PROJECT TITLE HERE

Jasen Zhang, Alex Zhu, Christian Pascual

Introduction

Exploratory Data Analysis

Our data is an aggregation of multiple sources including the American Community Survey, the Census, clinical trials and from cancer databases. The observational unit is a state county, and each observation contains various demographic, economic and educational factors. Our outcome is **cancer mortality rate**, measured as number of deaths per capita (100K people).

The data contains information on 3047 counties in the United States. The data is not fully descriptive of all counties in the US because 94 are missing. We surmise these counties are missing due to lack of data for them.

The average mortality per capita is 178.66 across the dataset. Cancer mortality has a nice bell shape, but there are a few outliers present.
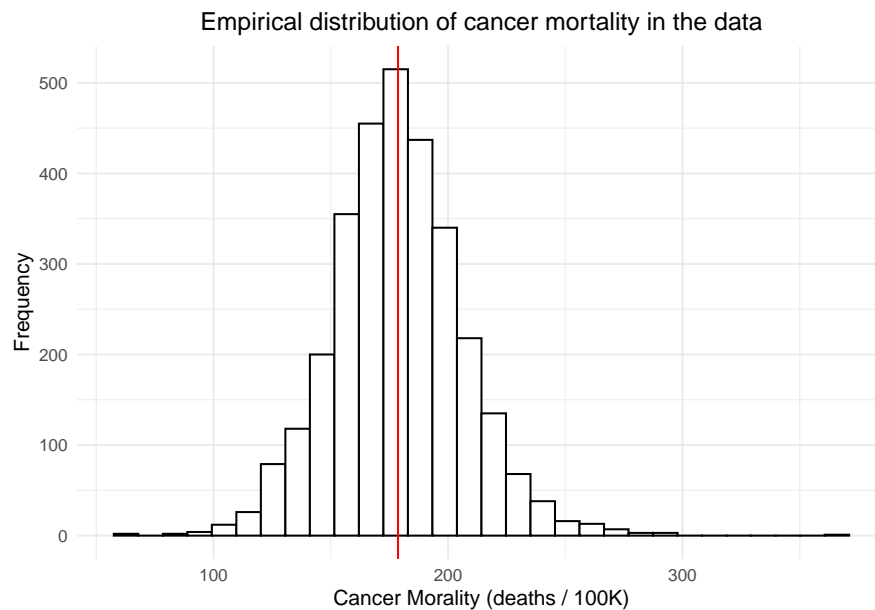
We explored how other variables in the dataset correlated with the cancer mortality in the counties. The data also includes information on the incidence of cancer in each county, so we checked how it related to mortality. As expected, higher incidence correlates highly with higher mortality, so this should be accounted for in a model. There was no apparent trend between cancer mortality per capita and either age or education level.

We found a relationship between insurance type and cancer mortality. Counties where a higher proportion of the population has a private insurance had lower cancer mortality. This trend was reversed for counties with a high proportion of public insurance. Race also seemed to have a relationship with cancer mortality. Counties with a higher proportion of African-Americans experienced higher rates of mortality, and the reverse trend was observed in white Americans. No obvious trend was seen in Asian-Americans or races defined as "other".

Cancer mortality by percentage of race in each county