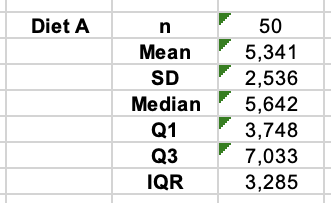
**Statistics Worksheets Comments**

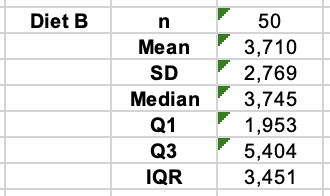
**by Vasilisa Lukashevich**

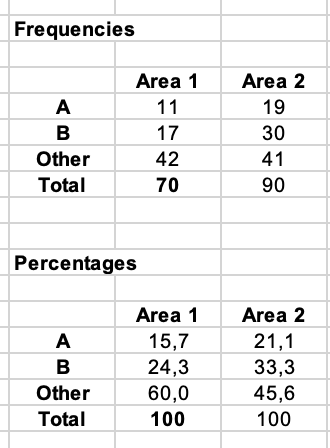
**(Research Methods and Professional Practice May 2023)**

**Exe 8.1B**

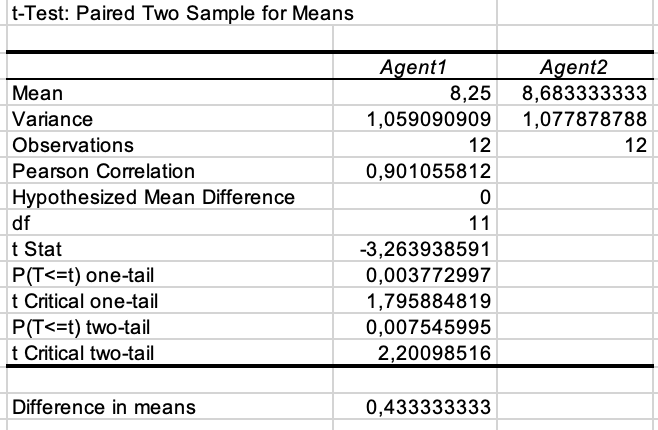
****In the 'Exe 8.1B' exercise on the statistical calculations of Diet A and Diet B, we can conclude that Diet A is more effective than Diet B in the weight reducing context. When examining the sample mean weight loss for Diet B, the result is 3.710, which is 1.631 kg less than average weight loss in the case of Diet A. Furthermore, the standard deviation for Diet B is 233 grams higher than the corresponding indicator for Diet A. This tells us that the variability of the values in the Diet B group is greater, suggesting that the participants in study B lost weight unevenly compared to those in study A.

**Exe 8.2B**

****Once again, the median of the Diet B sample dataset is 1.897 kg lower than in the case of Diet A, highlighting the superior effectiveness of Diet A. In the dataset for Diet B, the interquartile range (IQR) is also higher compared to the Diet A case, providing further evidence of the attribute's variability for asymmetric distributions.

**Exe 8.3D**

In Area 2, out of a total of 90 respondents, 21.1% preferred Brand A, 33.3% preferred Brand B, and the remaining 45.6% preferred another brand of breakfast cereal. This indicates that in Area 2, the Brand A and Brand B are more popular than other brands in contrast to Area 1. However, between Brand A and Brand B customers predominantly prefer Brand B in both areas.

**Exe 8.4G**

In this exercise, we conducted a comparison between two filtration agents, Agent1 and Agent2, to investigate the difference in effectiveness between them. As far as I understood, the lower impurity levels observed in the test batches after using the filter indicate better filtration results.

According to the research, the mean difference between the two filter systems is 0.43 (measured in parts per 1000 by weight).

The obtained data was analysed using a t-Test: Paired Two Sample for Means. The calculated t-value is -3.26 with 11 degrees of freedom.

The associated two-tailed p-value is 0.007, indicating that the observed t-value is significant at the 5% threshold.

Based on the sample mean impurity levels, it can be concluded that Agent1 performs better with a mean impurity level of 8.25 compared to Agent2 with a mean impurity level of 8.68.

