A study of Hepatitis B virus infection using chi-square statistic

A study of Hepatitis B virus infection using chi-square statistic

Oluwole A Odetunmibi[1,*], Adebowale O Adejumo[2] and Timothy A Anake[3]

[1,3]Department of Mathematics, Covenant University, Ota, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria

*Corresponding author e-mail: oluwole.odetunmibi@covenantuniversity.edu.ng

Abstract. Hepatitis B is caused by the hepatitis B virus (HBV) and it affects livers. It has been established that the disease is a serious medical condition caused by an overpowering immune response to infection. To this effect, there is a need for cross examination of records of patients on this disease to ascertain the factors that could be responsible for the survival or dying from this disease. Descriptive analysis of the data showed that sexually active age bracket (31 – 50) are greatly affected by the disease while female accounted for majority of those that are tested positive to the disease. Chi squared statistic was used to test for independence between age and gender of those who tested positive to disease between 2006 and 2015 in Lagos state, Nigeria. It was discovered that, both variables of age and gender are not independent which means there is association between the Age and Gender of HBV patients.

Keywords: Hepatitis B Virus, Infectious Disease, Liver, Chi Square Statistic

1. Introduction

Hepatitis B which affects livers is an infectious disease that is caused by virus called hepatitis B virus (HBV) [1]. Hepatitis has two major words combined together from the ancient Greek words which are *hepar* from the root word of *hepat* which means 'liver', and the Latin *itis* which simply means "inflammation". The disease could therefore simply be defined as liver injury which bring about inflammation of all the cells that are connected to the liver [1-2]. [3] Described HBV as circular genome hepadna virus which is composed of DNA that are double-stranded partially and in turns brings about replicated RNA through which intermediate form can be reverse transcription.

HBV disease mode of transmission has been established to be through having contact with the blood or infected person body fluids [2]. The prevalence of the disease in any particular population has been connected in general, to the average age at which individuals get infected with the virus [4]. The risk of getting infected with hepatitis B virus is primarily related to three major sources which are: perinatal, sexual and household exposure to infected individuals [5]. Finding a way of preventing HBV that is as a result of pernatal transmission is a major step in curtailing the spread of hepatitis B in an endemic environment. [6] reported that infection at the stage of perinatal by hepatitis B virus can result into about 90% increase of chronic stage of hepatitis. One of the major strategies of breaking the cycle of transmission of HBV is through the use of immunoprophylaxis. [6] Stated that, almost 90% of infants born to HBsAg-positive/HBeAg-positive women and 10% of infants born to HBsAg-positive/HBeAg-negative women become infected [6].

Anyone can contact Hepatitis B virus, but there are some certain set of individual that are at greater risk. These set of people includes: (i) Very young babies, (ii) Individuals who had been

hospitalized recently, (iii) Individuals who have very low functioning immune systems which is as a result of one major disease or the other and are daily usage of any drugs that may lead to the suppress their immune system. For example, those who are used to taking steroids for the purpose of preventing rejection of transplanted organs and (iv) The aged, that are already diagnose of having one health challenge or the other, [2][6][7]
In this article, the independence of age and gender on the spread of the HBV disease was investigated using Chi- square test statistic.

2.0 Material and Method

The data used in this article is a secondary data collected from Human Virology department of Nigeria Institute of Medical Research (NIMR), Yaba, Lagos State, Nigeria. The data contain patients that were tested positive to hepatitis B virus from January 2006 to December of 2015. Data on the date they were tested, the age of the patient and sex of the patient were collected. SPSS version 23 and Microsoft Excel were used to analyze the data for this study.

2.1 *Test Statistics χ^2 and G^2*

When considering multinomial sampling that has probabilities $\{\pi_{ij}\}$ attached to it in *I* x *J* table called contingency table, the statistical independence with null hypothesis of H_0 is $\pi_{ij} = \pi_{i.}\pi_{.j}$ $\forall$ *i* and *j*. When dealing with multinomial that is independent with samples in J column or I rows, then the homogeneity of each of the outcome probability that exist among the rows or the column is said to correspond to its independent. The situation described above works for multinomial with single sample, while the same tests apply when dealing with sample that are multinomial and as well are independent.

