**IE266: ENGINEERING STATISTICS**

**Case Study 1**

*“Academic integrity is expected of all students of METU at all times, whether in the presence or absence of members of the faculty. Understanding this, I declare that I shall not give, use, or receive unauthorized aid in this study.”*

Group 22

Işıl Çiçek - 2629509

Ekin Akçay - 2577419

Müjde Derya Ocaktan – 2629822

05.05.2024

**Part 2 – Statistical Inference**

With the increasing awareness of retailers to environmental issues, they have started to inspect the actions of their farmers more closely. The top management team of a large retailer, which has many stores in different regions and countries, wants their sustainability department to prepare a report so that the team can set new goals on GHG emission levels from agriculture.

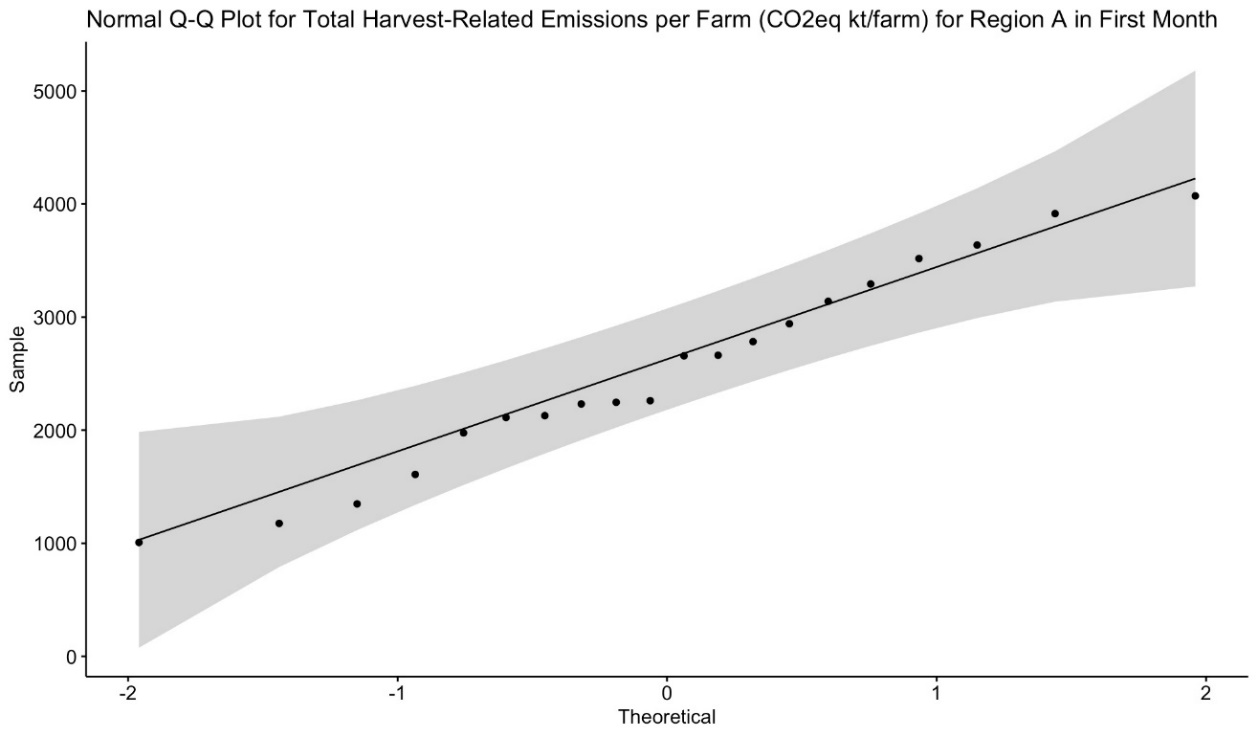
The department investigates the farmers in regions A and B with identical climate and earth conditions. Since a region contains more than 500 farms, samples of 20 farms from region A and 25 farms from region B are taken. It is assumed that the farms’ operations are independent of each other and each week. The data has 12 weeks, the duration of a harvest season. After harvesting one area, the farmer cannot crop the same area again.

