Statistical Analysis of Processing Data for Premier Bread Industry Using Chi-Squared and Goodness-of-Fit Test

[1]Okolie Paul Chukwulozie, [2]Ezeliora Chukwuemeka Daniel, [1]Chukwuneke Jeremiah Lekwuwa and [1]Achebe Chinonso Hubert
[1]Mechanical Engineering Department, Nnamdi Azikiwe University, P.M.B 5025 Awka, Nigeria
[2]Industrial and Production Engineering Department, Nnamdi Azikiwe University, P.M.B 5025 Awka, Nigeria
pc.okolie@unizik.edu.ng

***Abstract*– This research work was based on the statistical analysis of the data obtained from Premier Bread Industry using statistical tools. Line chart was used to understand how the processing data time varies with each bread size. While the statistical summary chart was used to extract information from the data which enable us to understand the situations these data portray. However, the probability plots were used to show the fitness of the data for modeling. Furthermore, data was also analyzed statistically using chi-squared and goodness-of-fit test.**

***Keywords*– Bread, Giant loaf, Long loaf, Small Loaf, Mixing, Matching, Molding, Baking, Goodness-of-Fit and Probability Plot**

I. INTRODUCTION

Statistics is the scientific discipline that deals with the collection, classification, analysis, and interpretation of numerical facts or data. Statistics carries the specific connotation of a quantitative, description and analysis of the various aspects of a state or other social or natural phenomena. Statistical concept and statistical thinking enable the user to solve problems in almost any domain, support decisions reached, and reduce guess work. The objective of statistical analysis is to extract information from data in order to better understand the situations that these data portray.

When the hypotheses come first the test is "prospective" and when the data come first the test is "retrospective". Traditionally, prospective tests have been required [1, 2]. However, there is a well-known generally accepted hypothesis test in which the data preceded the hypotheses [3].

A related question in use of statistics in the physical sciences is whether probability theory applies to the known past in the same way that it applies to the unknown future [4], although these questions have been discussed [5]. It hardly seems reasonable to accord the same status to a hypothesis that explains the results of an experiment after the results are known as to a hypothesis that predicts the results of an experiment before they are known. This is because it is well known that predicting an event before it occurs is more difficult than explaining it after it occurs.

In the sense of Fisher (but not of Neyman–Pearson), statistical significance is a statistical assessment of whether observations reflect a pattern rather than just chance. When used in statistics, the word significant does not mean important or meaningful, as it does in everyday speech: with sufficient data, a statistically significant result may be very small in magnitude.

The fundamental challenge is that any partial picture of a given hypothesis, poll or question is subject to random error. In statistical testing, a result is deemed statistically significant if it is so extreme (without external variables which would influence the correlation results of the test) that such a result would be expected to arise simply by chance only in rare circumstances. Hence the result provides enough evidence to reject the hypothesis of 'no effect'.

Researchers focusing solely on whether individual test results are significant or not may miss important response patterns which individually fall under the threshold set for tests of significance. Therefore along with tests of significance, it is preferable to examine effect-size statistics, which describe how large the effect is and the uncertainty around that estimate, so that the practical importance of the effect may be gauged by the reader.

The calculated statistical significance of a result is in principle only valid if the hypothesis was specified before any data were examined. If, instead, the hypothesis was specified after some of the data were examined, and specifically tuned to match the direction in which the early data appeared to point, the calculation would overestimate statistical significance.

An alternative (but nevertheless related) statistical hypothesis testing framework is the Neyman–Pearson frequentist school which requires that both a null and an alternative hypothesis be defined, and investigates the repeat sampling properties of the procedure, i.e., the probability that a decision to reject the null hypothesis will be made when it is in fact true and should not have been rejected (this is called a "false positive" or Type I error) and the probability that a decision will be made to accept the null hypothesis when it is in fact false (Type II error). Fisherian p-values are

philosophically different from Neyman–Pearson Type I errors. This confusion is unfortunately propagated by many statistics textbooks [6].

Popular levels of significance are 10%, 5%, 1%, 0.5% and 0.1%. if a test of significance gives a P-value lower than or equal to the significance level [7], the null hypothesis is rejected at that level. Such results are informally referred to as 'statistically significant (at the P = 0.05 level, etc.)'. The lower the significance level chosen, the stronger the evidence required. The choice of significance level is somewhat arbitrary, but for many applications, a level of 5% is chosen by convention [8, 9].

Statistical significance can be considered the confidence one has in a given result. In a comparison study, it is dependent on the relative difference between the groups compared, the amount of measurement and the noise associated with the measurement. In other words, the confidence one has in a given result being non-random (i.e., it is not a consequence of chance) depends on the signal-to-noise ratio (SNR) and the sample size.

Expressed mathematically, the confidence that a result is not by random chance is given by the following formula by Sackett [10].

$$Confidence = \frac{Signal}{Noise} \times \sqrt{(Sample\ Size)} \qquad (1)$$

For clarity, the above formula is presented in tabular form (Table 1).

In words, the dependence of confidence is high if the noise is low and/or the sample size is large and/or the effect size (signal) is large. The confidence of a result (and its associated confidence interval) is not dependent on effect size alone. If the sample size is large and the noise is low a small effect size can be measured with great confidence. Whether a small effect size is considered important is dependent on the context of the events compared.

In medicine, small effect sizes (reflected by small increases of risk) are often considered clinically relevant and are frequently used to guide treatment decisions if there is great confidence in them. Whether a given treatment is considered a worthy endeavor is dependent on the risks, benefits and costs [10].

