



An Analytical Study of The Cancer Patients of Jorhat Medical College Hospital

Pranjeet Borah^{1*}, Raktim Ranjan Borah², Nityaraj Chetia³

Abstract

Cancer is a very serious disease which may cause of death of the patients. Cancer is abnormal growth of cell in human body. It may occur almost any part of human body. Cancer is considered as second leading cause of death disease in the world. Though cancer is a very severe disease but it can be diagnosed. In this paper an attempt has been made to study the cancer patients in Jorhat Medical college and hospital by using various statistical test viz., test of proportion, Mann-Whitney U-test, run test, test of goodness of fit etc. by statistical software like SPSS, R etc.

Key words: Cancer, Run test, Mann-Whitney U-test, test of proportion, SPSS

DOI Number: 10.48047/NQ.2022.20.17.NQ880156

Neuroquantology 2022; 20(17):1217-1226

1. Introduction :

Cancer may be regarded as a group of diseases characterized by abnormal growth of cells, ability of invade adjacent tissues and even distant organs. Cancer can occur at any site or issue of the body and may involve any type of cells. It thrust a very heavy burden in public health sector and has become one of the leading disease worldwide nowadays. Cancers are caused by mutations that may be inherited, induced by environmental factors, or result from DNA replication errors (Tomasetti C et al.). Cancer, as a fatal disease, has a high level of associated misbeliefs and fear, and the patients have multiple cognitive representations of their illness (Lykins et al., 2008). Cancer is ranked as the first or second leading cause of death in 91 of 172 countries and is third or fourth in an additional 22 countries (Ferlay J et al.). The major categories of cancer are: **Carcinomas**, which arise from epithelial cells lining the internal surface of various organs (eg., mouth, oesophagus, intestine, uterus) and from the skin epithelium; **Sarcomas**, which arise from mesoderm cells constituting the various connective tissues (eg., fibrous tissues, fat and bone); **Lymphomas**, myeloma and leukemia's

arising from the cells of bone marrow and immune system. Total 13% of the annual deaths worldwide are from cancer or cancer-related and 70% of these deaths are in from low- and middle-income countries. In India cancer mortality has double from the year 1990 to 2016. India's cancer incidence is estimated at 1.15 million new patients in 2018 and is predicted to almost double as a result of demographic changes alone by 2040⁴. According to WHO, India has a cancer mortality rate of 79 per 100,000 deaths and accounts for over 6 percent of total deaths⁵. Further, the cancer mortality in India is projected to increase to over 900,000 deaths by the end of this decade⁶. Also, with higher burden of breast and uterine cancer, the cancer incidence in India is also identified with a significant gender dimension⁷⁻⁹. Most importantly, in India, and as elsewhere, the term cancer resonates shock and fear because of two concurrent reasons; first, very high treatment costs and second, poor chances of survival¹⁰.

In this paper an attempt has been made to test various objectives (which are mentioned below)

***Corresponding Author:** Pranjeet Borah

Address:^{1,2,3}Department of Statistics, Dibrugarh University, Dibrugarh-786004, Assam, India

¹Email: pranjeetborah123@gmail.com

²Email: raktimranjan2805@gmail.com

³Email: nityarajchetia454@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



on the cancer data which are collected from The Jorhat Medical College Hospital (JMCH). Jorhat Medical College & Hospital (JMCH) is a medical college cum hospital and medical research public university based in Jorhat, Assam, India. This is the fourth medical college of the state and it provides the healthcare needs of more than 1.2 million population of the entire Jorhat district, the neighboring districts of Golaghat Sivasagar and Majuli as well as the patients of neighboring states of Nagaland and Arunachal Pradesh. The college operates under the State Ministry of Health and Family Welfare, Assam. The foundation stone of the fourth Medical college of Assam was laid by then prime minister Dr Manmohan Singh on 25 August 2008. On 12 October 2009, the then Honourable Chief Minister of Assam, Tarun Gogoi inaugurated the Hospital wing of the Jorhat Medical College. It is said to be the best in Assam in terms of infrastructure and patient load as it is having a tremendous output following its inauguration in 2009.

2. Objectives and Methods :

The main objective of the present study are :

- To test the association between sex and various categories of cancer
- To test if the cancer patients are uniformly distributed over ages
- To test the significance of difference proportion of male and female cancer patients
- To test for association between Sex and Age of cancer patients
- To test for difference in median age of male and female cancer patients

2.1. Test of randomness :

There are many tests of randomness. We shall use the Wald Wolfowitz Run Test. A run is a sequence of identical letters preceded and followed by a different letter or no letter.

