HW 1

Chandan Suri – CS4090

1. @ Sample mean for the sample $x_1, \ldots, x_n$ ⇒

$$\bar{x}_n = \frac{\sum_{i=1}^{n} x_i}{n}$$ where n = sample size.

$$\Rightarrow \boxed{\sum_{i=1}^{n} x_i = n\bar{x}_n} \quad \text{---(1)}$$

When we add an additional observation x_{n+1} to the original sample as stated above,

Mean of the new sample with an additional observation can be written as ⇒

$$\bar{x}_{n+1} = \frac{\sum_{i=1}^{n+1} x_i}{n+1}$$

$$\Rightarrow \sum_{i=1}^{n+1} x_i = (n+1)\,\bar{x}_{n+1}$$

↓

We can break the L.H.S in 2 components as,

$$\Rightarrow \sum_{i=1}^{n} x_i + x_{n+1} = (n+1)\,\bar{x}_{n+1}$$

Using (1) from above,

$$\Rightarrow n\bar{x}_n + x_{n+1} = (n+1)\,\bar{x}_{n+1}$$

$$\therefore \boxed{\bar{x}_{n+1} = \frac{n\bar{x}_n + x_{n+1}}{(n+1)}}$$

Thus, $\bar{x}_{n+1}$ can be computed using $\bar{x}_n$ & x_{n+1}.

Ans.

(b.) According to the formula for variance,

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

for $(n+1)$ entries, we replace n with $n+1$ on both sides (LHS & RHS) :-

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

$$\Rightarrow \quad S_{n+1}^2 = \frac{1}{n} \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2$$

Multiplying both sides by n,

$$n S_{n+1}^2 = \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2$$

Adding and subtracting $\bar{x}_n$ on the RHS side inside the squared formula, (won't change anything)

$$n S_{n+1}^2 = \sum_{i=1}^{n+1} (\underbrace{x_i - \bar{x}_n}_{\text{Term 1}} + \underbrace{\bar{x}_n - \bar{x}_{n+1}}_{\text{Term 2}})^2$$

Using Binomial Expansion formula and treating term1 and term 2 as specified above,

$$\Rightarrow \quad n S_{n+1}^2 = \sum_{i=1}^{n+1} \left[(x_i - \bar{x}_n)^2 + 2 (x_i - \bar{x}_n)(\bar{x}_n - \bar{x}_{n+1}) + (\bar{x}_n - \bar{x}_{n+1})^2 \right]$$

As summation is distributive,

$$\Rightarrow \quad n S_{n+1}^2 = \sum_{i=1}^{n+1} (x_i - \bar{x}_n)^2 + 2 \underbrace{(\bar{x}_n - \bar{x}_{n+1})}_{\text{this doesn't have i so can take out of summation}} \sum_{i=1}^{n+1} (x_i - \bar{x}_n) + \sum_{i=1}^{n+1} (\bar{x}_n - \bar{x}_{n+1})^2$$

Similarly this can be taken out.

$$\Rightarrow n S_{n+1}^2 = \sum_{i=1}^{n+1} (x_i - \bar{x}_n)^2 + 2(\bar{x}_n - \bar{x}_{n+1}) \sum_{i=1}^{n+1} (x_i - \bar{x}_n) + (n+1)(\bar{x}_n - \bar{x}_{n+1})^2 \quad \text{——— (1)}$$

By taking out last term out of summation for 1st term on RHS, we can write it as,

$$\sum_{i=1}^{n+1} (x_i - \bar{x}_n)^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + (x_{n+1} - \bar{x}_n)^2 \quad \text{——— (2)}$$

& $\sum_{i=1}^{n+1} (x_i - \bar{x}_n)$ can be written as,

[Using the formula for mean]

$$\sum_{i=1}^{n+1} (x_i - \bar{x}_n) = (n+1)(\bar{x}_{n+1} - \bar{x}_n) \quad \text{——— (3)}$$

Substituting (2) and (3) in (1) above, we get:-

$$\Rightarrow n S_{n+1}^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 + (x_{n+1} - \bar{x}_n)^2 + 2(\bar{x}_n - \bar{x}_{n+1}).(n+1)(\bar{x}_{n+1} - \bar{x}_n) + (n+1)(\bar{x}_n - \bar{x}_{n+1})^2$$

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

$$\text{So, } \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = (n-1) S_n^2$$

Putting this value,

$$\Rightarrow n S_{n+1}^2 = (n-1) S_n^2 + (x_{n+1} - \bar{x}_n)^2 + 2(n+1).(\bar{x}_n - \bar{x}_{n+1}).(\bar{x}_{n+1} - \bar{x}_n) + (n+1)(\bar{x}_n - \bar{x}_{n+1})^2$$

Rewriting the 3rd term on RHS by taking out -1 and combining terms in it,

$$\Rightarrow n S_{n+1}^2 = (n-1) S_n^2 + (x_{n+1} - \bar{x}_n)^2 - 2(n+1)(\bar{x}_n - \bar{x}_{n+1})^2 + (n+1)(\bar{x}_n - \bar{x}_{n+1})^2$$

$$\Rightarrow nS^2_{n+1} = (n-1)S^2_n + (x_{n+1} - \bar{x}_n)^2 - (n+1)(\bar{x}_n - \bar{x}_{n+1})^2$$