II. METHODOLOGY

The data of Premier Bread Industry's operations were obtained. A statistical analysis of the data using graphical summary chart was carried out. The fitness of the data was validated using the normal probability plot. The confirmation of how good and fit the data are for modeling of the production process of Premier Bread Industry using the goodness-of-fit test for Poisson distribution was equally done. The process data for the bread Industry is as shown in Table 2.

Table 1: Dependence of confidence with noise, signal and sample size (tabular form)

Parameter	**Parameter increases**	**Parameter decreases**
Noise	Confidence decreases	Confidence increases
Signal	Confidence increases	Confidence decreases
Sample size	Confidence increases	Confidence decreases

Table 2: Process Data for Premier Bread Industry

Size of loaves_1	Giant loaf (x1)	long loaf (x2)	small loaf (x3)
Mixing (min)	2	1	2
Matching (min)	3	1	4
Molding (min)	3	2	2
Baking (min)	2	3	2
Profit per loaf (kobo)	1400	1100	400

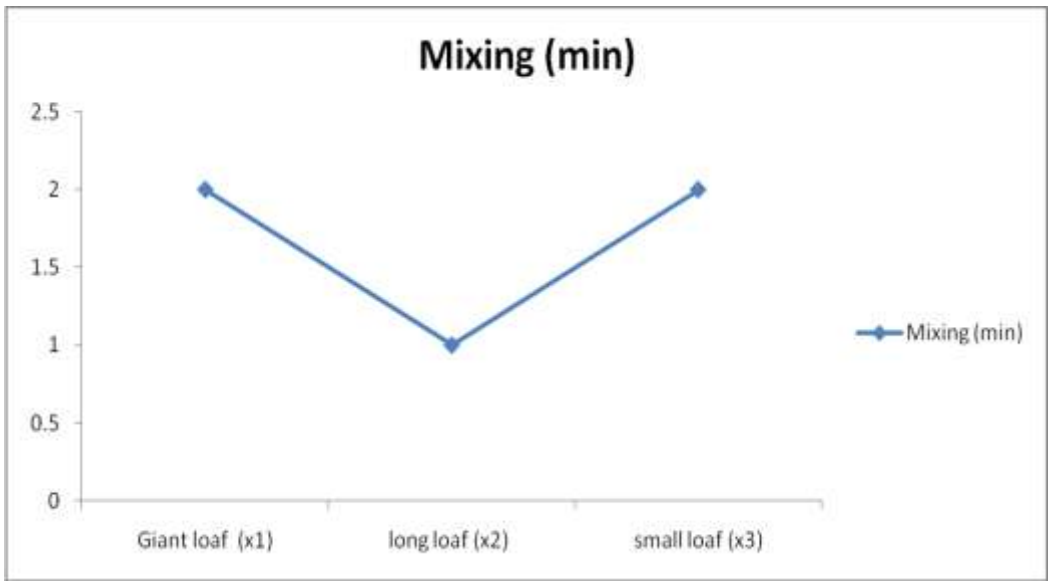

Fig. 1: Graphical representation of time for bread mixing

Fig. 1 shows the time it takes to mix a loaf of bread for its different sizes.

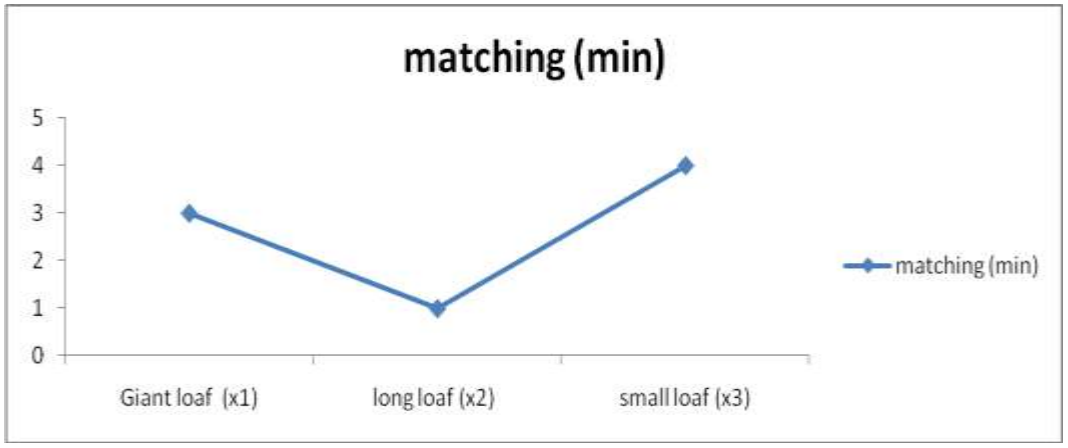

Fig. 2: Graphical representation of time for bread matching

Fig. 2 observes the time it takes to match a loaf of bread for its different sizes.

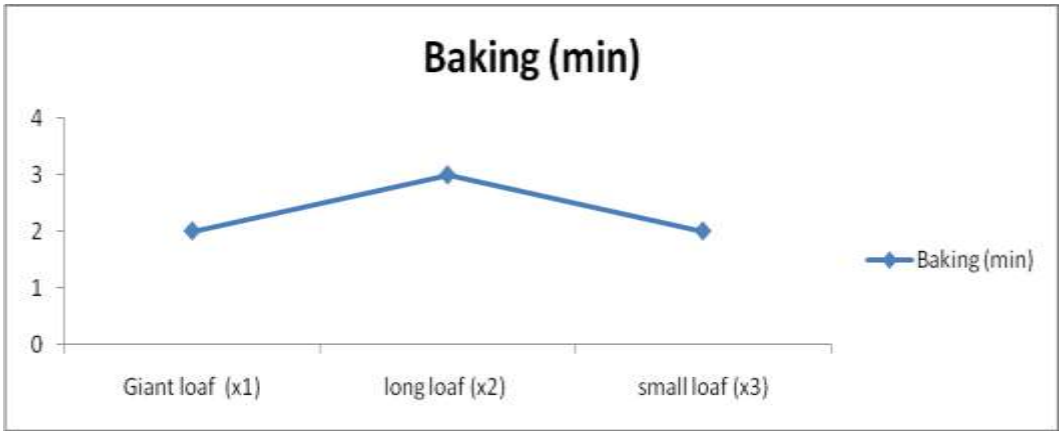

Fig. 3: Graphical representation of time for bread baking

Fig. 3 shows the time it takes to bake a loaf of bread for its different sizes.

