Research Journal of Mathematical and Statistical Sciences _____ISSN 2320–6047
Vol. **3(6),** 1-9, June **(2015)** *Res. J. Mathematical and Statistical Sci.*

Statistical Analysis of Breast Cancer Tumour Sizes

Mudunuru Venkateswara Rao and Chris P Tsokos
Department of Mathematics and Statistics, University of South Florida, Tampa, FL, USA

Available online at: www.isca.in, www.isca.me
Received 20th May 2015, revised 2nd June 2015, accepted 10th June 2015

Abstract

A tumour can be either malignant or benign. A benign tumour does not grow abnormally and is not harmful in the long run. But malignant tumours love to grow and conquer the surrounding area and therefore will require aggressive treatment methods. The object of the present study is to perform statistical analysis of malignant breast tumour with the tumour size being the response variable. We determined that the tumour sizes of White women, African American women and other race women are statistically different. The probability distribution that characterizes the behaviour of the response variable was obtained along with the confidence limits. The malignant tumour size is partitioned into age groups and we performed stage wise and race wise analysis of behaviour of breast tumours.

Keywords: Tumours, probability density function, Inverse Gaussian, Pair-wise t-tests.

Introduction

Any cancer that grows in our body is always dangerous. If it exists one must try to locate and get it out of our body immediately. Breast cancer is a signature disease of Western populations. Cancerous tumor in breast generally starts in its tissues. There are mainly two types of breast cancer, namely, Ductal carcinoma and Lobular carcinoma. The former starts in the ducts that move milk from the breast to the nipple and the latter starts is the lobules that produce milk[1-3]. It is very uncommon that the breast cancer can start in other areas of the breast. It is suggested by doctors in several occasions that by performing periodic breast self-examination, scheduled mammograms, annual clinical breast exams[4], a lady can keep a track of tumour growth in her breast.

Facts and Numbers

According to the recent cancer statistics by American Cancer Society (ACS), in 2015 there will be a total of 1.6 million new cancer cases expected to be diagnosed and close to six hundred thousand deaths due to cancer in America[5]. In the United States, approximately 13% of women will develop breast cancer in their lifetime. According to ACS, in 2015 there can be about 40,000 women who will die due to breast cancer. Breast and lung cancer death rates are higher in US women compared to other cancers. Next to breast cancer, skin cancer is the most commonly diagnosed cancer among U.S. women. Little more than 1 in 3 cancers in women (about 29%) are breast cancer. More than 231,000 women and 2,350 men will discover that they have breast cancer by the end of this year[6,7].

Questions of Interest: i. What is the probability distribution function (PDF) that best characterizes the behaviour of malignant tumours for Whites, African Americans and other races?. ii. Is there any statistical difference between mean tumour sizes between the three races (Whites, African Americans and Others) in the study?. iii. Is there any statistical difference between mean tumour sizes of any two races?. iv. If a lady feels a tumour while self-examining, what is the confidence interval estimation for the average tumour size based on her race?

Data Description: The data used in this research is obtained from the Surveillance, Epidemiology, and End Results (SEER)-Medicare database. SEER is a National Cancer Institute (NCI) funded program. In 1973, SEER started collecting the data from 9 states and metropolitan registries and by the end of 2001 it is expanded to 17 cancer registries covering 26% of the US population. SEER links data from NCI's cancer registry program, the federally funded insurance program for the US elderly. These data are made available to all researchers after they submit a signed request copy to ACS. Many valuable research articles and monographs on various cancers are published using this data[8-14].

Under SEER program, during 1973-2007 a total of 146 million person-years are covered with 3.1 million incident cancers on the basis of a positive and negative test. SEER database is a unique and reliable data source for researchers investigating the different aspects of cancer[15]. In this present study we have obtained breast cancer incidence data from the SEER database. We have used patient and population data from the SEER 9 and SEER 13 database registries[15,16]. Our breast cancer data obtained from SEER has information of breast cancer with the details including cancer site, tumour pathology, stage of cancer, and cause of death among many others[3, 4].

In this work, we pre-processed the SEER data for breast cancer to eliminate repetitions and missing data. The resulting data set had 47,167 malignant tumour records, which then pre-classified

into three groups of races. "Whites" (37,341; 79.15%), "African American or AA" (4,234; 9%) and "Others" (5,592; 11.85%) are given in table-1.

Table-1
Race and age details

Race	N	Percent	Minimum age	Median age	Maximum age
1	37341	79.17	21	62	102
2	4234	8.98	22	57	102
3	5592	11.86	21	53	99

In this work, demographic information included age, race, and marital status. Tumour characteristics like tumour size (1mm to 998mm), stage of cancer (I, II, III, IV), tumour grade (1, 2, 3, 4, or unknown), and tumour treatment (1, 2, 3, 4) are also included.

From table-1, median age at diagnosis in the White women is 62 years (range 21 to 102 years) compared with a median age of 57 years in the African American women (range 22 to 102 years) and a median age of 53 years in the Other races women (range 21 to 99 years). There is 62.15% survival and 37.35% of not survived patients in our data (table-2). More information about survival function can be found elsewhere[17]. From Table 3, majority of patients (about 92%) are diagnosed when they are in stages 1 and 2 and very few (about 8%) of them are diagnosed in advanced stage of breast cancer.

