

Assignment 01 (Solution)

Due Date: September 29th, 2023











Total marks: 80

For all of the questions, you **MUST** show all of the steps to earn full marks. Questions with the correct answer but without a proof are awarded a grade of zero. Don't skip the steps in the middle of the calculations!

Question 1 (10 marks)

Given the following table, identify the individuals, the data in the table, characterize the types of variables as well as the level of measurements (*i.e.*, nominal, ordinal, interval or ratio). Please explain your reasoning thoroughly.

Table 1.1: General summary of some countries

| Country | Population size ranking | Internet country code | Official language(s) | Land area (million km ²) | Median age (years) |
|---|-------------------------|-----------------------|----------------------------|--------------------------------------|--------------------|
|  Azerbaijan | 90th | .az | Azeri | 0.082629 | 32.6 |
|  Canada | 38th | .ca | English & French | 9.093507 | 41.8 |
|  Eritrea | 111th | .er | Tigrinya, Arabic & English | 0.101 | 20.3 |
|  Honduras | 95th | .hn | Spanish | 0.11189 | 24.4 |
|  Jamaica | 140th | .jm | English & Jamaican patois | 0.010831 | 29.4 |
|  Latvia | 153rd | .lv | Latvian | 0.062249 | 44.4 |
|  Montenegro | 172nd | .me | Montenegrin | 0.013452 | 39.6 |
|  Saint Lucia | 186th | .lc | English | 0.000606 | 36.9 |
|  Tonga | 191st | .to | Tongan | 0.000717 | 24.1 |
|  Zimbabwe | 73th | .zw | Shona | 0.386847 | 20.5 |

Source: [CIA The World Factbook \(2023\)](#)

Marking Scheme

- One mark for mentioning which is a qualitative variable (the population size ranking can be either or depending on how it is seen, as long as there is a valid explanation, it is good)
- One mark for mentioning which is a quantitative variable
- One mark for the correct sequence (0.5 if the sequence is correct by its first term isn't correct)
- One mark for correctly identifying the discrete variables (note that the population size ranking could be absent if they decided to treat it as a qualitative variable, which is fine)
- One mark for correctly identifying the continuous variables
- Five marks for correctly associating the level of measurement with the five variables (one for each variable)
- Award part marks whenever you deem appropriate

Solution

- The *individuals* are the countries because they are the objects being studied which contain the different characteristics (population size ranking, internet country code, official language(s), land area and median age) we are interested in
- The variable population size ranking can be considered qualitative (similar to letter grades A, B, C and D) or quantitative because the numbers refer to some order and is numeric. Since there is 'st', 'nd', 'rd' and 'th' ordinals attached, it is not completely numeric, either is correct (as long as an explanation has been given)
- The *qualitative* variables are internet country code and official language(s) because there is no order in which language is better and we cannot perform mathematical operations on languages. For instance, what does English + French give? Undefined behaviour. The same argument is applied to the internet country code label
- The *quantitative* variables are population size ranking, land area and median age
 - Discrete variables are population size ranking and land area. Land area is in millions, so, multiply the value in the cell by 1,000,000 to get a whole number to get Azerbaijan: 82,629, Canada: 9,093,507, Eritrea: 101,000, Honduras: 111,890, Jamaica: 10,831, Latvia: 62,249, Montenegro: 13,452, Saint Lucia: 606, Tonga: 717, and Zimbabwe: 386,847, all in km²
 - Continuous variable is median age
- Data is anything not in the first row nor first column
- The types of the variables are:
 - Population size ranking: ordinal level because order matters and 1st is 'higher' than 2nd
 - Internet country code: nominal level as there is no order and mathematical operations cannot be applied
 - Official language(s): nominal level as there is no order and mathematical operations cannot be applied
 - Land area (million km²): Ratio level because zero means there is no land area and we can perform +, −, × and ÷ on the values to get a meaningful result
 - Median age (years): Ratio level because zero means the median is zero years and we can perform +, −, × and ÷ on the values to get a meaningful result

Question 2 (10 marks)

Given the data in [Table 1.2](#), provide an experiment and an approach to choose a sample through simple random sampling, stratified sampling, systematic sampling and cluster sampling. There needs to be four distinct experiments (*i.e.*, you need to design the experiment and state its objective) and each experiments uses a distinct sampling method. Ensure to specify the individuals included in the sample by writing their ID number, the sample size and explain your approach thoroughly.

