

Dataset 7: Fecundity of fruitflies

Group members:

Vladyslav Horbatenko, vladyslav@ruc.dk

Layba Naeem, layba@ruc.dk

Maria Paz Buonomo Quiroga, paz@ruc.dk

Rishni Mahendran, rishni@ruc.dk

Maxim Gutu, stud-gutu@ruc.dk

Cojocarv Vlad, stud-cojocarv@ruc.dk

Group Project
Winter Semester 2024



Roskilde Universitet

Nov 2024

Contents

1	Introduction	2
2	Methods	2
2.1	Used methods overview	2
2.2	One way ANOVA	3
2.3	Post-Hoc Test: Tukey's HSD	3
2.4	T-test	3
3	Analysis	4
3.1	Summary statistics	4
3.1.1	Box Plots	4
3.1.2	Mean	5
3.1.3	Standard Deviation	5
3.2	Normality check	5
3.3	One-way ANOVA	5
3.3.1	Post-Hoc Test 1	6
3.3.2	Post-Hoc Test 2	7
3.4	T-test	7
4	Conclusion	8

1 Introduction

This study aims to analyze a dataset on the fecundity of a species of fruit flies called *Drosophila melanogaster*. The data is collected by R.R. Sokal in Biometry (Sokal & Rohlf, 1981) and contains the fecundity data of 25 fruit flies over 14 days. The dataset considers three genetic lines of the fruitfly *Drosophila melanogaster*, which were labeled in the following way: Resistant group (RS) that was selectively bred for resistance to DDT (insecticide), Susceptible group (SS) that was bred for susceptibility to DDT and the Nonselected group (NS) that is the control group.

From dataset we can see that the experimental unit in this study is identified as female fruit flies, and a total of 75 observations were analyzed, with 25 female fruit flies in each of the three groups.

The main aim of this study is to investigate the following two hypothesis:

1. Does the fecundity differ significantly between the selected lines (RS and SS) and the nonselected (NS) line?
2. Does the fecundity of the line selected for DDT resistance (RS) differ from the line selected for DDT susceptibility (SS)?

In order to have a better understanding on the way these hypotheses should be investigated, basic statistical analysis is first provided in the Analysis section (3) of this report. Following different approaches and methods are used to investigate the above stated hypothesis. An ANOVA test is conducted to determine differences among the three groups and so answering the first hypothesis. Based on the results given by the ANOVA test, a post-hoc test is conducted to compare the different groups. Finally, the second hypothesis is investigated through application of a t-test which determines the difference between the two selected groups.

** This project is based on literature from course material and the text book [1]*

2 Methods

2.1 Used methods overview

We visualized fecundity across the three groups using boxplots to examine distribution patterns and identify potential outliers. To test normality, we applied the Shapiro-Wilk test, chosen for its sensitivity to small samples ($n < 50$), its direct assessment of normality, and its superiority over alternatives like the Kolmogorov-Smirnov and D'Agostino-Pearson tests for small datasets.

To explore the first question, we need to compare three groups to determine if there is a difference between them. For this purpose, we could use several tests such as a One-Way ANOVA, Two-Way ANOVA, One-Way Randomized Blocks Design.

However, the One-Way ANOVA was chosen as it is specifically designed to compare the means of three or more independent groups. Our groups (RS, SS, and NS) are independent and do not have repeated measurements. The data satisfied ANOVA's assumptions of normality (via Shapiro-Wilk test).

The other tests were not chosen as they were not suitable to our case study. For instance, the *Two-Way ANOVA* test is used to evaluate the effects of two independent variables and their interaction; since our study involves only one independent variable (genetic line), it is not applicable. While, the *One-Way Randomized Blocks Design* test is used for repeated measurements or when blocking is required. As our groups are independent and have no repeated measures, it is not suitable.

We used Tukey's HSD because it identifies group differences after a significant ANOVA result, controls for Type I error in multiple comparisons, and is suitable for independent group means, as in our study.

On the other hand, we are not using Bonferroni Adjustment as its overly conservative and reducing statistical power, or Games-Howell Test, because it is meant for unequal variances, which did not apply to our data.