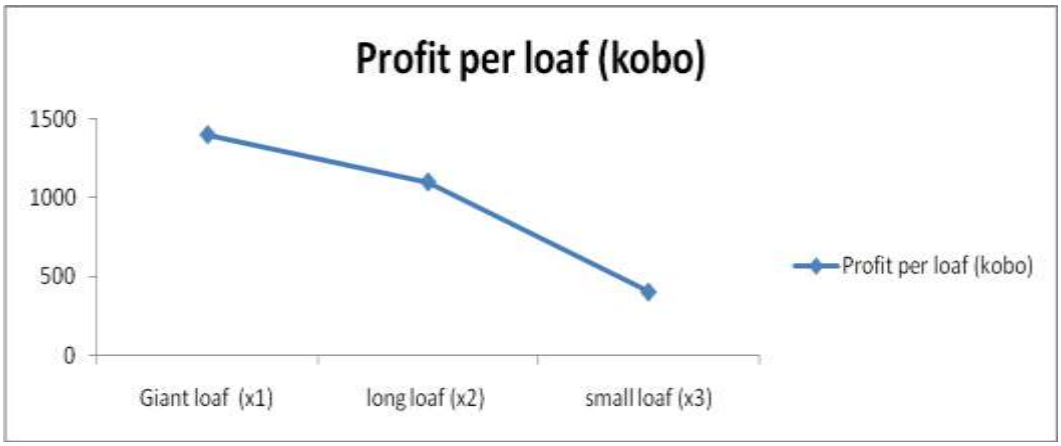

Fig. 4: Graphical representation of Profit per loaf

Fig. 4 shows the profit made (in kobo) per loaf of bread for its different sizes.

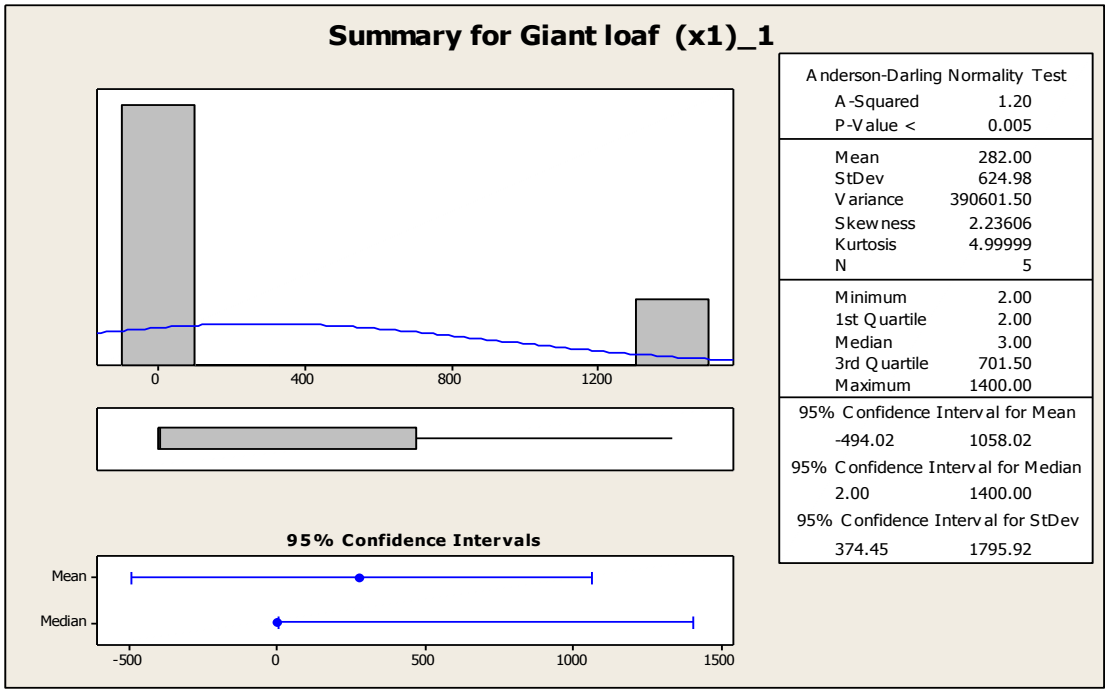

Fig. 5: Statistical Results for Graphical Summary for Giant loaf (x1) _1

Fig. 5 shows the statistical processing data for giant loaf of bread. It shows that the level of significance is strong and it rejects the null hypothesis test.

