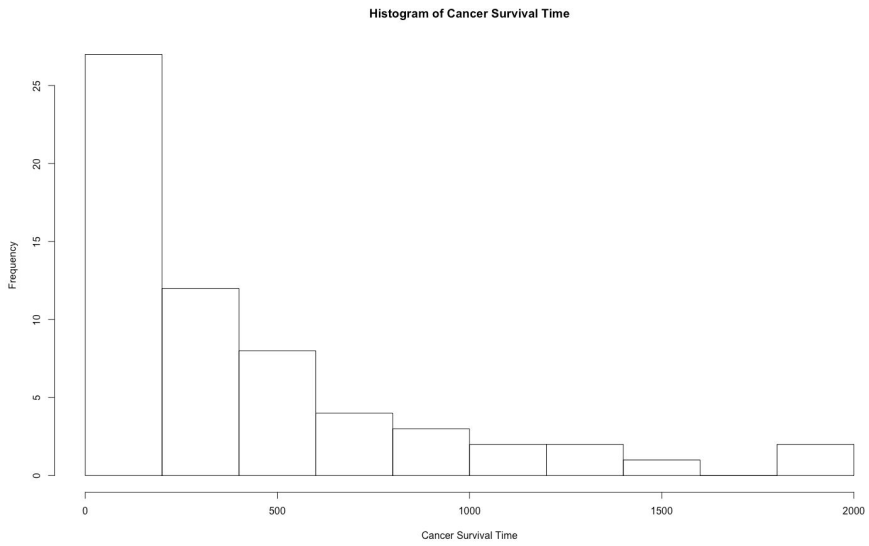
Introduction:

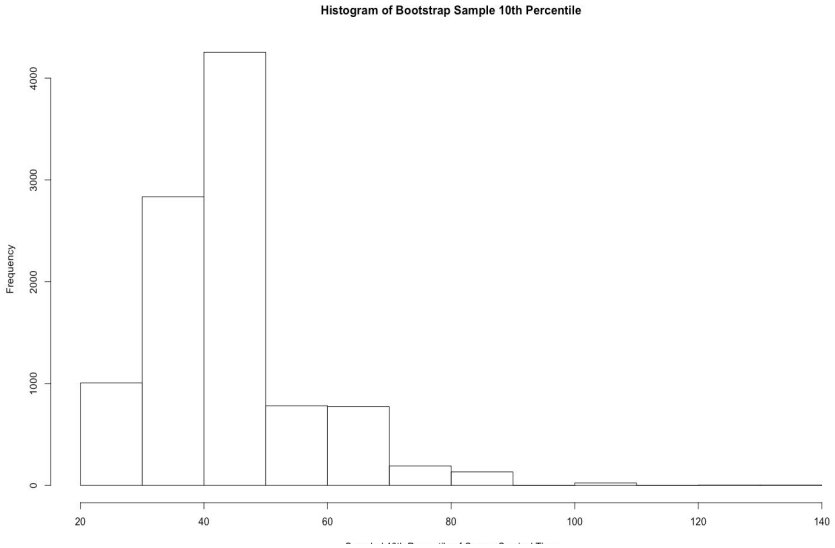
Cancer is a disease that is becoming more and more common in this day and age. It comes in many different forms and has many causes ranging from ones environment to their genetics. Although this disease is deadly, the length of time between diagnosis and death is variable. This project will attempt to find the approximate 10th and 90th percentiles of survival time, in days, of people diagnosed with cancer.

Although this subject may be of great interest to a lot of people, there is still a lot we do not know. For instance, we do not know the true distribution of survival time for people with cancer. Because of this, we can not perform traditional statistics, hypothesis testing, and confidence intervals. Instead, based on our original sample data, we will create a bootstrap distribution of size 10,000 to estimate the actual distribution of the 10th and 90th percentiles of survival time. The only assumptions for creating a bootstrap distribution is that our data is randomly sampled which we assume is true for our data.

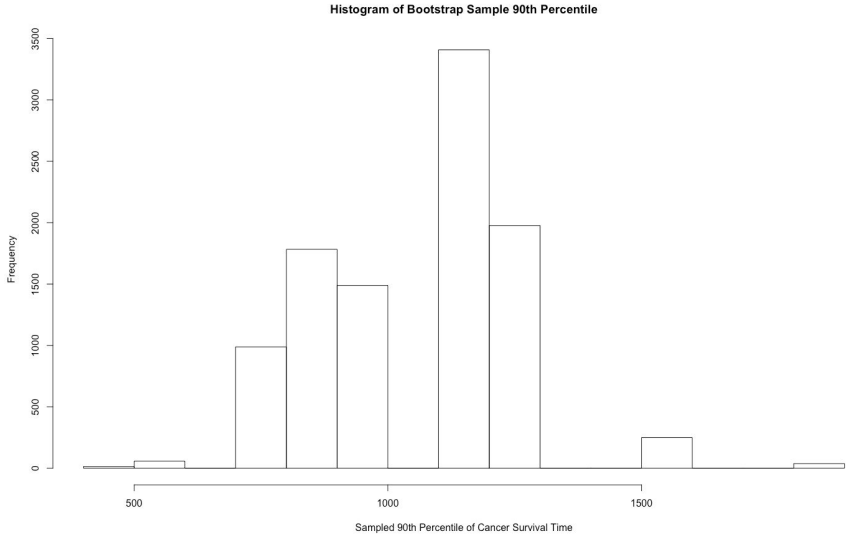
Summary:



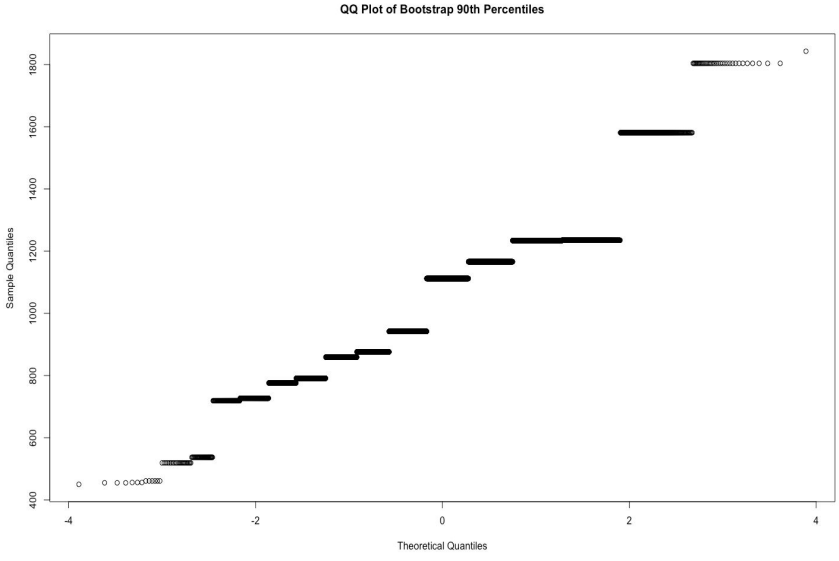
The first graph on the left, is the sample distribution of survival time of people with cancer. From this graph we can see that the original data is non-symmetric, right skewed, and has a heavy left tail.



To the left is the bootstrap distribution of the bootstrap sampled 10th percentiles. We can see that this data has a single peak around 40 days but is also skewed to the right.



The last histogram is the bootstrap distribution of the of the of the bootstrap sampled 90th percentiles. From the graph we can see the distribution is somewhat symmetrical with a single peak around 1200 and has many gaps between bins.



Following is a qq-plot of the bootstrap sampled 90th percentiles. Although the data looks symmetric a qq-plot was

done as well to better see the skew and symmetry of the data. We can see a slight tail on both ends and many large gaps between values.

Analysis:

10,000 bootstrap samples were taken from the original data and the following results were found.

Table of Sample And Bootstrap Sample Estimates

	θ^{Obs}	E_est[θ^{Est}]	SE_est(θ^{Est})	bias_est(θ^{Est})	θ^{BC}
10th Percentile	42	43.8616	11.28367	1.8616	40.1384
90th Percentile	1112	1052.675	189.995	-59.3248	1171.325

(All estimates are in days)

The 95% Bias Corrected Accelerated Bootstrap Confidence Interval for the 10th Percentile is:

CI = [20, 63] days

CI Width = 43 days

The 95% Bias Corrected Accelerated Bootstrap Confidence Interval for the 90th Percentile is:

CI = [727, 1235] days

CI Width = 508 days

Interpretation:

From these values, we found that the average 10th percentile of cancer survival time is 43.861 days with a standard error of 43.861 days. This estimate was also found to have a bias of 1.8616 days meaning our bootstrap estimate is greater than our observed sample 10th percentile of 42 days. Thus, our bootstrap corrected 10th percentile value is 40.1384 days.

Similarly, the average 90th percentile of cancer survival times is 1052.675 days with a standard error of 189.995. This estimate was found to have a bias of -59.3248 days meaning it underestimated our observe sample 90th percentile of 1112 days. Lastly, our bootstrap corrected 90th percentile value is 1171.325 days.

For both percentiles a bootstrap corrected accelerated interval (BCA) was made for both. This interval was used since it tends to perform well when there is skew in the data when compared to the empirical, t-method, and percentile intervals. An empirical bootstrap confidence may have been used for the 90th percentile. However based on the qqplot in the summary, this did not seem appropriate because of the slight skew on the ends and the disjointness in the plots. Thus the BCA interval was preferred.

The BCA interval for the 10th percentile was 20 and 63 days. This means we are 95% confident that the true population 10th percentile of cancer survival time is between 29 and 63 days, with a width of 43 days. Similarly, the BCA interval for the 90th percentile was 727 and 1235 days. This means that we are 95% confident that the true population 90th percentile of cancer survival time is between 727 and 1235 days with a width of 508 days.

Conclusion:

In conclusion, we estimated that the 10th percentile of cancer survival times is between 29 and 63 days while the 90th percentile is between 727 and 1235 days. This tells us that it is somewhat rare for someone to be diagnosed with cancer and only survive between 2-3 months. It is more realistic that people are in the in the later and survive for a couple of years after being diagnosed

with cancer. This data tells us that the time of survival for people with cancer is varied widely since the and there isn't a general number we can give to all cancer patients. However, if we were given more information such as the type of cancer, stage, age of the patient, etc, we may be able to make more meaningful and precise calculations and estimates. Nonetheless, these estimates can still be used as a general length of time to patients who have been diagnosed with cancer and can still give them some peace of mind knowing that most cancer patients live for more than a few months.