IDENTIFYING MANUFACTURING PROBLEMS WITH STATISTICS – A TUTORIAL USING SIMULATED DATA AND RAINCLOUD PLOTS

Hadi Akbar Dahlan

Leave a Nest Malaysia Sdn. Bhd., G-B, Block 2330, Century Square, Jalan Usahawan, Off Persiaran Multimedia, 63000 Cyberjaya, Selangor

Corresponding Author's Email: hadiukm@gmail.com

ABSTRACT: Engineers working in manufacturing will usually perform a routine statistical analysis to evaluate the performance of a production line. However, this article argues that such routine statistical analysis will only focus on conformity and not on developing problem-solving skills. Problem-solving skills involving statistics would lead to not only higher quality control during manufacturing processes, but also efficient problem diagnostics using manufacturing data. To encourage learning and gaining a deeper understanding of statistics, this article describes the application of raincloud plots and test of normality via a simulated case study. All analysis described in the simulated case study can be performed freely via open-source software and without programming knowledge.

KEYWORDS: *Raincloud plot; statistics; production line; engineering; case study.*

1.0 INTRODUCTION

Engineering and statistics are two inseparable fields. Engineering students will learn general statistics courses which include basic descriptive statistics and hypothesis testing [1, 2]. However, such statistical courses might be inadequate for real-life examples. One example of a limitation in teaching and performing engineering activities is about handling non-normal distribution data. Commonly, engineers in any field will be able to perform routine statistical analysis for the system in their respective fields. The routine statistical analysis usually includes descriptive statistics and variation of statistical control process practices. However, it is undeniable that such routine statistical analyses were analytical strategies based on predecessors and only focused on the conformity of the production line. An engineer must be equipped with more advanced statistical knowledge than the routine statistical analysis strategy in order to solve a problem that might occur outside the boundary of routine statistical analysis output.

One example is non-normal distribution data. Non-normal distribution occurs much more than normal distribution in actual engineering practices [3] One common strategy is to transform the non-normal data into a normal distribution with a variety of transformation strategies. However, does such a strategy teach the engineer about the system or this strategy had become an extended version of routine statistical analysis? This does not mean that the data transformation strategy is incorrect, but data transformation needs a deeper understanding of transformation strategy for correct execution [4]. Data transformation had been discouraged since it can lead to wrong interpretation of the actual data if precaution steps are not taken [5, 6].

The article intends to stress the importance of acquiring a deeper understanding of statistics outside the normal general introduction to statistics syllabus. There are anecdotes regarding engineering problems that had caused a detrimental loss in terms of human effort, resources and money due to the reason that the analyst does not realize that the data is non-normal in the first place and apply incorrect analysis to guide decision making [3]. Taking this into account, this article aims to showcase raincloud plots, a data visualization plot with minimal statistical interferences. Raincloud plot was developed by researchers in the field of neuroscience for the purposes of data transparency among researchers [7]. Although it is mainly utilized in basic science research, raincloud plots had also been utilized in applied science research. Examples of applied science research include detecting tempe sensory preferences among panelist with varied experiences to tempe [8] and identifying which working shift can contribute the most kinetic energy toward a wearable device [9]. This article would like to showcase that a raincloud plot could also help diagnose problems in engineering practices and promote its usage in any engineering or manufacturing-related industry.

2.0 METHODOLOGY

This article will utilize a simulated case study approach with simulated data obtained from a random number generator generated in Microsoft Excel. The simulated case study in this context means creating an artificial scenario to describe the problem and the solution to the reader. The simulated case study approach had been used in other engineering education such as in the design of experiments [10] and system analysis [11]. The statistical analysis described in the simulated case study was conducted using JASP software version 0.16 [12]. All simulated data and the JASP file had been uploaded to the Open Science Framework repository and can be accessed via this link (https://osf.io/2s9tu/) for data reproducibility and modification.

