Statistics for Data Science - 1

Practice assignment for Week 1

Syllabus covered:
1. Classification of statistics
2. Understanding the notion of sample and population
3. Classification of data
4. Understanding the notion of case/observation and variable
5. Classification of variables: Numerical and Categorical
6. Scales of measurement for variable: Nominal, ordinal, interval and ratio

1 Multiple choice questions

1. A quality engineer wants to check the quality of steel rods produced in a steel factory. For this, 40 pieces of steel rods are randomly selected from the steel factory to assess their quality. Based on this, choose the correct option below: [1]

    A. The population is all steel rods produced in all factories; the sample is the 40 steel rods selected from the given steel factory's production.

    B. The population is all steel rods produced in all factories; the sample is all the steel rods produced in the given steel factory.

    C. The population is all steel rods produced in the given steel factory; the sample is the 40 steel rods selected from the given steel factory's production.

    D. All the steel rods in the given steel factory are population; the sample is all steel rods in the given steel factory.

    Answer C

2. Values of temperature and humidity of a room are measured for 24 hours at a regular time interval of 30 minutes. Based on this, choose the correct option from below: [1]

    A. It is a cross-sectional data.

    B. It is time series data.

    C. None of the above.

    Answer B

3. **Pin code is a numerical variable.**

    A. True

B. False

Answer B

4. In the 2011 Cricket ODI World Cup quarter-final match between India and Australia, a media organization estimated that Australia would beat India by 50 runs if Australia bats first, based on the information of matches played between the two teams previously. Which branch of statistics does the above analysis belong to? [1]

   A. Descriptive Statistics

   B. Inferential Statistics

   Answer B

5. A class teacher wants to collect feedback from students of the class. The teacher hands out a blank sheet to each student to obtain descriptive input and suggestions on the class. The data collected by the class teacher is: [1]

   A. Structured Data

   B. Unstructured Data

   Answer B

6. Variables with an interval scale of measurement may be converted into a ratio scale of measurement by performing? [1]

   A. Addition operation

   B. Subtraction operation

   C. Multiplication operation

   D. Cannot be converted to ratio variables.

   Answer B

7. What is the scale of measurement for the amount of money you have? [1]

   A. Nominal

   B. Ordinal

   C. Ratio

   D. Interval

   Answer C

8. What is the scale of measurement for the military titles - Major, Captain, Colonel? [1]

   A. Nominal

   B. Ordinal

   C. Ratio

   D. Interval

   Answer B

2 Multiple Select Questions

9. Which of the following statements is (are) true: [2]
   A. All basic mathematical operations can be performed on some structured data.
   B. All basic mathematical operations can be performed on unstructured data.
   C. Email contents, text messages, and audio files are usually unstructured data.
   D. Height(cm), Weight(Kg), Age(years) are structured data.

   Answer A,C, D

10. Which of the following is(are) numerical variable(s)? [2]
    A. Height(cm)
    B. Day of the week
    C. Jersey number of sports player
    D. Mobile number
    E. Email address
    F. Age in years

    Answer A, F

11. Which of the following variables has (have) ratio scale of measurement? [2]
    A. Temperature in Kelvin
    B. Temperature in Centigrade
    C. Year
    D. Angle measured in degrees

    Answer A, D

12. Which of the following mathematical operation(s) can be performed on interval variables? [2]
    A. Addition
    B. Subtraction
    C. Multiplication
    D. Division

    Answer A, B

13. Which of the following is (are) expected while selecting a sample for a population? [2]
    A. Sample should be a subset of the population.
    B. Sample can contain data that is not from the population.

C. Sample should be representative of the characteristics of different elements in the population.

D. Sample need not be representative of the characteristics of different elements in the population.

Answer A, C

14. Which of the following operations can be valid for categorical variables? [2]

A. Addition

B. Subtraction

C. Comparison ($>$, $<$, $=$)

D. Multiplication

E. Division

Answer C

Statistics for Data Science - 1

Week 2 Practice Assignment

Graphical representation of categorical variables.

Syllabus covered:
1. Frequency table for categorical data.
2. Bar charts, Pareto charts, pie charts for representing categorical data.
3. Best practices while graphing.
4. Misleading graphs, area principle, and round off errors in pie charts.
5. Descriptive measures: Median and Mode

1 Multiple Choice Questions

1. Two biased dice A and B were tossed 20 and 10 times respectively, outcomes of which are given below.
   Die A: 1,2,3,4,4,5,5,6,2,3,3,3,5,3,5,6,5,5,1,1
   Die B: 1,2,3,4,5,6,3,4,1,3
   Which die has a higher relative frequency of getting 1 as outcome of die toss?

   [1 mark]

   (a) Die A

   (b) Die B

   Answer: b

2. A study conducted in the last 10 years shows how students choose different streams after completing 10th standard. The outcome of this study is given in Figure 2.1.P. What percentage of students have chosen Biology or Politics? [1 mark]

   (a) 10%

   (b) 20%

   (c) 25%

   (d) 30%

   Answer: b

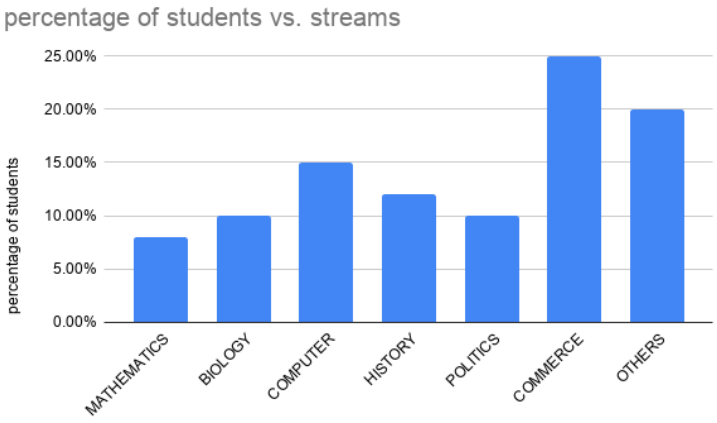


Figure 2.1.P: Streams

3. A poll was conducted by IMDb to rank 200 movies released in 2019 based on the revenue collected. Which graph is best suited to represent this data? [1 mark]

   (a) Pie chart

   (b) Bar chart

   (c) Pareto chart

   (d) None of the above

   Answer: c

4. The Union Budget of India is presented each year by the Finance Minister of India. Generally, subsidies are announced to boost the growth of key sectors such as food, petroleum, fertilisers, and others. What is the problem with the pie chart [Figure: 2.2.P] representing the share of subsidies of different sectors? [1 mark]

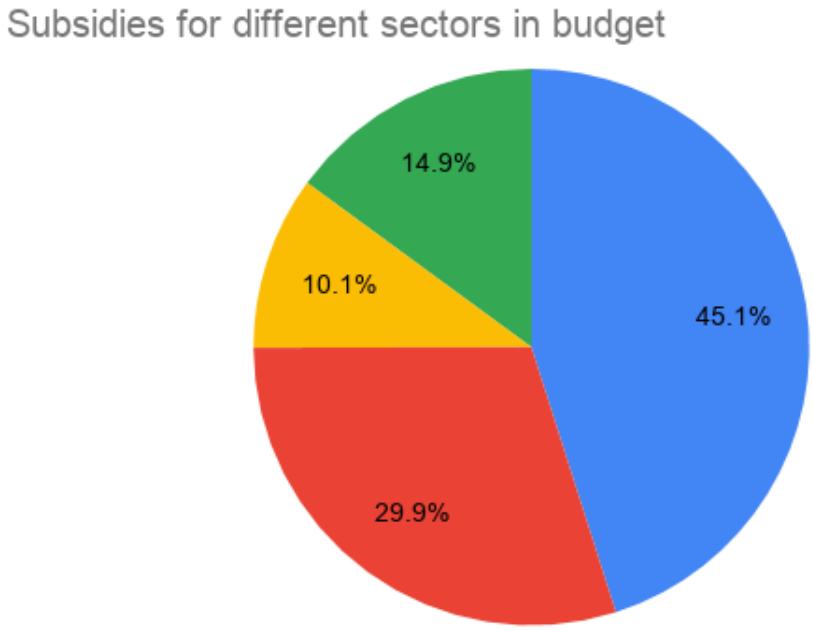


Figure 2.2.P: Subsidies for different sectors

(a) All slices have different colours.