To address the second question, we chose the independent samples *t*-test as it is specifically designed to test for differences between the means of two independent groups. The assumptions of normality (Shapiro-Wilk test) and

equal variances (Levene's test) were satisfied, and the t -test is more statistically powerful than non-parametric alternatives when these assumptions are met.

We did not use the Mann-Whitney U test because, although it does not require normality, it is less powerful than the t -test when parametric assumptions are satisfied, making the t -test the better choice for our data.

2.2 One way ANOVA

One way ANOVA is a statistical test to examine the difference in the means of more than two groups by focusing on one independent factor that characterizes the groups.

This test takes into account the variance between the samples in a group and also the variance between the groups that are being compared.

One way ANOVA is performed through hypothesis testing. The null and alternative hypothesis is formulated as:

- Null hypothesis(H_0) = The means of all the groups are equal
- Alternative Hypothesis(H_a) = The means of the groups are not equal

The next step is to calculate f-statistic given as:

$$F_s = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

where MS_{between} is the mean square between the groups while MS_{within} is the mean square within the group. Lastly, the calculated f-statistic is compared with the f critical from the F-distribution table to draw conclusion about whether to reject or accept the null hypothesis.

2.3 Post-Hoc Test: Tukey's HSD

Tukey's HSD (Honestly Significant Difference) Test is a **post-hoc analysis** used after ANOVA to determine which specific group means differ significantly from each other. It controls the **Type I error rate** when multiple comparisons are made.

1. **Compute the Standard Error (SE):** Based on the within-group variance and sample size.
2. **Determine the Critical Value (q):** From the Studentized Range Distribution, based on group counts and degrees of freedom.
3. **Calculate the HSD:** $HSD = q \cdot SE$.
4. **Compare Group Differences:** The mean differences between groups are tested against the HSD. If the difference exceeds the HSD, it is statistically significant.

2.4 T-test

T test is a statistical method that examines the difference in the means of two groups by hypothesis testing to observe if they are statistically different from each other. This test is used when the underlying population is considered to be normally distributed and the data points are not dependent on each other.

The following hypothesis are formulated in order to conduct a t test:

- H_0 : There is no difference in the means of the two comparing
- H_a : There is difference between the means of the comparing groups.

After formulating the hypothesis, the next step is the calculation of the t- statistic. The t-statistic can be calculated using the formula:

$$t_s = \frac{(y_1 - y_2) - 0}{SE(y_1 - y_2)}$$

where $y_1 - y_2$ is the difference between the sample means of two groups and $SE(y_1 - y_2)$ is the standard error of the difference between the two means. The degrees of freedom are calculated according to the t test being conducted and a level of significance is set for the test usually $\alpha = 0.005$. Lastly, the calculated t-statistic is compared with the t-critical and a decision is made about the rejection or acceptance of the null hypothesis.

3 Analysis

3.1 Summary statistics

In order to conduct the statistical analysis, it is important to have a concise overview of the data which can be done by calculating the appropriate summary statistics and representing the data graphically. These are useful to identify patterns in the dataset and any potential outliers. Furthermore, it helps to make decision about what kind of tests should be conducted for hypothesis testing.

The summary statistics below is provided by `df.describe()` method from `pandas` library, which provides us with overview of the data and some basic information:

Statistic	Resistant	Susceptible	Nonselected
Count	25.000	25.000	25.000
Mean	25.256	23.628	33.372
Standard Deviation (std)	7.772	9.768	8.942
Minimum (min)	12.800	10.800	14.900
25% Quantile	20.300	16.000	28.200
Median (50%)	23.600	22.500	34.400
75% Quantile	29.300	30.200	37.900
Maximum (max)	44.400	48.500	51.800

Table 1: Pandas Summary Statistics for Resistant, Susceptible, and Nonselected Groups

The following table provide general statistical information about the RS, SS and NS. It gives information regarding the sample size, mean, standard deviation, minimum and maximum points among the different groups, first, second and third quartile. These values will be essentials when we compare three groups and visualize their distributions.

