-1} = \frac{\Sigma[(x_i + c) - (\overline{x} + c)]^2}{n-1} = \frac{\Sigma(x_i - \overline{x})^2}{n-1}$$

Suppose the value of n is 6 and the observations are 46, 34, 82, 37, 83, 66, then

Mean $= \dfrac{46 + 34 + 82 + 37 + 83 + 66}{6} = 58$

Sample variance $(s^2) = \dfrac{\Sigma(x_i - \overline{x})^2}{n-1}$

$= \dfrac{(46-58)^2 + (34-58)^2 + (82-58)^2 + (37-58)^2 + (83-58)^2 + (66-58)^2}{5} = 485.2$

Suppose, we have n observations such that $x_1, x_2, ..., x_n$. Based on the given information, answer questions (5), (6), (7):

5. Calculate 10^{th}, 50^{th} and 100^{th} percentiles?

Solution:

To find the sample $100p$ percentiles of a dataset of size n;

(1) Arrange the data in ascending order.

(2) If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.

(3) If np is integer, then the average of the values in positions np and $np + 1$ is the sample $100p$ percentile.

For example,

Let $n = 7$ with observations $31, 36, 25, 34, 115, 108, 88$ and ascending order is $25, 31, 34, 36, 88, 108, 115$ then,

(i) $n = 7$ and $p = 0.1$, then $np = 0.7$.

Therefore, 10^{th} percentile will be 1^{st} observation = 25.

(ii) $n = 7$ and $p = 0.5$, then $np = 3.5$.

Therefore, 50^{th} percentile will be the 4^{th} observation = 36.

(iii) $n = 7$ and $p = 1$, then $np = 7$.
Therefore, 100^{th} percentile will be the last observation = 115.

6. Calculate the Inter Quartile Range (IQR) of the data.
Solution:
To find the sample $100p$ percentiles of a data set of size n;
(1) Arrange the data in ascending order.
(2) If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.
(3) If np is integer, then the average of the values in positions np and $np + 1$ is the sample $100p$ percentile.
For Q_1, $p = 0.25$
And, for Q_3, $p = 0.75$
Therefore, $IQR = Q_3 - Q_1$

For example,
Given, $n = 7$ and $p = 0.25$, then $np = 1.75$
Therefore, $Q_1 = 31$. and
$Q_3 = 75^{th}$ percentile.
Given, $n = 7$ and $p = 0.75$, then $np = 5.25$.
Therefore, $Q_3 = 108$.
Hence, $IQR = Q_3 - Q_1 = 108 - 31 = 77$.

7. How many outliers are there?
Solution:
We know, $IQR = Q_3 - Q_1$.
Outliers $< Q_1 - 1.5 \times IQR$ and Outliers $> Q_3 + 1.5 \times IQR$
For example,
$Q_1 = 25^{th}$ percentile of the data.
Given, $n = 7$ and $p = 0.25$, then $np = 1.75$
Therefore, $Q_1 = 31$. and
$Q_3 = 75^{th}$ percentile.
Given, $n = 7$ and $p = 0.75$, then $np = 5.25$.
Therefore, $Q_3 = 108$.
Hence, $IQR = Q_3 - Q_1 = 108 - 31 = 77$.
Since, Outliers $< Q_1 - 1.5 \times IQR$ and Outliers $> Q_3 + 1.5 \times IQR$
Now, $31 - (1.5 \times 77) = -84.5$ and $108 + (1.5 \times 77) = 223.5$
As there are no observations that satisfies the condition of outliers. Hence, there are no outliers for the given data.

8. In a deck, there are cards numbered 1 to n such that the number of cards of a given number is the same as the number on the card. Which of the following statement(s)

is/are true about the mean and mode of the numbers on this deck of card?

a. Mode is n.
b. Mean is $\frac{2n+1}{3}$.
c. Mode is $n-1$.
d. Mean is n.
e. Mean is $\frac{n+1}{2}$.
f. Mode is not defined for this data.

Answer: a, b

Solution:

Given that the number of cards of a number in the deck is the same as the number on the card. It means that:

Number (x_i)	Frequency (f_i)
1	1
2	2
...	...
...	...
n	n

Table 3.1

Hence, Mode $= n$.

Now, Total number of observations $= f_1 + f_2 + ... + f_n = 1 + 2 + ... + n = \frac{n(n+1)}{2}$

Sum of observations $= f_1x_1 + f_2x_2 + ... + f_nx_n = 1 \times 1 + 2 \times 2 + ... + n \times n$

So, $f_1x_1 + f_2x_2 + ... + f_nx_n = \frac{n(n+1)(2n+1)}{6}$

Therefore, Mean $= \frac{f_1x_1 + f_2x_2 + ... + f_nx_n}{f_1 + f_2 + ... + f_n} = \frac{\frac{n(n+1)(2n+1)}{6}}{\frac{n(n+1)}{2}} = \frac{2n+1}{3}$

Hence, options (a) and (b) are correct.

For example, $n = 42$

Given that the number of cards of a number in the deck is the same as the number on the card, it means that:

Number (x_i)	Frequency (f_i)
1	1
2	2
...	...
...	...
42	42

Table 3.2

Hence, Mode = 42.

Now, Total number of observations $= f_1 + f_2 + ... + f_{42} = 1 + 2 + ... + 42 = \frac{42(42+1)}{2}$

Sum of observations $= f_1x_1 + f_2x_2 + ... + f_{42}x_{42} = 1 \times 1 + 2 \times 2 + ... + 42 \times 42$

So, $f_1x_1 + f_2x_2 + ... + f_{42}x_{42} = \frac{42(42+1)(2(42)+1)}{6}$

$$\text{Mean} = \frac{f_1x_1 + f_2x_2 + ... + f_{42}x_{42}}{f_1 + f_2 + ... + f_{42}} = \frac{\frac{42(42+1)(2(42)+1)}{6}}{\frac{42(42+1)}{2}} = \frac{2(42)+1}{3}$$

Hence, Mean = 28.33

Figure 3.1.G shows a stem and leaf plot of the ratings (out of 100) of an actor's performance in different movies. Based on the given information, answer questions (9) and (10).

Stem	Leaf
5	3 9
7	2 2 5 8
8	7 7 7
9	9

Here 6 | 4 represents rating of 64.

Figure 3.1.G

9. What is the Inter Quartile Range (IQR) (Correct up to 1 decimal point accuracy)?
Solution:
To find the sample $100p$ percentiles of a data set of size n;
(1) Arrange the data in ascending order.
(2) If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.
(3) If np is integer, then the average of the values in positions np and $np + 1$ is the sample $100p$ percentile.

For Q_1, $p = 0.25$
And, for Q_3, $p = 0.75$
Therefore, $IQR = Q_3 - Q_1$

For example, $n = 10$
Number of observation; n = 10
$Q_1 = \left(\frac{10}{4}\right)^{th}$ observation $= 3^{rd}$ observation $= 72$
$Q_3 = \left(\frac{30}{4}\right)^{th}$ observation $= 8^{th}$ observation $= 87$

Therefore, $IQR = Q_3 - Q_1 = 87 - 72 = 15$

10. What is the median rating, if x points are added to all of his ratings and then converted to y points? (Correct up to 2 decimal point accuracy)
Solution:
There are 10 observations in the data. So, the Median of the given data will be the mean of 5^{th} and 6^{th} observation.

Median of given data $= \frac{75 + 78}{2} = 76.5$
Now, if x points are added to all of his ratings, the median becomes $76.5 + x$.
And, for conversion to y points, we have to multiply all the observations by $\frac{y}{100}$. Hence, the median for converted data $= (76.5 + x) \times \frac{y}{100}$.
Therefore, option b is correct.

Suppose, we substitute values of x and y as 3 and 40 respectively.
There are 10 observations in the data. So, the median of the given data will be the mean of 5^{th} and 6^{th} observation.

Median of given data $= \frac{75 + 78}{2} = 76.5$
Now, if 3 points are added to all of his ratings, the median becomes $76.5 + 3 = 79.5$.
And, for conversion to 40 points, we have to multiply all the observations by $\frac{40}{100}$.
Hence, the median for converted data $= (76.5 + 3) \times \frac{40}{100} = 31.8$.

Statistics for Data Science-1

Week-4 Graded Assignment Solution

The phone brands OnePlus, Vivo and Oppo are owned by BBK Electronics. Table 4.1.G represents the data for the sales (in Lakhs) of OnePlus and BBK Electronics by different dealers in Chennai and Punjab in the year 2010. Based on the given information, answer questions (1), (2), (3), (4), (5) and (6).

Dealer's Location	OnePlus	BBK Electronics
Chennai	a	b
Punjab	c	d
Chennai	e	f
Punjab	g	h
Chennai	i	j
Punjab	k	l
Chennai	m	n

Table 4.1.G

1. What is the population standard deviation of sales of OnePlus?(Enter the answer correct to 2 decimal accuracy)
   Solution:
   Let m_x and σ_x be the mean and population standard deviation of sales of OnePlus respectively.
   $m_x = \frac{a+c+e+g+i+k+m}{7}$
   $\sigma_x^2 = \frac{(a-m_x)^2+(c-m_x)^2+(e-m_x)^2+(g-m_x)^2+(i-m_x)^2+(k-m_x)^2+(m-m_x)^2}{7}$
   Therefore, Population standard deviation of sales of OnePlus= $\sqrt{\sigma_x^2}$

2. What is the sample standard deviation of sales of BBK Electronics?(Enter the answer correct to 2 decimal accuracy)
   Solution:
   Let m_y and S_y be the mean and sample standard deviation of sales of BBK Electronics respectively.
   $m_y = \frac{b+d+f+h+j+l+n}{7}$
   $S_y^2 = \frac{(b-m_y)^2+(d-m_y)^2+(f-m_y)^2+(h-m_y)^2+(j-m_y)^2+(l-m_y)^2+(n-m_y)^2}{7-1}$
   Therefore, Standard deviation of sales of BBK Electronics= $\sqrt{S_y^2}$

3. What is the sample co-variance between the sales of OnePlus and BBK Electronics?(Enter the answer correct to 2 decimal accuracy)

Solution:
Let X= Sales of OnePlus
Y= Sales of BBK Electronics.

Therefore, $Cov(X,Y) = \frac{\sum_{i=1}^{7}(x_i - m_x)(y_i - m_y)}{7-1}$
$= \frac{(a-m_x)(b-m_y)+(c-m_x)(d-m_y)+(e-m_x)(f-m_y)+(g-m_x)(h-m_y)}{6}$
$+ \frac{(i-m_x)(j-m_y)+(k-m_x)(l-m_y)+(m-m_x)(n-m_y)}{6}$

4. What is the correlation coefficient between the sales of OnePlus and BBK Electronics?(Enter the answer correct to 2 decimal accuracy)
Solution:
Let r be the correlation coefficient between the sales of OnePlus and BBK Electronics.
Therefore,

$$r = \frac{Cov(X,Y)}{S_x \times S_y}$$

where,
Cov(X,Y)= Sample Covariance between sales of OnePlus and BBK Electronics.
S_x= Sample standard deviation of sales of OnePlus=$\sqrt{\frac{n}{n-1}}\sigma_x = \sqrt{\frac{7}{6}}\sigma_x$
S_y= Sample standard deviation of sales of BBK Electronics.

5. What can you say about the linear relationship between the sales of OnePlus and BBK Electronics?(More than one option can be correct)

   (a) Strong
   (b) Positive
   (c) Weak
   (d) Negative
   (e) Absence of linear relationship
   (f) Moderate

Solution: x
If $0.75 \leq r \leq 1$, then there is a Strong and positive linear relationship between the sales of OnePlus and BBK Electronics.
If $0.5 \leq r < 0.75$, then there is a Moderate and positive linear relationship between the sales of OnePlus and BBK Electronics.
If $0.25 \leq r < 0.5$, then there is a Weak and positive linear relationship between the sales of OnePlus and BBK Electronics.
If $-0.25 < r < 0.25$, then there is absence of linear relationship between the sales of

OnePlus and BBK Electronics.
If $-0.5 < r \leq -0.25$, then there is a Weak and negative linear relationship between the sales of OnePlus and BBK Electronics.
If $-0.75 < r \leq -0.5$, then there is a Moderate and negative linear relationship between the sales of OnePlus and BBK Electronics.
If $-1 \leq r \leq -0.75$, then there is a Strong and negative linear relationship between the sales of OnePlus and BBK Electronics.

6. Is the sales of OnePlus strongly influenced by the location of dealer?

   a. Yes

   b. No

   Solution:
   Let Chennai be represented by 0 and Punjab by 1.
   The Point Bi-serial correlation coefficient between the sales of OnePlus and the location of dealer is given by:
   $r_{pb} = \dfrac{(\bar{Y_0} - \bar{Y_1})}{\sigma_x} \times \sqrt{p_0 \times p_1}$
   where, p_0 = Proportion of Chennai dealers $= \dfrac{4}{7}$
   p_1 = Proportion of Punjab dealers $= \dfrac{3}{7}$
   $\bar{Y_0}$ = Mean sales of OnePlus in Chennai $= \dfrac{a+e+i+m}{4}$
   $\bar{Y_1}$ = Mean sales of OnePlus in Punjab $= \dfrac{c+g+k}{3}$
   σ_x = Population standard deviation of sales of OnePlus

   $$r_{pb} = \frac{(\bar{Y_0} - \bar{Y_1})}{\sigma_x} \times \sqrt{\frac{4}{7} \times \frac{3}{7}} = \frac{(\bar{Y_0} - \bar{Y_1})}{\sigma_x} \times 0.49487$$

   If $0.75 \leq |r_{pb}| \leq 1$, then the sales of OnePlus is strongly influenced by the location of dealer.
   Else, the sales of OnePlus is not strongly influenced by the location of dealer.

N college students are classified according to their intelligence level and economic conditions and the results are given in Table 4.2.G.

Economic Conditions	Intelligence level			
	Bright	Average	Dull	Borderline
Good	a	b	c	d
Poor	e	f	g	h

Table 4.2.G

Based on the given information answer questions (7), (8), (9), (10) and (11).

7. What proportion of total students are dull?(Enter the answer correct to 2 decimal accuracy)
Solution:
Proportion of dull students $= \dfrac{c+g}{a+b+c+d+e+f+g+h}$.

8. What proportion of total students are poor economic conditions?(Enter the answer correct to 2 decimal accuracy)
Solution:
Proportion of students having poor economic conditions $= \dfrac{e+f+g+h}{a+b+c+d+e+f+g+h}$.

9. What proportion of students of good economic conditions are borderline?(Enter the answer correct to 2 decimal accuracy)
Solution:
Required proportion $= \dfrac{d}{a+b+c+d}$

10. What percentage of bright students are in poor economic conditions?(Enter the answer correct to 2 decimal accuracy)
Solution:
Required percentage $= \left(\dfrac{e}{a+e}\right) \times 100$

11. What percentage of average students are in good economic conditions?(Enter the answer correct to 2 decimal accuracy)
Solution:
Required percentage $= \left(\dfrac{b}{b+f}\right) \times 100$

Statistics for Data Science - 1

Sample Qualifier

1. A statistician, who is not good at math, wants to represent the relative frequencies of the 10 different categories of a categorical variable in a pie chart. He calculated the relative frequency of each category. In order to make a pie chart representing categories, he calculated the angle of slices for the first 9 categories using the wrong formula $\pi * r_i$ radians, where r_i represents the relative frequency of the i^{th} category. Since he knows that the sum of angles of slices must be 2π radians, he calculated the angle of the tenth slice as $2\pi - \pi * \sum_{i=1}^{9} r_i$.
   If the total frequency is equal to 500 and the angle of the tenth category using the above formula is 1.02π radians in the pie chart, then what is the actual frequency of the tenth category?
   Answer: 10 [3 marks]

Given 10^{th} category angle $= 1.02\pi$ radians.
10^{th} category angle from formula $= 2\pi - \pi * \sum_{i=1}^{9} r_i$.
$\Rightarrow 2\pi - \pi * \sum_{i=1}^{9} r_i = 1.02\pi$
$\Rightarrow \pi * \sum_{i=1}^{9} r_i = (2 - 1.02)\pi$

$$\Rightarrow \sum_{i=1}^{9} r_i = 0.98 \quad (1)$$

We know that sum of relative frequencies is equal to 1.
$\Rightarrow \sum_{i=1}^{10} r_i = 1$

$$\Rightarrow \sum_{i=1}^{9} r_i + r_{10} = 1 \quad (2)$$

Substituting equation (1) in equation (2), we will get
$\Rightarrow \sum_{i=1}^{9} r_i + r_{10} = 1$
$\Rightarrow 0.98 + r_{10} = 1$
$r_{10} = 0.02$

Given total frequency $= 500$.
Therefore frequency of 10^{th} category is $r_{10} * 500$
Substituting r_{10} value, we will get frequency of 10^{th} category as $0.02 \times 500 = 10$

2. The grade points achieved by the students in a Statistics exam is represented using a bar chart in Figure S.1. What is the box and whisker plot of the data given in Figure S.1? [3 marks]

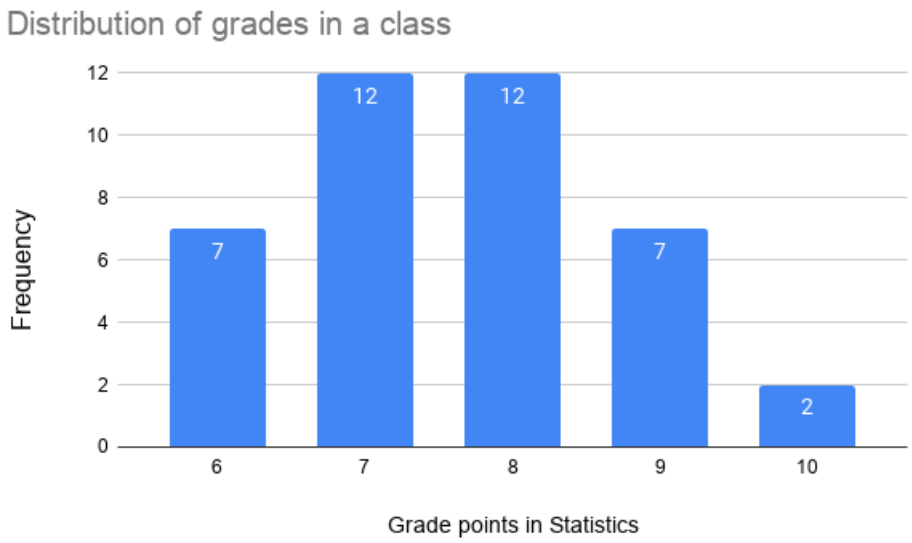
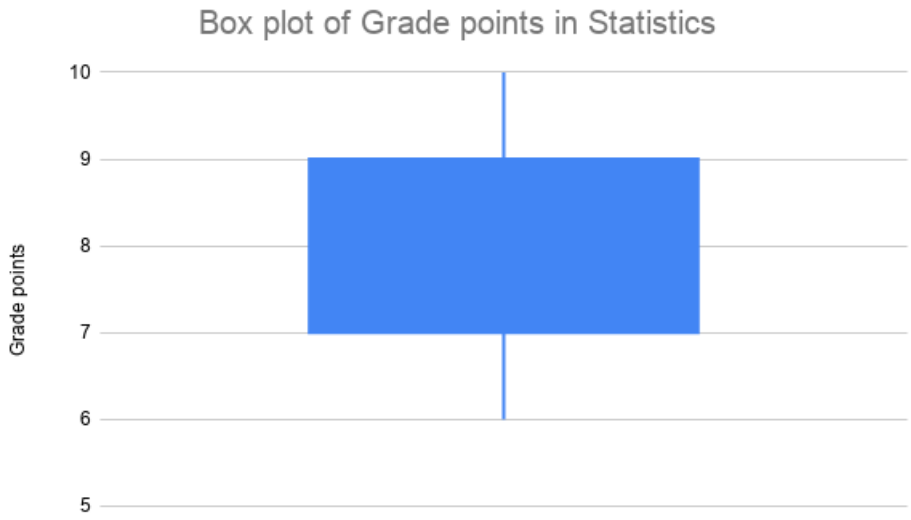
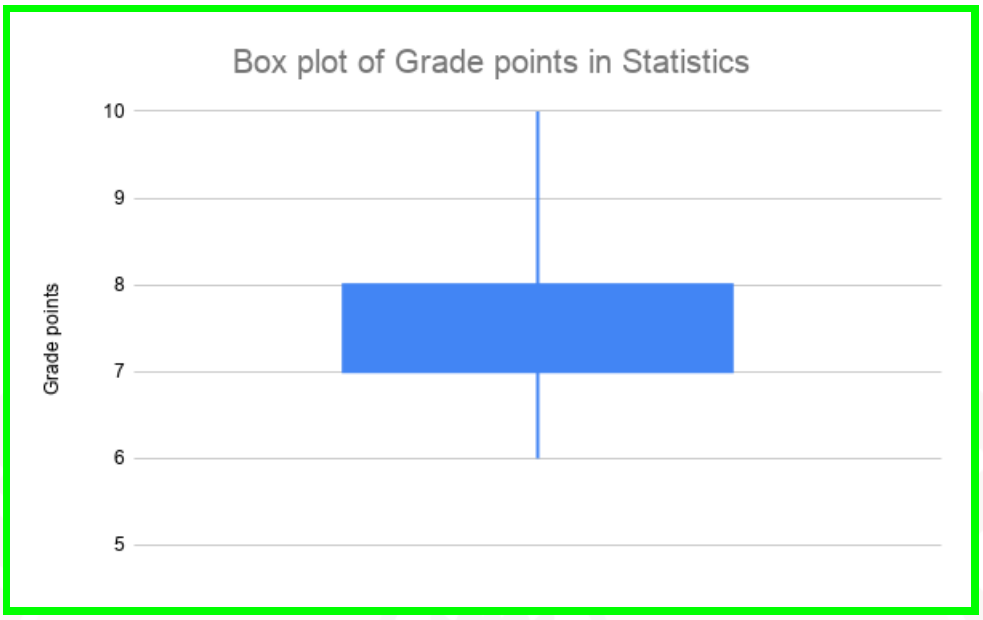


Figure S.1: Distribution of grade points in the Statistics exam dataset.

