Journal of Data Science 14(2016), 539-552

A Study of Age Distribution of Prostate Cancer Detection

Suparna Basu[1*], Sanjay Kumar Singh1 and Umesh Singh[1,2]

[1] *Department of Statistics, Banaras Hindu University*
[2] *DST - CIMS, Banaras Hindu University*

Abstract: This paper aims to propose a suitable statistical model for the age distribution of prostate cancer detection. Descriptive studies suggest the onset of prostate cancer after 37 years of age with maximum diagnosis age at around 70 years. The major deficiency of descriptive studies is that the results cannot be generalized for all types of populations usually having non-identical environmental conditions. The proposition follows by checking the suitability of the model through different statistical tools like Akaike Information Criterion, Kolmogorov-Smirnov distance, Bayesian Information Criterion and $\chi 2$ statistic. The Maximum likelihood estimate of the parameters of the proposed model along with their asymptotic confidence intervals have been obtained for the considered real data set.

Key words: Age distribution, Cumulative hazard rate, Prostate cancer, Statistical modelling, Q-Q plot.

1. Introduction

The prostate is a walnut-sized exocrine gland of the male reproductive system. Prostate cancer, also known as carcinoma of the prostate, is the development of cancer in the prostate gland. It is a disease of higher ages with incidence rarely before 37 and 67-70 being the highest risk age for a male (Li et al (2012); American Cancer Society (2014); Cancer research UK (2015)). The incidence of prostate cancer rapidly increases after the age of 55 years (Cancer research UK (2015); Ries et al. (2008); Howlader et al. (2014); Grőnberg (2003)). During the past few decades, the median age at diagnosis has been steadily decreasing due to the widespread introduction of Prostate-Specific Antigen (PSA) screening. In USA (Surveillance Epidemiology and End Result (SEER) data), the median age at diagnosis of prostate cancer dropped from 72 years in 1986 to 66 years in 2011 (Surveillance epidemiology and End Results program (SEER) (2014)). Factors that increases the risk of prostate cancer includes: older age, a family history of the disease, race, BMI, etc. . In United States, it is more common in the African American population than the Caucasian population and is the second leading cause of death among American men, next to lung cancer. The estimated new cases of prostate cancer in US in 2014 is 2,33,000, with estimated deaths being 29,480. Based on 2009-2011 data, it is concluded that approximately 15.0% of men

* Email : basusuparna91@gmail.com

will be diagnosed with prostate cancer at some point during their lifetime. Prevalence of this cancer in 2011, was estimated to be 2,707,821 (National Cancer Institute at the National Institute of Health (2014)). The early stages of prostate cancer are devoid of much symptoms. Yet, proper surveillance and screening can be of good help for a patients survival as early detection can improve life expectancy and quality of life (Moul et al. (2003); American Cancer Society (2014)). The most com- mon diagnostic techniques are Urine test, PSA test and prostate biopsy where only biopsy confirms prostate cancer with certainty (National Cancer Institute (2015)). The overall survival of young adults with respect to older population after intervention was observed to be significantly better (Lin et al. (2009)). Also, the disease progression was found to be inhibited due to early diagnosis and subsequent interventions (Moul et al. (2003); Anderson and Sternberg (2008)). The patients have better survival rates in absence of metastasis but they still remains at risk of death even after 15 years of disease detection (Aprikian et al. (1994)). The above studies clearly establishes the impact of age at diagnosis of prostate cancer and the patients course of life after detection. Earlier studies were emphasized on the descriptive and non parametric analysis of the effect of age on cancer grades and its severity with young age at diagnosis. Descriptive work certainly has its place in science but such inferences are (or should be thought of as) more shallow and tentative and intends to describe a big hunk of data with summary charts and tables only, but do not provide a platform to draw conclusions based on small samples about the population from which the samples come. Models play a critical role in sta- tistical data analysis since it helps in gaining approximate information speedily, cheaply and perhaps with greater accuracy about the population, contrary to de- scriptive studies. Once a model is identified, various forms of inferences such as prediction, control, information extraction, knowledge discovery, validation, risk evaluation, and decision making can be done within the framework of inductive or deductive arguments. In addition to the above qualities, models also wards off the requirement of large dataset for sound analysis and inference. Thus, the key to solving complex real-world problems lies in the development and construction of a suitable model. Keeping these points in mind, the present study is aimed at proposing a statistical model for the age distribution of prostate cancer where the random variable can be explained as the number of years lived by a male before being diagnosed with prostate cancer and obtain the mean and median age at diagnosis of the same for a given stage. Mathematically, if X denotes the age at which a patient is diagnosed with prostate cancer, we wish to develop a statistical model for X as cumulative distribution function (CDF) $F(\cdot)$ of X where :

$\mathbf{F(x) = P(X < x) = P}$(A male is diagnosed with prostate cancer before age x).