Table-2
Survival status details

Status	Frequency	Percent	Cumulative percent
Dead (0)	17853	37.85	37.85
Survived (1)	29314	62.15	100.00

Table-3
Breast cancer stage wise details

Stage	Frequency	Percent	Cumulative percent
1	23345	49.49	49.49
2	20017	42.44	91.93
3	2600	5.51	97.45
4	1205	2.55	100.00

Parametric Analysis of breast cancer tumour sizes

Most clinical research is driven by gathering of a good data. Many complex research problems involve designing of models using mathematical and statistical techniques. There are several methods of modelling a data. In order to analyze and model such a data set, one must initially need to make explicit remark on the underlying assumptions and the distribution of the observations[18]. The underlying assumptions include that the data gathered is random, the population of such data satisfies normality and there is homogeneity among variances. This gathered data further allows us to draw some conclusions about the specific characteristics of a larger population data based on this sample[19-23]. Parametric tests are more powerful compared to nonparametric tests as they are directly related to the parameters of the data. However, parametric tests are not so easier to conduct and hence researchers tend to nonparametric tests, which are comparatively weaker.

In our work we performed parametric analysis to determine the best fitted distribution that characterizes the behaviour of tumour size for each race by setting the hypothesis as follows:

H_0: The tumour size data followed a specific parametric model
H_1: The tumour size data did not follow a specific parametric model

After performing many trials, from the class of many parametric distributions, based on the results of minimum Anderson-Darling value, we identified that Inverse Gaussian distribution as the best probabilistic distribution function that characterizes the behaviour of the malignant tumours for all the three races considered in this study.

Inverse Gaussian distribution

Over a century, family of Inverse Gaussian distributions had attracted the attention of many researchers in many fields[24,25]. The Hazard rate function of Inverse Gaussian distribution is uni-modal which increases from zero to its maximum value and decreases asymptotically to a constant. The most differentiating fact is extreme values of outcomes can occur with almost all outcomes being small. It is a right-skewed distribution with long tail. For these reasons Inverse Gaussian distribution is often used in reliability and survival analysis. Various insurance problems and stock markets follow this distribution[25,26].

The distribution is described by two parameters. Mean or location ($\mu > 0$) and precision or shape ($\lambda > 0$). Let us suppose $x_1, x_2, ...x_n$ be *n* independent and random variables. If x_i follows the inverse Gaussian distribution, then probability density function of $x_i \sim IG(\mu, \lambda)$ is

$$f(x,\theta) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, x \geq 0; \theta = (\mu, \lambda)^T$$

The expected value is given by mean μ and variance is equal to $\frac{\mu^3}{\lambda}$. The cumulative distribution function is given by

$$F(y) = \phi(y) + \exp\left(\frac{2\lambda}{\mu}\right)\left(-\sqrt{\frac{4\lambda}{\mu} + y^2}\right); -\infty < y < \infty$$

Where ϕ is the standard normal distribution function

Fitted Probability Distribution Functions

Figure 1 is the fitted Inverse Gaussian PDF with estimated shape and location parameters as 43.933 and 32.756 respectively. From figure-2 the CDF graph explains how well the distribution fit to data and the PP plot in figure-3 is approximately linear and confirms about the fitted distribution.

Figure 4 is the fitted Inverse Gaussian PDF for AA women with estimated shape and location parameters as 66.614 and 39.611 respectively. From figure-5 the CDF graph explains how well the distribution fit to data and the PP plot in figure-6 is approximately linear and confirms about the fitted distribution.

Figure-7 below is the fitted Inverse Gaussian PDF for other race women with estimated shape and location parameters as 55.703 and 36.846 respectively. From figure-8 the CDF graph explains how well the distribution fit to data.

