**ADVANCED STATISTICS FOR ANALYTICS**

**Final Project**

“Statistical Analysis for Process Optimization, Quality Control, and Production Forecasting in Manufacturing”

Professor: Samira Ghorbanpour

Student: Ricardo Martinez

August 2025

**INTRODUCTION**

Manufacturing a product involves more than simply ensuring it works. It also requires producing it efficiently, meeting established quality standards, and preparing to meet future demand. This project follows a simulated but realistic example of a company producing a specialized polymer used in different industries.

The analysis was carried out in R and structured in three stages, each addressing a different aspect of the manufacturing process:

* **Stage 1:** Focuses on improving the process by using a ***Two-Way ANOVA*** to study how different *catalysts and temperatures* affect the amount of final product we can obtain.
* **Stage 2:** Uses a ***Chi-Square test*** to find out if the production shift has an impact on the types and frequency of defects found during trial runs.
* **Stage 3:** Uses ***time series forecasting*** to predict monthly production levels, helping the company prepare for changes in demand, plan raw material orders, and manage resources efficiently.

Together, these three stages provide a clear example of how data and statistical methods can support decision making, from optimizing processes in the lab to ensuring product quality and planning production for the future.

**Stage 1 – Process Optimization Study**  
*(Two-Way ANOVA)*

A manufacturing company is working to improve the production process for a high-performance material used in industries such as automotive and aerospace. The process involves combining specific raw materials in the presence of a catalyst, a substance that speeds up the chemical reaction.

The laboratory team tested three catalysts, **A, B, and C**, at three different temperatures: **150 °C, 170 °C , and 190 °C.** Each catalyst-temperature combination was tested through multiple production runs under identical operating conditions to ensure fair and reliable comparisons.

The yield percentage expressed as:

A black text on a white background

AI-generated content may be incorrect.

The yield percentage was measured to determine whether catalyst type, reaction temperature, or the interaction between the two had a significant effect on production efficiency.

To determine which combination produced the highest yield, the analysis was carried out in R using the following steps:

1. **Stated the hypotheses:** Defined the null and alternative hypotheses for the effects of catalyst type, temperature, and their interaction.
2. **Performed a Two-Way ANOVA:** Ran the statistical test and presented the results, including F-values and p-values.
3. **Interpreted the results:** Identified which factors had statistically significant effects on yield.
4. **Performed a Tukey’s test.**
5. **Visualized the interaction:** Created an interaction plot to illustrate how catalyst and temperature together affected yield.
6. **Compared mean yields:** Generated bar plots to compare average yields:

a) by catalyst type,  
b) by temperature, and  
c) grouped by both factors.

The analysis performed is shown below.

1. **State the Hypothesis:**

* **H₀A:** The mean yield is equal for all catalyst types (A, B, C).
* **H₀B:** The mean yield is equal for all temperature levels (150 °C, 170 °C, 190 °C).
* **H₀AB:** There is no interaction between catalyst type and temperature; the mean yield for each catalyst is the same regardless of temperature.
* **Hₐ:** At least one group mean is different.

1. **Perform a Two-Way ANOVA**

**CODE**

A blue background with white text

AI-generated content may be incorrect.

**RESULTS**

A screenshot of a computer screen

AI-generated content may be incorrect.

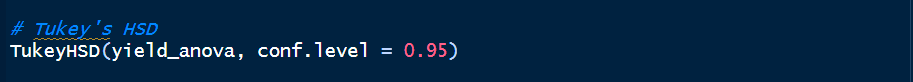
1. **Interpretation of the results**

* The p-value for Catalyst is < 2e-16, which is far below the 0.05 significance level. Therefore, the Null hypothesis **H₀A** is rejected and conclude that the mean yield differs significantly between the three catalyst types (A, B, C).
* The p-value for Temperature is also < 2e-16, Therefore, we reject **H0B**​ and conclude that at least one temperature level (150°C, 170°C, 190°C) produces a different mean yield from the others.
* For the Interaction Catalyst-Temperature The p-value is 0.12, which is greater than 0.05. This means we fail to reject **H0AB**​, suggesting that the effect of one factor on yield does not depend significantly on the level of the other factor.
* All the above implies that the optimal choice of catalyst and temperature can be determined independently.