**Exe 8.5**

If the lower means are better in this context, indicating a lower level of impurity in the testing samples after using the filter, then our hypothesis regarding the two filter agents should be formulated as follows:

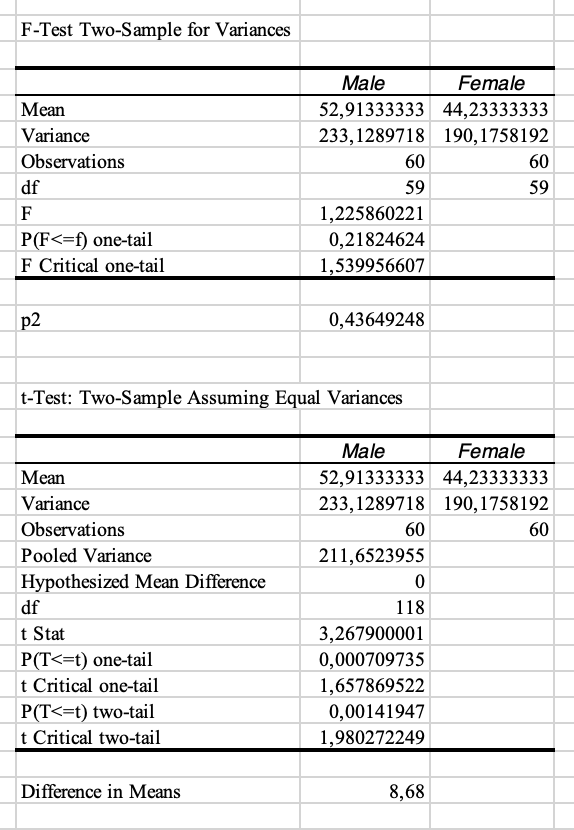
H0: 1 ≥ 2 against H1: 1 < 2

Null hypothesis (H0) states that sample mean of Agent1 is equal or higher than sample mean of Agent 2. On the other hand, our alternative hypothesis (H1) suggests that the mean of Agent1 is lower than mean of Agent2, indicating that Agent1 has a better cleaning effect than Agent2.

Once again, we conducted an analysis of the results using a t-Test: Paired Two Sample for Means. The t-value remains at -3.26, with 11 degrees of freedom.

The associated one-tailed p-value is 0.003, indicating that the observed t-value is statistically significant at the 1% level.

Therefore, we can reject the null hypothesis and conclude that Agent1 is more effective.

**Exe 8.6**

Based on the provided dataset, our conclusion is that the population mean income for males surpasses the corresponding indicator for females.

We compared the sample mean of males’ and females’ annual incomes, and found a difference in means of £8680.

Variance heterogeneity can lead to errors in rejecting the correct null hypothesis in a two-sample t-test (Keselman et al., 2004).

Firstly, we conducted the F-test to examine whether the variances of the two populations are equal.

H0: The variances of the two groups are equal.

H1: The variances of the two groups are not equal.

The sample variances for two groups are 233 (males) and 190 (females).

The F-test statistic for this case is F = ~233/~190 = 1.225, with 59 degrees of freedom for both groups.

Since F (~1.2) < F Critical one-tail (~1.5) and associated two-tailed p-value is p = 0.4364, which is not significant, we cannot reject the null hypothesis.

However, the F-test results does not address the difference in means between the groups. To arrive at a valid conclusion regarding the population mean income for males and females, the two-tailed t-test should be considered and interpreted appropriately (Berenson et al., 2015).

H0: The annual incomes of male and female are the same.

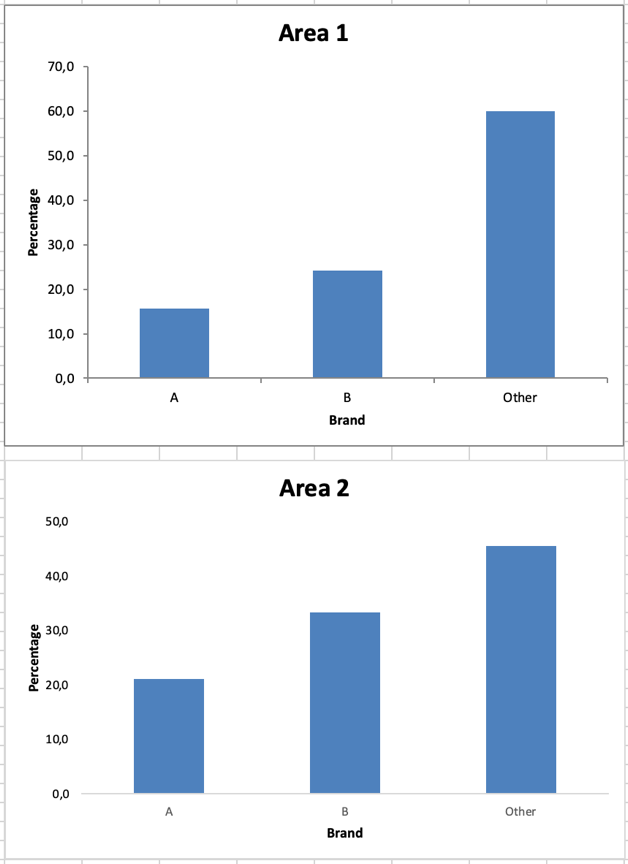
H1: The annual incomes of male and female are not the same.

