IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 05-11
www.iosrjournals.org

Parametric and Non-Parametric Approach On Breast Malignancy Patients

N. Paranjothi, A. Poongothai and G. Manimannan
Associate Professor, Department of Statistics, Annamalai University, Chidambaram.
Assistant Professor, Department of Statistics, Muthayammal Arts and Science college, Rasipuram, Salem.
Assistant Professor, Department of Statistics, Apollo Arts and Science College Chennai.

Abstract: *This study examines the efficacy of appropriate use of parametric and nonparametric statistical methods in research regarding the Health Sciences. Data from such research often fails to meet one or more of the assumptions of traditional parametric tests, thus necessitating the use of nonparametric techniques. To this end, this study analyzed the effects of using both parametric and nonparametric survival methods on a dataset of 300 patients from a private hospital in Bangalore. The data collected included socio-demographic parameters, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Oxygen Saturation (SPo2), and other breast cancer clinical parameters. To compare the results, the study employed the t and z tests of the parametric test and the log-rank test, Gehan Wilcoxon test, and Cochran Mandel Hensel test of the nonparametric survival test. It was found that the use of different statistical approaches led to different statistics, but the ASA grade clinical parameter produced similar results regardless of the methods applied.*
Keywords: *Parametric and Non Parametric Tets, t-test, z-test, SurvivalStatistics: log rank Test, Cohran Mantel Haenszel Test and Gehan Wilcoxon test.*

I. INTRODUCTION

Cancer is the result of healthy cells in an organ growing and changing uncontrollably, leading to the formation of a mass or sheet of cells known as a tumor. Tumors can be either cancerous (malignant) or benign. Malignant tumors can grow and spread to other parts of the body, whereas benign tumors are limited in their growth. Breast cancer is a malignant cell growth located in the breast, and if left untreated, it can spread to other areas of the body. This article will discuss non-invasive (Stage 0) and early and locally advanced invasive breast cancer (Stages I, II, and III). The stage of breast cancer describes the size and spread of the cancer. While it usually spreads to nearby lymph nodes, breast cancer is still classified as a local or regional disease if it has spread further to places like the bones, lungs, liver, and brain. This is referred to as metastatic, or Stage IV, breast cancer, which is the most advanced stage of the disease.

Breast cancer can be invasive or non-invasive. Invasive breast cancer is a type of cancer that has spread to surrounding tissues and/or distant organs, whereas non-invasive breast cancer does not spread beyond the milk ducts or lobules in the breast. There are various types of breast cancer and each is classified based on its appearance under a microscope. The most common type of breast cancer is non-invasive or tubal cancer. This is further divided into two categories: Ductal Carcinoma in Situ (DCIS) and invasive ductal carcinoma. DCIS is a non-invasive form of cancer (stage 0) which is confined to the duct and has not spread beyond it. Invasive ductal carcinoma, on the other hand, is a less common type of breast cancer which spreads outside of the ducts or lobules.

II. BACKGROUND OF THE STUDY

Parametric and nonparametric methods [1] identified a significant treatment effect in three of the five tests. Parametric analysis detected a significant difference in the likelihood of cure between treatment groups in only one of these trials; however, a significant difference in recurrence-free survival time was found between the treatment groups in the remaining two trials. Additionally, three-parameter survival models yielded similar results.

This article by Sujata [2] provides a comprehensive review of techniques used to analyze survival data from retinal health studies. Popular non-parametric techniques such as the Kaplan-Meier survival plot and the Cox proportional hazards model are discussed and illustrated using the Diabetic Retinopathy Study (DRS) data. Advanced topics, such as dimensional proportional hazards models and accelerated failure time models, a new measure of treatment effect, are also addressed. In this chapter, Jianqing Fan and Jiancheng Jiang [5] present a thorough analysis of non-parametric modeling methods, focusing on direction of variance smoothing using Cox-type models. In addition, model fitting, variable selection, and hypothesis testing issues are also discussed.

Sarada Ghosh et al. [12] investigated the risk factors associated with COVID-19 mortality, such as age, gender, number of comorbidities, and access to health care, using a non-parametric approach. They discovered that both gender and age had a major influence on mortality. Furthermore, they discussed how public health risks resulting from false assessments, conflicting messages, or misinformation can lead to increased awareness and panic outbreaks of COVID-19. When estimating conditional survival functions [15], non-parametric estimators are usually preferred over parametric and semi-parametric estimators due to their relaxed assumptions and robust estimation. However, at small sample sizes, parametric and semi-parametric estimators may have better performance properties than non-parametric estimators due to their lower variance. A simulation study showed that layered survival models consistently perform well in a wide range of scenarios by combining the strengths and weaknesses of different survival models. Additionally, stratified survival models performed better than the model selected by cross-validation. Finally, stratified survival models were used in the well-known German Breast Cancer Study.This research paper examined the selective effects of appropriate use of parametric inferential statistical methods and nonparametric survival methods.