3.0 RESULTS AND DISCUSSION

3.1 Simulated case study description

You are working for Water bottle company A and responsible for the production line of 500 ml bottled water. The production parameter of the 500 ml bottled water is shown in Table 1. Since this production line is producing continuously, a range of defect limits had been implemented. The lower and upper defect limits of the 500 ml bottled water are 498 ml and 500 ml respectively. Any bottled water within the 2 standard deviations is to be considered within the product guideline. Alternatively, any bottled water over or under the defect limit will be removed in the subsequent quality control check.

Table 1: Production parameter of the 500 ml bottled water

Production parameter		
Average volume (ml)	500	
Product defect limit (2 standard deviation)		
Lower limit	498	
Upper limit	502	
No. of Batch production	Day	Night
Accepted no. of defects	5% from total bottled water	

Then, the company received complaints that several customers were receiving boxes of bottled water where each bottled water weight lesser than 500 ml. Theoretically, this could happen if the box contained all water bottles packaged that weigh lesser than the average volume. Probability for this to occur should be very low because bottled water was packaged randomly from one shift in a production line. However, there was more than one complaint. This could indicate a systemic problem in the production line. As a quality control officer, you are instructed to investigate whether does the customer complaint was a fluke or a systematic error within the production line does exist. Based on the complaints, you were able to detect the problematic batch. You decided to analyse the data on one of the problematic batches. However, you could not determine which production shifts were at fault. It is imperative to detect which production shift is at fault so that the quality control team can identify the further problem and apply corrective measures at the production line. You were given the production report for both production shifts (Day and Night shift) which include individual bottle weight, average production, standard deviation, and a statistical process chart.

3.2 Diagnosing the production line data

The production report is shown in Table 2. Both production batches produced a total of 360 bottled water for each shift. There were slight differences in the average volume of bottled water. However, the differences are within the limit of 2 standard deviations as specified in the product guideline.

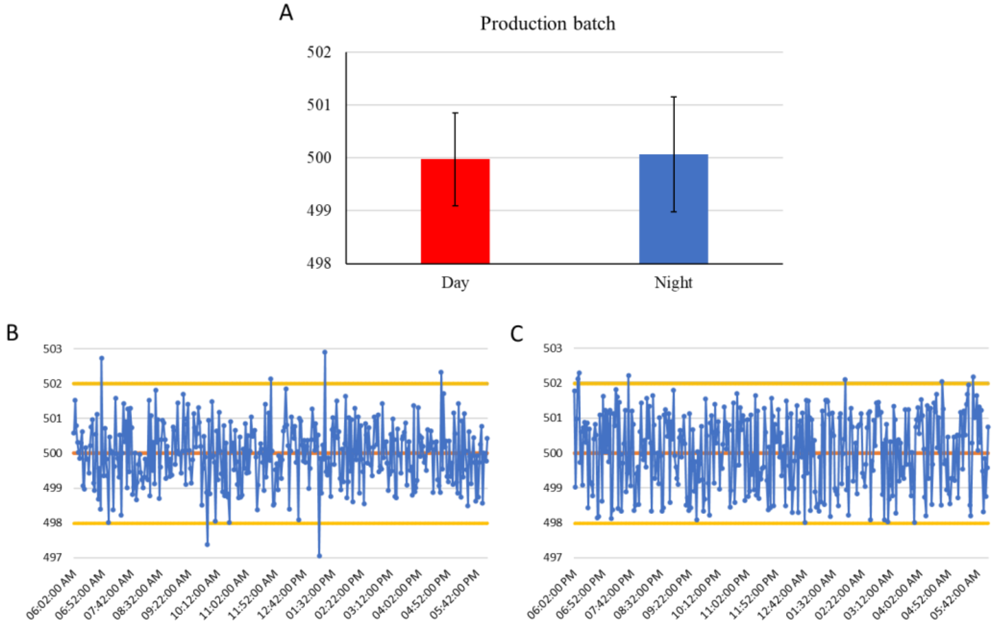
Table 2: Production report of the suspected problematic batch

Production batch	Day	Night
Number of bottled waters	360	360
Average volume (ml)	499.9776	500.0553
Standard deviation	0.879687	1.100428
Lower limit (ml)	498	
Upper limit (ml)	502	
No. of defects	6	6

From the information, you plotted a bar graph with standard deviation (Fig. 1A). However, the bar graph does not show any irregularities. The statistical process chart (SPC) for both day shift (Fig. 1B) and night shift (Fig. 1C) shows the individual bottled water volume throughout the production time. The statistical process chart only indicates the production time of bottled water defects. However, the SPC chart still could not answer whether is there a systematic problem or not in the production line.

