# Unit 8

# Data Analysis and Visualisation

# Workbook

**Weight loss dataset:**

For this activity, I revisited the weight loss dataset, which contains information on the weight losses achieved by two independent samples of 50 individuals. Each individual followed one of two distinct weight-reducing diets, Diet A or Diet B. The dataset includes two key variables: Diet, which indicates the type of diet undertaken (Diet A or Diet B), and Wtloss, which measures the individual’s weight loss (in kilograms) after a fixed period on the assigned diet.

As part of the analysis, I calculated descriptive statistics to summarize the data for each diet. This included determining the mean (average weight loss), median (middle value of weight loss), and standard deviation (a measure of variability in weight loss). Additionally, I computed the first quartile (Q1) and third quartile (Q3) to understand the distribution of weight loss, as well as the interquartile range (IQR), which represents the spread of the middle 50% of the data. These statistics provided a comprehensive overview of the weight loss outcomes for each diet.

To further analyze the data, I performed a t-test: Paired Two Sample for Means. This test was used to compare the average weight loss between the two diets and determine whether the observed difference was statistically significant. The results of this test will help assess whether one diet is more effective than the other in promoting weight loss.

|  |  |  |
| --- | --- | --- |
| t-Test: Paired Two Sample for Means |  |  |
|  |  |  |
|  | *A* | *B* |
| Mean | 5.3412 | 3.70996 |
| Variance | 6.429281 | 7.667594 |
| Observations | 50 | 50 |
| Pearson Correlation | -0.01002 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 49 |  |
| t Stat | 3.056933 |  |
| P(T<=t) one-tail | 0.001808 |  |
| t Critical one-tail | 1.676551 |  |
| P(T<=t) two-tail | 0.003616 |  |
| t Critical two-tail | 2.009575 |  |
|  |  |  |
| Difference in mean | 1.63124 |  |

I compared the weight loss outcomes for Diet A and Diet B and found that, on average, individuals lost more weight on Diet A. The results showed no correlation between the two diets, which is consistent with the experimental design. To assess the statistical significance of these findings, I conducted a t-test. The p-value for the one-tailed test was 0.0018, which indicates that the observed difference in weight loss between the two diets is statistically significant. This suggests that the null hypothesis—that there is no difference in weight loss between Diet A and Diet B—can be rejected. Similarly, the p-value for the two-tailed test was 0.0036, which accounts for the possibility of the difference being in either direction and further supports the conclusion that the difference is significant.

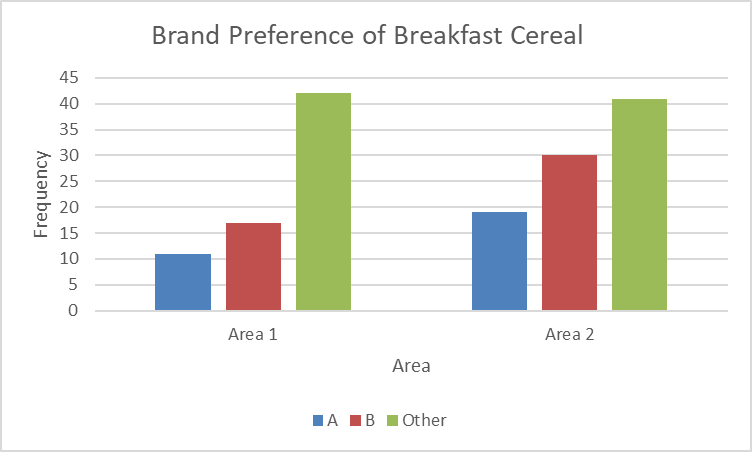
The t-statistic (3.0569) was greater than both the one-tailed and two-tailed critical values, providing strong evidence to reject the null hypothesis. This reinforces the conclusion that Diet A is more effective than Diet B in promoting weight loss. To visualize the results, I created a bar chart comparing the average weight loss for each diet, which clearly illustrates the difference in effectiveness between the two diets.

This chart clearly shows that diet A has more average weight loss than diet B.

**Brand Preference dataset:**

The next activity involved revisiting the brand preference dataset, which was collected as part of a marketing study. Individuals from two different demographic areas were asked to state their preferences for a specific type of breakfast cereal. The dataset focused on two key variables: Area, which indicates the demographic area (Area 1 or Area 2), and Brand, which represents the preferred brand (A, B, or Other). Of particular interest were the preferences for two specific brands, Brand A and Brand B, produced by a particular manufacturer.

In this analysis, I expanded the scope beyond Area 1 to include Area 2, as required. To better visualize the data, I began by creating a frequency table to summarize the brand preferences for both areas. I then used this table to construct a bar chart, which provided a clear and intuitive representation of the distribution of brand preferences across the two demographic areas. This visualization helped highlight any notable differences or patterns in brand preference between the two areas.

