EVALUATING THE PERFORMANCE OF THE LOG-RANK TEST, GENERALIZED WILCOXON TEST AND THE COX-MANTEL TEST FOR EQUALITY OF TWO SURVIVAL CURVES

ORAGWAM, O. H.[1], ARONU, C. O.[2], UKA, C. O.[3], EKWUEME, O. G.[4] AND SOL-AKUBUDE, V. I. P[5]

[1]Department of Field Services and Methodology, National Bureau of Statistics, Nigeria
[2]Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State
[3]Department of Statistics, Federal College of Agriculture, Ishiagu, Ebony State, Nigeria
[4]Department of Industrial Production Engineering, Nnamdi Azikiwe University, Awka, Nigeria
[5]Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State

ABSTRACT

This study examined the performance of the Log-rank test, Generalized Wilcoxon test and Cox-Mantel test for equality of two Survival curves. The objectives of the study includes: to determine among Log-rank test, Generalized Wilcoxon test and Cox-Mantel test for equality of two Survival curves, to determine the test that performs best in terms of relative efficiency of the test statistic measure and to determine the most powerful test in terms of rejecting the null hypothesis when it is true. Simulation from Bernoulli distribution and Poisson distribution with sample sizes of 5, 25, 35, 45, 55, 65, 75, 85, 95 and 100 were used to evaluate the methods. The findings of the study showed that the Cox-Mantel test performed best in terms of relative efficiency of the test statistic measure while the Generalized Wilcoxon test was found to be the most powerful test in terms of rejecting the null hypothesis when it is true assuming $\alpha = 0.05$ and $\alpha = 0.10$.

Keywords: Bernoulli distribution, Log-rank test, Cox-Mantel test, Poisson Distribution

1. INTRODUCTION

In many research disciplines, including engineering and medical science, researchers are often interested in estimating the time until an event of interest occurs. In the medical field, one may be interested in estimating the time until death of the subject from the beginning of observation, for example birth, onset of a disease, entrance or start of a clinical trial. In an engineering concept, Butler (2011), focused on time until failure of an object, such as a pot or a light bulb. She researched on estimation of survival time, or how long the pots function until they fail. Gardiner (2010) defined Survival Analysis as a collection of methods for the analysis of data that involve the time to occurrence of some event, and more generally, to multiple durations between occurrences of different events or a repeatable (recurrent) event. Such events are generally referred to as failures though the event may, for instance, be the performance of a certain task in a learning experiment in psychology or a change of residence in a demographic study.

From the extensive use over decades of survival times in clinical and health related studies and failure times in industrial engineering (reliability studies), these methods have evolved to special applications in several other fields, including demography (analyses of time intervals between successive child births), sociology (studies of recidivism, duration of marriages), and labor economics (analysis of spells of unemployment, duration of strikes) (Gardiner, 2010).

Survival analysis is a collection of statistical procedures for data analysis, for which the outcome variable of interest is time until event occurs. It is also the study about Survival data. Survival data include survival time, response to a given treatment and patient characteristics related to response, Survival and the development of a disease. A common feature of survival data is censoring, it

means that the exact failure times of a number of subjects are not known. In analyzing survival data, the survival function and the hazard function are the two functions that are dependent on time. The survival function S (t) is the probability of surviving at least to time, t. While the hazard function h (t) is the conditional probability of dying at time t having survived to that time. The graph of S (t) (survival time) against t is called the survival curve. Survival curves are generated by Kaplan-Meier method or the product limit. Kaplan-Meier method is used for finding survival probabilities for censored and non-censored observations; it estimates the curve from the observed survival times. Survival to any time point is calculated as the product of the conditional probabilities of surviving at each time interval.

The aim of this study is to compare the performance of the Log-rank test, Generalized Wilcoxon test and Cox-Mantel test for equality of two Survival curves with the following specific objectives: to ascertain among Log-rank test, Generalized Wilcoxon test and Cox-Mantel test which performs best in terms of Relative Efficiency, to determine the most powerful test among Log-rank test, Generalized Wilcoxon test and Cox-Mantel test and to determine the effect of sample size on the performance of the methods.