2. Data Description

The data for the present study was obtained from The Surveillance Epidemiology and End Result (SEER) program. In this study, we have considered the Black patients diagnosed during the year 1988-2003 with Malignant prostate cancer, prostate being the primary site of cancer invasion. The different stages of a prostate cancer behaves unlikely to the others and thus the stages were tackled separately. The stages are defined as stage I, stage II, stage III, stage IV by the SEER modified AJCC stage 3rd edition. The variable of interest is the age at diagnosis of the

disease for a patient diagnosed through some standard medical technique. The data is assumed to be the complete prostate cancer population of the nine cancer registries of SEER (SMSA, Connecticut, Metropolitan Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah and Metropolitan Atlanta). The num- ber of persons diagnosed with prostate cancer in stage I is 7769, 3277 in stage II, 2778 in stage III and 4194 in stage IV respectively.

3. Methodology and Estimation

It has been earlier studied that all the stages if pooled together for the survival analysis of a patient for a particular course of treatment yields non-conforming results (Chan and Tsokos (2013); Osei et al. (2013)). These studies suggested stage-wise division as a necessity for any trivial analysis of life time data related to Prostate Cancer. Keeping in mind these conclusions of the past studies, we have modelled and analysed the age at diagnosis for a prostate cancer patient at different stages namely stage I, stage II, stage III and stage IV respectively. The data in hand is the number of years of survival of a person before he is diagnosed with Prostate cancer which can be considered as disease free life time data. To propose a suitable model for such data, firstly we need to identify an appropriate probability density function (PDF) that characterizes the age at diagnosis of the disease. Model formulation demands explicit study of the nature of data. The data characteristics like the histogram (Silverman (1986)), Empirical cumulative distribution function(ECDF), five point summary measures, quantile-quantile plot (Wilk and Gnanadesikan (1968)), empirical hazard rate (Kapur and Lamberson (1977)) were studied, although the initial guess of a suitable model rested solely on the nature of observed hazard. Histogram is very powerful exploratory technique for empiri- cal study, but the choice of the class intervals plays crucial role in determining the nature of data. Often, the empirical hazard fails to clearly distinguish between a non-decreasing and a non-increasing hazard. We obtained the Nelson-Aalen estimate (Lawless (2011); Nelson (1972)) at each age (commonly known as the Cumulative hazard function) for a given stage since it clearly discerns the nature of hazard of a dataset.

The cumulative hazard rate in figure 1 indicates data to be of Increasing Failure Rate (IFR) with respect to age which implies with increasing age,the detection of prostate cancer increases or succinctly the chance of prostate cancer in a male increases. With this information, several IFR models namely Weibull, Gamma, Exponentiated exponential (Gupta and Kundu (2001)), Lindley (Ghitany et al. (2008)), Generalized exponential distribution (Gupta and Kundu (1999)), etc. were narrowed down from literature . All of the afore mentioned distributions exhibits IFR nature for shape parameter greater than one. However, in accordance with the phenomenon, the onset age of prostate cancer plays a vital role. Due to this age lag, a displaced model shall render a better explanation of the event (upta and Kundu (1999)) where the lowest age at diagnosis is the assumed displacement factor in each stage. Our objective now is to obtain the model of best fit among the class of all considered IFR distributions. Thus, our testing criterion leads to the following null and alternative hypotheses :

$\mathbf{H_0}$: The data is governed by the assumed distribution ;

H_1 : The data is governed by some other distribution.

