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?

20	A	6.449						
21	A	9.019						
22	A	-1.715						
23	A	4.718		**Diet B**	**n**	50		
24	A	4.007			**Mean**	3.710		
25	A	7.241			**SD**	2.769		
26	A	2.128						
27	A	6.968						
28	A	4.853						
29	A	0.055						

In this exercise, we can see that the same size is n = 50, which is large enough to do further hypothesis testing to better understand the initial results of the mean and standard deviation. The mean of the sample is 3.710kg, not as high a mean as diet A, which may indicate diet B is less effective. The standard deviation is 2.769kg, which indicates positive weight loss for most subjects, but with more variation than diet A. If the two diets were to be compared by these numbers alone, diet A seems the better choice.

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 $\mathrm{M} = 5.642$ kg, so the diet appears to have been effective.

The sample interquartile range of the weight loss for Diet A is $\mathrm{IQR} = 3.285$ 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?

22	A	-1.715						
23	A	4.718		**Diet B**	**n**	50		
24	A	4.007			**Mean**	3.710		
25	A	7.241			**SD**	2.769		
26	A	2.128			**Median**	3.745		
27	A	6.968			**Q1**	1.953		
28	A	4.853			**Q3**	5.404		
29	A	0.055			**IQR**	3.451		
30	A	2.680						

The median weight loss value for diet B is 3.745, indicating that a high proportion of participants had positive weight loss. The mean and median are very close in value, which indicates a normal distribution of data across the sample. But with an IQR of 3.451 and Q1 of 1.953, diet B appears to be less effective than diet A. In addition, the median value of diet A (5.642) is higher than the Q3 value (5.404) for diet B. With these factors alone, it appears diet A is the more effective of the two.

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?

1	**Area**	**Brand**					
2	1	B					
3	1	Other		**Frequencies**			
4	1	A					
5	1	B			**Area 1**	**Area 2**	
6	1	Other		**A**	11	19	
7	1	A		**B**	17	30	
8	1	Other		**Other**	42	41	
9	1	Other		**Total**	**70**	**90**	
10	1	Other					
11	1	Other					
12	1	B		**Percentages**			
13	1	Other					
14	1	Other			**Area 1**	**Area 2**	
15	1	A		**A**	15.7	21.1	
16	1	A		**B**	24.3	33.3	
17	1	A		**Other**	60.0	45.6	
18	1	B		**Total**	**100**	**100.0**	
19	1	A					
20	1	Other					

Of the 90 participants in Area 2, results shows an increase in preference for both ‘Product A’ (21.1%) and ‘Product B’ (33.3%) and a decrease in preference for ‘Other’ (45.6%) when compared to Area 1. It should be noted that frequency of choice ‘Other’ stayed fairly constant between Area 1 and 2, but the increase in frequency of product A and B may be influenced by a larger sample size. Further comparative measures should be undertaken to be certain, but based on the percentages, it appears that ‘Product B’ is preferred to ‘Product A’, while ‘Other’ is preferred to both products.

Hypothesis Testing Using LibreOffice

The Related Samples T Test

The Two -Tailed Test

Example 8.4

Consider the container design data in Data Set F (see the Data Annexe). Notice that the two variables Con1 and Con 2 indeed measure the same characteristic (the number of items sold), but under two different “conditions” (the two different container designs).

We conduct a two-tailed related samples t test of whether the underlying (population) mean number of items sold differs between the two container designs.

Strictly speaking, before undertaking the test we should calculate the differences

D = Con1 – Con 2

for each observation. A normal plot of these differences (i.e. of the values of the variable D) should then be constructed in order to check whether the data are acceptably near-normally distributed.

We will assume for now that the data are indeed so distributed so that the resulting t test is valid. You might want to construct the normal plot as an additional exercise.

1. Open the Excel workbook **Exa 8.4F.xlsx** from the Examples folder. This contains the relevant data.

2. From the **Data** menu bar tab, select **Statistics** and from the ensuing dialogue box, select **Paired t-test**. A new dialogue box appears.

3. In the **Variable 1 Range** box, enter the cell range where the data for the first variable (Con1) can be found, that is, the range B2:B11. In the **Variable 2 Range** box, enter the cell range where the data for the second variable (Con2) can be found, that is, the range C2:C11.