**a)** We are required to comment on the data based on the total harvest-related emissions per farm in region A for the first and the second month with a significance level of 0.05. Before conducting the analysis and deciding on confidence intervals, we must check the variances and the similarity of the variances. After summing the total harvest related emissions per farm and filtering the data accordingly we observe that result of the var test as follows:

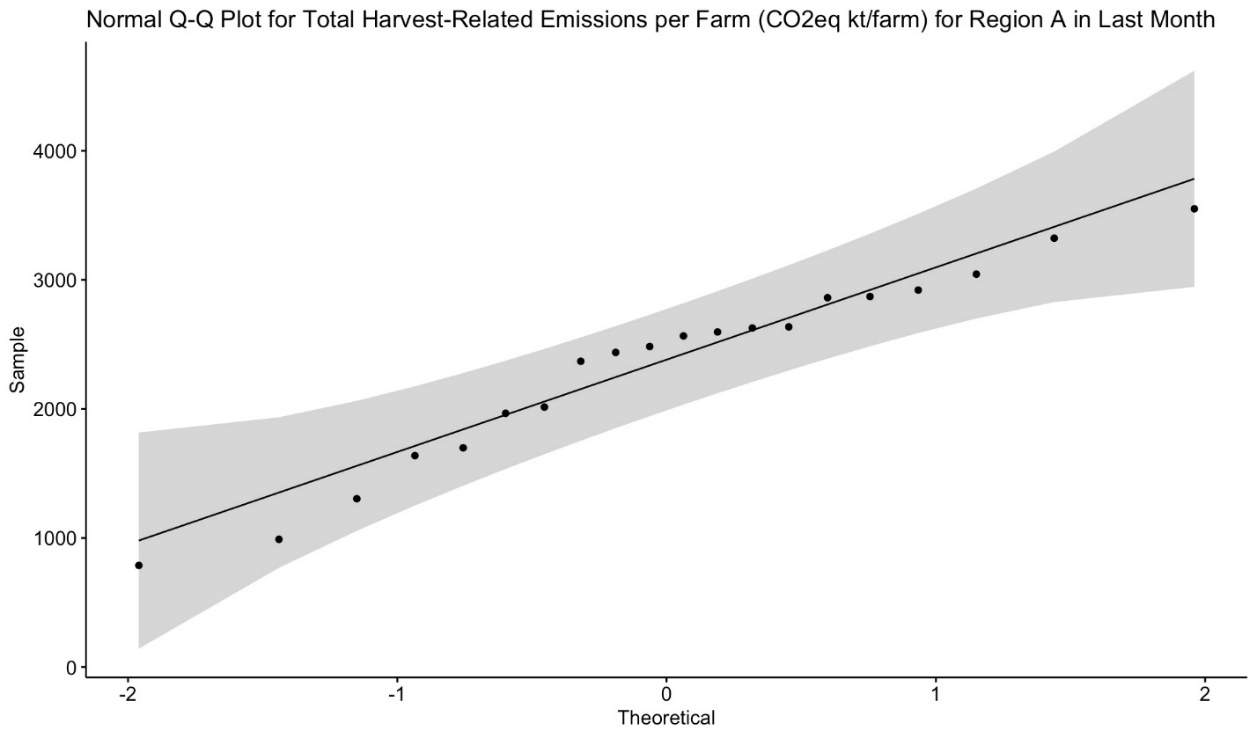
As 1 falls into the interval, we fail to reject the null hypothesis and conclude that the two population variances are identical with 0.05 significance level.

We are given the information that farms’ operations are independent of each other and each week. Therefore, we consider first and last months independent from each other.

Before conducting the t-test, we also have to check the normality of the data.



*Figure 1: Normal Q-Q Plot for Total Harvest-Related Emissions per Farm (CO2eq kt/farm) for Region A in the first month*

**

*Figure 2: Normal Q-Q Plot for Total Harvest-Related Emissions per Farm (CO2eq kt/farm) for Region A in the last month*

According to Figure 1 and Figure 2, we observe points close to the normality line and our data mostly follows a normal distribution. Therefore, we can assume that Total Harvest Related Emissions per Farm for Region A in the first month and Total Harvest Related Emissions per Farm for Region A in the last month are normally distributed.