Using the result from part (a) of this question and substituting the value of $\bar{x}_{n+1}$ in the above equation, we get:-

$$nS^2_{n+1} = (n-1)S^2_n + (x_{n+1} - \bar{x}_n)^2 - (n+1)\left[\bar{x}_n - \frac{n.\bar{x}_n + x_{n+1}}{n+1}\right]^2$$

$$\Rightarrow nS^2_{n+1} = (n-1)S^2_n + (x_{n+1} - \bar{x}_n)^2 - (n+1)\left[\frac{(n+1)\bar{x}_n - n\bar{x}_n - x_{n+1}}{n+1}\right]^2$$

$$\Rightarrow nS^2_{n+1} = (n-1)S^2_n + (x_{n+1} - \bar{x}_n)^2 - (n+1)\left[\frac{n\bar{x}_n + \bar{x}_n - n\bar{x}_n - x_{n+1}}{n+1}\right]^2$$

$$\Rightarrow nS^2_{n+1} = (n-1)S^2_n + (x_{n+1} - \bar{x}_n)^2 - (n+1)\left[\frac{\bar{x}_n - x_{n+1}}{n+1}\right]^2$$

$$\Rightarrow nS^2_{n+1} = (n-1)S^2_n + (x_{n+1} - \bar{x}_n)^2 - \frac{1}{(n+1)}(\bar{x}_n - x_{n+1})^2$$

we can rewrite $(\bar{x}_n - x_{n+1})^2$ as $\Rightarrow (x_{n+1} - \bar{x}_n)^2$

as $(a-b)^2 = (b-a)^2$,

$$nS^2_{n+1} = (n-1)S^2_n + (x_{n+1} - \bar{x}_n)^2 - \frac{1}{(n+1)}(x_{n+1} - \bar{x}_n)^2$$

$$nS^2_{n+1} = (n-1)S^2_n + \left[1 - \frac{1}{(n+1)}\right](x_{n+1} - \bar{x}_n)^2$$

$$nS^2_{n+1} = (n-1)S^2_n + \left[\frac{n+1-1}{n+1}\right](x_{n+1} - \bar{x}_n)^2$$

$$\boxed{nS^2_{n+1} = (n-1)S^2_n + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2}$$

Hence, Proved. Ans

(c.)

Sample size = 15

Mean ($\bar{x}_n$) of 15 strands = 12.58 mm.

Sample standard deviation (S_n) = 0.512 mm

When a 16th strand (x_{16}) with value 11.8 is added, (x_{n+1})

New mean ($\bar{x}_{n+1}$) ⇒

$$\bar{x}_{n+1} = \frac{x_1 + x_2 + \dots + x_{n+1}}{(n+1)}$$

can be re-written as,

$$(n+1)(\bar{x}_{n+1}) = \underbrace{\sum_{i=1}^{15} x_i}_{\hookrightarrow (n\bar{x}_n)} + x_{16}$$

So,

$$\bar{x}_{n+1} = \frac{n\,\bar{x}_n + x_{n+1}}{(n+1)} \qquad (n=15).$$

So,

$$\bar{x}_{n+1} = \frac{15 \times 12.58 + 11.8}{16}$$

$$\boxed{\bar{x}_{n+1} = 12.53 \text{ mm}}$$

Ans

for calculating S_{n+1}, we can use the result from part (b) of the question,

Variances are related as ⇒

$$n S^2_{n+1} = (n-1) S^2_n + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2$$

Substituting the values in the above equation,

$$15 \cdot S^2_{n+1} = 14 \times (0.512)^2 + \frac{15}{(15+1)}(11.8 - 12.58)^2$$

Dividing both sides by 15,

$$S^2_{n+1} = \frac{1}{15}\left[14 \times (0.512)^2 + \frac{15}{16}(11.8 - 12.58)^2\right]$$

$$S^2_{n+1} = \frac{14}{15}(0.512)^2 + \frac{1}{16}(11.8 - 12.58)^2$$

$$S^2_{n+1} = 0.24 + 0.04$$

$$\boxed{S^2_{n+1} = 0.28}$$

Using the above value for variance,

Standard deviation (new) $\Rightarrow \sqrt{S^2_{n+1}}$

$$S_{n+1} = \sqrt{0.28} = 0.53$$

$$\boxed{S_{n+1} = 0.53 \text{ mm}}$$

Ans.

2.

(a) The variables are as follows:-

(i) Borough :- This is a categorical variable. This is because it's values are some defined set of categories where it is not suitable to add, subtract or take the average of those values / categories.

Possible values are:- Manhattan, Bronx, Brooklyn, Queens & Staten Island.

(ii) Monthly Rent :- firstly, this is a numerical variable. This is because it is suitable to add, subtract or take the average of the values.

furthermore, this is a continuous variable. This is because the range of values of this variable span the entire number line (on the positive side) and it's mostly the result of some measurement.

Possible values :- Any reasonable positive value. (potentially spans the full positive range including decimal values).

(iii) Monthly income :- firstly, this is a numerical variable. This is because it is suitable to add, subtract or take the average of the values.

furthermore, this is a continuous variable. This is because the range of values of this variable can potentially span the entire numbers line (on the positive side) and it's mostly the result of some measurement.