2.2 *Pearson (χ^2) and Likelihood-Ratio (G^2) Chi-Squared Tests*

A test of independence null hypothesis (H_0) uses χ^2 with n_{ij} instead of n_i and $n_{ij} = n\pi_{i.}\pi_{.j}$ in place of e_i. Here $e_{ij} = E(n_{ij})$ under H_o.for all *i* and *j*. Usually, $\{\pi_{i.}\}$ and $\{\pi_{.j}\}$ are not known.
The maximum likelihood (ML) estimates for them are the sample marginal proportions $\hat{\pi}_{i.} = \frac{n_{i.}}{n}$ and $\hat{\pi}_{.j} = \frac{n_{.j}}{n}$, so estimated expected frequencies are $\hat{e}_{ij} = n\hat{\pi}_{i.}\hat{\pi}_{.j} = n_{i.}n_{.j}$. Then X^2 equals

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{r}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \tag{1}$$

Pearson (1900, 1904, 1922), stated that changing $\{e_{ij}\}$ by estimates $\{\hat{e}_{ij}\}$ would not have any effect on the distribution of X^2. He reported that X^2 may be asymptotically chi-squared which has degree of freedom = *IJ*-1 if the contingency table has *IJ* categories,

$$\begin{aligned} df &= (IJ-1) - (I-1-(J-1) \\ &= (I-1)(J-1) \end{aligned} \tag{2}$$

Fisher (1922) stated that there is an error in Pearson's theory and corrected the error. He developed the concept of degrees of freedom in his article. (He reported that Pearson had worked on indexed family of chi-squared distributions which had not explained explicitly with when ''degrees of freedom'' is introduced). His articles presented that the test produces the X^2 statistic while a different result is being produced by likelihood-ratio test.
The likelihood kernel for multinomial sampling is given as:

$$\prod_i \prod_j \pi_{ij}^{n_{ij}} \tag{3}$$

Where all $\pi_{ij} \geq 0$ and $\sum_i \sum_j \pi_{ij} = 1$.

In general case, $\widehat{\pi}_{ij} = \frac{n_{ij}}{n}$. The likelihoods ratio is given as:

$$\Lambda = \frac{\prod_i \prod_j \left(n_{i.} n_{.j}\right)^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}} \tag{4}$$

The ratio of the likelihood of the chi-squared statistic is -2 logΛ. This is denoted by G^2, which equals

$$G^2 = -2\text{loijg}\,\Lambda = 2\sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\widehat{\theta}_{ij}} \tag{5}$$

The higher the value of G^2 and X^2, the more reason why independence does not exist.
In situation where parameter space involves $\{\pi_{ij}\}$ subject to a restriction $\sum_i \sum_j \pi_{ij} = 1$ that is linear, the dimension is (IJ – 1). When dealing with null hypothesis (H_0), $\{\pi_{i.}\}$ and $\{\pi_{.j}\}$ are used in determining the value of $\{\pi_{ij}\}$ with the following dimension (I - 1) + (J - 1).

These dimensions has a difference that is equal to (I - 1) (J - 1). When considering large samples, the degree of freedom for G^2 with chi-squared null distribution is given as: (I – 1) (J – 1). G^2 and X^2 is said to have the same chi-squared distribution that has limited null hypothesis. Under this condition, X^2 - G^2 is said to be asymptotically equivalent and converges to zero in probability. These also holds with other sampling schemes for limiting results for multinomial sampling [8-11].