The obtained independent samples t = 3.267 with 118 degrees of freedom.

The associated two-tailed p-value is p = 0.0014, indicating that the result is statistically significant at the 1% level. This suggests that the difference in means of male and female income is not due to random chance. Therefore, we reject the null hypothesis.

The data provide strong evidence that the underlying mean annual income for this group of males is £8.68 (in thousands) higher than that of the group of females. Hence, the results strongly suggest that males have a higher income than females.

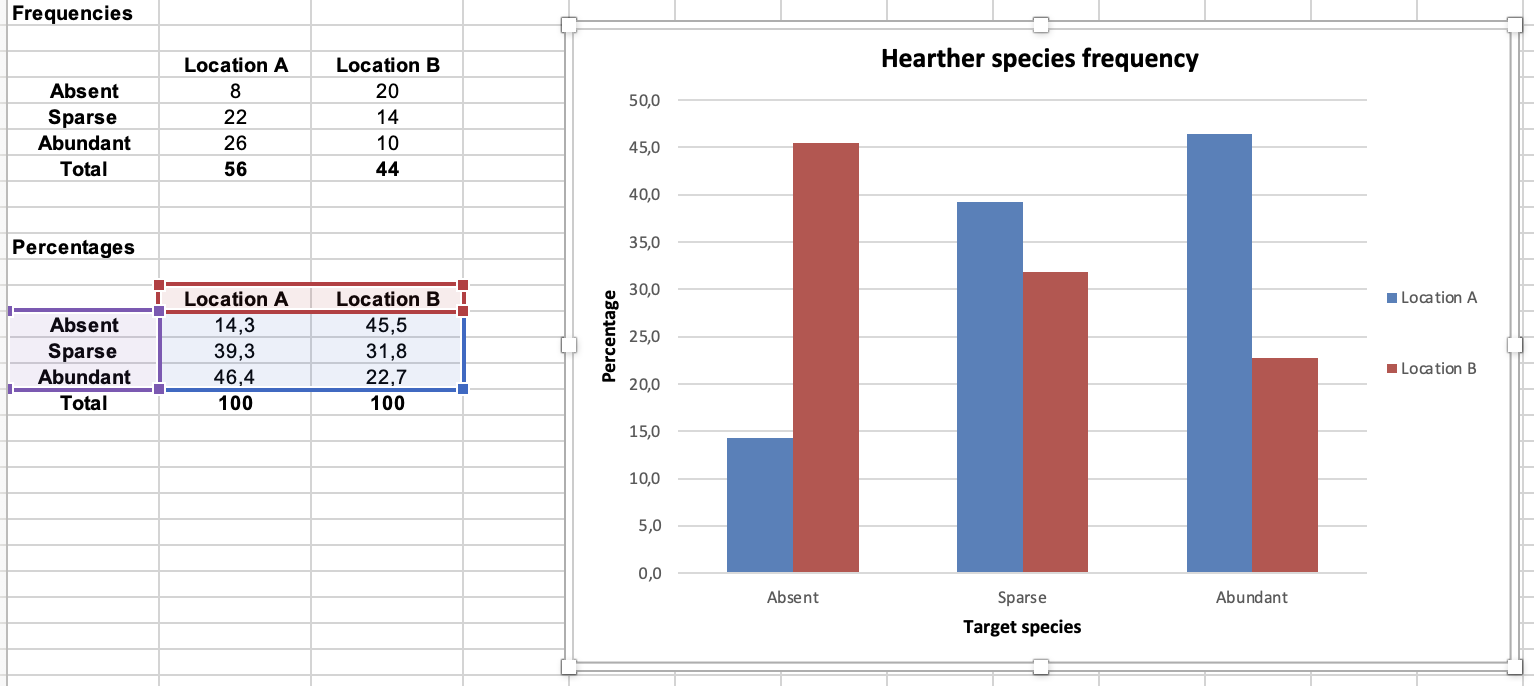
**Exe 9.1**

****

Contrasting demographic Area 1, in demographic Area 2, Brands A and B of breakfast cereal were more commonly chosen compared to other brands, accounting for over half of the responses when combined (54.4%). However, when comparing Brand A and Brand B, customers in both areas predominantly preferred Brand B.