1. **Perform a Tukey’s test.**

Since the null hypotheses for Catalyst and Temperature were rejected in the Two-Way ANOVA, a Tukey’s HSD test was performed to determine which specific group means differ significantly. As the interaction term was not significant, the post-hoc analysis was carried out only for the main effects of Catalyst and Temperature.

**CODE**



**RESULTS**

A computer screen with numbers and letters

AI-generated content may be incorrect.

The Tukey HSD results show that all catalyst types differ significantly from each other, with Catalyst C producing the highest yields, followed by B, then A. All temperature levels also differ significantly from each other, with yield increasing from Low to Medium to High temperature.

1. **Visualize the interaction**

**CODE**

A screen shot of a computer code

AI-generated content may be incorrect.

**PLOT**

A graph of a graph

AI-generated content may be incorrect.

Figure 1: Interaction plot showing the effect of Catalyst type and Temperature on yield (%)

**Interpretation of Interaction Plot**

* The interaction plot shows the effect of **Catalyst type** and **Temperature level** on yield percentage. The lines for different temperatures are approximately parallel which confirms that there is no strong interaction between Catalyst and Temperature.
* It is also noticeable that at all temperatures, catalyst C consistently produces better yields.
* Another observation is that higher temperatures produce better yields for all three catalysts.
* From the above then we conclude that Catalyst C at High temperature achieves the highest yield.

1. **Compare mean yields**
   1. by catalyst,
   2. by temperature,
   3. grouped by both factors.

**CODE**

**A computer screen shot of text

AI-generated content may be incorrect.**

**RESULTS**

A graph showing a number of blue rectangular objects

AI-generated content may be incorrect.

Figure 2: Mean yield (%) by Catalyst type with error bars representing 95% confidence intervals

Interpretation: The differences between catalysts are visually noticeable, suggesting that catalyst type has a strong effect on yield. This agrees with the results from ANOVA.

A screenshot of a graph

AI-generated content may be incorrect.

Mean yield (%) by Temperature level with error bars representing 95% confidence intervals

Interpretation: High temperature (190 °C) consistently leads to higher yields. The gap between High and Low is substantial, reinforcing the ANOVA.

A graph of different colored bars

AI-generated content may be incorrect.

Figure 4: Grouped bar plot showing the combined effect of Catalyst type and Temperature on yield (%)

|  |
| --- |
|  |

The grouped bar plot shows the combined effect of Catalyst and Temperature on yield. Catalyst C performs best across all temperatures. It is also noticeable that the differences between catalysts remain relatively consistent at each temperature level, which supports ANOVA results that there is no strong interaction between Catalyst and Temperature.

**CONCLUSION FOR STAGE 1**

The Two-Way ANOVA results indicated that both **Catalyst type** and **Temperature** have a statistically significant effect on production yield, but their interaction is not statistically significant. In other words, the choice of catalyst and the operating temperature independently influence yield, but the effect of one does not depend strongly on the other.

Based on these findings, the optimal condition for maximizing yield is the combination of Catalyst C at High temperature (190 °C).

**Stage 2 – Quality Control Assessment***(Chi-Square Test)*

Following the process optimization trials from Stage 1, the manufacturing company proceeded with initial production runs of the high-performance material. During this phase, the Quality Control (QC) team monitored product quality closely.

