A GOODNESS OF FIT TEST: USING A CHI-SQUARED APPROACH TO FIT THE LOGNORMAL PROBABILITY MODEL TO THE WEIGHTS OF STUDENTS OF AKWA IBOM STATE UNIVERSITY

Michael, Itoro Tim*, Iseh, Matthew Joshua, Ikpang, Ikpang Nkereuwem,
George, Ekemini Udoudo, and Bassey, David Ita

Department of Statistics, Akwa Ibom State University,Mkpat Enin, Nigeria

IJASR 2021
VOLUME 4
ISSUE 6 NOVEMBER – DECEMBER **ISSN: 2581-7876**

Abstract: This paper fits a lognormal probability model to the weights of Students of the Akwa Ibom State University. A sample of 983 Students was drawn from the Medical Centre of the Institution's Main Campus, Ikot Akpaden, Akwa Ibom State. Some exploratory data analyses were carried out to observe the behavior of the data set graphically. A chi-square test is used to ascertain whether or not the weights of students are log-normally distributed. From the graphical displays and the chi-squared test results, it is observed that the weights follow lognormal distribution even though the maximum likelihood estimates of the parameters are quite influential on the results at $\alpha \geq 0.11\%$ significance level.

Keywords: Chi-Square Test, Lognormal Distribution, Maximum Likelihood Estimates, Weights of Students

1. Introduction

Statistical techniques often rely on observations that come from a population and has a distribution of a specific form (e.g., normal, lognormal, gamma, exponential, etc.), thereby making it a necessity to fit the assumed probability distribution to the observation of interest to show how well it can give adequate information about the observation. It is important to note that just as there are different datasets, there are also different probability distributions.

Goodness of fit tests indicate whether or not it is reasonable to assume that a random sample comes from a specific probability distribution. According to Michael, Usoro, Ikpang, Ekemini and David (2019), measures of goodness of fit typically summarize the discrepancy between observed and expected values under the model considered and such measures can be used in statistical hypothesis testing to test for normality of residuals. That is, whether two samples are drawn from identical distributions or whether outcome frequencies follow a specified distribution and others.

Data for the frequentist test is tested against the null hypothesis that it follows the distribution of interest. Several authors, based on the frequentist test have invented several goodness of fit tests and according to Michael, Ikpang and Isaac (2017), Anderson and Dar-ling (1952) introduced the Anderson-Darling test, a statistical test of whether a given sample data is drawn from a given probability distribution with no parameter to be estimated.

Shapiro and Wilk (1965) introduced the Shapiro-Wilk test to test the null hypothesis that the random samples constituting a random variable comes from a normally distributed population. D'Agostino (1970) introduced the D'Agostino's K^2 test, a goodness of fit measure of departure from normality; the test aims to establish whether or not the given sample comes from a normally distributed population.

Observance to laid down conditions and techniques is necessary to ascertain whether or not a given data set follows a defined probability mode, since datasets do not just follow a given probability model. Till date, many probability models have been developed and used in fitting various datasets and many authors have contributed and defined various techniques to verify the normality of datasets and other distributions tests.

The graphical methods, frequentist tests and the Bayesian tests are just some of these techniques. The graphical methods involve the use of graphical tools to display box plots, histogram, Q-Q plots of the given data sets and comparing same with that of the theoretical distributions.

Pearson (1900) investigated the properties of Pearson's chi-squared test. Pearson chi-squared test tests a null hypothesis that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. Lilliefors (1967) introduced the Lilliefors test, a normality test based on the Kolmogorov-Smirnov test. It is used to test the null hypothesis that data come from a normally distributed population, when the null hypothesis does not specify which normal distribution.

Recently, Michael, Ikpang and Isaac (2017) fitted a normal distribution to the weights of students of the Akwa Ibom State University and Michael , Usoro , Ikpang, Ekemini and David (2019) fitted a Gamma probability model to the height of students of Akwa Ibom State University, using the Chi-squared approach by splitting the students' weights into different cells to obtain the observed values and using the raw data for the maximum likelihood estimation of model parameters mean and standard deviation, thereafter, calculating the cells probability and the chi-squared value. Datasets need not follow only one probability model.

This work fits the Lognormal distribution to the weights of Akwa Ibom State University Students using the Chi-Squared test. The Weights of 983 students of the Akwa Ibom State University was collected from the Medical Centre, Main Campus, Ikot Akpaden.