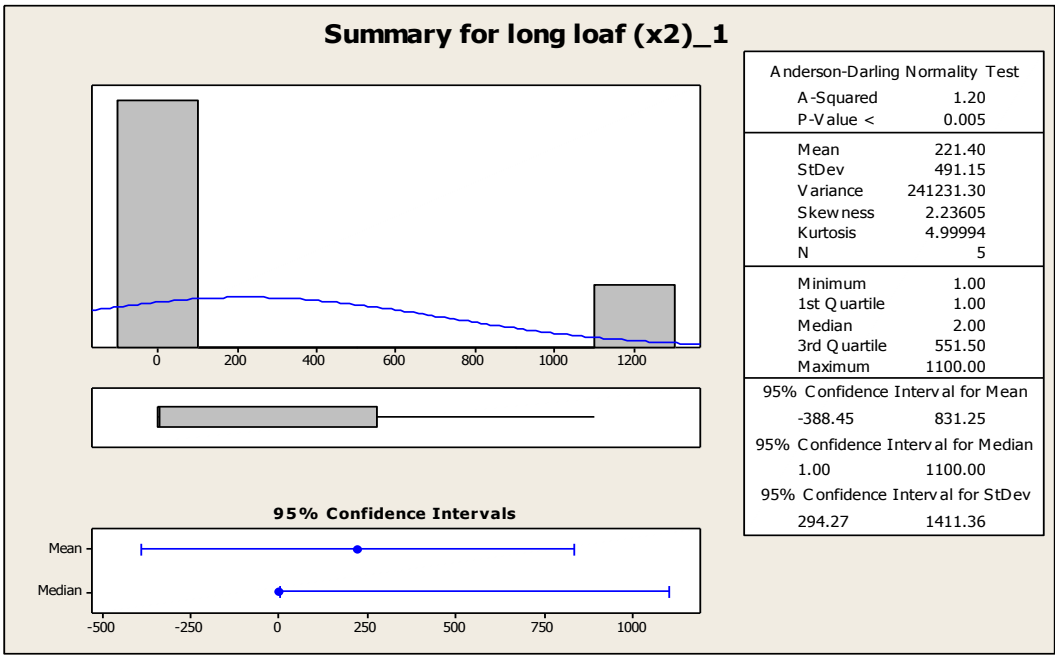

Fig. 6: Statistical Results for Graphical Summary for long loaf (x2)_1

Fig. 6 observes the statistical processing data for long loaf of bread. It shows that the significance level is high and its modeling is adequate for the data.

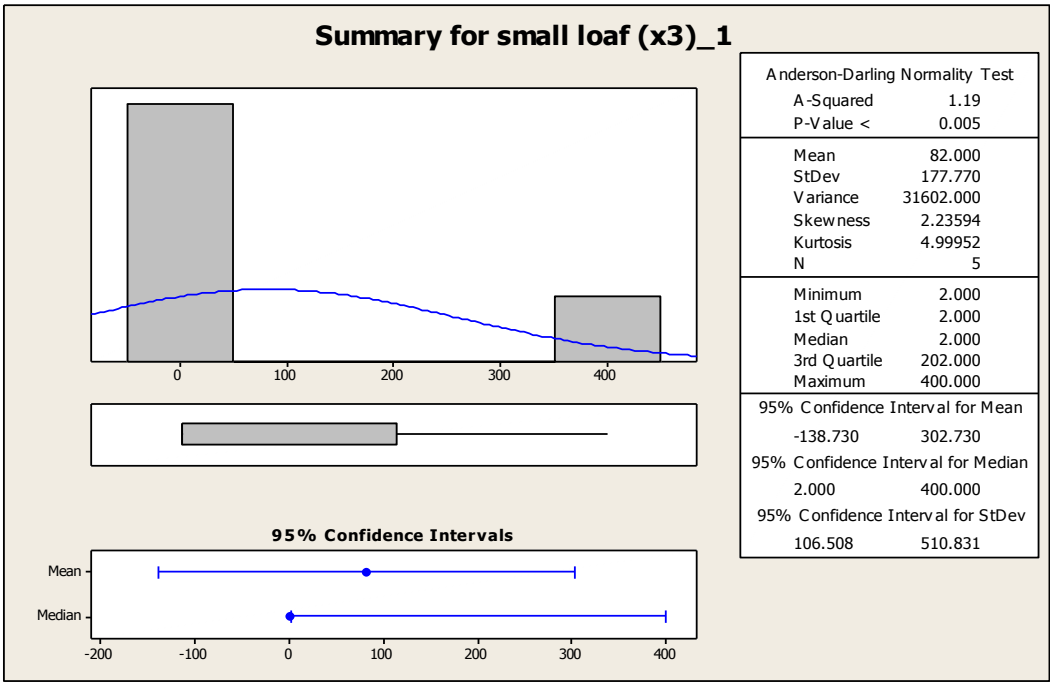

Fig. 7: Statistical Results for Graphical Summary for small loaf (x3) _1

Fig. 7 confirms the statistical processing data analysis for small loaf of bread. It shows that the level of significance is strong and it rejects the null hypothesis test.

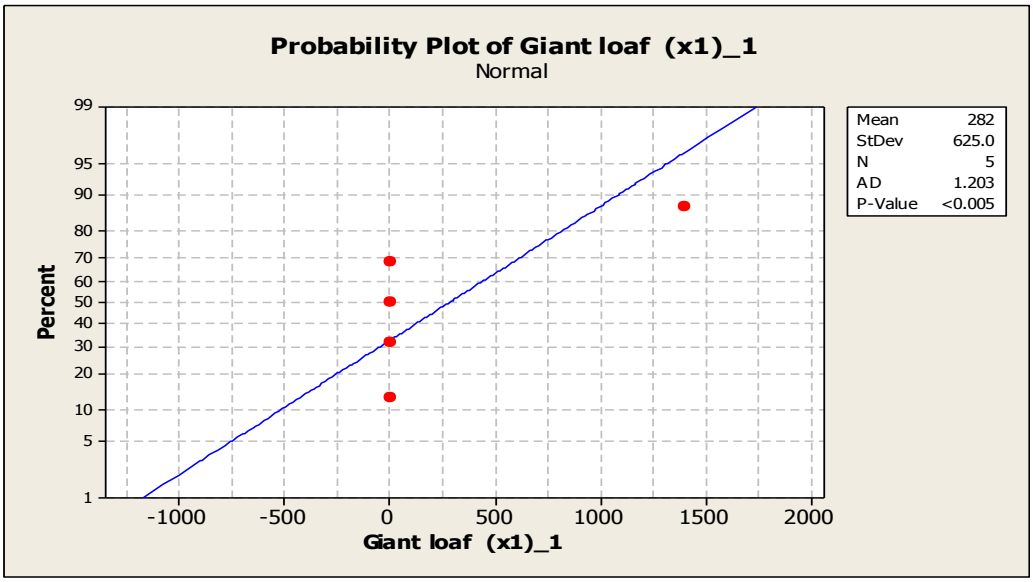

Fig. 8: Probability Plot of Giant loaf (x1) _1

Fig. 8 was used to validate and to confirm the model adequacy in figure 5. It's also used for the fitness of the data.