Over a one-month period, the team recorded the types of defects found in products from each of the three production shifts: **Morning, Afternoon, and Night.** Every product was inspected and classified into one of four categories: **Surface Defect, Color Issue, Strength Failure, or No Defect.**

The results for each combination of production shift and defect type were summarized in a contingency table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Shift** | **Color Issue** | **No Defect** | **Strength Failure** | **Surface Defect** |
| Morning | 10 | 70 | 5 | 15 |
| Afternoon | 15 | 50 | 10 | 25 |
| Night | 25 | 25 | 20 | 30 |

To determine whether defect type is related to production shift, the analysis was carried out in R using the following steps:

1. **Stated the hypotheses** for the Chi-Square Test of Independence.
2. **Created the contingency table** showing the frequency of each defect type by production shift.
3. **Performed the Chi-Square Test of Independence** using the contingency table.
4. **Reported the Chi-Square statistic,** degrees of freedom, and p-value.
5. **Made the decision** to reject or fail to reject the null hypothesis based on the p-value.
6. **Calculated and examine the standardized residuals** to identify which shift–defect combinations contribute most to the Chi-Square statistic.
7. **Created and interpreted an association plot** to visually assess the relationship between defect type and production shift.
8. **Generated a mosaic plot with shading** and interpreted what it indicates about the relationship between variables.
9. **Created a grouped bar plot** comparing the observed and expected counts for each defect type by production shift. Comment on any patterns or deviations observed.

The analysis performed is shown below.

1. **State the hypotheses for the Chi-Square Test of Independence**

H0: Defect type is independent of production shift.

Ha: Defect type is not independent of production shift.

1. **Construct the contingency table**

**CODE**

A blue background with white text

AI-generated content may be incorrect.

1. **Perform the Chi-Square Test of Independence using the contingency table.**

**CODE**

A blue screen with white text

AI-generated content may be incorrect.

1. **Report the Chi-Square statistic, degrees of freedom, and p-value.**

A blue background with white text

AI-generated content may be incorrect.

1. **Based on the p-value, decide whether to reject or fail to reject the null hypothesis.**

Since the p-value is less than 0.05, we reject the null hypothesis, and we can conclude that there is enough evidence to conclude that defect type is related to production shift.

1. **Obtain and examine the standardized residuals.**

**CODE**

A blue and white text on a dark background

AI-generated content may be incorrect.

**RESULTS**

A blue screen with white text

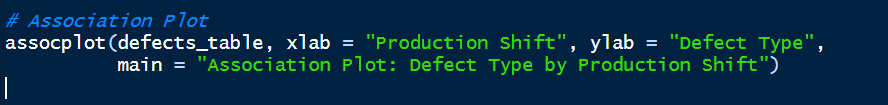
AI-generated content may be incorrect.

**Iterpretation of residuals:**

* **Morning-No Defect** with a residual value of 3.116 has a strong positive residual, meaning the morning shift produced significantly more “No Defect” items than expected.
* **Night-No Defect** with a residual value of -3.356 has a strong negative residual, meaning the night shift produced significantly fewer “No Defect” items than expected.
* **Night-Color Issue and Night-Strength Failure** with values of 2.04 and 2.44 respectively, Both are positive and above 2, indicating these defects occurred more often than expected during the night shift.
* **Morning-Strength Failure** and **Morning-Surface Defect with values of -1.95 and -1.73 respectively, have m**oderately negative, suggesting these defects were less frequent than expected during the morning shift.
* All the above suggests that product quality is highest in the morning shift and lowest in the night shift.