(b) There is no legend containing the name and share of each group.

(c) Chart subtitle is missing.

(d) No slice has an offset from the center.

Answer: b

5. A garments retailer keeps record of profits on all his products. Profits of the last two years are shown in Figure 2.3.P. Choose the correct statement based on the data given in Figure 2.3.P.

[1 mark]

(a) Pie chart is misleading because it disobeys area principle.

(b) Pie chart is misleading because profit has been rounded off.

(c) Pie chart is not labelled well.

(d) Pie chart is not misleading.

Answer: b

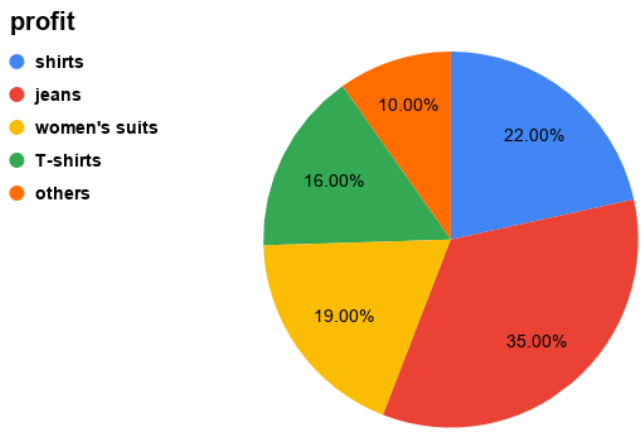


Figure 2.3.P: Profits of last two years

6. Which of the following measures of central tendency is applicable for a categorical variable with a nominal scale of measurement?

[1 mark]

(a) Mean

(b) Median

(c) Mode

(d) None of the above

Answer: c

Use the following information to answer the questions 7, 8, and 9.
Rajdhani trains connect the national capital of India with state capitals of India and/or the biggest cities of the states. The Rajdhani express (12437) running from Secunderabad Jn to H Nizamuddin Jn has one AC first-class coach (1A), five AC second-class (2A) coaches, and nine AC third-class (3A) coaches. Each of the 1A, 2A, and 3A coaches has a maximum capacity of 18 berths, 48 berths, and 64 berths respectively. On a given day, 10 1A berths, half of the 2A berths, and 275 3A berths were filled.

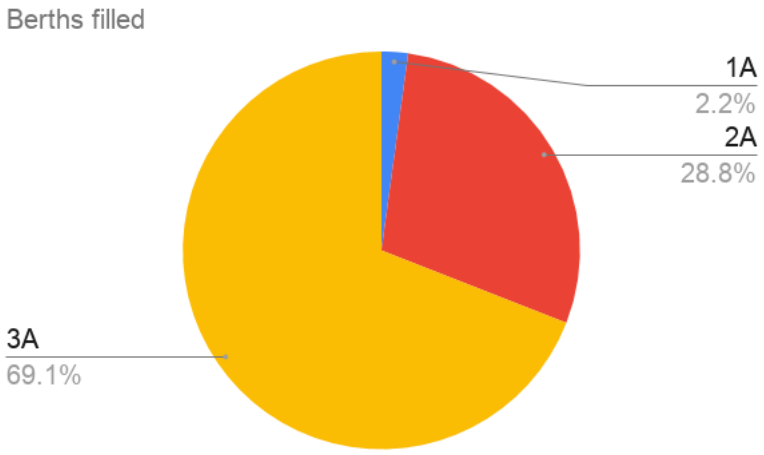
7. What is the percentage of berths filled in the train on the given day? [1 mark]
Answer: 48.56 Accepted range [48.5, 48.6]

8. What is the relative frequency of 3A berths among total berths unfilled in the train on the given day? [1 mark]
Answer: 0.7016 accepted range [0.701, 0.702]

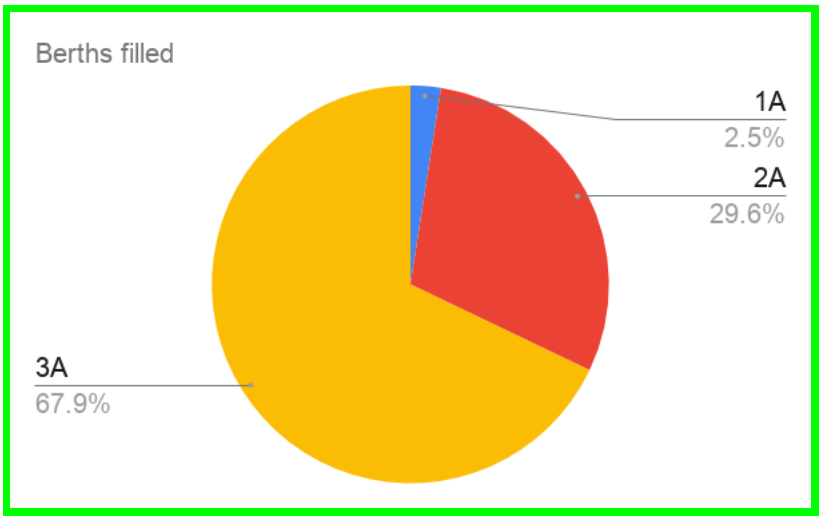
9. What is the pie chart representation of the berths filled in the train on the given day?

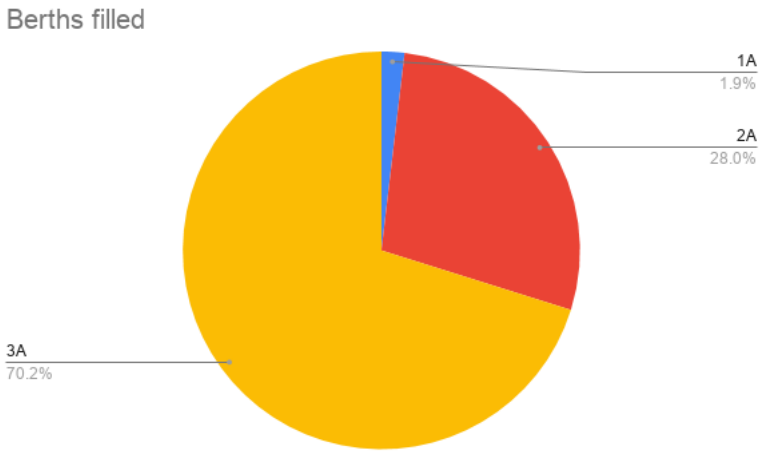
[1 mark]

(a)



(b)





(c)

Answer: b

10. The sector-wise division of income of a country for the year 2012 is shown in Figure 2.4.P. Figure 2.5.P represents the combined income of years 2012 and 2013 put together. In order to obey the area principle, the radius of the pie chart in Figure 2.5.P is made twice the radius of the pie chart in Figure 2.4.P. Based on this information, which of the following could be true? [5 marks]

