Data interpretation

|  |
| --- |
| Exercise 6.1 |
| Open the Excel workbook **Exa 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? |

Given the fact both diets have same sample size of 50, The relative effectiveness of the two diets is that Diet A has a higher mean of weight loss than diet B with the following means respectively A to B, 5.341 kg to 3.710 kg. As for standard deviation comparison between the two diets, diet A has a higher difference between the mean and the standard deviation as s= around 2.536 kg and diet 2 is 2.769 kg, showing a higher significance of effectiveness in diet A than B.

|  |
| --- |
| Exercise 6.2 |
| Open the Excel workbook **Exa 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.  The sample median weight loss for Diet B is M =3.745 kg, so the diet appears to have been effective, not as effective as diet 1 with M= 5.642 kg, however, both have positive effect and values.  The sample interquartile range of the weight loss for Diet B is 3.451 kg suggesting positive weight loss emphasising the effectiveness of the diet. |

|  |
| --- |
| Exercise 6.3 |
| Open the Excel workbook **Exa 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? |

From converting the raw data to frequencies, we can conclude that in area 2, more people prefer brand B than brand A, the remaining 41 had other preferences.

Then we converted the frequencies to percentages as the two areas had different total number of responders. Area A had 70 responders, and area B had 90.

In both demographic areas, more people preferred brand B than brand A, and even more people had other preference. As in area 1, out of 70 responders, 24.3% preferred brand B, and in area 2 out of 90 responders 33.3% preferred brand B of breakfast cereal. As for brand A, it was more preferred in area 2 with 21.1% compared to 15.7% in area 1. Finally, people who preferred other brands of cereal were higher in area 1 of 60% than area 2 of 45.6%.

|  |
| --- |
| **Exercise 7.1** |
| Recall that in the previous unit exercises, a two-tailed test was undertaken whether the population mean impurity differed 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? |

The obtained related samples t = -3.264 with 11 degrees of freedom.

The associated one-tailed p-value is p = 0.004, so the observed t is significant at the 1% level (one-tailed). The data therefore constitute strong evidence (on a one-tailed test) that the underlying mean number of impurities parts per 1000 were greater for Agent2, by an estimation of 8.683 parts per 1000 compared to agent1 which has a mean of 8.250 parts per 1000. An estimate of 8.683 - 8.250 = 0.433 parts per 1000, agent2 by estimation would clean more effectively by 0.433 parts per 1000 more than agent 1 . A higher level of significance was obtained with the one-tailed test.

|  |  |
| --- | --- |
| *Agent1* | *Agent2* |
| 8.25 | 8.683333 |

Assuming null hypothesis H0= mu1 ≥ mu2 against alternative H1: mu1 < mu2.

As the p-value is less than 0.05 we reject the null hypothesis because the data is significant. In other words, agent1 was a more effective agent because agent2 had more impurities.

|  |
| --- |
| **Exercise 7.2** |
| Consider the bank cardholder data of Data Set C. Open the Excel workbook **Exa8.6C.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? |

Given the sample variances for the two incomes are, respectively 𝑠1=233.1289 and 𝑠2=190.1758 The observed F test statistic is F = 1.22586 with 59 and 59 associated degrees of freedom, giving a two tailed p-value of p = 0.436492.

The observed F ratio is thus *not significant*. The data are consistent with the assumption that the population variances underlying the income between male and female do not differ, and we therefore proceed to use the *equal variances*form of the unrelated samples t test.

The obtained independent samples t =3.2679 with 118 degrees of freedom. The associated two-tailed p-value is p =0.001419, so the observed t is significant at the 1% level (two-tailed). The sample mean income for Males and Females were, respectively, 52.913 and 44.233.

The data therefore constitute strong evidence that the underlying mean income was greater for Males, by an estimated 52.913 – 44.233 = 8.68. The results strongly suggest that Males have higher income than females.

|  |
| --- |
| **Exercise 7.3** |
| Consider the filtration data of Data Set G. Open the Excel workbook **Exa8.4G.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. |

Given Null Hypothesis of two-tailed test H0: 𝜇 1 = 𝜇 2 against Alternative Hypothesis H1: 𝜇 1 ≠ 𝜇 2. Interprating the data acording to a two-tailed test to see if there is a significant difference between the two agents. The obtained related samples t = -3.264 with 11 degrees of freedom. The associated two-tailed p-value is p = 0.008, so the observed t is significant at the 5% level (two-tailed). he two-tailed test shows strong evidence of a difference between the means of Agent1 and Agent2. Given the p-value is less than 0.05, we reject the null hypothesis, indicating a significant difference in effectiveness between the two agents. Looking at the two means agent2 and agent1, an estimate of 8.683 - 8.250 = 0.433 parts per 1000. Agent2 by estimation would clean more effectively by 0.433 parts per 1000 more than agent 1.

|  |  |  |
| --- | --- | --- |
| |  | | --- | | Exercise 9.1 | | Open the Excel workbook in **Exa 9.1D.xlsx**from the Exercises folder. This contains the percentage frequencies together with the bar chart just created in the above example. Add a percentage frequency bar chart showing the brand preferences in Area 2, using the same format as that employed for the Area1 results in the above example. Drag your new chart so that it lies alongside that for Area 1. Briefly interpret your findings. What do these results tell you about the patterns of brand preferences for each of the two demographic areas?  In both demographic areas, the majority of people had an other prefrence than the given A and B. Moreover, brand B of breakfast cereals was more preffered than brand A in both areas 1 and 2. The bar chart shows more people preffere brand B’s cereals in area 2 than area1. Also, people in area2 preferred brand A more than people did in area1. However, area1 was the area were people voted for other brand the most. |   Exercise 9.2 |
| Open the Excel workbook in **Exa 9.2E.xlsx**from the Exercises folder. This contains the frequency distributions for Data Set E (see the Data Annexe) to which has been added the corresponding percentage frequency distributions. Complete a percentage frequency clustered column bar chart showing the heather species prevalence in the two different locations. Briefly interpret your findings. |

Data set E represents the prevalence of a certain species of Heather in two different species.

The bar charge shows the distribution pattern of Heather in percentage in two different locations. Location A has the highest abundance. Location B has the highest absence of Heather. Both locations A and B had slightly similar prevalence of sparse.

|  |
| --- |
| Exercise 9.3 |
| Open the Excel workbook in **Exa 9.3B.xlsx**from the Exercises folder. This contains the relative frequency histogram for the Diet A weight loss produced in Example 9.3 together with some of the Diet B weight loss summary statistics. Add a relative frequency histogram of the weight loss for Diet B, where possible using the same classes as those employed for the Diet A results in the above example.  Briefly interpret your histogram. What do these results tell you about the patterns of weight loss for each of the two diets? |

As for Diet B, the histogram shows a unimodal distribution that is positively skewed. As the peak of 3kg was more on the left side, the shift of right to left means there are more people that lost less weight going from 11kg to 3kg.

Diet A is also a unimodal distribution, but with a negative skew, showing more of the population of diet A losing more kilograms than diet B as the mean is around 5 kg here compared to 3 kg in diet B. Comparing between the two histograms, diet A shows more weight lost than diet B, showing more effectiveness.