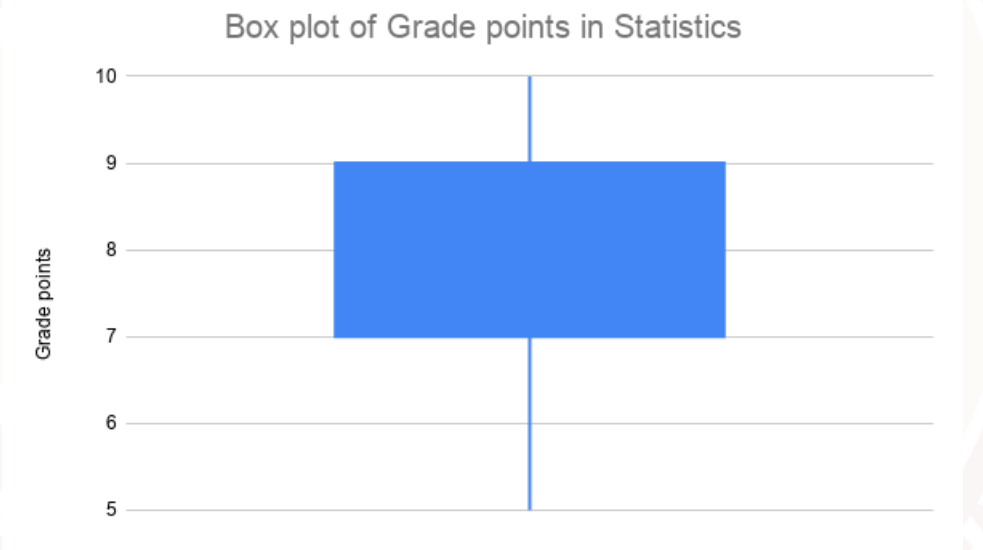


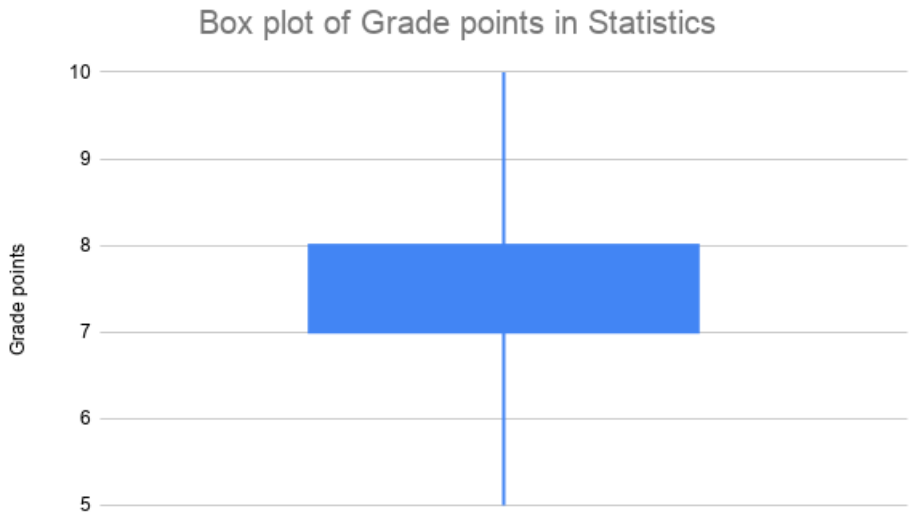
(a)

(b)


Box plot of Grade points in Statistics
Grade points
10
9
8
7
6
5

(c)


Box plot of Grade points in Statistics
Grade points
10
9
8
7
6
5



(d)

Answer: b

Solution:

Constructing a frequency table from the given bar chart, we get,

Grade point	Frequency
6	7
7	12
8	12
9	7
10	2

Grade points dataset

Let total number of students be n.
$n = 7 + 12 + 12 + 7 + 2 = 40$.
For box plot, we need to find the first quartile, third quartile, minimum, maximum, and median values.
Minimum grade point = 6
Maximum grade point = 10
First quartile(Q_1) is 25^{th} percentile value.
n= 40, p= 0.25
$\Rightarrow$ np= 10

Therefore, Q_1 is the average of the 10th and the 11th observation in an ascending ordered dataset.
Therefore $Q_1 = 7$.

Third Quartile (Q_3) is the 75^{th} percentile value.
n=40, p= 0.75 $\Rightarrow$ np= 30
Therefore, Q_3 is the average of the 30th and the 31st observation in an ascending ordered dataset.
Therefore $Q_3 = 8$.

Now, minimum = 6, Q_1 =7, Q_3=8, maximum =10.
Based on the above, the correct answer is option (b).

Use the following information and data given in Table S.1 to answer the questions 3 and 4
In an organization, data from a sample of 10 employees is collected. This data includes gender, age, and salary of employees and is given in Table S.1.

S.No	Gender	Age (in years)	Salary (in ₹lakhs)
1	F	27	7
2	F	33	8
3	F	39	8
4	F	35	13
5	M	46	25
6	M	49	32
7	M	53	35
8	M	66	38
9	M	60	40
10	M	54	44

Table S.1: Employee dataset

3. Which of the following is true about salary of employees whose data was collected? [5 marks]

   (a) The mean of the salary is approximately ₹25 lakhs.
   (b) The sample standard deviation of salary is ₹16.71 lakhs approximately.
   (c) The sample variance of salary is 216.66 (lakh rupee)^2 approximately.
   (d) Interquartile range of salary is equal to ₹30 lakhs.

Solution:
$n = 10$
The mean of the salary is

$$\frac{\sum_{i=1}^{10} x_i}{10}$$

$$\Rightarrow \frac{7+8+8+13+25+32+35+38+40+44}{10} = 25$$

Therefore, mean is 25.
Sample variance

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

$$\Rightarrow s^2 = \frac{(7-25)^2 + (8-25)^2 + (8-25)^2 + (13-25)^2 + (25-25)^2 + (32-25)^2 + (35-25)^2}{9} + \frac{(38-25)^2 + (40-25)^2 + (44-25)^2}{9}$$

$$\Rightarrow s^2 = 216.66(lakhrupee)^2 \text{ approximately}$$
$$\Rightarrow s = \sqrt{216.66} = 14.719 \text{ lakh rupees}$$

Q_1 is 25th percentile value.
n=10, p=0.25 $\Rightarrow$ np=2.5
Q_1 is 3rd observation in the ascending ordered data.
Therefore, $Q_1 = 8$

Q_3 is 75th percentile value.
n=10, p=0.75 $\Rightarrow$ np=7.5
Q_3 is the 8th observation in ascending ordered data. Therefore, $Q_3 = 38$

Interquartile range $= Q_3 - Q_1 = 38 - 8 = 30$.
Therefore, options (a), (c), and (d) are correct.

4. What is the absolute value of point bi-serial correlation coefficient between gender and salary? Enter the answer up to 2 decimals accuracy. [3 marks]
Answer:0.935 accepted range 0.87 to 0.96

Solution:
Point bi-serial correlation coefficient formula (r) is

$$\frac{(\bar{Y_0} - \bar{Y_1})\sqrt{p_0 \times p_1}}{\sigma}$$

$$\text{Population standard deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

$$\text{Population standard deviation } (\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} * \frac{n-1}{n}}$$

$$\sigma = s \times \sqrt{\frac{n-1}{n}}$$

Substituting s and n we get,

$$\sigma = 14.719 \times \sqrt{\frac{9}{10}} = 13.96 \text{ approximately}$$

Let females be encoded as 1 and males be encoded as 0.
Therefore,

$$\bar{Y_0} = \frac{25 + 32 + 35 + 38 + 40 + 44}{6} = 35.666$$

and

$$\bar{Y_1} = \frac{7+8+8+13}{4} = 9$$

$$p_0 = \frac{6}{10} = 0.6, p_1 = \frac{4}{10} = 0.4$$

$$\Rightarrow r = \frac{(\bar{Y_0} - \bar{Y_1})\sqrt{p_0 \times p_1}}{\sigma}$$

$$\Rightarrow r = \frac{(35.66 - 9)\sqrt{0.6 \times 0.4}}{13.96} = 0.9355$$

5. Table S.2 shows the outcomes obtained in 250 rolls of a die.

Outcome	Frequency	Relative frequency
1	x	
2	10	
3	y	0.2
4	35	
5	z	
6	60	

Table S.2: Frequency table

The difference between the relative frequency of value 1 and the relative frequency of value 5 is 0.06. Find the median of the outcomes of rolling the die. [3 marks]

4

Solution:

Total number of observations, $n = 250$

Let r_i and f_i be the relative frequency and frequency respectively for the outcome i of a die.

$r_3 = 0.2$

$f_3 = 0.2 \times 250 = 50$

We know that

$$\sum_{i=1}^{6} f_i = 250$$

$$\Rightarrow x + 10 + 50 + 35 + z + 60 = 250$$

$$\Rightarrow x + z = 95$$

Since, it is given that $r_1 - r_5 = 0.06$.

Therefore,

$$r_1 = r_5 + 0.06$$

$$\Rightarrow f_1 = f_5 + 0.06 * 250$$

We know that $f_1 = x, f_5 = z$

$$\Rightarrow x = z + 15$$

Substituting $z + 15$ in place of x, we get

$$2z + 15 = 95 \Rightarrow z = 40, x = 55$$

$n = 250$

Median is the average of the 125th and 126th observation in ascending ordered dataset. 125th and 126th observations are outcome 4 in ordered dataset. So, the median of the outcomes is 4.

6. Choose the correct statements among the following. [3 marks]

   (a) While the range and variance are affected by outliers, the interquartile range is not.

   (b) The range of the sample dataset is never greater than the range of the population.

   (c) Median is more affected by outliers than mean.

   (d) The units of the variance of a variable is same as the units of the variable.

Solution:
Range = maximum observation - minimum observation.
The presence of outliers affects the maximum and minimum, so it is affected by outliers. Interquartile range is the difference between the 75th percentile and the 25th percentile, so it is not affected by extreme values (outliers). Hence, option (a) is correct.

Sample is a subset of the population. So, the minimum and maximum values of the sample will also be there in the population. There will also be some observations that is in the population but will not be in the sample. So, the range of the sample dataset is never greater than the range of the population. Hence option (b) is correct.

Median is the 50th percentile or mid value, it is less affected by outliers while mean, which is the average of all observations, is more affected by outliers as outliers are included in the calculation of mean. Hence option (c) is incorrect.

The units of the variance of a variable is square of the units of the variable. Hence, option (d) is incorrect.

7. A group of 10 friends have an average of 8 fruits. Then five friends left the group with some fruits. The remaining friends have an average of 6 fruits. How many fruits did the friends who left the group take with them? [2 marks]

Answer : 50

Total number of friends = 10
Average number of fruits = 8
Let x_1, x_2,$\cdots$,x_{10} be the number of fruits each of the 10 friends have.
Now, from the definition of mean

$$\frac{x_1 + x_2 + x_3 + \cdots + x_{10}}{10} = 8$$

Therefore, total number of fruits are 80.
Now, 5 friends left the group and remaining 5 friends have an average of 6 fruits.
Therefore,

$$\frac{x_6 + x_7 + \cdots + x_{10}}{5} = 6$$
$$x_6 + x_7 + \cdots + x_{10} = 30$$

Therefore, total number of fruits remaining 5 friends have is 30
Hence, number of fruits taken by the group of friends who left is $80 - 30 = 50$.

8. If each value of the dataset 12, 15, 18, 18, 27, 32 is increased by 7, then which numerical summaries will not change? [3 marks]

(a) Mean
(b) Median
(c) Mode
(d) Range
(e) Standard deviation

Answer: Multiple Select Question: D, E

As per the properties of mean, if we add a constant to each data value,

New mean = Old mean + constant

So, mean will change after addition of a constant, hence option (a) is incorrect.
As per the properties of median, if we add a constant to each data value,

New median = Old median + constant

So, median will change after addition of a constant, hence option (b) is incorrect.
As per the properties of mode, if we add a constant to each data value,

$$\text{New mode} = \text{Old mode} + \text{constant}$$

So, mode will change after addition of a constant, hence option (c) is incorrect.
Range = Maximum data value - Minimum data value
Now, if we add a constant to maximum and minimum value then,
New range = (Maximum data value + constant) - (Minimum data value + constant)
= Maximum data value - Minimum data value
= Old range
Therefore, range is not affected by addition of a constant.
The formula for population standard deviation is,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{5}(x_i - \bar{x})^2}{n}} \tag{3}$$

The mean is changed by addition of a constant as discussed earlier and each data value is also increased by the same amount.
Now, the new population standard deviation is,

$$\sigma_{new} = \sqrt{\frac{\sum_{i=1}^{5}((x_i + constant) - (\bar{x} + constant))^2}{n}} \tag{4}$$

From equation (3) and equation (4), $\sigma = \sigma_{new}$
Therefore, standard deviation does not change by the addition of a constant.

9. Table S.3 gives the heights and weights of eight friends.

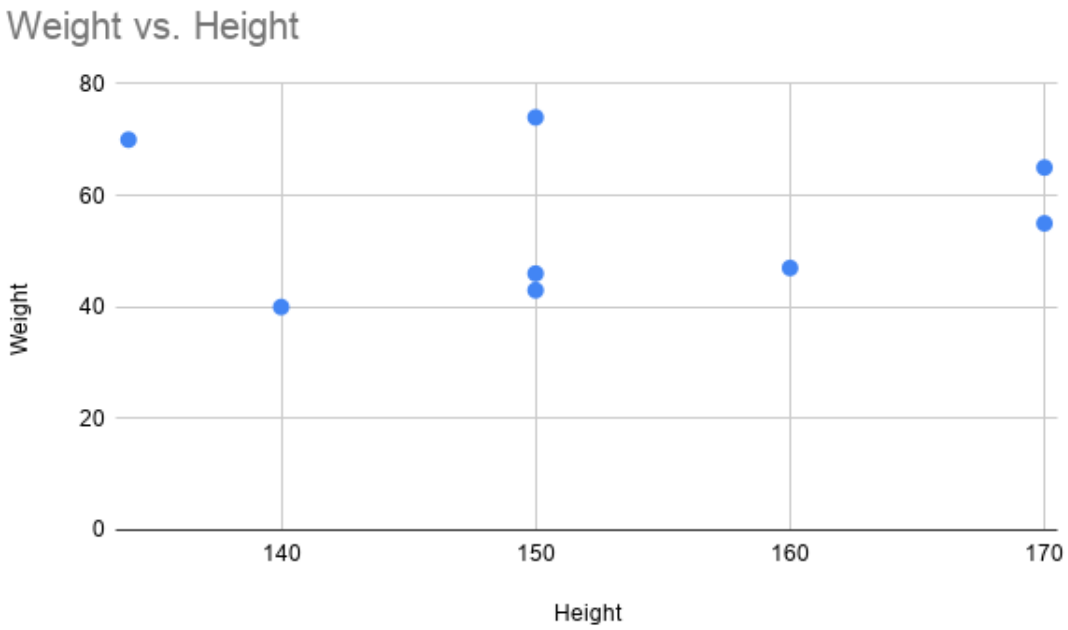
Name	Height(cm)	Weight(kg)
Anmol	140	40
Sujata	150	43
Subashish	170	55
Deepti	134	70
Rajesh	150	74
Kalpana	160	47
Nagarjuna	170	65
Shruti	150	46

Table S.3: Height and weight dataset

Which one of the following best describes the correlation between their heights and weights? [3 marks]

a) Low negative correlation
b) Low positive correlation
c) High negative correlation
d) High positive correlation

To find the solution of this answer, we can use both analytical as well as graphical method. But for such kind of problems, graphical method is faster.



From the scatter plot, it is clearly understood that there is a low positive correlation.

10. Nitin did a survey of the number of bikes owned by his friends, the result of which is represented in Table S.4.

Number of bikes owned	Frequency
0	5
1	8
2	4
3	2
4	1

Table S.4: Number of bikes

What is the sample standard deviation of the number of bikes owned by his friends? Enter the answer upto 2 decimal place accuracy.
Answer : 1.128 (Accepted range: 1.1 -1.2)

The mean for the given data is given by

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i * f_i}{\sum_{i=1}^{5} f_i} \tag{5}$$

Therefore, mean is

$$\frac{0*5+1*8+2*4+3*2+4*1}{5+8+4+2+1} = 1.3$$

Now, the sample variance is,

$$s^2 = \frac{\sum_{i=1}^{5} f_i * (x_i - \bar{x})^2}{(\sum_{i=1}^{5} f_i) - 1} \tag{6}$$

Therefore,

$$\frac{5*(0-1.3)^2+8*(1-1.3)^2+4*(2-1.3)^2+2*(3-1.3)^2+1*(4-1.3)^2}{5+8+4+2+1-1} = 1.273$$

The sample standard deviation is

$$s = \sqrt{1.273} = 1.128 \tag{7}$$

11. Figure S.2 shows the distribution of the household items expenditures used in a house throughout the year. Based on this information, choose the correct option(s) from below. [4 marks]

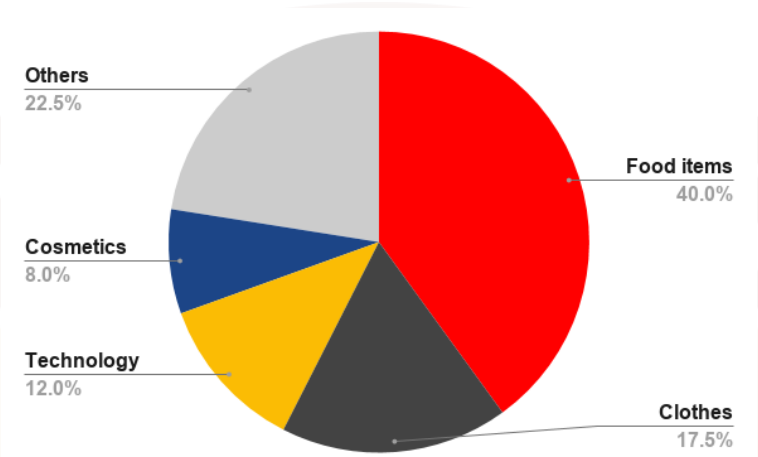


Figure S.2: House budget

(a) If the budget of the house is ₹30,000, then the expenditure on food items is 50% less than the rest of the expenditure.

(b) If the budget of the house is ₹30,000, then the expenditure on food items is approximately 33.3% less than the rest of the expenditure.

(c) If the budget of the house is ₹32,500, then the expenditure on technology is ₹3,900.

(d) If the budget of the house is ₹32,500, then the expenditure on clothes is ₹5,500.

(e) Cosmetics and Food items expenditures are in the ratio of 1:5.

The pie chart shown in Figure S.2 gives the information about percentage relative frequency of different household items expenditures throughout the year.
From the pie chart, the expenditure on food is 40 %. If the total expenditure is ₹30,000, then the food expenditure will be,

$$\frac{40}{100} \times 30,000 = 12,000 \tag{8}$$

So, the rest of the expenditure = 30,000 - 12,000 = ₹18,000

$$\frac{12,000 - 18,000}{18,0000} \times 100 = -33.33 \tag{9}$$

From above equation, it is clear that the expenditure on food items is less than 33.33 % than the rest of the expenditure.

Negative sign indicates that the expenditure on food items is less than rest of the expenditure.
Hence, option (a) is incorrect and option (b) is correct.
The percentage share of expenditure on technology is 12 %. If the budget is 32,500 then the expenditure on technology will be

$$\frac{12}{100} \times 32,500 = 3,900 \tag{10}$$

Therefore, option (c) is the correct option.
The percentage share of expenditure on clothes is 17.5 %. If the budget is 32,500 then the expenditure on technology will be

$$\frac{17.5}{100} \times 32,500 = 5,687.5 \tag{11}$$

Hence, option (d) is incorrect option.
The percentage share of expenditure on cosmetics is 8 %. If the budget is 30,000 then the expenditure on technology will be

$$\frac{8}{100} \times 30,000 = 2,400 \tag{12}$$

From equation (8) and equation (12), the expenditure on food items ₹12,000 and the expenditure on cosmetics is ₹2,400.
Therefore the ratio of expenditures on cosmetics and food items is

$$\frac{2,400}{12,000} = \frac{1}{5}$$

Therefore, option (e) is the correct option.

12. The mean and median of five non zero natural numbers are 8 and 6 respectively. 15 is the only mode of these five non zero natural numbers. What will be the range and the interquartile range of these five non zero natural numbers? [3 marks]

(a) 14 and 12

(b) 20 and 12

(c) 20 and 8

(d) 14 and 8

Let x_1, x_2, x_3, x_4, and x_5 be five non- zero natural numbers.
Median of these data points is given to be 6.
Let these numbers are sorted in ascending order.
Therefore, x_3 will be equal to 6.
Also, mode of the data is given as 15 and it is unimodal. Therefore, 15 must occur more than once in the dataset.
15 is greater than 6 and hence last two data points, x_4 and x_5 will be 15.
Now, mean for the given data is 8.

$$\frac{x_1 + x_2 + 6 + 15 + 15}{5} = 8$$

Therefore,

$$x_1 + x_2 = 4 \tag{13}$$

There are only two possibilities for equation (13) because x_i can take only non-zero natural numbers.
$x_1 = 1$ and $x_2 = 3$ or $x_1 = 2$ and $x_2 = 2$
But, the only mode for given data is 15 and hence, $x_1 = 2$ and $x_2 = 2$ is not possible for given data.
Therefore, $x_1 = 1$ and $x_2 = 3$
Now 1, 3, 6, 15, 15 are the data points.
Therefore, $Range = 15 - 1 = 14$
To find interquartile range, we need to find the first quartile and third quartile.
n = 5
$Q_1 = 0.25 * 5 = 1.25$
The next highest integer is 2 and hence, $Q_1 = 3$.
$Q_3 = 0.75 * 5 = 3.75$
The next highest integer is 4 and hence, $Q_1 = 15$.
Therefore,
$IQR = Q_3 - Q_1 = 15 - 3 = 12$
Hence, option (a) is correct.