1. **Create and interpret an association plot**

**CODE**



**PLOT**

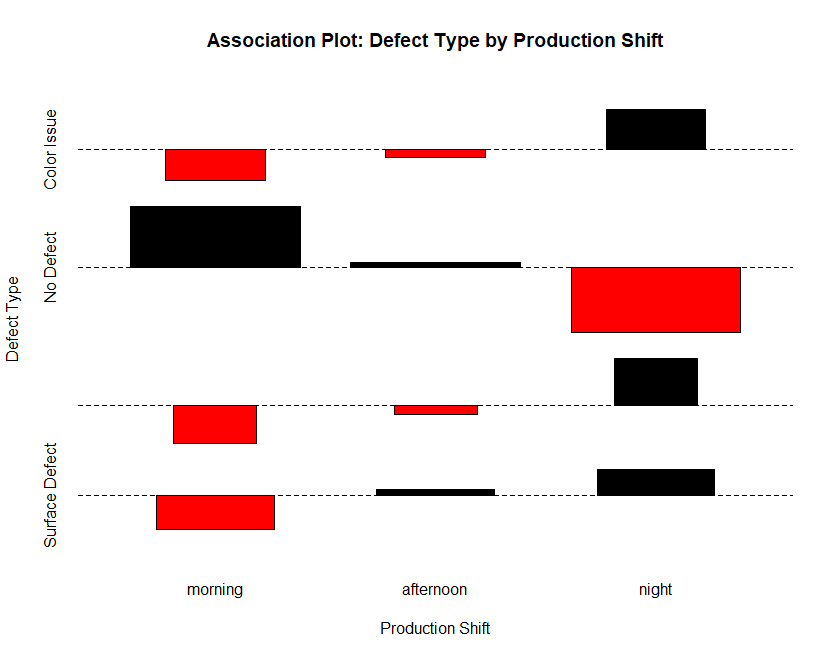
****

Figure 5: Association plot of defect type by production shift

**Interpretation of Association Plot:**

* **During the Morning Shifts** “No Defect” occurs more frequently than expected, and all other defects are less frequent than expected. These two observations from the association plot suggest a strong performance during that shift.
* **During the Afternoon shift** most categories are close to the expected frequency (bars are small), indicating no strong deviations.
* **During Night shift** “No defect” is less frequent than expected and all other defects occur more often than expected, highlighting potential quality control issues in the night shift.

1. **Produce a mosaic plot with shading and interpret what it indicated about the relationship between variables.**

**CODE**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**PLOT**

**A mosaic plot with different colored squares

AI-generated content may be incorrect.**

Figure 6: Mosaic plot of defect type by production shift with shading for residual significance

**Interpretation of Mosaic Plot:**

The plot suggests that production quality varies significantly by shift, with the morning shift showing fewer defects than expected, and the night shift showing more defects, especially in certain categories. This aligns with the Chi-Square results indicating a significant association between production shift and defect type.

1. **Create a grouped bar plot comparing the observed and expected counts for each defect type by production shift. Comment on any patterns or deviations observed.**

**A computer screen shot of a code

AI-generated content may be incorrect.**

**PLOT**

**A graph of different colored squares

AI-generated content may be incorrect.**

Figure 7: Grouped bar plot comparing observed and expected defect counts by production shift

**Interpretation of Grouped Bar Plot**

Overall, the most prominent deviations occur in the No Defect category, where the morning shift outperforms expectations and the night shift underperforms. These results support the Chi-Square test finding of a significant association between shift and defect type, with quality performance varying by shift.

**CONCLUSIONS FOR STAGE 2**

* Based on the Chi-Square Test for Independence the distribution of defects is not uniform across shifts, and certain shifts tend to have higher or lower frequencies of specific defect types.
* Residual analysis and visualization tools suggest that the morning shift has the stronger quality performance. In contrast night shift is the one that presents more potential quality challenges. Also, the afternoon shift’s defect distribution is generally closer to expected values, showing no major deviations.
* The quality team should investigate operational, staffing, or process factors that may be influencing defect rates, particularly during the night shift.

**Stage 3 – Production Forecasting***(Time Series Analysis)*

Following the process optimization in Stage 1 and the quality control assessment in Stage 2, the company began tracking monthly production output over several consecutive years. The data, measured in metric tons of the final product, provides a basis for understanding production patterns over time.

The production planning team understands that recognizing these patterns is essential for anticipating future needs and ensuring efficient operations. For this purpose, they decided to use time series analysis, applying the *moving average method*, to generate forecasts for future monthly output.

Performing this analysis will provide the production planning team with actionable forecasts to guide decision-making, and this stage focuses on generating those forecasts using the available production data.

This stage aims to produce forecasts that can support better operational planning and resource management. The analysis was carried out in R following these steps:

1. **Created a time series object** for the production data and plot.
2. **Decomposed** the time series into its trend, seasonal, and random components and **provided** a brief commentary.
3. Fitted and compared three moving average models using window sizes of 3 months, 6 months, and 12 months.
4. **Plotted each model** and described which best captured the underlying trend.
5. **Calculated** the following error metrics for each model:
6. Mean Squared Error (MSE)
7. Mean Absolute Deviation (MAD)
8. Mean Absolute Percentage Error (MAPE)
9. **Identified** the model that best fit the data based on the error metrics and explained the reasoning behind the selection.