Table 1.2: The data required to perform sampling analysis on

| ID | Name | Gender | Age | Education | Experience (years) | Salary | Number of children |
|----|-----------------|--------|-----|-----------|--------------------|---------|--------------------|
| 01 | Amanda Thomas | Female | 27 | Bachelor | 11 | 94,631 | 0 |
| 02 | Luke Smith | Male | 30 | Bachelor | 4 | 42,768 | 4 |
| 03 | Lucia Moore | Female | 28 | Bachelor | 14 | 166,470 | 0 |
| 04 | Kelsey Stewart | Female | 20 | Doctoral | 7 | 62,866 | 0 |
| 05 | Edwin Reed | Male | 25 | Bachelor | 11 | 148,221 | 0 |
| 06 | Olivia Perry | Female | 24 | Master | 6 | 89,519 | 2 |
| 07 | Evelyn Douglas | Female | 29 | Master | 4 | 117,232 | 0 |
| 08 | Arianna Farrell | Female | 26 | Doctoral | 0 | 154,352 | 0 |
| 09 | Paul Payne | Male | 23 | Master | 0 | 45,872 | 5 |
| 10 | Ashton Taylor | Male | 29 | Bachelor | 8 | 92,041 | 2 |
| 11 | Lydia Parker | Female | 21 | Primary | 10 | 80,423 | 5 |
| 12 | Lilianna Jones | Female | 27 | Doctoral | 5 | 143,957 | 3 |
| 13 | William Perry | Male | 19 | Doctoral | 3 | 108,867 | 5 |
| 14 | Vincent Grant | Male | 21 | Master | 0 | 45,785 | 2 |
| 15 | Lyndon Foster | Male | 19 | Bachelor | 4 | 57,247 | 3 |

Marking Scheme

- For simple random sampling and systematic sampling, each is worth two marks. One mark for a valid experiment and another mark for correctly creating the sample
- Both stratified sampling and cluster sampling are worth three marks each. They need to divide the individuals correctly (stratified based on some label and cluster randomly). So, one mark for valid experiment, another for correctly dividing the data and a third for correctly selecting the sample
- For each experiment, they need to explicitly write the value of n and IDs selected (take off 0.5 if they don't)

Solution

There are multiple experiments to choose from. Typical ones including finding the average/median age, experience, salary and number of children, or something that is more complex, such as finding if there is a difference between genders when it comes to wages:

- Simple random sampling: we can conduct the experiment of finding the average age by selecting five individuals where each individual has the same probability of being selected and each sample has the same probability of being selected. For example, the sample of size $n = 5$ contains IDs 01, 04, 05, 10 and 12 has been selected.
- Systematic sampling: we can calculate the median salary where we select every k th person. Suppose $k = 4$, then we can select the following individuals in our sample with $n = 3$: 04, 08 and 12.
- Stratified sampling: requires us to divide our data into categories to perform the experiment. Suppose we want to see if there is a difference between the salary between males and females. We can divide our data so that stratum A and stratum B contains all males and females, respectively:

| Stratum | IDs | Selected IDs |
|-----------|--------------------------------|--------------|
| Stratum A | 02, 05, 09, 10, 13, 14, 15 | 02, 10, 13 |
| Stratum B | 01, 03, 04, 06, 07, 08, 11, 12 | 06, 08, 11 |

Now, suppose that our sample size is $n = 3$, then we randomly select three males from stratum A and three females from stratum B. For instance, suppose we select the males with IDs 02, 10 and 13, and the females with IDs 06, 08 and 11.