The hypothesis of interest here are:

H_0 : The sequence is random.

H_1 : The sequence is not random.

Let, $U = \text{No. of runs} = r_1 + r_2$

Then using critical values of U (or p -value) we can take decision accordingly. When sample size n is large ($n \geq 25$), we can use approximate test for U . In Wald Wolfowitz Run Test,

$$E(U) = \frac{2n_1 + 2n_2}{n_1 + n_2}$$

$$V(U) = \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}$$

$$Z = \frac{U - E(U)}{\sqrt{V(U)}} \dots (1.4.1)$$

Since it is a two tailed test, we have to reject H_0 with α , if $|Z| > Z_{\alpha/2}$ or p -value $< \frac{\alpha}{2}$

2.2. Approximate tests of normality:

In many situations, especially when sample size is small or finite, normality test becomes very essential. For application of t , F or chi-square test also the normality assumptions need to be verified. There is an important theorem in statistics is known as the Central Limit Theorem, which states that

“if X_i ($i = 1, 2, \dots, n$) be independent random variables such that $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, then under certain general conditions, the random variable $S_n = X_1 + X_2 + \dots + X_n$ is asymptotically normally distributed with mean μ and variance σ^2 , where $\mu = \sum_{i=1}^n \mu_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ ”

There are many methods of testing normality. One of the methods is using the probits. Here, first the value of the variable X of interest are divided into suitable intervals and corresponding frequencies are calculated. Then from the cumulative frequencies F , percentage cumulative frequencies $100F/N$ ($N =$ total frequency) are computed.

If $A = 100F/N$ is the percentage area upto the ordinate ‘ x ’ then

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{t^2}{2}\right) dt$$

5 is added to avoid negative values which may occur in practice.

$A = 100 \phi(x)$ and the probit p corresponding to A to $5+x$.

2.3. Test of single proportion :

Suppose, p_0 is the assumed proportion.

To test, $H_0: p = p_0$

Against $H_1: p \neq p_0$

The test statistic is $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1) \dots (1)$



Where, \hat{p} =estimated proportion
 p_0 = assumed proportion
 Reject H_0 if $|Z| > Z_{\alpha/2}$.

$$F(x,p) = p^x(1-p)^{1-x}; x=0,1; 0 \leq p \leq 1$$

MLE of p is given by :

$$\frac{\partial \log L}{\partial p} / p = \hat{p} = 0 \text{ where } \frac{\partial^2 \log L}{\partial p^2} < 0$$

$$\hat{p} = \frac{\sum x}{n}, n = \text{sample size}$$

2.4. Chi-square test of goodness fit :

A very powerful test for testing the significance of discrepancy between theory and experiment was given by Prof. Karl Pearson in 1990 and is known as "Chi-square test of goodness of fit". It enables us to find if the deviation of the experiment from the theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If f_i ($i=1,2,\dots,n$) is a set of observed (experimental) frequencies and e_i ($i=1,2,\dots,n$) is the corresponding set of expected (theoretical or hypothetical) frequencies, then Karl Pearson's Chi-square is given by:

$$\chi^2 = \sum_{i=1}^n \frac{(f_i - e_i)^2}{e_i} \left(\sum_{i=1}^n f_i = \sum_{i=1}^n e_i \right) \dots \dots \dots (2)$$

Follows chi-square distribution with $(n-1)$ d.f.

2.5. Method of Maximum likelihood estimator :

From theoretical point of view, the most general method of estimation known is the method of Maximum Likelihood Estimators (M.L.E) which was initially formulated by C.F. Gauss but as a general method of estimation was first introduced by Prof. R.A. Fisher and later on developed by him in a series of papers

The principle of maximum likelihood consists in finding an estimator for the unknown parameter $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, say, which maximize the likelihood function $L(\theta)$ for variation in parameter, i.e., we wish to find $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ so that $L(\hat{\theta}) > L(\theta)$

Thus if there exists a function $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ of the sample values which maximises Likelihood Estimator (M.L.E)

2.5.1. MLE for Bernouli distribution:

Let X be a random variable
 Let $X=1$, success
 $= 0$, otherwise

With probability of success p
 Then $X \sim B(p)$ and

2.5.2. MLE for Normal distribution

When $x \sim N(\mu, \sigma^2)$, then

$$L = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

$$\log L = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

The likelihood equations for simultaneous estimation of μ and σ^2 are:

$$\frac{\partial \log L}{\partial \mu} = 0 \text{ and } \frac{\partial \log L}{\partial \sigma^2} = 0$$