2. LITERATURE REVIEW

Butler (2011) explained that Survival analysis is the study of lifetime distributions. This was demonstrated by employing the lifetime of pots, where lifetime is used to illustrate the period from the beginning of observation until failure. The time when the pots fail is referred to as the failure time. The failure time in her study is the day the pot stopped working. She used Akaike Information Criterion (AIC) to measure the relative goodness of fit of four models (Weibull model, Weibull change-point model, Gompertz model and Gompertz-Makeham model). From the results of the AIC, she found that the Gompertz Makeham distribution is considered the best fit for the distribution of the failure times for the pot data. However, the likelihood ratio test indicated that the Gompertz-Makehm model does not provide a significantly better fit than the Gompertz distribution, which was ranked second when looking at the AIC values. She concluded that the best model for the data is the Gompertz survival distribution, since the AIC value for the Gompertz model is smaller than the Weibull change-point model. She collected various covariates during her observation of the pots, such as average temperature, average voltage, and number of times the ratio in the bath dropped below threshold.

Silva and Vieira (2011) explaining the survival analysis technique said that "Survival analysis is a technique that allows study of the time passed before a certain event. It determines not only whether an event will occur but also when it should occur. The survival analysis results in two types of information: the survival function – proportion of the population that survives a certain time period, and the hazard function – proportion of the population which survived a certain period of time and will likely reach the terminal event in this period. Calculating the survival model uses the method of maximum likelihood with censored data left and right. They noted that the period of low income last up to t_i (duration observed in the sample) and the likelihood of this event is given by the survival function $S(t_i|X_i)$, with the vector of covariates being X_i.

Ramakrishnan and Ravanan (2013a) used the survival curve to give an idea about survival probabilities of male and female patients. Survival curve for male differs from the survival curve for female. They checked the significant difference between the sexes using the Log-rank test. Using Log-rank test regarding sex of the patients, the P-value is 0.367 which is greater than 0.05.

So, they accepted the null hypothesis that the survival distribution of male patients is same as that of female patients. Using Log-rank test, regarding treatment type of the patients in their study, the P-value was 0.811 which is greater than 0.05, they accepted the null hypothesis that the survival distribution of treatments is the same. They noted that Kaplan-Meier method of estimating survival curves only gives pictorial representation of the survival distributions but it does not take whole follow-up period into account. They also concluded that survival curves related to sex and to type of treatment which shows slight difference in the survival curves of corresponding survival distributions but Log-rank test determine there is no significant difference between these curves for both related to sex and treatment type.

Ramakrishnan and Ravanan (2013b), from the result of their study observed that the probability of survival of male patients is greater than that of female patients in the early period as described by the survival curve. But in the late period probability of survival of both are same and Median survival time for both groups are observed as 8.5 weeks. They stressed that the survival approach does not provide a comparison of the total survival experience of male and female groups. When using Log rank test and Cox-Mantel test, they observed that there is no significant difference between male and female patients survival. They explained further that the Log rank test and Cox-Mantel test are most likely to detect a difference between groups when the risk of an event is consistently greater for one group than another. Survival curves only gives pictorial representation but using non-parametric tests like Log rank test and Cox-Mantel test, researchers can easily identify the relation of survival distribution.

3. METHODOLOGY

3.1 Source of Data Collection and Description of data

The data used in this work were simulated using a Mintab 14.0 and STATASE version 9 for the distribution of different sample sizes. The sample sizes simulated were 5, 25, 35, 45, 55, 65, 75, 85, 95, and 100. The time of event (death) occurrence follows Poisson distribution since such events as death are rare (Bewick, et al., 2004, Lagakos, 2006) while the group (sex) follows the Bernoulli distribution because it's either male or female (Ramakrishnan and Ravanan, 2013a).

3.2 Survivor function and Kaplan-Meier estimator

Survival analysis is used to analyse data in which the time until the event of interest. The response is often referred to as a failure time, survival time or event time.Survivor function is the cumulative survival probability that an event will occur. Survivor function estimates the times up to time t that survive beyond the later time only if it survives until the earlier time and then survives the interval between the two times. Letting $S(t)$ to denote the survival function of T, then

$$S(t)=P(T\geq t)=1-F(t)=\int_{t}^{\infty} f(T)dT \qquad (1)$$

Where F (t) is the distribution function of the time parameter t and f(T) is the density function of the survival time T (Ramakrishnan and Ravanan, 2013b). The survival time T is the failure time with distribution F, density$f$. The Kaplan-Meier method of estimating survival curves gives a pictorial representation of the survival distributions but does not take the whole follow-up period into account (Butler, 2011).

