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]

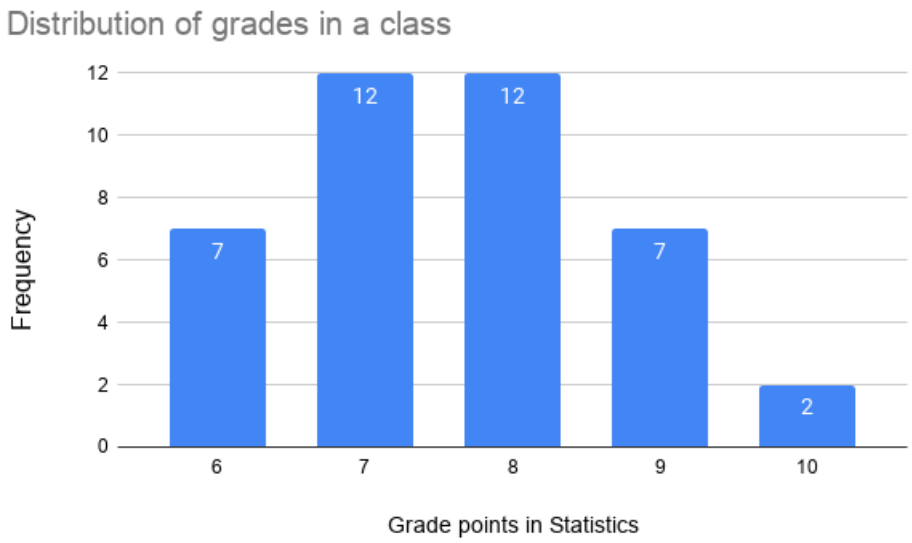
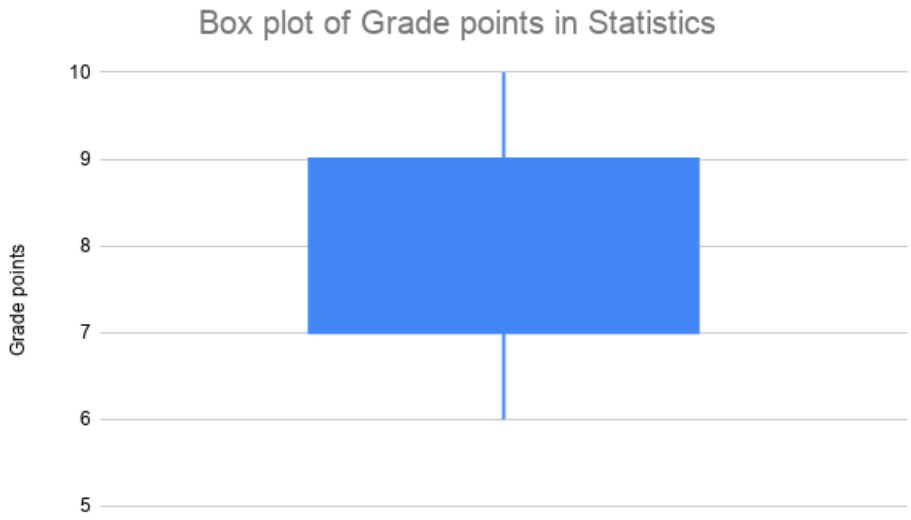
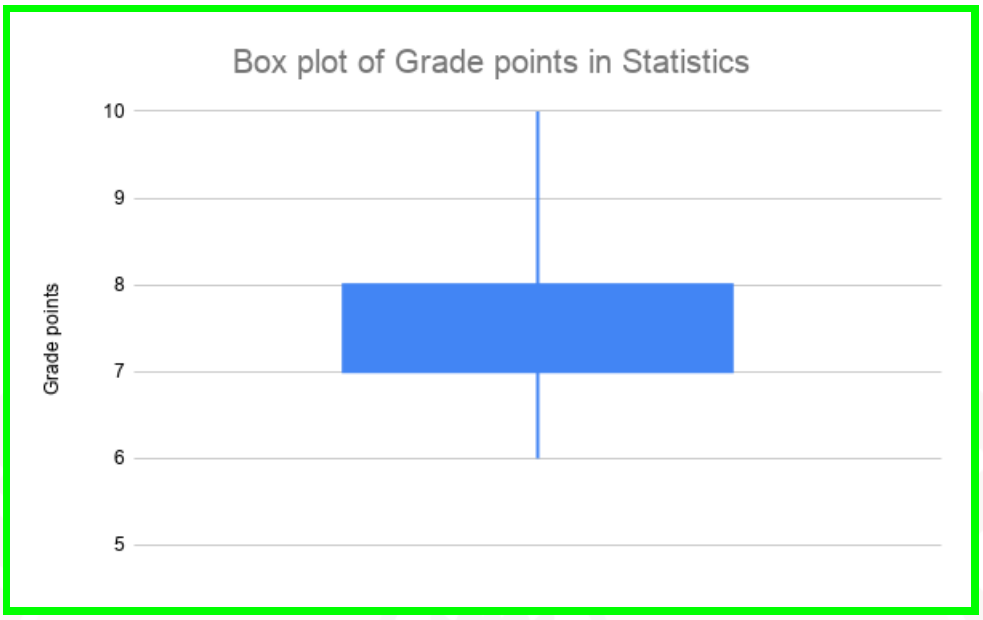