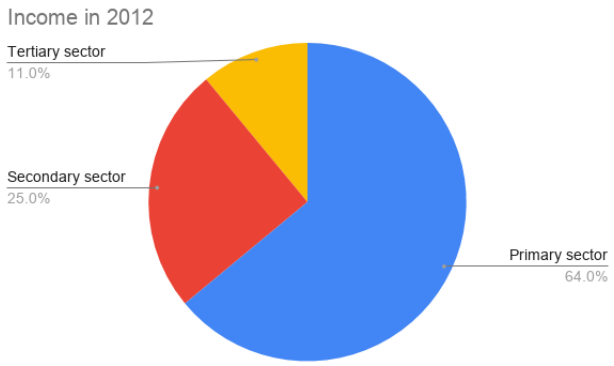


Figure 2.4.P: Income in 2012

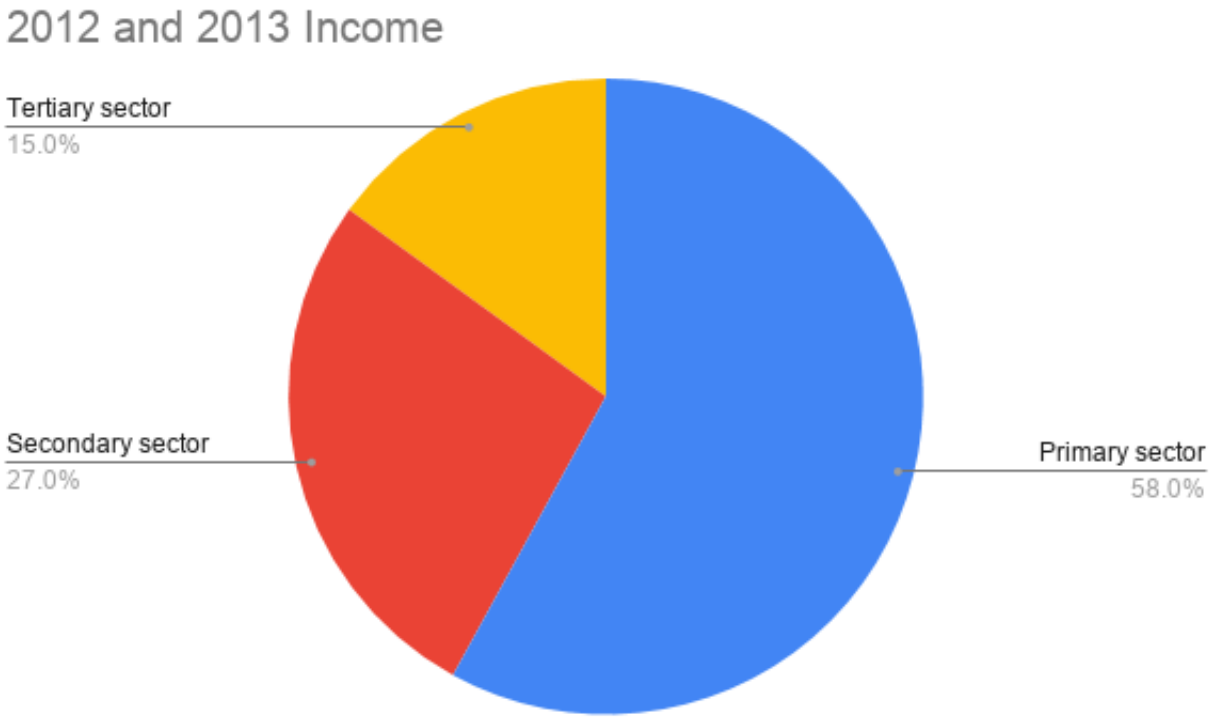


Figure 2.5.P: Cumulative incomes of 2012 and 2013

(a) The share of each sectors is same in year 2012 and 2013.

(b) The net income increased by threefold from the year 2012 to 2013.

(c) The share of tertiary sector in the 2013 income is 16.33% approximately.

(d) The primary sector income increased by 173.5%.

(e) The primary sector income increased by 162.5%.

(f) The income of secondary sector is exactly twice that of tertiary sector in the year 2013.

Answer: b, c, e

Use the following information to answer the questions 11, 12, 13, and 14. A contract for sports jerseys of 45 players in a team is being given to a vendor. All the players, except two, gave their preferred size from among the categories XL, L, M, and S. Two players A and B do not know which size will fit them. Assume the data is unimodal.

11. If player A knows that most of the players have a similar physique to him, which jersey can he order? [1 mark]

    (a) Size M jersey if it is the mode of the size of the jerseys.

    (b) Size S jersey if it is the mode of the size of the jerseys.

    (c) Size M jersey if is the median of the size of jerseys.

    (d) Size S jersey if it is the median of the size of jerseys.

    Answer: a,b

12. The 43 chosen jersey sizes are arranged in the increasing order of sizes. If player B knows that his jersey size is approximately equal to the jersey size of the player with the 22nd ranked jersey size in this arrangement, which jersey size should player B order? [1 mark]

    (a) Size M jersey if it is the mode of the size of the jerseys.

    (b) Size S jersey if it is the mode of the size of the jerseys.

    (c) Size M jersey if it is the median of the size of jerseys.

    (d) Size S jersey if it is the median of the size of jerseys.

    Answer: c, d

13. Players A and B have identified their preferred jersey sizes, based on the information on data from questions 11 and 12 and their preferences have been added to the existing dataset. What can be about the mode of the new dataset? [2 marks]

    (a) The mode does not change.

    (b) The mode is not necessarily equal to the jersey size of player A.

    (c) The mode is not necessarily equal to the jersey size of player B.

    (d) It cannot be determined whether the mode has changed or not.

    Answer: a, c

14. Players A and B have identified their preferred jersey sizes, based on the information on data from questions 11 and 12 and their preferences have been added to the existing dataset. What can be said about the median of the new dataset? [2 marks]

    (a) The median does not change.

(b) The median is not necessarily equal to the jersey size of player A.

(c) The median is not necessarily equal to the jersey size of player B.

(d) It cannot be determined whether the median has changed or not.

Answer: a, b

Statistics for Data Science - 1

Week 3 Practice Assignment

Describing numerical variable

1. The stem and leaf plot shown in Figure 3.1.P represents the temperatures in centigrade. What is the stem and leaf representation of temperatures 65^0C, 68^0C, and 67^0C? [Easy]

Stem	Leaf
5	1 2 4 5 7
6	4 7 8 9
7	1 3 5 6 7 8
8	4 5 6 8 9

Here 5 | 1 represents 51^0C

Figure 3.1.P: Temperature stem and leaf plot

(a) 6 | 5 8 7

(b) 6 | 5 7 8

(c) 5 | 6
7 | 6
8 | 6

(d) 6 | 8 7 5

2. If the sample variance of a data of size 10 is 23, then what is the population variance of this data? [Easy]

(a) 19

(b) 20.7

(c) 22

(d) 23

3. For a particular data, the value for the 10th percentile is 33.5, 25th percentile is 45, 50th percentile is 84.5, and 100th percentile is 102. What is the median of this data?

[Easy]

(a) 33.5

(b) 45

(c) 84.5

(d) 102

4. The marks scored by a group of students in an exam is given below.
110 20 50 60 45 30 42 21 15 62 26 33 17 32 27
What is the median of the marks scored? [Easy]

(a) 39.33

(b) 32

(c) 32.5

(d) No median is available for this data.

5. If each value of a numerical discrete variable is squared, then the mean of the new data is: [Easy]

(a) Equal to the square of the old mean.

(b) Less than the square of the old mean.

(c) Twice the old mean.

(d) The mean does not change.

(e) Can not be determined from the given data.

6. Which of the following is not a measure of dispersion? [Easy]

(a) Variance

(b) Standard Deviation

(c) Mean

(d) Range

7. If the mean and sample standard deviation of the data 1, 6, 10, 14, 4, x_1, x_2 are 7 and $\sqrt{106/6}$ respectively, then $|x_1 - x_2|$ is? [Medium]

   Answer: 2

8. If the mean of the observations $x_1, x_2, x_3, \cdots, x_n$ is 15, the mean of the observations $x_1, x_2, x_3, \cdots, x_{2n}$ is 45, and the mean of the observations $x_1, x_2, x_3, \cdots, x_{3n}$ is 34, then what is the mean of the observations $2x_{n+1}, 2x_{n+2}, 2x_{n+3}, \cdots, 2x_{2n}, 3x_{2n+1}, 3x_{2n+2}, 3x_{2n+3}, \cdots, 3x_{3n}$?

   [Medium]

   Answer: 93

9. If the mean and standard deviation of observations $x_1, x_2, x_3, \cdots, x_n$ are 20 and 5 respectively, and the mean and standard deviation of observations $a(3x_1 + b), a(3x_2 + b), a(3x_3 + b), \cdots, a(3x_n + b)$ are 94.5 and 22.5 respectively, then what is the sum of a and b? [Medium]

   Answer: 4.5

10. The five-number summary of 99 observations of a numerical variable is 25, 35, 47, 56, 78. Based on this information, which of the following statements could be true? [Hard]

    (a) The mean of the data is 41.1333

    (b) The range of the data is 53.

    (c) The interquartile Range is 21.

    (d) The mean of the data can be 35.

    (e) The mean of the data can be 56.

    (f) The mean of the data could be any value in the range of [35, 56].

    (g) The mean of the data could be any value in the range of [41.133, 53.65].

