8.1 Summary Measures

The instructions here are given for **Excel**, but the same commands work in **LibreOffice** and the data sets can also be downloaded in LibreOffice. Ensure you save your answers in the Exercise sheets for your submission.

Example 8.1

Here, we consider the dietary data from Data Set B (see the Data Annexe). We calculate the sample size, sample mean and sample standard deviation of the weight loss for those individuals who undertook Diet A.

1. Open the Excel workbook **Exa8.1B.xlsx** from the Examples folder. This contains the relevant data, together with an added text template.

2. We calculate the sample size for Diet A (the number of non-blank data entries for WtLoss) using the statistical function **COUNT()**. In cell F3, enter the formula **=COUNT(B2:B51).**

3. We calculate the sample mean weight loss for Diet A using the statistical function **AVERAGE()**. In cell F4, enter the formula **=AVERAGE(B2:B51)**.

4. We calculate the sample standard deviation of the weight loss for Diet A using the statistical function **STDEV()**. In cell F5, enter the formula **=STDEV(B2:B51)**

5. Highlight cells F4 and F5 and format them to 3 decimal places.

Note that the range B2:B51 includes the Wtloss data only for those individuals on Diet A.

Thus, the sample size for Diet A is $n = 50$ (50 individuals undertook Diet A)

The sample mean weight loss for Diet A is $\bar{x} = 5.341$. The average weight loss for those individuals who undertook Diet A is 5 341 kg, so the diet appears to have been effective.

The sample standard deviation of the weight loss for Diet A is $s = 2.536$ kg. Since the mean weight loss is a little larger than 2s, then a high proportion of those individuals on Diet A had a positive weight loss, again emphasising the effectiveness of the diet.

Exercise 8.1

Open the Excel workbook in **Exe 8.1B.xlsx** from the Exercises folder. Obtain the sample size, sample mean weight loss and the sample standard deviation of the weight loss for Diet B. Place these results in the block of cells F23 to F25, using the same format as that employed for the Diet A results in the above example.

Briefly interpret your findings. What do these results tell you about the relative effectiveness of the two weight-reducing diets?

- **Diet A** has a mean weight loss of 5.341 with a standard deviation of 2.536 among a sample of 50 individuals.
- **Diet B** displays a lower mean weight loss of 3.710 with a slightly higher standard deviation of 2.741, also based on a sample of 50 individuals.

Interpretation:

1. **Mean Weight Loss:** Diet A shows a higher average weight loss compared to Diet B. On average, individuals following Diet A lost more weight than those following Diet B.
2. **Variability:** Diet B has a slightly higher standard deviation compared to Diet A, indicating that while Diet B's mean weight loss is lower, the individual weight loss outcomes are more spread out or varied compared to Diet A.

In summary, based on these sample statistics, Diet A appears to be relatively more effective for weight reduction compared to Diet B. However, it's essential to consider the broader context, such as long-term effectiveness, potential side effects, and individual variations, before definitively determining the overall superiority of one diet over the other. These statistics only represent the sample and may not capture the entire population's responses to these diets.

Example 8.2

Here, we again consider the dietary data from Data Set B. We calculate the sample median, sample quartiles and sample interquartile range of the weight loss for those individuals who undertook Diet A.

1. Open the Excel workbook **Exa 8.2B.xlsx**. from the Examples folder. This contains the relevant data and previous work, together with an added text template.

2. We calculate the median weight loss for Diet A using the statistical function **MEDIAN()**. In cell F6, enter the formula **=MEDIAN(B2:B51).**

3. We calculate the first sample quartile weight loss for Diet A using the statistical function **QUARTILE()**. In cell F7, enter the formula **=QUARTILE(B2:B51,1)**.

4. We calculate the third sample quartile weight loss for Diet A using the statistical function **QUARTILE()**. In cell F8, enter the formula **=QUARTILE(B2:B51,3)**.

5. We calculate the interquartile range of the weight for Diet A by simply differencing the above two quartiles. In cell F9, enter the formula **= F8–F7**.

6. Highlight cells F6 to F9 and format them to 3 decimal places.

Note that the range B2:B51 includes the Wtloss data only for those individuals on Diet A.

The sample median weight loss for Diet A is $M = 5.642$ kg, so the diet appears to have been effective.

The sample interquartile range of the weight loss for Diet A is IQR = 3.285 kg. A high proportion of those individuals on Diet A had a positive weight loss, again emphasising the effectiveness of the diet.

Exercise 8.2

Open the Excel workbook in **Exe 8.2B.xlsx** from the Exercises folder. Obtain the sample median, first and third quartiles and the sample interquartile range of the weight loss for Diet B. Place these results in the block of cells F26 to F29, using the same format as that employed for the Diet A results in the above example.