Ramakrishnan and Ravanan (2013b), defined the Kaplan-Meier estimator by letting n be the total number of individuals whose survival time, censored or not, are available. Relabeling the survival times in order of increasing magnitude such that $t_1\leq t_2\leq\cdots\leq t_n$ and the values of r (risk set) are

consecutive integers 1, 2,.., (n-1) if there are no censored observation. The survival probabilities are calculated using

$$S(t) = \prod_{t_r \le t}^{n} \frac{(n-r)}{(n-r+1)} \quad (2)$$

Where r runs through those positive integers for which $t_r \le t$ and t_r is uncensored. The variance of S (t) is approximated by

$$\text{var}(S(t)) = [S(t)]^2 \sum_{r=1}^{n-1} \frac{1}{(n-r)(n-r+1)} \quad (3)$$

Here r includes those positive integers for which $t_{(r)} \le t$ and $t_{(r)}$ corresponds to death or event. (Ramakrishnan and Ravanan, 2013b).

3.3 Survival Tests

The three survival technique considered were Log-Rank test, Generalized Wilcoxon test and Cox-Mantel test in determining the most efficient test for equality of two survival curves. These tests were selected based on the fact that they are commonly used survival techniques by most researchers and the robustness of the techniques in determining the equality of two survival curves (Ramakrishan and Ravanan 2013a, 2013b; Gehan and Breslow, 1965).

3.3. 1. The Log Rank Test

The Log rank test is used to test the null hypothesis that there is no difference between the population survival curves (that is, the probability of an event occurring at any time point is the same for each population), (Bewick et al., 2004). This is a form of Chi-square test. It calculates a test statistic for testing the null hypothesis that the survival curves are the same for all groups, in other words, to test the null hypothesis that there is no difference between the populations in the probability of an event (death) at any time point. The Log-rank test statistic compares the two groups at each observed event time. It is constructed by computing the observed and expected number of events in one of the groups at each observed event time and adding them to obtain an overall summary across all the time points. For each time point the observed number of deaths in each group and the number expected if there was no difference are calculated. The number of expected in each is calculated as the proportion of subjects who are at risk at a given time point multiplied by the total number of events at that point. This test is most appropriate when the hazard functions are thought to be proportional across the groups if they are not equal. This test statistic is constructed by giving equal weights to the contribution of each failure time to the overall test statistic. The test statistic is given as;

$$\chi^2 (\text{Log-rank}) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \quad (4)$$

Equation (4) has approximately the Chi-square distribution with one degree of freedom, (Mantel, 1967; Bewick et al., 2004; Ramakrishnan and Ravanan, 2013a). A large Chi-square value would lead to the rejection of the null hypothesis in favor of the alternative that the two treatments are not equally effective at $\alpha = 0.05$.

Ramakrishnan and Ravanan (2013b) explaining the method noted that letting d_t be the number of deaths at time t and n_{1t} n_{2t} be the numbers of patients still exposed to risk of dying at time up to time t in the two treatment groups. The expected deaths for groups 1 and 2 at time t are

$$e_{1t} = \frac{n_{1t}}{n_{1t}+n_{2t}} \times d_t \quad (5)$$

$$e_{2t} = \frac{n_{2t}}{n_{1t}+n_{2t}} \times d_t \quad (6)$$

Then the total number of expected deaths in the two groups is given as $E_1 = \sum e_{1t}$ and $E_2 = \sum e_{2t}$ respectively. Let O_1 and O_2 be the observed numbers and E_1 E_2 the expected numbers of death in two treatment groups. The Log rank test is based on the same assumptions as the hazard ratio since the survival probabilities are the same for subjects early and late in the study, and the events happened at the time specified. The test is more likely to detect a difference between groups when the risk of an event is consistently greater for one group than another.