    (h) 80th percentile value is 78.

    Answer: a, b, c, g, h

Statistics for Data Science - 1

Week 4 Practice Assignment

Association between two variables

Syllabus covered:

- Use of two-way contingency tables to understand association between two categorical variables.
- Understand association between numerical variables through scatter plot; compute and interpret correlation.
- Understand relationship between a categorical and numerical variable.

1 Numerical answer type

1. What is the correlation coefficient between temperature measured in Celsius and temperature measured in Fahrenheit?
   Answer: 1

2. Calculate the correlation coefficient between marks obtained by a set of ten students in their class 10th(X) and class 12th(Y) exams as given in Table 4.1.P.

X	355	487	526	590	428	398	555	320	450	510
Y	300	340	400	450	300	325	450	400	375	400

Table 4.1.P: Marks obtained in class 10th and class 12th

   Answer: Accepted range 0.65 to 0.66

3. Table 4.2.P shows the scores for ten students in Statistics and Mathematics exam papers.

Statistics	55	87	56	90	84	98	55	45	51	75
Mathematics	85	65	80	60	70	63	76	68	75	60

Table 4.2.P: Marks obtained in Statistics and Mathematics

   a) Find the average score for the Statistics exam paper.
   Answer: 69.6
   b) Find the range for the Mathematics exam paper.
   Answer: 25
   c) Select which of the following describe the correlation between both exam papers.

a) weak

b) strong

c) no correlation

d) positive

e) negative

2 Multiple Choice Questions (MCQ)

1. What can be said about the correlation coefficient r of x and y where $y = x^2 + 8x + 16$, x takes the values of the first ten positive integers?

   a) $r = 1$

   b) $0 < r < 1$

   c) $-1 < r < 0$

   d) $r = -1$

2. *Simpson's Paradox* is defined as the reversal of conclusions in disaggregated and aggregated cross-tabulation.
   Table 4.3.P shows the disaggregated performances (Home matches and Away matches), and aggregated performances (both Home and Away matches) of two Indian captains A and B in ODI cricket.
   Aggregated data is combination of groups of data into single summary statistics. In the given cricket data Home and Away data is aggregated data, since the place where matches are played is combined into single variable.
   Disaggregated data is the data that is divided into detailed sub categories. In the given data, Home and Away matches are disaggregated data since the total matches played is divided based on the location (Home or Away) it is played.

	Total Matches	Home			Away			Home and Away		
		Won	Draw	Lost	Won	Draw	Lost	Won	Draw	Lost
Captain A	152	9	0	1	76	1	65	85	1	66
Captain B	260	150	0	50	25	0	35	175	0	85

Table 4.3.P: Disaggregated performances and aggregated performances

   (a) Based on the data given in Table 4.3.P, which captain has the better record in ODI captaincy in terms of win percentage?

      a) A

b) B
c) Both the captains have a similar record.

(b) Which captain has the highest loss percentage in ODI cricket among the two?

a) A
b) B
c) Both the captains have a similar record.

(c) Which captain has the best win percentage in Home ODI matches?

a) A
b) B
c) Both have similar record

(d) Which captain has the best win percentage in Away ODI matches?

a) A
b) B
c) Both have similar record

(e) Does the above data illustrate Simpson's Paradox?

a) No, because the conclusions in aggregated and disaggregated cross-tabulation are the same.
b) Yes, because the reversal of conclusions occurs in draw percentages of both the captains.
c) Yes, because the reversal of conclusions occurs in win percentages of both the captains.
d) Yes, because the reversal of conclusions occurs in loss percentages of both the captains.

3. Ajay has a large number of relatives who have different incomes and expenditures. In order to find out if there is any relation between the income and the expenditure of some of his relatives, he collects the data in Table 4.4.P.

(a) What is the sample mean (in INR lakhs) and sample standard deviation (in INR lakhs) of income of relatives of Ajay?

a) 9.2, 3.67
b) 9.7, 3.63
c) 9.7, 3.67
d) 9.2, 3.63

(b) What is the sample covariance(in INR lakh^2) of income, expenditure from the above data?

a) 6.28

Relative	Income in INR Lakhs	Expenditure in INR Lakhs
Relative 1	6	4.4
Relative 2	5	4.3
Relative 3	11	7.4
Relative 4	12	8.1
Relative 5	7	5
Relative 6	13	8.4
Relative 7	14	9

Table 4.4.P: Income and expenditure dataset

b) 7.33
c) 8.2
d) 5.6

(c) What is the unit of correlation coefficient for the given data?

a) Rupee^2
b) Lakh^2
c) Rupee
d) No units
e) Lakh

(d) What is the value of the correlation coefficient between income and expenditure?

a) 0.91
b) 0.78
c) 0.99
d) 0.85

4. Let r be the sample correlation coefficient for the data pairs $(x_i,\ y_i),\ i = 1, ..., n$. Then, the sample correlation coefficient for the data pairs $(a+bx_i,\ c+dy_i),\ i = 1, ..., n$, provided that b and d have the same sign, is

a) $(ac + bd)r$
b) $-r$
c) r
d) 1

5. Let the correlation coefficient r between two variables X and Y be zero. Then, the variables X and Y are:

a) linearly related.

b) not linearly related.

c) same.

6. Consider the set of points (2, 6), (3, 8), (4, 10), (5, 14), (10, n) in the XY - plane. What should the value of n be so that the correlation between the X-values and Y-values is 1?

   a) 23

   b) 26

   c) 29

   d) A value different from any of the above.

   e) No value for n can make $r = 1$.

3 Multiple Select Questions (MSQ)

1. If the correlation coefficient r between two variables X and Y is negative, then the relation between X and Y can be described as:

   a) when Y increases X does not change

   b) when Y increases X increases

   c) when Y increases X decreases

   d) when Y decreases X decreases

   e) when Y decreases X increases

2. Table 4.5.P represents the IQ score of eight fathers and their daughters.

Father's IQ	Daughter's IQ
130	108
100	117
124	112
89	115
132	100
99	102
100	75
129	135

Table 4.5.P: IQ dataset

Choose which of the following describe the correlation between the IQs of fathers and daughters.

a) strong

b) weak

c) no correlation

d) positive

e) negative

Statistics for Data Science - 1

Week 5 Practice Assignment

Counting Principles

1. There are 42 ways to select a captain and a vice captain from a team of n members. What is the value of n? [Easy]

   Answer: 7

2. If ${}^nC_r = 220$ and ${}^nP_r = 1320$, then what is the value of r? [Easy]

   Answer: 3

3. There are 30 students in a class who are supposed to be seated in 6 rows containing 6 seats each. In how many ways can they occupy seats if two students A and B are asked to sit in the same row? [Easy]

   (a) $6 \times 6 \times 5 \times {}^{34}P_{28}$
   (b) $6 \times 5 \times {}^{34}P_{28}$
   (c) $6 \times 6 \times 5 \times {}^{34}C_{28}$
   (d) $6 \times 5 \times {}^{34}C_{28}$

   Answer: a

4. How many five digit numbers (without repetitions) are there containing exactly two even digits? [Medium]

   (a) 12000
   (b) 11040
   (c) 16000
   (d) 13560

   Answer: b

5. Shweta wants to go from Jammu to Bangalore by taking halts at Delhi and Nagpur during the journey. She has options of 2 flights and 3 trains from Jammu to Delhi, 2 trains and 3 buses from Delhi to Nagpur, 2 flights, 2 trains, and 4 buses from Nagpur to Bangalore. In how many ways can she travel if she takes exactly one train during the whole journey? [Easy]

   Answer: 90

6. How many positive integers of at most 3 digits are there such that the product of their digits is 30? [Medium]

   Answer 14

7. A man has forgotten the password of his locker but he remembers that the password is an even number between 800 and 2250. He also knows that the password does not contain the digits 3, 5, and 7. What is the maximum number of attempts he will require to open his locker? [Medium]

   Answer: 405

8. A pizza shop has four types of pizzas and each of them comes in three different sizes: large, medium, and small. A customer is also given the option of adding extra cheese or extra paneer or extra vegetables where the customer can choose to add any one or any two or all or none of them. In how many ways can a pizza be ordered? [Easy]

   Answer: 96

9. A family of three members met five people at a function and they were all seated around a circular table. Choose the correct options [Medium]

   (a) They can occupy their seats in 120 ways if all three family members decided to sit together.