Thus giving ,

$$\hat{\mu} = \bar{x} \text{ and } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2 \dots \dots \dots (3)$$

2.5.3. MLE of multinomial cell probabilities

: X_1, X_2, \dots, X_m are counts in cell from 1 to m, each cell has a different probability and we fix the number of cells that fall to be $n: x_1 + x_2 + \dots + x_m = n$. the probability of each box is p, with also a constraint $p_1 + p_2 + \dots + p_m = 1$, this is a case in which the X_i 's are not independent, the joint probability of a vector x_1, x_2, \dots, x_m is called the multinomial, and has the form

$$f(x_1, x_2, \dots, x_m | p_1, p_2, \dots, p_m) = \frac{n!}{\prod x_i!} \prod p_i^{x_i} = \binom{n}{x_1, x_2, \dots, x_m} p_1^{x_1} p_2^{x_2} \dots p_m^{x_m}$$

Each cell taken separately against all the other boxes is a binomial.

Now, the log-likelihood of this is given by

$$l(p_1, p_2, \dots, p_m) = \log n! - \sum_{i=1}^m \log x_i! + \sum_{i=1}^m x_i \log p_i$$

Now we have to take the constraint into account, so we have to use the Lagrange multiplier.

We use ,



$$L(p_1, p_2, \dots, p_m, \lambda) = l(p_1, p_2, \dots, p_m) + \lambda(1 - \sum_{i=1}^m p_i)$$

$$\text{Implies, } (A_i B_j)_0 = N \cdot P[A_i B_j] = \frac{A_i B_j}{N} \dots \dots \dots (5)$$

By posing all the derivatives to be 0, we get the most natural estimate

$$\hat{p}_i = \frac{x_i}{n} \dots \dots \dots (4)$$

By using this formula, we can find out expected frequencies for each of the cell frequencies (A_iB_j) (i=1,2,...,r; j=1,2,...,s), under the null hypothesis of independence of attributes.

Maximizing log-likelihood, with and without constraint, can be unsolvable problem in closed form, then have to use iterative procedures.

The exact test for the independence of attributes is very complicated but fair degree of approximation is given, for large samples, by the Chi-square test of goodness of fit, viz.,

2.6. Test of independence of attributes:

Let us consider two attributes A and B. A is divided into r classes A₁, A₂, ..., A_r and B is divided into s classes B₁, B₂, ..., B_s. Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the following table known as (r x s) contingency table where (A_i) is the number of persons possessing the attribute A_i (i=1,2,...,r), (B_j) is the number of persons possessing the attribute B_j (j=1,2,...,s) and (A_i B_j) is the no. of persons possessing both the attributes A_i and B_j, (i=1,2,...,r ; j=1,2,...,s).

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \left[\frac{\{(A_i B_j) - (A_i B_j)_0\}^2}{(A_i B_j)_0} \right] = \sum_{i=1}^r \sum_{j=1}^s \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \quad (6)$$

Where,
 f_{ij} = observed frequency for contingency table category in column i and row j,
 e_{ij} = expected frequency for contingency table category in column i and row j,
 which is distributed as χ² variate with (r-1)(s-1) d.f.

Also,
 $\sum_{i=1}^r A_i = \sum_{j=1}^s B_j = N$, where N is the total frequency.

2.7. Test for proportion :
2.7.1. Test for single proportion : To investigate the significance of the difference between and assumed proportion p₀ and an observed proportion \hat{p} , this test is used.

Table 2.6.1: r x s contingency table:

A \ B	A ₁	A ₂	A _j ...	A _r	Total
B ₁	(A ₁ B ₁)	(A ₂ B ₁)	(A _j B ₁)	(A _r B ₁)	(B ₁)
B ₂	(A ₁ B ₂)	(A ₂ B ₂)	(A _j B ₂)	(A _r B ₂)	(B ₂)
⋮						⋮
B _j	(A ₁ B _j)	(A ₂ B _j)	(A _j B _j)	(A _r B _j)	(B _j)
⋮						⋮
B _s	(A ₁ B _s)	(A ₂ B _s)	(A _j B _s)	(A _r B _s)	(B _s)
Total	(A ₁)	(A ₂)		(A _j)	(A _r)	N

Suppose, p₀ is the assumed proportion.
 To test, H₀: p=p₀
 Against H₁: p≠p₀

The test statistic is

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dots \dots \dots (7)$$

Where, \hat{p} = estimated proportion
 p₀ = assumed proportion
 Reject H₀ if |Z| > Z α /2 .

