Journal of Balkumari College
ISSN : 2467-9321 Website: http://www.nepjol.info/index.php/jbkc
Volume : 9, Issue : 1, June 2020, Page No.: 84-88

Prediction of Haemoglobin Level by Some Probability Distribution

Govinda Prasad Dhungana[1*] Laxmi Prasad Sapkota[2]

ABSTRACT

Hemoglobin level is a continuous variable. So, it follows some theoretical probability distribution Normal, Log-normal, Gamma and Weibull distribution having two parameters. There is low variation in observed and expected frequency of Normal distribution in bar diagram. Similarly, calculated value of chi-square test (goodness of fit) is observed which is lower in Normal distribution. Furthermore, plot of PDFof Normal distribution covers larger area of histogram than all of other distribution. Hence Normal distribution is the best fit to predict the hemoglobin level in future.

Key Words: Normal, Log-Normal, Gamma, Weibull, Probability distribution

INTRODUCTION

Hemoglobin (Hb) is contained in red blood cells and is capable of combining with oxygen in the lungs, transporting it to the tissues and releasing it there(Thews, G; 1983). It is very important for human body especially women and under 5 year old children. Anemia is a major concern among women, which leads to increased maternal morbidity and mortality and poor birth outcomes, as well as reductions in work productivity (NDHS, 2016). Therefor measurement of Hb is essential for our body. It is measure in mg/dl unit as numerically which is continuous in nature. To predict continuous data, several probability model has been introduced. Some continuous probability model are:

The Probability Density Function (PDF) of normal(or Gaussian) distribution (Gupta S.C & Kapoor V.K, 2000) having parameters, mean (μ) and variance (σ^2) is

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right], -\infty < x < \infty \qquad (1.1)$$

Where "μ" is the location parameter and "σ" is the scale parameter. The case where $\mu = 0$ and $\sigma = 1$ is called the Standard Normal Distribution (SND). This curve has the smooth symmetric bell shape whose highest ordinate is the mode corresponds to the mean of the population. The any of probability distribution id fit after observed and expected frequency is almost equal. The expected frequency of normal distribution is calculated as:

$$Y_i = N \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma} \right)^2 \right], -\infty < x < \infty \qquad (1.2)$$

Where, Y_i = Expected frequency i^{th} class interval
N = Total number of sample or cases
x_i = Limit value of i^{th} class interval

Log-normal (or lognormal) distribution (Sukubhattu N; 2017) is another continuous probability distribution with parameters, mean (μ) and variance (σ^2). Thus, the random variable y=ln(x) is log-normally distributed having PDF

* Corresponding author: dhunganagovindana2012@gmail.com

[1, 2] PhD scholars, Deen Dayal Upadhyaya GorakhpurUniversity, Gorakhpur, India

$$f_x(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma} \right)^2 \right], 0 < x < \infty \quad (1.3)$$

Where, x is only positive real value. It is a convenient and useful model for measurements in exact and engineering sciences as well as medicine,economics and other fields. Similarly, the expected value of lognormal distribution can be calculated by using the relation

$$Y_i = N \times \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma} \right)^2 \right], x > 0 \quad (1.4)$$

Another continues probability distribution is Gamma (Sukubhattu N; 2017) having two parameters,$\alpha>0$, $\beta>0$, having the PDF

$$f_x(x) = \frac{(x/\beta)^{\alpha-1} \exp(-x/\beta)}{\beta\ \Gamma(\alpha)}, x > 0, \alpha > 0, \beta > 0 \quad (1.5)$$

Where α is shape parameter and β scale parameter. The main use of the Gamma distribution is in modeling situations involving continuous change. It is also used in calculus, differential equations, complex analysis, and statistics. The expected frequency can be calculated by using the relation

$$Y_i = N \times \frac{(x/\beta)^{\alpha-1} \exp(-x/\beta)}{\beta\ \Gamma(\alpha)}, x > 0, \alpha > 0, \beta > 0 \quad (1.6)$$