III. DATABASE

This study focuses on the application of Survival Analysis to breast cancer, particularly in Primary Health Checkups (PHC). Data was collected from a secondary source containing 300 patients from a Private Hospital in Bangalore. This data included socio-demographic variables, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Oxygen Saturation (SPo2), and breast cancer related clinical parameters. The PHC data was obtained from the Out Patients Department (OPD) of Private Hospital, Bangalore. The case sheet for each patient included socio-demographic characteristics, SBP, DBP, SPo2, ASA grade. An overview of the database, selected PHC parameters, and survival analysis models will be presented in the following section.

IV. METHODOLOGY

Statistical hypothesis testing is a pivotal concept in statistics, utilized by statisticians, machine learning experts and data scientists alike. This type of testing involves using statistical tests to examine whether the null hypothesis can be rejected or not. Generally, these tests assume there is no relationship or difference between groups. Parametric tests are based on the set parameters used to generate the probability model. For example, the T-test, Z-test, F-test, and ANOVA all adhere to the assumption that the population is normally distributed, or can be approximated to a possible normal distribution through the Central Limit Theorem. On the other hand, non-parametric tests do not suppose any prior knowledge of the population distribution. These tests are also known as 'distribution-free' tests due to the lack of fixed parameters and available distributions, such as the normal distribution. Examples of non-parametric tests are the chi-square test, Mann-Whitney U-test, Kruskal-Wallis H-test, log rank test, Gehan Wilcoxon test, and Cochran Mantel Henszel test. In the following flow diagram shows that parametric and non-parametric test and their statistical tests (Figure 1).

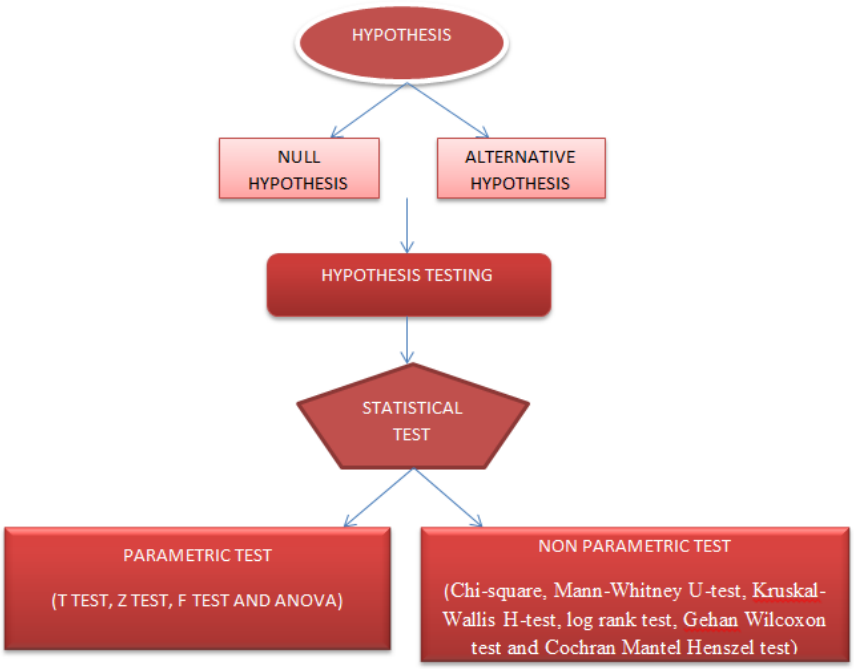

Figure 1. Parametric and Non parametric Flowchart

Parametric testing involves assuming parameters and being aware of the population distribution. Conversely, non-parametric testing does not make any assumptions. This study will compare the results of both tests and explain any differences.