Here, important are the mean and standard deviation values. We can observe that the NS group has the highest mean value compared to the other two groups. On the other hand the std is much higher in the SS group.

3.1.1 Box Plots

To visualize some of the data from table (1), box plots can be used for RS,SS and NS. It help visually compare the fecundity of the different groups. Box plots show the spread of the data and out-liners present in the datasets.

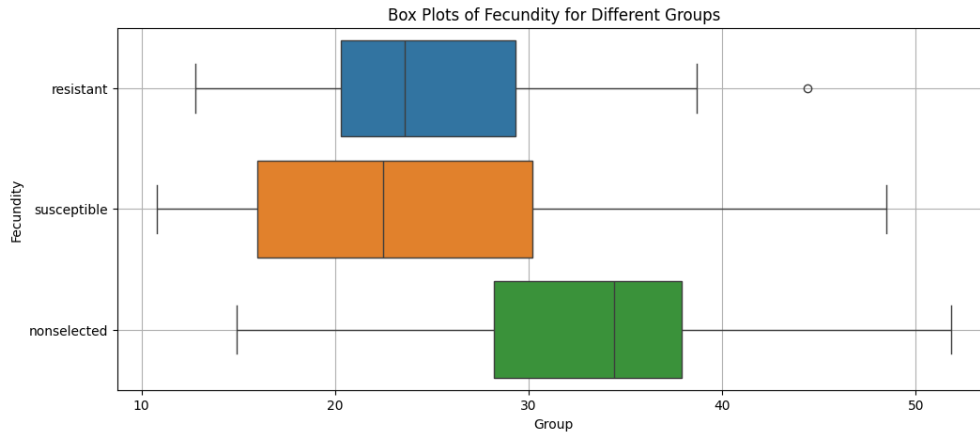


Figure 1: Enter Caption

From the box plot, it can be observed that the NS group has the highest median. This suggests that on average the NS group has higher fecundity compared to the other groups. The lowest median is found in the SS group, that consequently is considered the one to have the lowest fecundity.

A visible difference is seen in the SS group, where the box is comparatively wider and presents longer whiskers. This suggests that there is much more variability, therefore higher fecundity value. While on the other hand, the

boxes from RS and SS look more similar to each other with a difference in the whiskers that are longer in the NS group and much shorter in the RS. This indicates that the RS has the smallest range of fecundity. An outlier is then also present above the upper whisker on the RS. This point represents an individual showing high fecundity compared to the rest of the group

3.1.2 Mean

Mean is one of the most important and frequently used metric in statistics which is the point around which the data is distributed and is calculated by adding all the data points and dividing by the total number of the data points.

- RS= 25.25
- SS=23.62
- NS=33.3

3.1.3 Standard Deviation

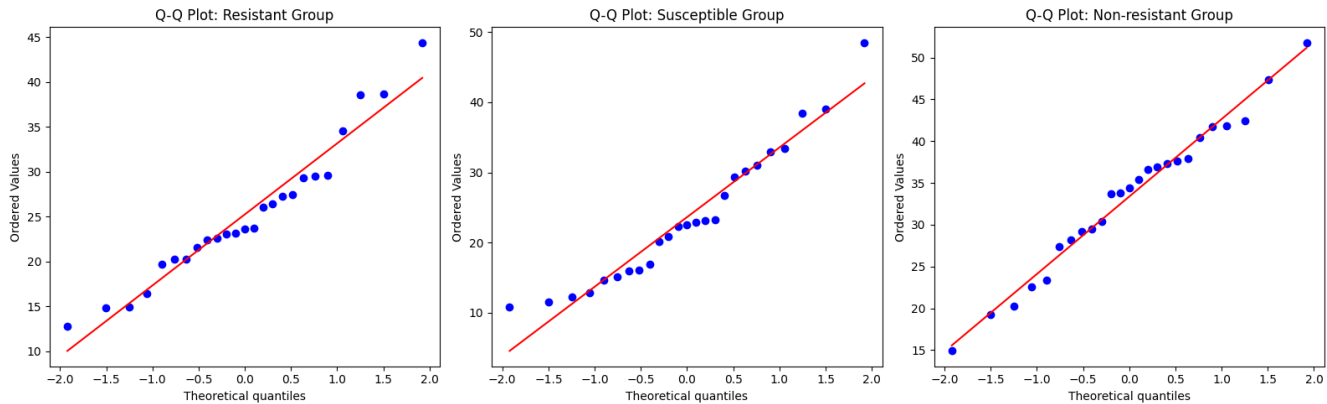
Standard deviation of a data set is also an important metric that gives the measure of how much the data points in the given sample or group are deviated or dispersed from their mean. Higher values of standard deviations indicate more spread of data around the mean and vice versa.

