Statistical Analysis for a Life Insurance Company

John Wensink

MIS450 - Data Mining

Colorado State University-Global Campus

Dr. Steven Chen

6/14/2020

Statistical Analysis for a Life Insurance Company

For this project, our organization has decided to analyze data from a recently conducted heart health study in order to set premiums for high-risk individuals. We will use several statistical analysis techniques to get a better understanding of the information contained in this data set. For this study, we will use the Statistical Analysis Software (SAS) University Edition Studio, which comes pre-programmed with Procedure (PROC) tasks including data exploration, summary statistics, distribution analysis, table analysis, and a substantial collection of other tools available for statistical analysis.

**SAS Heart Library**

A preliminary look at the data contained within the SASHELP.HEART library database, we see instances for status, deathcause, agechddiag, sex, ageatstart, height, weight, diastolic, systolic, mrw, smoking, ageatdeath, and cholestatus. The data types used are a combination of numerical, categorical, binary, and sequential data. For the purposes of performing actuarial analysis on premium prices for life insurance policies, the columns ageatdeath and deathcause will likely be used for our class/response variables, and the columns agechddiag, sex, height, weight, systolic, diastolic, mrw, smoking, cholesterol, and chol\_status will be acting as predictor variables. Fechtner (2018), while conducting similar research has taken the time to include an extra data step in order to preprocess the tables to customize her organization’s definition of a smoker, as well as simplify the CHD diagnosis:

data heart;

set sashelp.heart(where=(Cholesterol ne . and Weight\_Status ne ""));

if Smoking <= 5 then Smoker=0; /\*Create variable ‘Smoker’\*/

else if Smoking > 5 then Smoker=1;

if AgeCHDdiag = . then CHD=0; /\*Create variable ‘CHD’ (1=Yes, 0=No)\*/

else if AgeCHDdiag ne . then CHD=1;

keep Sex AgeAtStart Diastolic Systolic Chol\_Status BP\_Status Weight\_Status

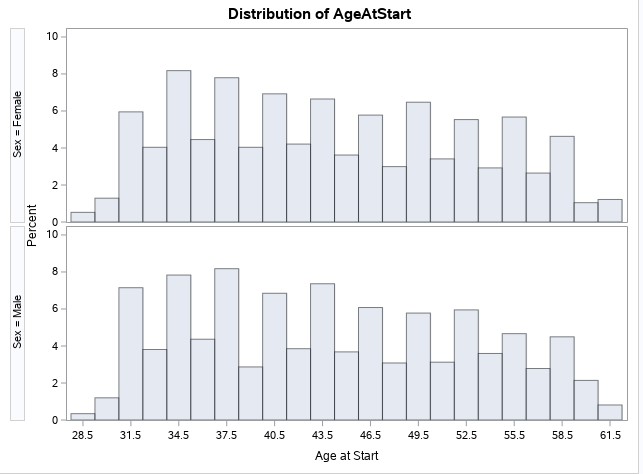
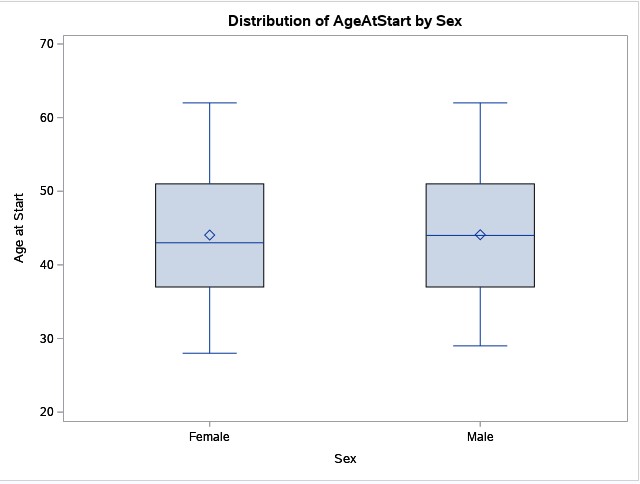
Smoker CHD; /\*Keep only variables which are needed for the analysis\*/

run;