After reviewing the production report, you were tempted to perform a statistical analysis. The routine analysis will usually involve performing the hypothesis test using an appropriate statistical test such as the Student's T-test or Analysis of Variance (ANOVA). However, there is one test that is usually missing from routine analysis and that is the test of normality.

Figure 1. The bar chart (A) of the average volume of bottled water for



both Day and Night shift. Figure B and C shows Statistical Process Chart (SPC) for day shift and night shift respectively.

3.3 Further diagnosis with normality test and raincloud plot

Test of normality is a test to determine whether your data is following the normal distribution plot or not. The normal distribution plot is also known as the Gaussian plot or bell curve. A normal data following a normal distribution plot means that when we are distributing the individual data plot into a continuous range, there will be a tendency that a singular average curve will be formed as shown in Fig. 2A. This kind of data is termed homogenous data. Any individual data point that is different from the singular curve means that there is a standard deviation between the individual point with the singular curve.

A non-normal data not following the normal distribution plot means that there might be no singular curve when we are distributing the individual data points into a continuous range. This kind of data is termed heterogeneous data. A non-normal distribution plot can describe the trend of the individual data point. Fig. 2B shows an example of a non-normal distribution plot where 2 curves were formed. This distribution plot is termed a bimodal plot. This indicates that there is a possibility an incident that affect the collection of data systematically.

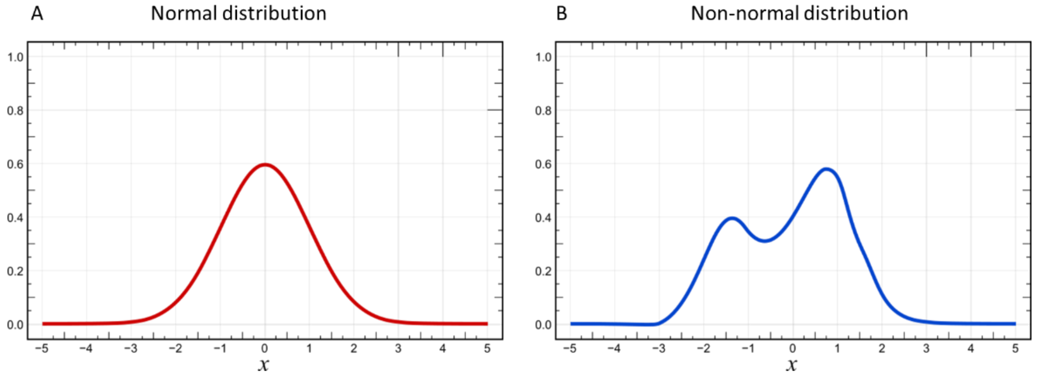


Figure 2. Normal distribution plot (A) and non-normal distribution plot (B).

A dataset that is not normal does not mean that there was a problem with data collection. It only indicates that the data are not homogenous, and this requires you to use alternative statistical tests for analysis. This is because the common hypothesis test such as Student's T-test and ANOVA requires the data is following a normal distribution. There are many types of tests to determine the test of normality. For example, the Shapiro-Wilk test, Levene's test and Jarque-Bera test can be used to determine whether a data set is normally distributed or not. However, the requirements and interpretation of the result are different for each test. For this case study, you chose to use the Shapiro-Wilk test because it is already included in the JASP software.

