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

I. T. Michael[1], I. N. Ikpang[2], A. A. Isaac[3]

Department of Mathematics & Statistics, Akwa Ibom State University, NIGERIA.

praisetim@yahoo.com, ikpangnkereuwem@gmail.com , tonyisaac4jesus@gmail.com

ABSTRACT

This paper fits a normal probability model to the Weights of Students of the Akwa Ibom State University. A sample of 970 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 normally distributed. From the graphical displays and the chi-squared test results, it is observed that the weights are normally distributed 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, Normal Distribution, Maximum Likelihood Estimates, Weights of Students

INTRODUCTION

For the frequentist tests, data are tested against the null hypothesis that it follows the distribution of interest. Based on the frequentist test, several goodness of fit tests has been invented by various authors. Anderson and Darling (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 come from a normally distributed population. D'Agostino (1970) introduced the D'Agostino's K2 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.

Overtime, many probability models have been used in fitting various datasets. Datasets do not just follow a given probability model, therefore, adherence to laid down conditions and techniques is necessary to ascertain whether or not a given data set follows a defined probability model. Many authors have contributed and defined various techniques to verify the normality and other distributions tests.

These techniques include but not limited to the following; The graphical methods, frequentist tests and the Bayesian tests. 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 $\left(\chi^2\right)$ 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.

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル, 筑西市,日本
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

This paper fits the normal distribution to the weights of Akwa Ibom State University Students using the Chi-squared test. The Weights of 970 Students of the Akwa Ibom State University was collected from the Medical Centre, Main Campus, Ikot Akpaden.

METHODOLOGY

This study employs two methods for testing normality are employed; the graphical method and the chi-squared methods.

The Graphical Method

The box plot, the histogram and density plot and the normal Q-Q plot for the given dataset are displayed.

The Chi-squared Method

According to Wackerly (2008), Karl Pearson 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,

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

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

Hogg et al., (2013), notes that the random variable X represented by the space $\{x: -\infty < x < \infty\}$ can be partitioned into k mutually disjoint sets $A_1, A_2, \ldots, A_k$, so that the events $A_1, A_2, \ldots, A_k$ are mutually exclusive and exhaustive. Let H_0 be the hypothesis that X is $N(\mu, \sigma^2)$ with μ and σ^2 unspecified, then each p_i is a function of the unknown parameters μ and σ^2 as seen in equation (2).

$$p_i = \int_{A_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(x-\mu)^2/2\sigma^2\right] dx, \quad i = 1,2,\ldots,k \quad (2)$$

Suppose that we take a random sample $Y_1, Y_2, \ldots, Y_k$ of size n from this distribution and If we let X_i denote the frequency of $A_i, i = 1,2,3, \ldots, k$, so that $X_1 + X_2 + \cdots + X_k = n$, then the random χ^2_{k-1} variable in (1) cannot be computed once $X_1, X_2, \ldots, X_k$ have been observed, since each p_i , and hence χ^2_{k-1}, is a function of μ and σ^2.

The values of μ and σ^2 that minimize χ^2_{k-1} are difficult to compute therefore, their maximum likelihood estimates, $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ are used to evaluate p_i 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.

RESULTS AND DISCUSSION

Various graphical displays are shown to demonstrate the behavior of the dataset and a chi-square test is carried out to ascertain through a statistical test if the dataset is distributed normally or not.

Graphical Displays

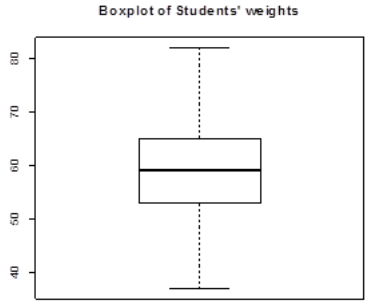

Figure 1. Box plot of the weights of students

Figure 1 evidently shows that the box plot is not skewed with the lower fence, the box and the upper fence of about the same size. Outliers are not present hence the weights are normally distributed.