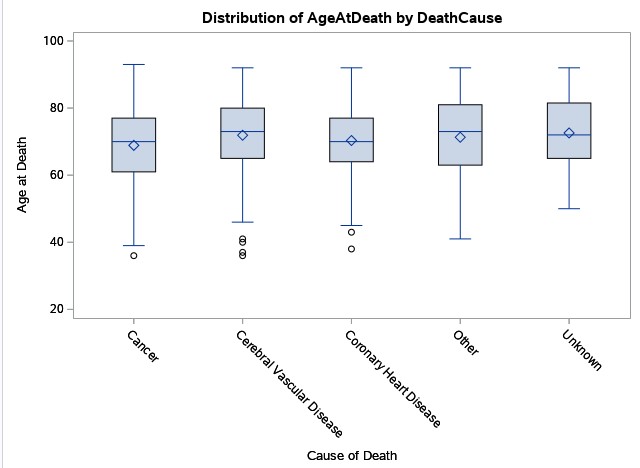
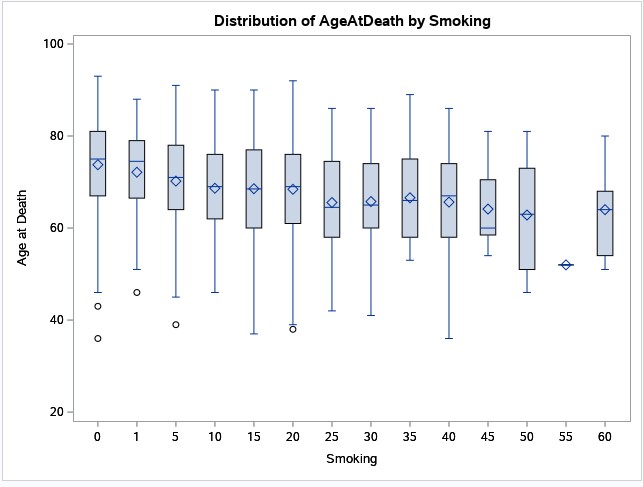
(Fechtner, 2018).

Although we will not be modifying the smoking threshold for our research as well as keeping all the columns in our data set, we will find it useful to simplify the coronary heart disease column using the researcher’s suggestion.

**Data Exploration**

One of the tasks included in SAS University Edition is from the SAS/STAT series called Data Exploration. The SAS documentation for data exploration states that it creates graphical views to examine relationships between variables (SAS, 2020). Taking a look at our dataset we will start by comparing the ages and genders of the sample group: 

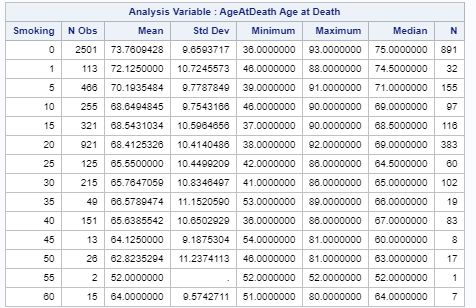
From these charts, we can see that the distribution of men and women is quite similar with some minor differences where the mean age of women is slightly higher than that of their male counterparts, and the age range extends a bit lower for females as well. Using the same tool it is interesting to examine the relationship ageatdeath has with smoking and deathcause:



From these box plots, we can see that no heavy smokers in this study lived to see 90 years old, and the median age for all causes of death seems to be in the early ‘70s with the second quartile beginning in the early ’60s. It may make sense for our organization to start collecting more premiums for individuals who are smokers as well as individuals who are advancing in age. It is my recommendation that we start collecting additional surcharges around age 60 increasing slightly until the policyholder’s death.

**Summary Statistics**

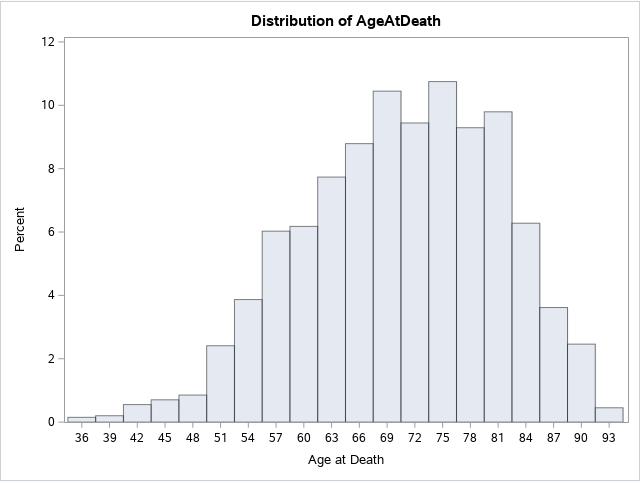
The SAS documentation describes Summary Statistics tasks as one which “provides descriptive statistics for variables across all observations and within groups of observations. You can also summarize your data in a graphical display, such as histograms and box plots.” (SAS, 2020). We will use the summary statistics tool to create tables and charts that describe the relationship between smoking and ageatdeath:



From this table, we can see that as people smoke at a higher rate, their mean age at death drops off substantially. It would be my recommendation to charge smokers a hefty surcharge on their premiums regardless of quantity. It seems that the data shows that smokers are making a harmful choice that substantially increases their likelihood of early death.