2. Methodology

As stated above, many authors have contributed and defined various techniques to verify or test many distributions. These techniques include but not limited to the following; the graphical methods and frequentist test.

This work employs the two methods, for testing or verifying if lognormal distribution fits the weights of Akwa Ibom State University Students; the graphical method and the chi-squared methods.

The graphical method

The graphical methods involve the use of graphical tools to display box plots, histogram and density plot of the given data sets and comparing same with that of the theoretical distribution. In this research work, we display the box plot, the histogram and density plot and the normal Q-Q plot for the raw and the simulated datasets.

The chi-squared method

According to Wackerly, Mendenhall and Scheaffer (2008), Karl Person in 1900 proposed the following test statistics, which is a function of the deviations of the observed counts from their expected values, weighted by the reciprocals of their expected values. Thus,

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{[X_i - E(n_i)]^2}{E(n_i)} = \sum_{i=1}^{k} \frac{[X_i - np_i]^2}{np_i} \qquad (1)$$

called the Pearson chi-squared test and denoted by χ^2_{k-1} with $k-1$ degrees of freedom.

Where:

X_i= an observed frequency (i.e.count) for n_i

$E(n_i)$= an expected (theoretical)frequency for n_i asserted by the null hypothesis.

n = the sample size

Sahoo (2013) noted that a random variable X is said to have a lognormal distribution with parameters μ and δ^2 written as $X \sim \Lambda(\mu, \delta^2)$ if its probability density function is given by:

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x) - \mu}{\sigma}\right)^2} & , if\ 0 < x < \infty \\ 0, & otherwise \end{cases} \qquad (2)$$

Where $-\infty < \mu < \infty$ and $0 < \delta^2 < \infty$ are arbitrary parameters

Suppose that we take a random sample $Y_1, Y_2, \dots, Y_k$ of size n from this distribution and If we let X_i denote the frequency of $A_i, i = 1,2,3,\dots,k$, so that $X_1 + X_2 + \cdots + X_k = n$,, then the random χ^2_{k-1} variable in (1) cannot be computed once $X_1, X_2, \cdots, X_k$ have been observed, since each pi , and henceχ^2_{k-1}, is a function of μ and δ^2. The values of μ and δ^2 that minimize χ^2_{k-1} are difficult to compute therefore, their maximum likelihood estimates;

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln(X_i)}{n} \tag{3}$$

and

$$\widehat{\delta^2} = \frac{\sum_{i=1}^{n}\left(\ln(X_i) - \frac{\sum_{i=1}^{n}\ln(X_i)}{n}\right)^2}{n} \tag{4}$$

are used to evaluate pi and χ^2_{k-1}. Using maximum likelihood estimates of the parameters in place of minimum chi-square estimates tend to lead to the rejection of the null hypothesis since the χ^2_{k-1} value is not minimized by maximum likelihood estimates, and as such the computed value is somewhat greater than it would be if minimum chi-square estimates are used.

3. Results and Discussion

Various graphical displays are shown to demonstrate the behavior of the dataset as seen in Figure 1 and Figure 2 below while a chi-square test is carried out to ascertain through a statistical test if the dataset follows a lognormal distribution or not.

