JMP Academic Case Study 037

Performance of Food Manufacturing Process – Part 1

Descriptive Statistics and Graphical Summaries

Produced by

Kevin Potcner, JMP Global Academic Team

kevin.potcner@jmp.com

jmp STATISTICAL DISCOVERY

Performance of Food Manufacturing Process – Part 1

Descriptive Statistics and Graphical Summaries

Key Ideas

This case study requires the use of descriptive statistics (e.g., mean, median, standard deviation) as well as graphical analysis techniques (histograms, box plots, time series) to uncover features in the data and to evaluate performance to specifications of a food manufacturing process.

Background

Constant monitoring of a manufacturing process is critical to ensure that a consistent product is produced and is able to meet the desired specifications. Food manufacturing is one such process where process quality is essential to ensure the consumer receives a consistent and safe product. Oakville Dairy is a food processing company that produces a variety of dairy-based products. A series of quality issues has recently surfaced in one of its yogurt products. In particular, some of the technicians in the testing lab have noticed rather excessive variation in the acidity levels (pH) between the batches of product. The Quality Engineering Team has been tasked with studying this through a more formal data collection and analysis process.

The Task

The facility runs eight separate production lines that make the yogurt. Ideally, each line should be producing product that is very similar between batches, as well as between the production lines. To get an overview of the current state of the process, the quality engineers take a sample of yogurt from each of the next 30 batches of product made by each of the eight production lines. The samples are taken to the lab and a variety of product characteristics including Acidity (pH) is measured.

The specifications for Acidity (pH) level for the product is: target = 4.35; lower specification limit (LSL) = 4.30; upper specification limit (USL) = 4.4; and a standard deviation ≤ 0.015.

They then analyze these data to describe the current state of the process and to determine if product specifications are being met.

The Data **Yogurt_1.jmp**

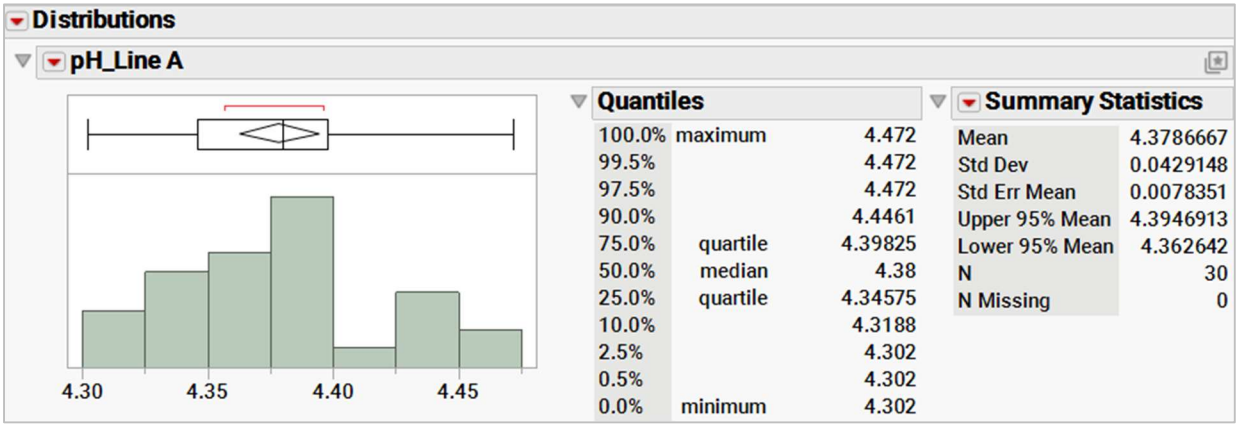
Batch	Batch number (1, … ,30) of the sample taken
pH Line A	Acidity (pH) of sample taken from each batch produced by Line A
pH Line B	Acidity (pH) of sample taken from each batch produced by Line B
pH Line C	Acidity (pH) of sample taken from each batch produced by Line C
pH Line D	Acidity (pH) of sample taken from each batch produced by Line D
pH Line E	Acidity (pH) of sample taken from each batch produced by Line E
pH Line F	Acidity (pH) of sample taken from each batch produced by Line F
pH Line G	Acidity (pH) of sample taken from each batch produced by Line G
pH Line H	Acidity (pH) of sample taken from each batch produced by Line H

Analysis

We will demonstrate the types of analyses that would be useful for assessing variation in pH by analyzing the data for Line A. The exercises at the end of the study will ask you to perform similar analyses for Lines B-H and to compare results across the eight lines.

We'll begin our analysis by summarizing the data graphically and numerically. Exhibit 1 displays a histogram, box plot and a variety of summary statistics.

Exhibit 1 Distribution



(To create, Analyze>Distribution. Select pH_Line A as the Y Variable.)

