International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Statistical Analysis of Cancer Data

Dr. S. N. Prasad[1], **Ajay Kumar Diwakar**[2]

[1]Professor & Head of Department Radiotherapy, JK Cancer Institute, Kanpur -208002 (Uttar Pradesh) , India

[2]Research Scholar & Project Technical Officer, JK Cancer Institute, Kanpur -208002 (Uttar Pradesh) , India

Abstract: ***A set of illnesses known as cancer [1] are characterised by abnormal cell proliferation and have the ability to infiltrate or spread to different bodily regions. It is the body's unchecked cellular proliferation of aberrant cells. It appears when the body's regular regulatory system malfunctions. Instead of dying, old cells proliferate uncontrollably to produce new, aberrant cells. These excess cells could aggregate into a tissue mass known as a tumour. The project's goal is to identify the greatest number of people impacted by cancer in each case while also analysing data on various types of cancer. Additionally, to forecast a cancer case in India, to increase public awareness, and to evaluate the findings using statistical analysis.***

Keywords: Statistical analysis, Experimental Design, One way Anova

1. Introduction

A category of illnesses known as cancer [2] involves aberrant cell proliferation and has the potential to invade or spread to different bodily regions. It is the body's unchecked cellular proliferation of aberrant cells.

Tumors that are cancerous or malignant [3] have cells that are unable to stop growing. They have become independent and won't stop dividing, to put it another way. To put it more simply, almost every cell in the body has some capacity for growth and cell division. It is crucial for all living things to do this. However, there is an issue with the DNA when cells behave erratically. Mutations, which are modifications to the DNA sequence, make cells forget how to cease dividing when they take place. The cell clump eventually develops into a tumour. The tumour may be benign, which indicates it is not cancerous, or malignant, which implies it is cancer.

The distribution of various cancer cases by state from 2019 to 2021 is included in the analysis of the cancer [4] data in this project effort. The purpose of this study is to determine, using Analysis of Variance [5], whether there is a significant difference between the mean number of cancer patients in various types of groups. We determine which states, sex groups, and age groups are disproportionately afflicted by cancer if there is a discernible difference. Finally, the likelihood of developing cancer is predicted by fitting an appropriate trend equation. The results are displayed graphically and explained using statistical analysis.

2. Proposed Analysis

a) The DATA set analysed in this project work comprises of various distribution of cancer data from the period 2019 to 2021.

Table I: Shows sample data set of various cancer cases

Type of Cancer	2019	2020	2021
Head and Neck Cancer	2476	1217	1176
Breast Cancer	480	252	276
Gallbladder	331	121	148
Cervix Cancer	287	161	199
Lung Cancer	250	123	131
Ovary	113	71	83
Liver	108	34	20
Chronic Myeloid Leukaemia	60	17	28
Colon and Rectum	59	30	37
Stomach	54	22	17
Other	1107	443	363

3. Methodology

3.1 Analysis of Variance

Analysis of variance (ANOVA) [6], is the hypothesis testing used for more than two

Analysis of variance (ANOVA) [6], is the hypothesis testing used for more than two samples. The data has only one Independent variable so One way Anova (or) Complete randomised design (CRD) [7] is applicable. It test the hypothesis that there is a significant difference between the population (more than two) means. F test [8] is used for analysis of variance.

The hypothesis is given by the Equation (1) and (2)

Ho:m1=m2=…=mn (1)

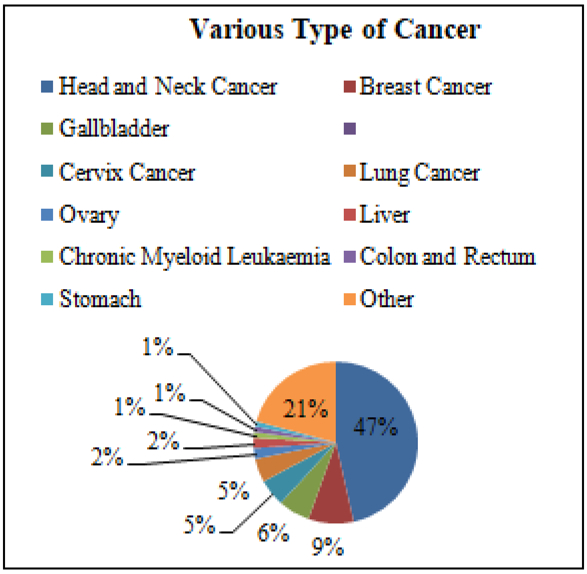
H1: At least two means are different (2)

where, Ho is null hypothesis and it tells that there is no significant difference between the population means (means are same); H1 is alternative hypothesis and it tells that is a significant difference between the population. If the test statistics calculated is greater than the critical value then Ho is rejected and H1 is accepted else Ho is accepted.

4. Result and Discussions

a) Various Type of cancer distribution of cancer cases

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
2019	11	5325	484.0909	529750.1		
2020	11	2491	226.4545	124030.5		
2021	11	2478	225.2727	112109.2		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	489004.1	2	244502	0.957718	0.395191	3.31583
Within Groups	7658898	30	255296.6			
Total	8147902	32				



5. Conclusion

Thus, by analysing the Cancer data from the year 2019 to 2021, by the above results and discussions we have reached a conclusion that the various cancer cases which is maximum affected by Cancer is Uttar Pradesh. UP has to take some serious effects in their lifestyle to overcome this problem.

Finally, the prediction of cancer cases follows a linear pattern (ie) decreasing year by year. This research work can be further extended by analysing the cancer dataset by other statistical and machine learning techniques to discover the hidden and innovative results.

References

[1] www.cancerresearchuk.org/about-cancer/what-is-cancer
[2] https://en.wikipedia.org/wiki/Cancer
[3] https://www.cancer.org/
[4] https://www.cancer.net/cancer-types
[5] https://en.wikipedia.org/wiki/Analysis_of_variance
[6] www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/
[7] https://www.statisticssolutions.com/anova-analysis-of-variance/
[8] https://www.youtube.com/watch?v=ITf4vHhyGpc
[9] https://www.statisticssolutions.com/what-is-linear-regression/
[10] https://en.wikipedia.org/wiki/Linear_regression
[11] https://onlinecourses.science.psu.edu/stat501/node/250
[12] https://www.youtube.com/watch?v=zPG4NjIkCjc