3.2 Normality check

To check if the data satisfies the conditions for the chosen Statistical tests a normality check of the samples is carried out. The code to do so is provided below:

```
1 stats.probplot(RS, dist="norm", plot=axes[0]) # Q-Q plot for resistant group
2 stats.probplot(SS, dist="norm", plot=axes[1]) # Q-Q plot for susceptible group
```

The plots obtained are the following:



The quantile plots show that the data aligns closely to a straight line for both cases. From this, we can infer that the data follows a normal distribution.

To double-check this, A Shapiro-Wilk test is also performed (Code is provided in the appendix). The obtained p-value for RS is 0.245, and the p-value for SS is 0.145. Both values are greater than 0.05, therefore our data is indeed normal.

3.3 One-way ANOVA

To perform one-way ANOVA, certain values need to be extracted from the dataset. The number of data points (n_i) and the mean (\bar{y}_i) for each group can be calculated using methods such as `df.column.count()` and `df.column.mean()`. The overall mean (\bar{y}) is then calculated as the weighted mean of the group means. These values serve as the basis

for calculating SS_{between} , SS_{within} , and SS_{total} .

First, we calculate SS_{between} , which measures the variability between group means relative to the overall mean. The formula for SS_{between} is:

$$SS_{\text{between}} = \sum_{i=1}^I n_i (\bar{y}_i - \bar{y})^2$$

Here, n_i represents the number of data points in the i -th group, \bar{y}_i is the mean of the i -th group, and \bar{y} is the overall mean. SS_{between} measures the variability between group means relative to the overall mean.

Next, we calculate SS_{within} , which measures the variability within each group around their respective group means. The formula for SS_{within} is:

$$SS_{\text{within}} = \sum_{i=1}^I (n_i - 1) s_i^2$$

where s_i^2 is the variance of the i -th group. For this dataset, the calculated SS_{within} value is 5432.66, representing the variability within the groups.

Finally, we calculate the Total Sum of Squares (SS_{total}), which represents the total variability in the data and is the sum of SS_{between} and SS_{within} . Thus:

$$SS_{\text{total}} = SS_{\text{between}} + SS_{\text{within}} = 1362.21 + 5432.66 = 6794.872$$

With SS_{between} , SS_{within} , and SS_{total} calculated, we can compute the F-statistic for the one-way ANOVA test. The F-statistic is used to determine whether the differences between group means are statistically significant and is calculated as:

$$F = \frac{\frac{SS_{\text{between}}}{k-1}}{\frac{SS_{\text{within}}}{N-k}} = \frac{\frac{1362.21}{2}}{\frac{5432.66}{72}} = 9.0268$$

Here, k is the number of groups, and N is the total number of observations.

This completes the required calculations for performing the one-way ANOVA.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-statistic	p-value
Between Groups	1362.211467	2	681.105733	9.026812	0.000318
Within Groups	5432.661504	72	75.453632		
Total	6794.872971	74			

Table 2: ANOVA Table

The test yields an F-statistic of 9.027, which is a relatively large number, pointing us in the direction that the variance between the groups means is much larger than the variance within groups. To verify this, we look at the p-value, which is equal to 0.000318. If we choose $\alpha = 0.05$, we can say that the p-value is much smaller than alpha, which allows us to reject our null hypothesis. Therefore we can be sure that there is significant difference between the groups. This leads us to perform a post-hoc test, to find out where that difference is.