Figure S.1: Distribution of grade points in the Statistics exam dataset.

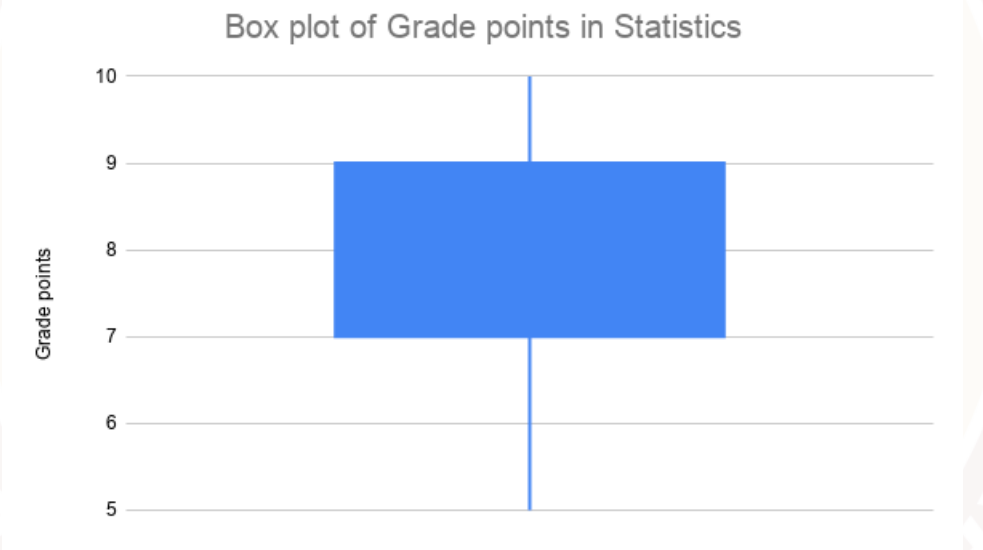


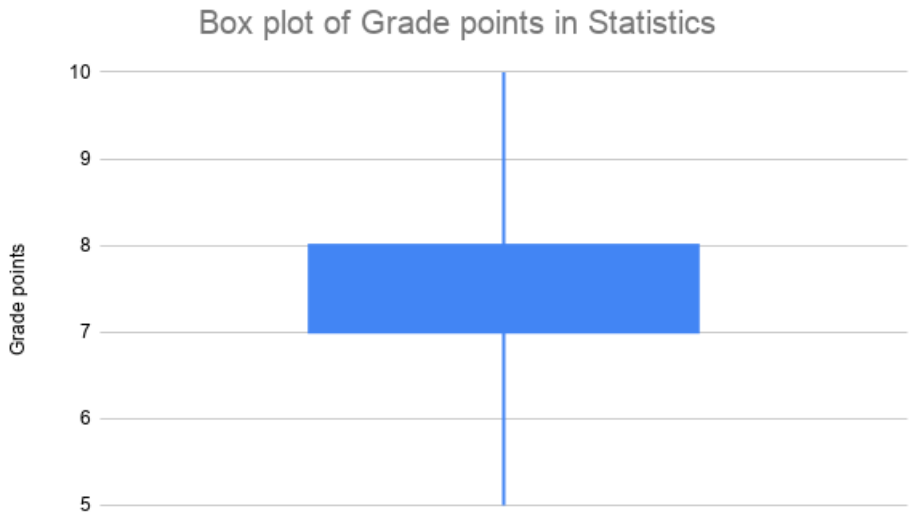
(a)

(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.