The problem is to test if the two attribute A and B under consideration are independent or not. Under the null hypothesis that the attributes are independent, the theoretical cell frequencies are calculated as follows:

2.7.2. Test for equality of proportion: This test is to test for the significance of difference between two sample proportion (p₁ and p₂) i.e. the null hypothesis to be tested here is H₀= p₁= p₂
 Against,
 H₁= p₁≠ p₂

P[A_i] = probability that a person possesses the attribute A_i = (A_i)/N ; i = 1,2,...,r
 P[B_j] = probability that a person possesses the attribute B_j = (B_j)/N ; j = 1,2,...,s

The test statistics to test this hypothesis is

$P[A_i B_j] = \frac{A_i B_j}{N}$, i=1,2,...,r; j=1,2,...,s and (A_iB_j)₀ = expected number of persons possessing both the attributes A_i and B_j



$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \quad \dots\dots\dots (8)$$

Where,

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

And $\hat{q} = 1 - \hat{p}$

The critical value of Z at 5% level of significance are 1.96

2.8. Inference by odds ratio:

Let us define,

X : exposure Y: age group status
 X= 1, if exposed Y= 1, if with the 45<age group
 = 0, otherwise =0, if with the 45>age group

Thus, odd ratio is defined as,

$$OR = \frac{pr(X = 1|Y = 1)/pr(X=0|Y=1)}{pr(X = 1|Y = 0)/pr(X=0|Y=0)}$$

$$= \frac{p_{00}p_{11}}{p_{01}p_{10}} \quad \dots\dots\dots (9)$$

Where, p_{00} is the probability that a randomly selected person does not have the risk and also does not have the disease, p_{11} is the probability that a randomly selected person have both risk and the disease. Similarly, p_{01} and p_{10} are defined.

Odds ratio can also be defined w.r.t. cross table as follows:

Table 1.4.2: cross table to find odds ratio:

	Y=1	Y=0	Total
X=1	a	c	a+c
X=0	b	d	b+d
Total	a+b	c+d	N

$$OR = \frac{ad}{bc}$$

Test for OR:

To test the hypothesis : $H_0: OR=1$ or $H_0: OR=0$

$H_1: OR>1$ $H_1: \log_e OR>0$

We have the test statistic,

$$Z = \frac{\log_e \widehat{OR}}{SE(\log_e \widehat{OR})} \sim N(0,1) \quad \dots\dots\dots (10)$$

Where,

$$SD(\log_e \widehat{OR}) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$SE(\log_e \widehat{OR}) = \frac{SD(\log_e \widehat{OR})}{\sqrt{N}}$$

Reject H_0 if $Z > Z\alpha$

2.9. Mann-Whitney U-test :

This non - parametric test for two samples was described by Wilcoxon and studied by Mann and Whitney . It is the most widely used test as an alternative to the parametric t - test ; when we do not make the t test assumptions about the parent population . The null hypothesis to be tested here is that there is no difference between the two groups.

We first combine the values of the two samples , arrange them in order of increasing size and then rank them . In the ranking, the algebraic sign is to be considered i.e. the lowest ranks are assigned to the largest negative numbers , if any .

For two samples with m and n members ($m < n$)

$$U_1 = mn + \frac{m(m+1)}{2} - R_1 \quad \dots\dots\dots (11)$$

$$U_2 = mn + \frac{n(n+1)}{2} - R_2 \quad \dots\dots\dots (12)$$

Then Mann - Whitney U statistics is $U = \min (U_1, U_2)$ (13)

Where ,

R_1 = Sum of the ranks assigned to the group whose size is m

R_2 = Sun of the ranks assigned to the group whose size is n

Then using tables of sigel , we can either reject or accept the hypothesis .When either of M or N is larger than 29 the sampling distribution of U approximately approaches the Normal distribution with

$$\text{Mean} = E(U) = \frac{mn}{2} \text{ and}$$

$$\text{Variance} = V(U) = \frac{mn(m+n+1)}{12}$$



$$Z(U) = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$$

Hence the table of standard normal distribution can be used to test the significant difference between the two populations .

