

## DAT 475 5-1 Project Two

In our previous analysis, we determined the production lines with the highest percentage of defects now we will analyze and compare these three lines specifically to each other. We will be looking at the line with the highest percentage of defects to determine if there is a statistically significant difference in the percentage of defects compared to the other two production lines. If the hypothesis test indicates there are statistically significant differences, this will help to justify to stakeholders that the lines with the highest number of defects should have those defects corrected. There are several steps needed to create a valid hypothesis test, we will go into detail for each of these steps below as it applies to this scenario.

### Creating a hypothesis test

1. **Formulate the Hypotheses:** Given the scenario, we can establish our null and alternative hypothesis as:
  - Null Hypothesis ( $H_0$ ): The production line for model 1 has no statistically significant difference in the percentage of defects compared to the other two production lines.
  - Alternative Hypothesis ( $H_1$ ): The production line for model 1 has a statistically significant difference in the percentage of defects compared to the other two production lines.
2. **Choose the Level of Significance:** This is also called alpha ( $\alpha$ ). Commonly used levels are 0.05 or 0.01. This is the probability of rejecting the null hypothesis when it is true. In other words, it is the risk of making a Type I error.
3. **Choose the Test Statistic:** Since we are comparing proportions from different production lines, a suitable statistical test would be the chi-square test for independence (if categories are nominal) or ANOVA (Analysis of Variance) if categories are ordinal or interval. You may need to ensure the data meets the assumptions for the test you choose.

4. **Collect Data:** In this case, we will need to gather data on the percentage of defects from each production line.
5. **Calculate the Test Statistic and Corresponding P-Value:** The test statistic will indicate how much the sample data deviate from what would be expected under the null hypothesis. The P-value is the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true.
6. **Draw a Conclusion:** If the P-value is less than or equal to the level of significance, you reject the null hypothesis in favor of the alternative. This would mean there is a statistically significant difference in the percentage of defects among the production lines. If the P-value is greater than the level of significance, you fail to reject the null hypothesis. This means that there's not enough evidence to suggest a statistically significant difference in the percentage of defects.
7. **Communicate the Results:** Whether you reject or fail to reject the null hypothesis, it is crucial to communicate your findings clearly and in context. This step is especially important if you need to justify changes to stakeholders. Make sure to explain your methods, your results, and what those results mean in a practical sense for the production lines.

For this analysis, I used IBM SPSS for a one-way ANOVA comparison of the percentage of defects between model 1, model 2, and model 3. I started by loading the data into the interpreter pictured below. For SPSS I found it was necessary to modify the Model column from string to numeric for it to be selectable for one-way ANOVA though while the variable is now numeric, it should still be treated as nominal, not scale, because these numbers represent categories (different production lines) rather than quantities on a scale. I did this in the variable view tab where you can manually modify the data types and I deleted the string model from the cells. Next, you can perform ANOVA by opening the analyze tab and selecting compare means. From there you can access One-Way ANOVA. I found the analysis to have much more meaningful results when the Model is selected as the independent variable and the Percentage is selected as the dependent variable since you're interested in the difference in defect percentages across different production lines.

SPSS Statistics Data Editor window showing a dataset with 16 rows and 10 columns. The columns are labeled: Model, Defects, Percentage, and five unlabeled 'var' columns. The data is as follows:

	Model	Defects	Percentage	var	var	var	var	var	var
1	1	10	30.00						
2	1	11	14.00						
3	1	12	11.50						
4	1	13	8.00						
5	1	14	5.00						
6	2	10	6.67						
7	2	11	3.11						
8	2	12	2.56						
9	2	13	1.78						
10	2	14	1.11						
11	3	10	7.23						
12	3	11	3.37						
13	3	12	2.77						
14	3	13	1.93						
15	3	14	1.20						
16									

The bottom of the window shows tabs for 'Data View' and 'Variable View', with 'Data View' currently selected.

SPSS Statistics Data Editor window with the 'One-Way ANOVA' dialog box open. The dialog box shows 'Defects' in the Factor list and 'Percentage' in the Dependent List. The Factor is set to 'Model'.