Among the competing models mentioned above, there are copious techniques for discriminating the best suited model for the above discussed phenomenon. To mention a few, based on which this study has been organized are Akaike Information Criterion (AIC), Kolmogorov-Smirnov distance (KS dist.), Bayesian Information Criterion (BIC) and the $\chi 2$ statistic. AIC and BIC quantifies the information loss due to the use of the assumed model with only difference between these being the strictness of BIC which penalizes the likelihood strongly due to additional parameters as compared to AIC (Akaike (1974); Schwarz et al. (1978)). This Rigorous nature of BIC always leads to greater information loss as compared to AIC for a given model and its corresponding estimates and that can also be verified from our results in tables 1. The KS distance measures the maximum deviation between the observed and assumed distribution function (Gun et al. (2010); Clauset et al. (2009)). If an assumed statistical distribution function is able to cover the maximum information in the data, accordingly, the KS distance shall be small in magnitude and vice-versa. The $\chi 2$ statistic is a well-known technique used for verifying goodness of fit between the estimated and observed density functions. The two best competing models in light of the above discussion are 2P-Gamma and 3P-Weibull distributions with lowest AIC, BIC, $\chi 2$ and KS values as given in table 1. The mathematical forms of these models are explained below where μ denotes the minimum age at diagnosis or the location factor and p and θ are the unknown shape and scale parameters respectively.

Model 1 : Displaced 3P Weibull distribution

The PDF, theoretical mean, distribution function, likelihood function and normal equations are given in equation 1-6.

$$f(x;p,\mu,\theta)=\frac{p}{\theta}\left(\frac{x-\mu}{\theta}\right)^{p-1}e^{-\left(\frac{x-\mu}{\theta}\right)^{p}} \qquad x>\mu;\ p>0;\theta>0 \quad (1)$$

$$E(X)=\mu+\theta\Gamma\left(\frac{1}{p}+1\right) \quad (2)$$

$$F(x;p,\mu,\theta)=1-e^{-\left(\frac{x-\mu}{\theta}\right)} \qquad x>\mu \quad (3)$$

$$L(p,\mu,\theta|x)=e^{-\sum_{i=1}^{n}\left(\frac{xi-\mu}{\theta}\right)^{p}}\prod_{i=1}^{n}\left(\frac{p}{\theta}\right)\left(\frac{xi-\mu}{\theta}\right)^{p-1} \qquad x_i>\mu \quad (4)$$

$$\frac{\partial \log L}{\partial p}=\frac{n}{p}+\sum_{i=1}^{n}log\left(\frac{xi-\mu}{\theta}\right)-\sum_{i=1}^{n}\left(\frac{xi-\mu}{\theta}\right)^{p}log\left(\frac{xi-\mu}{\theta}\right) \quad (5)$$

$$\frac{\partial \log L}{\partial \theta}=\frac{n(p-2)}{\theta}+\frac{p}{\theta^{2}}\sum_{i=1}^{n}\left(\frac{xi-\mu}{\theta}\right)^{p-1} \quad (6)$$

Model 1 : Displaced 3P Gamma distribution

The PDF, theoretical mean, distribution function, likelihood function and normal equations are given in equation 7-12.

$$f(x;p,\mu,\theta)=\frac{1}{\theta\Gamma(p)}\left(\frac{x-\mu}{\theta}\right)^{p-1}e^{-\left(\frac{x-\mu}{\theta}\right)}\ x>\mu;p>0;\theta>0 \tag{7}$$

$$E(X)=\mu+\theta p \tag{8}$$

$$F(x;p,\mu,\theta)=\int_{\mu}^{x}\frac{1}{\theta\Gamma(p)}\left(\frac{x-\mu}{\theta}\right)^{p-1}e^{-\left(\frac{x-\mu}{\theta}\right)}dx\ \ x>\mu \tag{9}$$

$$L(p,\mu,\theta|x)=\frac{1}{\theta^{n}\Gamma(p)^{n}}e^{-\sum_{i=1}^{n}\left(\frac{xi-\mu}{\theta}\right)}\prod_{i=1}^{n}\left(\frac{xi-\mu}{\theta}\right)^{p-1}\ xi>\mu \tag{10}$$

$$\frac{\partial\, Log\, L}{\partial\theta}=\frac{1}{\theta^{2}}\sum_{i=1}^{n}\left(\frac{xi-\mu}{\theta}\right)-\frac{np}{\theta} \tag{11}$$

$$\frac{\partial\, Log\, L}{\partial p}=\sum_{i=1}^{n}log(xi-\mu)-n\,log\,\theta-n\frac{\partial}{\partial p}log\,L\,(p) \tag{12}$$

On substituting μ = 0 in the above equations, we obtain the standard 2 parametric counterparts of the same models. The normal equations in all the above cases do not admit explicit solution and solutions have been evaluated numerically to estimate the parameters (using R software). From table 2, it is apparent that the minimum AIC and BIC along with $\chi 2$ statistic and KS distance for stage I are observed for the 3P-Weibull distribution. Consequently, in the remaining stages the 3P-Weibull outperforms the others except in stage III where we obtained the lowest AIC and BIC value for 3P-Weibull but the $\chi 2$ and KS value is slightly higher in comparison to 2P-Gamma distribution.