13. The average marks of all the students of four sections A, B, C, D taken together is 60 while the average marks of students of each sections is 45, 50, 72, and 80 respectively. If the average marks of sections A, and B together is 48 and of B and C together is 60 and the number of students in section D is 35, then find the number of students in section B. [3 marks]

Answer: 70

Solution:

Let the total marks obtained by students in section A, B, C, and D be a, b, c, and d respectively and total number of students in the respective sections be n_1, n_2, n_3, and n_4.
Let the total number of students be n. Therefore

$$\frac{a+b+c+d}{n} = 60 \tag{14}$$

Since the average marks of students of section A, B, C, and D is $45, 50, 72$, and 80 respectively. Therefore,

$$\frac{a}{n_1} = 45$$
$$\frac{b}{n_2} = 50$$
$$\frac{c}{n_3} = 72$$
$$\frac{d}{n_4} = 80$$

Since the average marks of section A and B is 48 and of section B and C is 60.
Therefore

$$\frac{a+b}{n_1+n_2} = 48$$
$$\Rightarrow \frac{45n_1+50n_2}{n_1+n_2} = 48$$
$$\Rightarrow 45n_1+50n_2 = 48n_1+48n_2$$
$$\Rightarrow 3n_1 = 2n_2$$
$$\Rightarrow n_1 = \frac{2n_2}{3} \tag{15}$$

and

$$\frac{b+c}{n_2+n_3} = 60$$
$$\Rightarrow \frac{50n_2+72n_3}{n_2+n_3} = 60$$

$$\Rightarrow 50n_2 + 72n_3 = 60n_2 + 60n_3$$
$$\Rightarrow 12n_3 = 10n_2$$
$$\Rightarrow n_3 = \frac{5n_2}{6} \tag{16}$$

Given number of students in section D is 35, therefore $n_4 = 35$.
Hence, total marks obtained by all students in section D is $80 \times 35 = 2800$.

From (14),

$$\frac{48n_1 + 48n_2 + 72n_3 + 2800}{n_1 + n_2 + n_3 + n_4} = 60$$

From (15) and (16),

$$32n_2 + 48n_2 + 60n_2 = 60(n_1 + n_2 + n_3 + n_4) - 2800$$
$$\Rightarrow 140n_2 = 60(\frac{15n_2}{6} + 35) - 2800$$
$$\Rightarrow -10n_2 = -700$$
$$\Rightarrow n_2 = 70$$

Therefore, number of students in section B is 70.

14. Based on the data published in the Statistical Hand Book (SHB) – 2020 by the Department of Economics and Statistics, Government of Tamil Nadu, the average rainfall recorded in Tamil Nadu during the period 2005-06 to 2018-19 are as follows:(all values are given in mm) 1034.6, 1078.9, 1304.1, 859.7, 1164.8, 1023.1, 937.8, 1165.1, 937.1, 743.1, 790.6, 987.9, 1138.8, 598.1

What scale should we use in stem and leaf plot such that there are exactly 9 stems in the plot?

Note: Also include the stems which do not have leaves. [2 marks]

(a) 0 | 5981 means 598.1 mm

(b) 05 | 981 means 598.1 mm

(c) 059 | 81 means 598.1 mm

(d) 0598 | 1 means 598.1 mm

Solution:

The first option 0 | 5981 means 598.1 mm, according to this we will have two stems. The minumum value is 0 | 5981, maximum value is 1 | 3041.

The second option is 05 | 981 means 598.1 mm. The stems are 05 |, 06 | , 07 |, 08 |, 09 | , 10 | , 11 | , 12 | , and 13 |. There are total of 9 stems. So, option (b) is correct.

The third option is 059 | 81 means 598.1 mm. We will have stems more than 9 in this way. So, option (c) is incorrect.

The fourth option is 0598 | 1 means 598.1 mm. It has more stems than 9. So, option (d) is incorrect.

15. A supermarket mailed 3020 uniquely identifiable coupons to homes in local residential communities. The number of coupons that were redeemed for each of the next six weeks was counted and shown in Figure S.3. Based on this information, choose the correct option(s) from below. [2 marks]

Figure S.3: Number of coupon redeemed in six week

(a) Given data is time series data.

(b) Given data is cross sectional data.

(c) Number of coupons redeemed is a continuous variable.

(d) Median of the given data is week 5.

(e) Relative frequency of number of coupons redeemed in week 2 is 360.

If the data varies with respect to time for a particular entity in space, it is time series data. If the data varies with respect to space, with the time being constant, it is cross-sectional data.
The number of coupons redeemed for a local residential community is varying over the weeks. Hence, option (a) is a correct and option (b) is incorrect.
Number of coupons redeemed is a discrete variable and hence option (c) is incorrect.

Total number of coupons redeemed in 6 weeks is 3020.
$1510th$ and $1511st$ observation is from week 5. Therefore, median will be week 5.
Hence, (d) is a correct option.
360 is the frequency of the number of coupons redeemed in week 2. Relative frequency for week 2 will be 360/3020. Hence, option (e) is incorrect.

16. A gym chain owner wants to know about the percentage of fat (%fat) in the body of people who have joined his gym centers for at least three months. The scatter plot of the data obtained from a sample of 20 people who visit a particular gym he owned is given in Figure S.4. Choose the correct options for the given data. [2 marks]

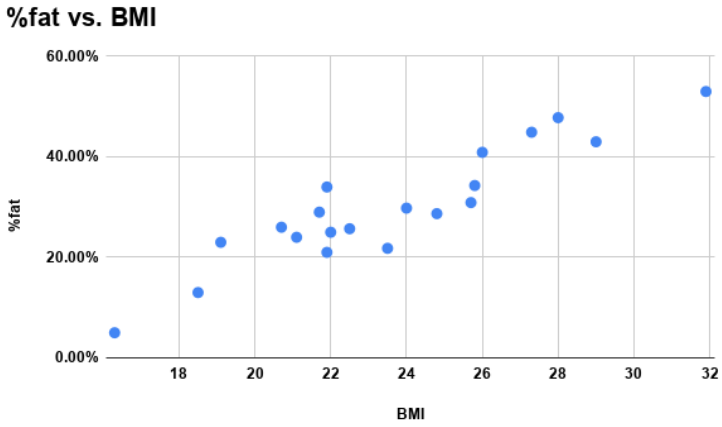


Figure S.4: %fat vs BMI

(a) The given sample is a good representative of the population.

(b) Association between BMI and %fat is non linear, strong and positive.

(c) Association in Figure S.4 shows that people having high BMI tends to have more fat percentage.

(d) Value of correlation coefficient r is more likely to be close to zero.

A gym chain owner has collected the data of 20 people from only a particular gym, and hence the given sample is not a good representative of the population.
Hence, option (a) is incorrect.
From the scatter plot, it is visible that the association is linear and hence option (b) is incorrect.
From the scatter plot, for larger values of BMI we are getting larger values of fat percentage.
Hence, option (c) is the correct.
As per the above discussion, there is a strong association between %fat and BMI and hence the correlation coefficient can not be close to zero. Hence, option (d) is incorrect.

17. Consider the box plot in Figure S.5 and choose the correct options. [3 marks]

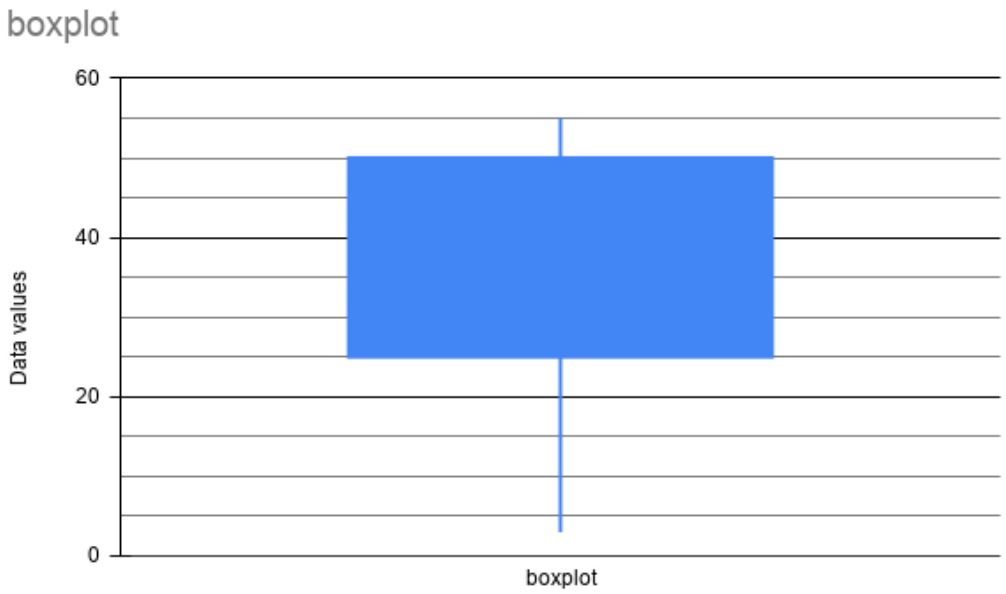


Figure S.5: Box plot

(a) Median of the data is necessarily equal to 37.5.

(b) The interquartile range of the given data is 25.

(c) There is no outlier in the dataset plotted in the box plot.

(d) Mode of the dataset will necessarily lie in [50, 55].

From the box plot shown in Figure S.5,
Maximum value = 55
$Q_1 = 25$
$Q_3 = 50$
Median of the data set must lie between Q_1 and Q_3 i.e., between 25 and 50. But it is not necessarily equal to 37.5.
Hence, option (a) is incorrect.

IQR = Q_3 - Q_1 = 50 - 25 = 25.
Hence, option (b) is correct.

As we know,

$$outlier < Q_1 - 1.5 \times IQR$$

and

$$1.5 \times IQR + Q_3 < outlier$$

$Q_1 - 1.5 \times IQR = 25 - 37.5 = -12.5$
$1.5 \times IQR + Q_3 = 37.5 + 50 = 87.5$
Therefore, there is no outlier in the dataset.
Hence, option (c) is correct.

Mode is that value in the dataset that has the maximum frequency.
Mode can lie anywhere in the dataset but it should not necessarily be in [50,55].
Hence, option (d) is incorrect.

Statistics for Data Science-1

Week-5 Graded Assignment

1. Vinod has n registers and m cover papers of different colours. In how many ways can he cover all the registers with cover papers?
   Answer: $m \times (m-1) \times (m-2) \times ... \times (m-n+1)$
   Solution:
   For the 1^{st} register, he can choose any of the m covers. So, number of ways to cover the first register with cover paper $= m$ ways.
   For the 2^{nd} register, he can choose any of the remaining $(m-1)$ covers. So, number of ways to cover the second register with cover paper $= m-1$ ways.
   Similarly, for n^{th} register, he can choose any of the remaining $(m-n+1)$ covers. So, number of ways to cover the second register with cover paper $= m-n+1$ ways.
   Hence, total number of ways in which he can cover all the registers with colour papers=$m \times (m-1) \times (m-2) \times ... \times (m-n+1)$ ways.

   For example:
   Suppose, we substitute values of n and m as 6 and 10 respectively.
   For the 1^{st} register, he can choose any of the 10 covers. So, number of ways to cover the first register with cover paper $= 10$ ways.
   For the 2^{nd} register, he can choose any of the remaining $(10-1)$ covers. So, number of ways to cover the second register with cover paper $= 10-1=9$ ways.
   Similarly, for 6^{th} register, he can choose any of the remaining $(10-6+1)$ covers. So, number of ways to cover the second register with cover paper $= 5$ ways.
   Hence, total number of ways in which he can cover all the registers with cover papers is
   $= 10 \times 9 \times 8 \times 7 \times 6 \times 5 = 151200$ ways.

2. n classmates could not agree on who would stand in the group photo along with the teacher for the yearbook. How many possible groups can be made such that there is at least one student with the teacher in the photo?
   Answer: $2^n - 1$
   Solution:
   Either a student be in that group photo or not. So, each student has 2 possibilities.
   The possibility that there may be 0 students along with the teacher in the group photo=1, as there is only 1 possible case of not selecting any of the n students.
   Hence, the total number of possible sets for a group photo where there is at least one student with the teacher=$2^n - 1$

For example:
Suppose, we substitute values of n as 7.
Either a student be in that group photo or not. So, each student has 2 possibilities.
The possibility that there may be 0 students along with the teacher in the group photo=1, as there is only 1 possible case of not selecting any of the 7 students.
Hence, the total number of possible sets for a group photo along with a teacher such that there is at least one student with a teacher in the photo $= 2^7 - 1 = 127$ ways.

Jay bought a new car in New York where a license plate can be created with alphabets A, B, C, D, E, W, X, Y, Z and numbers 0 to 9. He can either select a normal license plate or a VIP license plate. The VIP license plate begins with m alphabets followed by n numbers with repetition allowed. The normal license plate begins with a numbers followed by b alphabets without repetition. Based on this information, answer questions (3) and (4):

3. In how many ways can he select the VIP license plate?
Answer: $9^m \times 10^n$
Solution:
Total choice of alphabets = 9 i.e. A,B,C,D,E,W,X,Y and Z
Number of ways to select m alphabets with repetition $= 9^m$
Number of ways to select n numbers with repetition $= 10^n$
Hence, number of ways he can select VIP License Plate $= 9^m \times 10^n$

For example:
Suppose, we substitute values of m and n as 2 and 4 respectively.
Number of ways to select 2 alphabets with repetition $= 9^2$
Number of ways to select 4 numbers with repetition $= 10^4$
Hence, number of ways he can select VIP License Plate $= 9^2 \times 10^4 = 81000$ ways.

4. In how many ways can he select the license plate(normal or VIP)?
Answer: $[9^m \times 10^n] + [9(9-1)...(9-(b-1)) \times 10(10-1)...(10-(a-1))]$
Solution:
For the normal license plate:
Number of ways to select b alphabets without repetition $= 9 \times (9-1) \times ... \times (9-(b-1))$
Number of ways to select a numbers without repetition $= 10 \times (10-1) \times ... \times (10-(a-1))$
Hence, number of ways he can select the normal license plate
$=[9 \times (9-1) \times ... \times (9-(b-1))] \times [10 \times (10-1) \times ... \times (10-(a-1))]$
Therefore, total number of ways he can select the license plate = Number of ways he can select VIP License Plate + Number of ways he can select the normal license plate
$=[9^m \times 10^n] + [9(9-1)...(9-(b-1)) \times 10(10-1)...(10-(a-1))]$

For example:
Suppose, we substitute values of m, n, a and b as 2, 4, 2 and 3 respectively.
For the normal license plate
Number of ways to select 3 alphabets without repetition $= 9 \times (9-1) \times (9-2)$
Number of ways to select 2 numbers without repetition $= 10 \times (10-1)$
Hence, number of ways he can select the normal license plate is
$=[9 \times 8 \times 7] \times [10 \times 9] = 45360$ ways.
Therefore, total number of ways he can select the license plate = Number of ways he can select VIP License Plate + Number of ways he can select the normal license plate
$=[9^2 \times 10^4] + [9 \times 8 \times 7 \times 10 \times 9] = 810000 + 45360 = 855360$ ways.

5. Ram has n trophies that he wishes to place in his main cabinet, which has space only for two trophies. If the number of trophies is increased by 3, then the number of possible ways to arrange the trophies in the main cabinet becomes 5 times the number of ways to arrange n trophies. How many trophies does Ram have?
Answer: 3
Solution:
Number of ways Ram can place his n trophies in the main cabinet $= n(n-1)$
Number of ways Ram can place his $(n+3)$ trophies in the main cabinet $= (n+3)(n+2)$
It is given that:

$$5n(n-1) = (n+3)(n+2)$$
$$5n^2 - 5n = n^2 + 6n - n + 6$$
$$4n^2 - 10n - 6 = 0$$
$$(4n+2)(n-3) = 0$$

Therefore, $n = 3, -\frac{1}{2}$
Since, n can not be negative.
Hence, $n = 3$

6. There are N students in a class. The class teacher announced that the first n students who completes a given project within two days will be awarded. What are the possible number of ways the students will be awarded?

(a). $\frac{N!}{(N-n)!}$
(b). $\frac{N!}{(N-n)!n!}$
(c). $N!$
(d). $(N-n)!$

Answer: b

Solution:

Since first n students are to be awarded, i.e., there are n awards which are to be distributed among the N students in class. Therefore, for the distribution of the first award there are N ways and, for the second award, there are $(N-1)$ ways because one award is already distributed. Similarly, 3^{rd}, 4^{th}, . . . , n^{th} awards can be distributed in $(N-2)$, $(N-3)$, ... $(N-(n-1))$ ways.

Since here, events are occurring simultaneously and order doesn't matter (because, we are awarding the first n students and there will be no difference in awarding 3^{rd} student as first and then 5^{th} student as second or 5^{th} student as first and then 3^{rd} student as second etc.)

Thus, the total number of possible ways the student will be awarded,

$$= \frac{N \times (N-1) \times (N-2) ... \times (N-(n-1))}{n!} = \frac{N!}{(N-n)! \times n!}$$

For example;

Suppose, we substitute values of N and n as 40 and 5 respectively.

Since first 5 students are to be awarded, i.e., there are 5 awards which are to be distributed among the 40 students in class. Therefore, for the distribution of the first award there are 40 ways and, for the second award, there are 39 ways because one award is already distributed. Similarly, 3^{rd}, 4^{th} and 5^{th} awards can be distributed in 38, 37 and 36 ways.

Since here, events are occurring simultaneously and order doesn't matter (because, we are awarding the first 5 students and there will be no difference in awarding 3^{rd} student as first and then 5^{th} student as second or 5^{th} student as first and then 3^{rd} student as second etc.)

Thus, the total number of possible ways the student will be awarded,

$$= \frac{40 \times 39 \times 38 \times 37 \times 36}{5!} = \frac{40!}{35! \times 5!}$$

7. N students watched a patriotic movie. An analyst wishes to ask each student whether they liked the movie or not. Each student can either answer the question or refuse to respond. In how many ways, can the analyst get responses from the students?

   Answer: 3^N

   Solution:

   Since each student can choose to answer like or dislike or prefer not to answer. If he gives response then he has two options like or dislike the movie and another option is, he does not give any response. Thus, analyst can get feedback from each student in 3 ways.

From N students, analyst can get feedback in 3^N ways.

For example
Suppose, we substitute value of N as 6.
Analyst can get feedback from each student in 3 ways.
From 6 students, analyst can get feedback in $3^6 = 729$ ways.

8. If the value of sum of first n non-zero natural numbers is equal to $\frac{x(n+1)!}{2z}$, then find the value of $\frac{1}{x}$?.

Answer: $\frac{(n-1)!}{z}$
Solution:
As we know, the sum of first n non-zero natural numbers is $= \frac{n(n+1)}{2}$.
Given, $\frac{n(n+1)}{2} = \frac{x(n+1)!}{2z}$
$n(n+1) = \frac{x(n+1)n(n-1)!}{z}$
$1 = \frac{x(n-1)!}{z}$
$\frac{1}{x} = \frac{(n-1)!}{z}$

For example:
Suppose, we substitute values of n and z as 7 and 3 respectively.

The sum of the first 7 non-zero natural numbers is $= \frac{7(7+1)}{2}$.
Given, $\frac{7(7+1)}{2} = \frac{x(7+1)!}{2 \times 3}$
$7(7+1) = \frac{x(7+1)7(7-1)!}{3}$
$1 = \frac{x(7-1)!}{3}$
$\frac{1}{x} = \frac{(7-1)!}{3} = \frac{6!}{3} = \frac{720}{3} = 240$

9. Adam wrote down a n-digit university roll number on a piece of paper. On his way home from office, it rained heavily and the paper got wet. Later, he saw that the first m digits of the roll number had disappeared. In how many ways can Adam complete this university roll number if repetition of digits is allowed?

a. $m!$

b. 10^m

c. $10^{m-1} \times 9$

d. 9^m

Answer: b

Solution:

Adam has to complete the first m digits of a n-digit university roll number with repetition of digits allowed.

Now, the first digit can be filled with any number as it is a categorical variable and does not have any numeric meaning. For example: Roll number '000257' is also a 5-digit roll number.

Hence, number of ways to fill first digit =10 ways.

For the second digit, he can fill it with any number 0-9. Hence, number of ways to fill second digit = 10 ways.

Similarly, for the $3^{rd}, 4^{th}, ..., n^{th}$ digit also, he can fill it with any number 0-9.

Hence, the number of ways to fill any of these digits = 10

Therefore, the number of ways Adam can complete the n-digit university roll number $= 10^m$.

Hence, option (b) is correct.

For example:

Suppose, we substitute values of n and m as 10 and 3 respectively.

Adam has to complete the first three digits of a 10-digit university roll number with repetition of digits allowed.

Now, the first digit can be filled with any number as it is a categorical variable and does not have any numeric meaning. For example: Roll number '000257' is also a 5-digit roll number.

Hence, number of ways to fill first digit =10 ways.

For the second digit, he can fill it with any number from 0-9. Hence, number of ways to fill second digit = 10 ways.

Similarly, for the digit also, he can fill it with any digit 0-9. Hence, number of ways to fill third digit = 10 ways.

Therefore, the number of ways Adam can complete the 10-digit university number $= 10 \times 10 \times 10 = 10^3$.

Hence, option (b) is correct.

10. Let $x = \dfrac{5!}{4 \times 3!}$. Which of the following expressions is/are equal to x?

a. $5 \times 0!$

b. $5 \times \dfrac{1}{0!}$

c. $5 \times \dfrac{1}{0}$

d. $5 \times \dfrac{6}{3! + (3 \times 2) + (3 \times 2 \times 1)}$

e. $5 \times \dfrac{18}{3! + (3 \times 2) + (3 \times 2 \times 1)}$

Answer: a, b, e

Solution:

Let us first solve for x:

$x = \dfrac{5!}{4 \times 3!} = \dfrac{5 \times 4 \times 3!}{4 \times 3!} = 5$

Now,

a. $5 \times 0! = 5 \times 1 = 5 = x$

Hence, option a is correct.

b. $5 \times \dfrac{1}{0!} = 5 \times \dfrac{1}{1} = 5 = x$

Hence, option b is correct.

c. $5 \times \dfrac{1}{0} \neq x$

$\dfrac{1}{0}$ is not defined.

Hence, option c is incorrect.

d. $5 \times \dfrac{6}{3! + (3 \times 2) + (3 \times 2 \times 1)} = 5 \times \dfrac{6}{6 + 6 + 6} = 5 \times \dfrac{1}{3} \neq 5 \neq x$

Hence, option d is incorrect.

e. $5 \times \dfrac{18}{3! + (3 \times 2) + (3 \times 2 \times 1)} = 5 \times \dfrac{18}{6 + 6 + 6} = 5 = x$

Hence, option e is correct.