3.3.2. The Cox-Mantel Test

The Cox-mantel test is used to test the null hypothesis that there is no difference between the population survival curves. Let $t_1 \leq t_2 \leq \cdots \leq t_k$ be the distinct failure times in the two groups together and m_i the number of failure times equal to t_i or the multiplicity of t_i. Let R (t) be the set of individuals still exposed to risk of failure at time t, whose failure or censoring times are at least t. Let n_{1t} and n_{2t} be the number of patients in R (t) belonging to group 1 and 2, respectively (Ramakrishnan and Ravanan, 2013a, Ramakrishnan and Ravanan, 2013b). The total number of observations, failure or censored in $R(t_i)$ is $r_i = n_{1t} + n_{2t}$ and A_i is the proportion of r_i that belongs to group 2. Ramakrishnan and Ravanan (2013b) defined the Cox-Mantel test statistic as;

$$|U| = r_{(2)} - \sum_{i=1}^{k} m_i Ai \quad (7)$$

Where $r_{(2)}$ is the number of patients that belongs to the second group

$$I = \sum_{i=1}^{k} \frac{m_i(r_i - m_i)}{r_i - 1} A_i(1 - A_i) \quad (8)$$

Then from Equation (7) and (8), C the test statistic which is a standard normal variate under the null hypothesis is given as;

$$C = \frac{|U|}{\sqrt{I}} \quad (9)$$

3.3.3. The Generalized Wilcoxon Test (Breslow & Gehan Test)

The Wilcoxon test statistic is constructed by weighting the contribution of each failure time to the overall test statistic by the number of subjects at risk. Thus, it gives heavier weights to earlier failure times when the number at risk is higher which results to the test being susceptible to differences in the censoring pattern of the groups (Gehan, 1965, Breslow, 1970). Let n_{it} be the risk set (individuals exposed to risk at time t in the ith population (i = 1, 2)) and e_{it} be the observed event at time t in the ith population. Shen et al... (2013) defined that if $d_t = e_{1t} + e_{2t}$ and

$n_t = n_{1t} + n_{2t}$, then the expected events in the first and second population are computed as that of Equation (3.5) and (3.6) respectively. The test statistic is written as

$$\chi^2 (\text{Wilcoxon}) = \frac{O_1(O_1 - E_1)^2}{E_1} + \frac{O_2(O_2 - E_2)^2}{E_2} = \frac{\sum_{i=1}^{2} Oi(O_i - E_i)^2}{E_i} \quad (10)$$

The expected deaths in the two groups is given as $E_1 = \Sigma e_{1t}$ and $E_2 = \Sigma e_{2t}$ respectively. Let O_1 and O_2 be the observed numbers and E_1 E_2 the expected numbers of death in two treatment groups. The Wilcoxon test statistic defined above follows a Chi-square distribution with one degree of freedom (Gehan, 1965, Breslow, 1970).

4. DATA ANALYSIS AND RESULT

This section deals with data analysis and result of the analysis.

Table 1: Power of the techniques at 95 % and 90% confidence levels

SAMPLE SIZE	LOG RANK TEST		COX MANTEL TEST		WILCOXON TEST	
	α=0.05	α= 0.10	α=0.05	α=0.10	α=0.05	α=0.10
5	0.30	0.30	0	0	0.3	0.4
25	0	0.1	0	0	0.6	0.6
35	0.1	0.3	0	0	0.9	0.9
45	0	0	0	0	0.7	0.7
55	0	0	0	0	0.9	0.9
65	0.2	0.3	0	0	0.9	0.9
75	0	0	0.1	0.1	0.8	0.8
85	0.1	0.1	0	0.1	0.9	0.9
95	0.2	0.3	0	0	1	1
100	0	0	0	0	0.8	0.8

The result obtained in table 1 represents the power of the techniques at 95% and 90% confidence level, the reason for testing the power at 0.05 and 0.10 levels is because it was observed that the Cox-Mantel test and Log-rant test were insignificant at 95% confidence level and significant at 90% confidence level.