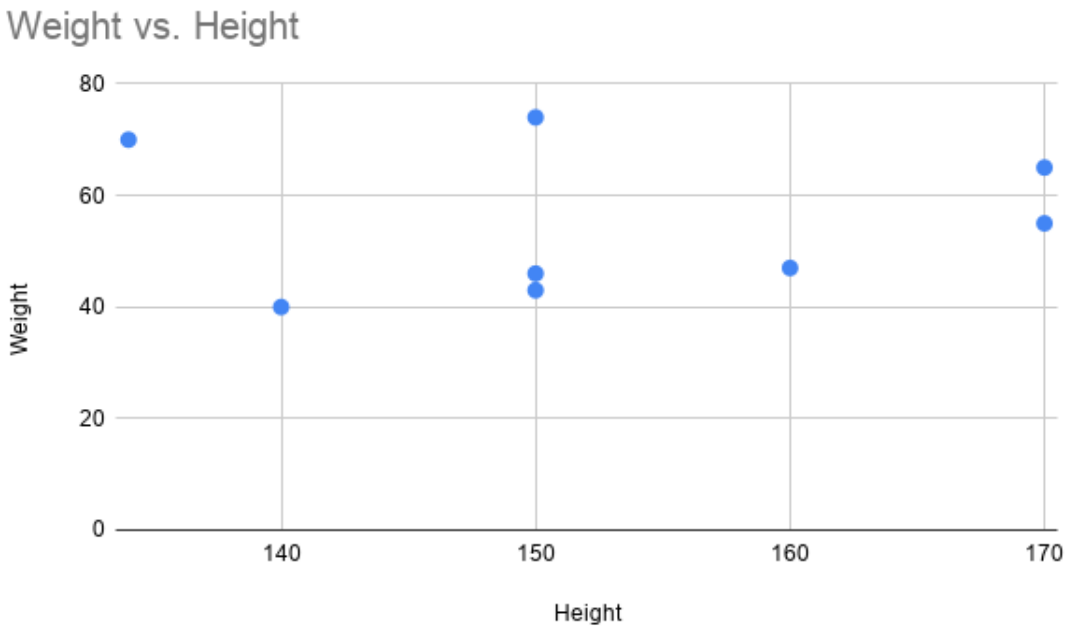
Name	Height(cm)	Weight(kg)
Anmol	140	40
Sujata	150	43
Subashish	170	55
Deepti	134	70
Rajesh	150	74
Kalpana	160	47
Nagarjuna	170	65
Shruti	150	46

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]