Statistics for Data Science-1

Week 6 Graded Assignment

1. How many 5-digit numbers can be formed from the numbers $0, 2, 4, 5, 7$ and 9 (without repetition), such that it is divisible by 4?

   a. 120

   b. 144

   c. 132

   d. 104

   Answer: b

   Solution:

   By the divisibility rule of 4, for a number to be divisible by 4, the last two digits of the number should be divisible by 4.

   Using the divisibility rule of 4, any 5-digit number formed from the numbers $0, 2, 4, 5, 7$ and 9 will be divisible by 4 if its last two digits are: 04, 20, 24, 40, 52, 72 or 92.

   Case 1- Number ends with 04:

   The last two digits of a five digit number is fixed (i.e. 04). We have to fill the first three digits using the remaining four numbers (as repetition is not allowed) i.e. 2,5,7 and 9.

   Now, the three places can be filled with these 4 numbers in ${}^4P_3$ ways.

   Similarly, in the cases where the five digit number ends with 20 and 40, the number of ways to fill in the last three digits is ${}^4P_3$ in each case.

   Case 2- Number ends with 24:

   The last two digits of a five digit number is fixed (i.e. 24). We have to fill the first three digits using the remaining four numbers (as repetition is not allowed) i.e. 0,5,7 and 9.

   0 cannot be the first digit as it will make the number a four digit number. Therefore, the number of ways to fill in the first place is 3.

   Now, 0 can be used along with the numbers 5,7 and 9 in the remaining places. Hence, the remaining two places can be filled with these 4 digits in ${}^3P_2$ ways.

   Therefore, total number of ways to complete a five-digit number ending with 24 $= 3 \times {}^3P_2$

   Similarly, in the cases where the five digit number ends with 52, 72 and 92, the number of ways to fill in the last three digits is $3 \times {}^3P_2$ in each case.

Hence, the total number of 5-digit numbers that can be formed from the digits $0, 2, 4, 5, 7$ and 9 (without repetition), such that it is divisible by 4 are:
$=^4 P_3 +^4 P_3 +^4 P_3 + 3 \times^3 P_2 + 3 \times^3 P_2 + 3 \times^3 P_2 + 3 \times^3 P_2$
$= 3 \times^4 P_3 + 4 \times 3 \times^3 P_2$
$= 3 \times 24 + 4 \times 3 \times 6$
$= 144$

2. There are n train stops between Chennai and Assam. How many train tickets are to be printed, so that a person can travel between any of the two stations (irrespective of direction of travel)?
Solution:
There are a total of $(n + 2)$ stations including Assam and Chennai along with the n stations between them.
Now, one can travel from any one station to another (irrespective of direction of travel).
Hence, total number of tickets to be printed $=^{(n+2)} P_2 = (n + 2)(n + 1)$ tickets.
Example: n=7
There are a total of 9 stations including Assam and Chennai along with the 7 stations between them.
Now, one can travel from any one station to another (irrespective of direction of travel).
Hence, total number of tickets to be printed $=^9 P_2 = 72$ tickets.

3. A man desires to throw a party for some of his friends. In how many ways can he select m friends from a group of n friends, if the two of his friends (say 'A' and 'B') will not attend the party together?
Solution:
Total number of ways in which he can invite his friends to the party (without any condition) $=^n C_m$ ways.
Now, number of ways in which he can invite his friends so that 2 of his friends will attend the party $=^{n-2} C_{m-2}$
Hence, Number of ways in which two of his friends will not attend the party together $=^n C_m -^{n-2} C_{m-2}$
Example: n=10 and m=7
Total number of ways in which he can invite his friends to the party (without any condition) $=^{10} C_7$ ways.
Now, number of ways in which he can invite his friends so that 2 of his friends will attend the party $=^8 C_5$
Hence, Number of ways in which two of his friends will not attend the party together $=^{10} C_7 -^8 C_5$
$= 120 - 56$
$= 64$ ways.

4. Suman has m clothes of different types, say, $C_1, C_2, ..., C_m$ and she wants to wear all these clothes at different days, say, $D_1, D_2, ..., D_m$. Due to some reason, C_1 must be used either at D_{m-2} or at D_{m-1} and C_2 can be used either at D_{m-1} or at D_{m-2} and

at D_m. Every cloth is to be used at only one day, in how many ways can clothes be used?

Solution:

C_1 must be used at either D_{m-2} or D_{m-1}, thus there are 2 ways in which C_1 can be used.

C_2 can be used in only 2 ways ,i.e, at D_{m-2} or D_{m-1} and D_m. (As C_1 is already used at anyone of D_{m-2} or D_{m-1})

Now , for C_3, $(m-2)$ days are available, i.e., there are $(m-2)$ ways for using C_3, $(m-3)$ ways for using C_4, $(m-4)$ ways for using C_5,..., 1 way for using C_m

Therefore, total number of ways clothes can be used will be $= 2 \times 2 \times (m-2) \times (m-3) \times (m-4), ..., \times 3 \times 2 \times 1$

For example:

If $m = 6$, then we will have 6 clothes and 6 days.

Corresponding number of ways will be;

C_1 must be used at D_4 or $D_5 = 2$ ways

C_2 can be used in 2 ways i.e., 1 way will be using at D_4 or D_6 (if C_1 is used at D_5) and 1 way will be at D_5 or D_6 (if C_1 is used at D_4) = total 2 ways for using C_2.

for C_3 four days are available= 4 ways.

for C_4 three days are available= 3 ways.

for C_5 two days are available= 2 ways.

for C_6 only one day is left = 1 way.

Therefore, total number of ways $= 2 \times 2 \times 4 \times 3 \times 2 \times 1 = 96$ ways.

5. How many n-digit numbers can be formed such that they read the same way from either of the side (i.e. the number should be a palindrome)?

   a. $9 \times 10^{\left(\frac{n-1}{2}\right)}$

   b. $9 \times 10^{\left(\frac{n+1}{2}\right)}$

   c. $9 \times 10^{n-1}$

   d. 10^n

Answer: a

Solution:

For a n-digit number to read the same way from either side, the $1^{st}, 2^{nd}, ...$ and $\left(\frac{n-1}{2}\right)^{th}$ digits needs to be the same as $n^{th}, (n-1)^{th}, ...$ and $\left(\frac{n+3}{2}\right)^{th}$ digits respectively.

So, number of ways to fill first place = 9, as the number will become a $(n-1)$ digit number if 0 is in the first place.

And, $2^{nd}, 3^{rd}$..., $\left(\frac{n-1}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ can be filled in $10^{\left(\frac{n-1}{2}\right)}$ ways as repeti-

tion is allowed.
$n^{th}, (n-1)^{th}, ...$ and $\left(\frac{n+3}{2}\right)^{th}$ digit should be same as $1^{st}, 2^{nd}, ...$ and $\left(\frac{n-1}{2}\right)^{th}$ digit respectively, so number of ways to fill that place is 1
Hence, the total number of n-digit numbers that can be formed such that they read the same way from either side $= 9 \times 10^{\left(\frac{n-1}{2}\right)}$ ways.
Hence, option a is correct.

Example: $n = 5$
For a 5-digit number to read the same way from either side, the first and second digit needs to be same as the fifth and fourth digit respectively.
So, number of ways to fill first place $= 9$, as the number will become a 4-digit number if 0 is in the first place.
And, second and third place can be filled in 10×10 ways as repetition is allowed.
Fourth and fifth digit should be same as first and second digit respectively, so number of ways to fill that place is 1×1
Hence, the total number of 5-digit number that can be formed such that they read the same way from either side $= 9 \times 100 = 9 \times 10^2$
Hence, option a is correct.

6. Find the total number of ways to form a m digit number (without repetition) from the digits $0, 1, 2, ..., n$.

   (a) $n \times {}^nP_{m-1}$
   (b) ${}^{n+1}P_m$
   (c) $(n-1) \times {}^{n-1}P_{m-1}$
   (d) $n + {}^nP_{m-1}$

   Answer: a
   Solution:
   We have to form m digit number with the digits $0, 1, 2, ..., n$ and we have $n+1$ digits in total.
   For the first digit we have n ways (as 0 can't be considered as first digit). Now, we have n digits remaining (because one digit is already used for first digit) and $m-1$ places. Therefore, number of ways to arrange n digits at $m-1$ places is ${}^nP_{m-1}$ ways.
   Hence, total number of ways to form a m digit number from the digits $0, 1, 2, ..., n$ (without repetition) is $n \times {}^nP_{m-1}$.
   For example: $m = 6$ **and** $n = 7$
   We have to form 6 digit numbers from digits $0, 1, 2, 3, 4, 5, 6, 7$ and we have 8 digits in total.
   For the first digit we have 7 ways(as 0 can't be considered as first digit). Now, we have 7 digits remaining (because one digit is already used for first digit) and 5 places.

Therefore, number of ways to arrange 7 digits at 5 places is ${}^7P_5$ ways.
Hence, total number of ways to form a 6 digit number from the digits $0, 1, 2, ..., 7$ (without repetition) is $7 \times {}^7P_5$.

7. In a restaurant, x men and y women are seated on $(x + y)$ chairs at a round table. Find the total number of possible ways such that x men are always sitting next to each other.

   (a) $x! \times y!$
   (b) $(x - 1)! \times (y - 1)!$
   (c) $x! \times (y + 1)!$
   (d) $(x + y - 1)!$

   Answer: a
   Solution:
   Considering the x men $M_1, M_2, ..., M_x$ who sit together as one, we get $(y + 1)$ persons in all, who can be seated at a round table in $y!$ ways. Further, since $M_1, M_2, ..., M_x$ can interchange their positions in $x!$ ways, the total number of possible ways of getting $M_1, M_2, ..., M_x$ together is $y! * x!$
   For example: $y = 2, x = 3$
   Considering the 3 men M_1, M_2 and M_3 who sit together as one, we get $2 + 1 = 3$ persons in all, who can be seated at a round table in 2! ways. Further, since M_1, M_2 and M_3 can interchange their positions in 3! ways, the total number of possible ways of getting M_1, M_2 and M_3 together is $2! * 3! = 12$.

8. In how many ways can a group of $n - m$ players be formed from n state level players and m district level players such that the group contains exactly 1 district level player?
   Answer: $\dfrac{m \times n!}{(m + 1)!(n - m - 1)!}$

   Solution:
   The group must have $(n - m)$ players and must contain exactly 1 district level player. Hence, we will select $(n - m - 1)$ persons from n state level players and 1 from m district level players.
   The total number of ways to form a group of $(n - m)$ players is:
   $= {}^nC_{n-m-1} \times {}^m C_1$

   $$= \frac{m \times n!}{(n - m - 1)! \times (m + 1)!}.$$

   For example: $n = 10, m = 6$
   The group must have 4 players. But the group must contain exactly 1 district level

player. Hence, we will select 3 players from 10 state level players and 1 from 6 district level players.

Total number of ways to form a group of 4 players $= {}^{10}C_3 \times {}^{6}C_1 = \dfrac{6 \times 10!}{3! \times 7!} = 720.$

9. Find the value of r such that the ratio of ${}^{3}P_r$ and ${}^{4}P_{r-1}$ will be $\dfrac{1}{2}$?

Answer: 3

Solution:

Given,

$$\frac{3P_r}{{}^{4}P_{r-1}} = \frac{1}{2}$$

$$\frac{3!/(3-r)!}{4!/(4-(r-1))!} = \frac{1}{2}$$

$$\frac{3!/(3-r)!}{4!/(5-r)!} = \frac{1}{2}$$

$$\frac{3!/(3-r)!}{4 \times 3!/(5-r)(4-r)(3-r)!} = \frac{1}{2}$$

$$\frac{(5-r)(4-r)}{4} = \frac{1}{2}$$

$$(5-r)(4-r) = 2$$

By solving above equation,
$r = 6$ and $r = 3$.
$r = 3$ is the answer since $r = 6$ is greater than $n = 3$.

10. Choose the incorrect option/s for $n > 2$:

a. ${}^{n}C_r + {}^{n}C_{r-1} = {}^{n+1}C_r$

b. ${}^{n}C_r = 1$ for $r = 0$ and $r = n$

c. ${}^{n}C_r = {}^{n-1}C_{r-1} + {}^{n-1}C_r$

d. None of the above

Answer: d

Solution:

Going through the options and solving for given conditions, options a, b, c are the true relations and hence, are incorrect. Therefore, the correct option is (d).

Statistics for Data Science-1

Week 7 Graded assignment

1. m boys and 2 girls are to be placed next to each other in the school ground for morning assembly. What is the probability that there are exactly 4 boys between the 2 girls?

   a. $\frac{2m-5}{{}^{m+2}P_2}$

   b. $\frac{2m-6}{{}^{m+2}P_2}$

   c. $\frac{2m-6}{{}^{m+3}P_2}$

   d. $\frac{2m-4}{{}^{m+2}P_2}$

   Answer: b

   Solution:

   There are a total of $(m+2)$ places to arrange the 2 girls.

   Therefore, the number of ways in which 2 girls can be arranged $= {}^{m+2}P_2$

   Positioning of the 2 girls such that there are exactly 4 boys between them can be done in the following ways:

   Case 1: First girl at 1^{st} place and second girl at 6^{th} place and vice-versa, i.e. 2 ways.

   Case 2: First girl at 2^{nd} place and second girl at 7^{th} place and vice-versa, i.e. 2 ways.

   Similarly,

   Case (m-3): First girl at $(m-3)^{th}$ place and second girl at $(m+2)^{th}$ place, and vice-versa, i.e. 2 ways.

   Hence, Number of possible ways such that there are exactly 4 boys between the 2 girls $= 2 \times (m-3) = 2m-6$ ways.

   Therefore, P[There are exactly 4 boys between the 2 girls] $= \frac{2m-6}{{}^{m+2}P_2}$

   Hence, option (b) is correct.

   Example: $m = 8$

   There are a total of 10 places to arrange the 2 girls.

   Therefore, the number of ways in which 2 girls can be arranged $= {}^{10}P_2$

   Positioning of the 2 girls such that there are exactly 4 boys between them can be done in the following ways:

   Case 1: First girl at 1^{st} place and second girl at 6^{th} place and vice-versa, i.e. 2 ways.

Case 2: First girl at 2^{nd} place and second girl at 7^{th} place and vice-versa, i.e. 2 ways.
Case 3: First girl at 3^{rd} place and second girl at 8^{th} place and vice-versa, i.e. 2 ways.
Case 4: First girl at 4^{th} place and second girl at 9^{th} place and vice-versa, i.e. 2 ways.
Case 5: First girl at 5^{th} place and second girl at 10^{th} place and vice-versa, i.e. 2 ways.

Hence, Number of possible ways such that there are exactly 4 boys between the 2 girls $= 2 \times 5 = 10$ ways.
Therefore, P[There are exactly 4 boys between the 2 girls] $= \frac{10}{{}^{10}P_2} = \frac{10}{90} = \frac{1}{9}$
Hence, option (b) is correct.

2. In a Multiple Select Question, there are m options, of which one or more can be correct. Let us define an event E that the option 'A' is correct. What is the cardinality of E?
Solution:
Case 1: Only option A is correct.
Number of possible elements = 1 i.e. {(A)}
Case 2: Two options are correct and A is one of them.
Number of possible elements $= {}^{m-1}C_1 = m - 1$
Case 3: Three options are correct and A is one of them.
Number of possible elements $= {}^{m-1}C_2$
Similarly,
Case m: All options are correct.
Number of possible elements 1
Hence, Cardinality of $E = 1 + {}^{m-1}C_1 + {}^{m-1}C_2 + ... + 1 = 2^{m-1}$

Example: m=4
Case 1: Only option A is correct.
Number of possible elements = 1 i.e. {(A)}
Case 2: Two options are correct and A is one of them.
Number of possible elements $= {}^{3}C_1 = 3$ i.e. {(A,B),(A,C),(A,D)}
Case 3: Three options are correct and A is one of them.
Number of possible elements $= {}^{3}C_2 = 3$ i.e. {(A,B,C),(A,B,D),(A,C,D)}
Case 4: All options are correct.
Number of possible elements 1 i.e. {(A,B,C,D)}
Hence, Cardinality of $E = 1 + 3 + 3 + 1 = 2^3 = 8$

3. A person predicts daily whether the price of stocks of wrist watch companies will go up or down. If his prediction on stock price of Titan is correct a times out of b, for Rolex it is correct p times out of q and for Fossil it is correct x times out of y, then what is the probability that at least two of his predictions are correct on a given day?

a. $\left[\frac{a}{b} \times \frac{p}{q} \times \left(1 - \frac{x}{y}\right)\right] + \left[\frac{a}{b} \times \left(1 - \frac{p}{q}\right) \times \frac{x}{y}\right] + \left[\left(1 - \frac{a}{b}\right) \times \frac{p}{q} \times \frac{x}{y}\right] + \left[\frac{a}{b} \times \frac{p}{q} \times \frac{x}{y}\right]$

b. $\left[\frac{a}{b}\times\frac{p}{q}\times\left(1-\frac{x}{y}\right)\right]+\left[\frac{a}{b}\times\left(1-\frac{p}{q}\right)\times\frac{x}{y}\right]+\left[\left(1-\frac{a}{b}\right)\times\frac{p}{q}\times\frac{x}{y}\right]$

c. $\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]+\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]+\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]+\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]$

d. $\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]+\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]+\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]$

Answer: a

Solution:

Let us define the following events:

A : Prediction for Titan is correct.

B : Prediction for Rolex is correct.

C : Prediction for Fossil is correct.

We are given that :

$P(A)=\frac{a}{b}$, $P(B)=\frac{p}{q}$ and $P(C)=\frac{x}{y}$

Case 1: Prediction for only Titan and Rolex is correct

$P(A\cap B\cap C^c)=\frac{a}{b}\times\frac{p}{q}\times\left(1-\frac{x}{y}\right)$

Case 2: Prediction for only Titan and Fossil is correct

$P(A\cap B^c\cap C)=\frac{a}{b}\times\left(1-\frac{p}{q}\right)\times\frac{x}{y}$

Case 3: Prediction for only Rolex and Fossil is correct

$P(A^c\cap B\cap C)=\left(1-\frac{a}{b}\right)\times\frac{p}{q}\times\frac{x}{y}$

Case 4: All predictions are correct.

$P(A\cap B\cap C)=\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}$

Hence, P(At least two predictions are correct)

$=\left[\frac{a}{b}\times\frac{p}{q}\times\left(1-\frac{x}{y}\right)\right]+\left[\frac{a}{b}\times\left(1-\frac{p}{q}\right)\times\frac{x}{y}\right]+\left[\left(1-\frac{a}{b}\right)\times\frac{p}{q}\times\frac{x}{y}\right]+\left[\frac{a}{b}\times\frac{p}{q}\times\frac{x}{y}\right]$

Hence, option (a) is correct.

Example: $a=4, b=5, p=, q=6, x=3, y=4$

Let us define the following events:

A : Prediction for Titan is correct.

B : Prediction for Rolex is correct.

C : Prediction for Fossil is correct.

We are given that :

$P(A)=\frac{4}{5}$, $P(B)=\frac{5}{6}$ and $P(C)=\frac{3}{4}$

Case 1: Prediction for only Titan and Rolex is correct

$P(A \cap B \cap C^c) = \frac{4}{5} \times \frac{5}{6} \times \frac{1}{4} = \frac{20}{120}$

Case 2: Prediction for only Titan and Fossil is correct

$P(A \cap B^c \cap C) = \frac{4}{5} \times \frac{1}{6} \times \frac{3}{4} = \frac{12}{120}$

Case 3: Prediction for only Rolex and Fossil is correct

$P(A^c \cap B \cap C) = \frac{1}{5} \times \frac{5}{6} \times \frac{3}{4} = \frac{15}{120}$

Case 4: All predictions are correct.

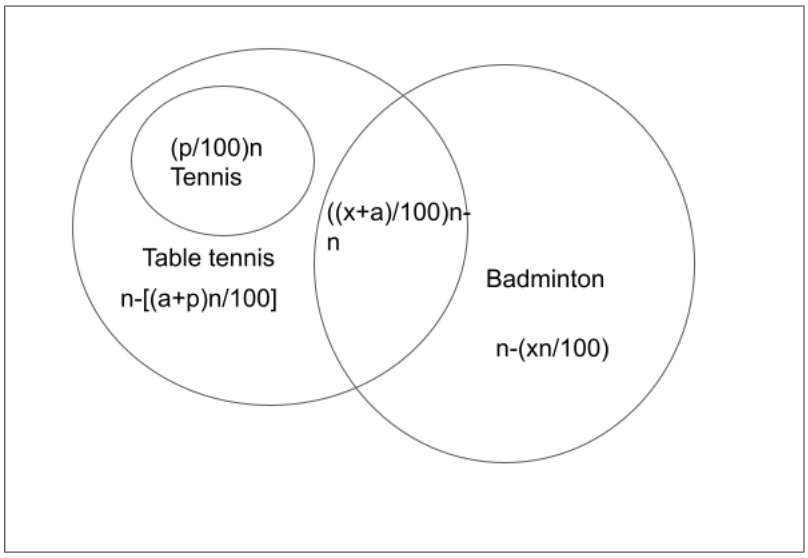
$P(A \cap B \cap C) = \frac{4}{5} \times \frac{5}{6} \times \frac{3}{4} = \frac{60}{120}$

Hence, P(At least two predictions are correct)$= \frac{20}{120} + \frac{15}{120} + \frac{12}{120} + \frac{60}{120} = \frac{107}{120}$

Hence, option (a) is correct.

