International Journal of Advanced Statistics and Probability, 7 (2) (2019) 42-46

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP

Research paper

Fitting a Gamma Distribution using a Chi-squared Approach to the Heights of Students of Akwa Ibom State University, Nigeria

Itoro Tim Michael [1] *, Anthony Effiong Usoro [1], Ikpang Nkereuwem Ikpang [1], Ekemini Udoudo George [1], David Ita Bassey [1]

[1] Department of Statistics, Akwa Ibom State University, Nigeria
**Corresponding author E-mail: itoromichael@aksu.edu.ng*

Abstract

This paper fits a gamma probability model to the heights of Students of the Akwa Ibom State University. A sample of 998 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 heights of students are gamma distributed. From the graphical displays and the chi-squared test results, it is observed that the heights follow gamma distribution even though the maximum likelihood estimates of the parameters are quite influential on the results at $\alpha \geq 0.01\%$ significance level.

Keywords: *Chi-Square Test; Gamma Distribution; Heights of Students; Maximum Likelihood Estimates.*

1. Introduction

Fitting a probability model to a given dataset is necessary to show how well that probability model can give adequate information about the dataset. Different datasets exist as well as different probability models.

The goodness of fit of a probability model describes how well it fits a set of observations or datasets. Measures of goodness of fit typically summarize the discrepancy between observed and expected values under the model considered. Such measures can be used in statistical hypothesis testing to test for; normality of residuals, whether two samples are drawn from identical distributions or whether outcome frequencies follow a specified distribution and others.

Several goodness of fit tests has been invented by various authors. According to Michael, Ikpang and Isaac (2017), 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 comes 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.

Till date, many probability models have been developed and used in fitting various datasets. Datasets do not just follow a given probability model, therefore, observance 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 are 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 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 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.

This work fits the Gamma distribution to the heights of Akwa Ibom State University Students using the Chi-Squared test. The Heights of 998 students of the Akwa Ibom State University was collected from the Medical Centre, Main Campus, Ikot Akpaden.

2. Methodology

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, frequentist test and the Bayesian tests.

This work employs two methods for testing or verifying if gamma distribution fits the heights of Akwa Ibom State University Students; the graphical method and the chi-squared methods. More attention will be given to the latter since it is the main method demanded for this work.

2.1. 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 density plot for the raw and the simulated datasets.

2.2. 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,

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

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

Hogg, McKean and Craig (2013), noted that the random variable X represented by the space $\{x: -\infty < x < \infty\}$ can be partitioned into k mutually disjoint sets $M_1, M_2, \ldots M_k$, so that the events $M_1, M_2, \ldots M_k$ are mutually exclusive and exhaustive.
Let H_0 be the hypothesis that X ~GAMMA(α, β) with β and α unspecified, then each is a function of the unknown parameters β and α as seen in the following equation (2).

$$p_i = \int_{M_i} \frac{1}{\Gamma(\beta)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}} dx , i = 1, 2, \ldots, k \tag{2}$$

Suppose that we take a random sample $Y_1, Y_2, \ldots, Y_k$ of size n from this distribution and if we set X_i to denote the frequency of M_i i = 1,2,3, ..., k, so that $X_1 + X_2 + \cdots + X_k = n$, then the random χ^2_{k-1}variables cannot be computed once $X_1, X_2, \ldots, X_k$ have been observed, since each pi, and hence χ^2_{k-1}, is a function of α and β.

The values of α and β that minimize χ^2_{k-1} are difficult to compute therefore, their maximum likelihood estimates 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 that it would be if minimum chi-square estimates are used. We avoid this by varying our level of significance denoted by α.

To ascertain the validity of our estimates, the method of moment estimates of the parameters given by $\hat{\beta} = \frac{\frac{\sum_{i=1}^{n} x^2}{n} - \bar{X}^2}{\bar{X}}$ and $\hat{\alpha} = \frac{\bar{X}^2}{\frac{\sum_{i=1}^{n} x^2}{n} - \bar{X}^2}$ can be used.