Table 3 shows the descriptive statistics of the 2 production batches. Based on the table, the t-test shows that there are no significant different ($p > 0.05$) between the day and night shift. The p-value of Shapiro-Wilk indicates that the Day shift is following a normal distribution plot. However, the night shift indicates that the production data do not follow a normal distribution plot. Therefore, you decided to visualize both datasets. Since the common data visualization strategy (Fig. 1) failed to show any discrepancy, you decided to plot a raincloud plot in JASP.

The idea of raincloud plot was conceived back in 2018 with the main objective to visualize raw data with minimal statistical interference [7]. The raincloud plot is commonly performed in a programming language such as R and Python. Fortunately, the JASP platform had also offer raincloud plots in their software starting from version 0.15. This enables analysts with no experience in a programming language to also use raincloud plots in their analysis. Fig. 3 shows the raincloud plot of the two production batches. Each batch contains a dot plot, box plot and distribution plot. From the dot plot, we can see that there are outliers in the Day batch. However, the distribution plot of the day batch is following a normal distribution. The middle line in the box plot indicates the median which is the middle point among the datasets. The median can be an approximate average in a normal distribution plot. In the context of the day batch, the median is aligned with the central curve in the distribution plot.

Table 3: Descriptive statistics of the production data

	Volume of bottled water	
	Day	**Night**
Valid	360	360
Mean	499.978	500.068
Std. Deviation	0.880	1.090
T-test Statistic	0.219	
Shapiro-Wilk	0.997	0.956
P-value of Shapiro-Wilk	0.663	< .001
Minimum	497.062	498.006
Maximum	502.898	502.294

The dot plot of the Night batch shows that the outlier is less severe than the day batch. However, the distribution plot is skewed to the right indicating that there is a spike in bottled water production that exceed 500 ml. The long "slump" distribution to the left side indicates that most of the bottled water is lesser than 500 ml. Combining these, the results show that there is some problem in the production line during the night shift. The problem was not visible if you just view the results based on normal descriptive statistical reports. (Table 2 and Fig. 1).

From the analysis, you were able to determine 2 issues: the test of normality indicate that the production line is homogenous during the day but changed as the night shift began. If we just followed the result of t-test, we would never detect the problem. The raincloud plot visualizes the individual bottled water data and showed that most of the bottled water during the night shift were lesser than 500 ml and the remaining has more than 500 ml. After reviewing the report for this production line, you proceed to obtain the production report for subsequent days.

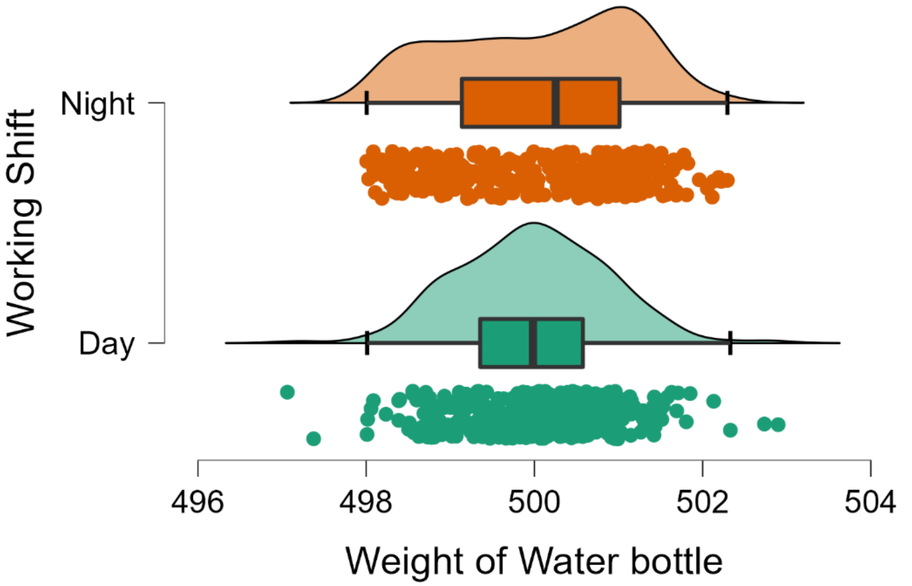


Figure 3. Raincloud plot of both day and night shift.