- Cluster sampling: suppose we want to find the average number of children. We can arbitrarily divide our individuals into, say five clusters (each cluster has three individuals), or, based on some criteria, such as the level of education as it is there by default, and select two of the clusters to obtain a sample of size $n = 6$. Here is an arbitrary division of the data:

| Cluster | IDs | Selected? |
|-----------|------------|-----------|
| Cluster A | 01, 09, 11 | ✓ |
| Cluster B | 02, 08, 13 | |
| Cluster C | 07, 10, 14 | ✓ |
| Cluster D | 04, 05, 06 | |
| Cluster E | 03, 12, 15 | |

Hence, the sample of size $n = 6$ that has been selected contains IDs: 01, 07, 09, 10, 11 and 14

Question 3 (10 marks)

Use the data in [Table 1.3](#) to generate a table of the frequency and relative frequency by

hand, as well as draw the histogram (also by hand). Ensure the classes themselves using the interval notation. Choose 6 as the initial class width.

| | | | | |
|--------|--------|--------|--------|--------|
| 10.084 | 11.130 | 12.552 | 19.628 | 20.010 |
| 31.522 | 44.216 | 47.344 | 49.036 | 50.913 |
| 51.592 | 51.664 | 56.654 | 59.312 | 62.817 |
| 66.186 | 79.142 | 79.654 | 79.711 | 81.328 |
| 86.964 | 92.242 | 93.448 | 97.866 | 98.727 |

Table 1.3: Question 3 data

Marking Scheme

- Five marks for correct (relative) frequency and class division (award as you deem fit)
- Five marks for the correct histogram. A mark for the main title, another for x/y -axes titles, two marks for the interval notation in x -axis and one last mark for y -axis values.

Solution

There are two solutions that are equivalent. The first is using the value six as the number of classes or, as mentioned, the class width. Solution one:

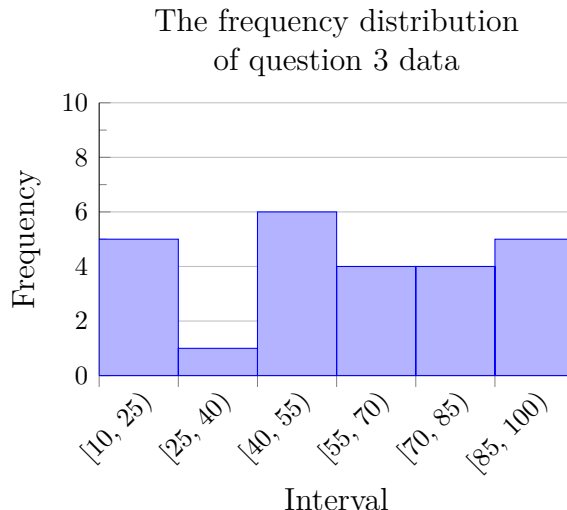
$$\begin{aligned}
 \text{Class width} &\approx \frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}} \\
 &\approx \frac{98.727 - 10.084}{6} \\
 &\approx 14.77383 \\
 &= 15
 \end{aligned}$$

The solution is to start from 10 increment by 15 to obtain the six classes. It wouldn't be 'wrong' to start at, say 5 or 6 or even 3. Starting at 10 seems to be the appropriate choice as it is the closest whole number to the smallest value in the data.

| | | | | |
|--------|--------|--------|--------|--------|
| 10.084 | 11.130 | 12.552 | 19.628 | 20.010 |
| 31.522 | 44.216 | 47.344 | 49.036 | 50.913 |
| 51.592 | 51.664 | 56.654 | 59.312 | 62.817 |
| 66.186 | 79.142 | 79.654 | 79.711 | 81.328 |
| 86.964 | 92.242 | 93.448 | 97.866 | 98.727 |

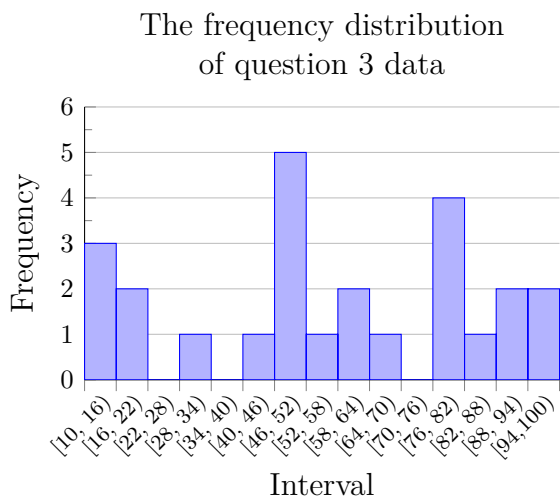
| Classes | Frequency | Relative frequency |
|-----------|-----------|-----------------------|
| [10, 25) | 5 | $\frac{5}{25} = 0.20$ |
| [25, 40) | 1 | $\frac{1}{25} = 0.04$ |
| [40, 55) | 6 | $\frac{6}{25} = 0.24$ |
| [55, 70) | 4 | $\frac{4}{25} = 0.16$ |
| [70, 85) | 4 | $\frac{4}{25} = 0.16$ |
| [85, 100) | 5 | $\frac{5}{25} = 0.20$ |

