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?

First of all, the sample size of group B is also 50, which makes for a fair comparison with group A. I note that the average weight loss for group B is 3.710kg with a standard deviation of 2.769. Unlike in group A, the fact that the mean is less than 2 times the standard deviation means that comparatively much fewer participants ended up with a positive weight loss. At first sight it appears that diet plan A is better than diet plan B, but further statistical testing is required to strengthen this belief.

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?

The sample median weight loss for Diet B is 3.745 kg, so the diet also appears to have been generally effective. The sample interquartile range of the weight loss for Diet B is 3.451 kg indicating that a good proportion of participants on this diet plan experienced a positive weight loss, supporting the belief that this diet plan is also effective.

Comparatively speaking, Diet A appears to be more effective than Diet B given that it has a much higher median with a smaller IQR indicating that a larger proportion of participants experienced positive weight losses with less variability across participants. Once again, this is a belief that needs to be tested and confirmed statistically.

Of the 90 respondents in Area 2, 21.1% preferred Brand A, 33.3% preferred Brand

B, and the remaining 45.6% preferred a different brand of cereal.

Comparatively speaking, brands A and B were both preferred by more participants in

Area 2 than in Area 1, and relatively fewer participants preferred some other brand of

cereal in Area 2 than in Area 1; 60.0% of participants in Area 1 versus 45.6% of

participants in Area B preferred some other brand of cereal. It is also observed that

in both areas 1 and 2, brand B was more popular than brand A. However, the

difference in the sample size is important to keep in mind; Area 2 had about 28%

more samples than Area 1 in this data set.

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.

The related samples t = -2.841 with 10 degrees of freedom.

The two-tailed p-value is p=0.018 which implies that the observed t quoted above is

significant at the 5% confidence level since it is p < 0.05.

Having obtained a p < 0.05 implies that the difference in means is statistically

significant.

Therefore, there is evidence that the underlying mean impurities present after

filtration (parts per 1000) with Agent 1 were less than those with Agent 2 by an

estimated mean of 0.4 (parts per 1000).

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?

In this case, given that μ_1 and μ_2 are the means of filter Agents 1 and 2 respectively, indicating the mean impurities present after filtration (parts per 1000) with each Agent, it is important to note that a lower mean value is preferable. The test here is therefore to determine whether the mean impurities $\mu_2 > \mu_1$. For this reason, we construct the following hypotheses:

H_0: $\mu_2 \leq \mu_1$ and H_1: $\mu_2 > \mu_1$

A preliminary check of whether the data are consistent with the one-tailed alternative hypothesis reveals that this is, in fact, the case, since the mean impurities present with Agent 2 were 8.7 which is great than those present with Agent 1 of 8.3. So the data are consistent with H_1 above.

The related samples t = -2.841 with 10 degrees of freedom is still the same.

The one-tailed p-value, however, is p=0.009 which means that the observed t quoted above is significant at the 1% confidence level since it is $p < 0.01$.

This is strong evidence that the null hypothesis can be rejected with 99% confidence, and that with the same confidence, the underlying mean amount of impurities (parts per 1000) with Agent 1 are smaller than with Agent 2, by an estimated 0.4 parts per 1000. Agent 1 should be preferred.

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?

I first conducted an F-test of variances on the two data sets to determine whether there is evidence of their variances being different.

The sample variances for the two groups (males and females) obtained, respectively, were $s_1^2 = 233.129$ and $s_2^2 = 190.176$. The F test statistic value determined is $F = 1.229$, and the associated degrees of freedom for males and females were both 59, yielding a two tailed p-value of $p = 0.4365$. This value is p < 0.05 indicating that the F ratio is not significant. This implies that data provide evidence of the assumption that the population variances underlying the incomes across the two groups are not different. We are therefore justified to proceed with a t-test of unrelated samples with equal variances. Since the intent is to determine

whether the population mean income of males exceeds that of females, a one-tailed t test is appropriate.

Assuming that μ_1 and μ_2 are the means of males and females respectively, the hypotheses are as follows:

H₀: $\mu_1 \leq \mu_2$ and H₁: $\mu_1 > \mu_2$

First a check is performed to determine whether the data are consistent with the one-tailed alternative hypothesis above. Accordingly, the sample mean incomes of males and females were 53.1 and 44.4 respectively; the data are in fact consistent with H₁. The difference in means is 8.7.

The obtained related samples t = 3.239 with 116 degrees of freedom.

The associated one-tailed p-value is p = 0.0008. This p < 0.001, so the observed t is significant at the 0.1% level (one-tailed).

This is very strong evidence that the null hypothesis can be rejected with 99.9% confidence, and that with the same confidence, the underlying mean income of males exceeds that of females by an estimated 8.7 (in £'000's).

The following are some assumptions made in the analysis:

- The incomes within each group follow a roughly normal distribution. This can be validated by plotting the data on a histogram and verifying this.

- The samples were obtained randomly and are representative of the populations of males and females. This can be validated by obtaining more information on the data collection strategy employed.

- Each income value is independent of others. This may not be the case given that, for example: the data may have been collected primarily targeting high-income groups, within specific occupational clusters e.g. primarily doctors, regions, etc. The presence of such factors can introduce dependence between samples. This can be validated by obtaining further information on the participants.