**Capstone Project I – Inferential Statistics**

**Data Mining and Data Wrangling:**

Data is obtained from United States Cancer Statistics, 1999-2013 provided at the Centers for Disease Control and Prevention.

<https://wonder.cdc.gov/CancerMort-v2013.html>

Data was downloaded from the above website into 60 text files in which 30 files were downloaded for females and 30 files for males, each having same feature variables like cancer sites, states, year, age, death, crude rate, population and race. These 60 files were concatenated after adding new variable name ‘sex’ and assigning female and male values to the corresponding files.

Subsequently, data wrangling was performed to deal with redundant columns likes State code, notes, etc. These columns were deleted. Also rows with Nans/Null values were deleted. Some rows had ‘Not Applicable’ or ‘Suppressed’ values for Death variable these rows were also deleted. Numeric features were checked for datatypes and were converted into numeric datatype for further calculations.

**Inferential Statistics on data:**

Inferential statistics were used to answer various questions related to the data:

***Is there a significant difference in deaths due to cancer between males and females?*** To answer this question male and female datasets were formed from original data. These datasets were explored for variable deaths to see the distribution using histogram and checked for normality using probability plots. The data was not normally distributed, so log transformations of the data was performed, however even after log transformation the data was not normally distributed. Therefore, Mann Whitney U test was performed to test for significance between deaths in both datasets. It was found that there is significant difference in deaths due to cancer between males and females.

***Is there a significant effect of race on deaths due to cancer? Do whites have significantly higher death than other races?*** To answer this datasets was obtained for each race from the original data. The crude rate were plotted with deaths for each race. It was seen that there was correlation for crude rate and race for each race. Also various statistics were obtained for each race. Further the data was divided into white race Vs other races. And t-test was performed to obtain the significance in difference. It showed that white population had significantly higher death as compared to other populations.

***To explore the affect of age group on death due to cancer? Do people of age 50 and above have significantly higher death due to cancer?*** To answer this, datasets was obtained for each age group from the original data. Mean deaths were plotted for each age group and it was seen that with increase in age there is significant increase in the death due to cancer. Further the data was divided into people of age 50 and above Vs other age groups. And t-test was performed to obtain the significance in difference. It showed that people with age 50 and above had significantly higher death as compared to other age groups.

***To explore if cancer site affects the death due to cancer? And if there is significant difference in deaths due to respiratory system cancer vs death due to digestive system cancer?*** To answer this, datasets was obtained for each cancer site from the original data. Mean deaths were plotted for each cancer site and it was found that Digestive system cancer and respiratory system cancer resulted in most deaths due to cancer. Further the data was divided into deaths due to respiratory system cancer and digestive system cancer. And t-test was performed to obtain the significance in difference. It showed that there was significant difference in deaths due to Digestive system cancer and respiratory system cancer.

***To explore the distribution of death due to cancer in various states of US ? Do people of California have significantly higher death due to cancer?*** To answer this, datasets was obtained for each State from the original data. Mean deaths were plotted for each State and it was found that California resulted in most deaths due to cancer. Further the data was divided into deaths in California and other states. And t-test was performed to obtain the significance in difference. It showed that there was significantly higher deaths in California as compared to other states.

***To further explore the distribution of death due to cancer in California ? And if there is significantly higher deaths due to digestive system cancer than due to respiratory system cancer in California?*** To answer this, datasets was obtained for only California and mean death was calculated for each Cancer site in California. Mean deaths were plotted for each cancer site and it was found that Digestive system cancer and respiratory system cancer resulted in most deaths due to cancer in California as well. Further the data for California was divided into deaths due to respiratory system cancer and digestive system cancer. And t-test was performed to obtain the significance in difference. It showed that in California deaths due to Digestive system cancer were not significantly higher than deaths due to respiratory system cancer.

**Conclusion:**

In California the deaths due to digestive cancer is not significantly higher than deaths due to respiratory system. There is no significant different in the deaths due to two cancers in California. However, when considering the whole USA deaths due to res system was significantly higher than that due to dig system. This suggests that different features (independent variables) like states, cancer sites and others might interact with each other to influence the deaths due to cancer, hence it is important to find out which of these features are most important factors to consider when trying to predict the deaths due to cancer. Also, this suggests that we should use the algorithms which are least affected by multi-collinearity to make reliable predictions.