3. Results and discussion

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

3.1. Graphical displays

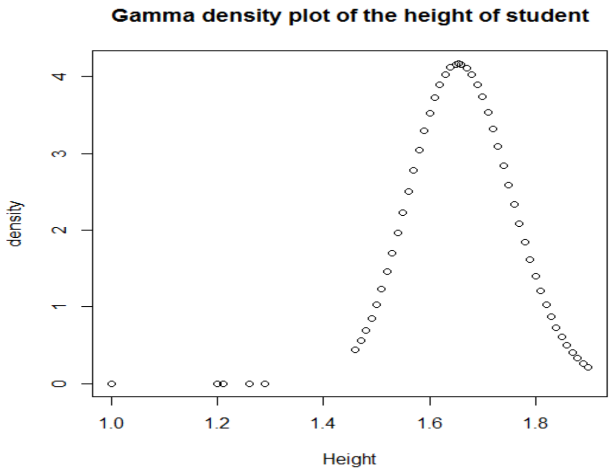

Fig. 1: Gamma Density Plot of the Heights of Students.

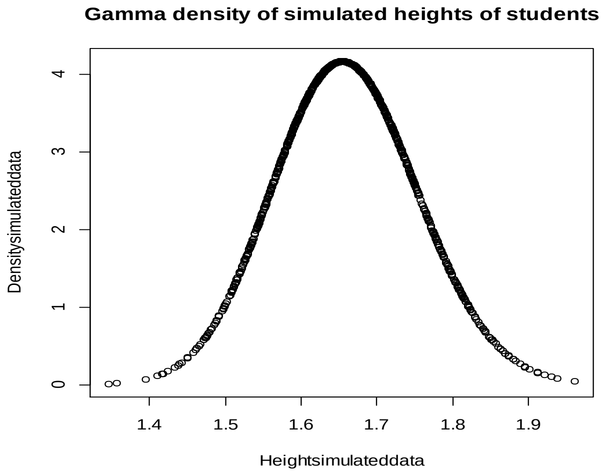

Fig. 2: Gamma Density Plot of the Heights of Students for the Simulated Data.

3.2. Chi-square test results

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

3.3. Research hypothesis

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

3.4. Estimation of parameters for the gamma distribution using maxlik in R

According to Wackerly, Mendenhall and Scheaffer (2008), a random variable ***X*** is said to have gamma distribution with parameters α > 0 and β > 0 if and only if the density function of X is

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\beta)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, 0 \le x < \infty \\ 0, otherwise \end{cases} \tag{3}$$

Where; $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$; α and β are the parameters of the distribution.
The log maximum likelihood function, (ℓ) of the gamma distribution is defined as;

$$\ell = -n \log \Gamma(\alpha) - \alpha n \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{\sum_{i=1}^{n} x_i}{\beta} \tag{4}$$

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

3.5. R Codes for obtaining the maximum likelihood estimate of the parameters

```
X = The Height of students
Library (maxLik)
gamma1<-function (statistics) {
alpha<-statistics [1]
beta<-statistics [2]
sum(dgamma(X, alpha,scale=beta, log=TRUE))
}
mle<-maxLik(logLik=gamma1, start=c(alpha=0.75, beta=2.21366))
Result <-summary(mle)
Result##alpha = α = 300.22048, beta = β = 0.00553
```

3.6. Computation of the respective probabilities

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

$M_1 = \{x: 0 < x < 1.50\}$, $M_2 = \{x: 1.5 \le x < 1.52\}$, $M_3 = \{x: 1.52 \le x < 1.54\}$,

$M_4 = \{x: 1.54 \le x < 1.56\}$, $M_5 = \{x: 1.56 \le x < 1.58\}$, $M_6 = \{x: 1.58 \le x < 1.60\}$, $M_7 = \{x: 1.60 \le x < 1.65\}$, $M_8 = \{x: 1.65 \le x < 1.70\}$, $M_9 = \{x: 1.70 \le x < 1.75\}$, $M_{10} = \{x: 1.75 \le x < 1.80\}$, $M_{11} = \{x: 1.80 \le x < 1.85\}$, $M_{12} = \{x: 1.85 \le x < \infty\}$

Let $p(M_i) = p_i, i = 1,2,\dots,k$, where p_i is the probability that the outcome of the random experiment is an element of the set M_i from the gamma probability distribution. The probabilities are obtained as follows:

$$p_i = \int_a^b \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} dx \, , i = 1,2,\dots,12 \qquad (5)$$

Where a and b are the lower and upper limits for each $M_i; i = 1,2,\dots,12$.