Possible values:- Any reasonable positive value.
(potentially spans the full positive range including decimal values)

(iv.) Number of bedrooms:- firstly, this is a numerical variable. This is because it is suitable to add, subtract or take the average of the values.

furthermore, this is a discrete variable. This is because it can potentially have only some finite number of values.

Possible values:- 1, 2, 3, ..., some reasonable n.
(set of natural numbers) (positive).

(v.) Presence of person elevators:- firstly, this is a categorical variable. This is because it's values are some defined set of categories where it is not suitable to add, subtract or take the average of the values.

Possible values:- Yes, No.
If it's unknown, another potential category could be "Unknown" or "Maybe".

(b.) In a survey or such a study as mentioned here, the procedure to draw a representative sample should essentially be able to avoid the bias. So, to make sure that the households were, at least, approximately selected according to a simple random sampling, it should be made sure that every case spanning different kind of necessary attributes in the population is randomly picked with the same probability.

To achieve this, one of the best ways is to use the most recent census data. This kind of data generally takes care of the biases and is representative.

C.

1. Ask randomly selected Columbia students & faculty:-

Just taking into consideration the students and faculty of the Columbia University won't be a good representative sample because of the inherent bias in the sampling technique. This is mainly because it is quite probable that most of the Columbia students and faculty might be living in close proximity to the college campus. Thus, mostly or a vast majority would be in Manhattan. This would not cover or span the data for all the boroughs. Thus, the mix of residents won't be a good representative sample.

Also, the monthly rent could also be a bit higher (or the average monthy rent) in comparison to the other boroughs. Additionally, the monthy income range would also be very limited as we are not taking a good representation of society here. Thus, this sample will very likely be biased & not a good representative sample.

2. Ask randomly selected NYC residents on facebook & Instagram:-

At first, this might look like a good procedure to get a representative sample that is unbiased but that's not the case!

This is primarily because it could be possible that people belonging to certain strata of the society are not active on social media. furthermore, it's quite likelly that aged residents staying in certain boroughs may not even have accounts on facebook or Instagram. Also, it's quite probable that some younger or middle aged residents may not have accounts who are not interested in social media

Also, there might be some people who are socially active on these accounts but may not even respond to the survey online. Thus, this may lead to a sample which is not representative and has an inherent bias in it.

(3.) Ask randomly selected people at randomly selected NYC supermarkets :-

For this case, it might be possible that some households rely on delivery for their grocery needs or they generally order online & thus, may never visit the grocery stores. Also, it's generally the case where only certain members of the family come to the grocery stores. for example, it might be possible that small children might never visit the grocery stores. It could also be gender-biased for some households. All these factors in combination might result in a biased sample that is not a good representative.

(3.)

Data on concentration (EU/mg) in settled dust :-
(for urban homes — U)
(for farm homes - F).

U :- 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0
17.0 23.0

F :- 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0
8.9 21.0 9.2 3.0 2.0 0.3

(a) Sample mean for 2 samples above :-

for urban homes ⇒

$$\text{Sample Mean} = \frac{\text{Sum of all concentrations in sample dust}}{\text{Total number of data points for concentration of urban homes.}}$$

$$\Rightarrow \frac{(6.0 + 5.0 + 11.0 + 33.0 + 4.0 + 5.0 + 80.0 + 18.0 + 35.0 + 17.0 + 23.0)}{11}$$

$$\Rightarrow \frac{237.0}{11} = \boxed{21.55} \text{ EU/mg}$$

Ans.

for farm homes ⇒

$$\text{Sample Mean} = \frac{\text{Sum of all concentrations in sample dust for farm homes}}{\text{Total number of data points for concentration of farm homes.}}$$

$$\Rightarrow (4.0 + 14.0 + 11.0 + 9.0 + 9.0 + 8.0 + 4.0 + 20.0 + 5.0 + 8.9 + 21.0 + 9.2 + 3.0 + 2.0 + 0.3) / 15$$

$\Rightarrow \frac{128.4}{15} = \boxed{8.56}$ EU/mg

Ans.

Comparison of the sample means for both the samples:-

The sample mean endotoxin concentration in settled dust for urban homes is almost 2.5 times of the sample mean endotoxin concentration in settled dust for farm homes. Thus, the sample mean concentration for urban homes is much larger in comparison to that of the farm homes.

Ans.

(b) Sample median for 2 samples above:-

for urban homes $\Rightarrow$

As a first step, we will sort the data points (endotoxin concentration values) in order to find the median.

Ordered data points $\Rightarrow$ (Ascending order)

$\Rightarrow$ 4.0 5.0 5.0 6.0 11.0 17.0 18.0 23.0 33.0 35.0 80.0

There are 11 (n) elements in our data set above,

So,

Median concentration for sample of urban homes $\Rightarrow$

$\left(\frac{n+1}{2}\right)^{th}$ element in sorted list

(Middle element in a sorted list containing odd number of elements)

$\Rightarrow$ Middle$^{(6th)}$ element is 17.0.

$\Rightarrow \boxed{17.0}$ EU/mg

Ans.

for farm homes ⇒

As a first step, we will sort the data points in ascending order to find the median.