Briefly interpret your findings. What do these results tell you about the relative effectiveness of the two weight-reducing diets?

- Interpretation:
  - Median: Diet A has a higher median of 5.642 compared with Diet B whose median is 3.745. This means that the median weight loss outcome for Diet A is bigger than that of Diet B.
  - Quartiles: Compared to diet B, the distribution spread for diet A is seen in its upper and lower quartiles. Quartiles for diet A are mostly higher than those for diet B meaning most people experienced more weight loss in diet A compared to diet B.
  - Interquartile Range (IQR): Diet's A IQR of 3.2845 is marginally less than Diet B's IQR of 3.4505. Weight loss values for diet A concentrate on a more restricted range with narrow spread, although it has high quartiles as opposed to diet B with narrow spread, but wide band.
- Effectiveness of Diets:
  - Compared to diet B, diet A seems to have a higher median and smaller interquartile range which indicates better consistency in weight loss within the central 50% of the sample and overall, more effective results among people.
  - Diet B is represented by a lesser median, but a wider range of outcomes for weight loss. The cases of overweight reduction in different persons are wider in this diet.
- These indications seem to point towards Diet A being more successful for weight loss as it would provide greater levels and consistently higher reductions of weight loss in majority of people.

Example 8.3

Consider the brand preference data of Data Set D (see the Data Annexe).

1. Open the Excel workbook **Exa8.3D.xlsx** from the Examples folder. This contains the relevant data, together with an added text template.

We are interested in seeing if the pattern of preferences for the various brands of breakfast cereal differs between the two demographic areas. However, the data are at an "individual" level, so it's impossible to obtain any meaningful information by simply inspecting this "raw" data.

We now calculate the frequencies and percentage frequencies of the occurrences of the nominal variable Brand for the first demographic area (i.e. for Area = 1).

2. In cell E6, enter the formula **=COUNTIF(B2:B71,"A")**. This counts the number of times that A occurs in the Brand data for Area 1, so gives the frequency of the outcome A for Area 1.

3 In cell E7, enter the formula **=COUNTIF(B2:B71,"B")**. This counts the number of times that B occurs in the Brand data for Area 1, so gives the frequency of the outcome B for Area 1.

4. In cell E8, enter the formula **=COUNTIF(B2:B71,"Other")**. This counts the number of times that Other occurs in the Brand data for Area 1, so gives the frequency of the outcome Other for Area 1.

5. In cell E9, enter the formula **=SUM(E6:E8)**. This just gives the total number of observations for Brand in Area 1. Embolden this cell.

Thus 11 out of 70 respondents in Area 1 preferred Brand A, 17 preferred Brand B, and the remaining 42 preferred some other brand of breakfast cereal. This is far more meaningful than the original listing of the raw data!

We now convert these frequencies to percentage frequencies.

6. In cell E15, enter the formula **=100*E6/E$9**. This expresses the original frequency (11) for Brand A as a percentage of the total number of observations (70).

7. Now copy cell E15 and paste into cells E16:E17. The Brand B and Other frequencies for Area 1 are now also expressed as percentages of the total number of observations for this Area.
8. Copy cell E9 and paste into cell E18. This constitutes a check that the three percentage frequencies indeed add up to 100%!

9. Format cells E15:E17 to one decimal place.

Thus, of the 70 respondents in Area 1, 15.7% preferred Brand A, 24.3% preferred Brand B, and the remaining 60.0% preferred some other brand of breakfast cereal.

Exercise 8.3

Open the Excel workbook in **Exe 8.3D.xlsx** from the Exercises folder. Obtain the frequencies and percentage frequencies of the variable Brand, but this time for the Area

2 respondents, using the same format as that employed for the Area1 results in the above example.

Briefly interpret your findings. What do these results tell you about the patterns of brand preferences for each of the two demographic areas?

- Comparing the brand preferences between Area 1 and Area 2:
  - About 21.11 % of brand A, which is more significant as compared with that in Area 1 (15.7%), are preferred in Area 2.
  - However, Brand B recorded a remarkable 33.33% for its performance in Area 2 as compared to 24.3% recorded among Area 1.
  - Nevertheless, there is a marked difference between 45.56% of “Other” brand consumers in Area 2 as compared with 60% in Area 1.
- This indicates that in Area 2, most people prefer both Brand A and Brand B as compared to those in Area 1. The preference for specific named brands (A&B) increase while preference for “Other” brands decrease in Area 2 in comparison with Area 1 where preference is diversified across different groups.