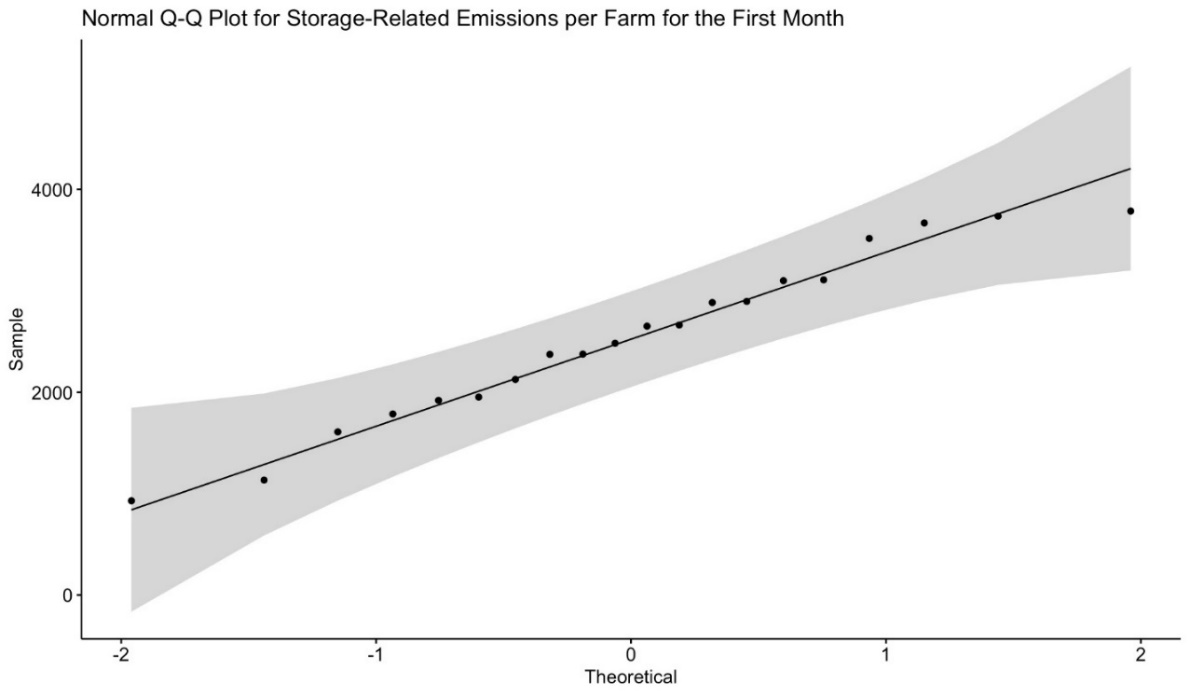
Proceeding with the t-test to decide on the similarity of the means for the first month and the last month. Since we assumed the normality of the data and equality of the variances our null and alternative hypothesis are described as follows:

And our test statistic is described as:

