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$= \dfrac{T'}{N} = \dfrac{109}{8} = 13.625$

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

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

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

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

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, $A = \left(8^2 + \dfrac{8 \times 13^2}{7}\right) \times (8-1) = 1800$
Therefore, $B = 1800 - 13^2 + 18^2 = 1955$
Hence, sample variance for the original dataset $= \dfrac{1955}{8-1} - \dfrac{109^2}{8 \times 7} = 67.125$

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 $\dfrac{2n+1}{3}$.
c. Mode is $n-1$.
d. Mean is n.
e. Mean is $\dfrac{n+1}{2}$.
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 $= f_1 + f_2 + ... + f_n = 1 + 2 + ... + n = \dfrac{n(n+1)}{2}$
Sum of observations $= f_1x_1 + f_2x_2 + ... + f_nx_n = 1 \times 1 + 2 \times 2 + ... + n \times n$

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

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

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 $= f_1 + f_2 + ... + f_{42} = 1 + 2 + ... + 42 = \dfrac{42(42+1)}{2}$

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

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

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

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 $= (76.5 + 3) \times \frac{40}{100} = 31.8$.