Ordered data points :-

⇒ 0.3 2.0 3.0 4.0 4.0 5.0 8.0 8.9 9.0 9.0 9.2 11.0 14.0 20.0 21.0

There are 15 elements in our data set above,

So,

Median concentration for sample of farm homes ⇒

$\left(\frac{n+1}{2}\right)^{th}$ element in sorted list containing odd number of elements

⇒ Middle element (8th) is 8.9.

⇒ **8.9** EU/mg

Ans.

Comparison :-

The sample median (endotoxin concentration) for urban homes is much larger (almost twice) than that for the farm homes.

Comparison of mean & median for urban homes ⇒

Mean for urban homes = 21.55 EU/mg

Median for urban homes = 17 EU/mg

As we can see above, mean for the urban homes data sample is larger than the median for the same sample by 4.55 EU/mg. This can be clearly attributed to the fact that the

data for the endotoxin concentration in settled dust for urban homes contains an outlier with the value of 80.0. As mean can be quite severely affected by the presence of outliers, the mean value is larger in comparison to the median. Median value is not affected by the presence of outliers as for computing the median value, all the data points are not computationally used (use just find the center). Thus, the presence of the outlier in the sample is the reason for the larger value of mean. Although both mean and median are the measures of the center, still both are computed in very different ways. Ans.

(4.)

Sample size = 100.

Number of contaminating particles on a silicon wafer ⇒

Number of particles	frequency
0	1
1	2
2	3
3	12
4	11
5	15
6	18
7	10
8	12
9	4
10	5
11	3
12	1
13	2
14	1
Total ⇒	100

(a) Proportion of the sampled wafers that had atleast 1 particle ⇒

$$\frac{\text{Total Number of Sampled wafers} - \text{Number of sampled wafers with zero number of particles}}{\text{Total number of sampled wafers}}$$

$$\Rightarrow \frac{100 - 1}{100} \Rightarrow \frac{99}{100}$$

$$\Rightarrow \boxed{0.99} \quad (99\%)$$

Ans.

Proportion of the sampled wafers with atleast 5 particles ⇒

$$\frac{\text{Total number of sampled wafers} - [\text{Number of sampled wafers with less than 5 particles}]}{\text{Total number of sampled wafers}}$$

⇒ $$\frac{\text{Total number of sampled wafers} - [\text{Number of sampled wafers with number of particles as } 0, 1, 2, 3, 4]}{\text{Total number of sampled wafers.}}$$

⇒ $$\frac{100 - [1 + 2 + 3 + 12 + 11]}{100}$$

$$= \frac{100 - 29}{100} = \frac{71}{100} = \boxed{0.71} \quad (71\%)$$

Ans.

(b) Proportion of the sampled wafers that had between 5 and 10 particles ⇒ (inclusive case):-

$$\frac{\text{Number of sampled wafers with number of particles as } 5, 6, 7, 8, 9, 10}{\text{Sample Size}}$$

⇒ $$\frac{[15 + 18 + 10 + 12 + 4 + 5]}{100}$$

⇒ $$\frac{64}{100} \Rightarrow \boxed{0.64} \quad (64\%)$$

Ans.

Proportion of sampled particles that had strictly between 5 and 10 particles =>

=> Number of sampled wafers with 6 number of particles +
Number of sampled wafers with 7 number of particles +
Number of sampled wafers with 8 number of particles +
Number of sampled wafers with 9 number of particles

Sample Size.