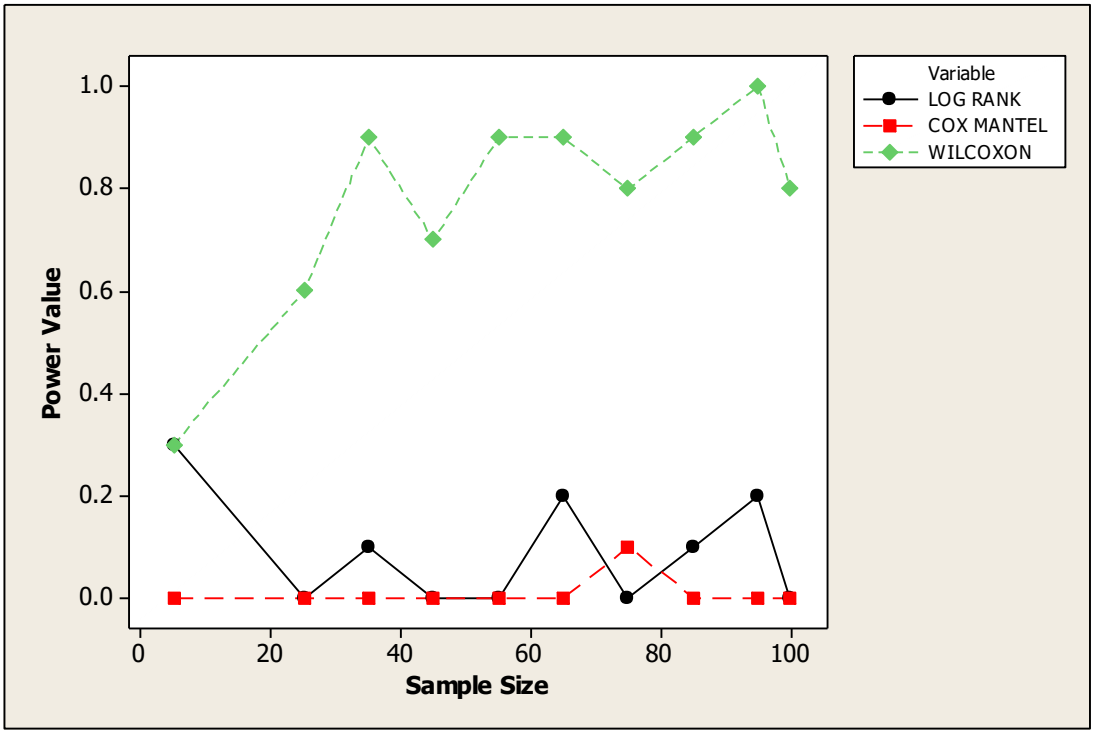

Figure 1 Distribution of Power of the test at α=0.05

Figure 1 revealed that Wilcoxon and Log-rank has the same power at sample size 5 while from sample size 25 through 100 the Wilcoxon test was found to be more powerful. It was equally observed that the Log-rank test has better power than the Cox-Mantel test at sample points 5, 35, 65, 85 and 95 while the Cox Mantel test has a better power over the Log-rank test at only sample size 75. Based on the observation, the Wilcoxon test was found to be the most powerful test followed by the Log-ranks test at α=0.05.