4.1 PARAMETRIC METHODS

4.1.1 t -TEST

In statistical tests, the probability distribution of statistics is important. When samples are drawn from a population N (μ,σ^2) with sample size n , the distribution of the sample mean $\bar{X}$ must be a normal distribution N ($\mu,\sigma^2/n$).Under the null hypothesis $\mu = \mu 0$, the distribution of the statistic $z=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$should be standardized to a normal distribution. When the variance of the population is unknown, it is possible to substitute with the sample variance s^2.In this case, the statistic $\frac{\bar{X}-\mu_0}{s/\sqrt{n}}$follows a t distribution (n−1 degrees of freedom). An independent-group t test can be performed to compare means between two independent groups, along with a paired t test for paired data.Since the t test is a parametric test, the samples must meet certain prerequisites such as normality, equal variances, and independence.Sampling variance is instead of population variance to determine the sampling distribution of the mean. The sample variance is defined as:

$$s^2=\frac{\sum(Xi-\bar{X})^2}{n-1}$$

In cases where sampling variance is used, the sampling distribution follows a t distribution, which depends on each sample's 0 degrees of freedom rather than a normal distribution.If two samples show a normal distribution and have equal variance, the t statistic is:

$$t=\frac{(\bar{X1}-\bar{X2})-(\mu_1-\mu_2)}{s(1+2)\sqrt{\frac{1}{n1}+\frac{1}{n2}}} \sim t(n_1+n_2-2)$$

The population mean difference (μ1 - μ2) is assumed to be 0; Thus:

$$t=\frac{(\bar{X1}-\bar{X2})}{s(1+2)\sqrt{\frac{1}{n1}+\frac{1}{n2}}}$$

The population variance was unknown, so the pooled variance of the two samples was used:

$$s_{(1+2)}{}^2=\frac{(n1-1)s_1{}^2+(n2-1)s_2{}^2}{n1+n2-2}$$

However, if the population variances are not equal, the t statistic of the t test will be

$$t=\frac{(\bar{X1}-\bar{X2})}{\sqrt{\frac{s_1{}^2}{n1}+\frac{s_2{}^2}{n2}}}$$

4.1.2 Z- Test

A Z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large (N<30).The test statistic is assumed to have a normal distribution, and the nuisance parameters such as the standard deviation must be known for an accurate z-test to be performed.Z-tests are closely related to t-tests, but t-tests perform better when t-tests have a small sample size, i.e. less than 30. Also, t-tests assume that the standard deviation is unknown, while z-tests assume that it is.

$$z=\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$$

The z statistic follows a normal distribution.

4.2NON PARAMETRIC APPROACH

4.2.1. Log Rank Test

A log-rank test is a non-parametric hypothesis test to compare the survival trend of two or more groups when there are censored observations. It is widely used in clinical trials to compare the effectiveness of interventions when the outcome of an event is timed.This test was first proposed by Nathan Mandel and named the log-rank test by Richard and Julian Peto. It is also known as the Mandel-Cox test and can be thought of as a time-adjusted version of the Cochran-Mantel-Haenszel chi-square test.The null hypothesis for the test is that there is no difference in the survival experience of the subjects in the different groups being compared. Its name derives from its association with a test that uses the logarithm of the ranks of the data.

To calculate log rank statistics:

$$Z=\frac{(Oi-Ei)2}{Var\,(Oi-Ei)}$$

4.2.2 Gehan Wilcoxon Test

Gehan's generalized Wilcoxon test (1965) is a generalizationWilcoxon's two-sample rank test for censored data. The Wilcoxon statistic is often introduced from different perspectives and is mainly considered as a linear standard statistic or U-statistic. For survival data analysis, it is also considered as a weighted Mandel-Haenszel statistic.Different approaches to censoring have led to different versions of the generalized Wilcoxon statistic. Although the Gehan–Wilcoxon statistic is widely used in practice and available in many statistical software packages, it is inappropriate for censored data and the Wilcoxon statistic has been criticized as inferior to the Peto–Peto generalization.Wilcoxontest is based on statistics$U_w=\sum_{j=1}^{r} nj(d_{ij}-e_{1j})$ with variance $V_w=\sum_{j=1}^{r} n_j^2 v_{ij.}$

Hence the GehanWilcoxon test statistic

$$\frac{U_w^2}{V_w}$$

It has an asymptotic chi-square distribution with one degree of freedomUnder the null hypothesis.

4.2.3 Cochran Mantel Haenzel Test

To test for conditional independence, examine the associations in the partial table 2X 2 X K Tables. Note that the null hypothesis levels of conditional independence are equivalent to the statement of all conditional odds ratios given equal to 1.

$$H_0: \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(k)} = 1$$

Cochran-Mantel-Haenszel (CMH) test statistic

$$M^2=\frac{[\sum_k (n_{11k}-\mu_{11k})]2}{\sum_k Var\,(n_{11k})}$$

Where $\mu_{11k}=E(n_{11})=\frac{n1+kn+1k}{n++k}$ is the expected frequency of the first cell in the kth partition table holds the conditional independence and variance of the cell (1, 1) is.

$$Var\,(n_{11k})=\frac{n1+kn2+kn+1kn+2k}{n++k2(n++k-1)}$$

V. RESULT AND DISCUSSION

The normal distribution is a widely-used probability function in statistics which describes how data values are distributed. When a collection of data is randomly gathered from independent sources, it is usually seen to be distributed normally. This is shown on a bell curve graph, where the peak represents the mean of the data set and half of the values lie to the left of the mean and the other half to the right. The density function of the normal distribution curve, cumulative density function and the generation of a vector of normally distributed random numbers are all demonstrated in Figure 2.