As n grows, these results also apply and hence ($e_{ij} = n\pi_{ij}$) grow when considering cells that has a fixed number. Multinomial distribution that has $\{n_{ij}\}$ can be approximated as they grow while G^2 and X^2 are said to tend towards chi-squared distributions which is known as multivariate normal distribution. X^2 converges faster to chi-squared quicker than G^2. When $\frac{n}{ij} < 5$, the approximation when considering G^2 is usually not very good. When the value of I and J is considerably large and some of the expected frequencies are very small say as low as 1 while some of the values exceeded 5 then, it can be decent for X^2.
Chi-squared tests of independence may not be adequate sometimes to answer all the questions about a particular data set but may indicate the extent to which the evidence of association exists between or among variables under consideration. Investigating the nature of association through the use of: breaking into components of chi-squared, residuals study and parameter estimate such as: describing the strength of association through the use of odds ratios which is rather more efficient. [10-13]

3.0 Data Analysis and Results

Table 1 Distribution of Patients by Age

Age Interval	Frequency	Percentage	Cumulative Percentage
11-20	81	0.47	0.47
21-30	1708	10.11	10.58
31-40	7371	43.65	54.23
41-50	5153	30.52	84.75
51-60	2001	11.85	96.6
61-70	475	2.81	99.41
71-80	86	0.51	99.92
81-90	10	0.06	99.98

Above 90	2	0.02	100
Total	**16887**	**100**	

Table 1 showed that, patients' age were classified into 9 categories. It was discovered from the table that 31- 40 and 41-50 age bracket accounted for 74.17% of the patients that were tested positive within the period under considerations. The descriptive statistics for the table 1 is presented in table 2 while the bar graph for the table is presented in figure 1.

Table 2: Descriptive Analysis Hepatitis B Patients Age Distribution

Mean	Variance	Standard Deviation	Skewness	Kurtosis
40.9137	**87.94**	**9.3776**	**0.717**	**0.877**

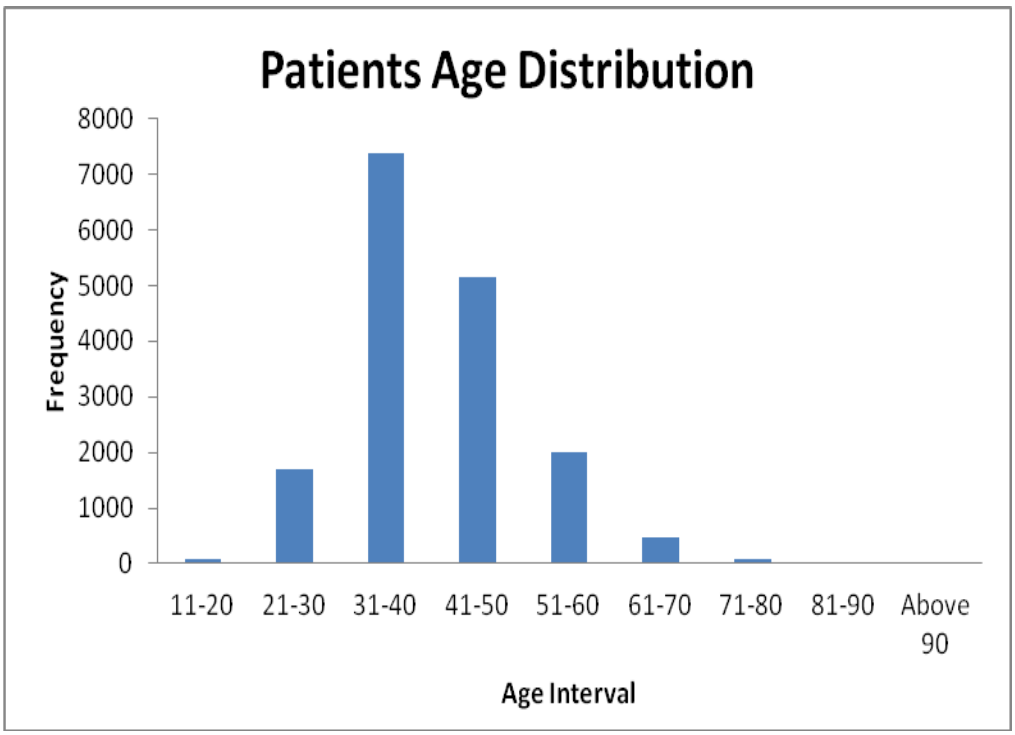

Figure 1: Bar Chart for Hepatitis B Patients Age distribution

Table 3 Distribution of Patients by Gender

	Frequency	Percent	Valid Percent	Cumulative Percent
Male	5747	34	34	34
Female	11140	66	66	100
Total	16887	100	100	