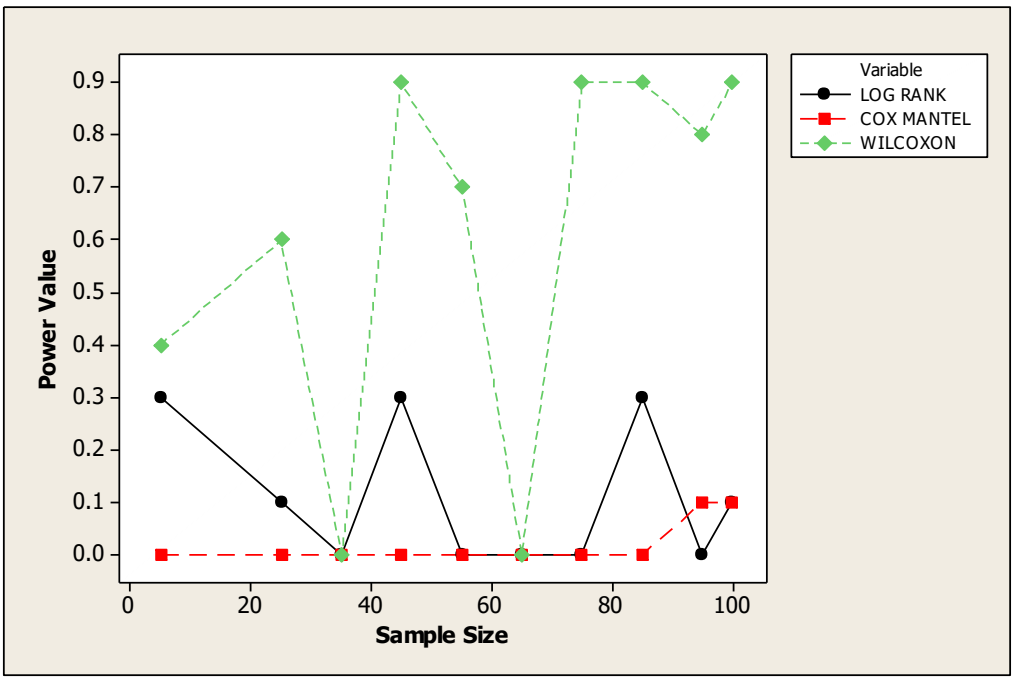

Figure 2: Distribution of Power of the test at α=0.10

Figure 2 revealed that Wilcoxon was found to have a greater power than the Log-rank test and the Cox-Mantel test across the sample sizes. It was observed that the Log-rank test has better power than the Cox-Mantel test at sample points 5, 25, 35, 65 and 95 while the Cox Mantel test has a better power over the Log-rank test at only sample size 75. Based on the observation, the Wilcoxon test was found to be the most powerful test followed by the Log-ranks test at α=0.10.

Table 2: Relative Efficiency of the Test Statistic values of the three methods

Sample size	Log-rank	Cox-Mantel	Wilcoxon
5	1.95657	0.71892	3.943
25	1.07564	0.54267	12.92
35	3.65774	0.4631	59.702
45	0.85741	0.78782	19.629
55	0.3771	0.45384	8.759
65	2.60096	0.4017	90.681
75	0.96373	0.6264	38.69
85	1.55674	1.0361	174.956
95	2.66152	0.46761	122.871
100	1.06994	0.64444	57.116