The table 1 shows the calculated probabilities (p_i) obtained from (4) for each set M_i with cells $= 1,2,\dots,12$, observed and expected frequencies, X_i and np_i, respectively.

Table 1: Cells, Calculated Probabilities, Observed and Expected Frequencies

Cells (i)	sets(M_i)	Observed frequencies (X_i)	Probabilities (p_i)	Expected frequencies np_i
1	(0,1.5)	25	0.0437	43.5673
2	[1.5,1.52)	27	0.0248	24.7537
3	[1.52,1.54)	33	0.0341	34.0620
4	[1.54,1.56)	46	0.0447	44.5452
5	[1.56,1.58)	69	0.0555	55.4364
6	[1.58,1.6)	63	0.0659	65.7342
7	[1.6,1.65)	201	0.1965	196.0590
8	[1.65,1.7)	243	0.2017	201.3117
9	[1.7,1.75)	145	0.1595	159.2111
10	[1.75,1.8)	87	0.0986	98.3785
11	[1.8,1.85)	40	0.0482	48.1233
12	[1.85,∞)	19	0.0269	26.8177

The Test Statistic

$$\chi^2_{k-3} = \sum_{i=1}^{n} \frac{(X_i - npi)^2}{npi} \qquad (6)$$

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

Table 2: Computation of the Chi-Squared Value

Cells (i)	Observed frequencies (X_i)	Expected frequencies. np_i	$(X_i - npi)^2$	$\frac{(X_i - np_i)^2}{np_i}$
1	25	43.5673	344.7446	7.912922
2	27	24.7537	5.045864	0.203843
3	33	34.0620	1.127844	0.033112
4	46	44.5452	2.116443	0.047512
5	69	55.4364	183.9712	3.3186
6	63	65.7342	7.47585	0.113728
7	201	196.0590	24.41348	0.124521
8	243	201.3117	1737.914	8.632953
9	145	159.2111	201.9554	1.268475
10	87	98.3785	129.4703	1.316042
11	40	48.1233	65.988	1.371228
12	19	26.8177	61.11643	2.278959
Total				26.62189

So that;

$$\chi^2_{k-3} = \sum_{i=1}^{n} \frac{(X_i - npi)^2}{npi} = 26.62189$$

3.7. Significance levels and critical values

The degree of freedom (df) = n – k – 1 = 9, where n represents the number of cells and k, the number of parameters estimated. Table 3 presents some significance levels with their corresponding critical values.

Table 3: Significance Levels with Their Corresponding Critical Values for Df=9

Significance levels	Critical values	Degree of freedom
0.0001	33.71995	9
0.0011	27.62883	9
0.0021	25.92691	9
0.0031	24.88592	9
0.0041	24.1303	9
0.0051	23.53515	9
0.0061	23.04321	9
0.0071	22.62334	9
0.0081	22.25672	9
0.0091	21.93108	9

3.8. The decision rule

Reject H_0 if $|\chi^2_{k-3}| > \chi^2_{Crit.}$, where χ^2_{k-3} is the computed value of the test statistic and $x^2_{Crit.}$ is the critical value obtained from the table above.

4. Conclusion

It is observed from the results above that χ^2_{k-3} = 26.62189 < 27.62883 = $\chi^2_{Crit.}$ when the significance level $\alpha \geq 0.01\%$. Hence, the height of students of Akwa Ibom State University follow a gamma distribution at $\alpha \geq 0.01\%$. 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 help 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. https://doi.org/10.1214/aoms/1177729437.

[2] 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.

[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] 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.

[6] 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.

[7] 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.

[8] 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/biomet/52.3-4.591.

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