Graphical displays

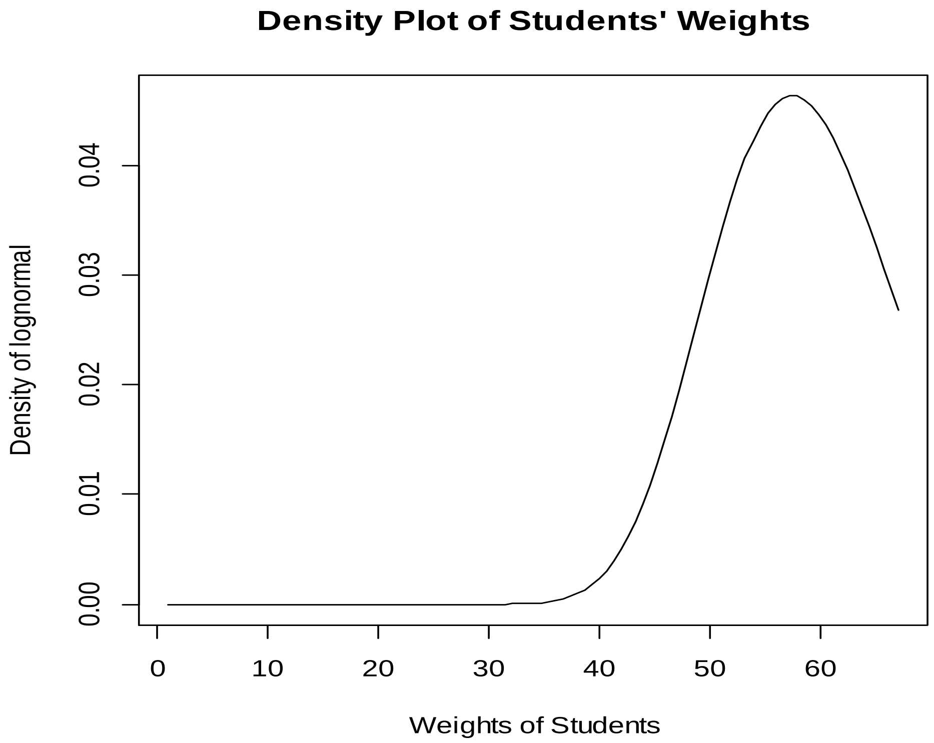

Figure 1: Density plot of student's weights

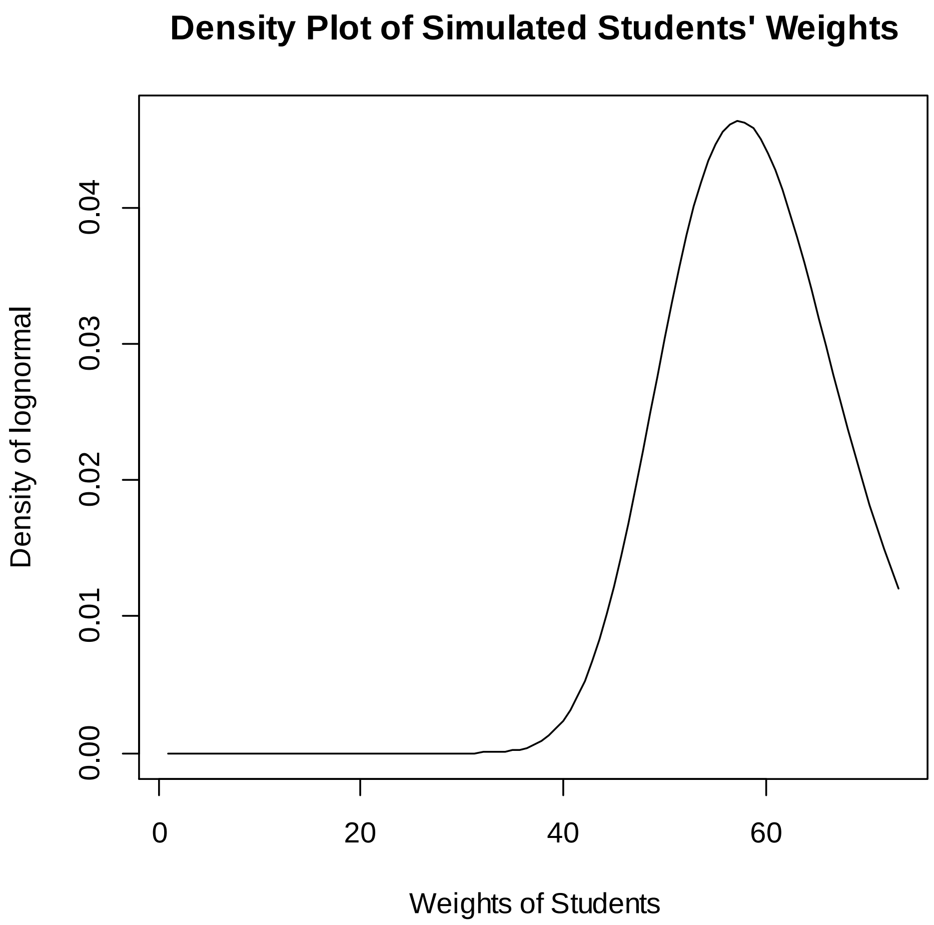

Figure 2: Density plot of simulated student's weights

Chi-square test results

The chi-square test is employed to ascertain whether or not the data follow the distribution of interest.

Research hypothesis

The Null hypothesis (H_0): The weight of students follows a lognormal distribution.
The Alternative Hypothesis (H_1): The weight of students does not follow a lognormal distribution.