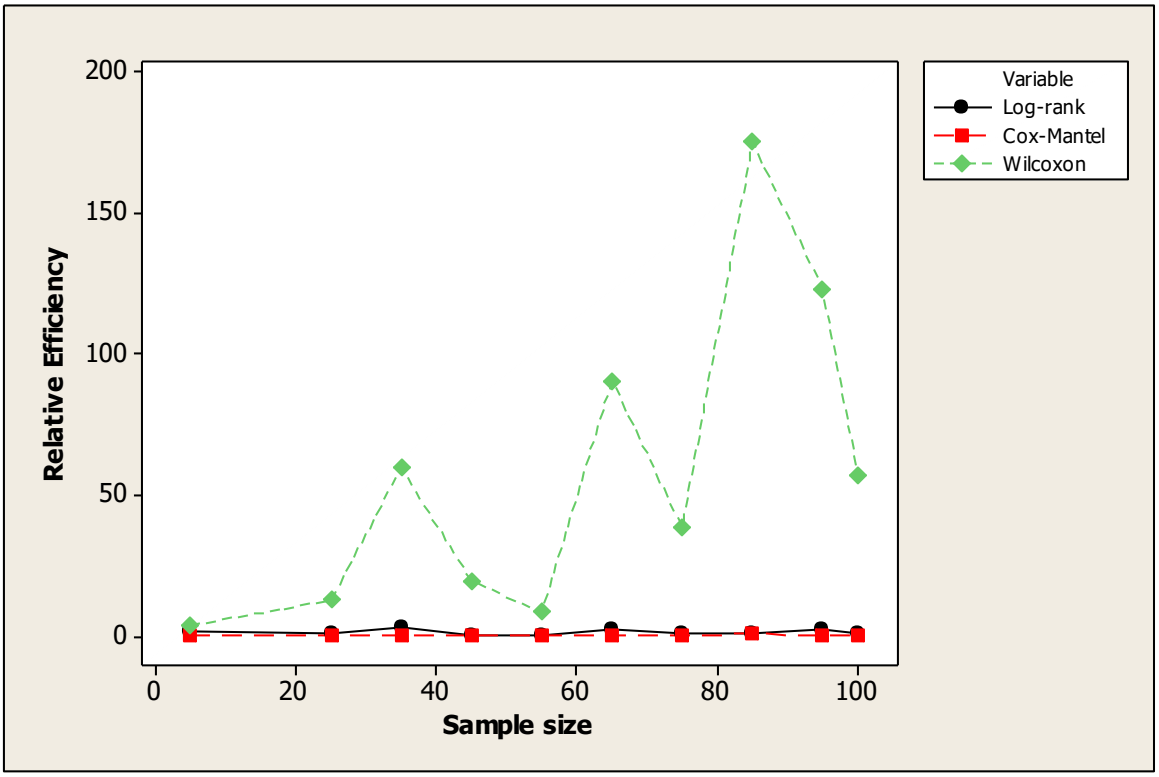

Figure 3: Distribution of Relative Efficiency for test statistic measure for the Log-rank, Cox-Mantel and Wilcoxon test.

Figure 3 revealed that Cox-Mantel test is more efficient than the Log-rank test at sample points 5, 25, 35, 65, 75, 85, 95 and 100 since it has the least standard deviation at the aforementioned

sample points. Also, figure 3 showed that Cox-Mantel and Log-rank test are relatively more efficient than the Wilcoxon test since the Cox-Mantel test and the Log-rank test recorded the least standard deviation. This implies that the Cox-Mantel test is the most relatively efficient followed by the Log-rank test.

5. CONCLUSION

This study compared the efficiency of the Log-rank test, Generalized Wilcoxon test and Cox-Mantel test for equality of two Survival curves. Findings of the study showed that the Cox-Mantel test performed best in terms of relative efficiency of the test statistic measure.

In addition, the generalized Wilcoxon test was found to be the most powerful test in terms of rejecting the null hypothesis when it is true. Hence, we conclude that users should employ the Cox-Mantel test when interest is on attaining the relative efficiency and the Generalized Wilcoxon test when interest is on employing a more powerful test in rejecting the null hypothesis when it is true for equality of two survival curves.

The present study compared the Log-rank test, Generalized Wilcoxon test and Cox-Mantel test for equality of two Survival curves. We recommend comparison of the discussed methods for equality of more than two survival curves as area for further research.

REFERENCES

Bewick, V., Cheek, L. and Ball, J. (2004). Statistics Review 12: Survival Analysis. Critical Care, 8(5): 389-394.

Breslow, N. E. (1970). A Generalized Kruskal–Wallis Test for Comparing K Samples Subject to Unequal Patterns of Censorship. Biometrika 57: 579–594.

Butler, E. L. (2011). Estimating the Survival Distribution of Aluminum Processing Pots, Dietrich College Honors Theses, Pape 113.

Gardiner, J. C. (2010). Survival Analysis: Overview of Parametric, Nonparametric and Semi parametric Approaches and New Developments. Statistics and Data Analysis, SAS Global Forum, Paper 252.

Gehan, E. A. (1965). A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples, Biometrika, 52(1/2), 203-223.

Lagakos, S. W. (2006). Time-to-Event Analyses for Long-Term Treatments: The APPROVe Trial. N Engl J Med 355, 113-117.

Mantel, N. (1967) Evaluation of survival data and two new rank order statistics arising in its consideration, Cancer Chemotherapy Rep., 50, 163-170.

Ramakrishnan, M. and Ravanan, R. (2013a). Estimation of Survival Distribution Using R Software. International Journal of Scientific and Research publication, 3(4), 1-5.

Ramakrishnan, M. and Ravanan, R. (2013b). Non-Parametric Methods for Comparing Two Survival Distributions, International Referred research Journal, 4(2), 121-125.

Shen, C., Hauang, J. and Lee, C. (2013). Weighted Wilcoxon-Type Rank Test for Interval Censored Data. Journal of Applied Mathematics, 10(1): 2-9.

Silva, J.F. and Vieira, A.C. (2011). Duration of Low Wage Employment: A study Based on Survival model. IZA Discussion Paper No. 5972, pages 5-11.