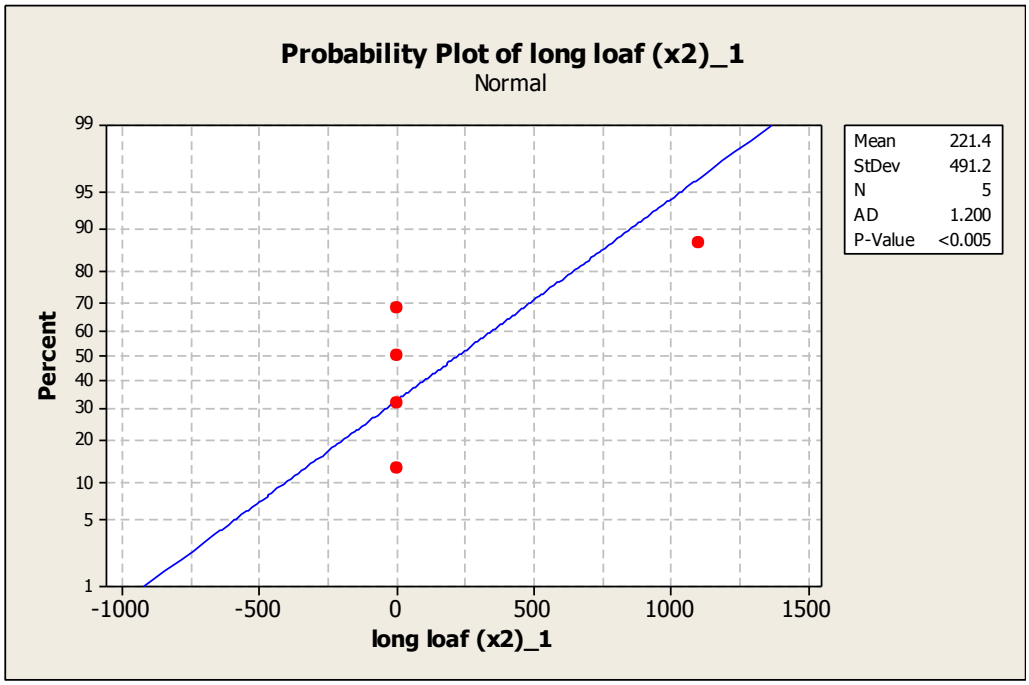

Fig. 9: Probability Plot of long loaf (x2) _1

Fig. 9 was used to validate and to confirm the model adequacy in figure 6. It's also used to test for the fitness of the processing data.

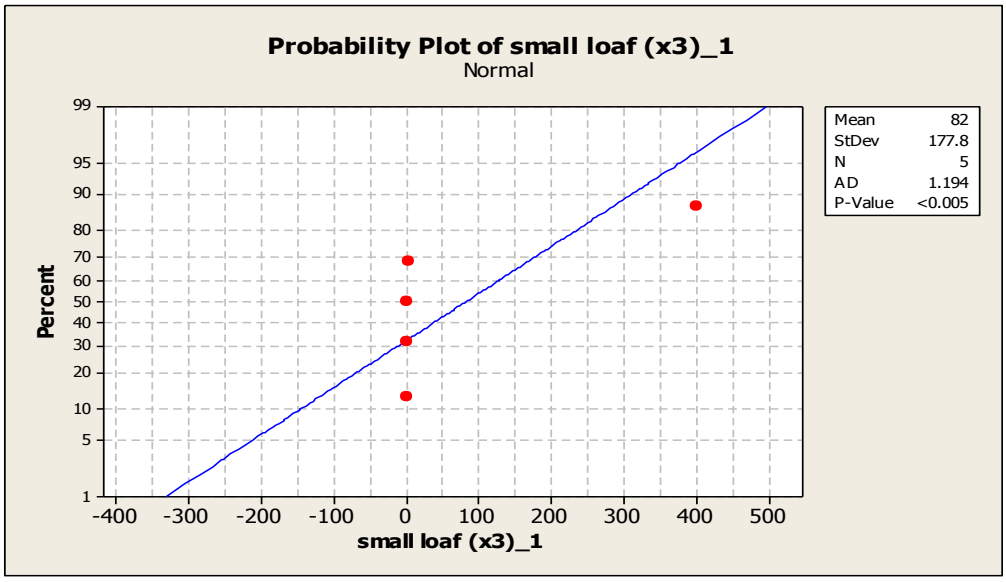

Fig. 10: Probability Plot of small loaf (x3) _1

Fig. 10 was used to validate and to confirm the model adequacy in figure 7. It's also used to test for the fitness of the processing data.