One-Way ANOVA Dialog Box:

- Dependent List: Percentage
- Factor: Model
- Buttons: OK, Paste, Reset, Cancel, Help, Contrasts..., Post Hoc..., Options...

The background data table is the same as in the first image.

## Output Analysis

The screenshot displays the IBM SPSS Statistics Viewer interface. On the left, a tree view shows the output structure: Output > Log > Oneway > Title > Notes > Descriptives > ANOVA > Post Hoc Tests > Title > Multiple Corr > Homogeneous > Title > Percent. The main window shows the 'Oneway' results for 'Percentage'.

### Descriptives

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	5	13.7000	9.73139	4.35201	1.6169	25.7831	5.00	30.00
2	5	3.0460	2.16359	.96759	.3595	5.7325	1.11	6.67
3	5	3.3000	2.34615	1.04923	.3869	6.2131	1.20	7.23
Total	15	6.6820	7.50760	1.93845	2.5244	10.8396	1.11	30.00

### ANOVA

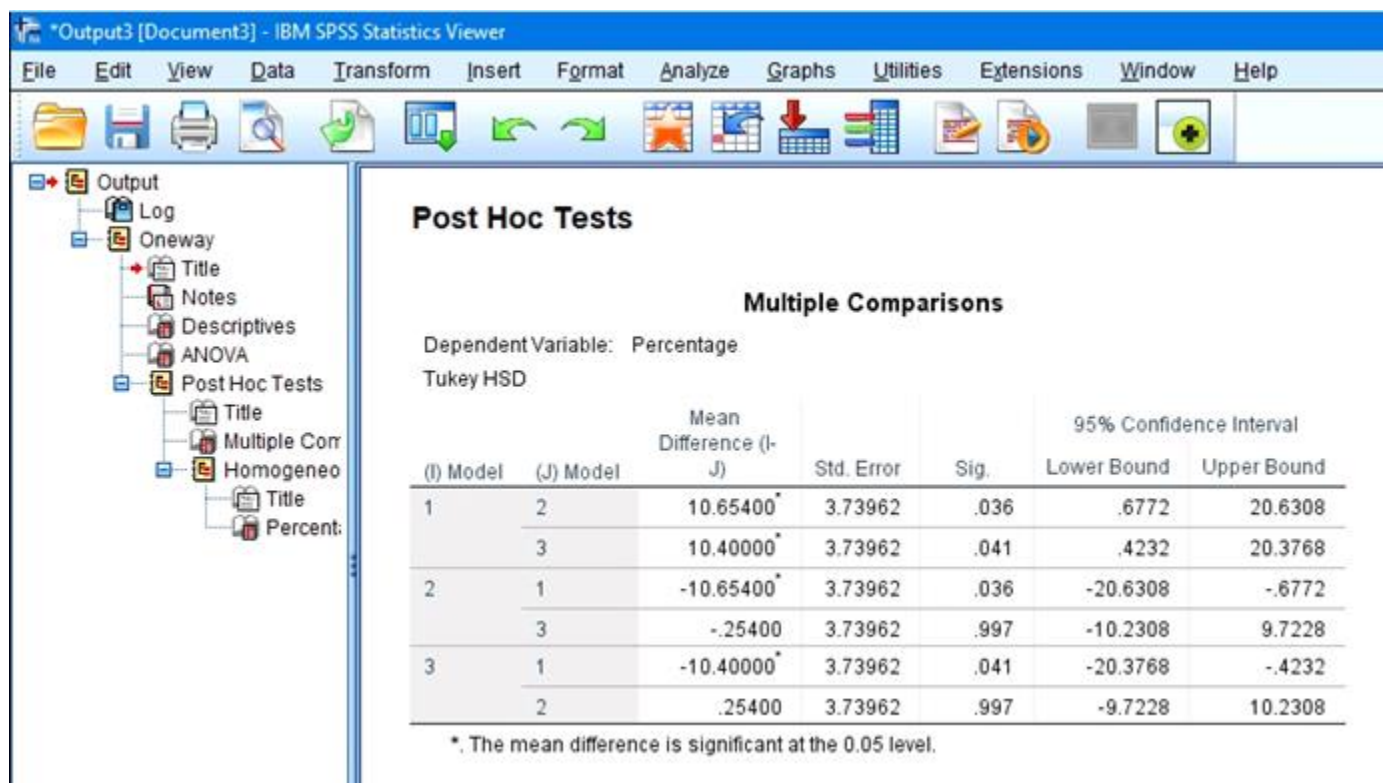
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	369.554	2	184.777	5.285	.023
Within Groups	419.542	12	34.962		
Total	789.096	14			