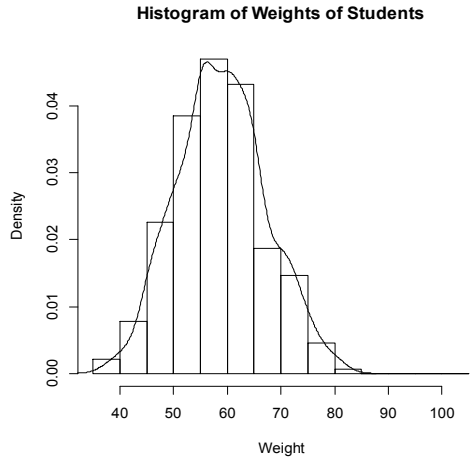

Figure 2. Histogram and Density Plots of the Weights of Students

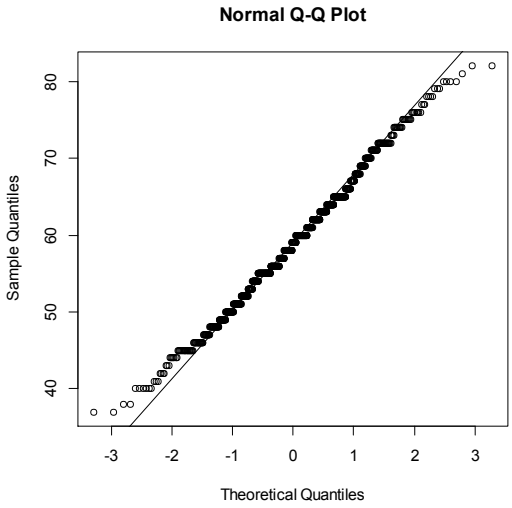

Figure 3. Q-Q norm and Q-Q line of the Weights of Students

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル, 筑西市,日本
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

Figure 2 shows the summary of the overall pattern to be approximately normal and finally, from Figure 3, it is observed that the points on the plot form a nearly linear pattern indicating that the weights are approximately normally distributed.

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 normal distribution

The alternative hypothesis (H_1): The weight of students does not follow a normal distribution

Estimation of Parameters for the Normal Distribution Using Maxlik in R

According to Wackerly *et al*., (2008) , a random variable X is said to have a normal probability distribution if and only if, for $\sigma > 0$, and $-\infty < \mu < \infty$, the density function of X is

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$, $-\infty < x < \infty$ where μ and σ are the parameters of the distribution.

The log maximum likelihood function ,(ℓ) of the normal distribution is defined as

$$\ell = \frac{-n}{2}\log(2\pi) - nlog(\sigma) - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} \quad (3)$$

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

Computation of the Respective Probabilities

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

$A_1 = \{x: -\infty < x \leq 40\}$, $A_2 = \{x: 40 < x \leq 45\}$, $A_3 = \{x: 45 < x \leq 50\}$,

$A_4 = \{x: 50 < x \leq 55\}$, $A_5 = \{x: 55 < x \leq 60\}$, $A_6 = \{x: 60 < x \leq 65\}$

$A_7 = \{x: 65 < x \leq 70\}$, $A_8 = \{x: 70 < x \leq 75\}$, $A_9 = \{x: 75 < x < \infty\}$

Let $p(A_i) = p_i, i = 1,2,\ldots,k$, 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:

$$p_i = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(x-\mu)^2/2\sigma^2\right]dx, \quad i = 1,2,\ldots,9 \quad (4)$$

Where a and b are the lower and upper limit for each $A_i; i = 1,2,\ldots,9$. Table 1 shows the calculated probabilities obtained from (4) with observed and expected frequencies.

