# A statistical analysis on the survival of breast cancer using Haberman’s data set

### Introduction

It is estimated that about 12% of the world’s female population are affected by breast cancer nowadays. There are many factors that would contribute to one having breast cancer for example overdose of alcohol, obesity and late pregnancy. The difficulties in treating breast cancer also varies and due to this reason, 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 (a.k.a Haberman’s survival data). By analyzing the result of the study, one might be able to obtain an insight into what factors would affect the chances of survival hence helping medical staff to better inform the patient and also more efficiently allocate medical resources.

### Research Questions and Hypothesis

The Haberman’s data include four attributes as follows:

1. Age of patient when operation is carried out
2. The year that the operation was carried out
3. Number of positive cased axillary nodes found
4. Survival status 5 years after the operation

All attributes in the list apart from number 3 are self explanatory and straightforward. The third item in the list describes the number of positive cased axillary nodes found in the patient. This is an important metric because about 75% of the lymph in breast drains to the axillary nodes. If cancer cells are discovered in the axillary nodes, it is very probable that breast cancer is present. And it is estimated that the more axillary nodes are affected, the more severe the situation is.

In this investigation, a total of 2 hypotheses will be tested against the data. They are described as follows:

1. Due to the common understanding that our bodies get weaker as we get older, it is reasonable to hypothesize that the older the patient is during the operations, the less likely for he/she to survive in the future 5 years
2. Technological advances and gathering of experiences should allow doctors to handle cancer cases more efficiently and hence improves the survival chance of patients. Therefore it is hypothesized that the later in the decade that the operation was carried out, the more likely for a patient to survive the future 5 years

### Experimental Design

Chi-square tests are used in this investigation because we are trying to see the relationships between two variables i.e. age/year at operations and survival status.

1. The first test is to examine the relationship between the age when operations were carried out and the survival status 5 years later. The null hypothesis would be there is no relationship between those variables.

In order to reduce the number of categories in the ‘age’ variable, it is decided that age groups should be set up and separate the patients into various groups. Figure. 1 shows the distribution of ages within the data.

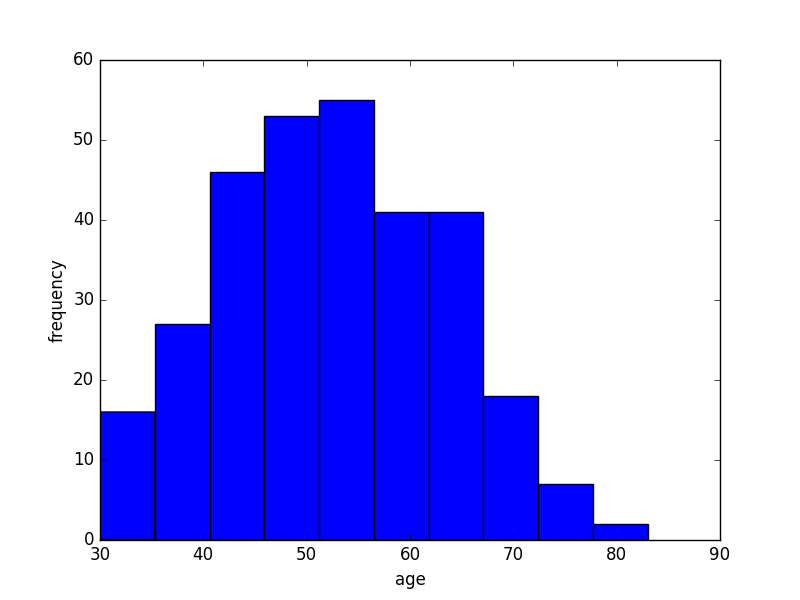


Figure.1 Age distribution within the data set

The data can approximately represent a slightly positive skewed distribution with a mean age of 52.46 and a standard deviation of 10.79. The maximum age in the group is 83 and the minimum age is 30. In order to fulfill some test requirements that will be discussed later, a group size of not smaller than 10 is desired. Therefore three age groups will be set up namely 30-45, 46-60 and >60.