The total frequency is 25 (which is $n = 25$) and total relative frequency is 1 (or 100%)



Solution two (six is the class width):


| Classes | Frequency | Relative frequency |
|--|-----------|-----------------------|
| [10, 16) | 3 | $\frac{3}{25} = 0.12$ |
| [16, 22) | 2 | $\frac{2}{25} = 0.08$ |
| [22, 28) | 0 | $\frac{0}{25} = 0.00$ |
| [28, 34) | 1 | $\frac{1}{25} = 0.04$ |
| [34, 40) | 0 | $\frac{0}{25} = 0.00$ |
| [40, 46) | 1 | $\frac{1}{25} = 0.04$ |
| [46, 52) | 5 | $\frac{5}{25} = 0.20$ |
| [52, 58) | 1 | $\frac{1}{25} = 0.04$ |
| [58, 64) | 2 | $\frac{2}{25} = 0.08$ |
| [64, 70) | 1 | $\frac{1}{25} = 0.04$ |
| [70, 76) | 0 | $\frac{0}{25} = 0.00$ |
| [76, 82) | 4 | $\frac{4}{25} = 0.16$ |
| [82, 88) | 1 | $\frac{1}{25} = 0.04$ |
| [88, 94) | 2 | $\frac{2}{25} = 0.08$ |
| [94, 100) | 2 | $\frac{2}{25} = 0.08$ |
| The total frequency is 25 (which is $n = 25$) and total relative frequency is 1 (or 100%) | | |



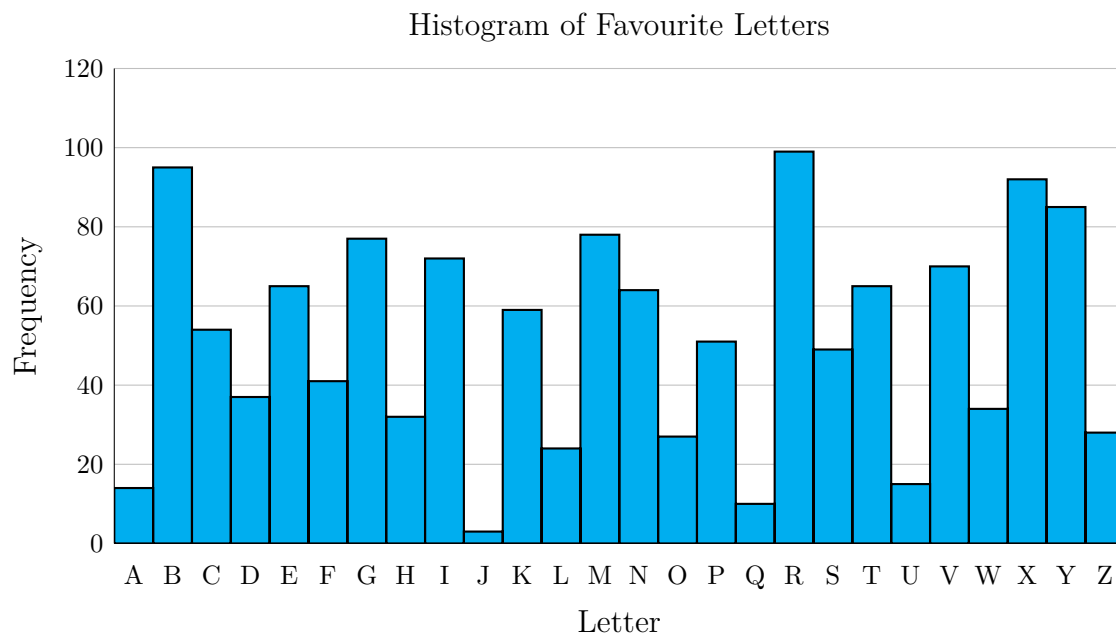
Question 4 Excel (5 marks)