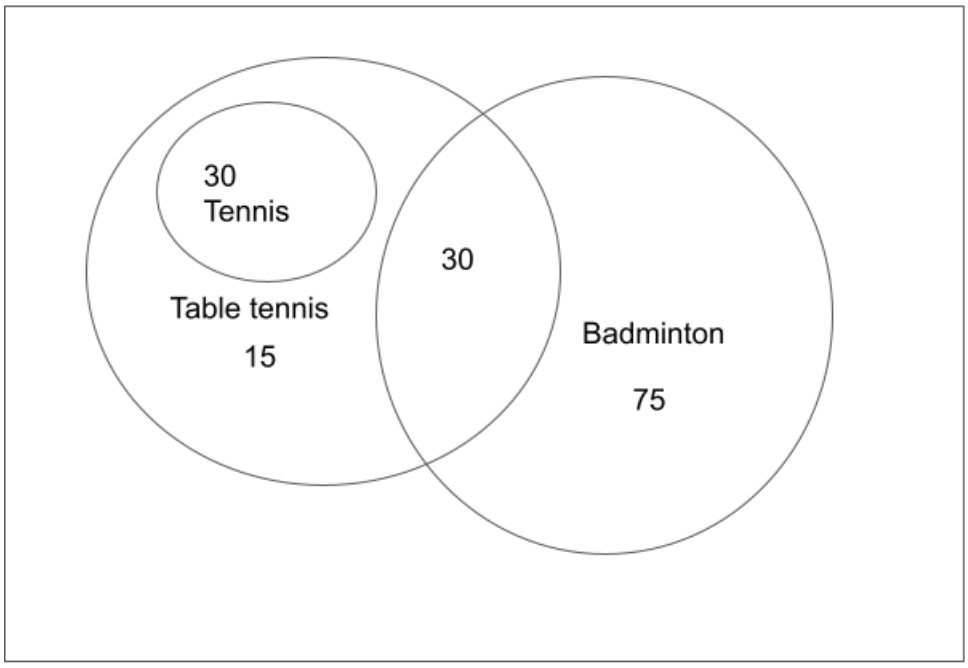
4. There are a total of n students who are part of badminton, table tennis and tennis team of the college. Of which $x\%$ of students play table tennis, $p\%$ play tennis and $a\%$ play badminton. It is also noticed that all students who play tennis also play table tennis, but not badminton. Now a student is selected at random, what is the probability that he/she is the part of table tennis team only? (Enter the answer correct to 1 decimal place.)

Solution:



Therefore, probability that he/she is the part of table tennis team only $= \left(1 - \frac{a+p}{100}\right)$

Example: $n = 150, x = 50, p = 20, a = 70$



Therefore, probability that he/she is the part of table tennis team only $= \frac{15}{150} = \frac{1}{10}$

5. The chance that a student will clear the quiz 1 paper is a and the chance that he will clear both quiz 1 and quiz 2 papers is b. The chance that he will clear at least one quiz paper is c. What is the chance that he will clear quiz 2 paper? (Enter the answer correct to 2 decimal accuracy)
   Solution:
   Let us define the following events:
   A: Student will clear the quiz 1 paper ; B: Student will clear the quiz 2 paper.
   We are given that:
   $P(A) = a$, $P(A \cap B) = b$, $P(A \cup B) = c$ and we want, $P(B)$
   Now, $P(A \cup B) = P(A) + P(B) - P(A \cap B) \implies c = a + P(B) - b$
   $P(B) = d = c - a + b$

For example: $a = 0.4$**,** $b = 0.3$**,** $c = 0.5$
Let us define the following events:
A: Student will clear the quiz 1 paper ; B: Student will clear the quiz 2 paper.
We are given that:
$P(A) = 0.4$, $P(A \cap B) = 0.3$, $P(A \cup B) = 0.5$ and we want, $P(B)$
Now, $P(A \cup B) = P(A) + P(B) - P(A \cap B) \implies 0.5 = 0.4 + P(B) - 0.3$
$P(B) = 0.5 - 0.4 + 0.3 = 0.4$
Therefore, the chance that the student will clear the quiz 2 paper is 0.4

6. If $P(A) = x$ and $P(B) = y$ and probability of the complement of $(A \cup B)$ is z, then calculate $P(A \cup B)$? (Enter the answer correct to 2 decimal point accuracy)
Solution:
$P(A \cup B) = 1 - P(A \cup B)^c = 1 - z$
For example: $x = 0.2$, $y = 0.5$, and $z = 0.4$
$P(A \cup B) = 1 - P(A \cup B)^c = 1 - 0.4 = 0.6$

7. a cards are drawn at random (without replacement) from a pack of 52 cards. Find the probability that b are black and c are red. (Enter the answer correct to two decimal places)
Solution:
Since there are 26 black cards (of spades and clubs) and 26 red cards (of diamonds and hearts) in a pack of cards, the required probability$=\frac{^{26}C_b \times ^{26}C_c}{^{52}C_a}$
Example: $a = 4, b = 2$ and $c = 2$
Since there are 26 black cards (of spades and clubs) and 26 red cards (of diamonds and hearts) in a pack of cards, the required probability$=\frac{^{26}C_2 \times ^{26}C_2}{^{52}C_4} = 0.39$

Pramod goes to a shop to buy some clothes. Shopkeeper shows him x shirts, y pants and z t-shirts. If he selects three clothes at random, then based on the information, answer the questions 8, 9 and 10.

8. Find the probability that the randomly chosen clothes are of different type. (Enter the answer correct to three decimal places)
Solution:
The total number of cases are $^{x+y+z}C_3$.
Since the number of favourable cases of getting one cloth of each type is
$= {}^xC_1 \times {}^yC_1 \times {}^zC_1$
Therefore, the required probability$=\frac{^xC_1 \times ^yC_1 \times ^zC_1}{^{x+y+z}C_3}$

For example: $x = 5$**,** $y = 4$**,** $z = 10$
The total number of cases are $^{19}C_3$.
Since the number of favourable cases of getting one cloth of each type is

$= {}^5C_1 \times {}^4C_1 \times {}^{10}C_1$

Therefore, the required probability$= \dfrac{{}^5C_1 \times {}^4C_1 \times {}^{10}C_1}{{}^{19}C_3} = \dfrac{200}{969} = 0.2064$

9. Find the probability that the randomly chosen clothes does not contain pant. (Enter the answer correct to two decimal places)

   Solution:

   The total number of cases are ${}^{x+y+z}C_3$.

   If randomly chosen clothes does not contain pant, then all the three clothes must be from shirts and t-shirts, i.e., from $x+z$ clothes. Hence, the number of favourable cases for this event is ${}^{x+z}C_3$.

   Therefore, the required probability$= \dfrac{{}^{x+z}C_3}{{}^{x+y+z}C_3}$

   For example: $x = 5$**,** $y = 4$**,** $z = 10$

   The total number of cases are ${}^{19}C_3$.

   If randomly chosen clothes does not contain pant, then all the three clothes must be from shirts and t-shirts, i.e., from $5+10 = 15$ clothes. Hence, the number of favourable cases for this event is ${}^{15}C_3$.

   Therefore, the required probability$= \dfrac{{}^{15}C_3}{{}^{19}C_3} = \dfrac{2730}{5814} = 0.4695$

10. Find the probability that at least one of the clothes is a shirt. (Enter the answer correct to two decimal places)

    Solution:

    The total number of cases are ${}^{x+y+z}C_3$.

    P(at least one of the clothes is shirt) = 1-P(none of the three clothes is shirt)

    In order that none of the 3 clothes is shirt, all the 3 clothes must be from pants and t-shirts,i.e, from $y + z$ clothes and the number of favourable cases for this event is ${}^{y+z}C_3$.

    P(none of the three clothes is shirt)$= \dfrac{{}^{y+z}C_3}{{}^{x+y+z}C_3}$

    Hence, P(at least one of the clothes is shirt)$= 1 - \dfrac{{}^{y+z}C_3}{{}^{x+y+z}C_3}$

    For example: $x = 5$**,** $y = 4$**,** $z = 10$

    The total number of cases are ${}^{19}C_3$.

    P(at least one of the clothes is shirt) = 1-P(none of the three clothes is shirt)

    In order that none of the 3 clothes is shirt, all the 3 clothes must be from pants and t-shirts, i.e, from $4+10 = 14$ clothes and the number of favourable cases for this event is ${}^{14}C_3$.

    P(none of the three clothes is shirt)$= \dfrac{{}^{14}C_3}{{}^{19}C_3} = \dfrac{2184}{5814}$

    Hence, P(at least one of the clothes is shirt)$= 1 - \dfrac{2184}{5814} = 0.624$

11. An urn contains 3 balls numbered $1, 2$ and 3. The co-efficients of the equation

$px^2+qx+c=0$ is determined by drawing the numbered balls with replacement. What is the probability that the equation will have imaginary roots?

Nature of roots:

Consider a quadratic equation: $ax^2+bx+c=0$

Compute $D=b^2-4ac<0$

$$\text{Roots} = \begin{cases} D<0 & \text{imaginary roots} \\ D\geq 0 & \text{real roots} \end{cases}$$

a. $\frac{4}{27}$

b. $\frac{23}{27}$

c. $\frac{16}{27}$

d. None of the above

Answer: b

Solution:

Since each coefficient in equation $px^2+qx+c=0$ is determined by drawing a numbered ball from the urn, each of the coefficients p, q and c can take values from 1 to 3.

Therefore, total number of possible outcomes $=3\times 3\times 3=27$

P[Imaginary roots] $=1-P$[Real roots]

For a Quadratic equation to have real roots, the equation $q^2-4pc\geq 0$ should be satisfied.

The number of favourable cases for real roots are:

pc	p	c	$4pc$	q (such that $q^2-4pc\geq 0$)	No. of cases
1	1	1	4	2, 3	2
2	1	2	8	3	1
	2	1	8	3	1

Hence, Total number of favourable cases for real roots $=2+1+1=4$

Therefore, P[Real roots]=$\frac{4}{23}$

P[Imaginary roots]=$1-\frac{4}{23}=\frac{23}{27}$

Hence, option (b) is correct.

12. If A and B are mutually exclusive or disjoint events, then which of the following is/are always true:

a. $P(A)=P(B)$

b. $P(A) < P(B)$

c. $P(A) \leq P(B^c)$

d. $P(A) \geq P(B^c)$

Answer: c

Solution:

Since, A and B are mutually exclusive or disjoint events.

Therefore, $A \cap B = \phi$

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

Therefore, $A \subseteq B^c \Rightarrow P(A) \leq P(B^c)$

Statistics for Data Science-1

Week 8 Graded Assignment

1. Zaheer Khan has taken m five-wicket hauls in his last n matches. His match records are selected at random, one by one, and analyzed. If none of the match records is analyzed more than once, then what is the probability that the k^{th} one analyzed is his last five-wicket haul match?

   a. $\frac{{}^{m}C_{m-1} \times {}^{n-m}C_{k-m}}{{}^{n}C_{k-1}} \times \frac{1}{n-k+1}$

   b. $\frac{{}^{m}C_{m-1} \times {}^{n-m}C_{k-m}}{{}^{n}C_{k-1}}$

   c. $\frac{1}{n-k+1}$

   d. $\frac{{}^{n-m}C_{k-m}}{{}^{n}C_{k-1}}$

   Answer: a

   Solution:

   Let A be the event of getting exactly $(m-1)$ five-wicket haul matches when the first $(k-1)$ matches are analyzed.

   Let B be the event that k^{th} match analyzed is five-wicket haul match.

   Now, P(A)= $\frac{{}^{m}C_{m-1} \times {}^{n-m}C_{k-m}}{{}^{n}C_{k-1}}$

   $P(B|A)$ = Probability that the k^{th} match analyzed is five-wicket haul match given that the analysis of the first $k-1$ matches shows $m-1$ five-wicket haul matches.

   $= \frac{1}{n-k+1}$

   Required Probability = $P(A \cap B) = P(A) \times P(B|A) = \frac{{}^{m}C_{m-1} \times {}^{n-m}C_{k-m}}{{}^{n}C_{k-1}} \times \frac{1}{n-k+1}$

   Hence, option (a) is correct.

   Example: n=15,k=9, m=4

   Let A be the event of getting exactly 3 five-wicket haul matches when the first 8 matches are analyzed.

   Let B be the the event that 9^{th} match analyzed is five-wicket haul match.

   Now, P(A)= $\frac{{}^{4}C_{3} \times {}^{11}C_{5}}{{}^{15}C_{8}}$

   $P(B|A)$ = Probability that the 9^{th} match analyzed is a five-wicket haul match given that the analysis of the first 8 matches shows 3 five-wicket haul matches = $\frac{4-3}{15-8} = \frac{1}{7}$

Required Probability $= P(A \cap B) = P(A) \times P(B|A) = \frac{{}^4C_3 \times {}^{11}C_5}{{}^{15}C_8} \times \frac{1}{7} = \frac{8}{195}$

2. A and B predicts the outcomes of a cricket match and their chances of predicting the runs scored by a specific batsman correctly are $\frac{a}{b}$ and $\frac{c}{d}$ respectively independent of each other. If the probability of them predicting the same wrong score is $\frac{p}{q}$. Given that they predicted the same score, find the probability that their answer is correct.

   a. $\frac{q-p}{q}$

   b. $\frac{p(b-a)(d-c)}{qac+p(b-a)(d-c)}$

   c. $\frac{qac}{qac+p(b-a)(d-c)}$

   d. $\frac{p(b-a)(d-c)}{q}$

   Answer: c

   Solution:

   Let us define the following events:
   E_1: Both A and B predicted the score correctly.
   E_2: Exactly one of them predicted the score correctly.
   E_3: Neither of them predicted the score correctly.
   E: They predicted the same score.

   Now, $P(E_1) = \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}$; $P(E|E_1) = 1$

   $P(E_2) = \frac{a}{b} \times \frac{d-c}{d} + \frac{b-a}{b} \times \frac{c}{d} = \frac{a(d-c)+c(b-a)}{bd}$; $P(E|E_2) = 0$

   $P(E_3) = \frac{b-a}{b} \times \frac{d-c}{d} = \frac{(b-a)(d-c)}{bd}$; $P(E|E_3) = \frac{p}{q}$

   Hence, By Bayes' Rule:

   $$P(E_1|E) = \frac{P(E_1) \times P(E|E_1)}{P(E_1) \times P(E|E_1) + P(E_2) \times P(E|E_2) + P(E_3) \times P(E|E_3)}$$

   $$= \frac{\frac{ac}{bd} \times 1}{\frac{ac}{bd} \times 1 + \frac{a(d-c)+c(b-a)}{bd} \times 0 + \frac{(b-a)(d-c)}{bd} \times \frac{p}{q}}$$

   $$= \frac{qac}{qac+p(b-a)(d-c)}$$

   Hence, option (c) is correct.

For example: a=1, b=3, c=1, d=4, p=1 and q=228

Let us define the following events:

E_1: Both A and B predicted the score correctly.

E_2: Exactly one of them predicted the score correctly.

E_3: Neither of them predicted the score correctly.

E: They predicted the same score.

Now, $P(E_1) = \frac{1}{3} \times \frac{1}{4} = \frac{1}{12}$; $P(E|E_1) = 1$

$P(E_2) = \frac{1}{3} \times \frac{3}{4} + \frac{2}{3} \times \frac{1}{4} = \frac{5}{12}$; $P(E|E_2) = 0$

$P(E_3) = \frac{2}{3} \times \frac{3}{4} = \frac{6}{12}$; $P(E|E_3) = \frac{1}{228}$

Hence, By Bayes' Rule:

$$P(E_1|E) = \frac{P(E_1) \times P(E|E_1)}{P(E_1) \times P(E|E_1) + P(E_2) \times P(E|E_2) + P(E_3) \times P(E|E_3)}$$

$$= \frac{\frac{1}{12} \times 1}{\frac{1}{12} \times 1 + \frac{5}{12} \times 0 + \frac{6}{12} \times \frac{1}{228}} = \frac{\frac{1}{12} \times \frac{38}{38}}{\frac{1}{12} \times \frac{38}{38} + 0 + \frac{1}{12} \times \frac{1}{38}} = \frac{38}{39}$$

An item is produced in three factories A, B and C. Factory A produces x times the number of items produced by factory B, and the factories B and C produces the same number of items. It is known that $p\%$, $q\%$, $r\%$ of the items produced by factories A, B and C respectively are defective. All items produced in the three factories are stocked, and an item is selected at random. On the basis of given information, answer the questions (3) and (4).

3. What is the probability that the selected item is defective?(Enter the answer correct to two decimal places)

Solution:

Let the number of items produced by each of the factories B and C be n. Then the number of items produced by factory A is xn.

Let us define the events:

A= Item produced is defective.

E_1= Item is produced by factory A.

E_2= Item is produced by factory B.

E_3= Item is produced by factory C.

Now,

$P(E_1) = \frac{xn}{xn+n+n} = \frac{x}{x+2}$ and, $P(E_2) = \frac{n}{xn+n+n} = \frac{1}{x+2} = P(E_3)$

Also it is given in the question that,

$P(A|E_1) = \frac{p}{100}$, $P(A|E_2) = \frac{q}{100}$ and, $P(A|E_3) = \frac{r}{100}$

Hence,
P(Selected item at random is defective)= $P(A) = \sum_{i=1}^{3} P(A \cap E_i)$
$= \sum_{i=1}^{3} P(A|E_i) \times P(E_i)$
$= P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)$

$$= \left(\frac{x}{x+2} \times \frac{p}{100}\right) + \left(\frac{1}{x+2} \times \frac{q}{100}\right) + \left(\frac{1}{x+2} \times \frac{r}{100}\right)$$

For Example: $x = 5$**,** $p = 8$**,** $q = 7$ **and** $r = 4$
Let the number of items produced by each of the factories B and C be n. Then the number of items produced by factory A is $5n$.
Let us define the events:
A= Item produced is defective.
E_1= Item is produced by factory A.
E_2= Item is produced by factory B.
E_3= Item is produced by factory C.
Now,
$P(E_1) = \frac{5n}{5n+n+n} = \frac{5}{7}$ and, $P(E_2) = \frac{n}{5n+n+n} = \frac{1}{7} = P(E_3)$
Also it is given in the question that,
$P(A|E_1) = \frac{8}{100}$, $P(A|E_2) = \frac{7}{100}$ and, $P(A|E_3) = \frac{4}{100}$

Hence,
P(Selected item at random is defective)= $P(A) = \sum_{i=1}^{3} P(A \cap E_i)$
$= \sum_{i=1}^{3} P(A|E_i) \times P(E_i)$
$= P(E_1)P(A|E_1) + P(E_2)P(A|E_2) + P(E_3)P(A|E_3)$

$$= \left(\frac{5}{7} \times \frac{8}{100}\right) + \left(\frac{1}{7} \times \frac{7}{100}\right) + \left(\frac{1}{7} \times \frac{4}{100}\right) = \left(\frac{51}{700}\right) = 0.07$$

4. If an item selected at random is found to be defective, what is the probability that it was produced by factory B?(Enter the answer correct to two decimal places)
Solution:
P(Item is produced by factory B | Item is defective)= $P(E_2|A)$
By using the Bayes' Rule, we get:

$$P(E_2|A) = \frac{P(E_2)P(A|E_2)}{P(A)} = \frac{\frac{1}{x+2} \times \frac{q}{100}}{\left(\frac{x}{x+2} \times \frac{p}{100}\right) + \left(\frac{1}{x+2} \times \frac{q}{100}\right) + \left(\frac{1}{x+2} \times \frac{r}{100}\right)}$$

For Example: $x = 5$**,** $p = 8$**,** $q = 7$ **and** $r = 4$
P(Item is produced by factory B | Item is defective)= $P(E_2|A)$
By using the Bayes' Rule, we get:

$$P(E_2|A) = \frac{P(E_2)P(A|E_2)}{P(A)} = \frac{\frac{1}{7} \times \frac{7}{100}}{\left(\frac{5}{7} \times \frac{8}{100}\right) + \left(\frac{1}{7} \times \frac{7}{100}\right) + \left(\frac{1}{7} \times \frac{4}{100}\right)} = \frac{7}{51} = 0.14$$

5. A particular task is given to three persons, Manoj, Kalpana and Ananya whose probabilities of completing it are $\frac{a}{b}$,$\frac{c}{d}$ and $\frac{e}{f}$ respectively, independent of each other. What is the probability that the task will be completed? (Enter the answer correct to two decimal places)

   Solution:

   Let A,B,C denote the events that the task is completed by Manoj, Suresh and Kapil respectively. Then

   $P(A) = \frac{a}{b}$, $P(B) = \frac{c}{d}$, and $P(C) = \frac{e}{f}$

   The task will be completed if at least one of them completes the task.Thus, we have to calculate the probability of occurrence of at least one of the three events A,B,C, i.e, $P(A \cup B \cup C)$.

   Now, $P(A \cup B \cup C) = 1 - P(A \cup B \cup C)^c = 1 - P(A^c \cap B^c \cap C^c)$

   Since A,B,C are mutually independent $\implies$ A^c,B^c and C^c are mutually independent.

   Therefore, $P(A \cup B \cup C) = 1 - P(A^c)P(B^c)P(C^c)$

   $P(A \cup B \cup C) = 1 - (1 - \frac{a}{b})(1 - \frac{c}{d})(1 - \frac{e}{f})$

   For example:

   Let A,B,C denote the events that the task is completed by Manoj, Suresh and Kapil respectively. Then

   $P(A) = \frac{1}{4}$, $P(B) = \frac{2}{6}$, and $P(C) = \frac{1}{5}$

   The task will be completed if at least one of them completes the task. Thus, we have to calculate the probability of occurrence of at least one of the three events A,B,C, i.e, $P(A \cup B \cup C)$.

   Now, $P(A \cup B \cup C) = 1 - P(A \cup B \cup C)^c = 1 - P(A^c \cap B^c \cap C^c)$

   Since A,B,C are mutually independent $\implies$ A^c,B^c and C^c are mutually independent.

   Therefore, $P(A \cup B \cup C) = 1 - P(A^c)P(B^c)P(C^c)$

   $P(A \cup B \cup C) = 1 - (1 - \frac{1}{4})(1 - \frac{1}{3})(1 - \frac{1}{5}) = 1 - \frac{2}{5} = \frac{3}{5} = 0.6$

6. If A and B are two independent events such that $P(A^c) = m$ and $P(B^c) = x$ and $P(A \cup B) = n$, then calculate the value of x. (Enter the answer correct to two decimal places)

   Solution:

   We are given that $P(A^c) = m$ and $P(B^c) = x$ and $P(A \cup B) = n$.