As can be seen in the summary statistics table, one of the 30 batches has a pH of 4.302 (minimum), which is within specification. One batch has a pH of 4.472 (maximum), which is beyond the USL of 4.40.

The sample mean is 4.379 and the sample standard deviation is .043. How close is this sample mean to the target of 4.35? In terms of pH, it is 4.379 – 4.350 = 0.029 pH units away. A standard way to statistically measure distance between two values is in terms of the number of standard deviations, so that the distance is expressed relative to the variation in the data. Here we would say that the sample mean of 4.379 is (4.379 – 4.350)/0.043 = 0.668 standard deviations higher than the target of 4.35. Less than one standard deviation is typically considered to not be very far apart.

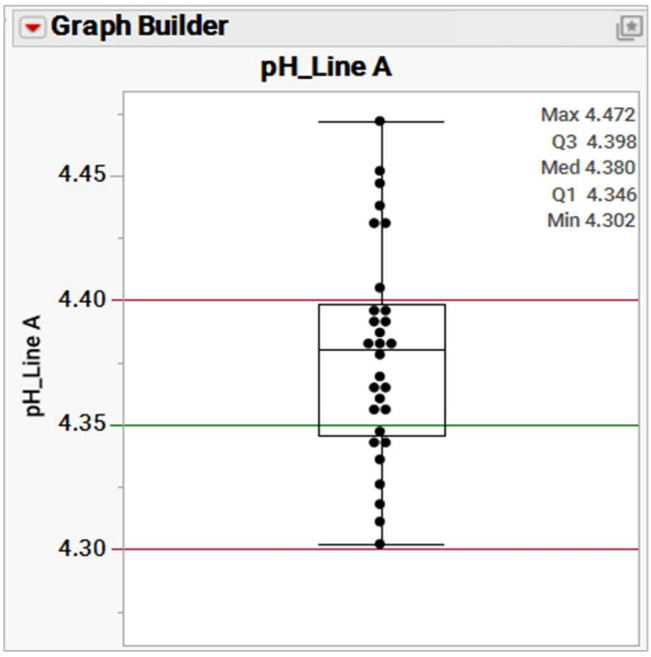
Since this process has specification limits (LSL = 4.30 and USL = 4.40), it is also quite valuable to measure distance in terms of the size of the specification window. Here we would say that the distance the sample mean of 4.379 is from target of 4.35 is |100*(4.379 – 4.350)/(4.400 – 4.300)| = 28.7% the width of the specification limits. We can so express this as the distance of the closest specification limit from the target. Using this method, we see that the sample of mean is |100*(4.379 – 4.350)/(4.400 – 4.350)| = 57.3% the distance to the USL from the target. This is undesirable as it means that the process is on average almost 60% the distance to the nearest specification from the target.

The median value of 4.38 tells us that half of the batches have a pH reading above 4.38 and half are below. Ideally, both the mean and the median would be very close to 4.35. It is important to realize that the mean and median are merely summary statistics based on a sample of 30 batches. More formal statistical techniques are used to determine if these data provide enough statistical evidence for us to conclude that the process in general is on average different than the target of 4.35.

The sample standard deviation of 0.043 is well beyond the desired specification of 0.015. It is 0.043/0.015 = 2.87 times larger. In addition, some of the batches have pH that exceed the USL of 4.40. Clearly, there is too much variation in the pH for this production line.

A very useful graph to display data and how they compare to specifications is a box plot, as shown in Exhibit 2. The box is constructed from five summary statistics: minimum, 25th quantile (Q1), median, 75th quantile (Q3), and maximum.