3.4 Analysis of subsequent production days

For subsequent production reports (Day 2 and Day 3), you produced a descriptive statistics table with an additional analysis: the test of normality using the Shapiro-Wilk test. (Table 4). The average volume for all bottled water is close to 500 ml and the standard deviation for each production shift are within the accepted range. There are several product defects detected but this is nothing unusual. However, the Shapiro-Wilk test shows a similar trend with the previous day for day 2. The t-test shows significant different ($p<0.05$) for day 3, indicating that the problem in night shift had become more apparent. The production line data for the night shift is not homogenous ($p < 0.05$ in the Shapiro-Wilk test) on both days.

You plotted the data in a bar graph and raincloud plots for both day and night shifts on day 2 and day 3 (Fig. 4). You tried to plot the bar graph for both days (Fig. 4A & 4B) to visualize the average and standard deviation of the production line. The bar graph can only show that the average volume of bottled water is lesser than 500 ml. If this were a routine analysis, you might had considered this as a normal result and proceeded to perform statistical analysis. However, this time you have known better and plot the raincloud plot as well (Fig. 4C &

4D).

Table 4: Descriptive statistics of the production data in day 2 and 3

Production batch	Day 2		Day 3	
	Day	Night	Day	Night
Number of bottled waters	360	360	360	360
Average volume (ml)	499.9298	499.9165	500.0698	499.8831
Standard deviation	0.784411	1.088368	0.758792	1.112429
T-test	0.850		0.009	
P-value of Shapiro-Wilk	0.886	< .001	0.518	< .001
Lower limit (ml)	498			
Upper limit (ml)	502			
No. of defects	6	5	1	4

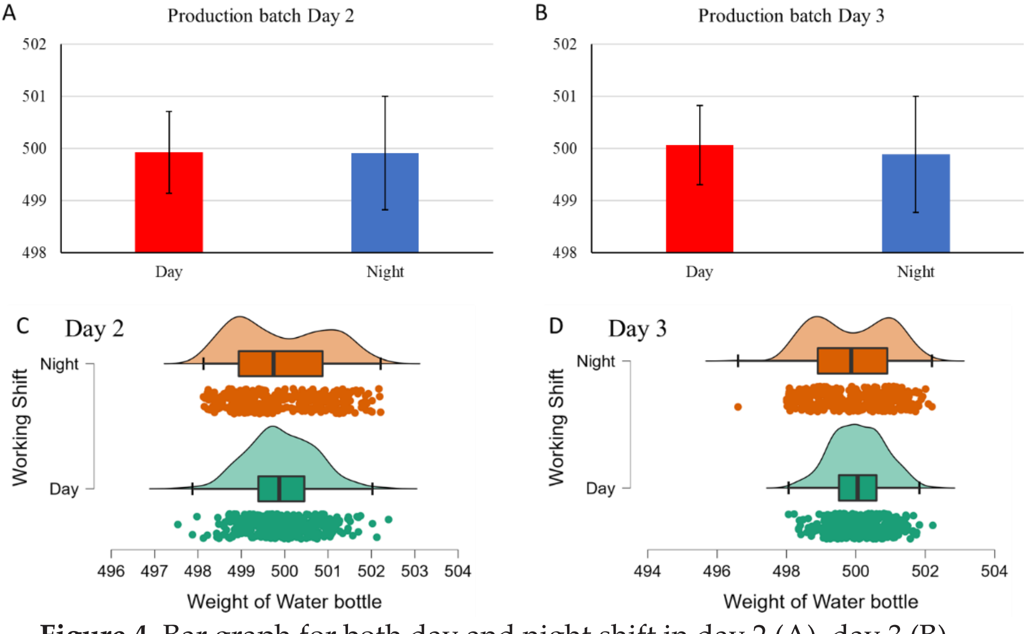


Figure 4. Bar graph for both day and night shift in day 2 (A), day 3 (B) followed by raincloud plots for day 2 (C) and day 3 (D).