Above we have the results of the one-way analysis I have summarized the results below.

- Sum of Squares:** The "between-groups" sum of squares is 369.554, which indicates there's some variability in the 'Percentage' means across the different 'Models'. The "within-groups" sum of squares is 419.542, showing that there's variability within the 'Percentage' measurements within each 'Model'. The "total" sum of squares is the sum of the "between" and "within" sums of squares.
- Degrees of Freedom (df):** The degrees of freedom (df) for between-groups is 2 (which is the number of groups minus one), and for within-groups is 12 (which is the total number of observations minus the number of groups), both of which are as expected.
- Mean Square:** This is the sum of squares divided by the degrees of freedom. Here, the between-groups mean square is 184.777 (369.554 divided by 2) and the within-groups mean square is 34.962 (419.542 divided by 12).

4. **F-Statistic:** The F-value (5.285) is the ratio of the between-groups mean square to the within-groups mean square. A larger F-value indicates greater difference between the group means relative to the variation within the groups.
5. **Sig (p-value):** The p-value is 0.023, which is less than 0.05, the common significance level. This means that the difference in 'Percentage' between the 'Models' is statistically significant at the 0.05 level. You would reject the null hypothesis and conclude that at least one 'Model' has a different mean 'Percentage' of defects than at least one of the others.

In summary, this ANOVA result suggests that there is a statistically significant difference in defect percentages among the three models, given the p-value of 0.023. We have also performed a Tucky post hoc test which can provide pairwise comparisons between the models. The results are pictured below.



**Post Hoc Tests**

Dependent Variable: Percentage  
Tukey HSD

(I) Model	(J) Model	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	10.65400*	3.73962	.036	.6772	20.6308
	3	10.40000*	3.73962	.041	.4232	20.3768
2	1	-10.65400*	3.73962	.036	-20.6308	-.6772
	3	-.25400	3.73962	.997	-10.2308	9.7228
3	1	-10.40000*	3.73962	.041	-20.3768	-.4232
	2	.25400	3.73962	.997	-9.7228	10.2308

\*, The mean difference is significant at the 0.05 level.

- Model 1 vs Model 2: The mean difference is 10.654 (Model 1 has higher defect percentage), and this is statistically significant ( $p = 0.036$ , which is less than 0.05). The 95% confidence interval does not include 0 (0.6772 to 20.6308), which confirms the statistical significance.

- Model 1 vs Model 3: The mean difference is 10.4 (Model 1 has higher defect percentage), and this is also statistically significant ( $p = 0.041$ , which is less than 0.05). The 95% confidence interval does not include 0 (0.4232 to 20.3768), which again confirms the statistical significance.
- Model 2 vs Model 3: The mean difference is -0.254 (Model 3 has slightly higher defect percentage), but this difference is not statistically significant ( $p = 0.997$ , which is much greater than 0.05). The 95% confidence interval does include 0 (-10.2308 to 9.7228), which confirms the lack of statistical significance.

So, based on this post-hoc test, you can conclude that Model 1 has a statistically significantly higher defect percentage compared to both Model 2 and Model 3, while there is no statistically significant difference in defect percentage between Model 2 and Model 3. Given the results of these tests we will reject the null hypothesis, this means that there is a statistically significant difference in the percentage of defects on the production line with the highest percentage of defects compared to the other two lines. This would justify further investigation and potentially correcting the defects on this line.