A. Goodness-of-Fit Test for Poisson Distribution

```
Data column: Giant loaf  (x1)

Poisson mean for Giant loaf  (x1) = 282

Giant loaf                  Poisson             Contribution
(x1)          Observed  Probability  Expected     to Chi-Sq
<=2                  2     0.000000         *             *
3 - 285              2     0.586247   2.93124       0.29585
286 - 1399           0     0.413753   2.06876       2.06876
>=1400               1     0.000000         *             *


N  N*  DF  Chi-Sq  P-Value
5   0   2       *    0.000

Expected value is approximately 0.  Chi-Square value is extremely large and
     denoted as *.


2 cell(s) (50.00%) with expected value(s) less than 5.
```

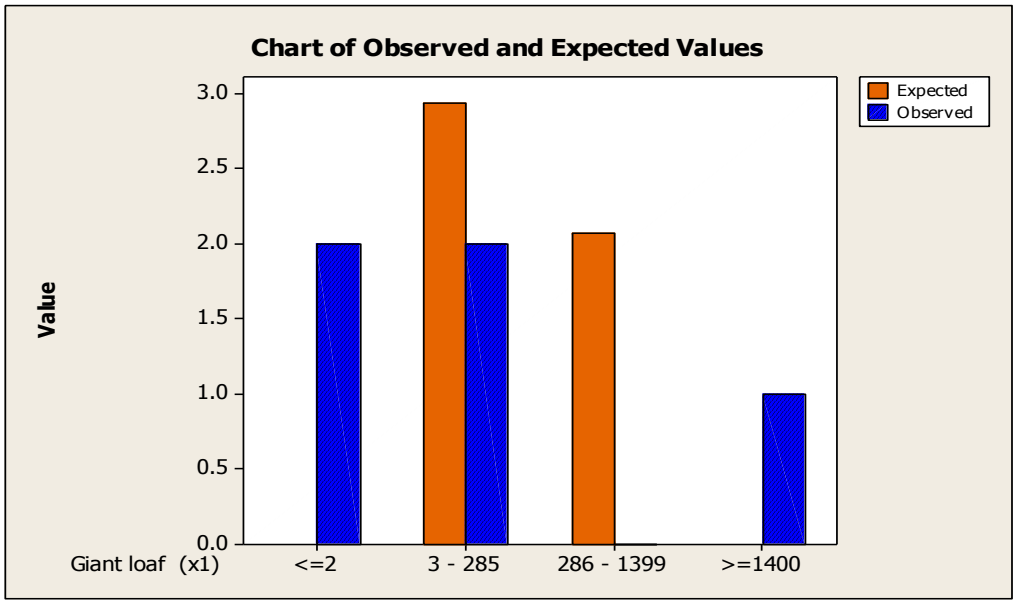

Fig. 11: Chart of Observed and Expected Values

Fig. 11 shows the number of observed and expected values of the processing data

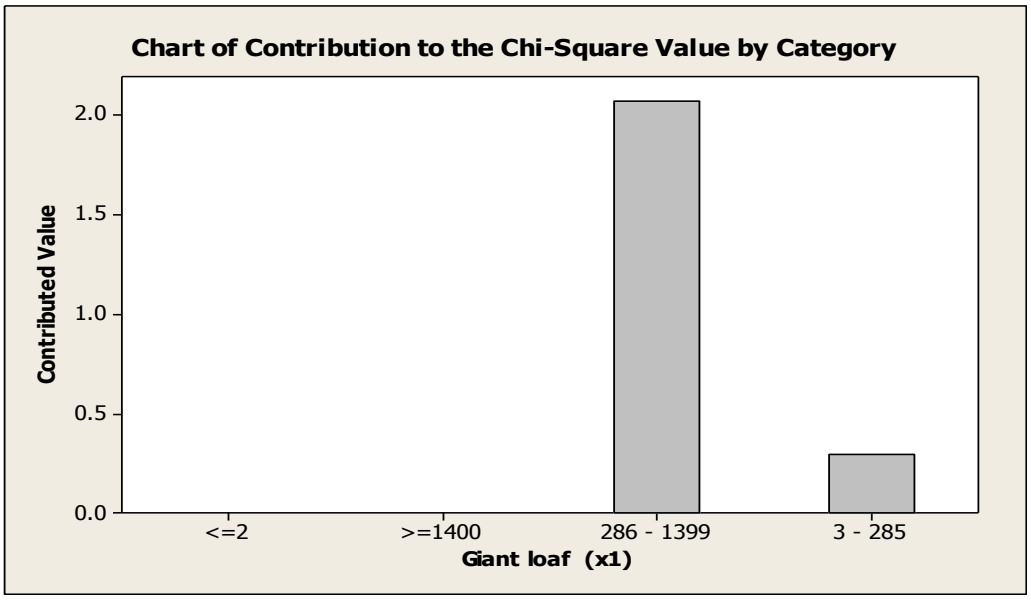

Fig. 12: Chart of Contribution to the Chi-Square Value by Category

B. Goodness-of-Fit Test for Poisson Distribution

```
Data column: long loaf (x2)

Poisson mean for long loaf (x2) = 221.4

                                Poisson              Contribution
long loaf (x2)  Observed  Probability  Expected     to Chi-Sq
<=2                    3     0.000000         *              *
3 - 224                1     0.586690   2.93345        1.27434
225 - 1099             0     0.413310   2.06655        2.06655
>=1100                 1     0.000000         *              *

N  N*  DF  Chi-Sq  P-Value
5   0   2       *    0.000
```

```
Expected value is approximately 0.  Chi-Square value is extremely large and
     denoted as *.
2 cell(s) (50.00%) with expected value(s) less than 5.
```