Although, all the above criterion are not comparable in absolute sense but with modesty, we choose that distribution which provides minimum information loss and explains the maximum possible variation in the data thereby incorporating the basic demand of the situation, which in this study is the onset age of Prostate Cancer in an African American. The phenomenon eventually rules out the preference of 2P-Gamma model with respect to a shifted Weibull model.

Requirement of a displaced model to draw inferences about such population is justified by the quantile-quantile or the Q-Q plot in figure 2 where the observed data has been plotted after displacing each observation of a stage by an appropriate value from the origin for that particular stage.

A Q-Q plot is used to plot theoretical quantiles on one axis and observed data quantiles on the other. The hypothesis is plausible if the observed data quantiles and the theoretical quantiles are reasonably well approximated by a straight line, while marked deviations from linearity provide evidence against this hypothesis (Wilk and Gnanadesikan (1968); Gibbons and Chakraborti (2014)). In this study, we noted linear relation between the two with slope of the line being the estimated scale parameter θˆ and the plotted line met ordinate at a point away from the origin which was found to be equivalent to the location factor considered for a given stage.

Thus, evidently in all the stages, 3P-Weibull distribution may be considered as a suited model for further analysis. The maximum likelihood estimates for the shape and scale parameters with reference to eq. 3-5 is summarized in table 2 for each stage accompanied by the 95% asymptotic confidence interval which can be expressed as

$$\text{C. I.} = \widehat{\Theta} \pm 1.96\sqrt{Var(\widehat{\Theta})}\,; \qquad \widehat{\Theta} = (\hat{p}, \hat{\theta})$$

where $Var(\Theta) = I(\Theta)^{-1}$ and $I(\Theta) = -\mathrm{E}(\frac{\delta^2}{\delta\theta^2} LogL(p, \mu, \theta\,|x))$.

The estimated PDF and Survival function plots shown in figure 3 reveals a closeness between the estimated model with the empirical one, yet statistically it is not justified to claim that the assumed model is suitable for the phenomenon, since in all the cases, the $\chi 2$ statistic is rejected at 5% as well 1% level of significance. One of the reason of rejection on the basis of $\chi 2$ may be due to large data. Authors in (Kunte and Gore (1992)) have discussed the effect of large sample on $\chi 2$ statistic. Here it becomes a necessity to discuss certain fallacies of large sample in statistical theories. Firstly, defining a sample as "small " or "large " has not been rigorously penned in statistical literature. Referring to the suggestion by (Kunte and Gore (1992)) sample sizes of order higher than 102 can be regarded as a "large " sample. Owing to large sample size in each stage, it is not justified to reject a model based on $\chi 2$ statistic since with a large data set, the goodness-of-fit test becomes sensitive to even very small, inconsequential departures from a distribution.

This occurs due to uncontrolled probability of type-II error (β) which approaches zero with increasing sample size, for a fixed type-I error (α =0.5, say) which is in contrary to the criterion led down for a good statistical test. Sensitivity of $\chi 2$ test has been demonstrated by (Kunte and Gore (1992)) through an example of sex ratio study in samples of small and large sizes respectively. The inferential problem for large data also persists in p-value interpretation. Several authors have criticized the dependency of p-value on sample size and its inconsistency pertaining to a large sample (Lin et al. (2013); Berkson (1938)). Theoretically, when the null hypothesis is true, an increase in sample size should result in a p-value close to 0 but the magnitude of p-value in such cases should be very carefully interpreted as it fails to vehemently portray evidence in support of the alternative hypothesis and merely expresses the consistency of the data with the null hypothesis.

Hence, rejecting the null hypothesis solely relying on the p-value of $\chi 2$ statistic shall render false conclusions about the observed phenomenon.

Another reason of disagreement of the null hypothesis for the proposed model might be due to the effect of "registry" data (Psoter and Rosenfeld (2013)). The problem at hand is possibly a victim of the large sample size and heterogeneity among the people reporting a registry.