$$\Rightarrow \frac{[18 + 10 + 12 + 4]}{100} \Rightarrow \frac{44}{100}$$

$$\Rightarrow \boxed{0.44} \quad (44\%)$$

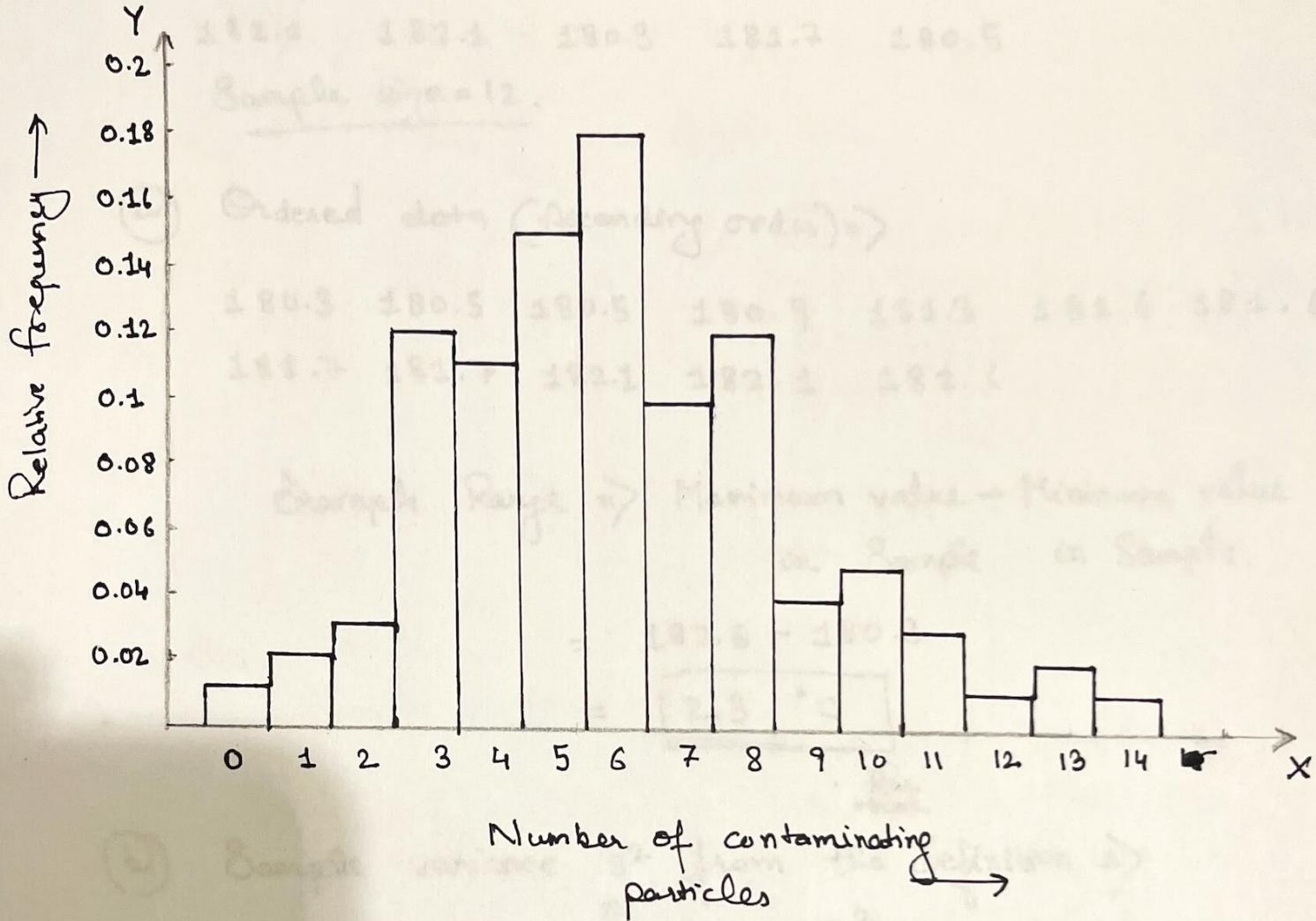
Ans.

(c.)

Number of particles	frequency	Relative frequency
0	1	1/100 = 0.01
1	2	2/100 = 0.02
2	3	3/100 = 0.03
3	12	12/100 = 0.12
4	11	11/100 = 0.11
5	15	15/100 = 0.15
6	18	18/100 = 0.18
7	10	10/100 = 0.10
8	12	12/100 = 0.12
9	4	4/100 = 0.04
10	5	5/100 = 0.05
11	3	3/100 = 0.03
12	1	1/100 = 0.01
13	2	2/100 = 0.02
14	1	1/100 = 0.01

(c.) Continued...

Histogram with Relative frequency (silicon wafers) vs. Number of contaminating particles =>



The histogram is almost symmetric (not exactly symmetric though) and seems to be unimodal. (with a mode at number of particles = 6).

However, the distribution is also a little positively skewed and has a few smaller modes as well. Positive skewness is quite evident by just looking at the frequencies too as for ≤ 6 number of particles, 62 is the cumulative frequency of the silicon wafers.

Ans.

(5.) Data on melting point (°C) for each of the 12 specimens of the polymer is:-

180.5 181.7 180.9 181.6 182.6 181.6 181.3
182.1 182.1 180.3 181.7 180.5

Sample size = 12.

(a) Ordered data (Ascending order) ⇒

180.3 180.5 180.5 180.9 181.3 181.6 181.6
181.7 181.7 182.1 182.1 182.6

Sample Range ⇒ Maximum value in Sample − Minimum value in Sample.

= 182.6 − 180.3

= **2.3 °C**

Ans

(b) Sample variance s^2 from the definition ⇒

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

As, if we subtract/add any constant to all the data points, the sample variance or the standard deviation shouldn't change,

So, for making the computations easier, we can subtract 180 from all the data points to make values smaller which in turns makes the process of calculating variance easier.

After deducting 180 ⇒ (from ordered data).

⇒ 0.3 0.5 0.5 0.9 1.3 1.6 1.6 1.7 1.7 2.1 2.1 2.6

(I have made the deductions above from the ordered data list, this can be done on original data set too, it doesn't impact the final calculations).

$n = 12$ (sample size).

$\bar{x}_n \Rightarrow$ Mean of the sample data

$$\Rightarrow \sum_{i=1}^{n} x_i \Big/ n \Rightarrow \frac{[0.3 + 0.5 + 0.5 + 0.9 + 1.3 + 1.6 + 1.6 + 1.7 + 1.7 + 2.1 + 2.1 + 2.6]}{12}$$

$$= \frac{16.9}{12} = \underline{\underline{1.41}} \ ^\circ C$$

So,

Sample variance (s^2) $\Rightarrow$

$$\Rightarrow \frac{1}{12-1}\left[(0.3-1.41)^2 + (0.5-1.41)^2 + (0.5-1.41)^2 + (0.9-1.41)^2 + (1.3-1.41)^2 + (1.6-1.41)^2 + (1.6-1.41)^2 + (1.7-1.41)^2 + (1.7-1.41)^2 + (2.1-1.41)^2 + (2.1-1.41)^2 + (2.6-1.41)^2\right]$$

$$\Rightarrow \frac{1}{11} \cdot [1.23 + 0.83 + 0.83 + 0.26 + 0.01 + 0.04 + 0.04 + 0.08 + 0.08 + 0.48 + 0.48 + 1.42]$$

$$\Rightarrow \frac{5.78}{11} = 0.5254 \approx \boxed{0.53} \ ^\circ C^2$$

Ans.