Estimation of parameters for the lognormal distribution using maxLik in R

According to Sahoo (2013), a random variable X is said to have a lognormal distribution with parameters μ and δ^2 written as $X \sim \Lambda(\mu, \delta^2)$ if its probability density function is given by:

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2} & \text{, if } 0 < x < \infty \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Where $-\infty < \mu < \infty$ and $0 < \delta^2 < \infty$ are arbitrary parameters

The log maximum likelihood function, $\mathcal{L}(\mu, \sigma^2|X)$ of the lognormal distribution is defined as;

$$\mathcal{L}(\mu,\sigma^2|X) = \frac{-n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\ln(X_i) - \frac{\sum_{i=1}^{n}(\ln(X_i))^2}{2\sigma^2} + \frac{\sum_{i=1}^{n}\ln(X_i)\mu}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \qquad (6)$$

and the maximum likelihood estimate of the parameters are obtained using maxLik package (Henningsen and Toomet, 2009) in R program.

R Codes for obtaining the maximum likelihood estimate of the parameters

```
library(maxLik)
lognormal1<-function(statistics){
  mu<-statistics[1]
  sigma<-statistics[2]
  sum(dlnorm(Weight, mu, sigma, log=TRUE))
}
mle<-maxLik(logLik=lognormal1, start=c(mu=40, sigma=25))
result<-summary(mle)
result
##
```

$$\boldsymbol{\hat{\mu} = 4.071475, \hat{\sigma} = 0.148381, \hat{\sigma}^2 = 0.02201692}$$

Computation of the respective probabilities

The random variable X, denoting the weights of students is partitioned into the following k mutually disjoint sets:

$A_1 = \{-\infty < x \leq 45\}$ $\qquad$ $A_2 = \{45 < x \leq 50\}$
$A_3 = \{50 < x \leq 55\}$ $\qquad$ $A_4 = \{55 < x \leq 60\}$
$A_5 = \{60 < x \leq 65\}$ $\qquad$ $A_6 = \{65 < x \leq 70\}$
$A_7 = \{70 < x \leq 75\}$ $\qquad$ $A_8 = \{75 < x \leq 80\}$
$A_9 = \{80 < x \leq 85\}$ $\qquad$ $A_{10} = \{85 < x \leq \infty\}$

Let $(A_i\ ) = \quad = 1,2, \ldots , k$, where p_i is the probability that the outcome of the random experiment is an element of the set A_i from the normal probability distribution. The probabilities are obtained as follows:

$$p_i = \int_a^b \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{1}{2}(\frac{\ln(x) - \mu}{\sigma})^2}; i = 1, 2, \ldots, 10 \qquad (7)$$

Where a and b are the lower and upper limit for each A_i, i=1, 2, …, 10

Table 1 shows the calculated probabilities obtained from (7) with observed and expected frequencies.

Table 1. Calculated Probabilities, Observed Frequencies and Expected Frequencies

Cell(i)	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	10
A_i	$(-\infty, 45]$	$(45,50]$	$(50,55]$	$(55,60]$	$(60,65]$	$(65,70]$	$(70,75]$	$(75,80]$	$(80,85]$	(85,∞)
X_i	27	102	145	230	222	131	78	28	10	10
p_i	0.0371	0.1041	0.1915	0.2285	0.1948	0.1276	0.0678	0.0305	0.0117	0.0621
np_i	36.511	102.323	188.218	224.588	191.481	125.396	66.635	29.974	11.795	6.081

The Test Statistic:

$$\chi^2_{k-3} = \sum_{i=1}^{k} \frac{[X_i - np_i]^2}{np_i} \qquad (8)$$

The test statistic in (8) where X_i and np_i denote the observed and expected frequencies respectively with $k-3$, the degree of freedom is used to obtain values in Table 2 so that

$$\chi^2_{k-3} = \sum_{i=1}^{k} \frac{[X_i - np_i]^2}{np_i} = 22.51428 \qquad (9)$$

Table 2. Ratio of Deviation of Observed from Expected Values to the Expected Values

$\boldsymbol{Cell(i)}$	$\boldsymbol{Observed(X_i)}$	$\boldsymbol{Expected(np_i)}$	$\boldsymbol{(X_i - np_i)^2}$	$\boldsymbol{(X_i - np_i)^2 / np_i}$
1	27	36.511	90.452	2.4774
2	102	102.323	0.105	0.0010
3	145	188.218	1867.760	9.9234
4	230	224.588	29.292	0.1304
5	222	191.481	931.399	4.8642
6	131	125.396	31.409	0.2505
7	78	66.635	129.161	1.9383
8	28	29.974	3.895	0.1299
9	10	11.795	3.220	0.2730
10	10	6.081	15.360	2.5261
Total:				**22.5143**

Significance Levels and Critical Values

The degree of freedom $(df) = n - k - 1 = 7$. Where n represents the number of cells and k the number of parameters estimated.