Exhibit 2 Box Plot

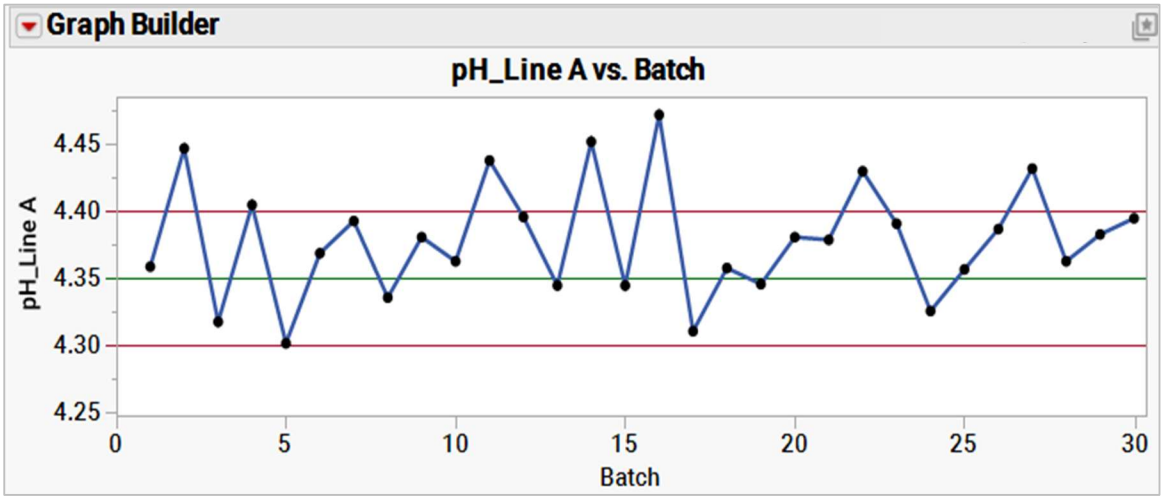


(To create, *Graph>Graph Builder. Select pH_Line A for Y. Select both the Points and Box Plot graphs in the palette at the top. To add reference lines, double click on the Y axis. Under Reference Lines, type in the value of the target 4.35. Choose a desired color and line style, and click Add. Do the same for the LSL 4.30 and USL 4.40. To display the summary statistics, select the 5 number summary check box.*)

This graph makes it clear that the pH values for the batches tend to be toward the higher values, with some batches having pH above the upper specification limit. We can easily see, for example, that seven of the 30 batches have pH that exceed the USL of 4.40. There are none that exceed the LSL, though a few are quite close (e.g., Batch 5 has a pH of 4.302). Clearly, this production line appears to be having problems with producing yogurt with a consistent level of pH that meets specification.

The batches are organized in the data table in the order they were produced. When data has a time order, such as the case here, it is very important to examine the data over time, as it will allow us to see a more descriptive view of the batch-to-batch variation and the overall performance of the process. Specifically, we will be able to examine if the average and variation in pH level is consistent over time or if there are distinct changes and patterns. Exhibit 3 is a time series plot of these data. Reference lines indicating the target of 4.35 and the LSL of 4.30 and USL of 4.40 have been included.

Exhibit 3 Time Series Graph



(To create, Graph>Graph Builder. Select pH_Line A for Y, Batch in X. Select both the Points and Line graphs in the palette at the top. To add reference lines, double click on the Y axis to bring up the Axis Settings. Under Reference Lines, type in the value of the target 4.35, choose a desired color and line style, and click Add. Do the same for the LSL 4.30 and USL 4.40.)

Though the average pH is above the target and the variation is too large, the performance seems to be relatively the same across the entire production time period (e.g., no sudden sustained shifts, trending upward or downward, cyclical patterns, outliers, etc.).

Summary

Statistical insights

We have learned that Line A does not meet the desired specification. Specifically, the sample standard deviation of 0.043 is much larger than the desired value of 0.015. The sample mean of 4.379 is quite a bit higher than the target of 4.35. In fact, it's 57% the distance to the USL from the target.

It is important to visualize data in a variety of ways. Here we used a box plot with individual values and a time series graph. One type of visualization may be best for illustrating a particular feature in the data, while another visualization is better to illustrate a different feature.

Implications

The Quality Engineering Team will need to study this production line to uncover sources that are causing the excessive variation in pH between the batches, as well as reasons the process is off target. Data was collected on the seven other production lines (B, C, D, E, F, G and H). In the following exercises, you will perform a similar set of analyses on these additional lines.

Exercises

Use the data in file **Yogurt_1.jmp** to perform the following analyses:

1. Generate summary statistics for each of the production lines. Create a table that displays the minimum and maximum values, the sample means and sample standard deviations, as well as how far the mean is from target using the methods that were illustrated for Line A. Provide a few observations regarding the performance of the production lines based on these numerical summaries.

2. Create comparative box plots for all eight production lines on one graph that also displays all the individual pH values for the 30 batches. Add reference lines for the target, LSL, and USL. Describe the performance of these production lines using this graphical display.

3. Create time series graphs for each production line, displaying all eight individual times series plots as separate panels in one graph. Include reference lines for the target, LSL, and USL.

   Hint: To create all the times series plots in one graph, the data will need to be stacked so that all pH values are in one column, another column identifies the batch, and a third column identifies the production line. Search for Stack Columns in the Help Menu to learn how to do this.

4. Which lines have consistent performance across the data collection time period? Do all those lines meet specifications? Are there any changes over time for any of the production lines that would indicate their performance is not consistent? If so, describe those patterns. Are there any unusual values (i.e., outliers)? Why would these outcomes be a cause for concern in using this sample of data to reach conclusions about the performance of the production lines in general?

5. Summarize the key results in just a few bullet points that can be communicated to management. Which one of the three displays created (summary table, comparative box plots, or time series plots) would you choose to show?