4. Discussion on sub-sample

To tackle the problem of large data, a number of techniques has been elicited by (Lin et al. (2013)). The simplest and general way is to draw sub-samples from the above population and check the goodness of fit on the basis of the sub-samples. A question that may arise at this stage would be "what should be the size of the sub sample?" There are a number of ways to choose the

sample size. The simplest way of deciding the minimum sub-sample size is based on specifying the margin of error and confidence level for the confidence interval. Here we have considered 5% margin of error and 95% confidence interval. The minimum sample size based on the above criterion was obtained as 367 in stage I, 344 in stage II, 338 in stage III and 352 in stage IV respectively. Hereafter the study was conducted seriatim for 100 samples of sizes 400, 500, 600, 700, 800, 1000 and 2000 each at every stages and it was noted that with increase in the sample size, rejection goodness of fit with reference to $\chi 2$ statistic escalated significantly. This corroborate the drawback of $\chi 2$ goodness-of-fit test as pointed out by (Kunte and Gore (1992)). The approximate rejection rate of the null hypothesis in terms of $\chi 2$ values for different sample sizes is summarized in table 3.

We observed that the rejection rate of sub-samples based on $\chi 2$ statistic remained fairly constant for 400-600 sized samples. Thus, our preferred sample size is 600, based on which we draw further inferences. It is worthwhile to mention here that graphical study also revealed that with increase in the sample size, the discrepancy between assumed model and observed values increases. Based on sub-samples, the average estimates of shape, scale, their corresponding confidence intervals and mean square errors (MSE) are summarized in table 4. The estimated probability density plot for all the stages are presented in figure 4. The MSE of the estimates in table 4 was computed by assuming the estimates obtained in table 2 as the true value of the parameters. It is quite clear that the estimates obtained from a sub-sample of 600 in table 4 is justly equivalent to the estimates of the initial sample in table 2 with a small MSE in all the cases. The above techniques were also implemented on single registries. In terms of $\chi 2$ goodness of fit test a similar trend was observed in rejection of the proposed model. The Connecticut registry reported 695 and Seattle (Puget Sound) reported 296 prostate cancer patients in stage I and the $\chi 2$ goodness of fit test failed to reject the null hypothesis of Weibull distribution governing the age distribution of prostate cancer detection among African American males both at 4% and 6% level of significance subsequently. In other registries of higher sample sizes Weibull distribution was found unsuitable for modelling the age distribution of prostate cancer detection with regards to $\chi 2$ test. This oddity in the behaviour of the $\chi 2$ statistic is evidently a result of large samples as well as increased heterogeneity in the "registry" due to large number of reporting. In either situations, the sample size is the key factor in deciding the rejection-non rejection of the null hypothesis under $\chi 2$ goodness of fit test. Thus, inferring on the basis of sub-sample for the whole African American residents of these above mentioned places shall not yield a larger type I error for the hypothesis under examination.

5. Result and Conclusion

In this study we proposed a suitable model governing the age at diagnosis of prostate cancer. 3P-Weibull distribution was zeroed onto to be the best model in each of the stages. The Maximum Likelihood Estimation technique was opted for the parametric estimation in which the location parameter was assumed to be the lowest age at diagnosis leaving the other two unknown parameters free for estimation. Even though graphically the Weibull distribution was good enough for the situation under study, yet the model lacked the statistical properties of a good

proposition in the sense that all the model fitting criterion were nullified by the Weibull density. Literature referred to such anomalies as the "Large sample paradox". Also, due to heterogeneity amongst and within registries the model failed to explain the phenomenon for larger sample sizes.

The estimation procedure was repeated for smaller samples and single registries with optimum sample size and the subsequent estimates obtained from the above re-sampling technique was close to the estimates obtained from complete sample and also complied with the $\chi 2$ test for goodness of fit. Thus, for further inferences, the average estimates and corresponding average asymptotic confidence interval based on the sub-sample has been considered which is presented in table 4. The mean age at cancer detection using eq.8 was estimated to be 67.6824, 65.5237, 64.4222, 69.1231 years and median age using eq. 9 being 68.0569, 65.5782, 64.2331, 71.2589 years respectively for stage I, stage II, stage III and stage IV. This study accomplishes one of the major benefits of statistical modelling of failure time data, that is, small samples being much effective in analysing and inferring about a much larger population only through such life time models. This model well depicts high risk ages along with the mean and median age at diagnosis and accordingly, the policies should be framed to aim surveillance and screening of males at these particular ages regularly. The suitability of the model rested solely on the phenomenon and we believe the model to perform best in situations considered in this study.