The raincloud plots show that the volume of bottled water during the day shift for both day 2 and day 3 are following a normal distribution plot. However, the distribution plot for the night shift for both day 2 and 3 are non-normal distribution. The bimodal distribution plot was exhibited by the night shift on both day 2 and 3. The transition from a skewed plot on day 1 to a bimodal plot on day 2 and 3 indicate that the problem in the night shift had exacerbated.

If you followed the routine analysis and only observed the bar chart (Fig. 1A, 4A & 4B), you might have performed the routine Student's t-test. If following this strategy, you will only realize of the problem in day 3 (Table 4). However, the t-test does not show which shift is the source of problem. The non-normal production of bottled water during the night shift might have not been detected and the company might continuously receive complaints. The reputation of the company A would be tarnished if this continues. Since you had detected that there is a problem on the production line during the night shift, you and your team will now be able to focus the effort to detect the actual problem and implement the corrective measure. The additional test of normality and raincloud plot had save you and your team the time and resources to detect the "source" of the problem in a production line.

4.0 CONCLUSION

It is imperative that engineers regardless of the field gain a deeper understanding of statistics. This is because the basic statistical analysis will not be enough to solve a complex problem that might occur during their work. This article stresses such importance by describing a simulated case study that shows how adding a test of normality and data visualization via raincloud plots could help in identifying possible problems in a production line. All statistical analyses and datasets described in the simulated case study are freely available and do not require any programming skills to perform. We hope that the application of raincloud plots and advanced statistical analytical skills will flourish in the engineering and manufacturing field.

SUPPLEMENTARY DATA

All simulated data described in the article are available at https://osf.io/2s9tu/ for learning, reproducibility, and modification. JASP software is an open-source software, and it is available to be downloaded at their website.

REFERENCES

[1] W. Zhan, R. Fink, and A. Fang, "Application of statistics in engineering technology programs," *American Journal of Engineering Education,* vol. 1, pp. 65-78, 2010.

[2] R. V. Hogg, "A core in statistics for engineering students," *The American Statistician,* vol. 48, pp. 285-287, 1994.

[3] T. Pyzde, "Why normal distributions aren't [all that normal]," *Quality Engineering,* vol. 7, pp. 769-777, 1995.
[4] D. K. Lee, "Data transformation: a focus on the interpretation," *Korean journal of anesthesiology,* vol. 73, pp. 503-508, 2020.
[5] M. W. Hattab, "On the use of data transformation in response surface methodology," *Quality and Reliability Engineering International,* vol. 34, pp. 1185-1194, 2018.
[6] A. P. St‐Pierre, V. Shikon, and D. C. Schneider, "Count data in biology—Data transformation or model reformation?," *Ecology and evolution,* vol. 8, pp. 3077-3085, 2018.
[7] M. Allen, D. Poggiali, K. Whitaker, T. Marshall, J. van Langen, and R. Kievit, "Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved] " *Wellcome Open Research,* vol. 4, 2021.
[8] H. A. Dahlan, Y. Nambu, S. P. Putri, and E. Fukusaki, "Effects of Soaking Tempe in Vinegar on Metabolome and Sensory Profiles," *Metabolites,* vol. 12, p. 30, 2022.
[9] C. Beach and A. J. Casson, "Impact of shift working on the potential for self-powering via kinetic energy harvesting in wearable devices," presented at the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2021.
[10] P. Goos and B. Jones, *Optimal design of experiments: a case study approach*: John Wiley & Sons, 2011.
[11] J. E. Gibson, W. T. Scherer, W. F. Gibson, and M. C. Smith, *How to do systems analysis: Primer and casebook*: John Wiley & Sons, 2016.
[12] JASP Team. (2021, December 23). *JASP (Version 0.16)*. Available: https://jasp-stats.org/