4. Put the results in cell

The resulting output is presented overleaf.

Not all this output is relevant, so it need not all be discussed.

The obtained related samples t = 2.875 with 9 degrees of freedom.

The associated two-tailed p-value is p = 0.018, so the observed t is significant at the 5% level (two-tailed).

Paired t-test		
Alpha	0.05	
Hypothesized Mean Difference	0	
	Variable 1	Variable 2
Mean	172.600	159.400
Variance	750.267	789.378
Observations	10.000	10.000
Pearson Correlation	0.863	
Observed Mean Difference	13.200	
Variance of the Differences	210.844	
df	9.000	
t Stat	2.875	
P (T<=t) one-tail	0.009	
t Critical one-tail	1.833	
P (T<=t) two-tail	0.018	
t Critical two-tail	2.262	

The sample mean numbers of items sold for Container Designs 1 and 2 were, respectively 172.6 and 159.4. The data therefore constitute significant evidence that the underlying mean number of containers sold was greater for Design 1, by an estimated 172.6-159.4 = 13.2 items per store. The results suggest that Design 1 should be preferred.

Exercise 8.4

Consider the filtration data of Data Set G. Open the Excel workbook **Exe8.4G.xlsx** which contains these data from the Exercises folder.

Assuming the data to be suitably distributed, complete a two-tailed test of whether the population mean impurity differs between the two filtration agents, and interpret your findings.

Paired t-test		
Alpha	0.05	
Hypothesized Mean Difference	0	
	Variable 1	**Variable 2**
Mean	8.25	8.68333333333333
Variance	1.05909090909091	1.07787878787879
Observations	12	12
Pearson Correlation	0.901055811772492	
Observed Mean Difference	-0.433333333333334	
Variance of the Differences	0.211515151515151	
df	11	
t Stat	-3.26393859147807	
P (T<=t) one-tail	0.00377299731515574	
t Critical one-tail	1.79588481870404	
P (T<=t) two-tail	0.00754599463031149	
t Critical two-tail	2.20098516009164	
Difference in means	-0.433333333333334	

In this scenario, the hypotheses for the data could be stated as the following:

H0 = Mean impurities with Agent 1 = Mean impurities with Agent two,

and,

H1 = Mean impurities with Agent 1 =/= Mean impurities with Agent 2.

With a critical value of 2.2009, the following decision rule could be determined:

If the tSTAT > + 2.2009, reject the null hypothesis,
or if the tSTAT < -2.2009, reject the null hypothesis;
otherwise, do not reject the null hypothesis.

The sampe mean of Agent 1 = 8.25, while the same mean of Agent 2 = 8.68. These observations come from a sample size of 12, so not very large, but their variance scores (1.059 and 1.078, respectively) are fairly similar so a normal distribution can be assumed. There is an observed mean difference of -0.434, showing Agent B to have slightly more impurities left in each batch overall. With a tSTAT of -3.264, and a p-Value of .0075 against an alpha of 0.05, statistical significance has been demonstrated. If H0 were not rejected, there would be a .0075 chance of there being a -0.434 difference between Agent 1 and Agent 2; therefore, H0 can be rejected. This does not mean the difference between the two agents is necessarily significant; -0.434 is not a large difference in impurities. But if choosing the best filtration agent, Agent 1 tested less in impurities and can be chosen.

The One-Tailed Test

Example 8.5

Recall that in Example 11.4, we conducted a two-tailed related samples t test of whether the underlying (population) mean number of items sold differs between the two container designs of data Set F.

However, now suppose that Container Design 1 is a new, hopefully more attractive design, whereas Container Design 2 is the design in current use. Presumably, the company will only go to the expense of implementing the new design if it can be shown to lead to higher sales than the current design. Thus, the investigators seek evidence that $\mu_1 > \mu_2$, so wish to test:

$$H_0: \mu_1 \quad \mu_2 \qquad \text{against} \qquad H_1: \mu_1 > \mu_2$$

The relevant t test is conducted exactly as before. However, this time, the results are interpreted a little differently.

We first of all check whether the data are consistent with the one-tailed alternative hypothesis. As before, the sample mean numbers of items sold for Container Designs 1 and 2 were, respectively 172.6 and 159.4, so that the data are indeed consistent with H_1.

As before, the obtained related samples t = 2.875 with 9 degrees of freedom.