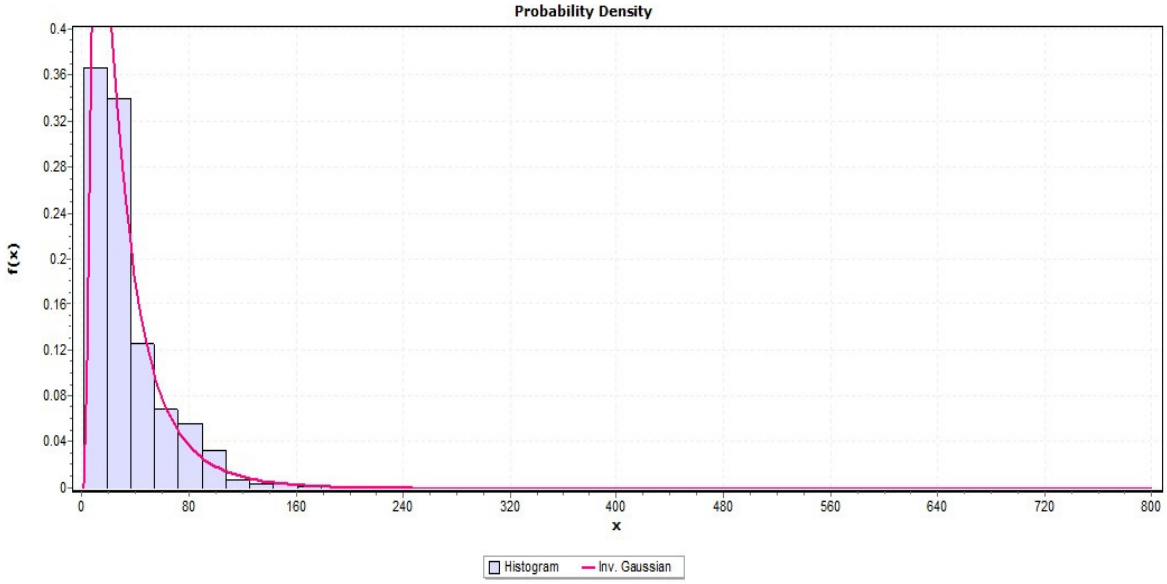

Figure-1
PDF for white women: Inverse Gaussian Distribution

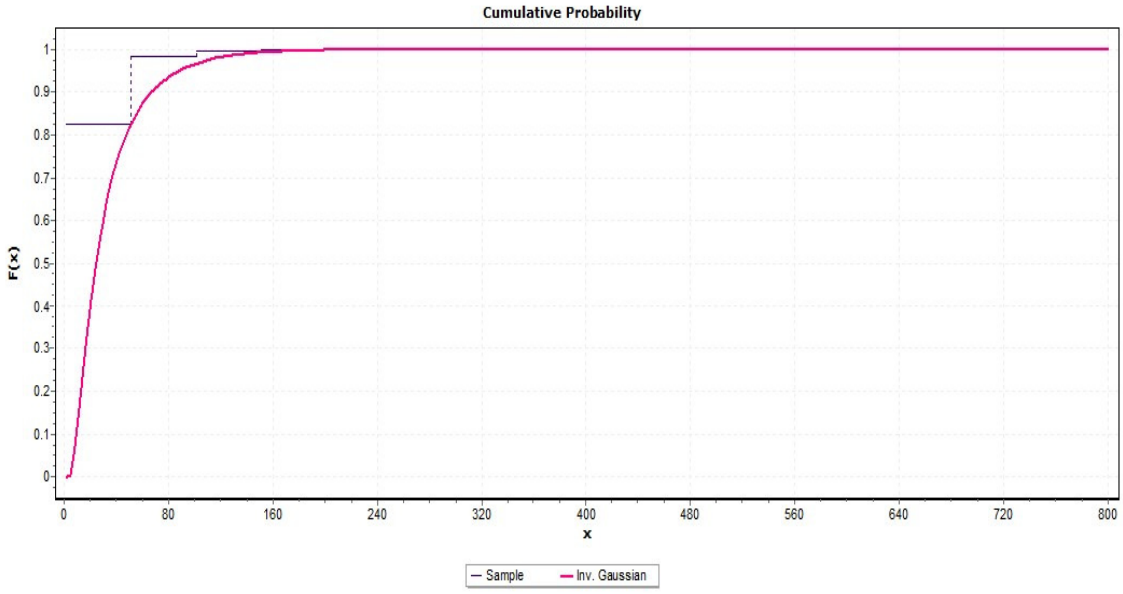

Figure-2
Inverse Gaussian CDF for White Women

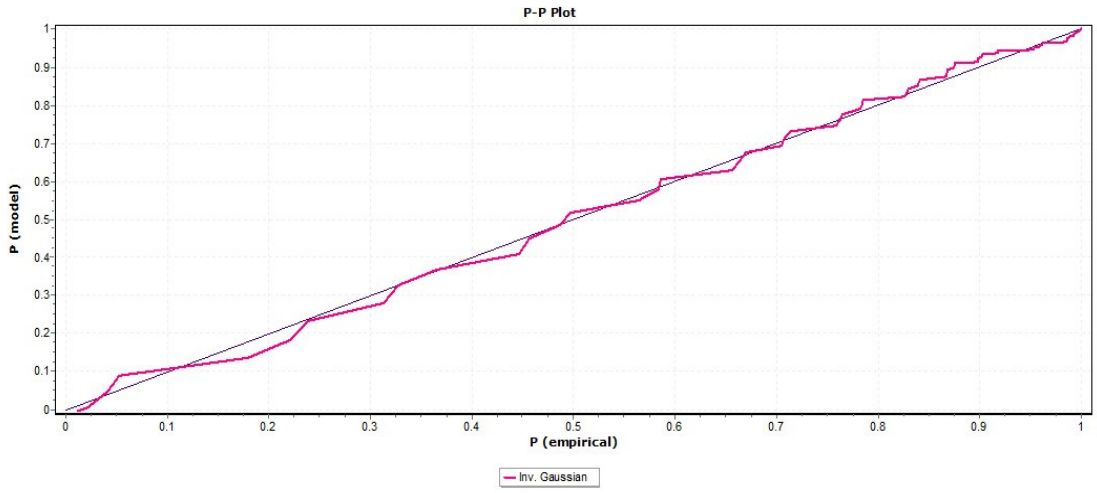

Figure-3
Inverse Gaussian PP Plot for White Women

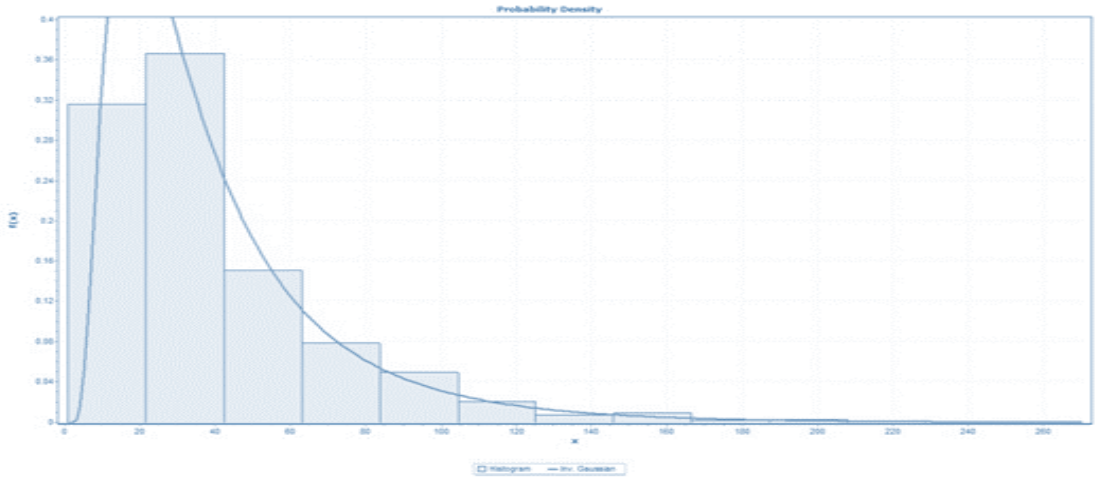

Figure-4
PDF for African American Women: Inverse Gaussian distribution

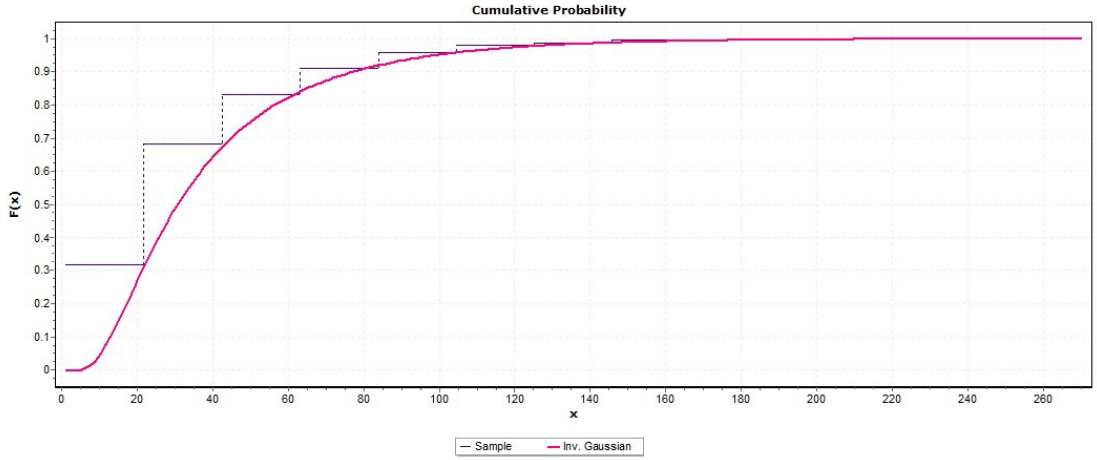

Figure-5
Inverse Gaussian CDF for AA Women

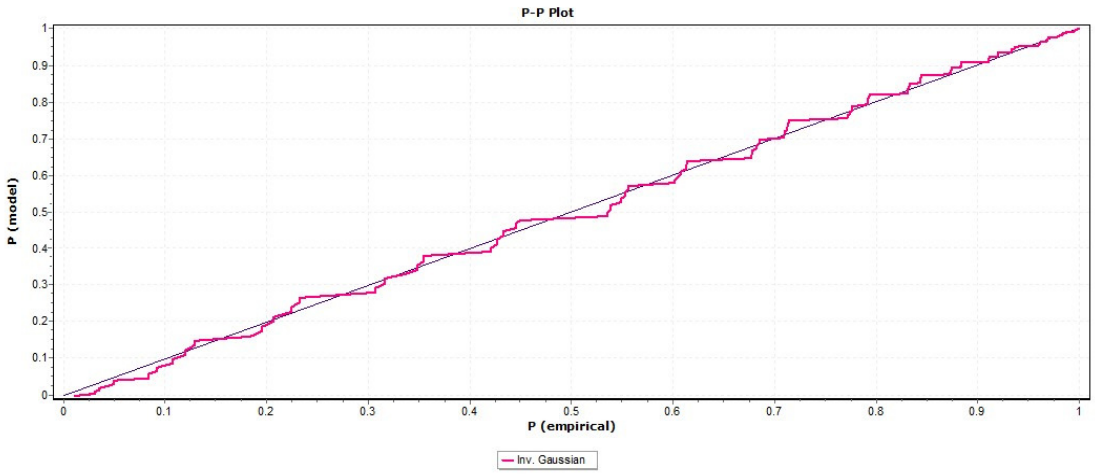

Figure-6
Inverse Gaussian PP Plot for AA Women

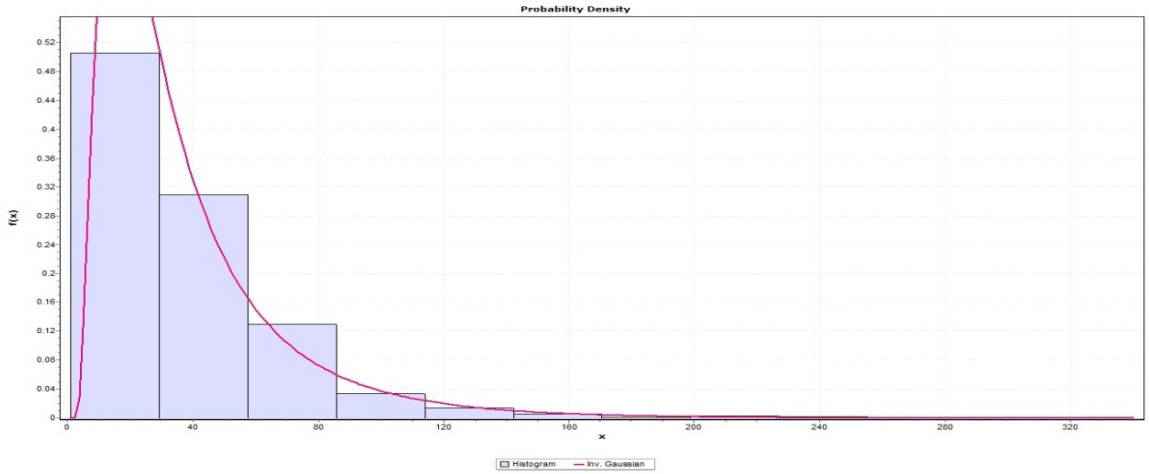

Figure-7
PDF for Other races: Inverse Gaussian Distribution

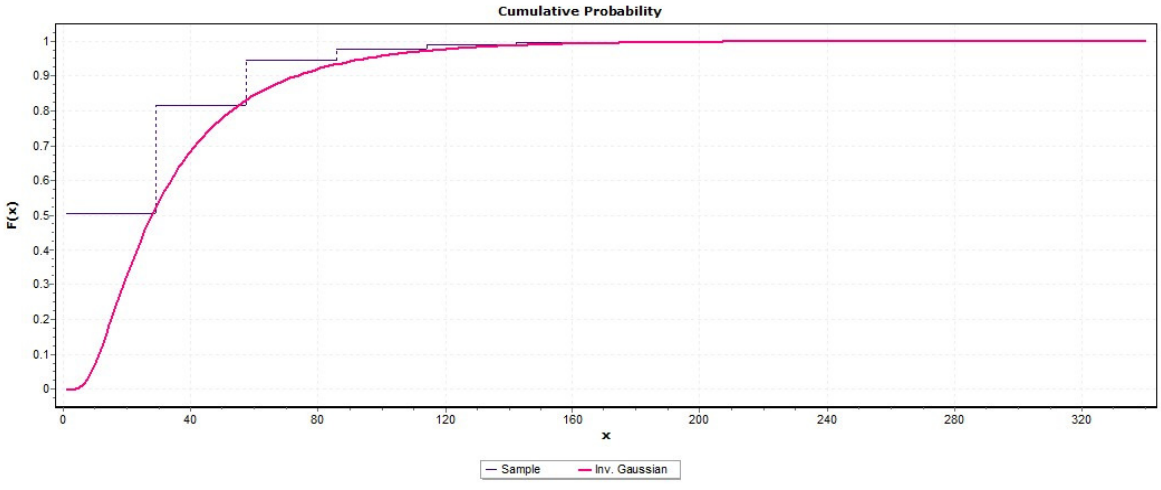

Figure-8
Inverse Gaussian CDF for Other race women

Table-4
PDF summary for three races