Ties: When tied observations occur, we give each of the tied observations, the average of the ranks they would have, if no ties have occurred. If this occur between two or more observation in the same group , the value of U is not affected . But if ties occur between two or more observations involving both the groups , the value U is affected . This effect changes the variability of the set of ranks. Hence correction for ties must be applied to the standard deviation of U. The standard deviation with this correction is as follows :

$$s.d = \left\{ \left(\frac{mn}{N(N-1)} \right) \left(\frac{N^3 - N}{12} - \sum T \right) \right\}^{\frac{1}{2}}$$

Where, N=m+n and T is the number of observations tied for a given rank.

$\sum T$ is found by summing the T's overall the group of tied observations.

3. Results and Discussion :

3.1. Test of Randomness :

First of all we shall see whether our data is random one. We shall use run test on our data with runs of age and diseases of the patient

H_0 : Sample is random

H_1 : Sample is not random

Let us fix $\alpha=0.01$. Using SPSS package we have-

Table 3.1.1: SPSS output for a Run test

	diseases	age
Test Value ^a	1	46
Cases < Test Value	90	200
Cases >= Test Value	310	200
Total Cases	400	400
Number of Runs	128	196
Z	-1.796	-5.01
Asymp. Sig. (2-tailed)	.072	.617

We have to reject H_0 if p-value < $\alpha/2$, i.e., if p-value < 0.025

From the table we have seen that all the p-values > 0.025. So, we do not reject H_0 and thus we

may conclude that our sample is random.

3.2. Test Of Normality :

For the test of normality, we use the data for present age of cancer patients. For this we use normal P-Plot(using SPSS) as shown in the figure 3.2.1

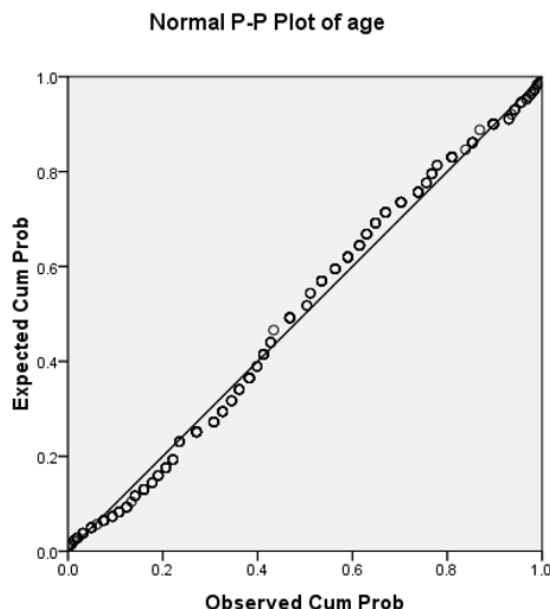


Fig 3.2.1 Normal Probability Plot For Age of cancer patients

From the figures 3.2.1 it can be said that the samples can be regarded as drawn from a normal population and this ensures the validity of tests based on normality.

3.3. Presentation of Age Data with Ogive :

From the data on age of the people, given in appendix, we have that, Lowest age = 9, highest age = 83 and n = total no. of people = 400. By using Sturges' Formula (1.4.5) we get, No. of classes, k = 9 and width i = 8. Then the frequency distribution table is given by,

Table 3.3.1: Distribution of patients by age:

Age	interval	Frequency	Cumulative	Percent
	9-17	10	10	2.5
	18-26	44	54	11.0
	27-35	66	120	16.5
	36-44	54	174	13.5
	45-53	90	264	22.5
	54-62	83	347	20.8
	63-71	41	388	10.2
	72-80	11	389	2.8
	81 & above	1	400	0.2



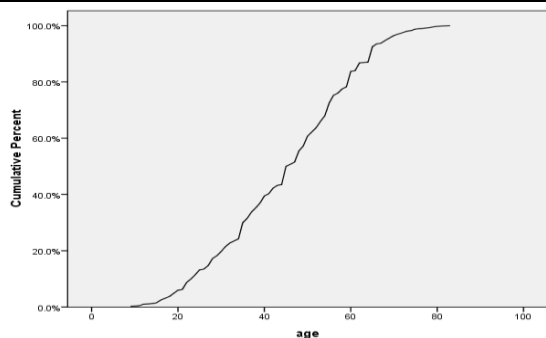


Fig: 3.3.1. Ogive for the age of the Cancer patients

From the ogive we have, $Q_1 = 34$ years, $Q_2 = 45.6$ years, $Q_3 = 55.76$ years

3.4. Test of Goodness of Fit:
3.4.1. Test for uniformity of cancer patients over different age group :

We feel that the cancer patients should be from all the age groups equally. But we have to test Whether the patients are equally distributed over different age group or not.