   Now, $P(A \cup B) = 1 - P(A \cup B)^c = 1 - P(A^c \cap B^c)$

   $P(A \cup B) = 1 - P(A^c).P(B^c)$ [As Events are independent]

$n = 1 - mx$
$mx = 1 - n$
Hence, $x = \dfrac{1-n}{m}$
For example:
We are given that $P(A^c) = 0.6$ and $P(B^c) = x$ and $P(A \cup B) = 0.7$.
Now, $P(A \cup B) = 1 - P(A \cup B)^c = 1 - P(A^c \cap B^c)$
$P(A \cup B) = 1 - P(A^c).P(B^c)$ [As Events are independent]
$0.7 = 1 - 0.6x$
$0.6x = 1 - 0.7 = 0.3$
Hence, $x = \dfrac{0.3}{0.6} = 0.5$
Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels.The results are given Table Q8.1.G: Answer the questions 7, 8 and 9.

Researcher	No. of students in each level		
	Below Avg.	Avg.	Above Avg
X	a	b	c
Y	d	e	f

Table 1: *
Table Q8.1.G: Intelligence Level

7. What is the probability that a student falls in below average level? (Enter the answer correct to 2 decimal accuracy).
Solution:
From the table we have
Number of students investigated by researcher X =$a + b + c$
Number of students investigated by researcher Y=$d + e + f$
Total number of students = a+b+c+d+e+f
Now, P(Student falls in below average level)=$\dfrac{(a+d)}{(a+b+c+d+e+f)}$
For example: a = 76, b=70, c=54, d=50, e=23 and f=27
From the table we have
Number of students investigated by researcher X =$76 + 70 + 54 = 200$
Number of students investigated by researcher $Y = 50 + 23 + 27 = 100$
Total number of students = 300
Now, P(Student falls in below average level)=$\dfrac{76+50}{200+100} = \dfrac{126}{300} = 0.42$

8. What is the probability that a student is of average level given that the investigation

is done by researcher Y?(Enter the answer correct to 2 decimal accuracy).

Solution:

Let us define events;

A: Student is of average level

B: Investigation is done by researcher Y.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{e}{a+b+c+d+e+f}}{\frac{d+e+f}{a+b+c+d+e+f}} = \frac{e}{d+e+f}$$

For example:

Let us define events;

A: Student is of average level

B: Investigation is done by researcher Y.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{23}{300}}{\frac{100}{300}} = \frac{23}{100} = 0.23$$

9. What is the probability that investigation is done by researcher X given that the student is of below average level?(Enter the answer correct to 2 decimal accuracy).

Solution:

Let us define events;

C: Investigation is done by researcher X

D: Student is of below average level

$$P(C|D) = \frac{P(C \cap D)}{P(D)} = \frac{\frac{a}{a+b+c+d+e+f}}{\frac{a+d}{a+b+c+d+e+f}} = \frac{a}{a+d}$$

For example:

Let us define events;

C: Investigation is done by researcher X

D: Student is of below average level

$$P(C|D) = \frac{P(C \cap D)}{P(D)} = \frac{\frac{76}{300}}{\frac{126}{300}} = \frac{76}{126} = 0.6032$$

10. During the monsoon, it rains one-third of the days and affects students travel to school. The probability that there will be heavy traffic on a rainy day is 0.5 and on a non-rainy day is 0.25. If it rains and there is heavy traffic, the probability of a student arriving late to school is 0.5. If it is a clear day and there is no traffic, this probability is reduced by $\frac{3}{8}$. In other possible situations, the probability of a student reaching school late 0.25. If on a randomly selected day, a student arrives late to school, then what is the

probability that it rained that day?
Hint: Consider the event as,
H= There is heavy traffic
H^c= There is no traffic.

a. $\frac{1}{8}$
b. $\frac{6}{11}$
c. $\frac{11}{48}$
d. None of the above

Answer: b
Solution:
Let us define the following events:
H: There is heavy traffic.
R: Rainy-day.
E: Student is late for school.
Now, we are given that, $P(R) = \frac{1}{3}$

$P(H|R) = 0.5$ and, $P(H|R^c) = 0.25$
$P(E|R \cap H) = 0.5$, $P(E|R^c \cap H^c) = 0.5 - \frac{3}{8} = \frac{1}{8}$ and $P(E|R^c \cap H) = P(E|R \cap H^c) = 0.25$
$P(R \cap H \cap E) = P(R)P(H|R)P(E|R \cap H) = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{12}$

Similarly,
$P(R^c \cap H \cap E) = P(R^c)P(H|R^c)P(E|R^c \cap H) = \frac{2}{3} \times \frac{1}{4} \times \frac{1}{4} = \frac{2}{48}$
$P(R \cap H^c \cap E) = P(R)P(H^c|R)P(E|R \cap H^c) = \frac{1}{3} \times \left(1 - \frac{1}{2}\right) \times \frac{1}{4} = \frac{1}{24}$
$P(R^c \cap H^c \cap E) = P(R^c)P(H^c|R^c)P(E|R^c \cap H^c) = \frac{2}{3} \times \frac{3}{4} \times \frac{1}{8} = \frac{3}{48}$

Therefore, $P(E) = P(R \cap H \cap E) + P(R^c \cap H \cap E) + P(R \cap H^c \cap E) + P(R^c \cap H^c \cap E)$
$= \frac{1}{12} + \frac{2}{48} + \frac{1}{24} + \frac{3}{48} = \frac{11}{48}$

Also, $P(R \cap E) = P(R \cap H \cap E) + P(R \cap H^c \cap E) = \frac{1}{12} + \frac{1}{24} = \frac{3}{24}$

Now, Required Probability; $P(R|E) = \frac{P(R \cap E)}{P(E)} = \frac{\frac{3}{24}}{\frac{11}{48}} = \frac{6}{11}$

Hence, option (b) is correct.

11. There are two shops, A and B, selling t-shirts in the market. Shop A has stock of n red and 2 black t-shirts and Shop B has a stock of 2 red and n black t-shirts. One of the shops is selected at random and two t-shirts are purchased from it. If both the t-shirts purchased are red and the probability that it was purchased from a shop A is $\frac{6}{7}$, find the value of n.

a. 3

b. 4

c. 5

d. 6

Answer:b

Solution:

Let us define the following events:
$E1$: Shop A is selected.
$E2$: Shop B is selected.
E: Two t-shirts purchased are of red color.

Now, $P(E1) = P(E2) = \frac{1}{2}$

$P(E|E1) = \frac{{}^{n}C_2}{{}^{n+2}C_2} = \frac{n(n-1)}{(n+2)(n+1)}$

and, $P(E|E2) = \frac{{}^{2}C_2}{{}^{n+2}C_2} = \frac{2}{(n+2)(n+1)}$

Using Bayes' Theorem, we get:

$$P(E1|E) = \frac{P(E1)P(E|E1)}{P(E1)P(E|E1) + P(E2)P(E|E2)} = \frac{6}{7} \qquad \text{(Given)}$$

$$\implies \frac{\frac{1}{2} \times \frac{n(n-1)}{(n+2)(n+1)}}{\frac{1}{2} \times \frac{n(n-1)}{(n+2)(n+1)} + \frac{1}{2} \times \frac{2}{(n+2)(n+1)}} = \frac{6}{7}$$

$$\implies \frac{n(n-1)}{n(n-1)+2} = \frac{6}{7}$$

$7n(n-1) = 6n(n-1) + 12 \implies n^2 - n - 12 = 0$
Therefore, $n = 4, -3$. Since, n cannot be negative. So, $n = 4$
Hence, option (b) is correct.

Statistics for Data Science-1

Week-9 Graded Assignment

1. A discrete random variable X can take the values $1, 2, 3, \cdots, n$. For these values the cumulative distribution function is defined by:
   $F(x) = P(X \leq x) = \dfrac{x^2 + k}{m}$; $x = 1, 2, 3, \cdots, n$
   Find the value of k.
   Answer: $k = m - n^2$
   Solution:
   $F(n) = P(X \leq n) = 1$
   $\implies \dfrac{n^2 + k}{m} = 1$
   Hence, $k = m - n^2$
   Suppose, we substitute values of n and m as 3 and 40 respectively, then

$$\frac{3^2 + k}{40} = 1$$
$$k = 31$$

2. An organization in Texas organizes a lucky draw this month. n thousand tickets are sold for $m\$$ each. Each has an equal chance of winning. x tickets will win $a\$$, y tickets will win $b\$$ and z tickets will win $c\$$. Let, the random variable X denote the net gain from purchase of one ticket. What is the probability that X takes a value less than b? (Enter the answer correct to 4 decimal place)
   Answer: $\dfrac{n \times 1000 - x}{n \times 1000}$
   Solution:
   X can take values $-m$, $c - m$, $b - m$ and $a - m$
   $P(X < b) = P(X = b - m) + P(X = c - m) + P(X = -m)$
   $P(X < b) = \dfrac{y}{n \times 1000} + \dfrac{z}{n \times 1000} + \dfrac{n \times 1000 - x - y - z}{n \times 1000}$
   $P(X < b) = \dfrac{n \times 1000 - x}{n \times 1000}$

   Suppose, we substitute values of n, m, x, a, y, b, z and c as 5, 1, 1, 1000, 2 500, 10 and 100 respectively, then
   $P(X < 500) = P(X = 499) + P(X = 99) + P(X = -1) = \dfrac{2}{5000} + \dfrac{10}{5000} + \dfrac{4987}{5000}$
   Therefore, $P(X < 500) = \dfrac{4999}{5000} = 0.9998$

3. In a group of n people, x are photographers and $n - x$ are journalists. m people are randomly picked from a group of these n people. Let, Y be a random variable which represents the number of photographers. How many possible values can the random variable Y take?
Answer: $m + 1$
Solution:
Possible values of Y are $0, 1, 2, ..., m$.
Hence, the number of possible values Y can take is $m + 1$.
Suppose, we substitute values of m, x and n as 8, 240 and 15 respectively, then possible values of Y are $0, 1, 2, ..., 8$
Hence, the number of possible values Y can take is 9.

4. Which of the following is/are discrete random variables?

   a. Number of tires produced in an automotive tire factory every 30 minutes.

   b. The number of kernels(pieces) of popcorn in a 1 kg container.

   c. The time between customers entering a checkout lane at a retail store.

   d. The amount of rain recorded at an airport one day.

   e. The amount of liquid in a 2 $litres$ bottle of soft drink.

   f. The number of no-shows for every 1000 reservations made with a commercial airline.

   Answer: a, b, f
   Solution:
   The number of tires produced in an automotive tire factory every 30 minutes can have countable possible values, and hence it denotes a discrete random variable.
   Hence, option (a) is correct.

   The number of kernels of popcorn in a 1 kg container also have countable possible values, it cannot take all values between some interval and hence it is a discrete random variable. So option (b) is correct.

   The time between customers entering a checkout lane at a retail store can take any values between some interval. Hence, it is a continuous random variable.
   So, option (c) is incorrect.

   Again, the amount of rain recorded at an airport one day and the amount of liquid in a 2 $litres$ bottle of soft drink can take any values between some interval. Hence, they are continuous random variable.
   So, option (d) and (e) are incorrect.

The number of no-shows for every 1000 reservations made with a commercial airline can have countable possible values, and hence it denotes a discrete random variable. Hence, option (f) is correct.

5. A biased coin with probability of heads 0.75 is tossed three times. Let X be a random variable that represents the number of head runs, a head run being defined as a consecutive occurrence of at least two heads. Then the probability mass function of X is given by:

a.
$$P(X = x) = \begin{cases} 0.375 & \text{for } x = 0 \\ 0.625 & \text{for } x = 1 \end{cases}$$

b.
$$P(X = x) = \begin{cases} 0.297 & \text{for } x = 0 \\ 0.703 & \text{for } x = 1 \end{cases}$$

c.
$$P(X = x) = \begin{cases} 0.016 & \text{for } x = 0 \\ 0.140 & \text{for } x = 1 \\ 0.422 & \text{for } x = 2 \\ 0.422 & \text{for } x = 3 \end{cases}$$

d.
$$P(X = x) = \begin{cases} 0.016 & \text{for } x = 0 \\ 0.844 & \text{for } x = 1 \\ 0.140 & \text{for } x = 2 \end{cases}$$

Answer: b
Solution:

Possible outcomes	X	$P(X = x)$
HHH	1	0.422
HHT	1	0.141
HTH	0	0.141
HTT	0	0.047
THH	1	0.141
THT	0	0.047
TTH	0	0.047
TTT	0	0.016

Table 9.1

Hence, the probability mass function of X is given by:

$$P(X = x) = \begin{cases} 0.297 & \text{for } x = 0 \\ 0.703 & \text{for } x = 1 \end{cases}$$

6. Nina has n music sessions in a week. She attends the sessions n days a week $x\%$ of the time, $n-1$ days $y\%$ of the time, one day $z\%$ of the time, and no days $p\%$ of the time. Let, X be a discrete random variable representing the number of sessions she attends in a week. Suppose one week is randomly selected, what is the probability that the random variable X takes the value at most $n-1$?(Enter the answer correct to 2 decimal places)

Answer: $1 - \frac{x}{100}$

Solution:

The pmf of random variable X is given by:

$$P(X = k) = \begin{cases} \frac{x}{100} & \text{for } k = n \\ \frac{y}{100} & \text{for } k = n-1 \\ \frac{z}{100} & \text{for } k = 1 \\ \frac{p}{100} & \text{for } k = 0 \end{cases}$$

$$\begin{aligned} P(X \leq n-1) &= P(X=0) + P(X=1) + P(X=n-1) \\ &= \frac{p}{100} + \frac{z}{100} + \frac{y}{100} \\ &= \frac{p+y+z}{100} \\ &= 1 - \frac{x}{100} \end{aligned}$$

Suppose, we substitute values of n, x, y, z and p as 5, 50, 20, 10 and 20 respectively, then

The pmf of random variable X is given by:

$$P(X = k) = \begin{cases} 0.5 & \text{for k=5} \\ 0.2 & \text{for k=4} \\ 0.1 & \text{for k=1} \\ 0.2 & \text{for k=0} \end{cases}$$

$$\begin{aligned} P(X \leq 4) &= P(X=0) + P(X=1) + P(X=4) \\ &= 0.2 + 0.1 + 0.2 \\ &= 0.5 \end{aligned}$$

7. Find the value of k for which $k\left(\frac{m}{n}\right)^x$ (x = 0, 1, 2, ...) is a pmf. (Enter the answer correct up to 2 decimal places)

   Answer: $\frac{n-m}{n}$

   Solution:

   For pmf: $k\left[\left(\frac{m}{n}\right)^0 + \left(\frac{m}{n}\right)^1 + \left(\frac{m}{n}\right)^2 + \cdots\right] = 1$

   $\implies k.\frac{1}{1-\frac{m}{n}} = 1$

   $\implies k.\frac{n}{n-m} = 1$

   Therefore, $k = \frac{n-m}{n}$.

   For example:

   Take $m = 3$ and $n = 8$. For pmf: $k\left[\left(\frac{3}{8}\right)^0 + \left(\frac{3}{8}\right)^1 + \left(\frac{3}{8}\right)^2 + \cdots\right] = 1$

   $\implies k.\frac{1}{1-\frac{3}{8}} = 1$

   $\implies k.\frac{8}{5} = 1$

   Therefore, $k = \frac{5}{8}$.

8. Using the information in the previous question, calculate $P(X = 2)$. (Enter the answer correct up to 2 decimal places)

   Answer: $\frac{(n-m)}{n}.\left(\frac{m}{n}\right)^2$

   Solution:

   $P(X=2) = \frac{(n-m)}{n}.\left(\frac{m}{n}\right)^2$.

   For example:

   Take $m = 3$ and $n = 8$. For pmf: $k\left[\left(\frac{3}{8}\right)^0 + \left(\frac{3}{8}\right)^1 + \left(\frac{3}{8}\right)^2 + \cdots\right] = 1$

$$\implies k.\frac{1}{1-\frac{3}{8}} = 1$$

$$\implies k.\frac{8}{5} = 1$$

Therefore, $k = \frac{5}{8}$.

And, $P(X=2) = \frac{5}{8}.\left(\frac{3}{8}\right)^2 = 0.09$.

9. From a box A containing 3 white and 6 black balls, 5 balls are transferred into an empty box B. Let X be a random variable that represents the number of white balls which are transferred from A to B. What value of random variable will have the least probability?

Answer: 0

Solution:

Let us define the following cases:
Transfer of 0 white and 5 black balls.
Transfer of 1 white and 4 black balls.
Transfer of 2 white and 3 black balls.
Transfer of 3 white and 2 black balls.
Probabilities for all cases:

(i) $P(X=0) = \frac{{}^6C_5}{{}^9C_5} = 0.048$

(ii) $P(X=1) = \frac{{}^3C_1\,{}^6C_4}{{}^9C_5} = 0.357$

(iii) $P(X=2) = \frac{{}^3C_2\,{}^6C_3}{{}^9C_5} = 0.476$

(iv) $P(X=3) = \frac{{}^3C_3\,{}^6C_2}{{}^9C_5} = 0.119$

.

Thus, $X=0$ has the least probability.

10. The probability mass function of a random variable X is given by:

$$P(X=x) = \begin{cases} 3k^2 - 3k & \text{for } x = 0 \\ 2k^2 - 1 & \text{for } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Determine the value of k given $k > 0$.

Answer:1

Solution:

From properties of pmf,

$$p(0) + p(1) = 1$$
$$3k^2 - 3k + 2k^2 - 1 = 1$$
$$5k^2 - 3k - 2 = 0$$
$$5k^2 - 5k + 2k - 2 = 0$$
$$5k(k-1) + 2(k-1) = 0$$
$$(5k+2)(k-1) = 0$$
$$k = \frac{-2}{5} \text{ or } k = 1$$

As $k > 0$, therefore $k = 1$.

Statistics for Data Science-1

Week-10 Graded Assignment

1. There are 2^n numbered cards in a deck among which ${}^nC_i$ cards bear the number i ; i=0,1,2,...,n. From the deck, m cards are drawn with replacement. What is the expectation of the sum of their numbers? (Enter the answer correct to one decimal accuracy)

   Answer: $\frac{mn}{2}$

   Solution:

   Let, X_j; i=1,2,...,m be the random variable representing the number on the j^{th} card drawn.

X_j	0	1	2	...	n
Number of cards	${}^nC_0$	${}^nC_1$	${}^nC_2$	...	${}^nC_n$
$P(X_j = x)$	$\frac{{}^nC_0}{2^n}$	$\frac{{}^nC_1}{2^n}$	$\frac{{}^nC_2}{2^n}$	...	$\frac{{}^nC_n}{2^n}$

Table 10.1

$$E(X_j) = \sum_{x=0}^{n} xP(X_j = x) = \sum_{x=0}^{n} x \times \frac{{}^nC_x}{2^n}$$

$$E(X_j) = \frac{1}{2^n} \times [(0 \times {}^nC_0) + (1 \times {}^nC_1) + (2 \times {}^nC_2) + ... + (n \times {}^nC_n)]$$

$$E(X_j) = \frac{1}{2^n} \times [0 + (1 \times n) + \left(2 \times \frac{n(n-1)}{2!}\right) + \left(3 \times \frac{n(n-1)(n-2)}{3!}\right) ... + (n \times 1)]$$

$$E(X_j) = \frac{n}{2^n} \times [1 + (n-1) + \left(\frac{(n-1)(n-2)}{2!}\right) + ... + 1]$$

$$E(X_j) = \frac{n}{2^n} \times [{}^{n-1}C_0 + {}^{n-1}C_1 + {}^{n-1}C_2 + ... + {}^{n-1}C_{n-1}]$$

$$E(X_j) = \frac{n}{2^n} \times 2^{n-1} = \frac{n}{2}$$

Therefore, Expectation of the sum of their numbers is given by:

$= \sum_{j=1}^{m} E(X_j)$

$= \sum_{j=1}^{m} \frac{n}{2}$

$= \frac{mn}{2}$

Suppose, we substitute values of n and m as 4 and 7 respectively, then

Let, X_j; i=1,2,...,7 be the random variable representing the number on the j^{th} card drawn.

X_j	0	1	2	3	4
Number of cards	${}^4C_0$	${}^4C_1$	${}^4C_2$	${}^4C_3$	${}^4C_4$
$P(X_j = x)$	$\frac{{}^4C_0}{2^4}$	$\frac{{}^4C_1}{2^4}$	$\frac{{}^4C_2}{2^4}$	$\frac{{}^4C_3}{2^4}$	$\frac{{}^4C_4}{2^4}$

Table 10.1

$E(X_j) = \sum_{x=0}^{4} xP(X_j = x) = \sum_{x=0}^{4} x \times \frac{{}^4C_x}{2^4}$

$E(X_j) = \frac{1}{2^4} \times [(0 \times {}^4C_0) + (1 \times {}^4C_1) + (2 \times {}^4C_2) + (3 \times {}^4C_3) + (4 \times {}^4C_4)]$

$E(X_j) = \frac{1}{2^4} \times [0 + (1 \times 4) + \left(2 \times \frac{4(4-1)}{2!}\right) + \left(3 \times \frac{4(4-1)(4-2)}{3!}\right) + (4 \times 1)]$

$E(X_j) = \frac{4}{2^4} \times [1 + (4-1) + \left(\frac{(4-1)(4-2)}{2!}\right) + 1]$

$E(X_j) = \frac{4}{2^4} \times [{}^3C_0 + {}^3C_1 + {}^3C_2 + {}^3C_3]$

$E(X_j) = \frac{4}{2^4} \times 2^3 = \frac{4}{2} = 2$

Therefore, Expectation of the sum of their numbers is given by:
$= \sum_{j=1}^{7} E(X_j) = \sum_{j=1}^{7} 2 = 14$

An unbiased die is thrown $n+2$ times. After each throw a '+' is recorded for 2 or 5 and '-' is recorded for 1,3,4 or 6, the signs forming an ordered sequence. To each, except the first and last sign, a random variable X_i; i=1,2,...,n is associated which takes the value 1 if both of its neighbouring sign differs from the one between them and 0 otherwise. If the random variable Y is defined as $Y = aS + b$ where, $S = \sum_{i=1}^{n} X_i$, then use the given information to answer question (2) and (3).