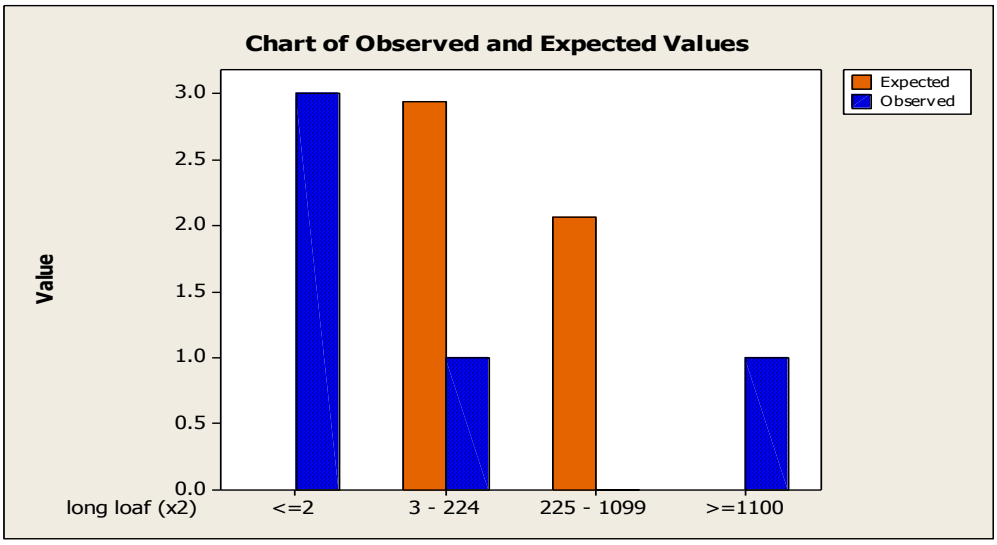

Fig. 13: Chart of Observed and Expected Values

Fig. 13 shows the number of observed and expected values of the processing data

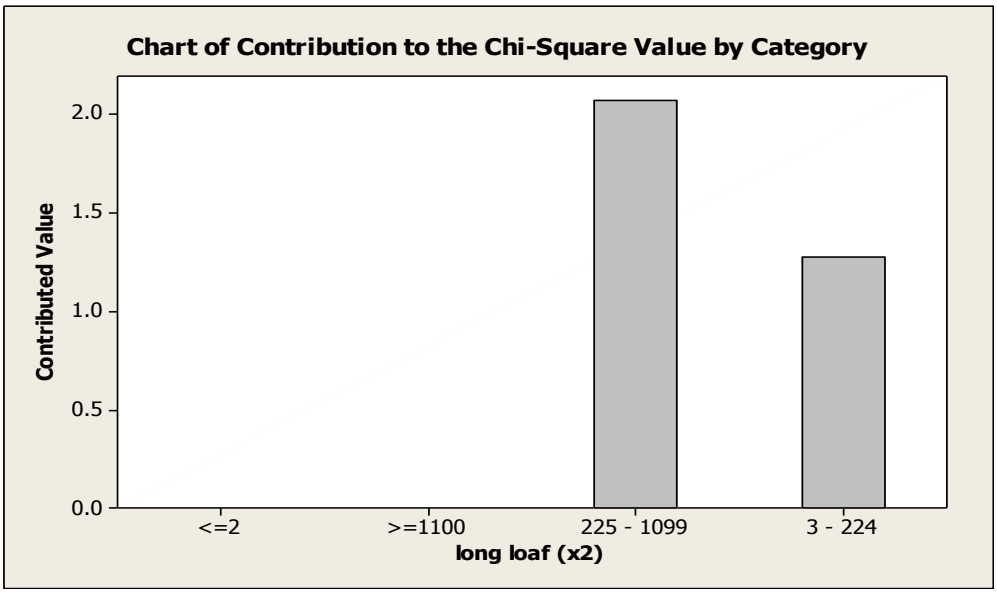

Fig. 14: Chart of Contribution to the Chi-Square Value by Category

C. Goodness-of-Fit Test for Poisson Distribution

```
Data column: small loaf (x3)

Poisson mean for small loaf (x3) = 82

small
loaf                  Poisson              Contribution
(x3)    Observed  Probability  Expected     to Chi-Sq
<=2            3            0         *             *
3              0            0         0           0.0
```

```
>=4           2            1         5           1.8

N  N*  DF  Chi-Sq  P-Value
5   0   1       *    0.000

Expected value is approximately 0.  Chi-Square value is extremely large and
     denoted as *.

WARNING:  1 cell(s) (33.33%) with expected value(s) less than 1.  Chi-Square
     approximation probably invalid.

1 cell(s) (33.33%) with expected value(s) less than 5.
```