   (b) They can occupy their seats in 720 ways if all three family members decided to sit together.

   (c) They can occupy their seats in 4320 ways if not all three members sit together.

   (d) They can occupy their seats in 4920 ways if not all three members sit together.

   Answer: b, c

10. Eight girls were playing a game during which they decided to sit around in two concentric circles as shown in Figure 5.1.P. Four out of them will sit on the inner circle and the remaining 4 will sit on the outer circle. In how many ways can they occupy their places if Priya, who is the game master, has to sit in the outer circle? [Medium]

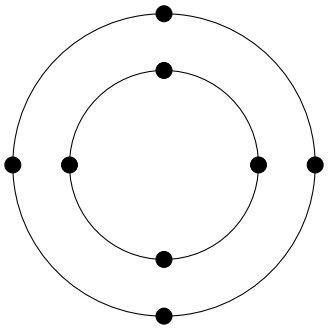


Figure 5.1.P: Sitting plan of girls

(a) ${}^7C_4 \times (3!) \times (5!)$

(b) ${}^7C_4 \times (3!) \times (4!)$

(c) ${}^7C_4 \times (3!) \times (3!)$

(d) ${}^8C_4 \times (3!) \times (4!)$

Answer: b

11. David wants to create an account on a learning application in which he has to provide a username and a password. The username must be such that the first 5 characters are alphabets, followed 2 or 3 digits. The password must contain 7 characters out of which the first 2 are alphabets followed by a 5-digit odd number.The last digit of password must be larger than second last digit and no digit should be zero. If he cannot use alphabets in the password which have already been used in the username, then which of the following is(are) true? [Hard]

(a) The number of ways David can generate his username is $26^5 \times (1100)$.

(b) The number of ways David can generate his username is $26^5 \times (1000)$.

(c) If he chooses his username as "David343", then he can generate his password in 1852200 ways.

(d) If he chooses his username as "David343", then he can generate his password in 7056720 ways.

(e) If David uses "David343" as his username and wants to use distinct digits (no two digits are same) in his password, then he can generate his password in 2032800 ways.

Answer: a,d,e

12. How many words (without repetition of letters) of at most 5 letters can be formed from the letters of the word 'CONQUER' such that the word always contains the string 'QR'. [Medium]

    Answer: 311

Statistics for Data Science - 1

Week 6

Practice Assignment

1 Multiple Choice Questions

1. Two fair dice are rolled. The sum of the outcomes of the tossing of the two dice are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.
   Let E be the event that the sum is less than or equal to 12.
   What is the complement of event E? [Easy]

   a. $E = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$
   b. $E = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
   c. Null event
   d. Cannot be determined.

   Based on the following information, answer question numbers 2 and 3.
   Raghav is selected for a school cricket club. He has to select a jersey number of his choice between 0 to 99, both inclusive.

2. What is the probability that he will select either an even number jersey or a jersey number divisible by 3? [Easy]
   Answer:67/100

3. What is the probability that he will choose a jersey number neither divisible by 3 nor by 5? [Easy]
   Answer:27/50

4. $E1$, $E2$, $E3$, $E4$ are disjoint events. The probabilities of $E2$ and $E4$ are 0.3 and 0.22 respectively. The probability of union of these four events is 1. Find $E1 \cup E3$. [Easy]
Answer: 0.48

5. If the intersection of two events is null set, then the two events are [Easy]

   a. Independent events
   b. Mutually exclusive
   c. Disjoint events
   d. Cannot say

6. 150 employees are working in an NPTEL office. 80 of them drink tea, 90 drink coffee. Some of them may drink both, but every person drinks at least one of these two beverages. What is the probability that a person chosen at random from this population drinks both tea and coffee (correct up to 4 decimal points)? [Easy]
Accepted Range: 0.13-0.14

2 Numerical Answer Type

Based on the following information, answer the questions (7), (8), (9).
90 students have participated in the annual cultural festival of a college comprising of singing, dancing and acting competitions. The number of students who did not participate in the acting or dancing competitions are 10. The number of students who participated only in dancing competition are 15. The total number of students who participated in the acting competition are 50. The number of students who participated in both the singing and the acting competition are 20. The number of students who participated in both the dancing and the acting are 15. The students who participated in all the three competitions are 10. [Medium]

7. What is the probability that a student selected at random participated only in the dancing competition?(Correct up to two decimal points)
Accepted Range: 0.16-0.17

8. What is the probability that a student selected at random participated in both the singing and the dancing competitions? (Correct up to two decimal points)
Accepted Range: 0.27-0.28

9. What is the probability that a student selected at random has participated in both the singing and the acting competitions but not participated in the dancing competition? (Correct up to two decimal points)
Accepted Range: 0.11-0.12

10. For the qualifier exam of the Statistics for Data Science course, the student has to solve five multiple choice questions. Each question has four options, a, b, c, and d. The student has to attempt all the questions.
What is the probability that the student will select either option a or b for all questions? [Medium]
Answer: 0.03125

Based on the following information, answer the questions (11) and (12).
According to a survey, 70 percent of all participants who participated in a marathon completed the race. Two participants are randomly selected from the participants of the race (order does not matter). [Medium]

11. What is the probability that neither of the participants completed the marathon?
Answer: 0.09

12. What is the probability that one of the participants completed the marathon and the other one did not?
Answer: 0.21

13. $P(C) = 0.3$, $P(A \cup B) = 0.6$, and $P(A \cup B \cup C) = 0.8$, then what is the value of $P((A \cup B)^c \cap C)$? [Hard]
Answer: 0.2

Statistics for Data Science - 1

Week 7 Practice Assignment

Conditional probability and Bayes' theorem

1. Let A and B be the two events of a random experiment. Probability that at least one of the two events A and B will occur is 0.6. Probabilities of event A occurring and B occurring are 0.3 and 0.4 respectively. Find the probability that event A will occur given that event B has occurred. [Easy]

   (a) $\frac{1}{2}$
   (b) 1
   (c) $\frac{2}{5}$
   (d) $\frac{1}{4}$

2. A random sample of 500 people are classified by gender and their level of education. The data is given in Table 7.P.1 .

Level of education	Male	Female
Elementary	25	30
Secondary	70	105
College	160	110

Table 7.P.1: Education level

   If a person is picked at random from this group, find the probability that the person is a male, given that the person has a secondary education. [Easy]
   Answer: 0.4

3. Among the customers who buy Samsung mobile phones at a particular shop, 40% also buy back cover for that phone. Suppose that 60% of the customers who visit the shop buy Samsung mobile phones. What is the probability that a customer at that shop purchased both Samsung mobile phone and a back cover for it? (Write your answer upto two decimal places.) [Easy]
   Answer: 0.24

4. Jasmine has two coins, out of which one is fair and the other results in a tail with 0.6 probability. She picks a coin randomly (probability of picking a coin is 0.5) and tosses it and it shows head. What is the probability that she picks the fair coin? [Medium]

1. $\frac{4}{9}$
2. $\frac{5}{6}$
3. $\frac{5}{9}$
4. $\frac{1}{3}$

Answer: $\frac{5}{9}$

5. Let A and B be two independent events of a random experiment. Then, which of the following is/are always true? [Easy]

   (a) $P(A \cup B) = P(A)P(B^C) + P(B)$

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

   (c) $P((A \cap B)|A) = P(A \cap B)$

   (d) A^C and B^C are independent.