2. Find the expected value of Y .

   (a) $\left(a \times \frac{2n}{9}\right) + b$

   (b) $\left(a \times \frac{2n}{9}\right)$

   (c) $\frac{2n}{9} + b$

   (d) $\frac{2n}{9}$

Answer: a
Solution:

$$X_i = \begin{cases} 1 & \text{if the pattern is either '+ - +' or '- + -'} \\ 0 & \text{otherwise} \end{cases}$$

$E(X_i) = 1 \times P(X_i = 1) + 0 \times P(X_i = 0) = P(X_i = 1)$
Now,
$P(X_i = 1)$ =P(Pattern = '+ - +') + P(pattern='- + -')

$$P(X_i = 1) = \left(\frac{4}{6}\right)^2 \times \left(\frac{2}{6}\right) + \left(\frac{4}{6}\right) \times \left(\frac{2}{6}\right)^2 = \frac{2}{9}$$

$E(S) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$

$$E(S) = \frac{2n}{9}$$

Therefore, $E(Y) = aE(S) + b$

$$E(Y) = \left(a \times \frac{2n}{9}\right) + b$$

Suppose, we substitute values of n, a and b as 10, 9 and 1 respectively then,

$E(S) = E(\sum_{i=1}^{10} X_i) = \sum_{i=1}^{10} E(X_i)$

$$E(S) = \sum_{i=1}^{10} \frac{2}{9} = \frac{20}{9}$$

Therefore, $E(Y) = 9E(S) + 1 = 9 \times \frac{20}{9} + 1$
$E(Y) = 20 + 1 = 21$

3. Which of the following statement(s) is/are true?

a. $V(Y) = a^2V(S) + b$
b. $V(Y) = a^2V(S)$
c. $V(Y) \neq a^2V(S)$
d. $E(Y) = a^2E(S) + b$
e. $E(Y) = aE(S) + b$

Answer: b,e
By the property of Expectation and Variance, we get:
$V(Y) = a^2V(S)$
$E(Y) = aE(S) + b$ (always)
Hence, option (b) and (e) is correct.

Suppose, we substitute values of a and b as 9 and 1 respectively then,

$V(Y) = (9)^2 \times V(S) = 81V(S)$
$E(Y) = 9E(S) + 1$ (always)

Amandeep is in the middle of a bridge of infinite length. He takes the unit step to the right with probability p and to the left with probability $1 - p$. Assume that the movements are independent of each other.
Hint: Consider the random variable X_i associated with the i^{th} step defined as:

$$X_i = \begin{cases} 1 & \text{if the step of Amandeep is towards the right} \\ -1 & \text{if the step of Amandeep is towards the left} \end{cases}$$

Using this information, answer question (4) and (5).

4. What is the expected distance between the starting point and end point of Amandeep after n steps?

(a). $2p - 1$

(b). $n(2p - 1)$

(c). $1 - 2p$

(d). $n(1 - 2p)$

Answer: *b*
Solution:
Let us associate a random variable X_i with the i^{th} step.

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ step of Amandeep is towards the right} \\ -1 & \text{if } i^{th} \text{ step of Amandeep is towards the left} \end{cases}$$

$E(X_i) = 1 \times P(X_i = 1) + (-1) \times P(X_i = -1)$
$E(X_i) = 1 \times p - 1 \times (1 - p) = 2p - 1$
$S = X_1 + X_2 + ... + X_n$ represents the random distance moved from the starting point after n steps.
Therefore, $E(S) = \sum_{i=1}^{n} E(X_i) = n(2p - 1)$
Hence, expected distance between the starting point and end point of Amandeep after n steps is $n(2p - 1)$

For example: p=0.6 , n=7

Let us associate a random variable X_i with the i^{th} step.

$$X_i = \begin{cases} 1 & \text{if } i^{th} \text{ step of Amandeep is towards the right} \\ -1 & \text{if } i^{th} \text{ step of Amandeep is towards the left} \end{cases}$$

$E(X_i) = 1 \times P(X_i = 1) + (-1) \times P(X_i = -1)$
$E(X_i) = 1 \times 0.6 - 1 \times 0.4 = 0.2$

$S = X_1 + X_2 + ... + X_7$ represents the random distance moved from the starting point after 7 steps.
Therefore, $E(S) = \sum_{i=1}^{7} E(X_i) = 7(0.2) = 1.4$
Hence, the expected distance between the starting point and end point of Amandeep after 7 steps is 1.4.

5. What is the variance distance between the starting point and end point of Amandeep after n steps?

   (a). $4np(1-p)$

   (b). $4p(1-p)$

   (c). 1

   (d). $n^2(2p-1)^2 - 1$

   Answer: a
   Solution:
   $E(X_i)^2 = 1^2 \times P(X_i = 1) + (-1)^2 \times P(X_i = -1)$
   $E(X_i)^2 = p + (1-p) = 1$
   $V(X_i) = E(X_i)^2 - [E(X_i)]^2$
   $V(X_i) = 1 - (2p-1)^2 = 4p(1-p)$
   Therefore, $V(S) = \sum_{i=1}^{n} V(X_i)$ (because, movements of steps are independent)
   Hence, $V(S) = \sum_{i=1}^{n} 4p(1-p) = 4np(1-p)$
   For Example: p=0.6 , n=7
   $E(X_i)^2 = 1^2 \times P(X_i = 1) + (-1)^2 \times P(X_i = -1)$
   $E(X_i)^2 = 0.6 + 0.4 = 1$
   $V(X_i) = E(X_i)^2 - [E(X_i)]^2$
   $V(X_i) = 1 - (0.2)^2 = 0.96$
   Therefore, $V(S) = \sum_{i=1}^{7} V(X_i)$
   Because, movements of steps are independent.
   Hence, $V(S) = \sum_{i=1}^{7} (0.96) = 7 \times (0.96) = 6.72$

6. A box contains a white and b black balls. c balls are drawn at random without replacement. Find the expected value of the number of white balls drawn?(Enter the answer correct to 2 decimal places).
   Solution:
   Let X denote the number of white balls drawn. The probability distribution of X is obtained as follows:

X	0	1	2	...	c
$p(x)$	$\frac{{}^{b}C_{c}}{{}^{a+b}C_{c}}$	$\frac{{}^{a}C_{1} \times {}^{b}C_{c-1}}{{}^{a+b}C_{c}}$	$\frac{{}^{a}C_{2} \times {}^{b}C_{c-2}}{{}^{a+b}C_{c}}$	...	$\frac{{}^{a}C_{c}}{{}^{a+b}C_{c}}$

Then expected number of white balls drawn is :
$E(X) = 0 \times \frac{{}^{b}C_{c}}{{}^{a+b}C_{c}} + 1 \times \frac{{}^{a}C_{1} \times {}^{b}C_{c-1}}{{}^{a+b}C_{c}} + 2 \times \frac{{}^{a}C_{2} \times {}^{b}C_{c-2}}{{}^{a+b}C_{c}} + ... + c \times \frac{{}^{a}C_{c}}{{}^{a+b}C_{c}}$

For example: a=7, b=4, c=2
Let X denote the number of white balls drawn. The probability distribution of X is obtained as follows:

X	0	1	2
$p(x)$	$\frac{{}^{4}C_{2}}{{}^{11}C_{2}} = \frac{6}{55}$	$\frac{{}^{7}C_{1} \times {}^{4}C_{1}}{{}^{11}C_{2}} = \frac{28}{55}$	$\frac{{}^{7}C_{2}}{{}^{11}C_{2}} = \frac{21}{55}$

Then expected number of white balls drawn is :
$E(X) = 0 \times \frac{6}{55} + 1 \times \frac{28}{55} + 2 \times \frac{21}{55} = \frac{70}{55} = 1.27.$

7. Rohit wants to open his door with 5 keys(out of which 1 will open the door) and tries the keys independently and at random. If unsuccessful keys are eliminated from further selection, then Find the expected number of trials required to open the door.
(Hint: Suppose Rohit gets the first success at x^{th} trial, i.e., he is unable to open the door in the first $(x-1)$ trials. And, P(he gets first success at second trial)=$(1-\frac{1}{5}) \times \frac{1}{4}$)

(a). 1

(b). 9

(c). 3

(d). 2

Answer:c
Solution:

If unsuccessful keys are eliminated from further selection, then the random variable X will take the values from 1 to n. In this case, we have
Probability of success at the first trial=$\frac{1}{5}$
Probability of success at the second trial=$\frac{1}{4}$
Probability of success at the third trial=$\frac{1}{3}$
Probability of success at the fourth trial=$\frac{1}{2}$
Probability of success at the fifth trial=1 and so on.
Hence probability of 1^{st} success at the 2^{nd} trial $= (1 - \frac{1}{5}) \times \frac{1}{4} = \frac{1}{5}$

probability of 1^{st} success at the 3^{rd} trial $= (1-\frac{1}{5}) \times (1-\frac{1}{4}) \times \frac{1}{3} = \frac{1}{5}$

and so on. In general,

$p(x)$= probability of 1^{st} success at the x^{th} trial=$\frac{1}{5}$; $x = 1, 2, 3, 4, 5$

Therefore,

$$E(X) = \sum_{x=1}^{5} xp(x) = 1 \times \frac{1}{5} + 2 \times \frac{1}{5} + 3 \times \frac{1}{5} + 4 \times \frac{1}{5} + 5 \times \frac{1}{5} = 3$$

8. X and Y are independent random variables with means m_1 and m_2, and variances v_1 and v_2 respectively. Find the variance of $aX + bY$?

   Answer: $a^2 \times v_1 + b^2 \times v_2$

   Solution:

   Since X and Y are independent random variables.

   Therefore, $V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y) = a^2 \times v_1 + b^2 \times v_2$

   For example: $m_1 = 10$, $m_2 = 20$, $v_1 = 2$ **and** $v_2 = 3$

   Since X and Y are independent random variables.

   Therefore, $V(3X + 4Y) = 9 \times V(X) + 16 \times V(Y) = 9 \times 2 + 16 \times 3 = 18 + 48 = 66$

9. Let X be a random variable with the following probability distribution:

X	a	b	c
$P(X = x)$	$\frac{1}{d}$	$\frac{1}{e}$	$\frac{f}{g}$

   Calculate the value of $E(2X + 1)^2$.(Enter the answer correct to 2 decimal places)

   Answer: $4 \times \left(a^2 \times \frac{1}{d} + b^2 \times \frac{1}{e} + c^2 \times \frac{1}{f}\right) + 4 \times \left(a \times \frac{1}{d} + b \times \frac{1}{e} + c \times \frac{1}{f}\right) + 1$

   Solution:

   $E(X) = \Sigma xp(x) = a \times \frac{1}{d} + b \times \frac{1}{e} + c \times \frac{1}{f}$.

   $E(X^2) = \Sigma x^2 p(x) = a^2 \times \frac{1}{d} + b^2 \times \frac{1}{e} + c^2 \times \frac{1}{f}$

   $E(2X + 1)^2 = E(4X^2 + 4X + 1) = 4E(X^2) + 4E(X) + 1$

   For example: $a = -3, b = 6, c = 9, d = 6, e = 2$ and $f = 3$

X	-3	6	9
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

   $E(X) = \Sigma xp(x) = (-3) \times \frac{1}{6} + 6 \times \frac{1}{2} + 9 \times \frac{1}{3} = \frac{11}{2}$.

$$E(X^2) = \Sigma x^2 p(x) = 9 \times \frac{1}{6} + 36 \times \frac{1}{2} + 81 \times \frac{1}{3} = \frac{93}{2}$$

$$E(2X+1)^2 = E(4X^2 + 4X + 1) = 4E(X^2) + 4E(X) + 1 = 4 \times \frac{93}{2} + 4 \times \frac{11}{2} + 1 = 209$$

10. Suppose that X is a random variable for which $E(X) = m$ and $Var(X) = v$. Find the positive values of a and b such that $Y = aX - b$, has expectation 0 and variance 1.

a. $\frac{-1}{\sqrt{v}}, \frac{-m}{\sqrt{v}}$

b. $\frac{1}{v}, \frac{m}{\sqrt{v}}$

c. $\frac{1}{\sqrt{v}}, \frac{m}{\sqrt{v}}$

d. $1, m$

Answer: c

Solution:

$E(Y) = aE(X) - b = 0 \implies ma - b = 0 \implies ma = b$ ...(1)

Now,$V(Y) = a^2V(X) = 1 \implies a^2 \times v = 1 \implies a^2 = \frac{1}{v} \implies a = \frac{1}{\sqrt{v}}$

Now putting value of a in equation(1) , we get $b = \frac{m}{\sqrt{v}}$

Hence, option(c) is correct.

For example: m=10 and v=25

$E(Y) = aE(X) - b = 0 \implies 10a - b = 0 \implies 10a = b$ ...(1)

Now,$V(Y) = a^2V(X) = 1 \implies a^2 \times 25 = 1 \implies a^2 = \frac{1}{25} \implies a = \frac{1}{5}$

Now putting value of a in equation(1) , we get $b = 2$

Statistics for Data Science-1

Week-11 Graded Assignment

1. A match predictor claims that he can predict the result of a match correctly $x\%$ of the time. It is agreed that his claim will be accepted if he correctly predicts the results of at least m of n matches. What is the probability that his claim gets rejected?

   (a) $\sum_{i=0}^{m-1} {}^nC_i \ p^i(1-p)^{n-i}$
   (b) $\sum_{i=0}^{m} {}^nC_i \ p^i(1-p)^{n-i}$
   (c) $\sum_{i=1}^{m-1} {}^nC_i \ p^i(1-p)^{n-i}$
   (d) $\sum_{i=1}^{m} {}^nC_i \ p^i(1-p)^{n-i}$

   Answer: a
   Solution:
   P(Correct match prediction)$= p = \dfrac{x}{100}$
   Let, X be a random variable representing the number of correct match predictions.
   Hence, $X \sim Binomial(n, p)$
   The probability that his claim gets rejected is given by:

$$\begin{aligned} P(X < m) &= \sum_{i=0}^{m-1} P(X = i) \\ &= \sum_{i=0}^{m-1} {}^nC_i \ p^i(1-p)^{n-i} \end{aligned}$$

   Suppose we substitute the values of x, m and n as 75, 5 and 6 respectively, then

   P(Correct match prediction)$= p = 0.75$
   Let, X be a random variable representing the number of correct match predictions.
   Hence, $X \sim Binomial(6, 0.75)$
   The probability that his claim gets rejected is given by:

$$\begin{aligned} P(X < 5) &= \sum_{i=0}^{5-1} P(X = i) \\ &= \sum_{i=0}^{4} {}^6C_i \ 0.75^i(1-0.75)^{6-i} \\ &= 0.466 \end{aligned}$$

   Therefore, the probability that his claim gets rejected is 0.47

2. If $X \sim Binomial(n, p)$, then which of the following statement/s is/are always true? ($n > 0$ and $0 < p < 1$)

   (a) $E(X) \leq Var(X)$

   (b) $E(X) < Var(X)$

   (c) $E(X) \geq Var(X)$

   (d) $E(X) > Var(X)$

   (e) $Var(X) \leq S.D(X)$

   (f) $Var(X) \geq S.D(X)$

   Answer: d

   Solution:

   If $X \sim Binomial(n, p)$, then

   $E(X) = np$ and $Var(X) = np(1-p)$

   When, $0 < p < 1$ and $n > 0$, $np > np(1-p)$(always)

   Thus, option(d) is always true.

   For example: $n = 1$ **and** $p = \frac{1}{2}$

   $np = 1 \times \frac{1}{2} = \frac{1}{2}$ and $np(1-p) = 1 \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

   Since, $\frac{1}{2} > \frac{1}{4}$, thus $E(X) > Var(X)$

   Option (e) and (f) are not always true.

   For example: $n = 3$ and $p = \frac{1}{2}$

   $Var(X) = np(1-p) = 3 \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{4} = 0.75$

   $S.D(X) = \sqrt{Var(X)} = \sqrt{0.75} = 0.87$, which is greater than variance in this particular example.

   But, for $n = 16$ and $p = \frac{1}{2}$

   $Var(X) = np(1-p) = 16 \times \frac{1}{2} \times \frac{1}{2} = \frac{16}{4} = 4$

   $S.D(X) = \sqrt{Var(X)} = \sqrt{4} = 2$, which is less than variance in this particular example.

   Hence, we can say that option (e) and (f) are not always true.

3. Two friends (say 'A' and 'B') could not decide whether to play a racing game or a boxing game on Xbox. They decide to play a card-game first. If 'A' wins at least x rounds out of the y rounds of the card game played, then the boxing game will be played. The chances of 'A' winning in any round of the card game is $a : b$. Find the probability that the boxing game will be played on Xbox? (Enter the answer correct

to 2 decimal places)
Hint: If the chances of happening of an event is $x : y$ then, the probability equals $\frac{x}{x+y}$

Answer: $\sum_{i=x}^{y} {}^{y}C_i \ p^i(1-p)^{y-i}$

Solution:

Number of rounds of card game played; $n = y$

P('A' winning a round of card game)$= p = \frac{a}{a+b}$

Let, X be a random variable representing the number of rounds of card game won by 'A'.

Hence, $X \sim Binomial(n, p)$

Now, the probability that the boxing game will be played is given by:

$$\begin{aligned} P(X \geq x) &= \sum_{i=x}^{y} P(X = i) \\ &= \sum_{i=x}^{y} {}^{y}C_i \ p^i(1-p)^{y-i} \end{aligned}$$

Suppose we substitute the values of x, y, a and b as 3, 5, 3 and 2 respectively, then

Number of rounds of card game played; $n = 5$

P('A' winning a round of card game)$= p = \frac{3}{3+2} = 0.6$

Let, X be a random variable representing the number of rounds of card game won by 'A'.

Hence, $X \sim Binomial(5, 0.6)$

Now, the probability that the boxing game will be played is given by:

$$\begin{aligned} P(X \geq 3) &= \sum_{i=3}^{5} P(X = i) \\ &= \sum_{i=3}^{5} {}^{5}C_i \ 0.6^i(1-0.6)^{5-i} \\ &= 0.68 \end{aligned}$$

Therefore, the probability that the boxing game will be played is 0.68.

4. Let $X \sim Binomial(n, p)$. If the probabilities of x and $x+1$ successes are a and, b respectively, then find the parameter 'p' of the distribution.(Enter the answer correct to 2 decimal places)

   Answer: $\frac{b(x+1)}{a(n-x)+b(x+1)}$

Solution:

$$\begin{aligned}
\frac{a}{b} &= \frac{P(X=x)}{P(X=x+1)} \\
\frac{a}{b} &= \frac{{}^nC_x p^x (1-p)^{n-x}}{{}^nC_{x+1} p^{x+1} (1-p)^{n-x-1}} \\
\frac{a}{b} &= \frac{{}^nC_x (1-p)}{{}^nC_{x+1} p} \\
ap \times {}^nC_{x+1} &= b(1-p) \times {}^nC_x \\
\frac{ap}{x+1} &= \frac{b(1-p)}{n-x} \\
ap(n-x) &= b(1-p)(x+1) \\
p &= \frac{b(x+1)}{a(n-x)+b(x+1)}
\end{aligned}$$

Suppose we substitute the values of n, x, a and b as 5, 1, 0.4096 and 0.2048 respectively, then

$$\begin{aligned}
p &= \frac{0.2048(1+1)}{0.4096(5-1)+0.2048(1+1)} \\
p &= \frac{0.2048(2)}{0.4096(4)+0.2048(2)} \\
p &= \frac{0.4096}{0.4096(4)+0.4096} \\
p &= \frac{1}{5} = 0.2
\end{aligned} \tag{1}$$

5. If the expected number of sixes hit by a batsman on n balls is e and the variance for the same is v, then what is the probability of him hitting at least one six on any randomly selected n balls? (Enter the answer correct to 4 decimal places)

   Answer: $1-(1-p)^n$

   Solution:

   Let, X be a random variable representing the number of sixes hit by a batsman in an over.

   Hence, $X \sim Binomial(n,p)$

   $E(X) = np$ and $Var(X) = np(1-p)$

Now,

$$Var(X) = E(X) \times (1-p)$$
$$(1-p) = \frac{v}{e}$$
$$p = 1 - \frac{v}{e}$$

Also, $e = np$
$\implies n = \frac{e}{p}$
Therefore, $P(X \geq 1) = 1 - P(X = 0) = 1 - (1-p)^n$

Suppose we substitute the values of e and v as 4 and $\frac{4}{3}$ respectively, then

$$Var(X) = E(X) \times (1-p)$$
$$(1-p) = \frac{\frac{4}{3}}{4}$$
$$p = 1 - \frac{1}{3} = \frac{2}{3}$$

Also, $4 = \frac{2n}{3}$
$\implies n = 6$
Therefore,
$P(X \geq 1) = 1 - P(X = 0) = 1 - \left(\frac{1}{3}\right)^6 = 0.9986$

6. The probability of a student clearing a competitive exam is $\frac{1}{m}$. If he gives the exam n times, then what is the probability of him clearing the exam at least twice? (Enter the answer correct to 2 decimal places)

Answer: $1 - \left[{}^nC_0 \left(\frac{1}{m}\right)^0 \left(1 - \frac{1}{m}\right)^{n-0} + {}^n C_1 \left(\frac{1}{m}\right)^1 \left(1 - \frac{1}{m}\right)^{n-1}\right]$

Solution:
p = Probability of clearing the exam = $\frac{1}{m}$
1-p = $1 - \frac{1}{m}$
$p(x)$ = Probability of clearing the exam x times out of a total of n
$= {}^nC_x \left(\frac{1}{m}\right)^x \left(1 - \frac{1}{m}\right)^{n-x}$; $x = 0, 1, ..., n$
Thus, Probability of clearing the exam at least twice

$P(X \geq 2) = 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - [p(0) + p(1)]$

$$=1 - \left[{}^{n}C_0 \left(\frac{1}{m}\right)^0 \left(1 - \frac{1}{m}\right)^{n-0} + {}^{n}C_1 \left(\frac{1}{m}\right)^1 \left(1 - \frac{1}{m}\right)^{n-1} \right]$$

For example: m=4 and n=7

p = Probability of clearing the exam = $\frac{1}{4}$

1-p = $1 - \frac{1}{4} = \frac{3}{4}$

p(x) = Probability of clearing the exam x times in total of 7

$= {}^{7}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{7-x}$; $x = 0, 1, ..., 7$

Thus, Probability of clearing the exam at least twice

$P(X \geq 2) = 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - [p(0) + p(1)]$

$$=1 - \left[{}^{7}C_0 \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{7-0} + {}^{7}C_1 \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^{7-1} \right] = \frac{4547}{8192} = 0.56$$

7. Choose the correct condition/s about binomial distribution.

   (a) The probability of success p keeps varying for each trial.