RACE	$\hat{\lambda}$	$\hat{\mu}$
White	43.933	32.756
African American	66.614	39.611
Others	55.703	36.846

Table 5 has the race wise details of 95% confidence interval estimation of true mean tumour size based on Inverse Gaussian distribution. After identifying the distribution functions that best characterizes the probability distribution of malignant tumours for the three races, we proceed to compare the differences of mean tumour sizes for the three races.

Table-5
Estimated mean tumour size and confidence intervals for all the three race women

Race	$\hat{\mu}$	$\hat{\lambda}$	**SD**	**95% CI for** μ
1	32.756	43.933	28.284	(32.47, 33.04)
2	39.611	66.614	30.545	(38.69, 40.53)
3	36.846	55.703	29.967	(36.06, 37.63)

Comparison of mean tumour sizes: Let $\mu_W, \mu_{aa}, and\ \mu_{oth}$ represent mean tumour sizes of whites, African Americans and other races respectively. Our interest is to test the hypothesis whether all the three races have same mean tumour size or otherwise.

$H_0 : \mu_W = \mu_{aa} = \mu_{oth}$ vs. H_1: At least one of them is not equal.

By performing a one way ANOVA at 5% level of significance, we obtained the p-value which is very low ($p < 0.0001$); leading us to the conclusion that there is significant difference between the average mean tumour sizes of the three races. So, we now proceed in pair wise testing of mean tumour sizes for all three races. The table 6 below has the details of the results after performing t-test for pair wise testing. Clearly, we conclude that the average tumour size is significantly different for all the three races in this study.

