**INTRODUCTION**

**Title of the data: Haberman's Survival Data**

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.

Analyzing cancer survival data over time allows researchers to track progress in cancer treatment and management. By comparing survival rates from different time periods, researchers can identify improvements or gaps in cancer care and develop strategies to address them.

5. Number of Instances: 306

6. Number of Attributes: 4 (including the class attribute)

7. 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

8. Missing Attribute Values: None

**Description of the data set.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Age* |  | *Year of Operation* |  | *No of positive axillary nodes detected* |  | *Survival Status* |  |
|  |  |  |  |  |  |  |  |
| Mean | 52.4575163 | Mean | 62.8529412 | Mean | 4.02614379 | Mean | 1.26470588 |
| Standard Error | 0.61759226 | Standard Error | 0.1857561 | Standard Error | 0.41100513 | Standard Error | 0.02526169 |
| Median | 52 | Median | 63 | Median | 1 | Median | 1 |
| Mode | 52 | Mode | 58 | Mode | 0 | Mode | 1 |
| Standard Deviation | 10.8034523 | Standard Deviation | 3.24940466 | Standard Deviation | 7.18965351 | Standard Deviation | 0.44189912 |
| Sample Variance | 116.714583 | Sample Variance | 10.5586307 | Sample Variance | 51.6911175 | Sample Variance | 0.19527483 |
| Kurtosis | -0.589393 | Kurtosis | -1.1188257 | Kurtosis | 11.7308769 | Kurtosis | -0.8566113 |
| Skewness | 0.14650506 | Skewness | 0.07875486 | Skewness | 2.9838229 | Skewness | 1.07192839 |
| Range | 53 | Range | 11 | Range | 52 | Range | 1 |
| Minimum | 30 | Minimum | 58 | Minimum | 0 | Minimum | 1 |
| Maximum | 83 | Maximum | 69 | Maximum | 52 | Maximum | 2 |
| Sum | 16052 | Sum | 19233 | Sum | 1232 | Sum | 387 |
| Count | 306 | Count | 306 | Count | 306 | Count | 306 |
| Confidence Level(95.0%) | 1.21528098 | Confidence Level(95.0%) | 0.36552572 | Confidence Level(95.0%) | 0.80876454 | Confidence Level(95.0%) | 0.04970926 |

**RESEARCH QUESTION & HYPOTHESIS**

1. **Are younger patients at the time of operation likely to survive?**

**H0**: There is no relationship between the age of the patient at the time of operation and the survival status. µ1-µ2 = 0

**Ha**: Patients with younger age at the time of the operation are more likely to survive. µ1-µ2<0

We would need to perform a two-sample t-test in order to determine if the difference between the two samples mean is statistically significant or not. The samples are independent so a two samples independent t-test would be appropriate.

|  |  |
| --- | --- |
| **Row Labels** | **Survival Status** |
| 1 | 225 |
| 2 | 81 |
| **Grand Total** | **306** |

The mean of the survival status is

|  |  |
| --- | --- |
| Mean (1) | 52.0177778 |
| Mean (2) | 53.6790123 |
|  |  |

The two samples have larger than 30 number of observations. The mean age of patients who survived is slightly smaller than those who did not.

I will now check if this is due to real difference between the two means or due to sampling variability. We are going to use 5% as our significance level.

|  |  |  |
| --- | --- | --- |
| t-Test: Two-Sample Assuming Equal Variances | | |
|  |  |  |
|  | Variable 1 | Variable 2 |
| Mean | 52.01777778 | 53.6790123 |
| Variance | 121.2675397 | 103.370679 |
| Observations | 225 | 81 |
| Pooled Variance | 116.5578395 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 304 |  |
| t Stat | -1.187499049 |  |
| P(T<=t) one-tail | 0.117978926 |  |
| t Critical one-tail | 1.649881428 |  |
| P(T<=t) two-tail | 0.235957851 |  |
| t Critical two-tail | 1.967798141 |  |

With a 5% level of significance and a p-value of 0.1179 and a t-statistic that is not extreme we can conclude that there is insufficient evidence to reject the null hypothesis.

1. **Are patients with lower number of positive axillary nodes detected before operation more likely to survive?**

**Ho:** There no relationship between the number of positive axillary nodes detected and survival status.µ1-µ2 = 0

**Ha:** Patients with lower number of positive axillary nodes detected before operation more likely to survive. µ1-µ2 < 2

I would need to perform two sample t-test in order to determine of the difference between the two-sample means is stat significant or not. A two-sample independent test would be appropriate since the samples are independent.

|  |  |
| --- | --- |
| Mean (1) | 2.79111111 |
| Mean(2) | 7.45679012 |

The mean number of positive axillary nodes of patients who survived is smaller than those who did not.

I would now check to see if these is due to real difference or due to sampling variability using 5% as the significance level

|  |  |  |
| --- | --- | --- |
| t-Test: Two-Sample Assuming Equal Variances | | |
|  |  |  |
|  | Variable 1 | Variable 2 |
| Mean | 2.79111111 | 7.45679012 |
| Variance | 34.4606349 | 84.3762346 |
| Observations | 225 | 81 |
| Pooled Variance | 47.596319 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 304 |  |
| t Stat | -5.2191674 |  |
| P(T<=t) one-tail | 1.6677E-07 |  |
| t Critical one-tail | 1.64988143 |  |
| P(T<=t) two-tail | 3.3354E-07 |  |
| t Critical two-tail | 1.96779814 |  |

Given an exceptionally high t-statistic value and an extremely low p-value, we have substantial evidence to reject the null hypothesis. Consequently, we can confidently conclude that patients with a lower number of positive axillary nodes detected before surgery have a higher likelihood of survival.

**CONCLUSION**

Considering the result from the test, we can agree that patients who are younger in age at the time of the operation are more likely to survive and with a 5% level of significance and a p-value of 0.1179 and a t-statistic that is not extreme we can conclude that there is insufficient evidence to reject the null hypothesis in the first question.