The associated one-tailed p-value is p = 0.009, so the observed t is significant at the 1% level (one-tailed).

t-Test: Paired Two Sample for Means		
	Con1	*Con2*
Mean	172.6	159.4
Variance	750.2666667	789.3777778
Observations	10	10
Pearson Correlation	0.863335004	
Hypothesized Mean Difference	0	
df	9	
t Stat	2.874702125	
P(T<=t) one-tail	0.009167817	
t Critical one-tail	1.833112923	
P(T<=t) two-tail	0.018335635	
t Critical two-tail	2.262157158	
Difference in Means	13.2	

The data therefore constitute strong evidence (on a one-tailed test) that the underlying mean number of containers sold was greater for Design 1, by an estimated 172.6 - 159.4 = 13.2 items per store. The results continue to suggest that Design 1 should be preferred.

Although broadly similar conclusions were reached as before, a higher level of significance was obtained with the one-tailed test.

Notice that if we had sought to test the alternative pair of one-tailed hypotheses

$$H_0: \mu_1 \geq \mu_2 \qquad \text{against} \qquad H_1: \mu_1 < \mu_2$$

we would have found the difference in sample means to be consistent with the *null hypothesis* that the population mean sales for Design 2 was no greater than that for Design 1. We would thus have declared the result to be not significant without even bothering to inspect the p-value.

Exercise 8.5

Recall that in Exercise 8.4, a two-tailed test was undertaken of whether the population mean impurity differs between the two filtration agents in Data Set G.

Suppose instead a one-tailed test had been conducted to determine whether Filter Agent 1 was the more effective. What would your conclusions have been?

Paired t-test		
Alpha	0.05	
Hypothesized Mean Difference	0	
	Variable 1	**Variable 2**
Mean	8.25	8.68333333333333
Variance	1.05909090909091	1.07787878787879
Observations	12	12
Pearson Correlation	0.901055811772492	
Observed Mean Difference	-0.433333333333334	
Variance of the Differences	0.211515151515151	
df	11	
t Stat	-3.26393859147807	
P (T<=t) one-tail	0.00377299731515574	
t Critical one-tail	1.79588481870404	
P (T<=t) two-tail	0.00754599463031149	
t Critical two-tail	2.20098516009164	
Difference in means	-0.433333333333334	

As in the two-tailed test, in this one-tail scenario, the hypotheses for the data could be stated as the following:

H0 = Mean impurities with Agent 1 = Mean impurities with Agent two,

and,

H1 = Mean impurities with Agent 1 =/= Mean impurities with Agent 2.

The one-tail critical value, however, is 1.796. Therefore, the following decision rule could be determined:

If the tSTAT > + 1.796, reject the null hypothesis,
or if the tSTAT < -1.796, reject the null hypothesis;
otherwise, do not reject the null hypothesis.

There is still an observed mean difference of -0.434, but with a tSTAT of -3.264 against a critical value of -1.796, there is a stronger demonstration of the statistic's validity. The p-Value indicates that if H0 were not rejected, there would be a 0.0037 chance of a -0.434 mean difference occuring. We can therefore conclude with more certainty that H0 is rejected in this scenario. The question of the relevance of sample mean difference remains as it does in the two-tailed test.

The INDEPENDENT Samples T Test

Example 8.6

Consider again Data Set B, the dietary data. Not unreasonably, we wish to test whether the population mean weight loss differs between the two diets. Since completely separate samples of individuals undertook the two diets (i.e. no-one underwent both diets), the independent samples t test is appropriate here.

We know that such a test (and the F test that precedes it) will yield valid results, as we have already completed normal plots for the weight loss data for each of the two diets, and have found both sets of data to exhibit acceptable near-normality (see Example 3.4 and Exercise 3.4).

1. Open the Excel workbook **Exa 8.6B.xlsx** from the Examples folder. This contains the relevant data, together with some of the previously calculated summary statistics for the weight loss on each diet.

We begin by performing the F test of variances.

2. From the **Data** menu bar tab, select **Statistics** and from the ensuing dialogue box, choose **F-test**. A further dialogue box opens.

3. In the **Variable 1 Range** box, enter the cell range where the Diet A weight losses can be found (B2:B51), and in the **Variable 2 Range** box, enter the cell range where the Diet B weight losses can be found (B52:B101).