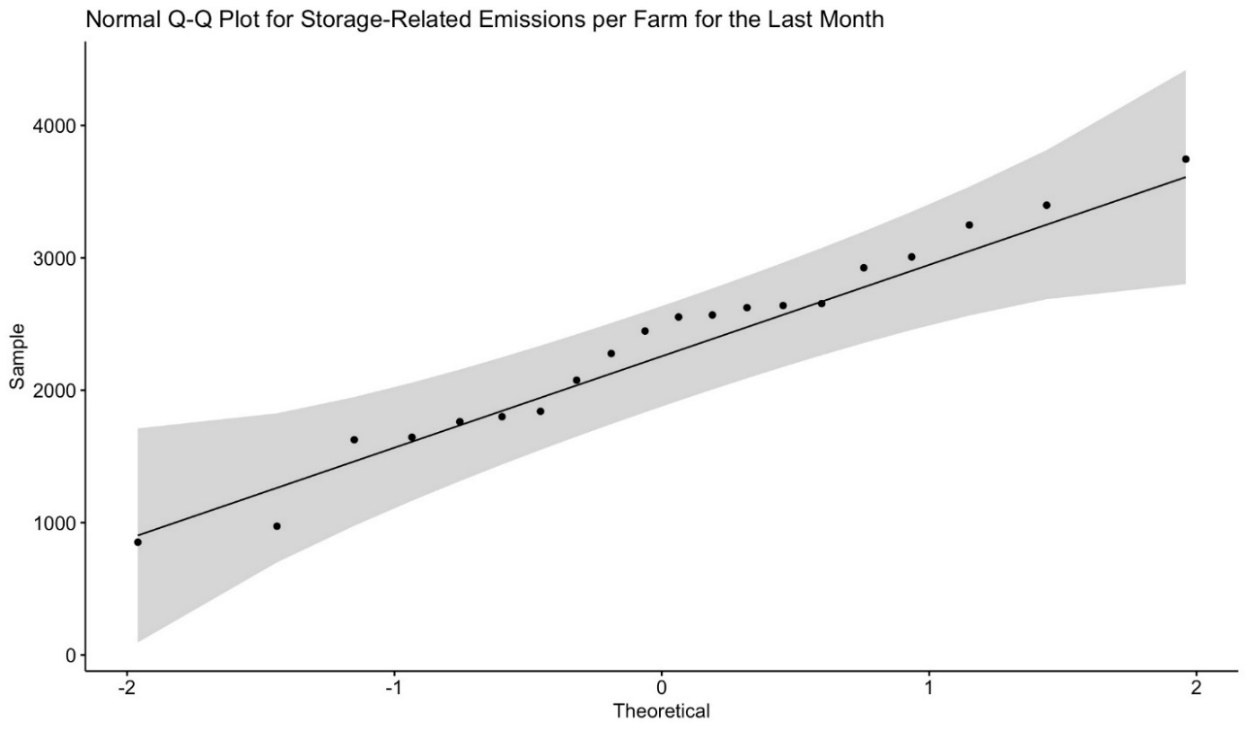
Therefore, we fail to reject the null hypothesis and conclude that the means of first and the last month are equal.

Additionally, we are required to conclude the equality of total storage related emissions per farm for the first and the last month in Region A, with a significance level of 0.05. Before conducting the analysis and deciding on confidence intervals we must check the variances and the similarity of the variances. After summing the total storage related emissions per farm and filtering the data accordingly, we observe that the result of the var-test as follows:

As 1 falls into the interval, we can assume that the two populations has the same variance with a confidence coefficient of 0.05.



*Figure 3: Normal Q-Q Plot for Total Storage-Related Emissions per Farm for Region A in the first month*

**

*Figure 4: Normal Q-Q Plot for Total Storage-Related Emissions per Farm for Region A in the last month*

Referring to Figure 3 and Figure 4, we visualize points near to the normality line and our data mostly follows a normal distribution. Therefore, we can assume that Total Storage Related Emissions per Farm for Region A in the first month and Total Storage Related Emissions per Farm for Region A in the last month are normally distributed.

We are moving on with the t-test to decide on the similarity of the means for the first month and the last month. Since we assumed the normality of the data and equality of the variances, our null and alternative hypothesis can be described as follows:

And our test statistic is described as:

Therefore, we fail to reject the null hypothesis and conclude that the means of first and the last month are equal.

**b)** Firstly, we are required to compare the average Harvest Amount of Region A per farm for each week with Region B at 0.05 significance level. For this purpose, a two-sample t-test is applied, and the result of this test proves that the averages of regions A and B are similar to each other. Normality assumption is also made as we have used t-test for all three cases in b part. By applying the variance test for each week separately, we find out that the variances are also identical. For the first part, each week’s average harvest amounts for Region A and B are compared at a 0.05 significance level with the test statistic:

The following test statistic values and p-values found are:

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Test Statistic Value** | **p-value** | **Status** |
| 1 | 0.85256 | 0.3986 | Fail to reject |
| 2 | 1.4221 | 0.1622 | Fail to reject |
| 3 | 6.0187 | 3.434e-07 | Reject |
| 4 | -0.9664 | 0.3392 | Fail to reject |
| 5 | 3.5504 | 0.0009466 | Reject |
| 6 | 2.2951 | 0.02667 | Reject |
| 7 | 4.5377 | 4.528e-05 | Reject |
| 8 | 2.4451 | 0.01865 | Reject |
| 9 | -1.6739 | 0.1014 | Fail to reject |
| 10 | 3.3338 | 0.001771 | Reject |
| 11 | 3.6376 | 0.000732 | Reject |
| 12 | -0.08992 | 0.9288 | Fail to reject |

*Table 1: Average Harvest Amount of Region A per farm for each week with Region B at 0.05 significance level*

From the results of the table above, we observe that the averages are distinct values in some weeks where is rejected. To compare their values, a one-sided 95% lower CI is established for those weeks with a new null and alternative hypothesis:

**Week 3:**

**Week 5:**

**Week 6:**

**Week 7:**

**Week 8:**

**Week 10:**

**Week 11:**

Since 0 does not fall into any CI, we reject which shows that the average of Region A is higher than Region B in these weeks.