6. A target is shot by Anu and Abhishek at the same time. Anu hits the target with probability 0.5 and Abhishek hits the target with probability 0.8. It is known that the target is hit by at least one of them, then what is the probability that it is hit by Abhishek? [Medium]

   (a) $\frac{5}{9}$

   (b) $\frac{4}{9}$

   (c) $\frac{9}{13}$

   (d) $\frac{8}{9}$

7. Bag A contains 2 red and 2 yellow balls and bag B contains 1 red and 2 yellow balls. A bag is selected randomly and a ball is drawn from the selected bag. Probability of selecting bag A is $\frac{1}{4}$. What is the probability that the ball drawn is yellow? (Write your answer upto three decimal places) [Easy]
   Answer: 0.625

8. There is a 40% chance that Ajay will go to school today. If Ajay does not go to the school then there is a 15% chance that Prachi, his friend, will go to school. What is the probability that at least one of them will go to school? [Medium]
   Answer: 0.49

9. In a particular court, it is found that 90% of the guilty suspects are properly judged while 10% of the guilty suspects are incorrectly found innocent. On the other hand, innocent suspects are misjudged 1% of the time. If the suspect was selected from a group of suspects of which 35% have ever committed a crime, and the judge declared that he is guilty, what is the probability that he is innocent? (Write the answer upto two decimal places.) [Medium]
   Answer: 0.02

10. In a manufacturing company, three machines, M_1, M_2, and M_3, make 30%, 45%, and 25%, respectively, of the products. It is known from past experience that 2%, 3%, and 2% of the products made by each machine, respectively, are defective. Which of the following options is (are) correct? (Write your answer upto two decimal places.) [Hard]

(a) There is 0.0245 probability that a randomly selected product is defective.

(b) There is 0.0456 probability that a randomly selected product is defective.

(c) If a product was chosen randomly and found to be defective, the probability that it was made by machine M_3 is $\frac{10}{49}$.

(d) If a product was chosen randomly and found to be defective, the probability that it was made by machine M_1 is the highest.

(e) If a product was chosen randomly and found to be defective, the probability that it was made by machine M_2 is the highest.

(f) If a product was chosen randomly and found to be defective, the probability that it was made by machine M_3 is the highest.

Statistics for Data Science - 1

Week 8 Practice Assignment

Random Variable

1 Multiple Choice Questions

1. 5 random alphabets were selected from the 26 English alphabets. Let the random variable be number of vowels in the selected 5 random alphabets. If the probability that the random variable will take value 4 is $x \times 10^{-3}$, then what is the value of x?. Enter the answer up to 2 decimals accuracy. [Easy 1 mark]
   Answer: 1.59622, accepted range 1.45 to 1.70

2. A guns manufacturing factory in Uttar Pradesh produced 100 guns out of which 5 are defective. Now a dealer named Munna wants to buy the 10 guns from the factory. Let a random variable be the number of guns Munna brought are defective. What value of random variable has the highest probability? [Medium 3 marks]
   Answer: 0

3. Probability distribution of a random variable X is given below.
   $P(X = x) = \frac{1}{2^x}, x \in N \ \ \& \ \ x \geq 1$
   $P(X = x) = 0$ for all other values of x
   Does the above probability distribution satisfy all the properties of pmf (Probability mass function)? [Easy 1 mark]
   Hint: Use the following formula if required

$$a + ar + ar^2 + ar^3 + ar^4 + \cdots = \frac{a}{1-r} \text{where } r < 1$$

   (a) True
   (b) False

4. In an exam there are 10 multiple choice questions, each question contains 4 options out of which one option is correct. Every student needs to attempt all the questions. Let the random variable be the number of questions that are correct. What is the probability that the random variable is equal to 7? [Easy 1 mark]

   (a) $\frac{^{10}C_7 \times 3^3}{4^{10}}$
   (b) $\frac{^{10}C_3 \times 3^3}{4^{10}}$
   (c) $\frac{^{10}C_7 \times 7^3}{4^{10}}$

(d) $\frac{^{10}C_3 \times 7^3}{4^{10}}$

5. In an experiment of tossing a fair coin, random variable X is defined as the number of tosses required to get the first head. Probability mass function ($P(X = i)$) is defined as the probability of getting first head in i_{th} number of tosses. For example, the probability that first head happens in the first toss ($P(X = 1)$) is $\frac{1}{2}$. What is the value of $P(X = i)$? [Easy 1 mark]

   (a) $\frac{1}{2^i}$
   (b) $\frac{2}{4^i}$
   (c) $\frac{4}{8^i}$
   (d) $\frac{1}{2^{i-1}}$

   Use the following information to answer question 6 and 7. Two friends Ramesh and Suresh wants to play book cricket game. The rules of the game are:

   - Your initial score is zero. You have maximum of 5 chances.
   - You need to randomly open a book page. The last digit of the right side page number will be added to your current score.
   - If you get zero, you are out and you won't have any more chances to play again even though your chances are not done.
   - The person who gets the highest score is the winner.

   The book Ramesh and Suresh are using to play book cricket contains 300 pages, numbered 1 to 300, with the right side pages containing even numbered pages. If the probability mass function, $P_R(X = x_i)$ and $P_S(X = x_i)$ be the probability that Ramesh and Suresh scores respectively x_i runs.

6. What is the probability that Ramesh will be not out at the end of his chances? [Easy 1 mark]

   (a) 0
   (b) $\frac{9^5}{10^5}$
   (c) $\frac{4^5}{1365}$
   (d) $\frac{4^5}{5^5}$
   (e) $\frac{4^5}{10^5}$
   (f) $\frac{1}{1365}$

7. If Ramesh has scored 6 runs in his chances what is the probability that Suresh will score more than 6 runs and win the game? [Hard 5 marks]

    1. $\frac{8}{5^5}$
    2. $\frac{1357}{1365}$
    3. $\frac{5^5-8}{5^5}$
    4. $\frac{8}{5^5}$
    5. $\frac{8}{1365}$

8. In a game of cricket, an over consists of 6 balls. The outcomes of each ball is either zero runs, one run, two runs, three runs, four runs, five runs, or six runs. The probability mass function $P(X = i)$ is defined as probability of scoring i runs in an over of 6 balls. Assuming that there is an equal probability that outcome of each ball is any value from 0 to 6 runs, what is the value of $P(X = 6)$. Enter the answer up to 4 decimals accuracy [Medium 3 marks]
    0.0099, accepted range: 0.0090 to 0.011

9. Which of the following is(are) probability mass functions? [Medium 3 marks]

    (a) $P(X = x) = \frac{3}{4} \times \frac{1}{4^x}, x \in \{0, 1, 2, 3, 4, \cdots\}$

    (b) $P(X = x) = \frac{1}{8} \times (x + 1) \times \frac{1}{2^x}, x \in \{0, 1, 2, 3, 4, \cdots\}$

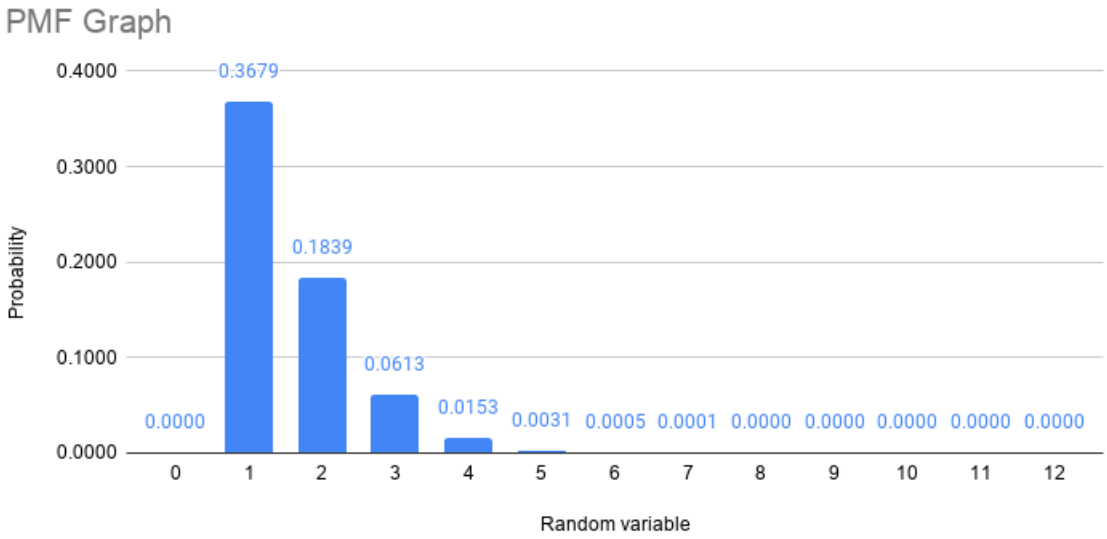
    (c) $P(X = x) = \frac{1}{2} \times \frac{1}{2^x}, x \in \{0, 1, 2, 3, 4, \cdots\}$