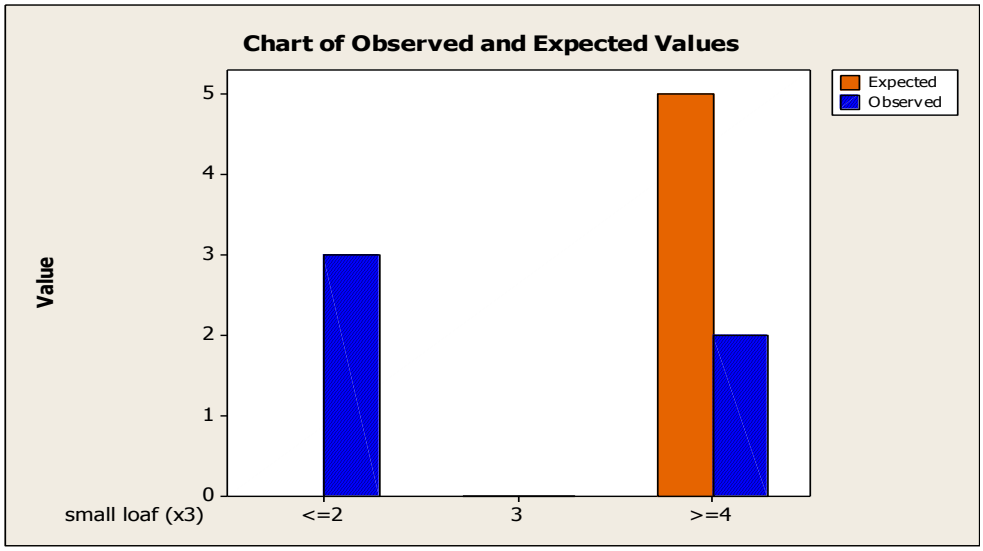

Fig. 15: Chart of Observed and Expected Values

Fig. 1 shows the number of observed and expected values of the processing data

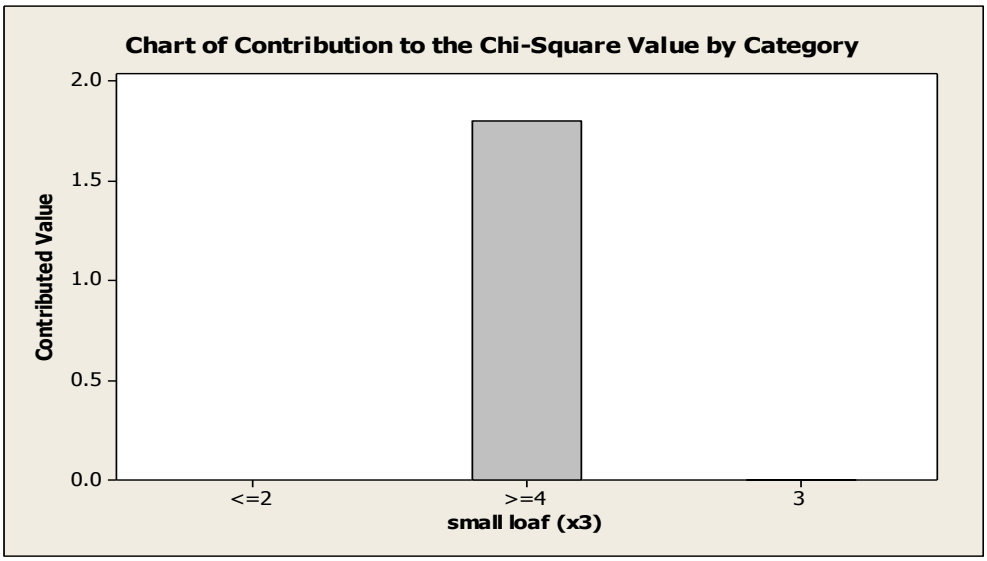

Fig. 16: Chart of Contribution to the Chi-Square Value by Category

III. DISCUSSION OF RESULTS

The line chart was used to show the production process time it takes each loaf of bread to be produced. This was analyzed using different sizes of bread in the case study industry. Statistical summary was performed using graphical summary to show that the results obtained was 95% confidence for different sizes of the bread analyzed. The P-value (i.e., 0.005) for each size of the bread shows that the

data collect were significant. The use of Anderson-darling test for normal Probability plot was used to show and to validate the modeling of the production process data. Their results show that the production process bread sizes were significant and fit for modeling. From the analysis of the result of observed and expected, the poisson distribution show that the p- value = 0.000, which still means that the production process data for the different sizes of the bread were significant. It showed how good and fit the data is for modeling of the production process of the bread sizes. The chi-square and the chart of observed and expected values were used to confirm the goodness-of-fit of the processing data using poisson distribution. Their results were use to show a strong recommendation of the processing data and fitness for the modeling of their production process.

IV. CONCLUSION

The statistical analyses of the processing data were done to understand the statistical behavior. It showed without any guess work the behavior of the process data and the adequacy and goodness-of-fit of the model produced for the process data. The statistical analysis also recommended how strong the data is (by using the level of significant) for modeling.

REFERENCES

[1]. Bacon, Francis (1952), Adler, Mortimer, ed. *Novum Organum*. Great Books of the Western World 30. Encyclopedia Britannica.

[2]. Boole, George (1958) [1854]. "22". *The Laws of Thought*. New York: Dover Publications Inc. p. 402. ISBN 0-486-60028-9.

[3]. USEPA (December 1992). *Respiratory Health Effects of Passive Smoking: Lung Cancer and other disorders*. Washington D. C.: U. S. Environmental Protection Agency. Retrieved Aug. 8, 2012.

[4]. Moore, David; McCabe, George P. "*Introduction to the practice of statistics",* New York: W.H. Freeman and Co. (2003).

[5]. Root, D.H. "Bacon, Boole, the EPA and Scientific Standards". *Risk Analysis* 23 (4): 663–668, (2003).

[6]. Hubbard, Raymond; Bayarri, M.J, "*P Values are not Error Probabilities*", a working paper that explains the difference between Fisher's evidential *p*-value and the Neyman–Pearson Type I error rate (α), (2003).

[7]. Fisher R. A, "The arrangement of field experiments". *Journal of the Ministry of Agriculture* 33: 504, (1926).

[8]. Fisher R. A. "*Statistical Methods for Research Workers"*, Edinburgh: Oliver and Boyd, 1925, p.43, (1925).

[9]. Higgs, M. D, "Do We Really Need the *S*-word?", *American Scientist* 101: 6–1. doi:10.1511/2013.100.6. edit. (2013).

[10]. Sackett DL, "Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!)". *CMAJ* 165 (9): 1226–37. PMC 81587. PMID 11706914, (2001).

[11]. Okolie, P.C, et al, "Optimal Production Mix for Bread Industries: A Case Study of Selected Bakery Industries in Eastern Nigeria", Journal of Engineering and Applied Sciences Volume 5 Issue 6, (2010)