Intro to Inferential Statistics

Final Project using Haberman Survival Dataset

1. Introduction:

The project uses the Haberman's Survival Dataset obtained from UCI repository. The Dataset contains the patient information from the study conducted on the survival of the patients who underwent surgery for breast cancer.

Data Set Information:

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Attribute Information:

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute)

-- 1 = the patient survived 5 years or longer

-- 2 = the patient died within 5 year

The dataset has 306 instances and three attributes and one class attribute. There are no null values and missing values in the dataset.

```
df.describe()
```

	Age	Year	No_nodes	Survival
count	306.000000	306.000000	306.000000	306.000000
mean	52.457516	62.852941	4.026144	1.264706
std	10.803452	3.249405	7.189654	0.441899
min	30.000000	58.000000	0.000000	1.000000
25%	44.000000	60.000000	0.000000	1.000000
50%	52.000000	63.000000	1.000000	1.000000
75%	60.750000	65.750000	4.000000	2.000000
max	83.000000	69.000000	52.000000	2.000000

Figure 1

The dataset is imbalanced as the class survival=1 has 225 patients who survived 5 years or longer and survival=2 has 81 patients who survived less than 5 years. The average number of positive axillary nodes is 4 and the maximum number of nodes is 52. The mean age is 52 and the max age is 83 as can be seen from Figure 1.

2. Research Question and Hypothesis:

The Haberman survival dataset has attributes pertaining to the age of the patient, the year the patient underwent operation, the number of positive axillary nodes present and were removed during the surgery and the class attribute Survival years. The research work pertains to determining the relation between the number of positive axillary nodes removed and the survival rate of the patient.

Research Question: Is there a relation between the number of positive axillary nodes removed and the survival rate of the patient.

Hypothesis H_0 : The number of positive axillary nodes removed in the surgery does not determine the survival rate of the patient.

Alternate Hypothesis H_A: The number of positive axillary nodes removed in the surgery determines the survival rate.

3. Experimental Design:

The sample size selected is 30 which is obtained by using the resample method without replacement. Pearson Correlation is used to determine the correlation between the positive axillary nodes and the survival rate of the patient. The pearson correlation is chosen as

- Both the features No_nodes and Survival are quantitative
- Both the features are normally distributed.

The alpha value chosen is 0.05 pertaining to the confidence interval of 95%. It is a two tailed test.

```
df.corr()
```

	Age	Year	No_nodes	Survival
Age	1.000000	0.089529	-0.063176	0.067950
Year	0.089529	1.000000	-0.003764	-0.004768
No_nodes	-0.063176	-0.003764	1.000000	0.286768
Survival	0.067950	-0.004768	0.286768	1.000000

Figure 2

The Pearson's Correlation value obtained for No_nodes and Survival is r = 0.2867 from Figure 2. Also, r^2=0.0821. Degrees of freedom df=58 (since N-2=58 where N=60 the sample size 30 for each attribute). The t-statistic value is calculated using the below formula.

$$t = r\sqrt{N-2} \div \sqrt{1-r^2} \qquad \text{.....(1)}$$

The t-statistic value calculated is **2.2787**. The t critical value obtained from the t table corresponding to α=0.05 and the degrees of freedom=58(N-2) is **1.671**. The t-statistic value obtained is greater than the t-critical value. The p-value is a statistical measure used to validate a hypothesis and measures the probability of obtaining the observed results. The lower the p value, greater is the statistical significance of the observed difference. The p value obtained from graphpad.com/quickcals is **p=0.1179.**

4. Results:

The t-statistic value obtained 2.2787> 1.671 and the null hypothesis is rejected. The p value obtained is 0.1179 which is greater than the alpha value of 0.05 ie. p>0.05. Hence, the deviation from the null hypothesis is not statistically significant and the null hypothesis is rejected. The histogram and PDF of the number of positive axillary nodes is given in Figure 3.

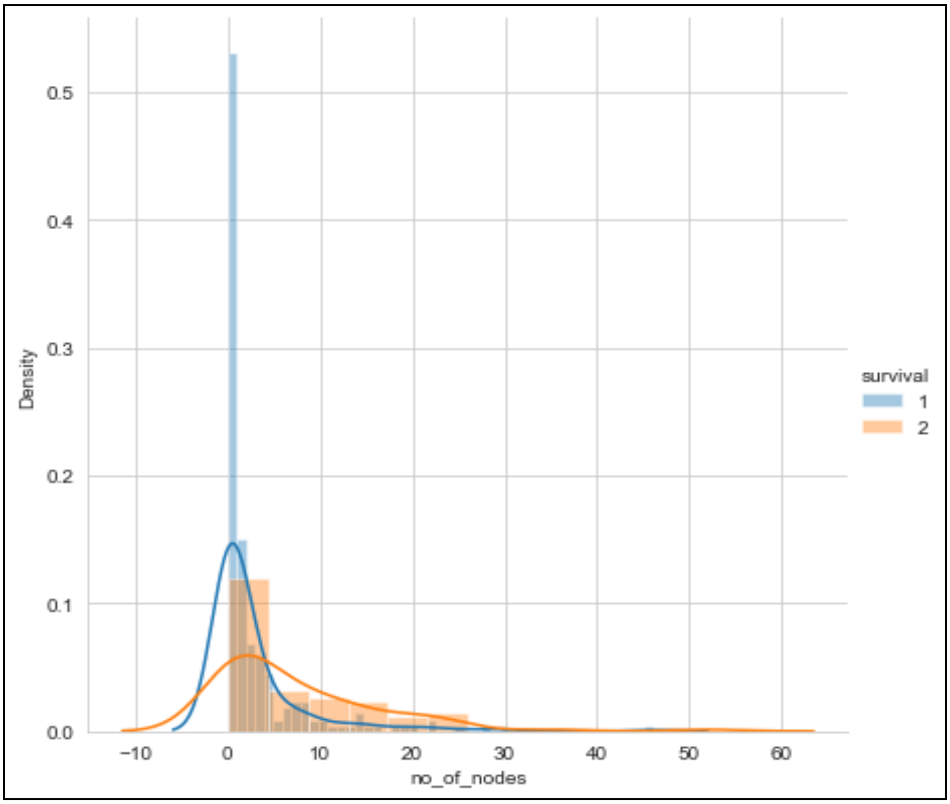

Figure 3

It is evident from Figure 3 that the survival rate of the patients is more when the no_of _nodes removed is less and the survival rate is less when the no_of_nodes removed is more. Hence if the number of axillary nodes is less then the survival of the patients for 5 years is more.

5. Conclusions:

The p value obtained is greater than the alpha value of 0.05 and hence the null hypothesis is rejected. The null hypothesis that 'the number of positive axillary nodes removed in the surgery does not determine the survival rate of the patient' is rejected. The correlation value between No of positive axillary nodes and the survival is 0.2867 indicating weak and positive association. Hence, it is concluded that **the number of positive axillary nodes removed in a surgery is weakly associated with the survival rate of the patient**.