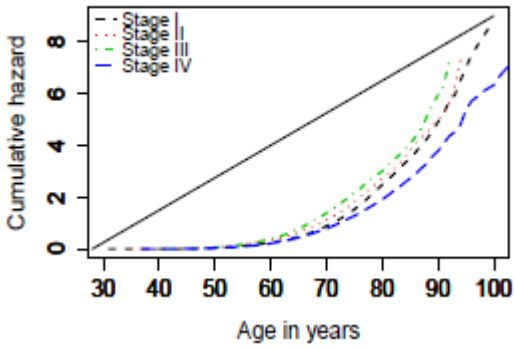

Figure 1: Cumulative Hazard for all stages

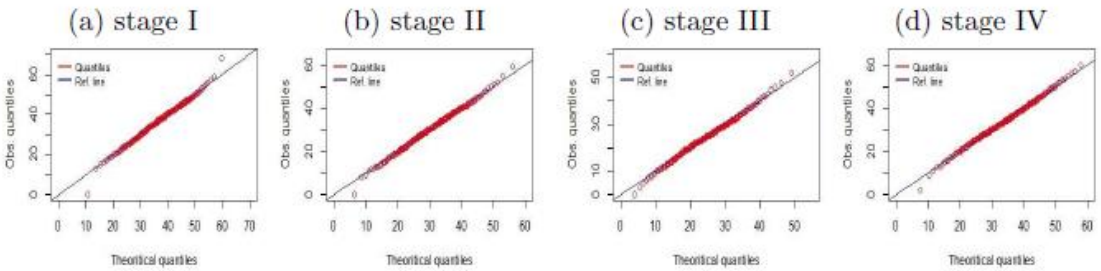

Figure 2: Q-Q plot of all the stages

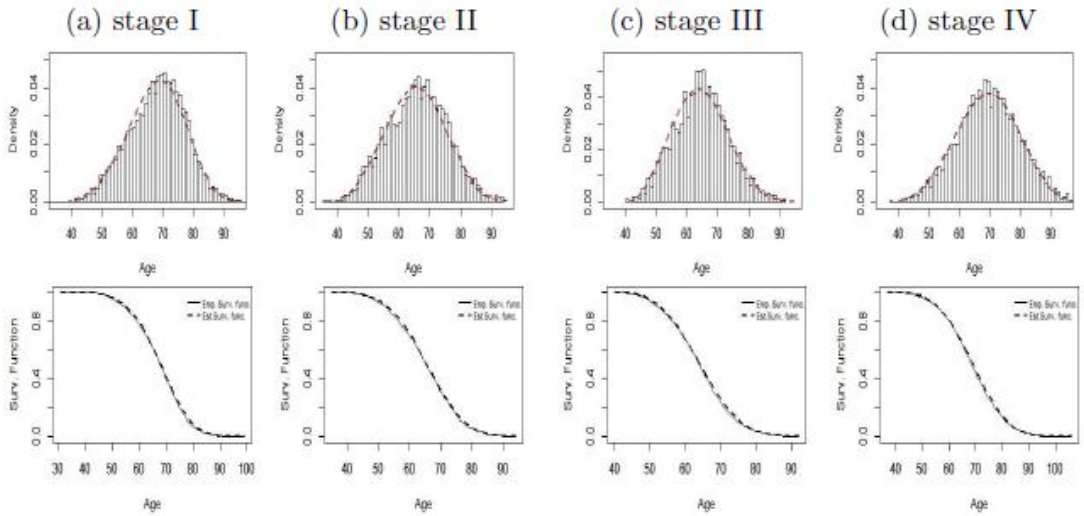

Figure 3: The estimated density function and survival function with reference to observed histogram and ECDF.