Table 3 showed the frequency distribution for gender of Hepatitis B patients recorded within the period of 2006 and 2015 under considerations. The table shows that female are greatly affected by the disease because 66% of the patients are female while just 34% of the patients are male. The descriptive statistics for table 3 is presented in table 4 while the bar graph for table 3 is presented in figure 2.

Table 4: Descriptive Analysis for Hepatitis B Patients Gender Distribution

Mean	Variance	Standard Deviation	Skewness	Kurtosis
1.6597	**0.225**	**0.47383**	**-0.674**	**-1.546**

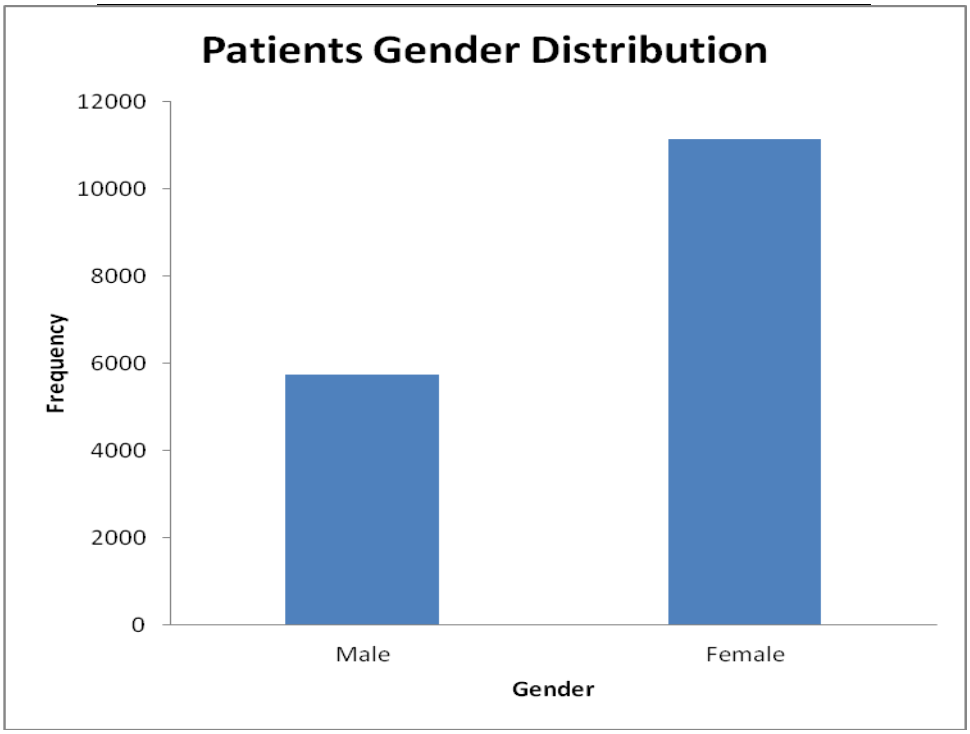

Figure 2: Bar Chart for Hepatitis B Patients Gender Distribution

3.1 *Chi-square Analysis*

Hypothesis:

H_o: Age and Gender of HBV patients are independent

H_a: Age and Gender of HBV patients are not independent.

Decision Rule: If p-value is less than the cut-off point of 0.05 significance level, reject the null hypothesis.

Table 5: Cross Tabulation of Outcome and Gender

Age Interval	Gender		Total
	F	M	
<20	54_a	27_a	81
21-30	1064_a	644_a	1708
31-40	4545_a	2826_a	7371
41-50	3461_a	1692_a	5153
51-60	1583_a	418_a	2001
61-70	105_a	370_a	475
>70	35_a	63_a	98
Total	11113	5747	16887

The subscript letter in each of the value denotes a subset of gender categories where the column proportion does not have significant different from each other at the .05 level.