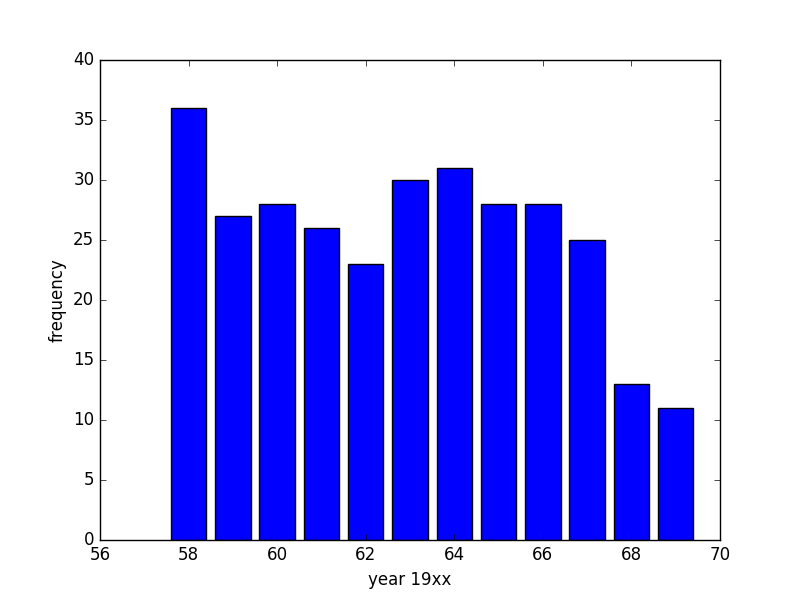
1. The second test will be performed to check whether operations carried out later in the decade would result in a lower death chance. The null hypothesis would be there is no difference in patient survival chance within around a decade (1958 to 1970). Figure.2 shows the distribution of operations carried out in during the 1958 – 1970 period.

Figure.2 Number of operations carried out in the respective year

The test period will be split into three equally long eras for the ease of analysis. Each era will include a 4-year period and test will be performed to determine whether are significant differences in survival status as time advances.

In order to justify the use of chi-square test in this investigation, the pre-assumption for chi-square tests will be discussed.

* The independence of each data variable: Every entry in the data set represents the information of one patient. And the information of a certain patient does not affect the information of other patients. Therefore we can conclude that the data are independent from each other.
* Reasonable sample sizes and expected values (a very minimum of 5): With 306 entries in the data it is reasonable to assume that the expected value of each categories would be greater than 5.

### Results

With regards to hypothesis one, the patients are divided into different age groups and their survival status is recorded. In Table.1, the counts for each category are recorded. From the total survival ratio, the expected values of each category are also determined. The Chi Square statistics can then be calculated and is shown below:

Χ2 = 1.69

P = 0.43

With the above statistics, the result of the first investigation is considered insignificant. This means it is not reasonable to reject the null hypothesis hence one cannot display there is a relationship between age at operations and their survival chances.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 30-45 | 46-60 | >60 | Total |
| Survived | 70 (65.44) | 100 (102.94) | 55 (56.62) | 225 |
| Not-survived | 19 (23.56) | 40 (37.06) | 22 (20.38) | 81 |
| Total | 89 | 140 | 77 | 306 |

Observed Value

(Expected Value)

Table.1 Result table corresponding to hypothesis one

Similar to the above analysis, a Chi Square test is also performed to validate hypothesis two. The patients were divided into different groups by the year that they took the operations and their survival status was also recorded. In Table.2, the counts for each era are recorded. From the total survival ratio, the expected values of each era are also determined. The Chi Square statistics was then calculated and shown below:

Χ2 = 3.00

P = 0.22

Comparing to the result of the first investigation, the P value for the second hypothesis is lower. Yet it is still not low enough to consider the statistics to be significant. This means it is not reasonable to reject the null hypothesis hence one cannot display there is a relationship between year of operations and the patients’ survival chances.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1958-1961 | 1962-1965 | 1966-1969 | Total |
| Survived | 89 (86.03) | 76 (82.35) | 60 (56.62) | 225 |
| Not-survived | 28 (30.97) | 36 (29.65) | 17 (20.38) | 81 |
| Total | 117 | 112 | 77 | 306 |

Observed Value

(Expected Value)

Table.2 Result table corresponding to hypothesis two

### Discussion

From the results shown in the previous session it can be concluded that the analysis does not support the hypothesis of this investigation. According to a research carried out by Cancer Research UK (2007 – 2011), it claimed that “Five-year net survival is highest in the youngest adults for nearly all cancers, with survival generally decreasing with increasing age.”[1]. The data of their study is obtained and the breast cancer related information is extracted and plotted in Figure. 3.