Table 1. Calculated probabilities, observed frequencies and expected frequencies

Cell	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
A_i	$(-\infty, 40]$	$(40,45]$	$(45,50]$	$(50,55]$	$(55,60]$	$(60,65]$	$(65,70]$	$(70,75]$	$(75,\infty)$
X_i	10	38	110	187	228	210	91	71	25
p_i	0.0112	0.0352	0.0942	0.1770	0.2336	0.2166	0.1412	0.0647	0.0263
np_i	10.864	34.144	91.374	171.690	226.592	210.102	136.964	62.759	25.511

The Test Statistic

$$\chi^2_{k-3} = \sum_{i=1}^{k} \frac{[X_i - np_i]^2}{np_i} \tag{5}$$

The test statistic in (5) 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

$$\chi^2_{k-3} = \sum_{i=1}^{k} \frac{[X_i - np_i]^2}{np_i} = 22.1925$$

Table 2. Ratio of deviation of observed from expected values to the expected values

Cells	*Observed(X_i)*	*Expected(np_i)*	$(X_i - np_i)^2$	$(X_i - np_i)^2/np_i$
1	10	10.864	0.746496	0.06871
2	38	34.144	14.86874	0.43547
3	110	91.374	346.9279	3.79679
4	187	171.690	234.3961	1.36523
5	228	226.592	1.982464	0.00875
6	210	210.102	0.010404	0.0000
7	91	136.964	2112.689	15.42514
8	71	62.759	67.91408	1.08214
9	25	25.511	0.261121	0.01024
Total				22.1925

Significance Levels and Critical Values

The degree of freedom (df) = $n - k - 1 = 6$. where n represent the number of cells and k, the number of parameters estimated. Table 3 presents some significance levels with their corresponding critical values.

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル, 筑西市,日本
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp

Table 3. Significance level and critical values

Significance level	0.0001	0.0011	0.0021	0.0031	0.0041	0.0051	0.0061	0.0071	0.0081
Critical value	27.856	22.230	20.672	19.724	19.038	18.499	18.053	17.674	17.344

The Decision Rule

Reject H_o if $|\chi^2_{k-3}| > \chi^2_{crit}$. where χ^2_{k-3} is the computed value of the test statistics and χ^2_{crit} is the critical value obtained from Table 3.

R codes for obtaining the maximum likelihood estimate of the parameters

x= The Weights of students

```
library(maxLik)
function1<-function(tim){
mu<-tim[1]
sigma<-tim[2]
a<-(x-mu)^2
b<-(sigma)^2
c<-a/b
d<-0.5*c
sum(-0.5*log(2*pi)-log(sigma)-d)}
results<-maxLik(logLik=function1,start=c(mu=40,sigma=25))
finalresults=summary(results)
finalresults ## mu= μ =58.9330,sigma=σ=8.2876
```

CONCLUSION

The maximum likelihood estimates of the paramaters does not miminise the chi-squared value, however, its value is somewhat higher than the chi-squared values obtainable using the chi-square table. This is avoided by varrying the significance level and using any statistical software to obtain the corresponding critical values.

It is observed from Table 3 that $\chi^2_{k-3} = 22.1925 < 22.230 = \chi^2_{critical\ value}$ when the significance level $\alpha \geq 0.11\%$. Hence, the weights of students of Akwa Ibom State University is normally distributed 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.

The box plot, the histogram and density plot and the normal Q-Q plot also show normality.

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] 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.

[2] D'Agostino, R. B. (1970). Transformation to normality of the null distribution of g_1. *Biometrika.* 57 (3): 679–681. JSTOR 2334794

[3] 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*

[4] Hogg, R. V., McKean, J. W. & Craig, A. T. (2013). *Introduction to Mathematical Statistics, 7th Ed.*, Boston: Pearson Education, Inc.

[5] Lilliefors, H. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*. 62; 399–402.

[6] 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.

[7] Shapiro, S. S. & Wilk,M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika.* 52 (3–4): 591–611.JSTOR 2333709

[8] Wackerly, D. D., Mendenhall, W. & Scheaffer, L. R. (2008). Mathematical statistics with applications, 7th Ed., USA: Thomson Higher Education, Inc.

Leena and Luna International, Chikusei, Japan.
(株) リナアンドルナインターナショナル, 筑西市,日本
ISSN: 2186-8476, ISSN: 2186-8468 Print
www.ajsc. leena-luna.co.jp