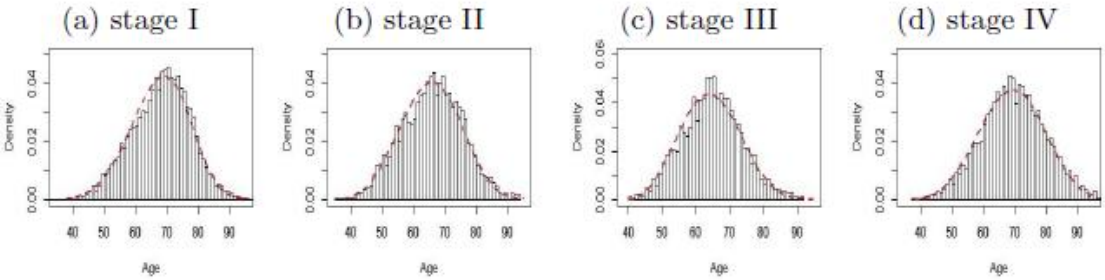

Figure 4: The estimated density function of sub-sample of size 600 with reference to observed histogram.

Table 1: AIC, BIC, $\chi 2$ and KS distance for all stages

Stage	Distribution	AIC	BIC	χ^2	KS Dist.
I	2P Weibull	56749.724	56763.640	132.811	0.0369
	3P Weibull	56530.927	56544.843	58.362	0.0241
	2P Gamma	56847.607	56861.523	320.754	0.0317
	3P Gamma	57249.301	57263.216	608.054	0.0437
II	2P Weibull	24238.603	24250.792	71.014	0.0445
	3P Weibull	24081.273	24093.462	18.218	0.0241
	2P Gamma	24169.491	24181.680	78.993	0.0274
	3P Gamma	24375.325	24387.514	217.868	0.0407
III	2P Weibull	20161.499	20173.358	197.879	0.0637
	3P Weibull	19905.995	19915.854	21.189	0.0279
	2P Gamma	19960.659	19972.518	20.948	0.0237
	3P Gamma	20147.153	20159.010	145.872	0.0432
IV	2P Weibull	31489.293	31501.975	128.796	0.0561
	3P Weibull	31265.402	31278.083	22.988	0.0272
	2P Gamma	31338.736	31351.417	46.083	0.0162
	3P Gamma	31598.863	31611.545	183.608	0.0282

Table 2 : Estimates and Confidence Intervals of 3P-Weibull

Stages	Parameters	Estimates	Confidence Interval/ length
I	$\hat{p}$	4.4985	(4.4207, 4.57628) 0.1555
	$\hat{\theta}$	40.2057	(39.9964, 40.4151) 0.4186
II	$\hat{p}$	3.5487	(3.4536, 3.6439) 0.1903
	$\hat{\theta}$	33.9096	(33.2208, 34.5948) 0.6888
III	$\hat{p}$	3.0348	(2.9468, 3.1228) 0.176
	$\hat{\theta}$	27.2727	(26.9209, 27.6244) 0.7035
IV	$\hat{p}$	3.5180	(3.4352, 3.6008) 0.1656
	$\hat{\theta}$	35.6997	(35.3762, 36.0232) 0.6470

Table 3: rejection rate (in%) of null hypothesis with respect to $\chi 2$ statistic

Sample size	600	800	1000	2000
rejection rate	5	7	9	52

Table 4: Average estimates and confidence interval of 3P-Weibull distribution based on sample size 600

Stages	Parameters	Estimates/ MSE	Confidence Interval/ length
I	$\hat{p}$	4.5231	(4.2415, 4.8047)
		(0.0203)	0.2816
	$\hat{\theta}$	40.1847	(39.4354, 40.9340)
		(0.1347)	0.7493
II	$\hat{p}$	3.5603	(3.3374, 3.7832)
		(0.0118)	0.2229
	$\hat{\theta}$	33.8938	(33.0912, 34.6964)
		(0.1251)	0.8026
III	$\hat{p}$	3.0478	(2.8579, 3.2377)
		(0.0090)	0.1899
	$\hat{\theta}$	27.3298	(26.5741, 28.0855)
		(0.0938)	0.7557
IV	$\hat{p}$	3.5120	(3.2931, 3.7309)
		(0.0097)	0.2189
	$\hat{\theta}$	35.6880	(34.8313,36.5446)
		(0.1818)	0.8567

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

[2] American Cancer Society (2014). Prostate Cancer Prevention and Early detection. http://www.cancer.org/cancer/prostatecancer/ moreinformation/prostatecancerearlydetection/ prostate-cancer-early-detection-acs-recommendations.