3.3.1 Post-Hoc Test 1

Tukey's HSD 1st method:

```

1 from statsmodels.stats.multicomp import pairwise_tukeyhsd
2 data_melted = df.melt(var_name='Group', value_name='Value')
3 tukey_test = pairwise_tukeyhsd(endog=data_melted['Value'], # Numerical values
4                               groups=data_melted['Group'], # Groups
5                               alpha=0.05)
6 print(tukey_test.summary())
7 import matplotlib.pyplot as plt

```

We receive the following summary:

Group1	Group2	MeanDiff	p-adj	Lower	Upper	Reject
Nonselected	Resistant	-8.116	0.0051	-14.1169	-2.1151	True
Nonselected	Susceptible	-9.744	0.0007	-15.7449	-3.7431	True
Resistant	Susceptible	-1.628	0.7934	-7.6289	4.3729	False

Table 3: Multiple Comparison of Means - Tukey HSD, FWER=0.05

And corresponding plot:

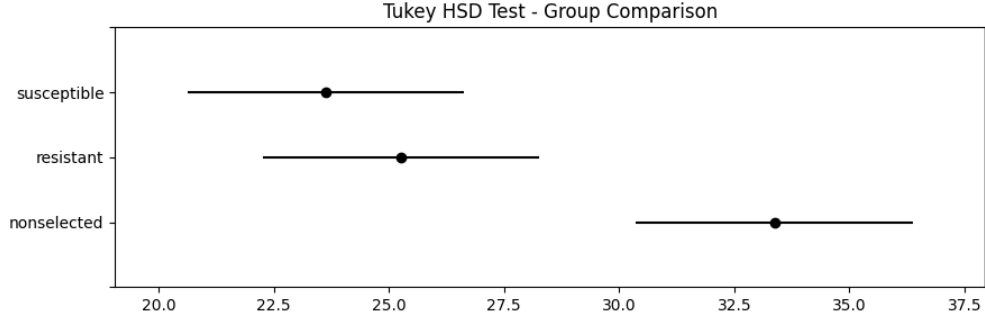


Figure 2: Enter Caption

3.3.2 Post-Hoc Test 2

To perform Tukey's HSD (2nd method), we begin by calculating the standard error (SE) of group differences using the formula $SE = \sqrt{\frac{MS_{within}}{n_{per\ group}}}$. The result is $SE = 1.7736$.

Next, we obtain the critical value (q) from statistical tables for $df_{between} = 2$, $df_{within} = 72$, and $\alpha = 0.05$, which gives $q = 3.5$.

The HSD is then calculated as $HSD = q \times SE$, resulting in $HSD = 6.206$.

The differences between group means are computed as

$$\text{Difference}_{NS,RS} = |\bar{x}_{NS} - \bar{x}_{RS}|,$$

$$\text{Difference}_{NS,SS} = |\bar{x}_{NS} - \bar{x}_{SS}|, \text{ and}$$

$$\text{Difference}_{RS,SS} = |\bar{x}_{RS} - \bar{x}_{SS}|.$$

Finally, these differences are compared to the HSD.

The difference between NS and RS is 8.116, which is **significant**.

The difference between NS and SS is 9.744, also **significant**.

However, the difference between RS and SS is 1.628, which is **not significant**.

3.4 T-test

To answer our second question: Did the line selected for resistance differ in fecundity from the line selected for susceptibility? A t-test is performed. The hypothesis to answer our question are the following:

H0: mean of RS = mean of SS

H1: mean of RS \neq mean of SS

The code used to carry out the test is presented below, with a step-by-step of the calculations of the quantities needed.


```

1 diff_means = rs_mean - ss_mean #Difference of the means
2 SE_diff = math.sqrt((rs_std_dev**2 / rs_count) + (ss_std_dev**2 / ss_count)) #standard error of
   the difference of the means
3 t_s = (diff_means-0)/SE_diff #t statistics
4
5 #Calculate the degrees of freedom
6 SE_rs = rs_std_dev / math.sqrt(rs_count)
7 SE_ss = ss_std_dev / math.sqrt(ss_count)
8 df = (SE_rs**2 + SE_ss**2)**2/(SE_rs**4/(rs_count-1)+SE_ss**4/(ss_count-1))
9
10 #Calculate p-value
11 p = stats.t.sf(np.abs(t_s), df)
12 p = p*2