A fictitious survey has been conducted to see which letter people like the most. The data is found in `Letter-Exercise.xlsx` which contains a single worksheet 'Letter Exercise' on Brightspace. Use Excel to generate a histogram of the data found. Ensure to have the 'bins' option to rely on category. Also, include the required titles and change the default colour. Take a screenshot¹ and attach it as a part of your submission. Ensure to use Google Chrome or Firefox to access [Excel online](#) (don't use Safari, it has issues)

Marking Scheme

- All Excel exercises (in this assignment and upcoming ones) have been shown in the lecture
- Three marks, and one mark for each: main title and x -axis title and y -axis title (they are not there by default)
- Two marks for having the correct number of bins/classes (probably all or nothing as Excel decides to show only three bins by default)
- Please give them a reminder that they have to change the colour in case the histogram has the following colour:  (which is not used in this document)

Solution



Question 5 (5 marks)

Suppose we currently have the following grades available from the course we are enrolled in. Find the weighted average of the entries.

¹Please don't use your camera phone, instead, take an actual screenshot

| Entry | Weight (%) | Grade |
|--------------|------------|------------------|
| Assignment 1 | 10 | $\frac{29}{40}$ |
| Assignment 2 | 15 | $\frac{36}{50}$ |
| Assignment 3 | 05 | $\frac{15}{23}$ |
| Assignment 4 | 05 | $\frac{8}{19}$ |
| Midterm 1 | 15 | $\frac{60}{80}$ |
| Exam 1 | 30 | $\frac{70}{120}$ |

Marking Scheme

- Five marks for getting the correct answer
- Four for getting something close to the correct answer
- 2.5 for correct numerator but wrong denominator
- Allocate as you deem fit for two marks or less

Solution

Note that

$$\frac{29}{40} = 0.725, \quad \frac{36}{50} = 0.72, \quad \frac{15}{23} = 0.65217, \quad \frac{8}{19} = 0.42105, \quad \frac{60}{80} = 0.75, \quad \frac{70}{120} = 0.58333$$

$$\begin{aligned}
 \bar{x}_w &= \frac{\sum_{i=1}^n (w_i \cdot x_i)}{\sum_{i=1}^n w_i} = \frac{(w_1 \cdot x_1) + (w_2 \cdot x_2) + (w_3 \cdot x_3) + \cdots + (w_n \cdot x_n)}{w_1 + w_2 + w_3 + \cdots + w_n} \\
 &= \frac{(0.10 \cdot 0.725) + (0.15 \cdot 0.72) + (0.05 \cdot 0.65217) + (0.05 \cdot 0.42105) + (0.15 \cdot 0.75) + (0.30 \cdot 0.58333)}{0.10 + 0.15 + 0.05 + 0.05 + 0.15 + 0.30} \quad (1) \\
 &= \frac{0.0725 + 0.108 + 0.03261 + 0.02105 + 0.1125 + 0.175}{0.8} \\
 &= \frac{0.52166}{0.8} \\
 &= 0.652075 \text{ or } 65\%.
 \end{aligned}$$

Question 6 (10 marks)

For each of the following samples, find the mean, median and mode:

Sample A: 72 64 57 25 59 80 80

Sample B: 79 50 68 21 66 67 23 29

Sample C: 42 68 76 42 24 49 24 68 37

Marking Scheme

- One mark for each of the mean, median and mode (there are three of each) and the mode of Sample B has two marks. They have to differentiate between none and zero.

Solution

It would be easier to sort the values first and then find the required metrics and also highlighting the **median** and the **modes** or **purple** if both:

Sample A: 25 57 59 **64** 72 **80** **80**

Sample B: 21 23 29 **50** **66** 67 68 79

Sample C: **24** **24** 37 **42** **42** 49 **68** **68** 76

The mean:

$$\text{Sample A: } \frac{25 + 57 + 59 + 64 + 72 + 80 + 80}{7} = 62.42857$$

$$\text{Sample B: } \frac{21 + 23 + 29 + 50 + 66 + 67 + 68 + 79}{8} = 50.375$$

$$\text{Sample C: } \frac{24 + 24 + 37 + 42 + 42 + 49 + 68 + 68 + 76}{9} = 47.77778$$

The median:

Sample A: 64

$$\text{Sample B: } \frac{50 + 66}{2} = 58$$

Sample C: 42

The mode:

Sample A: 80, which is present twice

Sample B: None, the mode doesn't exist (it is not zero!)