Table-6
Pair wise comparison of mean tumour sizes

H_{NULL}	$H_{ALTERNATE}$	**P-value**
$\mu_W = \mu_{aa}$	$\mu_W \neq \mu_{aa}$	0.001
$\mu_{aa} = \mu_{oth}$	$\mu_{aa} \neq \mu_{oth}$	0.0001
$\mu_W = \mu_{oth}$	$\mu_W \neq \mu_{oth}$	0.0002

Previous studies[27-29] have shown that breast cancer in these younger women is more aggressive, with higher rate of occurrence and recurrence rates compared with older women. In our study we have the median age of women for all the three races more than 50 years. In Table 8, we classified the tumour stage taking age group into consideration. The majority of women are in the ages from 45 to 79. From table 7 and Figure 10, African American women are the majority of population in all the age groups who are diagnosed with breast cancer. Table 7 gives the age group wise confidence interval for mean tumour size for all the three races. Very interestingly, from figure 9 majority of women in younger ages (20–44 years) are identified with stage-2 breast cancer.

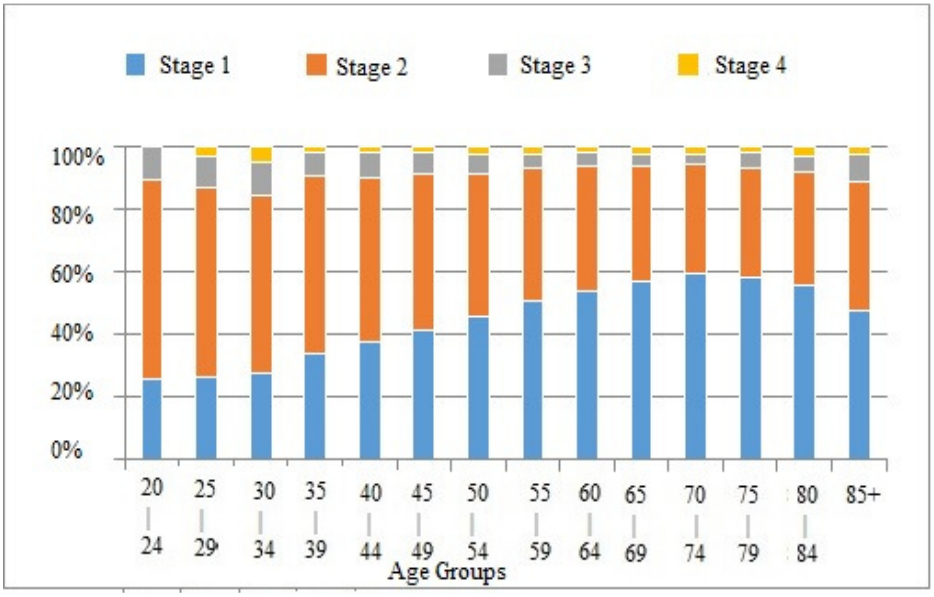

Figure-9
Stages vs. age group

Conclusion

The PDF for all the three races is identified as Inverse Gaussian and the details about mean tumour sizes along with 95% confidence intervals for mean tumour sizes for all the three races were tabulated in table-5. One way ANOVA was performed for comparing mean tumour sizes of three races and at 5% level of significance, we conclude that the average tumour size for all the three races is statistically not the same.

Later, we performed pair-wise testing between the races and the results are tabulated in table-6. From these results we conclude that the average tumour sizes are significantly different for all the three races. Also compared with Whites and other race women, African American women have comparatively a greater mean tumour sizes and Whites have the least. This is also supported by the results published in table-7. Finally from Table 8 grouping ages into groups of 5, we also stratified the number of women diagnosed with breast cancer in different stages which gives an insight for future studies.