Table 3.4.1 Distribution of patients by age group:

Cancer	Age interval									Total
	9-17	18-26	27-35	36-44	45-53	54-62	63-71	72-80	81and above	
	10	44	66	54	90	83	41	11	1	400

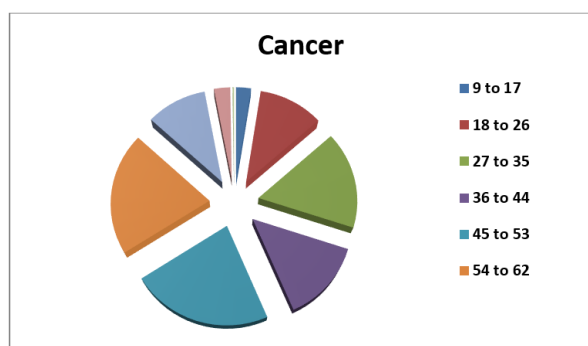


Fig 3.4.1: Pie diagram for cancer patients over different age group

Let us set the hypothesis,
 $H_0 =$ The patients are equally distributed over all the age interval i.e. patients are in the proportion 1:1:1:1:1:1
 $H_1 =$ The patients are not equally distributed

Table 3.4.2 : Table for expected frequencies:

Age interval	9-17	18-26	27-35	36-44	45-53	54-62	63-71	72-80	81 &above
O_i	10	44	66	54	90	83	41	11	1
E_i	44.4	44.4	44.4	44.4	44.4	44.4	44.4	44.4	44.4

Now calculate,
 $\chi^2 = 187.42$ (using 6)
 $\chi^2_{0.05,8} = 15.50$

Since,
 Calculated $\chi^2 >$ Tabulated χ^2 ,

So we reject our null hypothesis and conclude that the cancer patients are not equally distributed over different age group.

Now, consider $P_i =$ probability of persons belonging to i th age group ($i=1, \dots, 9$)

Then the MLE of P_1, P_2, \dots, P_9 for the multinomial distribution are-

$P_1=0.025, P_2=0.11, P_3=0.17, P_4=0.14, P_5=0.23, P_6=0.21, P_7=0.10, P_8=0.03, P_9=0.0025$

We have seen that the probability is highest for the age group (45-53). So on the basis of our sample we can say that the probability of a

patient will belong to the age group (45-53) will be highest(0.23) and the probability that a patient will belong to the age group (81&above) is lowest(0.0025)

3.4.2. Chi-Square Test for Independence of Attributes

We want to test the hypothesis that there is an association between sex and various category of cancer.

Here, sex: male and female

Category: Breast cancer, Colon cancer, Stomach cancer, Gallbladder cancer, Lung cancer, Cervix cancer, Larynx cancer, other cancer

$H_0:$ there is no association between sex and category of cancer

$H_1:$ there is an association



Table 3.4.2: 8×2 Contingency table for sex and category of cancer patients:

sex category	male		female		Total
	O _i	E _i	O _i	E _i	
Breast cancer	0	36.9	90	53.1	90
Colon cancer	23	18.04	21	25.96	44
Stomach cancer	32	21.32	20	30.7	52
Gallbladder cancer	21	18.04	23	25.96	44
Lung cancer	18	8.2	2	11.8	20
Cervix cancer	0	7.38	18	10.62	18
Larynx cancer	17	8.61	4	12.39	21
Other cancer	51	45.51	64	65.5	111
Total	162		238		400

Using SPSS

$$\chi^2=115.964 \text{ with } (8-1)(2-1)=7 \text{ d.f.}$$

Since tabulated value of χ^2 with 7 d.f. at 5% level of significance is 14.06

Here,

calculated $\chi^2 >$ tabulated χ^2

So, we reject the null hypothesis and we can conclude that there is an association between sex and category of cancer

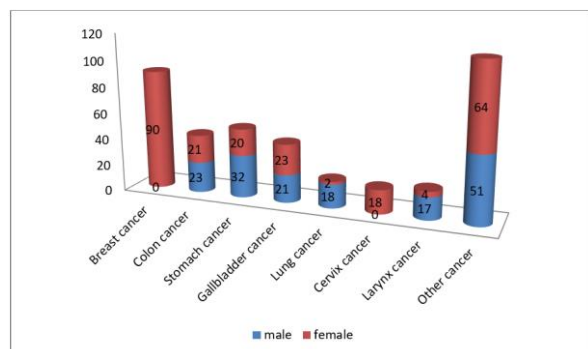


Fig: 3.4.2. Bar diagram for category of cancer w.r.t. sex

3.5. Test of significance of difference proportion of male cancer patients and female cancer patients:

Hypothesis: Proportion of male cancer patients is not equal to female cancer patients in age ≥ 45 years