    (d) $P(X = x) = \frac{1}{4} \times (x + 1) \times \frac{1}{2^x}, x \in \{0, 1, 2, 3, 4, \cdots\}$

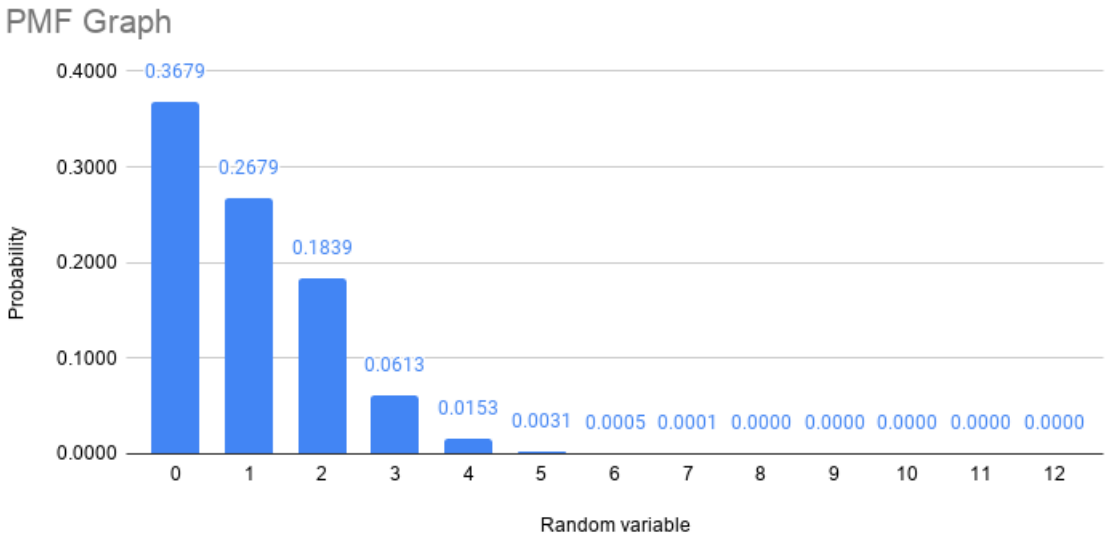
10. Which of the following is the correct graph of probability mass function? Note: The values are rounded off to 4 decimals. [Easy 1 mark]

$$P(X = x) = e^{-1} \times \frac{1}{x!}, x \in \{0, 1, 2, 3, \cdots\}$$

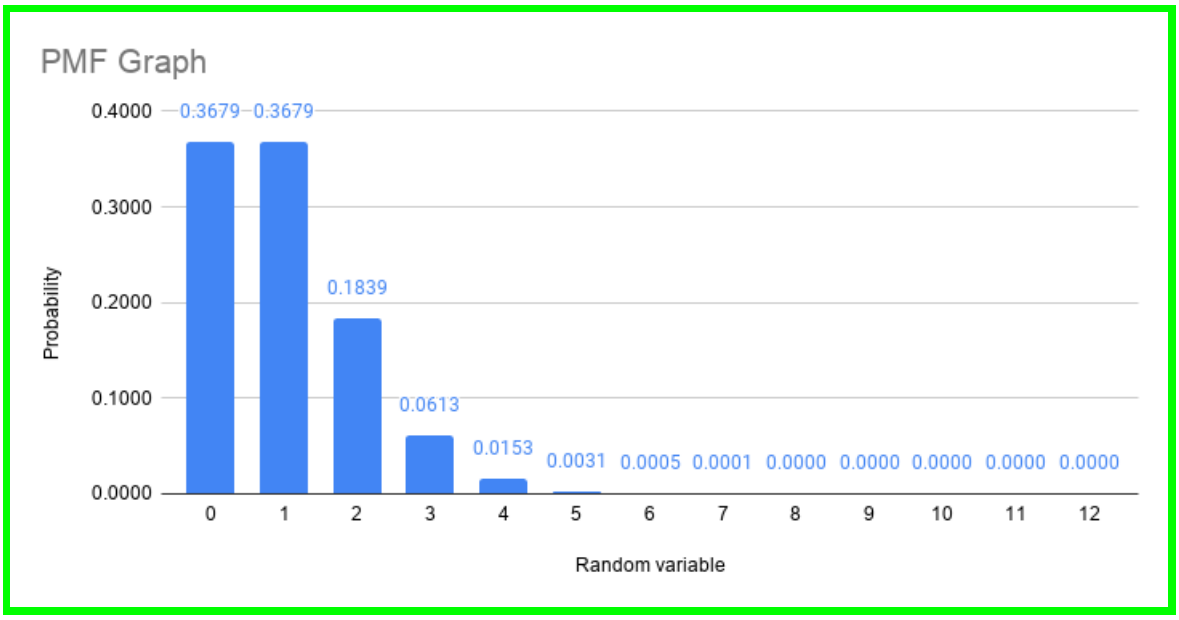
(a)



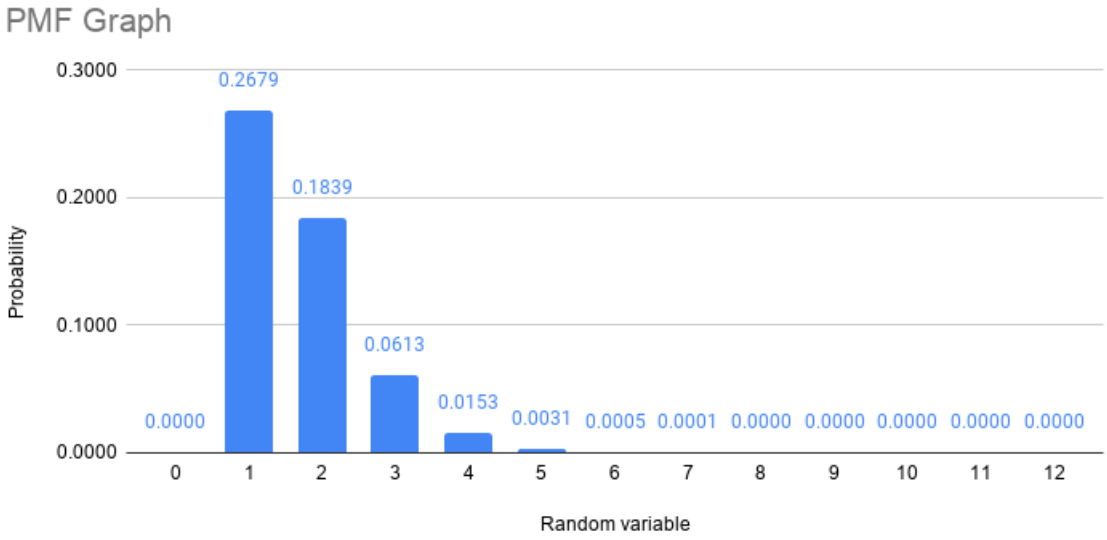
(b)



(c)


PMF Graph
0.4000
0.3000
0.2000
0.1000
0.0000
Probability
0.3679
0.3679
0.1839
0.0613
0.0153
0.0031
0.0005
0.0001
0.0000
0.0000
0.0000
0.0000
0.0000
0 1 2 3 4 5 6 7 8 9 10 11 12
Random variable

(d)


PMF Graph
0.3000
0.2000
0.1000
0.0000
Probability
0.0000
0.2679
0.1839
0.0613
0.0153
0.0031
0.0005
0.0001
0.0000
0.0000
0.0000
0.0000
0.0000
0 1 2 3 4 5 6 7 8 9 10 11 12
Random variable

Statistics for Data Science - 1

Week 9 Practice Assignment

Discrete random variables

Syllabus covered:

- Expectation of a random variable.
- Expectation of a function of random variable.
- Variance of a random variable.
- Variance of a function of a random variable.
- Standard deviation of a random variable.