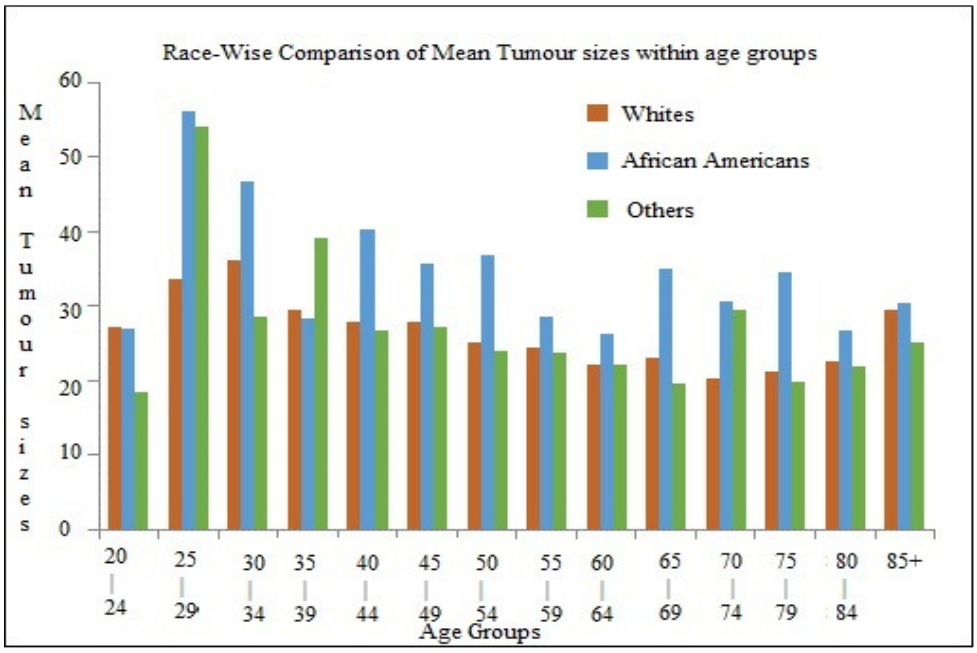

Figure-10
Race wise comparison of mean tumour sizes

Table-7
Age group based race wise confidence interval of tumour sizes

Age Group	Race 1				Race 2				Race 3			
	Mean	S.D	C.I (95%)		Mean	S.D	C.I (95%)		Mean	S.D	C.I (95%)	
			L.C.I	U.C.I			L.C.I	U.C.I			L.C.I	U.C.I
20-24	27.12	15.75	19.40	34.84	26.86	13.67	16.73	36.99	18.5	3.32	15.3	21.75
25-29	33.53	74.51	22.70	44.36	56.2	162.8	3.02	109.38	54.2	167.6	-2.14	110.54
30-34	36.24	82.99	29.51	42.97	46.8	124.2	24.94	68.66	28.55	22.25	25.00	32.11
35-39	29.44	60.16	26.36	32.52	28.46	22	25.57	31.35	39.19	114.9	27.44	51.01
40-44	27.79	59.2	25.58	30.00	40.23	110.4	29.35	51.11	26.84	55.2	22.71	30.94
45-49	27.81	70.22	25.57	30.05	35.8	89.67	27.88	43.72	27.11	67.33	22.7	31.51
50-54	25.06	62.78	23.15	26.96	36.92	95.54	28.89	44.95	24.01	38.44	21.4	26.58
55-59	24.44	68.85	22.31	26.57	28.56	64.53	23.03	34.09	23.81	55.75	19.6	28.00
60-64	22.20	59.49	20.36	24.04	26.27	51.07	21.63	30.91	22.21	44.1	18.6	25.84
65-69	23.03	67.36	20.99	25.07	35.11	104.3	25.51	44.71	19.52	16.65	18.1	20.98
70-74	20.41	53.55	18.81	22.00	30.67	78.75	22.47	38.87	29.45	101.5	19.2	39.72
75-79	21.31	48.48	19.75	22.86	34.58	98.68	23.52	45.64	19.84	14.66	18.1	21.58
80-84	22.65	51.02	20.62	24.68	26.81	21.21	23.79	29.83	21.89	25.49	17.5	26.29
85+	29.43	73.19	25.91	32.95	30.42	23.38	26.20	34.64	25.12	16.65	21.1	29.14

Table-8
Age group based stage wise confidence interval of tumour size

AGE	Stage 1		Stage 2		Stage 3		Stage 4		All
	Count	Row%	Count	Row%	Count	Row%	Count	Row%	Total
20-24	7	25.93	17	62.96	3	11.11	0	0.00	27
25-29	67	26.59	152	60.32	25	9.92	8	3.17	252
30-34	239	27.86	481	56.06	94	10.96	44	5.13	858
35-39	700	34.08	1157	56.33	152	7.40	45	2.19	2054
40-44	1456	37.89	1992	51.83	305	7.94	90	2.34	3843
45-49	2153	41.54	2560	49.39	348	6.71	122	2.35	5183
50-54	2561	45.78	2546	45.51	342	6.11	145	2.59	5594
55-59	2634	50.44	2209	42.30	250	4.79	129	2.47	5222
60-64	2723	53.92	1999	39.58	209	4.14	119	2.36	5050
65-69	2909	56.58	1892	36.80	199	3.87	141	2.74	5141
70-74	2997	59.35	1755	34.75	169	3.35	129	2.55	5050
75-79	2496	58.03	1509	35.08	199	4.63	97	2.26	4301
80-84	1526	55.49	993	36.11	143	5.20	88	3.20	2750
85+	877	47.61	755	40.99	162	8.79	48	2.61	1842
All	23345	49.49	20017	42.44	2600	5.51	1205	2.55	47167

References

1. Perou C.M., Sorlie T., Eisen M.B., van de Rijn M., Jeffrey S.S., Rees C.A. and Botstein D., Molecular portraits of human breast tumours, *Nature*, **406(6797)**, 747-752 **(2000)**

2. Garcia-Closas M., Hall P., Nevanlinna H., Pooley K., Morrison J., Richesson D.A. and Kropp S., Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics, *PLoS genetics*, **4(4), (2008)**

3. Colditz G.A., Rosner B.A., Chen W.Y., Holmes M.D. and Hankinson S.E., Risk factors for breast cancer according to estrogen and progesterone receptor status. *Journal of the National Cancer Institute*, **96(3)**, 218-228 **(2004)**

4. Berg W.A., Zhang Z., Lehrer D., Jong R.A., Pisano E.D., Barr R.G. and ACRIN 6666 Investigators. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk, *JAMA*, **307(13)**, 1394-1404 **(2012)**