Here, $H_0: p_1 = p_2$

Against $H_1: p_1 \neq p_2$

Since it is a two tailed test ,so we shall reject H_0 if $Z < -1.96$ or $Z > 1.96$

Let,

p_1 =Proportion of male cancer patients aged ≥ 45 years

p_2 = Proportion of female cancer patients aged ≥ 45 years

Here, $H_0: p_1 = p_2$

Against, $H_1: p_1 \neq p_2$

We have used random variable X and Y as age of male and female cancer patients respectively and

$X=1$, if age ≥ 45

$=0$, otherwise

$Y=1$, if age < 45

$=0$, otherwise

Then $X \sim \text{Ber}(p_1)$ and $Y \sim \text{Ber}(p_2)$ and we shall use MLE s of p_1 and p_2 i.e. \hat{p}_1 and \hat{p}_2 ,

3.5.1 Cross table for sex and age of patients

Sex	Age ≥ 45	Age < 45	Total
Male	112	52	164
Female	114	122	236
Total	226	174	400

Now,

$$\hat{p}_1=0.68, \quad \hat{p}_2=0.48$$

$$\hat{P}=0.562$$

$$\hat{Q}=0.438$$

To test this we shall use Z test (using formula 8)

$$Z = \frac{0.68 - 0.48}{\sqrt{(0.562 - 0.438) \left(\frac{1}{164} + \frac{1}{236} \right)}}$$

$$= 4.08$$

Here, $Z > 1.96$, so we reject H_0 to accept H_1 i.e. proportion of male cancer patients is not equal to female cancer patients in age ≥ 45 years

From our observations ,we have found that the proportion is male cancer patients is more than female cancer patients in age ≥ 45 years

So, we set our hypothesis as

$H_0: P_1 = P_2$ vs $H_1: P_1 > P_2$

We use Z test as (1.4.11) and $Z = 4.08$

Now, with $\alpha = 0.05$, $Z = -1.645$ (for right tailed test)

Since $Z > Z_\alpha$ i.e. $Z > 1.645$

So we reject H_0 to accept H_1 . And we conclude on the basis of our sample that proportion of male cancer patients is significantly higher in our population than the proportion of male cancer patients in age ≥ 45

3.6. Odds Ratios for Association Between Sex and Age of cancer patients :

Claim: There is no association between sex and



age group of cancer patients
 Let X and Y represents the attributes sex and age group of patients. And
 X=1, if male
 =0, if female
 Y=1, if age ≥45
 =0, if age <45

Table 3.6.1 Cross table for sex and age of patients

Sex	Age group		Total
	Y=1(if age ≥45)	Y=0 (if 45< age)	
Male	112	52	164
Female	114	122	236
Total	226	174	400

Here,

$$OR = \frac{\text{odds of male cancer patients in age group} \geq 45}{\text{odds of male cancer patients in the age group} < 45}$$

$$\widehat{OR} = 2.30$$

Since OR= 2.30, it seems ,there is an association between sex and age of patients. And cancer patients coming from the age ≥ 45 is 2.30 times more likely to be coming from the age<45.Next we are to test,

$$H_0: OR=1 \text{ or } H_0: \log_e OR=0$$

$$H_1: OR>1 \text{ or } H_1: \log_e OR>0$$

Using formula (9) we get

$$OR=2.30 \log_e OR=0.83 \quad SE(\log_e OR)=0.10$$

Using formula (1.4.13) we get

$$Z=8.3$$

Here, $Z > Z_\alpha$ the test is significant, so we reject the null hypothesis H_0 . Hence we conclude that there is a significant association between sex and age of cancer patients

3.7. Test for difference in Medians age of male and female cancer patients:

Assuming that we could collect only a small no. of data for our study. We have tried to test the hypothesis that –

There is a difference between male and female patients w.r.t their median age .