```

Listing 1: T-test

The results obtained are:

- The SE of the difference is: 2.4462189599461452 and the t statistic is: 0.6655168759038024
- The p value is: 0.5090656597345964

If we choose $\alpha = 0.05$ and we compare it with the p-value obtained, we can see that the p-value is larger than alpha. This means that there is no evidence to reject H_0 , meaning that we have no evidence that the RS is any different than the SS group.

4 Conclusion

This report aimed to do a statistical analysis of the Fecundity of fruitflies data set. With a particular goal of answering the following questions:

1. Does the fecundity differ significantly between the selected lines (RS and SS) and the nonselected (NS) line?
2. Does the fecundity of the line selected for DDT resistance (RS) differ from the line selected for DDT susceptibility (SS)?

To address the first question, a One-way ANOVA test was conducted among the three groups. revealing significant difference between groups. Since this test does not tell us where that difference is, we decided to perform a Tukey post hoc test. This led to the result that the significant difference was indeed between the non-selected and the selected lines. Giving us a clear answer to our first question.

Regarding the second question, we already had some information from the Tukeys test that told us there was no significant difference found between the SS and RS lines. To confirm this finding, we performed a t-test with these two groups. This test verified the previous information, showing there is no significant difference between the two selected lines.

References

- [1] Myra L Samuels, Jeffrey A Witmer, and Andrew A Schaffner. *Statistics for the life sciences*. eng. Fifth edition, Global edition. Always learning. Boston, [Massachusetts]: Pearson, 2016. ISBN: 1292101814.

Appendix

```
1 # -*- coding: utf-8 -*-
2 """MiniProjectStatModels.ipynb
3
4 Automatically generated by Colab.
5
6 Original file is located at
7     https://colab.research.google.com/drive/1250DV-fm_kGm7o8bDlAvCE2Fsd85U1NX
8
9 ## Imports
10 """
11
12 from google.colab import drive
13 drive.mount('/content/drive')
14
15 import pandas as pd
16 import numpy as np
17 import matplotlib.pyplot as plt
18 import seaborn as sns
19 from google.colab import files
20 import pandas as pd
21 import scipy.stats as stats
22 import math
23
24 uploaded = files.upload()
25
26 df = pd.read_csv('Dataset 7 Fecundity of fruit flies.csv')
27
28 """## Data analysis and pre-processing"""
29
30 df.drop(columns=['Unnamed: 3'], inplace=True)
31
32 print(df.info())
33 df.head()
34
35 # Separate data into RS, SS, and NS
36 RS = df['resistant']
37 SS = df['susceptible']
38 NS = df['nonselected']
39
40
41 df.describe()
42
43 # boxplot
44 plt.figure(figsize=(12, 5))
45 sns.boxplot(data=[RS, SS, NS], orient='h')
46 plt.title('Box Plots of Fecundity for Different Groups')
47 plt.ylabel('Fecundity')
48 plt.xlabel('Group')
49 plt.grid('True')
50 plt.show()
51
52 colors = ['blue', 'orange', 'green']
53
54 # histogram for each column with a unique color
55 for i, column in enumerate(df.columns):
56     plt.figure(figsize=(8, 6))
57     sns.histplot(df[column], kde=False, bins='auto', alpha=1, color=colors[i])
58     plt.title(f"Histogram of {column}")
59     plt.xlabel("Values")
60     plt.ylabel("Frequency")
61     plt.grid(True, linestyle='--', alpha=0.7)
62     plt.show()
```

```

63
64 # to count number of data inputs
65
66 rs_count = RS.count()
67 ss_count = SS.count()
68 ns_count = NS.count()
69
70 rs_count, ss_count, ns_count
71
72 """Now lets calculate some data for further analysis and save then into variables thus we can use
    them later"""
73
74 rs_mean = RS.mean()
75 ss_mean = SS.mean()
76 ns_mean = NS.mean()
77
78 #overall mean
79 overall_mean = (rs_mean + ss_mean + ns_mean)/3
80
81 rs_mean, ss_mean, ns_mean, overall_mean
82
83 """Now lets calculate the standart deviation and other metrics"""
84
85 def calculate_std_dev(data):
86     n = len(data)
87     if n == 0:
88         return 0 # handle empty data case
89
90     mean = sum(data) / n
91     variance = sum([(x - mean) ** 2 for x in data]) / n
92     std_dev = math.sqrt(variance)
93     return std_dev
94
95 rs_std_dev = calculate_std_dev(RS.dropna().tolist()) # handle potential NaN values
96 ss_std_dev = calculate_std_dev(SS.dropna().tolist())
97 ns_std_dev = calculate_std_dev(NS.dropna().tolist())
98
99 rs_std_dev, ss_std_dev, ns_std_dev
100
101 import math
102
103 k = df.shape[1] # number of groups
104 n = rs_count + ss_count + ns_count # total number of observations
105
106 # Variances for each group (square of standard deviation)
107 rs_variance = rs_std_dev ** 2
108 ss_variance = ss_std_dev ** 2
109 ns_variance = ns_std_dev ** 2
110
111 # Degrees of freedom for each group
112 df_rs = rs_count - 1
113 df_ss = ss_count - 1
114 df_ns = ns_count - 1
115
116 # Pooled standard deviation calculation
117 pooled_variance = (
118     (df_rs * rs_variance) +
119     (df_ss * ss_variance) +
120     (df_ns * ns_variance)
121 ) / (n - k) # n - k is the total degrees of freedom
122
123 s_pooled = math.sqrt(pooled_variance)
124
125 print(f"Pooled Standard Deviation (s_pooled): {s_pooled}")
126
127 SS_within = (
128     (rs_count - 1) * rs_variance +
129     (ss_count - 1) * ss_variance +
130     (ns_count - 1) * ns_variance
131 )
132

