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).

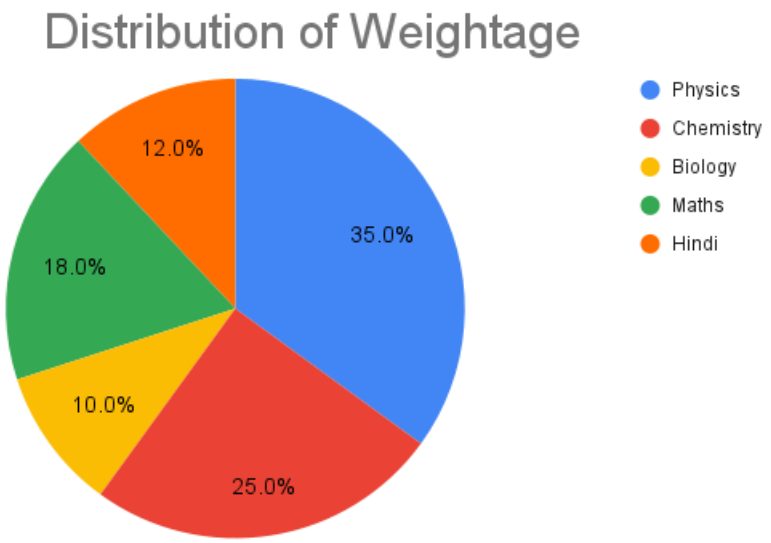


Figure 2.1.G: Weightage distribution of 5 different subjects

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.

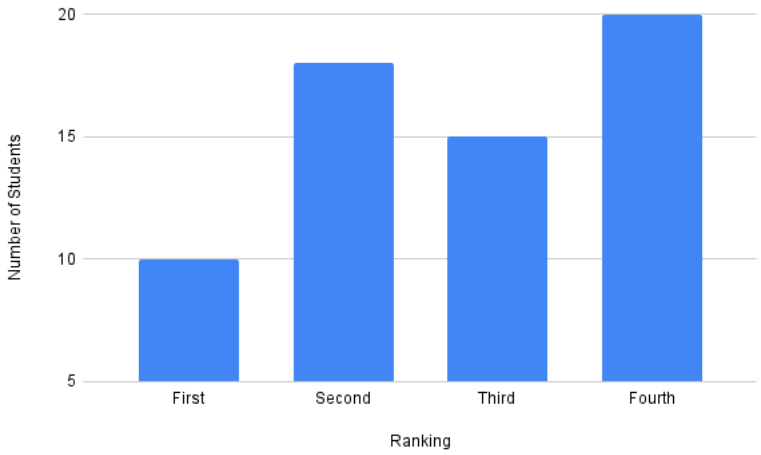
Also, the given data is nominal, which implies that we can arrange the bars in a specific order while plotting and it would be still appropriate for the representation of number of players.

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

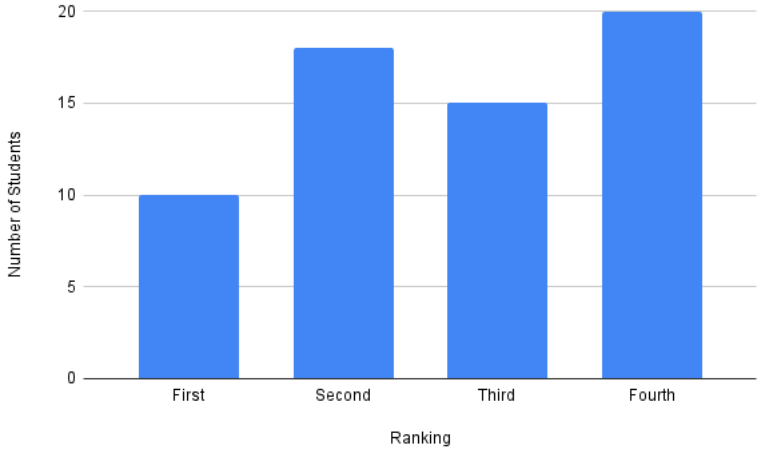
Hence, option (d) is correct.

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?

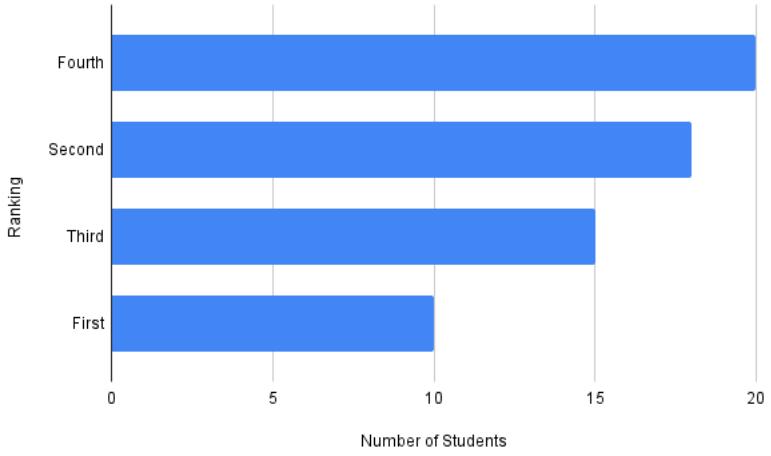
   a.



   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)

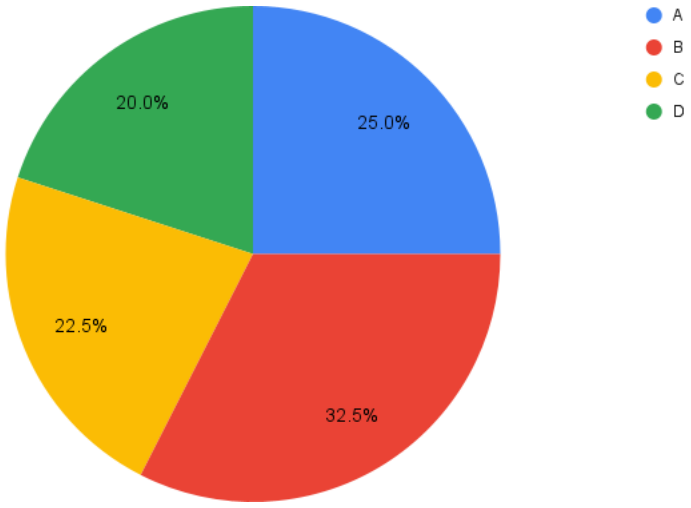


Figure 2.2.G: Distribution of grades in a Statistics class

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 $\frac{18}{20} = 0.9$