1 Numerical answer type (NAT)

1. Suppose that random variable X can take values 1, 2, and 3. If $E[X] = 2.2$ and $P(X = 2) = P(X = 3)$, then find $P(X = 1)$. [Easy]
   0.2

2. A six sided die is biased. The numbers from two to six are equally likely to land face up, but number one is thrice as likely to land face up as each of the other numbers. If random variable Y is the score shown on the uppermost face, calculate the expected value of random variable Y. Enter the answer correct upto 3 decimals accuracy. [Easy]
   2.875

3. An entrepreneur has invested in four startups. If the startup makes profit, it will yield respective profits of 10, 20, 30 and 40 (in units of lakh). On the other hand, for each startup the entrepreneur may not gain profit, he will incur a loss of 3(in units of lakh). If the probabilities that the entrepreneur will make profits from these startups are, respectively, 0.2, 0.3, 0.4, and 0.1, what is the expected total profit?
   15 [Medium]

4. Let X be how much you win (in Rupees) in a lottery. In the lottery we draw a random number (outcome) between 1 and 1,000 (including both endpoints). The prize is computed as follows:

$$X = \begin{cases} 0 & 1 \leq x \leq 700 \\ 4 & 700 < x \leq 950 \\ 10 & 950 < x \leq 990 \\ 100 & 990 < x \leq 999 \\ 1000 & 999 < x \leq 1000 \end{cases}$$

Assume that each outcome is equally probable. Find the expected winning. [Medium]
3.3

5. Let X be a random variable with probability mass function as [Easy]

$$P(X = k) = \begin{cases} 0.2 & \text{for } k = 0 \\ 0.3 & \text{for } k = 1 \\ 0.5 & \text{for } k = 2 \\ 0 & \text{otherwise.} \end{cases}$$

i) Find $\text{Var}(X)$. Enter the answer correct upto 2 decimals accuracy.
0.61
ii) If $Y = (X + 2)^2$, find $\text{Var}(Y)$. Enter the answer correct upto 2 decimals accuracy.
23.25

6. Let X and Y be two independent random variables. Suppose that we know $\text{Var}(2X - Y) = 3$ and $\text{Var}(X + 3Y) = 6$. Find the value of $\text{Var}(X) + \text{Var}(Y)$. [Easy]
1.2

7. By investing in a particular stock, a person can make a profit of ₹10,000 in one year with probability 0.45 or take a loss of ₹7,000 with probability 0.55. What is this person's expected gain?
[Easy]
650

8. Four fair coins are tossed. If random variable X represents the number of heads, what is the standard deviation of random variable X? [Medium]
1

2 Multiple choice questions (MCQ)

1. Rohan and Jyoti works for the same company. Rohan's Diwali bonus is a random variable with a standard deviation of ₹1,000. If the Jyoti's bonus is 20% of the Rohan's bonus. Find the standard deviation of the bonus received by the Jyoti. [Easy]

    1. ₹1,000

2. ₹200
3. ₹50
4. ₹20,000

2. Let X be a random variable with probability mass function as

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

Then the value of $E[\frac{1}{3^X}]$ is [Hard]

(a) $\frac{p}{1+p}$

(b) $\frac{p}{2+p}$

(c) $\frac{p}{1-p}$

(d) $\frac{1-p}{2+p}$

Statistics for Data Science - 1

Week 10

Practice Assignment

1. In a four match test series between India and Australia, what is the probability that an Indian captain will win at least two tosses? (Both the captains have equal chance of winning a toss) [Easy]

   (a) 0.5

   (b) 0.25

   (c) 0.6875

   (d) 0.75

2. If $n = 4$ and $p = 0.70$, then what is the probability distribution of a binomial random variable X?

   a.

$$P(X = x) = \begin{cases} 0 & \text{for } x = 0 \\ 0.0837 & \text{for } x = 1 \\ 0.2646 & \text{for } x = 2 \\ 0.4116 & \text{for } x = 3 \\ 0.2401 & \text{for } x = 4 \end{cases}$$

   b.

$$P(X = x) = \begin{cases} 8.1 \times 10^{-3} & \text{for } x = 0 \\ 0.0756 & \text{for } x = 1 \\ 0.2646 & \text{for } x = 2 \\ 0.4116 & \text{for } x = 3 \\ 0.2401 & \text{for } x = 4 \end{cases}$$

   c.

$$P(X = x) = \begin{cases} 0.0837 & \text{for } x = 1 \\ 0.2646 & \text{for } x = 2 \\ 0.4116 & \text{for } x = 3 \\ 0.2401 & \text{for } x = 4 \end{cases}$$

d.

$$P(X=x)=\begin{cases}8.1\times10^{-3} & \text{for } x=0\\ 0.0756 & \text{for } x=1\\ 0.2646 & \text{for } x=2\\ 0.4116 & \text{for } x=3\\ 0.1523 & \text{for } x=4\end{cases}$$

Answer: b

3. If $Var(X)=1.2$ and $p=0.60$ of a binomial random variable X, then find the cumulative distribution function of X?

(a)

$$F(x)=\begin{cases}0 & \text{for } x<0\\ 0.01024 & \text{for } 0\le x<1\\ 0.08704 & \text{for } 1\le x<2\\ 0.31744 & \text{for } 2\le x<3\\ 0.66304 & \text{for } 3\le x<4\\ 1 & \text{for } 4\le x<5\end{cases}$$

(b)

$$F(x)=\begin{cases}0 & \text{for } x<0\\ 0.01024 & \text{for } 0\le x<1\\ 0.0768 & \text{for } 1\le x<2\\ 0.2304 & \text{for } 2\le x<3\\ 0.3456 & \text{for } 3\le x<4\\ 0.2592 & \text{for } 4\le x<5\\ 0.07776 & \text{for } 4\le x\end{cases}$$

(c)

$$F(x)=\begin{cases}0 & \text{for } x<0\\ 0.01024 & \text{for } 0\le x<1\\ 0.08704 & \text{for } 1\le x<2\\ 0.31744 & \text{for } 2\le x<3\\ 0.66304 & \text{for } 3\le x<4\\ 0.92224 & \text{for } 4\le x<5\\ 1 & \text{for } 5\le x\end{cases}$$

(d) Cumulative distribution function is not possible for a binomial random variable.

Answer: c

4. Which of the following statement(s) is(are) true for the pmf for a binomial distribution for same number of n trials?

   a. The shape of pmf is symmetric if the probability of success is 0.5.
   b. The shape of pmf is symmetric if the probability of failure is 0.5.
   c. The shape of pmf is right skewed if $p > 0.5$.
   d. The shape of pmf is left skewed if $p > 0.5$ if n is small.

5. An Indian cricket player scores more than 50 runs in 45% of all the matches he played. Assume this percentage holds true for future matches. The runs he scored in a match is independent of every other match. Find the probability that he will score more than 50 runs exactly 3 times in next 8 matches.(Correct up to 2 decimal accuracy)
   0.2568 Accepted range: 0.24-0.26

6. The probability of finding defective castings in a sample is p. The probability of finding exactly 2 defective castings in a randomly selected 3 castings is 0.027. Find the value of p. (Assume finding the number of defective castings follow binomial distribution) Note: $(p < 0.5)$)

   a. 0.2
   b. 0.1
   c. 0.35
   d. 0.4

7. The probability that a python developer writes a line of code without any error is 0.98. He makes those errors randomly. Making of an error in each line of code is independent of each other. He writes a code of 100 lines. What is the probability that he makes exactly one error? (Correct up to 2 decimal points)
   0.27 Accepted range:0.26-0.28

8. A school teacher has marbles of 6 different colors. All the 6 colored marbles are in equal proportion. He gave 18 marbles each to every student in the class. If Vijay is a student from the class and he likes blue colored marbles, what is the probability that he would receive less blue colored marbles than he expected? (Assume probability of getting a blue marble is equal for all the trials) (Correct up to 2 decimal points)
   0.4026 Accepted range: 0.39-0.41

9. According to a survey, 58% of teenagers in India play online games more than 3 hours per day. Assume this survey is true for current population of India. 40 teenagers are selected at random from this population. Let X be a binomial random variable which represents the number of teenagers who spends more than 3 hours in playing online games. Find the standard deviation of the probability distribution of X. (Correct up to 2 decimal points)
3.12 Accepted range: 3.00-3.20

10. For a binomial random variable, $n = 5, (1 - p) = 0.6$, find standard deviation of a binomial random variable.

    (a) 0.2
    (b) 0.12
    (c) 0.547
    (d) 1.09