Table 6: Pearson Chi-Square and Likelihood Ratio Estimates for Table 5

	Value	Df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.001[a]	6	0.027		
Continuity Correction[b]	0.008	6	0.028		
Likelihood Ratio	0.002	6	0.026		
Fisher's Exact Test				0.008	0.046
N of Valid Cases	16887				

This implies that the two factors (Age and Gender) are not independent which means that there is association between them since the p-values are less than 0.05.

4.0 Conclusion

From the analysis presented above, the descriptive analysis together with the bar charts of all variable considered for this work were presented. It was discovered that sexually active age bracket (31- 50) were greatly affected by the disease. Also, the research established that females have higher risk of contracting the disease when compared with male since 66% of the recorded data were female.

Test of significance was carried out using Chi-square method in order to established whether age and gender of those who tested positive to the disease are independent or not. It was discovered from the Chi-square test of independence that both variables (age and gender) are not Independent. It simply means there is an association between the two variables when it comes to the spread of Hepatitis B virus.

Acknowledgement

The financial support from Covenant University, Ota, Nigeria is greatly appreciated.

References

[1] Karthikeyan T 2013 Analysis of classification algorithms applied to hepatitis patients. *International Journal of Computer Applications*. 62(15):2530. DOI: 10.5120/10157-5032.
[2] World Health Organization. 2008, hepatitis B. World Health Organization Fact Sheet N_ 204. http://www.who.int/mediacentre/factsheets/fs204/en/index.html
[3] Locarnini S 2004 Molecular virology of hepatitis B virus. Semin. *Liver Dis*. 24 (Suppl. 1), 3–10.
[4] Medley G F, Lindop N A, Edmunds W J and Nokes D 2001. Hepatitis-B virus endemicity: heterogeneity, catastrophic dynamics and control, *Nat. Med.* 7, 619–624.
[5] Shepard C W, Simard E P, Finelli L, Fiore A E and Bell B P 2006 Hepatitis B virus infection: epidemiology and vaccination. *Epidemiol. Rev*. 28, 112–125.
[6] Goldstein S T, Zhou F J, Hadler S C, Bell B P, Mast E E and Margolis H S 2005 A mathematical model to estimate global hepatitis B disease burden and vaccination impact. *Int. J. Epidemiol*. 34, 1329–1339.
[7] Adamu P I, Oguntunde P E, Okagbue H I and Agboola O O 2018 On the Epidemiology and

Statistical Analysis of HIV/AIDS Patients in the Insurgency Affected States of Nigeria, *Open Access Macedonian Journal of Medical Sciences*; 6 (7): 1315-1321.

[8] Oguntunde P E, Adejumo A O and Okagbue H I 2017 Breast Cancer Patients in Nigeria: Data exploration approach, *Data in Brief.* 15: 47-57.

[9] Agresti A 2002 "*Categorical Data Analysis*", New York. Wiley

[10] Agresti A 2002. "*Inference for Contingency Tables*" Willey Series in Probability and Statistics.

[11] Watson G S 1959 Some recent results in &i-square goodness-of-fit tests, *Biometrics* 15, 440-468.

[12] Odetunmibi O A, Adejumo A O, and Sanni O O M. 2013 Loglinear Modelling of Cancer Patients Cases in Nigeria: An Exploratory Study Approach. *Open Science Journal of Statistics and Application*. 1, No. 1, pp. 1-7.

[13] Adejumo A O, Suleiman E A, Okagbue H I, Oguntunde P E and Odetunmibi O A 2017 Quantitative Evaluation of Pregnant Women Delivery Status' Records in Akure, Nigeria. *Data in Brief*; 16: 127-34. https://doi.org/10.1016/j.dib.2017.11.041 PMid:29201979 PMCid:PMC5699871