(c) Sample Standard deviation (s) ⇒

$$= \sqrt{\text{Sample Variance}}$$

$$= \sqrt{s^2}$$

from previous part (b), we know $s^2 = 0.53$,

$$\therefore \Rightarrow \sqrt{0.53} = \boxed{0.73\ ^\circ C}$$

Ans.

(d) Sample variance (s^2) using the shortcut method ⇒

$$\frac{n}{n-1}\left[\overline{x^2} - \overline{x}^2\right]$$

$\overline{x^2}$: mean of squared data points; $\overline{x}^2$: (mean of data points)^2

Computing the mean of the squared data points ⇒

$$\overline{x^2} \Rightarrow \frac{1}{12}\left[(0.3)^2 + 2(0.5)^2 + (0.9)^2 + (1.3)^2 + 2(1.6)^2 + 2(1.7)^2 + 2(2.1)^2 + (2.6)^2\right]$$

$$= \frac{1}{12}[29.57] = \frac{29.57}{12}$$

Computing the mean of the data points ⇒ (this we computed before!)

$$\overline{x} \Rightarrow \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{1}{12} \times 16.9$$

$$= \frac{16.9}{12}$$

Let's substitute these values in the formula above ⇒

$$s^2 = \frac{12}{11}\left[\frac{29.57}{12} - \left(\frac{16.9}{12}\right)^2\right]$$

$$= \frac{12}{11}\left[2.464 - 1.983\right]$$

$$= 0.5247 \approx \boxed{0.53\ ^\circ C^2}$$

Ans.

(6.) Data on oxidation-induction time (min) ⇒

87 103 130 160 180 195 132 145 211 105 145
153 152 138 87 99 93 119 129

Sample size = Number of elements in sample data = 19

$\therefore \underline{n = 19}$

(a) As we need to calculate standard deviation & sample variance which aren't affected if we add/subtract a constant from all the data points in the sample. So, for making our calculations easy, we can subtract 80 from all the data points to deal with smaller values.

→ After Subtracting 80 from all the data points above ⇒

7 23 50 80 100 115 52 65 131 25 65
73 72 58 7 19 13 39 49

for calculating sample variance, we will use the shortcut method here ⇒

$$\frac{n}{n-1}\left(\overline{x^2} - \overline{x}^2\right)$$

$\overline{x^2}$ ↓ mean of squared data points

$\overline{x}^2$ ↓ mean of data points

n ⇒ sample size

Squared Mean of data points (using the data after subtracting 80 from all) ⇒

$$\overline{x}^2 = \left(\sum_{i=1}^{n} x_i / n\right)^2$$

$$\text{Mean} = \overline{x} = \left(\sum_{i=1}^{n} x_i\right) / n$$

$$= \frac{\text{Sum of all the data}}{\text{Sample size}}$$

$$\Rightarrow \frac{[7 + 23 + 50 + 80 + 100 + 115 + 52 + 65 + 131 + 25 + 65 + 73 + 72 + 58 + 7 + 19 + 13 + 39 + 49]}{19}$$

$$\Rightarrow \boxed{\frac{1043}{19}} = \underline{54.9}$$

So, $\bar{x}^2 = (54.9)^2 = \underline{3014.01}$

Now, for calculating mean of squared data points, Let's square the data points (using the data from which 80 has been subtracted) $\Rightarrow$

49 529 2500 6400 10000 13225 2704
4225 17161 625 4225 5329 5184 3364
49 361 169 1521 2401

$\therefore$ Mean of squared data points $(\overline{x^2}) \Rightarrow$

$$= \left(\sum_{i=1}^{n} x_i^2\right) / n$$

$$= \frac{[7^2 + 23^2 + 50^2 + 80^2 + 100^2 + 115^2 + 52^2 + 65^2 + 131^2 + 25^2 + 65^2 + 73^2 + 72^2 + 58^2 + 7^2 + 19^2 + 13^2 + 39^2 + 49^2]}{19}$$

Using squared data,

$$= \frac{[49 + 529 + 2500 + 6400 + 10000 + 13225 + 2704 + 4225 + 17161 + 625 + 4225 + 5329 + 5184 + 3364 + 49 + 361 + 169 + 1521 + 2401]}{19}$$

$$= \boxed{\frac{80021}{19}} = \underline{4211.63}$$

Putting the values of $\overline{x^2}$ and $\bar{x}^2$ in the formula for sample variance $\Rightarrow$

$$s^2 \Rightarrow \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)$$

$$\Rightarrow \frac{19}{18}\left(\frac{80021}{19} - \left(\frac{1043}{19}\right)^2\right)$$

$$\Rightarrow \frac{19}{18}\left(4211.6316 - 3013.4321\right)$$

$$\Rightarrow \frac{19}{18} \times 1198.1995$$

$$\Rightarrow 1264.766 \approx \boxed{1264.77} \text{ min}^2$$

Ans

Standard deviation $(s) \Rightarrow \sqrt{s^2}$

$$= \sqrt{1264.77} = 35.56 \text{ min}$$

Ans.