**Distribution Analysis**

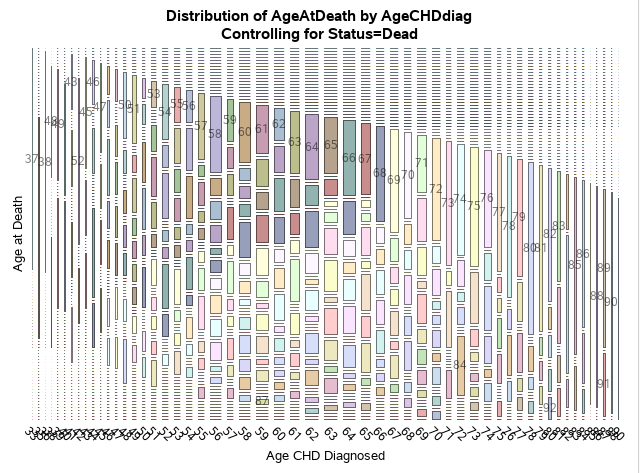
The SAS documentation describes this task as a tool to create different sorts of plots like histograms and probability plots representing the distribution of numeric variables (SAS, 2020). For our purposes, we will examine a histogram representing the distribution of ageatdeath:



This chart gives our organization additional information on the distribution of age at death and perhaps forces us to reconsider our previous observation that the age in which premium surcharges should start. This chart shows that in the early ‘50s likelihood of death for these high-risk individuals really begins to grow.

**Table Analysis**

The SAS documentation describes the Table Analysis task as a tool that “provides one-way to n-way frequency and contingency (crosstabulation) tables. This task also generates statistics about the association between rows and columns” (SAS, 2020). For our purposes, we will be using the Table Analysis task to examine the relationship between age at death and age CHD diagnosed:



This is an interesting plot which appears to show that the earlier one gets diagnosed with CHD, the higher their risk for early death. It would be my recommendation to start charging a premium surcharge upon diagnosis of CHD in accordance with all applicable laws.

**Conclusion**

The role of statistical analysis in actuarial science is profound. Calculating a competitive premium that will not lead the company to insolvency is a very difficult task that requires analysis of huge amounts of data. Running SAS on my computer was surprisingly slow for the limited amount of observations being processed. I would imagine our firm will be needing substantial investment of capital into computational resources whether on-premises or in the cloud. The techniques we explored today are certainly powerful indicators of risk factors affecting the possibility of early death. SAS is full of tools that can be used for this purpose, for example Jiang and Ni (2008) recommend using a survival analysis task for an individual undergoing certain specific risky behaviors and underlying health conditions. It seems a skilled statistical analyst will be only limited by his creativity and processing power budget in pursuit of information regarding relationships between healthcare datasets.

References

Fechtner, S. (2018). The Propensity Score Matching. *PhUSE EU Connect*. Retrieved from <https://www.lexjansen.com/phuse/2018/rw/RW03.pdf>

Jiang, X., & Ni, J. (2008). Statistical Analysis and Empirical Study for Life Insurance. Retrieved June 14, 2020, from <https://pdfs.semanticscholar.org/e5d5/a914b9b38323e80e35931548e6516884bb98.pdf>

SAS. (2020). Data Exploration Task. Retrieved June 14, 2020, from <https://documentation.sas.com/?docsetId=webeditorref&docsetTarget=n0s0baurvln78qn1vowygfyvv9aw.htm&docsetVersion=5.1&locale=en>

SAS. (2020). Summary Statistics Task. Retrieved June 14, 2020, from <https://documentation.sas.com/?docsetId=webeditorref&docsetVersion=5.1&docsetTarget=p0aheyslqgh8sxn1lfe9pj883r1m.htm&locale=en>

SAS. (2020). Table Analysis Task. Retrieved June 14, 2020, from <https://documentation.sas.com/?docsetId=webeditorref&docsetVersion=5.1&docsetTarget=n01urfic5gmsoen17pknic4daa6k.htm&locale=en>

SAS. (2020). Distribution Analysis Task. Retrieved June 14, 2020, from <https://documentation.sas.com/?docsetId=webeditorref&docsetVersion=5.1&docsetTarget=n0uiqaj7iu2zdkn13erg3gl4u83t.htm&locale=en>