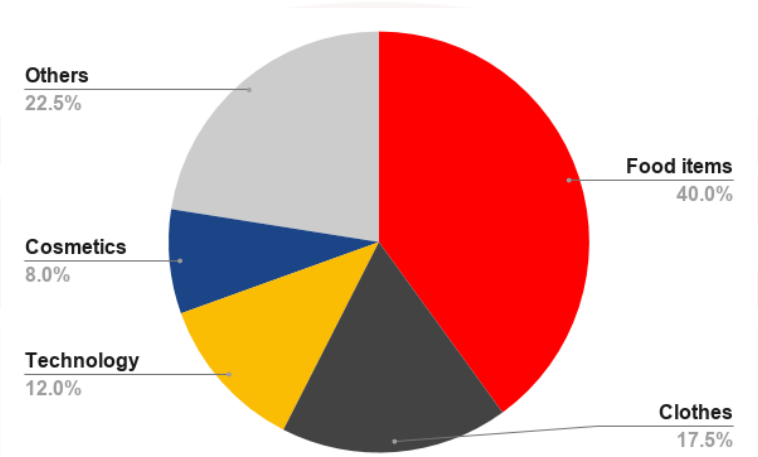


Figure S.2: House budget

(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]

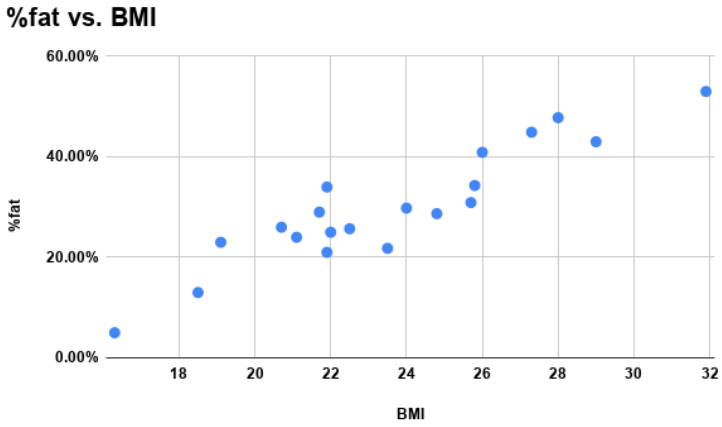


Figure S.4: %fat vs BMI

(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]

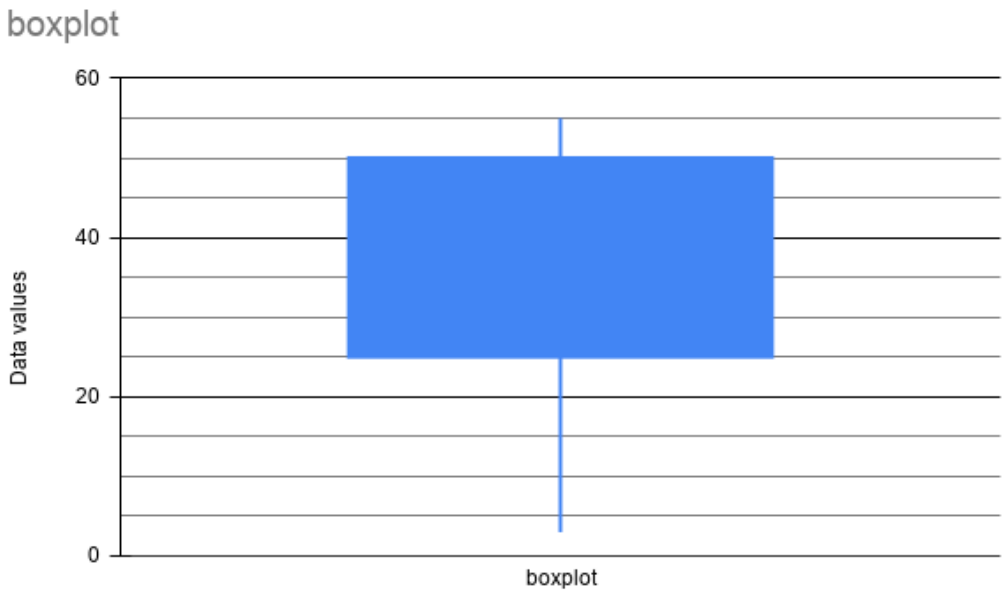


Figure S.5: Box plot

(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.