```

```

133 print((rs_count - 1) * rs_variance)
134 print((ss_count - 1) * ss_variance)
135 print((ns_count - 1) * ns_variance)
136
137 SS_within
138
139 """Now, as we have several column and their means, we can also try ANOVE one side test, for taht
    we'll also need to calculate SSR"""
140
141 # calculate SS_between
142 SS_between = rs_count * (rs_mean - overall_mean)**2 + ss_count * (ss_mean - overall_mean)**2 +
    ns_count * (ns_mean - overall_mean)**2
143 print("SS_between:", SS_between)
144
145 """Step 4: Calculate SST."""
146
147 SS_total = SS_between + SS_within
148 print("SS_toatl:", SS_total)
149
150 """Step 5: Fill in the ANOVA table."""
151
152 # the actual ANOVA one way table
153 from scipy.stats import f
154
155 k = df.shape[1] # number of groups
156 n = rs_count + ss_count + ns_count # total number of observations
157
158 # degrees of freedom
159 df_between = k - 1
160 df_within = n - k
161 df_total = n - 1
162
163 # mean squares
164 MS_between = SS_between / df_between
165 MS_within = SS_within / df_within
166
167 # F-statistic
168 F = MS_between / MS_within
169
170 # p-value is calculated using approximation, as we are not using scipy.stats
171 # yet in a more accurate way, we'd use a statistical table for accurate p-value.
172
173 p_value = f.sf(F, df_between, df_within)
174
175 # ANOVA table as a data frame
176 anova_table = pd.DataFrame({
177     'Source': ['Between Groups', 'Within Groups', 'Total'],
178     'Sum of Squares': [SS_between, SS_within, SS_total],
179     'Degrees of Freedom': [df_between, df_within, df_total],
180     'Mean Square': [MS_between, MS_within, ''], # Mean Square for Total is not defined
181     'F-statistic': [F, '', ''],
182     'p-value': [p_value, '', '']
183 })
184
185
186 anova_table
187
188 if p_value < 0.05:
189     print("The differences between the groups are statistically significant.")
190 else:
191     print("The differences between the groups are not statistically significant.")
192
193 """### Maria
194
195 Normality check:
196 1) Quantile plot
197 2) Shapiro test
198 """
199
200 import matplotlib.pyplot as plt
201 import scipy.stats as stats