The analysis performed is shown below.

* 1. **Create a time series object for the production data.**

**CODE**

A computer screen shot of a program

AI-generated content may be incorrect.

**PLOT**

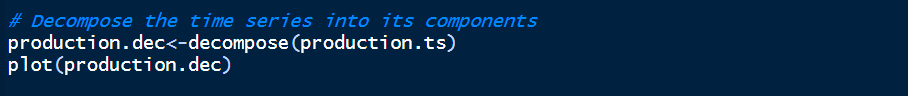
A graph showing the growth of a company

AI-generated content may be incorrect.

Figure 8: Time series plot of monthly production output (tons)

* 1. **Decompose the time series into its components and briefly analyze each.**

**CODE**



**PLOT**

**A graph of different types of lines

AI-generated content may be incorrect.**

Figure 9: Decomposition of the time series into trend, seasonal, and random components

* The trend component shows a steady upward pattern in production from 2019 to 2023.
* The seasonal component repeats annually, with peaks around mid-year and lows near the start of each year.
* The random component fluctuates moderately, indicating some month-to-month variation not explained by the trend or seasonality.
  1. **Fit and compare three moving average models using window sizes of 3 months, 6 months, and 12 months.**

A screenshot of a computer program

AI-generated content may be incorrect.**CODE**

* 1. **Plot each model and comment on which best captures the underlying trend.**

A graph with red lines and numbers

AI-generated content may be incorrect.

Figure 10: Monthly production output with 3-month moving average overlay

This model follows the short-term fluctuations closely, making it quick to reflect recent changes in production.

A graph with red lines and numbers

AI-generated content may be incorrect.

Figure 11: Monthly production output with 6-month moving average overlay

This model smooths the data more, reducing the impact of small monthly changes while still showing the main seasonal pattern and the overall trend.

A graph with lines and numbers

AI-generated content may be incorrect.

Figure 12: Monthly production output with 12-month moving average overlay

This model produces a very smooth curve that highlights the long-term trend. Most of the seasonal pattern is removed, which can make it harder to see short-term increases or decreases in production.

* 1. **For each model, calculate the following error metrics:**
     1. Mean Squared Error (MSE)
     2. Mean Absolute Deviation (MAD)
     3. Mean Absolute Percentage Error (MAPE)

**A computer screen shot of text

AI-generated content may be incorrect.CODE**

**RESULTS**

A computer screen shot of error

AI-generated content may be incorrect.

* 1. **Identify the model that best fits the data based on the error metrics and explain the reasoning behind the selection.**

Although MA-6 appears to follow the overall trend visually, its error metrics are the highest, which means it is less accurate. MA-12 produces an even smoother line that clearly shows the long-term trend, but it reacts slowly to actual changes in production and its three metrics are also higher than those of MA-3. The MA-3 moving average is the best fit, as it has the lowest error values among the three models, making it the most accurate option for forecasting in this case.

**CONCLUSION FOR STAGE 3**

The decomposition analysis confirmed a steady upward trend in production with a consistent seasonal pattern and moderate irregular variation. Based on the analysis of three moving average models with window sizes of 3, 6, and 12 months, the MA-3 model is the most suitable for short-term forecasting of monthly production output, providing timely and reliable information to support operational planning.

**FINAL CONCLUSION**

This project demonstrates how statistical techniques can be applied in an integrated way to support decision-making in a manufacturing environment. The use of Two-Way ANOVA in Stage 1 allowed the identification of process factors that significantly influence yield, guiding process optimization efforts. The Chi-Square Test in Stage 2 provided valuable insights into quality performance across shifts, helping target resources where they are most needed. Finally, time series forecasting in Stage 3 offered a practical tool for anticipating production levels, enabling better planning for materials, labor, and capacity. Together, these methods form a coherent analytical framework that can help a company move from reactive to proactive management, improving efficiency, product quality, and preparedness for future demand.