Table 3: Some Significance Levels with their Corresponding Critical Values.

Significance Level	Critical Value	Degrees of Freedom
0.0001	29.8775	7
0.0011	24.0868	7
0.0021	22.4784	7
0.0031	21.4971	7
0.0041	20.7861	7
0.0051	20.2268	7
0.0061	19.7651	7
0.0071	19.3715	7
0.0081	19.0280	7
0.0091	18.7232	7

The Decision Rule

Reject H_0 if $\chi^2_{k\text{-}3} > \chi^2_{crit}$. Where $\chi^2_{k\text{-}3}$ is the computed value of the test statistic and χ^2_{crit}. is the critical value obtained from Table 3.

4. Conclusion

It is observed from Table 3 that $\chi^2_{k-3} = 22.5143 < 24.0868 = \chi^2_{crit}$. when the significance level $\alpha \geq 0.11\%$. Hence, the weights of students of Akwa Ibom State University follow a lognormal distribution at $\alpha \geq 0.11\%$ using the chi-square test. This may be due to the fact that the maximum likelihood estimates of the parameters instead of the minimum chi-square estimates were used.

Acknowledgement

The first author appreciates the Director of Health Services, Dr. Geoffrey C. Udoh and the Head of Medical Laboratory Unit, Mr. Aniekan B. Okure of the Akwa Ibom State University Medical Centre for their effort during the data collection process from the Medical Centre for this study.

References

1. Shapiro, S. S. & Wilk,M. B. (1965). An Analysis of Variance Test for Normality (complete samples). *Biometrika*. 52 (3–4): 591–611. https://doi.org/10.1093/biometrika/52.3-4.591.
2. Michael, I. T., Usoro,A. E., Ikpang, I. N., George, E. U. & Bassey, D. I.(2019). Fitting a Gamma Distribution using a Chi-Squared Approach to the Heights of Students of Akwa Ibom State University, Nigeria. *International Journal of Advanced Statistics and Probability* 7(2) (2019) 42 – 46.
3. D'Agostino, R. B. (1970). Transformation to Normality of the Null Distribution of g1. *Biometrika*. 57 (3): 679–681. JSTOR 2334794 https://doi.org/10.1093/biomet/57.3.679.
4. Anderson, T. W. & Darling, D.A (1952). Asymptotic Theory of Certain "Goodness-of-Fit" Criteria Based on Stochastic Processes. *Annals of Mathematical Statistics*. 23: 193–212. https://doi.org/10.1214/aoms/1177729437.
5. Henningsen, A. & Toomet, O. (2009). MaxLik: Tools for Maximum Likelihood Estimation. R package version 0.5, Retrieved September 10, 2017 from *http://CRAN.R-project.org*.
6. Hogg, R. V., McKean, J. W. & Craig, A. T. (2013). *Introduction to Mathematical Statistics, 7th Ed.*, Boston: Pearson Education, Inc.
7. Wackerly, D. D., Mendenhall, W. & Scheaffer, L. R. (2008). *Mathematical Statistics with Applications*, 7th Ed., USA: Thomson Higher Education, Inc.
8. Michael, I. T., Ikpang, I. N. & Isaac, A. A (2017). Goodness of Fit Test: A Chi-Squared Approach to Fitting of a Normal Distribution to the Weights of Students of Akwa Ibom State University, Nigeria. *Asian Journal of Natural and Applied Sciences* 6(4) 107 – 113.
9. Lilliefors, H. (1967). On the Kolmogorov–Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*. 62; 399–402. https://doi.org/10.1080/01621459.1967.10482916.*z*
10. Pearson, K. (1990). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*. 50(302): 157-175. https://doi.org/10.1080/14786440009463897.
11. Sahoo, P. (2013). *Probability and Mathematical Statistics*, Louisville, KY 40292, USA