   (b) The number of trials n is finite.

   (c) The trials are dependent on each other.

   (d) The trials are independent of each other.

   Answer: b, d

   Solution:

   As we know, if a random variable, X $B(n, p)$ then:

   - The total number of trials (n) is fixed.
   - Each of the trials is independent and identically distributed.

   Independent trial means that each of the trials is independent of the other trials. And, identical trials mean that probability of success is same for each of the trials.

8. Rithika wants to test whether the coin she has is a fair coin or not. To test this, she conducted an experiment of tossing the coin 5 times. Binomial random variable X is defined as the total number of heads(i) after 5 tosses. The probability distribution of the binomial random variable is given in Table 11.1.G.

$X = i$	$P(X = i)$
0	${}^5C_0p^0(1-p)^5$
1	${}^5C_1p^1(1-p)^4$
2	${}^5C_2p^2(1-p)^3$
3	${}^5C_3p^3(1-p)^2$
4	${}^5C_4p^4(1-p)^1$
5	${}^5C_5p^5(1-p)^0$

Table 11.1.G

What is the approximate probability of getting a head in tossing the given coin?(Enter the answer correct to one decimal place)

Answer: p

Solution:

Given the coin is tossed 5 times, $n = 5$

Given $P(X = 5) = 0.0009765625 \implies P(X = 5) = {}^5C_5 \times p^5 \times (1-p)^0 = 0.0009765625$

$\implies p^5 = 0.0009765625 \implies p = 0.25$

Thus, probability of head: $p = 0.25$ or $\dfrac{1}{4}$

9. At a school function, It is noticed that $a\%$ of the students are not wearing polished shoes and $b\%$ of students are not wearing school ties. It is announced that the students who have committed any of the infractions will be punished, and that these two infractions are independent of one another. If a teacher selects 5 students at random, then find the probability that exactly three of the students will be punished for any of the infractions?

(a) ${}^5C_3\left(\dfrac{100 \times (a+b) - (a \times b)}{10000}\right)^3\left(\dfrac{10000 - 100 \times (a+b) + (a \times b)}{10000}\right)^2$

(b) ${}^5C_3\left(\dfrac{a \times b}{10000}\right)^3\left(\dfrac{10000 - a \times b}{10000}\right)^2$

(c) ${}^5C_3\left(\dfrac{a+b}{100}\right)^3\left(\dfrac{100-(a+b)}{100}\right)^2$

(d) ${}^5C_3\left(\dfrac{a}{100}\right)^3\left(\dfrac{100-a}{100}\right)^2 + {}^5C_3\left(\dfrac{b}{100}\right)^3\left(\dfrac{100-b}{100}\right)^2$

Answer: a

Solution:

Let X be a random variable representing the number of students punished for any one of the infractions.

Define events as follows:

A: Student is not wearing polished shoes.

B: Student is not wearing a school tie.
Given, $P(A) = \frac{a}{100}$, $P(B) = \frac{b}{100}$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(A \cup B) = \frac{a}{100} + \frac{b}{100} - \left(\frac{a}{100} \times \frac{b}{100}\right) = \frac{a+b}{100} - \left(\frac{a}{100} \times \frac{b}{100}\right)$

$X \sim Binomial(5, p)$
Where, $p = \frac{a+b}{100} - \left(\frac{a}{100} \times \frac{b}{100}\right)$ and $1 - p = 1 - \left(\frac{a+b}{100}\right) + \left(\frac{a}{100} \times \frac{b}{100}\right)$
Now,

$$P(X = 3) = {}^5C_3(p)^3(1-p)^2 = {}^5C_3 \left[\frac{a+b}{100} - \left(\frac{a}{100} \times \frac{b}{100}\right)\right]^3 \left[1 - \left(\frac{a+b}{100}\right) + \left(\frac{a}{100} \times \frac{b}{100}\right)\right]^2$$

For example: a =5 and b=10

Let X be a random variable representing the number of students punished for any one of the infractions.
Define events as follows:
A: Student is not wearing polished shoes.
B: Student is not wearing a school tie.

Given, $P(A) = \frac{5}{100}$, $P(B) = \frac{10}{100}$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(A \cup B) = \frac{5}{100} + \frac{10}{100} - \left(\frac{5}{100} \times \frac{10}{100}\right) = \frac{5+10}{100} - \left(\frac{5}{100} \times \frac{10}{100}\right)$
$\implies 0.15 - 0.005 = 0.145$

$X \sim Binomial(5, 0.145)$
Where, $p = \frac{5+10}{100} - \left(\frac{5}{100} \times \frac{10}{100}\right) = 0.15 - 0.005 = 0.145$ and $1 - p = 1 - 0.145 = 0.855$
Now,
$P(X = 3) = {}^5C_3(0.145)^3(0.855)^2 = 0.0223$

10. There are x black and y blue pens in a box. A pen is chosen at random, and its colour is noted. If the process repeats independently, n times with replacement, then calculate the expected number of black pens chosen.

(a) $n \times \left(\frac{x}{y}\right)$

(b) $n \times \left(\frac{x}{x+y}\right)$

(c) $n \times \left(1 - \frac{x}{y}\right)$

(d) $n \times \left(\frac{y}{x+y}\right)$

Answer: b

Solution:

Let X be defined as number of black pens in n independent trials.

p is the probability of success.

$p = \frac{x}{x+y}$, $1 - p = 1 - \frac{x}{x+y} = \frac{y}{x+y}$

Given, process is repeats n times with replacement.

$E(X) = np = n \times \frac{x}{x+y}$

For example: x=10, y=20 and n=10

Let X be defined as number of black pens in 10 independent trials.

p is the probability of success.

$p = \frac{10}{10+20} = \frac{1}{3}$, $1 - p = 1 - \frac{1}{3} = \frac{2}{3}$

Given, process is repeats 10 times with replacement.

$n = 10$

$E(X) = 10 \times \frac{1}{3} = \frac{10}{3}$

11. Number of calls received at an office follows a Poisson distribution with an average of 1 call per minute. Find the probability that no call will be received in two minutes at the office. (Enter the answer correct to 2 decimal accuracy.)

Answer: 0.14 Range: 0.10 to 0.16

Solution:

Let random variable X represents the number of calls received at an office and it is given that $X \sim$ Poisson (λ).

According to the question, λ gives the average number of calls per minute. Therefore, there will be 2 calls on an average in two minutes. So, $\lambda = 2$.

Thus, Probability that no calls will be received in two minutes at the office is given as:

$$P(X = 0) = \frac{e^{-2} \times 2^0}{0!} = e^{-2} = 0.14$$

A quiz team is to be chosen randomly from 6 boys and 4 girls. The team has 3 slots which are to be filled randomly. If X denotes the number of boys in the quiz team, then using the given information, answer questions (12) and (13).

12. Calculate $P(X = 2)$. (Enter the answer correct to 1 decimal place)
Answer: 0.5, Range: 0.4 to 0.6
Solution:
Here, random variable X represents the number of boys in the quiz team. And, we have total 10 members out of which 6 are boys and 4 are girls. Also, we have to select a sample of 3 members randomly to filled the slots.
Clearly, X follows hypergeometric distribution, where,
N = total number of members = 10.
m = number of boys = 6.
n = number of randomly selected members = 3.
Now, pmf of hypergeometric distribution is given as:

$$P(X = i) = \frac{{}^{m}C_i \times {}^{N-m}C_{n-i}}{{}^{N}C_n}; i = 0, 1, 2, 3$$

Thus, required probability will be:

$$P(X = 2) = \frac{{}^{6}C_2 \times {}^{4}C_1}{{}^{10}C_3}$$

$$= \frac{15 \times 4}{120} = \frac{60}{120} = 0.5$$

13. Calculate the value of $E(X)$. (Enter the answer correct to 1 decimal place)
Answer: 1.8, Range: 1.5 to 2.1
Solution:
As we know that if $X \sim$ Hypergeometic distribution, then $E(X) = \frac{mn}{N}$.
Thus, $E(X) = \frac{6 \times 3}{10} = \frac{18}{10} = 1.8$

Statistics for Data Science - 1

Week 12 Graded Assignment

Continuous random variables

1. Let a random variable is uniformly distributed over $[a, b]$ with expectation and variance e and v respectively. Find the value of ab.
   Answer: $e^2 - 3v$
   Solution:
   Given random variable is uniformly distributed from $[a, b]$.
   $E[X] = \frac{b+a}{2} = e \Rightarrow (b+a)^2 = e^2 \times 4 = 4e^2$
   $Var(X) = \frac{(b-a)^2}{12} = v \Rightarrow (b-a)^2 = 12v$
   $(b+a)^2 - (b-a)^2 = 4ab = 4e^2 - 12v$
   $\Rightarrow ab = e^2 - 3v$
   For example: e=6 and v=4
   Given random variable is uniformly distributed from $[a, b]$.
   $E[X] = \frac{b+a}{2} = 6 \Rightarrow (b+a)^2 = 6^2 \times 4 = 144$
   $Var(X) = \frac{(b-a)^2}{12} = 4 \Rightarrow (b-a)^2 = 48$
   $(b+a)^2 - (b-a)^2 = 4ab = 144 - 48 = 96$
   $\Rightarrow ab = 24$

2. Probability density function of a random variable X is given by

$$f(x) = \begin{cases} kx^2 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   Find $P(m < X < n)$?
   (Hint: $\int_a^b cx^2 dx = c\frac{b^3 - a^3}{3}$ where c is a constant)
   (a). $n^2 - m^2$
   (b). $\frac{n^3 - m^3}{9}$
   (c). $\frac{n^2 - m^2}{9}$
   (d). $n^3 - m^3$

   Answer: d
   Solution:

We know that $\int_{-\infty}^{\infty} f(x)dx = 1$

$\Rightarrow \int_0^1 kx^2 dx = 1$

$\Rightarrow k\frac{x^3}{3}|_0^1 = 1$

$\Rightarrow k\frac{1}{3} = 1 \Rightarrow k = 3$

We know that $P(m \leq X \leq n) = \int_m^n f(x)dx$

$\Rightarrow P(m \leq X \leq n) = \int_m^n 3x^2 dx$

$\Rightarrow P(m \leq X \leq n) = 3\frac{x^3}{3}|_m^n$

$\Rightarrow P(m \leq X \leq n) = n^3 - m^3$

For example: m =0.1 and n=0.7

We know that $\int_{-\infty}^{\infty} f(x)dx = 1$

$\Rightarrow \int_0^1 kx^2 dx = 1$

$\Rightarrow k\frac{x^3}{3}|_0^1 = 1$

$\Rightarrow k\frac{1}{3} = 1 \Rightarrow k = 3$

We know that $P(a \leq X \leq b) = \int_a^b f(x)dx$

$\Rightarrow P(0.1 \leq X \leq 0.7) = \int_{0.1}^{0.7} 3x^2 dx$

$\Rightarrow P(0.1 \leq X \leq 0.7) = 3\frac{x^3}{3}|_{0.1}^{0.7}$

$\Rightarrow P(0.1 \leq X \leq 0.7) = 0.7^3 - 0.1^3 = 0.343 - 0.001 = 0.342$

3. The amount of milk produced every day by a dairy is uniformly distributed between 100 litres and 120 litres. What is the probability that the amount of milk produced is more than m litres given that on that day, more than n litres of milk was produced?(Enter the answer correct to 2 decimal place)

   Answer: $\frac{120-m}{120-n}$

   Solution:

   Given that milk produced is uniform between 100 and 120 litres.
   Given that on a given day milk produced is more than n litres.
   Let X denote the quantity of milk produced.

   $$P(X > m|X > n) = \frac{1 - P(X \leq m)}{1 - P(X \leq n)} = \frac{1 - (\frac{m-100}{120-100})}{1 - (\frac{n-100}{120-100})} = \frac{120-m}{120-n}$$

   For Example: m=115 and n=110

Given that milk produced is uniform between 100 and 120 litres.
Given that on a given day milk produced is more than 110 litres.
Let X denote the quantity of milk produced.
$P(X > 115|X > 110) = \frac{120-115}{120-110} = \frac{5}{10} = 0.5$

4. The time that Jan shatabdi express will reach the Delhi station is uniformly distributed between x p.m. and y p.m. What is the probability that the train reaches Delhi exactly at z p.m?
Answer: 0
Solution:
The probability that train reaches exactly at z PM is $\int_z^z f(x)dx = 0$
Since the area under the curve at a particular instant of x value is zero.
For example: x=12, y=5 and z=2
The probability that train reaches exactly at 2 PM is $\int_2^2 f(x)dx = 0$
Since the area under the curve at a particular instant of x value is zero.

5. The length of time for one person to be served at a restaurant is exponentially distributed, with an expected time of t minutes. If 5 persons arrived at the restaurant, then what is the probability that at least 2 of them will be served in more than p minutes? (Assume that each person is served independently)

(a) $1 - \left(1 - e^{-(\frac{p}{t})}\right)^5 - 5.e^{-(\frac{p}{t})}\left(1 - e^{(-\frac{p}{t})}\right)^4$
(b) $1 - \left(e^{-(\frac{p}{t})}\right)^5 - \left[5\left(e^{-(\frac{p}{t})}\right)^4(1 - e^{-(\frac{p}{t})})\right]$
(c) $1 - (1 - e^{-pt})^5 - [5.e^{-pt}(1 - e^{-pt})^4]$
(d) $1 - (e^{-pt})^5 - [5.e^{-4pt}(1 - e^{-pt})]$

Answer: a
Solution:
Given expected time $\dfrac{1}{\lambda} = t$ minutes.
$\Rightarrow \lambda = \frac{1}{t}$
$P(X > p) = e^{-p\lambda} = e^{\frac{-p}{t}}$
$\Rightarrow P(X \leq p) = 1 - e^{\frac{-p}{t}}$

Let i be the event that i among 5 of them are served in more than p minutes.

$P(i) = {}^5C_i(P(X > p))^i(P(X \leq p))^{5-i}$

$P(i) = {}^5C_i(e^{\frac{-p}{t}})^i(1 - e^{\frac{-p}{t}})^{5-i}$

$P(i \geq 2) = 1 - P(i = 1) - P(i = 0)$

$P(i \geq 2) = 1 - {}^5C_1(e^{\frac{-p}{t}})^1(1-e^{\frac{-p}{t}})^4 - {}^5C_0(e^{\frac{-p}{t}})^0(1-e^{\frac{-p}{t}})^5$

$P(i \geq 2) = 1 - 5(e^{\frac{-p}{t}})(1-e^{\frac{-p}{t}})^4 - (1-e^{\frac{-p}{t}})^5$

Hence, option (a) is correct.

For Example: t=5 and p=8

Given expected time $\frac{1}{\lambda} = 5$ minutes.
$\Rightarrow \lambda = \frac{1}{5}$
$P(X > 8) = e^{-8\lambda} = e^{\frac{-8}{5}}$
$\Rightarrow P(X \leq 8) = 1 - e^{\frac{-8}{5}}$

Let i be the event that i among 5 of them are served in more than 8 minutes.

$P(i) = {}^5C_i(P(X > 8))^i(P(X \leq 8))^{5-i}$

$P(i) = {}^5C_i(e^{\frac{-8}{5}})^i(1-e^{\frac{-8}{5}})^{5-i}$

$P(i \geq 2) = 1 - P(i=0) - P(i=1)$

$P(i \geq 2) = 1 - {}^5C_0(e^{\frac{-8}{5}})^0(1-e^{\frac{-8}{5}})^5 - {}^5C_1(e^{\frac{-8}{5}})^1(1-e^{\frac{-8}{5}})^4$

$P(i \geq 2) = 1 - (1-e^{\frac{-8}{5}})^5 - 5(e^{\frac{-8}{5}})(1-e^{\frac{-8}{5}})^4$

6. Probability density function of a random variable X is given by

$$f(x) = \begin{cases} 0.2x & \text{if } 0 \leq x \leq 1 \\ 0.2 & \text{if } 1 \leq x \leq 2 \\ 0.2x - 0.2 & \text{if } 2 \leq x \leq 3 \\ 0.4 & \text{if } 3 \leq x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

Find the probability that $0 < X < 2.5$. (Answer upto three decimal points)
Hint: Try to find the area under pdf using integration, otherwise draw the graph of pdf of X
Answer: 0.425 accepted range: [0.421, 0.429]
Solution:
$P(0 < X < 2.5) = \int_0^1 0.2x dx + \int_1^2 0.2 dx + \int_2^{2.5}(0.2x - 0.2)dx$

$P(0 < X < 2.5) = 0.1x^2|_0^1 + 0.2x|_1^2 + (0.1x^2|_2^{2.5} - 0.2x|_2^{2.5})$

$P(0 < X < 2.5) = 0.1 + (0.2 * 2 - 0.2) + (0.1 * 2.5^2 - 0.1 * 2^2 - 0.2 * 2.5 + 0.2 * 2)$
$P(0 < X < 2.5) = 0.425$

7. In a certain exhibition, the time for the next visitor to come is exponentially distributed with a standard deviation of t minutes. What will be the expected time (in minutes) for two visitors to arrive if one comes after the other? (Note that the arrival of a visitor is independent of the arrival of the previous visitor)

   (a). $\frac{1}{t}$

   (b). $2t$

   (c). t

   (d). $\frac{2}{t}$

   Answer: b

   Solution:

   Given standard deviation $= t$ minutes.

   $\Rightarrow$ Variance $= t^2$

   Variance $= \frac{1}{\lambda^2} = t^2$

   $1/\lambda = t \Rightarrow E[X] = 1/\lambda = t$

   Expected time for two visitors $= 2 * E[X] = 2t$ minutes, since both are independent of other visitor.

   For Example: t=10

   Given standard deviation $= 10$ minutes.

   $\Rightarrow$ Variance $= 10^2 = 100$

   Variance $= \frac{1}{\lambda^2} = 100$

   $1/\lambda = 10 \Rightarrow E[X] = 1/\lambda = 10$

   Expected time for two visitors $= 2 * E[X] = 20$ minutes, since both are independent of other visitor.

8. The total duration (in minutes) of a badminton match in the Premier Badminton League (PBL) is uniformly distributed between, $[a, b]$ with variance 12 square minutes. The probability that a match will last at most m minutes is $\frac{1}{p}$. Find the expected time duration (in minutes) of a badminton match.(Enter the answer correct to 1 decimal place)

   Solution:

   Given Variance $= 12$ square minutes.

   $\Rightarrow \frac{(b-a)^2}{12} = 12$

   $\Rightarrow (b-a)^2 = 12^2 \Rightarrow b - a = 12$

   Given $P(X \leq m) = \frac{1}{p}$

   $\Rightarrow P(X \leq m) = \frac{m-a}{b-a} = \frac{1}{p}$

$\Rightarrow m - a = \dfrac{b-a}{p}$

$a = m - \left(\dfrac{b-a}{p}\right) \Rightarrow b = m - \left(\dfrac{b-a}{p}\right) + 12$

$E[X] = \dfrac{b+a}{2}$

For example: m=42 and p=6

Given Variance = 12 square minutes.

$\Rightarrow \dfrac{(b-a)^2}{12} = 12$

$\Rightarrow (b-a)^2 = 12^2 \Rightarrow b - a = 12$

Given $P(X \leq 42) = \dfrac{1}{6}$

$\Rightarrow P(X \leq 42) = \dfrac{42-a}{b-a} = 1/6$

$\Rightarrow 42 - a = 12/6 = 2$

$a = 40 \Rightarrow b = 52$

$E[X] = \dfrac{b+a}{2} = \dfrac{40+52}{2} = 46.$

9. The probability density function of the time X (in minutes) between calls at the customer care is given by

$$f(x) = \begin{cases} \frac{1}{m}.e^{\frac{-x}{m}} & \text{if } 0 \leq x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Find the probability that time between calls exceeds the mean time.

(a) $1 - \dfrac{1}{e}$

(b) $1 - e$

(c) $\dfrac{1}{e}$

(d) e

Answer: c

Solution:

Mean time is $\dfrac{1}{\lambda} = \dfrac{1}{\frac{1}{m}} = m$

$P(X \geq t) = e^{-\lambda t}$

$P(X \geq \frac{1}{\lambda}) = e^{-\lambda * \frac{1}{\lambda}} = e^{-1} = \dfrac{1}{e}$

For example: m=4

Mean time is $\dfrac{1}{\lambda} = \dfrac{1}{\frac{1}{4}} = 4$

$P(X \geq t) = e^{-\lambda t}$

$P(X \geq \frac{1}{\lambda}) = e^{-\lambda * \frac{1}{\lambda}} = e^{-1} = \frac{1}{e}$

10. The lifetime of a light bulb is exponentially distributed with a mean life of n months. If there are $p\%$ chances that a light bulb will last at most t months, then what is the value of t?

(a). $n \times \ln\left(\frac{100-p}{100}\right)$

(b). $n \times \ln\left(\frac{100}{100-p}\right)$

(c). $\frac{1}{n} \ln\left(\frac{100-p}{100}\right)$

(d). $\frac{1}{n} \ln\left(\frac{100}{100-p}\right)$

Answer: b

Solution:

Given mean of exponential random variable (life of light bulb)=n months.

$\Rightarrow \frac{1}{\lambda} = n, \Rightarrow \lambda = \frac{1}{n}$.

Given $P(X \leq t) = \frac{p}{100}$

$\Rightarrow 1 - e^{-\lambda t} = \frac{p}{100}$

$\Rightarrow e^{-\lambda t} = 1 - \frac{p}{100} = \frac{100-p}{100}$

$\Rightarrow e^{\lambda t} = \frac{100}{100-p}$

$\lambda t = \ln\left(\frac{100}{100-p}\right)$

$t = \frac{1}{\lambda} \ln\left(\frac{100}{100-p}\right)$

$t = n \times \ln\left(\frac{100}{100-p}\right)$

Hence, option b is correct.

For example: n=18 and p=60

Given mean of exponential random variable (life of light bulb)=18 months.

$\Rightarrow \frac{1}{\lambda} = 18, \Rightarrow \lambda = \frac{1}{18}$.

Given $P(X \leq t) = 0.6$

$\Rightarrow 1 - e^{-\lambda t} = 0.6$

$$\Rightarrow e^{-\lambda t} = 0.4$$
$$\lambda t = \ln 2.5$$
$$\frac{1}{18}t = \ln 2.5$$
$$t = 18 \ln 2.5$$