[3] Anderson, J. and Sternberg, C. N. (2008). Adapting treatment for prostate cancer according to risk of disease progression. *Critical Reviews in Oncology/Hematology*, 68(1, Supplement):S23 – S31.

[4] Aprikian, A. G., Zhang, Z.-F., and Fair, W. R. (1994). Prostate adenocarcinoma in men younger than 50 years: A retrospective review of 151 patients. *Cancer*, 74(6):1768–1777.

[5] Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33(203):526–536.

[6] Cancer research UK (2015). Cancer statistics. http://www. cancerresearchuk.org/health-professional/cancer-statistics/ statistics-by-cancer-type/prostate-cancer/incidence.

[7] Chan, Y. M. and Tsokos, C. P. (2013). Parametric survival analysis: A comparison of prostate cancer survivorship by race. *International Journal of Mathematical Sciences in Medicine*, pages 31–47.

[8] Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

[9] Ghitany, M., Atieh, B., and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and computers in simulation*, 78(4):493–506.

[10] Gibbons, J. and Chakraborti, S. (2014). Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded. Taylor & Francis.

[11] Grönberg, H. (2003). Prostate cancer epidemiology. *The Lancet*, 361(9360):859–864.

[12] Gun, A. M., Gupta, M. K., and Dasgupta, B. (2010). *An Outline of Statistical Theory*, volume 2. The World Press Pvt. ltd.

[13] Gupta, R. D. and Kundu, D. (1999). Theory and methods: Generalized exponential distributions. *Australian and New Zealand Journal of Statistics*, 41(2):173–188.

[14] Gupta, R. D. and Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and weibull distributions. *Biometrical journal*, 43(1):117–130.

[15] Howlader, N. et al. (2014). Seer cancer statistics review, 1975–2005. *Bethesda, MD: National Cancer Institute*, pages 1975–2011.

[16] Kapur, K. and Lamberson, L. (1977). *Reliability in engineering design*. Wiley.

[17] Kunte, S. and Gore, A. P. (1992). The paradox of large samples. *Current Science*, 62(5):393–395.

[18] Lawless, J. F. (2011). Statistical models and methods for lifetime data, volume 362. John Wiley & Sons.

[19] Li, J., Djenaba, J. A., et al. (2012). Recent trends in prostate cancer incidence by age, cancer stage, and grade, the united states, 2001–2007. *Prostate Cancer*, 2012.

[20] Lin, D. W., Porter, M., and Montgomery, B. (2009). Treatment and survival outcomes in young men diagnosed with prostate cancer. *Cancer*, 115(13):2863–2871.

[21] Lin, M., Lucas Jr, H. C., and Shmueli, G. (2013). Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.

[22] Moul, J., Anderson, J., et al. (2003). Early prostate cancer: prevention, treatment modalities, and quality of life issues. *European urology*, 44(3):283–293.

[23] National Cancer Institute (2015). PDQ prostate cancer treatment. http://www.cancer.gov/types/prostate/patient/prostate-treatment-pdq.

[24] National Cancer Institute at the National Institutes of Health (2014).

[25] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966.

[26] Osei, N., Bonsu, M., and Tsokos, C. P. (2013). Statistical evaluation of different prostate cancer treatments. *International Journal of Mathematical Sciences in Medicine*, pages 17–31.

[27] Psoter, K. J. and Rosenfeld, M. (2013). Opportunities and pitfalls of registry data for clinical research. *Paediatric Respiratory Reviews*, 14(3):141 – 145.

[28] Ries, L. A. et al. (2008). Seer cancer statistics review, 1975–2005. *Bethesda, MD: National Cancer Institute*, pages 1975–2005.

[29] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[30] Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

[31] Surveillance epidemiology and End Results program (SEER) (2014).

[32] Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, 55(1):pp. 1–17.

Received January 3, 2013; accepted July 21, 2016.

Co-author : Suparna Basu
Department of Statistics
Banaras Hindu University
Varanasi-221 005, Uttar Pradesh, India
Email: basusuparna91@gmail.com

Sanjay Kumar Singh
Department of Statistics
Banaras Hindu University
Varanasi-221005, Uttar Pradesh, India
Email: singhsk@bhu.ac.in

Umesh Singh
Department of Statistics and DST-CIMS
Banaras Hindu University
Varanasi-221 005, Uttar Pradesh, India
Email: umesh@bhu.ac.in