Weibull distribution (Lie, C D et al.,2006) is the most important distribution for reliability problems in survival modeling having two parameter $\alpha>0$ and $\beta>0$.The PDF of Weibull distribution is

$$f_x(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}; x > 0, \alpha > 0, \beta > 0 \quad (1.7)$$

Weibull family, having bathtub shaped and unimodal failure rates besides a broader class of monotone failure rates. Hence, these distribution are more applicable in engineering, actuarial, environmental science, medical sciences, biological studies, demography, economics for reliability and survival studies, Bebbington et al. (2007). The expected frequency can calculated as

$$Y_i = N \times \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}; x > 0, \alpha > 0, \beta > 0 \quad (1.8)$$

To calculate the expected frequency, firstly we estimate to unknown parameters each of distributions using *optim* () function in R software Braun, W. J. et al. (2016), & R Core Team (2019). After estimate the parameters value, we can calculate the probability and expected frequency of each of corresponding class.

In this paper, the expected frequency is calculated from four different probability model having equal parameters (i.e. two), i.e. one is shape and another one is scale parameter. Then we comparison of bar diagram of observed and expected frequency. The p –value is computed by chi square test of goodness fit. Therefore, the least calculated value of any distribution is best fit probability model among them as well as higher coverage the area of PDF in histogram.

METHODOLOGY

Nepal Demographic and Health Survey (NDHS) 2016 is a national representative quantitative population-based survey. Study was completed from 19 June 2016 to 31 January 2017 (NDHS,2016). The raw data is freely available and download from *https://www dhsprogram com/data/dataset_admin*. The NDHS 2016 measured the hemoglobin level for 97% of eligible women age 15-49 from the subsample households. Blood samples were drawn from a drop of blood taken from a finger prick and collected in a microcuvette. Hemoglobin analysis was carried out on-site using a battery-operated portable HemoCue analyzer. Results were provided verbally and in written as

a continue variable. Total of 6463 record of hemoglobin among reproductive age group women were recorded in available dataset. After download the dataset, only hemoglobin among reproductive age group was extracted from dataset and checked for completeness and accuracy. To detect the outliers using the following expression.

$$\text{Outliers} < \{Q_1 - 1.5\times (Q_3 - Q_1)\} \text{ and } \{Q_3 + 1.5\times (Q_3 - Q_1)\} < \text{Outliers} \qquad (2.1)$$

There were 124 (1.93%) outliers were identified from dataset. These outliers were eliminated and final sample were 6299 taken for study purpose.

RESULT AND DISCUSSION

Exploratory data analysis (EDA) is a method to analyzing data sets and it summarize their main characteristics. We have shown that the basic descriptive statistic whereas mean and median was same (i.e. 12.4 g/dl). The shape of distribution was slightly left skewed (β_1 = -0.11) as well as platykurtic (β_2 =2.73) in nature (table 1).

Table 1: Descriptive statistics of hemoglobin status (n=6299)

Min	Q_1	Mean	Median	Q_3	Max	Skewness	Kurtosis
8.6	11.4	12.4	12.4	13.3	16.1	-0.11	2.73

To fit the probability distribution, firstly we estimated the value of unknown parameters. The parameters value of different probability model like as Normal, Log-Normal, Gamma and Weibull distribution were presented as follows (table 2).

Table 2: Parameters value of different probability distributions

Distribution	Normal		Log-Normal		Gamma		Weibull	
Parameters	Mean	SD	Mean	SD	Shape	Rate	Shape	Scale
Values	12.354	1.405	2.507	0.116	75.448	6.107	9.687	12.978

After estimated the parameters value we find out the expected frequency of different probability model. The comparison of observed and expected frequency of each of distribution in figure 1 and figure 2. In figure 1, we demonstrated observed and expected frequency of Normal distribution (left panel) and Log normal distribution (right panel).Similarly, in figure 2, Gamma distribution (left panel) and Weibull distribution (right panel) were demonstrated of observed and expected frequency. In both figure, variation of observed and expected frequency of Normal distribution was lower as compare to other distributions which is contradictory finding with Yadav, R., & Pande, B. B. (1998). Finding of this study, Log normal distribution was best fit distribution for calculation of future flood at various site of river Rapti and its tributaries and other river system in Tarai region.