4. Put the results in H2

5. Some output appears. Widen columns H to J to render it legible. In cell H14, type: p2, and in cell I14, enter the formula: =2*I11 to obtain the required two-tailed p-value.

The relevant output is as follows:
And reduce the number of decimal places to 3

F-test		
Alpha	0.05	
	Variable 1	Variable 2
Mean	5.341	3.710
Variance	6.429	7.668
Observations	50.000	50
df	49.000	49
F	0.839	
P (F<=f) right-tail	0.730	
F Critical right-tail	1.607	
P (F<=f) left-tail	0.270	
F Critical left-tail	0.622	
P two-tail	0.540	
F Critical two-tail	0.567	1.7622

The sample variances for the two diets are, respectively

$$s_1^2 = 6.429 \text{ and } s_2^2 = 7.668$$

The observed F test statistic is F = 0.839 with 49 and 49 associated degrees of freedom, giving a two tailed p-value of p = 0.5399^{NS}

The observed F ratio is thus *not significant*. The data are consistent with the assumption that the population variances underlying the weight losses under the two diets do not differ, and we therefore proceed to use the *equal variances* form of the independent samples t test.

Since we wish to test if the population mean weight losses differ between the two diets, a two-tailed t test is appropriate here.

6. We will use the formula **=TTEST(data1;data2;mode;type**). Here the first two are self explanatory, **mode** indicates whether it is a 1 tailed test (1) or a two tailed test (2), **type** indicates whether it is a paired t test (1), an equal variances independent t test (2) or and unequal variances independent t test (3). This then returns the p-value for the chosen test.

7. As we have chosen a two tailed test then our formula will read **=TTEST(B2:B51,B52:B101,2,2)**
(we have shown above that we can assume equal variances)

The output is as follows (I have included the one tailed p-value for completeness):

Two-tailed 0.00275154 P-value
One-tailed 0.00137577 P-value

The associated two-tailed p-value is p = 0.0028, so the observed t is significant at the 1% level (two-tailed).

The sample mean weight losses for Diets A and B were, respectively, 5.341 kg and 3.710 kg.

Notice that these findings are consistent with the results of Example 3.1 and Exercise 3.1.

The data therefore constitute strong evidence that the underlying mean weight loss was greater for Diet A, by an estimated 5.314 – 3.710 = 1.631 kg. The results strongly suggest that Diet A is more effective in producing a weight loss.

Exercise 8.6

Consider the bank cardholder data of Data Set C. Open the Excel workbook **Exe8.6C.xlsx** which contains this data from the Exercises folder.

Assuming the data to be suitably distributed, complete an appropriate test of whether the population mean income for males exceeds that of females and interpret your findings. What assumptions underpin the validity of your analysis, and how could you validate them?

F-test			
Alpha	0.05		
	Variable 1	**Variable 2**	
Mean	52.913	44.233	
Variance	233.129	190.176	
Observations	60	60	
df	59	59	
F	1.226		
P (F<=f) right-tail	0.218		
F Critical right-tail	1.540		
P (F<=f) left-tail	0.782		
F Critical left-tail	0.649		
P two-tail	0.436		
F Critical two-tail	0.597	1.674	
Two-Tailed	0.00141947	P-Value	
One-Tailed	0.00070974	P-Value	
Mean difference	8.680	income difference	

In this scenario, the sample mean value for income of men (V1) is 52.913 and the sample mean value of income for women (V2) is 44.233, with a mean difference of 8.680. The variance for V1 is 233.129 and the variance for V2 is 190.176. Thus, an f-test will be performed to validate the distribution of the sample.

The Hypotheses of these samples are as follows:

H0: variance of V1 = variance of V2

H1: Variance of V1 =/= variance of V2

With an f test lower tail of 0.597 and an upper tail of 1.674, the following decision rule can thus be concluded:

If F-test < 0.597, reject H0
or, if F-test is <1.674, reject H0;
otherwise, do not reject H0.

Based on the result of the above F-test, the F-test is 1.226. Thus, the level of variance is *not significant* and the sample can be assumed to have a normal distribution. With a p-Value of 0.0014, there is a 0.0014 chance this variance is not normal. Thus the results of the sample are significant, and there is strong evidence to support the mean difference between men's and women's income levels.