(b) If the observations are re-expressed in hours, then all the data points would be divided by 60.

An easier method will be using the fact, when a constant is multiplied by each data point, then variance is multiplied by the square of that constant.

In this case, the constant is $\Rightarrow \underset{(c)}{\left(\frac{1}{60}\right)}$

$\Rightarrow \boxed{s^2_{new} = c^2 s^2}$ $\rightarrow$ Using this,

$\Rightarrow$

$$S^2_{new} = \left(\frac{1}{60}\right)^2 \times 1264.77$$

$$S^2_{new} = \frac{1264.77}{3600} = \boxed{0.35 \text{ hr}^2}$$

Ans.

&

New standard deviation will be $\Rightarrow$

$$\sqrt{S^2_{new}} = \sqrt{0.35} = 0.59 \text{ hr}$$

Ans.

(7.) Frequency Distribution data for discoveries ⇒

Sample size (n) = 100

Number of discoveries	Frequency
0	9
1	12
2	26
3	20
4	12
5	7
6	6
7	4
8	1
9	1
10	1
12	1
Total freq. ⇒	100

The data is inherently ordered in the table above.

∴ Median ⇒

As the number of data points is even, we take the mid element and the next to mid element & divide their sum by 2.

$$\Rightarrow \frac{\left(\frac{n}{2}\right)\text{th term} + \left(\frac{n}{2}+1\right)\text{term}}{2}$$

∵ n = 100

$$\Rightarrow \frac{50\text{th term} + 51\text{st term}}{2}$$

$$\Rightarrow (3+3)/2 = \underline{3}$$

for determining the lower and upper quartiles of the data,

We divide the data in 2 halves.

[0 (9 times), 1 (12 times), 2 (26 times), 3 (3 times)]

50 data points

↓

(1st half of the list)

[3 (17 times), 4 (12 times), 5 (7 times), 6 (6 times), 7 (4 times), 8 (1 time), 9 (1 time), 10 (1 time), 12 (1 time)]

50 data points

(2nd half of the list)

∴ Lower quartile (Q_1) ⇒ Take middle element of the 1st half of the list.

As there are even number of elements, we take the middle and next to middle element (both are middle elements)

$$\Rightarrow \frac{2+2}{2} = \frac{4}{2} = \underline{\underline{2}}.$$

∴ Upper quartile (Q_3) ⇒ Take middle element of the 2nd half of the list.

⇒ Mean of the 2 middle elements there as number of elements is even.

$$\Rightarrow \frac{4+4}{2} = \underline{\underline{4}}.$$

∴ Inter-quartile Range (IQR) ⇒

$$= Q_3 - Q_1$$

$$= 4 - 2 = \underline{\underline{2}}$$

$\therefore$ LOB $\Rightarrow$

$$Q_1 - 1.5IQR$$

$= 2 - 1.5 \times 2 = 2 - 3 = \underline{\underline{-1}}$.

$\therefore$ UOB $\Rightarrow$

$$Q_3 + 1.5IQR$$

$= 4 + 1.5 \times 2 = 4 + 3 = \underline{\underline{7}}$.

When we look at the data, the outliers would be with values such that,

value < LOB OR value > UOB

There are no values with value < LOB.!

for outliers on the upper side (value > UOB)

$\Rightarrow$ 8, 9, 10, 12.

$\therefore$ Outliers $\Rightarrow$ 8, 9, 10, 12

Smallest data point that is not an outlier

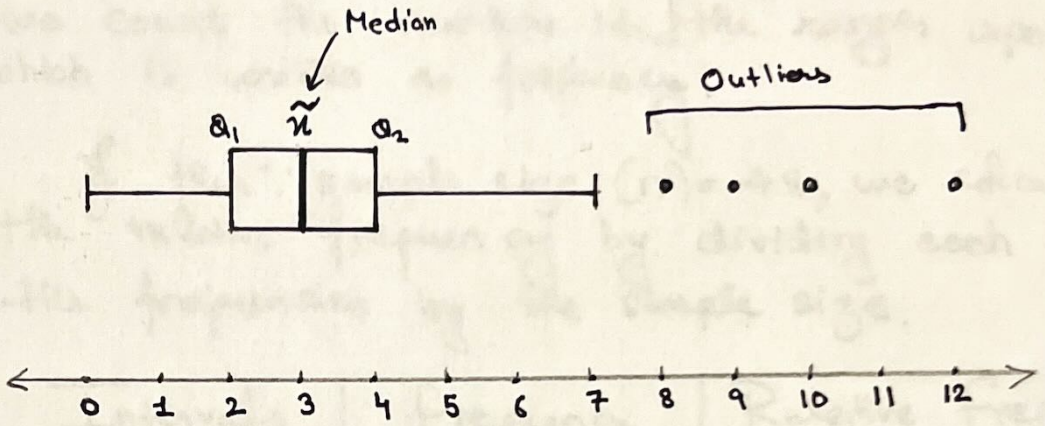
$\Rightarrow$ 0 (as there were no outliers on the lower end & this is the minimum value in the sample).