5. American Cancer Society. *Cancer Facts and Figures 2013*. Atlanta: American Cancer Society, **(2013)**

6. American Cancer Society. *Cancer Facts and Figures 2015*. Atlanta: American Cancer Society, **(2015)**

7. Siegel, Rebecca, et al., Cancer statistics, 2014, *CA: a cancer journal for clinicians,* 64.1 **(2014)**

8. Anderson L.A., Pfeiffer R., Warren J.L., Landgren O., Gadalla S., Berndt S.I. and Engels E.A., Hematopoietic malignancies associated with viral and alcoholic hepatitis. *Cancer Epidemiology Biomarkers and Prevention*, **17(11)**, 3069-3075, **(2008)**

9. Anderson L.A., Landgren O. and Engels E.A., Common community acquired infections and subsequent risk of chronic lymphocytic leukemia, *British journal of haematology*, **147(4)**, 444-449. **(2009)**

10. Anderson L.A., Gadalla S., Morton L.M., Landgren O., Pfeiffer R., Warren J.L. and Engels E.A., Population-based study of autoimmune conditions and the risk of specific lymphoid malignancies, *International Journal of Cancer*, **125(2)**, 398-405. **(2009)**

11. Anderson L.A., Pfeiffer R.M., Landgren O., Gadalla S., Berndt S.I. and Engels E.A., Risks of myeloid malignancies in patients with autoimmune conditions, *British journal of cancer*, **100(5)**, 822-828 **(2009)**

12. Quinlan S.C., Morton L.M., Pfeiffer R.M., Anderson L.A., Landgren O., Warren J.L. and Engels E.A., Increased risk for lymphoid and myeloid neoplasms in elderly solid-organ transplant recipients, *Cancer Epidemiology Biomarkers and Prevention*, **19(5)**, 1229-1237, **(2010)**

13. Chang, C.M., Quinlan S.C., Warren J.L. and Engels E.A., Blood transfusions and the subsequent risk of hematologic malignancies, *Transfusion*, **50(10)**, 2249-2257, **(2010)**

14. Lanoy E. and Engels E.A., Skin cancers associated with autoimmune conditions among elderly adults, *British journal of cancer*, **103(1)**, 112-114 **(2010)**

15. SEER-13. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence-SEER 13 Regs Research Data, Nov 2010 Sub (1992–2008) —Linked To County Attributes—

Total U.S., 1969–2009 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April **2011**

16. SEER-9. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence-SEER 9 Regs Research Data, Nov 2009 Sub (1973–2008) Katrinia/Rita Population Adjustment> —Linked To County Attributes—Total U.S., 1969–2007 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2011, based on November 2010 submission. (**2011**). http://seer.cancer.gov/

17. Singh Kusum Lata and R.S. Srivastava, Inverse Maxwell Distribution as a Survival Model, Genesis and Parameter Estimation, *Research Journal of Mathematical and Statistical Sciences*, **2(7)**, 23-28 **(2014)**

18. Masuku, Ajay S. Singh and Micha B., Applications of Modeling and Statistical Regression Techniques in Research, *Research Journal of Mathematical and Statistical Sciences,* **1(6)**, 14-20 **(2013)**

19. Siegel S. and Castellan N.J., Nonparametric statistics for the behavioural sciences (McGraw-Hill, New York), **(1988)**

20. Portney L.G. and Watkins M.P. Foundations of clinical research: applications to practice, *Appleton and Lange, East Norwalk*, 148 **(1993)**

21. Mattson D.E., Statistics: Difficult Concepts, *Understandable Explanations (CV Mosby, St. Louis, 1981,* **(1981)**

22. Sukla M.K., A.K. Mangaraj and L.N. Sahoo, An Investigation on the Stochastic Modeling of Daily Rainfall Amount in the Mahanadi Delta Region, India, *Research Journal of Mathematical and Statistical Sciences,* **2(9)**, 1-8 **(2014)**

23. Colton T., Statistics in medicine. *Little, Brown, Boston*, 164, **(1974)**

24. Chhikara R., *The Inverse Gaussian distribution: Theory: Methodology, and Applications* (Vol. 95), CRC Press, **(1988)**

25. Gjerde T., Eidsvik J., Nyrnes E. and Bruun B.T., Normal Inverse Gaussian Error Distributions Applied for the Positioning of Petroleum Wells

26. Folks J.L. and Chhikara R.S., The inverse Gaussian distribution and its statistical application-a review, *Journal of the Royal Statistical Society, Series B (Methodological)*, 263-289 **(1978)**

27. El Saghir N.S., Seoud M., Khalil M.K., Charafeddine M., Salem Z.K., Geara F.B. and Shamseddine A.I., Effects of young age at presentation on survival in breast cancer, *BMC cancer*, **6(1)**, 194 **(2006)**

28. Shannon C. and Smith I.E., Breast cancer in adolescents and young women, *European Journal of cancer*, **39(18)**, 2632-2642, **(2003)**

29. Anders C.K., Hsu D.S., Broadwater G., Acharya C.R., Foekens J.A., Zhang Y. and Blackwell K.L., Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression, *Journal of Clinical Oncology*, **26(20)**, 3324-3330. **(2008)**

30. http://www.seer.cancer.gov