Sample C: 24, 42 and 68, which were present twice

To conclude:

| Sample | Mean \bar{x} | Median (M) | Mode |
|----------|----------------|----------------|---------------|
| Sample A | 62.42857 | 64 | 80 |
| Sample B | 50.375 | 58 | None |
| Sample C | 47.77778 | 42 | 24, 42 and 68 |

Question 7 (10 marks)

Find the quartile ranges (*i.e.*, Q_1 , Q_2 , Q_3 and Q_4) for each of the three samples given in question 6. Also, conclude whether there are minor outliers².

Marking Scheme

- Note that the answers might be a little different as there are multiple ways of finding the quartiles
- One grade for each of the Q_1 , Q_3 and Q_4 and the last mark is ensuring that the Q_2 is also valid and the method of including/excluding the median has been explicitly specified

Solution

We will use the exclusion of median approach when finding the quartiles. It would be easier to sort the values first; and then find the required metrics and highlight Q_1 with red, $M = Q_2$ with purple, Q_3 with blue and Q_4 with green (Section 1.13 in Component 01 goes through this extensively):

Sample A: 25 57 59 64 72 80 80

Sample B: 21 23 29 50 66 67 68 79

Sample C: 24 24 37 42 42 49 68 68 76

The conclusion is:

| Sample | Q_1 | Q_2 | Q_3 | Q_4 |
|----------|--------------------------|------------------------|--------------------------|-------|
| Sample A | 57 | 64 | 80 | 80 |
| Sample B | $\frac{23+29}{2} = 26$ | $\frac{50+66}{2} = 58$ | $\frac{67+68}{2} = 67.5$ | 79 |
| Sample C | $\frac{24+37}{2} = 30.5$ | 42 | $\frac{68+68}{2} = 68$ | 76 |

The interquartile range formula is $IQR = Q_3 - Q_1$ and the minor outlier range is anything less than $Q_1 - (1.5 \cdot IQR)$ or larger than $Q_3 + (1.5 \cdot IQR)$:

| Sample | Q_1 | Q_3 | IQR | $Q_1 - (1.5 \cdot IQR)$ | $Q_3 + (1.5 \cdot IQR)$ |
|----------|-------|-------|--------------------|------------------------------------|------------------------------------|
| Sample A | 57 | 80 | $80 - 57 = 23$ | $57 - (1.5 \cdot 23) = 22.5$ | $80 + (1.5 \cdot 23) = 114.5$ |
| Sample B | 26 | 67.5 | $67.5 - 26 = 41.5$ | $26 - (1.5 \cdot 41.5) = -36.25$ | $67.5 + (1.5 \cdot 41.5) = 129.75$ |
| Sample C | 30.5 | 68 | $68 - 30.5 = 37.5$ | $30.5 - (1.5 \cdot 37.5) = -25.75$ | $68 + (1.5 \cdot 37.5) = 124.25$ |

There are no minor outliers because all data in Sample A, B and C are within the range $(22.5, 114.5)$, $(-36.25, 129.75)$, $(-25.75, 124.25)$, respectively.

Question 8 (10 marks)

Find the standard deviation and variance of the three samples given in question 6.

²This question doesn't ask about major outliers, just focus on minor outliers

Marking Scheme

- Three marks for each sample and the last is for correct variance
- They have to show all of the steps here! In case they refer to their work and say the calculations yield both the variance and standard deviation, then this is okay. They don't have to rewrite the entire solution twice to find the variance and standard deviation.
- In case only the results are given, they receive a grade of zero