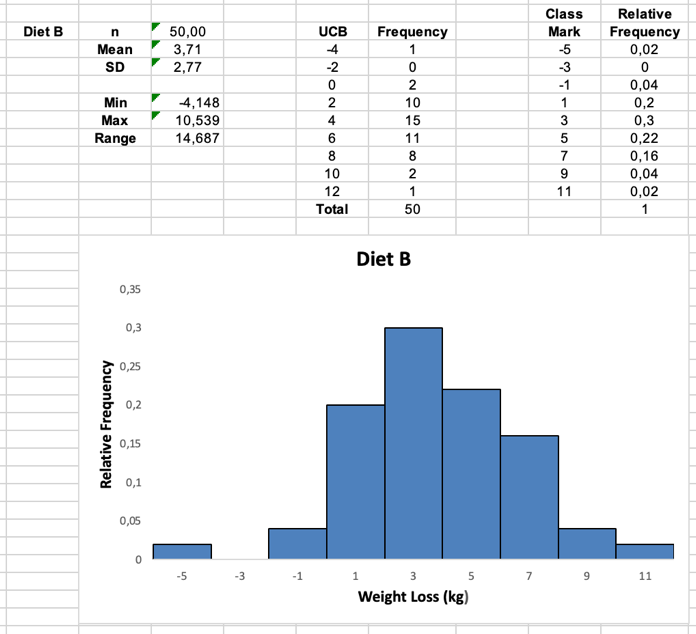
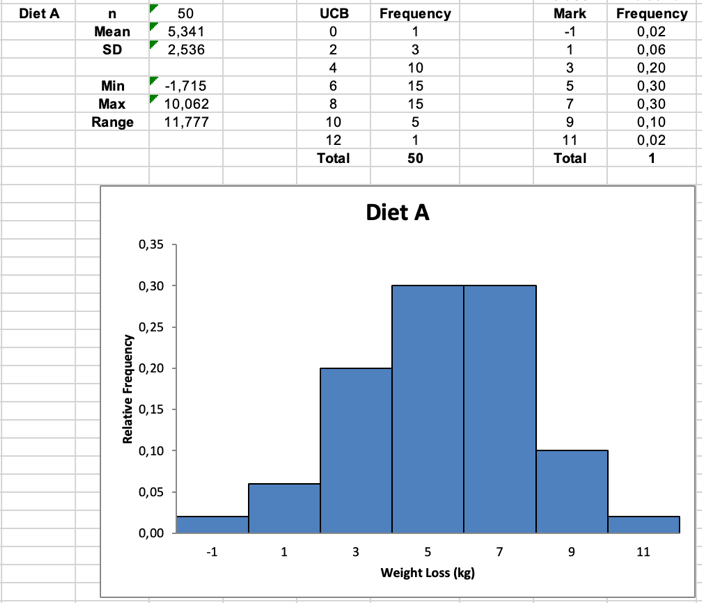
The results may suggest some insights in the context of marketing. For instance, that the promotion efforts for both Brand A and Brand B have been relatively effective in Area 2. And since customers answers that they predominantly prefer Brand B, this brand may have been promoted more effectively or resonated better with the target audience in both areas, than Brand A.

**Exe 9.2**

****

Here is a table on the prevalence of a certain heather species in two heathland locations, A and B. In the percentage frequency clustered column bar, we can see that the two locations differ significantly. In Location A, the most represented species of heather is Abundant (46.4%), while in Location B, the Absent species of the plant are most popular (45.5%). The prevalence of Sparse heather plants is relatively similar in both locations; however, this percentage indicator is 7.5% higher in Location A than in Location B.

**Exe 9.3**

****

As we learned from the previous exercises, the Diet B group exhibits greater variability in the values. In order to maintain consistency with the Diet A histogram, we used a similar interval between classes (in kgs). At first sight, the histogram for Diet B appears to be almost symmetrical but exhibits a slight left-skewness, with a gap at the -3 mark on the X-axis.

However, it is evident that a significant portion of the histogram is located towards the negative side. This observation highlights why Diet B is less effective compared to Diet A, as some participants actually gain weight instead of losing it.

(Therefore, if we exclude in mind the negative weight loss values, the results for Diet B become positively skewed).

The most frequent weight loss results in Diet A were 5 and 7 kilos, which were equally represented in the histogram. On the other hand, in the case of Diet B, the most popular weight loss value appears to be 3 kilos.

**References:**

Berenson, L., Levine, D. & Szabat, K. (2015) *Basic Business Statistics: Concepts and Applications.* 13th Ed. Pearson

Bruce, P., Bruce, A. & Gedeck, P. (2020) *Practical statistics for data scientists: 50+ essential concepts using R and Python.* O'Reilly Media.

Holmes, A., Illowsky, B. & Dean, S. (2017) *Introductory business statistics*. Rice University.

Keselman, H.J., Othman, A.R., Wilcox, R.R. & Fradette, K. (2004) The new and improved two-sample t test. *Psychological Science, 15*(1), pp.47-51.