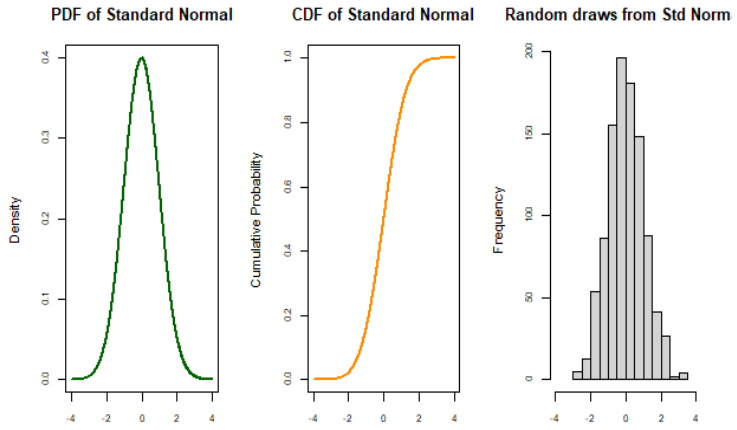

Figure 2. PDF, CDF and A Vector Construction of a Histogram of Normally Distributed Random Numbers

Histograms are an effective tool for understanding the shape of a data distribution. They provide an unbiased view of the frequency of each value found in the data set. Histograms are a graphical representation of a data set that displays how often each value appears. The x-axis is divided into bins, and the height of the bar reflects the number of values in the data set that fall into that bin. Figure 3 shows a histogram with a normal distribution overlay, depicting light blue bands with a black border and a dark blue thick density plotline. This highlights the shape of the data set.

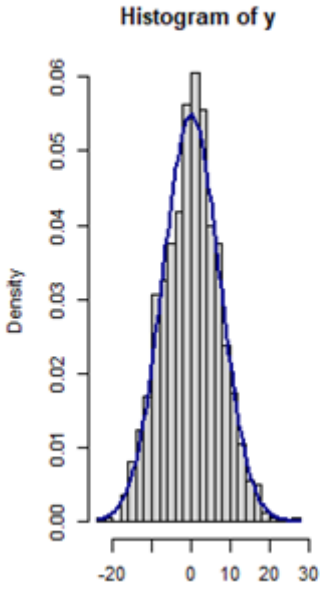

Figure 3. Histogram with a Normal Distribution

Hypothesis testing is a fundamental concept in statistics, which is frequently implemented by statisticians, machine learning professionals, and data scientists. This technique uses statistical tests to determine whether or not the null hypothesis (H_0) is accepted or rejected. H_0 states that there is no significant difference between ASA Grade among the Breast Malignancy Patients. Alternatively, if H_1 is accepted, it implies that there is a significant difference between ASA Grade among the Breast Malignancy Patients. Statistical tests used in hypothesis testing can be either parametric or non-parametric.

Parametric Tests	Test Value	P-Valie	95% Confidence Interval
t - test	-91.839	<0.022	-51.76977 to -4960199
Z - test	-41.385	<0.022	-53.08634 to -48.28542

The mean value of t test and Z test for ASA grade I was 99.96401 and ASA grade II was 150.64989. The test value for t test is -91.839 and Z test is -41.385. The calculated value of p for t test and Z test is 0.022 which is much smaller than the tabulated value of p is 0.05 so reject the null hypothesis. There is a significant difference between ASA grades among breast cancer patients.

Non-Parametric Test	Chi-Square Value	P-Value
Log Rank Test	59.7	0.01
Gehan Wilcoxon Test	44.4	0.02
Cochran Mantel Hanszel Rest	272.64	0.22

Chi-square values for non-parametric test are 59.7, 44.4 and 272.64 respectively. The calculated p value for all three non-parametric tests (0.01, 0.02 and 0.022) is smaller than the tabulated p value of 0.01. So reject H_0 and accept the alternative hypothesis that there is a significant difference between ASA Grade among patients with breast cancer.