Solution

Let us find the sample mean of the three samples individually:

$$\begin{aligned}
 a : \text{ Sample A} &= \frac{25 + 57 + 59 + 64 + 72 + 80 + 80}{7} = 62.42857 \\
 b : \text{ Sample B} &= \frac{21 + 23 + 29 + 50 + 66 + 67 + 68 + 79}{8} = 50.375 \\
 c : \text{ Sample C} &= \frac{24 + 24 + 37 + 42 + 42 + 49 + 68 + 68 + 76}{9} = 47.77778
 \end{aligned} \quad (2)$$

$$\begin{aligned}
 s_a &= \sqrt{\frac{(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2 + (\bar{x}_3 - \bar{x})^2 + (\bar{x}_4 - \bar{x})^2 + (\bar{x}_5 - \bar{x})^2 + (\bar{x}_6 - \bar{x})^2 + (\bar{x}_7 - \bar{x})^2}{n - 1}} \\
 s_a &= \sqrt{\frac{(25 - 62.42857)^2 + (57 - 62.42857)^2 + (59 - 62.42857)^2 + (64 - 62.42857)^2 + (72 - 62.42857)^2 + (80 - 62.42857)^2 + (80 - 62.42857)^2}{7 - 1}} \\
 s_a &= \sqrt{\frac{1400.89785 + 29.46937 + 11.75509 + 2.46939 + 91.61227 + 308.75515 + 308.75515}{6}} \quad (3) \\
 s_a &= \sqrt{\frac{2153.71429}{6}} \\
 s_a &= \sqrt{358.95238} \\
 s_a &= 18.94604
 \end{aligned}$$

The sample standard deviation of sample A is $s_a = 18.94604$, which makes the variance

to be $s_a^2 = 18.94604^2 = 358.95243$. Now, finding Sample B's standard deviation:

$$\begin{aligned}
 s_b &= \sqrt{\frac{(\mathbf{x}_1 - \bar{x})^2 + (\mathbf{x}_2 - \bar{x})^2 + (\mathbf{x}_3 - \bar{x})^2 + (\mathbf{x}_4 - \bar{x})^2 + (\mathbf{x}_5 - \bar{x})^2 + (\mathbf{x}_6 - \bar{x})^2 + (\mathbf{x}_7 - \bar{x})^2 + (\mathbf{x}_8 - \bar{x})^2}{\mathbf{n} - 1}} \\
 s_b &= \sqrt{\frac{(\mathbf{21} - 50.375)^2 + (\mathbf{23} - 50.375)^2 + (\mathbf{29} - 50.375)^2 + (\mathbf{50} - 50.375)^2 + (\mathbf{66} - 50.375)^2 + (\mathbf{67} - 50.375)^2 + (\mathbf{68} - 50.375)^2 + (\mathbf{79} - 50.375)^2}{\mathbf{8} - 1}} \\
 s_b &= \sqrt{\frac{862.89062 + 749.39062 + 456.89062 + 0.14062 + 244.14062 + 276.39062 + 310.64062 + 819.39062}{7}} \quad (4) \\
 s_b &= \sqrt{\frac{3719.875}{7}} \\
 s_b &= \sqrt{531.41071} \\
 s_b &= 23.05235
 \end{aligned}$$

The sample standard deviation of sample B is $s_b = 23.05235$, which makes the variance to be $s_b^2 = 23.05235^2 = 531.41084$. Now, finding Sample C's standard deviation:

$$\begin{aligned}
 s_c &= \sqrt{\frac{(\mathbf{x}_1 - \bar{x})^2 + (\mathbf{x}_2 - \bar{x})^2 + (\mathbf{x}_3 - \bar{x})^2 + (\mathbf{x}_4 - \bar{x})^2 + (\mathbf{x}_5 - \bar{x})^2 + (\mathbf{x}_6 - \bar{x})^2 + (\mathbf{x}_7 - \bar{x})^2 + (\mathbf{x}_8 - \bar{x})^2 + (\mathbf{x}_9 - \bar{x})^2}{\mathbf{n} - 1}} \\
 s_c &= \sqrt{\frac{(\mathbf{24} - 47.77778)^2 + (\mathbf{24} - 47.77778)^2 + (\mathbf{37} - 47.77778)^2 + (\mathbf{42} - 47.77778)^2 + (\mathbf{42} - 47.77778)^2 + (\mathbf{49} - 47.77778)^2 + (\mathbf{68} - 47.77778)^2 + (\mathbf{68} - 47.77778)^2 + (\mathbf{76} - 47.77778)^2}{\mathbf{9} - 1}} \\
 s_c &= \sqrt{\frac{565.38282 + 565.38282 + 116.16054 + 33.38274 + 33.38274 + 1.49382 + 408.93818 + 408.93818 + 796.4937}{8}} \quad (5) \\
 s_c &= \sqrt{\frac{2929.55556}{8}} \\
 s_c &= \sqrt{366.19444} \\
 s_c &= 19.13621
 \end{aligned}$$