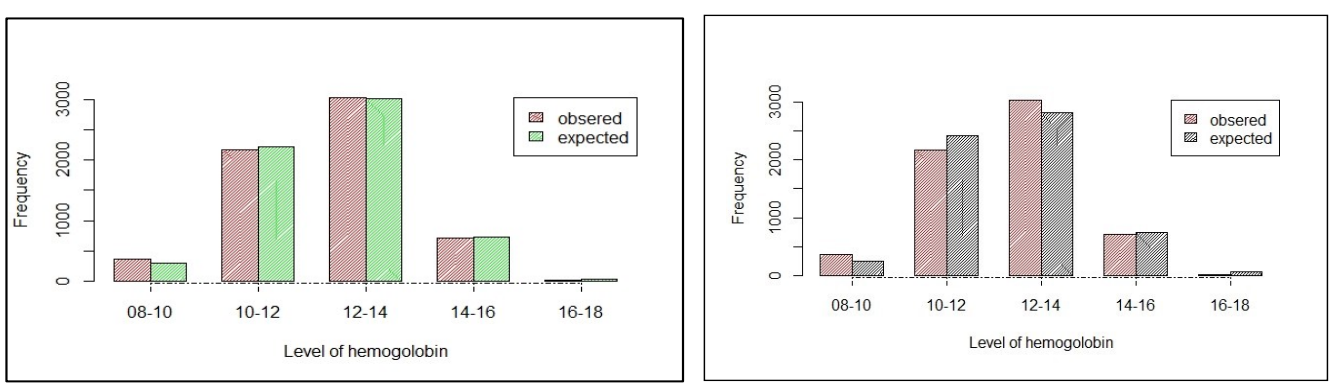

Fig 1: Observed and expected frequency of Normal distribution (left panel), Log-normal distribution (right panel)

The observed end expected frequency Gamma and Weibull distribution was present in following figure.

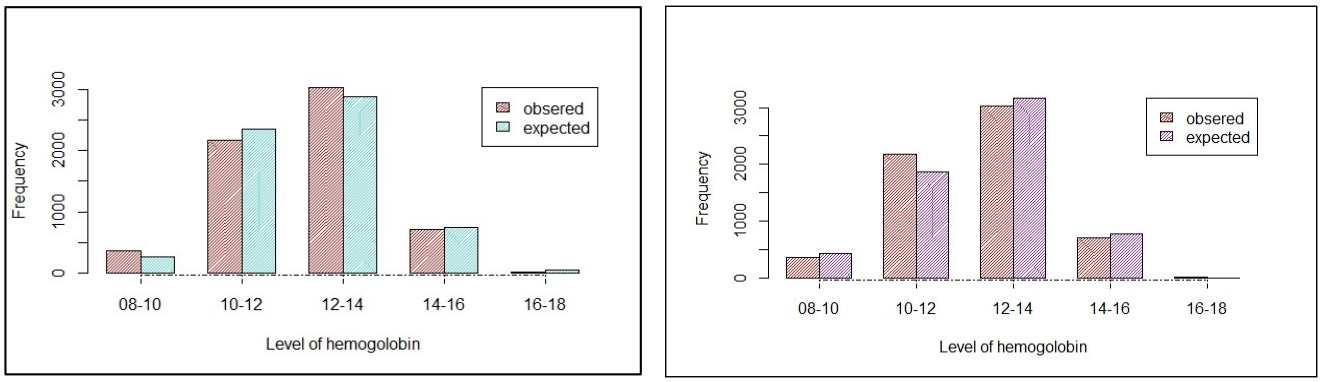

Fig 2: Observed and expected frequency of Gamma distribution (left panel), Weibull distribution (right panel)

Then after, calculate the p-value as well as calculated value of χ^2 test (Machin D et al. 2013) of goodness of fit by using following relation.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (3.1)$$

The observed and expected frequency of haemoglobin level was warries from (8-10) g/dl to (16-18) g/dl. The p-value of each of distribution was statistical significant (p-value<0.001) but calculated value of χ^2 was lower in Normal distribution as compare to other distributions which is contradictory finding with Yadav, R., & Pande, B. B. (1998). Finding of this study, calculated value of Log normal distribution was higher as compare to other distribution.