****

From the analysis, I observed that Area 1 showed a clear preference for the “Other” brand of breakfast cereal, followed by Brand B and then Brand A. Interestingly, Area 2 exhibited a similar pattern, with “Other” being the most preferred brand, followed by Brand B and Brand A. This suggests that there is no significant difference in the hierarchy of breakfast cereal preferences between the two areas, as “Other” remains the most popular choice across both demographics.

To further investigate whether there is a statistically significant association between Area and Brand preference, I plan to perform a Chi-Square Test of Independence. The hypotheses for this test are as follows:

* Null Hypothesis (H0*H*0​): Brand preference is independent of demographic area.
* Alternative Hypothesis (H1*H*1​): Brand preference is associated with demographic area.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Frequencies** | |  |  |  | **Expected** |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | **Area 1** | **Area 2** | row total |  |  | **Area 1** | **Area 2** | row total |
| **A** | 11 | 19 | 30 |  | **A** | 13.13 | 16.88 | 30.01 |
| **B** | 17 | 30 | 47 |  | **B** | 20.56 | 26.44 | 47 |
| **Other** | 42 | 41 | 83 |  | **Other** | 36.31 | 46.69 | 83 |
| **Total** | **70** | 90 | 160 |  | **Total** | **70** | 90.01 | 160.01 |

As part of the analysis, I created an expected frequency table by calculating the expected frequencies for each combination of Area and Brand preference. This was done using the formula for expected frequencies in a Chi-Square test. Once the expected frequencies were determined, I used Excel’s CHISQ.TEST function to calculate the p-value, which was found to be 0.1927.

Since this p-value is greater than the significance level of 0.05, the results are not statistically significant. This means we fail to reject the null hypothesis and conclude that brand preference is independent of demographic area. In other words, there is no significant association between the area an individual resides in and their preference for breakfast cereal brands. The observed patterns in brand preference are likely due to random variation rather than any meaningful relationship between the two variables.

**Filtration Dataset:**

The final stage in the production of a chemical product involves a filtration process designed to remove impurities, which are unwanted side products. To evaluate the effectiveness of two potential filter agents, **Agent 1** and **Agent 2**, the production manager conducted an experiment. A total of **12 batches** of the product were prepared, and each batch was divided into two equal parts. One part was filtered using **Agent 1**, and the other part was filtered using **Agent 2**. After filtration, the amount of impurity remaining in the product was measured in **parts per 1000 by weight**.

The dataset includes the following variables:

* **Batch**: The identification number of the batch (ranging from 1 to 12).
* **Agent1**: The amount of impurities (in parts per 1000) remaining after filtration with Agent 1.
* **Agent2**: The amount of impurities (in parts per 1000) remaining after filtration with Agent 2.

This setup allows for a direct comparison of the effectiveness of the two filter agents within the same batch, controlling for any batch-specific variations.

**Null Hypothesis (H0*H*0​):** There is **no significant difference** in the mean amount of impurities remaining after filtration using Agent 1 and Agent 2.

**Alternative Hypothesis (H1*H*1​):** There is **a significant difference** in the mean amount of impurities remaining after filtration using Agent 1 and Agent 2.

I started by performing a t-test.

|  |  |  |
| --- | --- | --- |
| t-Test: Paired Two Sample for Means |  |  |
|  |  |  |
|  | *Variable 1* | *Variable 2* |
| Mean | 8.25 | 8.683333 |
| Variance | 1.059091 | 1.077879 |
| Observations | 12 | 12 |
| Pearson Correlation | 0.901056 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 11 |  |
| t Stat | -3.26394 |  |
| P(T<=t) one-tail | 0.003773 |  |
| t Critical one-tail | 1.795885 |  |
| P(T<=t) two-tail | 0.007546 |  |
| t Critical two-tail | 2.200985 |  |
|  |  |  |
| Difference in mean | -0.43333 |  |

The analysis revealed that the mean impurity levels for the two filter agents were relatively close, with a difference of only 0.43 parts per 1000. However, Agent 1 consistently showed slightly lower impurity levels compared to Agent 2, suggesting that it may be more effective in reducing impurities. This difference was clearly demonstrated in the bar chart I created, which visually compared the average impurity levels for both agents across the 12 batches. The chart highlights the trend that Agent 1 outperforms Agent 2, albeit by a small margin.

The p-value obtained from the statistical test was 0.003773, which is less than the significance level of 0.05. This indicates that the results are statistically significant, allowing us to reject the null hypothesis. In practical terms, this means there is a significant difference in the mean amount of impurities remaining after filtration when using Agent 1 compared to Agent 2. This finding supports the conclusion that one filter agent is more effective than the other in reducing impurities in the chemical product.