Largest data point that is not an outlier

$\Rightarrow$ 7 (As the outliers on the upper end are 8, 9, 10, 12 and the largest value other than those in our sample is 7.).

Boxplot =>

(Using median, Q_1, Q_3, LOB, UOB and outliers, we draw the boxplot below) =>



(8.)

Given data,

Median = $\tilde{x}$ = 3.71

Q_1 = 2.99

Q_3 = 5.21.

(a) Based on the information given, we count the numbers in the ranges specified which is written as frequency.

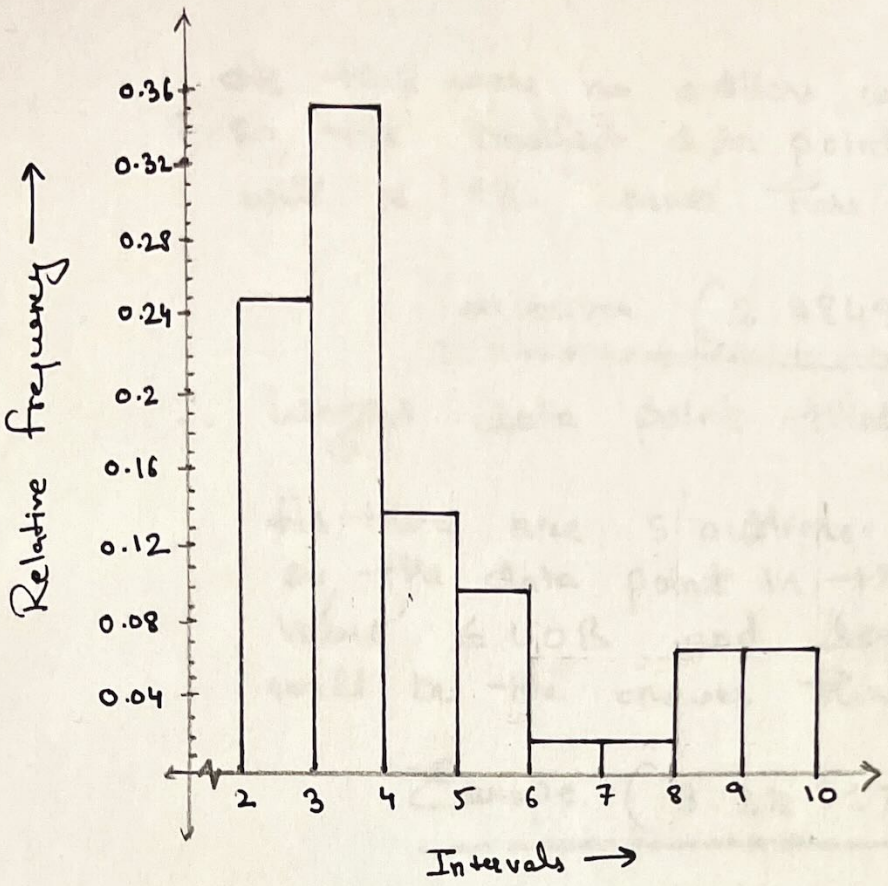
& then ∴ sample size (n) = 48, we calculate the relative frequency by dividing each of the frequencies by the sample size.

Intervals	frequency	Relative frequency
[2, 3]	12	12/48 = 0.25
[3, 4]	17	17/48 = 0.35
[4, 5]	6	6/48 = 0.125
[5, 6]	5	5/48 = 0.1
[6, 7]	1	1/48 = 0.02
[7, 8]	1	1/48 = 0.02
[8, 9]	3	3/48 = 0.0625
[9, 10]	3	3/48 = 0.0625
	48	

Using the data above,

we will draw a histogram with relative frequency vs. intervals.

Histogram =>



(b) $\because Q_1 = 2.99$ & $Q_3 = 5.21$

$\therefore$ IQR $\Rightarrow Q_3 - Q_1 = 5.21 - 2.99 = \underline{2.22}$

$\therefore$ LOB $\Rightarrow Q_1 - 1.5 \times IQR$

$= 2.99 - 1.5 \times 2.22$

$= \underline{-0.34}$

$\therefore$ UOB $\Rightarrow Q_3 + 1.5 \times IQR$

$= 5.21 + 1.5 \times 2.22$

$= \underline{8.54}$

$\therefore$ Outliers (with value < LOB or value > UOB) $\Rightarrow$ Africa (9.350624), Antarctica (8.612503), North America (9.147401), South America (8.823942), Asia (9.740262).

∴ Smallest data point that is not an outlier ⇒

As there were no outliers with value < LOB, so, the smallest data point of the full sample will be the answer here which is ⇒

Vancouver (2.484907).

∴ Largest data point that is not an outlier ⇒

As there are 5 outliers with value > UOB, so, the data point in the sample with value ≤ UOB and largest in the sample will be the answer here which is ⇒

Europe (8.228177)

Ans

(c) Boxplot ⇒