Table 3: Observed and expected frequency of different probability models

Hemoglobin level (g/dl)	Observed frequency	Expected frequency			
		Normal	Log-normal	Gamma	Weibull
08-10	368	291	245	260	427
10-12	2180	2229	2422	2359	1871
12-14	3033	3017	2823	2890	3161
14-16	710	732	739	737	781
16-18	8	31	68	53	4
$\chi_{cal}^2 - value,\ df = 4$		**40.81**	**160.81**	**105.68**	**115.71**

Furthermore, the graphical presentation of PDF of two parameters family distribution like as Normal, Log-normal, Gamma and Weibull. The PDF of Normal distribution was best fitted distribution than Log-normal, Gamma and Weibull distribution to predict the hemoglobin level (figure 3).

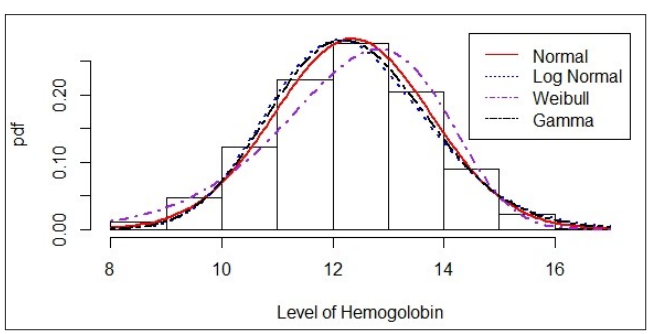

Figure 3: PDF of different distributions to predict the hemoglobin level

CONCLUSION

Hemoglobin is crucial component of human body. It is measured in numerically (g/dl) as continuous variable. So, it follows some theoretical distribution like as Normal, Log-normal, Gamma and Weibull having same parameters (i.e. two parameters). The chi-square value of goodness of fit of Normal distribution is lower as compare to other distributions. There is low variation in observed and expected frequency of Normal distribution as compare to other theoretical distribution. Finally, the plot of PDF of Normal distribution is best fit than other distribution. Hence, Normal distribution is best fitted distribution among Log-normal, Gamma and Weibull distribution to predict the hemoglobin level in future.

REFERENCE

Braun, W. J., & Murdoch, D. J. (2016). *A first course in statistical programming with R*. Cambridge University Press.

Bebbington, M., Lai, C. D., & Zitikis, R. (2007). A flexible Weibull extension. *Reliability Engineering & System Safety*, *92*(6), 719-726.

Gupta, S.C., Kapoor, V.K., (2000). Fundamentals of Mathematical Statistics. Eds -10 Sultan Chand & Sons.

Lai CD., Murthy D., Xie M. (2006). Weibull Distributions and Their Applications. In: Pham H. (eds) Springer Handbook of Engineering Statistics. Springer Handbooks. Springer, London

Machin, D.,Campbell,J. M., Walter, S. J.,(2013). Medical statistics Fourth Edition, John Wiley &Sons, ltd.

Ministry of Health, Nepal; New ERA; and ICF. (2017). *Nepal Demographic and Health Survey 2016.* Kathmandu, Nepal: Ministry of Health, Nepal

R Core Team R (2019). *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sukubhattu,N.P (2013). Probability and Inference –I. Theory techniques and Applications, Asmita Books Publishers and Distributors (p) LTD.

Thews, G. (1983). Blood gas transport and acid-base balance. In *Human Physiology* (pp. 489-507). Springer, Berlin, Heidelberg.

Yadav, R., & Pande, B. B. (1998). Best fitted distribution for estimation of future flood for Rapti river systems in Eastern Uttar Pradesh.

www.e-safe-anaesthesia.org - sessions – pdf-Sickle cell disease; downloaddate-20/02/2020