```

```

202
203 # Create Q-Q plots for resistant and susceptible groups
204 fig, axes = plt.subplots(1, 2, figsize=(12, 5))
205
206 # Q-Q plot for resistant group
207 stats.probplot(RS, dist="norm", plot=axes[0])
208 axes[0].set_title("Q-Q Plot: Resistant Group")
209
210 # Q-Q plot for susceptible group
211 stats.probplot(SS, dist="norm", plot=axes[1])
212 axes[1].set_title("Q-Q Plot: Susceptible Group")
213
214 plt.tight_layout()
215 plt.show()
216
217
218 from scipy.stats import shapiro, levene, ttest_ind
219
220 # Check for normality using Shapiro-Wilk test
221 normality_resistant = shapiro(RS)
222 normality_susceptible = shapiro(SS)
223
224 normality_resistant, normality_susceptible
225
226
227 diff_means = rs_mean - ss_mean #Difference of the means
228 SE_diff = math.sqrt((rs_std_dev**2 / rs_count) + (ss_std_dev**2 / ss_count)) #standard error of
    the difference of the means
229 t_s = (diff_means-0)/SE_diff #t statistics
230 print("The SE of the difference is: ", str(SE_diff), " and the t statistic is: ", str(t_s))
231
232 #Calculate the degrees of freedom
233 SE_rs = rs_std_dev / math.sqrt(rs_count)
234 SE_ss = ss_std_dev / math.sqrt(ss_count)
235 degFreedom = (SE_rs**2 + SE_ss**2)**2/(SE_rs**4/(rs_count-1)+SE_ss**4/(ss_count-1))
236
237 #Calculate p-value
238 p = stats.t.sf(np.abs(t_s), degFreedom)
239 p = p*2
240 print("The p value is: ", str(p))
241
242
243 from statsmodels.stats.multicomp import pairwise_tukeyhsd
244 import matplotlib.pyplot as plt
245
246 data_melted = df.melt(var_name='Group', value_name='Value')
247
248 tukey_test = pairwise_tukeyhsd(endog=data_melted['Value'], # Numerical values
    groups=data_melted['Group'], # Groups
    alpha=0.05)
249
250
251
252 print(tukey_test.summary())
253
254
255 fig = tukey_test.plot_simultaneous()
256 fig.set_size_inches(10, 3)
257 plt.title('Tukey HSD Test - Group Comparison')
258 plt.show()
259
260 from statsmodels.stats.multicomp import pairwise_tukeyhsd
261
262 import math
263 n_per_group = 25 # Assuming 25 observations per group
264
265 # 1. Calculate the Standard Error (SE)
266 SE = math.sqrt(MS_within / n_per_group)
267
268 # 2. Critical value q (approximate value from Studentized Range Distribution table)
269 q_value = 3.5 # value for df_between=2, df_within=72, and alpha=0.05
270
271 # 3. Calculate HSD

```

```

272 HSD = q_value * SE
273
274 # 4. Display HSD
275 print(f"Calculated HSD: {HSD}")
276
277 # Calculate differences between group means
278 diff_NS_RS = abs(ns_mean - rs_mean)
279 diff_NS_SS = abs(ns_mean - ss_mean)
280 diff_RS_SS = abs(rs_mean - ss_mean)
281
282 # Compare differences to HSD
283 print(f"Difference between NS and RS: {diff_NS_RS} - {'Significant' if diff_NS_RS > HSD else 'Not
    Significant'}")
284 print(f"Difference between NS and SS: {diff_NS_SS} - {'Significant' if diff_NS_SS > HSD else 'Not
    Significant'}")
285 print(f"Difference between RS and SS: {diff_RS_SS:.3f} - {'Significant' if diff_RS_SS > HSD else '
    Not Significant'}")

```

Listing 2: All code