For the second part, we apply the same analysis for each week’s average On-Farm Loss/Harvest Amount ratio per farm. We start by using the F test to compare the variances at a 0.05 significance level.

By applying var-test, we observe 1 falls into the CI and fail to reject for all weeks except for week 10. For week 10, at 95% CI:

Now that we have found out variances are identical for almost each week, we use two-sample t-test to determine whether the averages are equal or not for each week separately. However, for week 10, we use Welch two-sample t-test. The test statistic we use for that case specifically is:

At a 95% confidence interval:

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Test Statistic Value** | **p-value** | **Status** |
| 1 | 0.84486 | 0.4029 | Fail to reject |
| 2 | -0.021224 | 0.9832 | Fail to reject |
| 3 | 0.053315 | 0.9577 | Fail to reject |
| 4 | -0.75012 | 0.4573 | Fail to reject |
| 5 | 0.4108 | 0.6833 | Fail to reject |
| 6 | -0.42897 | 0.6701 | Fail to reject |
| 7 | -1.9992 | 0.05194 | Fail to reject |
| 8 | 0.014677 | 0.9884 | Fail to reject |
| 9 | 0.16053 | 0.8732 | Fail to reject |
| 10 | -0.37296 | 0.7112 | Fail to reject |
| 11 | -0.3547 | 0.7245 | Fail to reject |
| 12 | 0.10294 | 0.9185 | Fail to reject |

*Table 2: Comparison of average On-Farm Loss/Harvest Amount ratio of Region A per farm for each week with Region B at 0.05 significance level*

Based on the results of the table above, we can say that the average values are the same for each week.

For the last part, the same analysis is made to inspect whether Harvest-related emissions are greater than Storage-related emissions of Region A.

Since both emissions depend on the amount of harvested rice, they are dependent on each other. Therefore, a paired t-test must be used for comparison:

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Test Statistic Value** | **p-value** | **Status** |
| 1 | 0.0038096 | 0.4985 | Fail to reject |
| 2 | 0.27591 | 0.3928 | Fail to reject |
| 3 | -0.9597 | 0.8254 | Fail to reject |
| 4 | 0.70099 | 0.2459 | Fail to reject |
| 5 | 0.34404 | 0.3673 | Fail to reject |
| 6 | -1.1509 | 0.868 | Fail to reject |
| 7 | 1.2784 | 0.1083 | Fail to reject |
| 8 | -0.062675 | 0.5247 | Fail to reject |
| 9 | 2.2853 | 0.01698 | Reject |
| 10 | -0.43531 | 0.6659 | Fail to reject |
| 11 | -0.22456 | 0.5876 | Fail to reject |
| 12 | -1.0141 | 0.8384 | Fail to reject |

*Table 3: Comparison of Storage-related emissions and Harvest-related emissions of Region A at 0.05 significance level*

As a result, Storage-related emissions of Region A are greater than Harvest-related emissions except for week 9. It may be equal or Harvest-related emissions may be greater for that week only.

**c)** The farm's harvest and storage-related emissions/yield is seen as a good indicator of the emission efficiency by retailers. A farm is indicated as efficient if its total emission/total yield ratio in a harvest season is less than 1.25. The alternative hypothesis suggests that the proportion of efficient suppliers in Region A is greater than Region B and we are suggested to check whether there’s sufficient evidence to reject this hypothesis at a 0.05 significance level. Firstly, we start by assigning the value 1 to data with efficient status and 0 to those not efficient for both regions. Then, we obtain the ratios of farms with weeks in efficient status to the total number of farms and weeks in a region: Proportion of Region A = 0.65 and Proportion of Region B = 0.44.

As we are checking whether the ratio is less than 1.25, we have used a one-sided confidence interval with lower bound:

Since this interval includes 0, we fail to reject which means the proportion of efficient suppliers in Region A is greater hence more efficient than Region B.