Since our sample size is small so we have to take help o a non- parametric method to test this hypothesis .We write

Rank	6	11	14	21	9	19	12	4.5	18	22	17	3	1	16	10	2	4.5	13	7.5	15	20	7.5
Observation	20	36	42	73	25	70	38	18	65	75	58	16	9	54	32	15	18	39	22	46	71	22

In our study we have found that minimum age of the cancer patients is 9 and maximum is 83. So the range is 74. But the patients are not

$$H_0: M_X=M_Y$$

$$H_1: M_X \neq M_Y$$

Where,

M_X =Median age of male patient

M_Y = Median age of female patient

we have taken a random sample of size 12 for male and 10 for female which is shown in observation **table 3.6.1** of age for male and female-

Table 3.6.1: observation table on age of male and female patients:

male	20, 36, 42, 73, 25, 70, 38, 18, 65, 75, 58, 16
female	9, 54, 32, 15, 18, 39, 22, 46, 71, 22

Here,

n_1 = size of the small sample = 10

n_2 =size of the large sample = 12

The combined sample are given below-

R_1 = Sum of the ranks assigned to the sample whose size is small = 96.5

R_2 = Sum of the ranks assigned to the sample whose size is large = 156.5

Using formula (11)&(12)

We get, $U = 41.5$ using formula(13)

Here,

$$U(10,12) < U$$

Hence, we accept the null hypothesis and we can conclude that median age of male and female cancer patients are equal.

4. Conclusions:

Our study is based on a sample of size 400 collected from JMCH out of which 40.8% patients are male and 59.2% are female. Again 22.5% Breast cancer, 13% Stomach cancer, 11% Gallbladder cancer, 5.2% Larynx cancer, 5% Lung cancer, 4.5% Cervix cancer and 27.8% other cancer, where others included Liver cancer, Penile cancer , Anal cancer, Acute lymphoblastic cancer, Glioblastoma cancer, Testicular cancer, Ovarian cancer, Head And Neck cancer, Lip cancer, Tongue cancer.

uniformly distributed over the age groups (9-17), (18-26), (27-35), (36-64), (45-53), (54-62), (63-71), (72-80), (81&above), as the test for



uniformity was found to be significant. Also the maximum likelihood estimates of the probabilities shows that the probability that a cancer patients will belong to the age group(45-53) is highest with probability 0.23 and the probability is lowest to the age group(81& above) with probability 0.0025 .

The ogive shows that the median age of the patient is 45.6 years i.e. 50% of the cancer patients are of age< 45.6 years. Again 25% of the patient are of age less than 34 years($Q_1=34$); 75% Patient's age is less than 55.76 years in our study population.

We have tried to study the association between sex and age of patients with the help of odds ratio, we have found that there is a significant association between these two attributes and cancer patients coming from age ≥ 45 is 2.30 times more than the cancer patients coming from age<45. The test from equality of proportion also shown that proportion male cancer patient is significantly higher than the proportion of female cancer patient in age ≥ 45 . In our study we have also found that though there is a difference in the median age of male and female cancer patients, the difference is not statistically significant as shown by the median test.

References :

- Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331), 1330-1334.
- Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., ... & Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International journal of cancer*, 144(8), 1941-1953.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6), 394-424.
- WHO: Global Cancer Observatory, International Agency for Research on Cancer
- WHO. NCD Country Profile. Geneva: World Health organization; 2011
- Takiar, R., Nadayil, D., & Nandakumar, A. (2010). Projections of number of cancer cases in India (2010-2020) by cancer groups. *Asian Pac J Cancer Prev*, 11(4), 1045-1049.
- Okonkwo, Q. L., Draisma, G., der Kinderen, A., Brown, M. L., & de Koning, H. J. (2008). Breast cancer screening policies in developing countries: a cost-effectiveness analysis for India. *JNCI: Journal of the National Cancer Institute*, 100(18), 1290-1300.

- Bloom, D. E., Cafiero-Fonseca, E. T., Candeias, V., Adashi, E., Bloom, L., Gurfein, L., ... & Saxena, A. (2014). Economics of non-communicable diseases in India: the costs and returns on investment of interventions to promote healthy living and prevent, treat, and manage NCDs. In *World Economic forum, Harvard school of Public health* (pp. 22-23).
- Confortini, C. C., & Krong, B. (2015). Breast cancer in the global south and the limitations of a biomedical framing: a critical review of the literature. *Health Policy and Planning*, 30(10), 1350-1361.
- Brown, M. L., Lipscomb, J., & Snyder, C. (2001). The burden of illness of cancer: economic cost and quality of life. *Annual review of public health*, 22, 91.