The sample standard deviation of sample C is $s_c = 19.13621$, which makes the variance

to be $s_c^2 = 19.13621^2 = 366.19453$. To conclude:

| Sample | n | Mean (\bar{x}) | Standard deviation (s) | Variance (s^2) |
|----------|-----|--------------------|----------------------------|--------------------|
| Sample A | 7 | 62.42857 | 18.94604 | 358.95238 |
| Sample B | 8 | 50.375 | 23.05235 | 531.41071 |
| Sample C | 9 | 47.77778 | 19.13621 | 366.19444 |

Question 9 (10 marks)

Given the following population data, find the mean, median, mode, standard deviation and variance:

Population A: 61 81 77 80 72 30 39

Marking Scheme

- One grade for each: mean, median, mode and variance
- Four marks for standard deviation by showing all steps
- Award two marks for correct results (they might divide by $n - 1$ instead of N)

Solution

Let us sort the data first to find the median and highlight it:

30 39 61 **72** 77 80 81

The mean is

$$\begin{aligned}
 \mu &= \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} \\
 &= \frac{30 + 39 + 61 + 72 + 77 + 80 + 81}{7} \\
 &= 62.85714
 \end{aligned} \tag{6}$$

The mode is none or doesn't exist as all of the data values are unique. Let us find the

population standard deviation and population variance:

$$\begin{aligned}
 \sigma &= \sqrt{\frac{(\mathbf{x}_1 - \mu)^2 + (\mathbf{x}_2 - \mu)^2 + (\mathbf{x}_3 - \mu)^2 + (\mathbf{x}_4 - \mu)^2 + (\mathbf{x}_5 - \mu)^2 + (\mathbf{x}_6 - \mu)^2 + (\mathbf{x}_7 - \mu)^2}{N}} \\
 \sigma &= \sqrt{\frac{(\mathbf{30} - 62.85714)^2 + (\mathbf{39} - 62.85714)^2 + (\mathbf{61} - 62.85714)^2 + (\mathbf{72} - 62.85714)^2 + (\mathbf{77} - 62.85714)^2 + (\mathbf{80} - 62.85714)^2 + (\mathbf{81} - 62.85714)^2}{7}} \\
 \sigma &= \sqrt{\frac{1079.59165 + 569.16313 + 3.44897 + 83.59189 + 200.02049 + 293.87765 + 329.16337}{7}} \quad (7) \\
 \sigma &= \sqrt{\frac{2558.85714}{7}} \\
 \sigma &= \sqrt{365.55102} \\
 \sigma &= 19.11939
 \end{aligned}$$

The population standard deviation of population C is $\sigma = 19.11939$, which makes the population variance to be $\sigma^2 = 19.11939^2 = 365.55107$. In conclusion:

| Population | n | Mean (μ) | Median (M) | Mode | Standard deviation (σ) | Variance (σ^2) |
|--------------|-----|----------------|----------------|------|---------------------------------|-------------------------|
| Population A | 7 | 62.85714 | 72 | None | 19.11939 | 365.55107 |

Submission:

- Please ensure to submit all of your work through Crowdmark (more on that during the lecture on Tuesday September 26)
- You are encouraged to use pencil and eraser when solving the assignment
- Insert your name, student number and page number (x of total) on the top of each page
- Upload images and name them `q1.zzz`, `q2.zzz`, etc., where `zzz` is the file extension, like `png`, etc. In case a question spans two or more pages (say Q1), name it `Q1P1.zzz`, `Q1P2.zzz`, etc. where P2 stands for page 2.
- For your Excel work, ensure to submit a single `.xlsx` file with the name of the file as your student ID (seven-digit number), for instance `4824931.xlsx`
- The Excel output is submitted as an image as well as the file (both are required)