VI. CONCLUSION

This study explores the applications of survival analysis in health, specifically the Primary Health Check (PHC) of Breast Malignancy patients, by using parametric inferential statistical methods and non-parametric survival methods. Data was collected from a sample of 300 patients at St. John's Hospital, Bangalore and involved socio-demographic characteristics, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Oxygen Saturation (SPo2) and breast cancer relatedclinical parameters. The normal distribution is a probability function used to understand the shape of a data set, illustrated by a bell-curve graph showing the mean of the data set and the number of values to the left and right of the mean. Histograms are then used to identify which values are more common and which are less common. The aim of this research was to evaluate the comparison between parametric, non-parametric and statistical tests for breast malignancy patients. The mean values of t test and Z test for ASA grade I was 99.96401 and ASA grade II was 150.64989. The calculated value of p for t test and Z test was 0.022, which was much smaller than the tabulated value of p (0.05), so the null hypothesis was rejected. Chi-square values for non-parametric test were 59.7, 44.4 and 272.64 respectively, with a calculated p value of 0.011 and 0.022, which was smaller than the tabulated p value of 0.01. The alternative hypothesis was accepted, concluding that there is a significant difference between ASA Grade among patients with breast malignancy.

REFERENCES

[1]. J W Gamel, R L Vogel, P Valagussa, G Bonadonna, Parametric survival analysis of adjuvant therapy for stage II breast cancer, National Library of Medicine, 1994 Nov 1;74(9):2483-90.
[2]. Sujatha.v, Parametric and non-parametric Approaches on Survival Analysis for Diabetic Retinpathy Data, Research Journal of Pharmacy and Technology (RJPT), Volume 9, issue 4, 2016.
[3]. WHO, Coronavirus disease 2019 (COVID-19) situation report-61 (2020)
[4]. Gamel JW, Vogel RL, Valagussa P, BonadonnaG.Cancer. Parametric survival analysis of adjuvant therapy for stage II breast cancer.,1994 Nov 1;74(9):2483-90. 3.PMID: 7923004
[5]. Jianqing Fan and Jiancheng Jiang, Non and Semi Parametric Modeling in Survival Analysis, Research Gate, April 2009.
[6]. H. Akaike. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60 (1973), 255–65.
[7]. P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, Statistical Models Based on Counting Processes, Springer-Verlag, New York 1993.
[8]. M. Aitkin and D. G. Clayton. The fitting of exponential, Weibull and ex- treme value distributions to complex censored survival data using GLIM. Appl. Statist. 29 (1980), 156–163.
[9]. O. E. Barndorff-Nielsen and D. R. Cox. Asymptotic Techniques for Use in Statistics. Chapman & Hall, 1989, page 252.
[10]. J. M. Begun, W. J. Hall, W.-M. Huang, and J. A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist. 11 (1982), 432–452.
[11]. J W Gamel [1], R L Vogel, Comparison of parametric and non-parametric survival methods using simulated clinical data,Stat Med, 1997 Jul 30;16(14):1629-43.
[12]. Sarada Ghosh, GuruprasadSamanta, Juan J. Nieto, Application of non-parametric models for analyzing survival data of COVID-19 Patients, Journal of Infection and Public Health, 14 (2021), 1328-1333.
[13]. K. Kelland, Italy sewage study suggests COVID-19 was there in December 2019 HEALTHCARE & PHARMA (2020)
[14]. Gamel JW, Vogel RL.Br J Cancer. A model of long-term survival following adjuvant therapy for stage 2 breast cancer.1993 Dec;68(6):1167-70.
[15]. Andrew Wey, John Connett, Kyle Rudser**,** Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models**,** *Biostatistics*, Volume 16, Issue 3, July 2015, Pages 537–549
[16]. HaybittleJL.ClinOncol (R CollRadiol). Life expectancy as a measurement of the benefit shown by clinical trials of treatment for early breast cancer. 1998;10(2):92-4.
[17]. Multi-agent chemotherapy for early breast cancer.Early Breast Cancer Trialists' Collaborative Group.Cochrane Database Syst Rev. 2002;(1):CD000487.
[18]. Schmid P, PossingerK.MedKlin (Munich). [High-dose chemotherapy in breast cancer].2002 Nov 15;97(11):677-86.

[19]. Boonstra P. S. Taylor J. M. G. Mukherjee B. (2013). Incorporating auxiliary information for improved prediction in high-dimensional datasets: an ensemble of shrinkage approaches Biostatistics 14, 259 – 272.
[20]. Bou-Hamad I.Larocque D. Ben-Ameur H. (2011). A review of survival trees Statistics Surveys 5 44–71.