Figure.3 Relationship between age and survival rate

From Figure.3 it can be seen that the survival rate is the highest for the age group of 60 – 69. This is not the expected observation. Figure. 3 shows that the lowest survival rate age group is the eldest group which is between 80 – 99 years old. This is reasonable because as patients are older their immune system gets weaker and their metabolism becomes slower. These would all negatively impact their recovery progress hence lower the chances of survival. However from the statistics it can also be seen that the survival rate of the age group 15 – 39 is very similar to that of 70 – 79. The 70 – 79 age group’s survival rate, similar to the 80 – 99 group is a result of aging. However the low survival rate of the younger age group is caused by the difference in cancer itself. According to a research carried out by Partridge et al, prognosis tends to be worse in women under 40 than in older women. Breast cancers in younger women are more likely to be fast growing, higher grade and hormone receptor-negative. [2] Therefore this could be the case why fatalities in younger ages are amongst the highest.

On the other hand, treatment year seem to not affect the survival rate either. According to an article from cancer.org, the appearance of chemotherapy took place after Second World War. The first metastatic cancer was healed in 1956[3]. Therefore it can be deduced that the period of 1960 to 1970 is still considered early in the treatment of cancer. Scientific research and development did not take place in a fast enough pace to show the difference in survival rate. However if we compare the survival data with later researches, it can clearly demonstrate that the more modern the era, the more likely the patient is to survive.

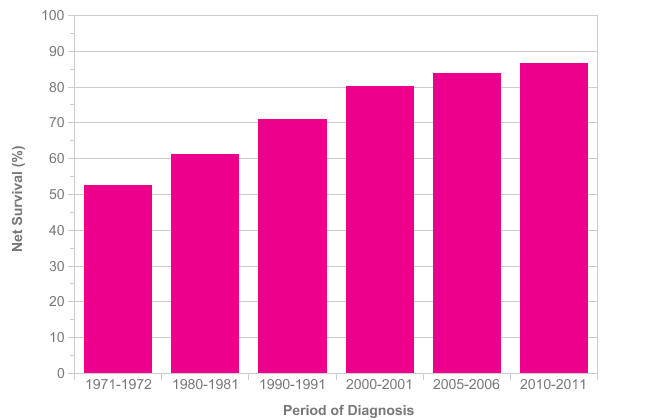


Figure. 4 Breast cancer 5-year survival rate in different decades

Figure. 4 shows the 5 year survival rate of breast cancer patient in the past decades. It shows an obvious trend of increasing survival rate as years go by. This can be best explained by technological advancement so treatments can are more effective.

### Conclusion

Although neither the hypothesis were supported by the analysis, it is still interesting to discover surprising facts that don’t entirely agree with common sense. In a typical situation it would be a fair guess that mortality would increase with age but this is not true for the case of breast cancer. This is due to the difference in speed and grade of cancer cells development. The investigation has also shown that the advancement in technology between 1960 and 1970 is perhaps not so quick that it did not significantly affect the 5-year survival rate of breast cancer patients.

### References

1.

http://www.cancerresearchuk.org/health-professional/cancer-statistics/survival/age#heading-Zero

2. http://ww5.komen.org/BreastCancer/YoungWomenandBreastCancer.html#sthash.plWcDkcP.dpuf

Partridge AH, Goldhirsch A, Gelber S, Gelber RD. Chapter 85: Breast Cancer in Younger Women, in Harris JR, Lippman ME, Morrow M, Osborne CK. Diseases of the Breast, 5th edition, Lippincott Williams & Wilkins, 2014.

3.

http://www.cancer.org/cancer/cancerbasics/thehistoryofcancer/the-history-of-cancer-cancer-treatment-chemo