**Emily Lynch**

**Public Health 223**

**Data Assignment 1**

**1. How many observational units (cases) are in this dataset? (1 point)**

1,000 cases

Found with: total id

**2. How many variables are in this dataset? Identify which variables in the dataset are categorical and which are quantitative. (4 points)**

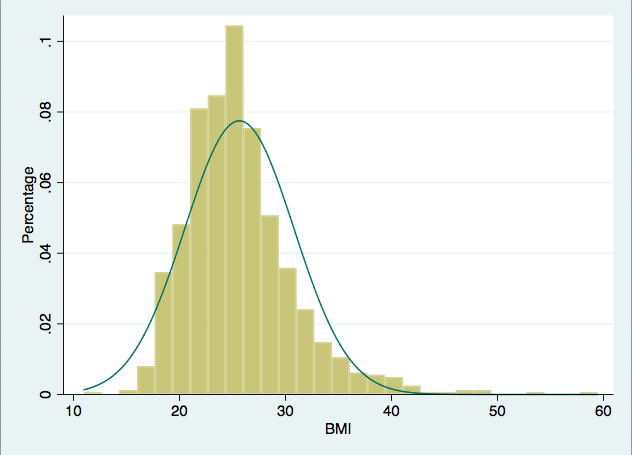
There are 9 variables in this dataset. Smoker, Hospitalized, Race, and Asthma are all categorical variables. ID Number, Age, BMI, Weight, and Systolic Blood Pressure are all quantitative variables.

**3. For BMI:**

* **a. Describe numerically the center and spread. (4 points)**

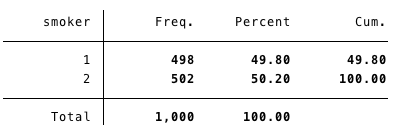
The center for BMI for this dataset is 25.64937. The spread if using range (max-min=spread) is 48.46, if using IQR (3rd Quartile-1st Quartile=spread) is 31.39, or if using standard deviation is 5.147525.

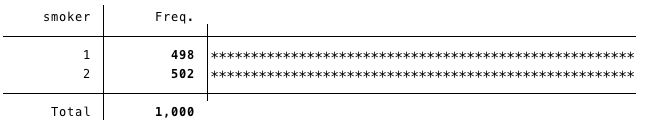
Found with: summarize bmi, detail

* **b. Create a graphical summary to describe these data. Make sure you give a title and label the axes of your graph. (4 points)**
* Found with: histogram bmi, normal ytitle(Percentage) xtitle(BMI) (bin=29, start=10.98, width=1.6710345)
* **c. Write 2-3 sentences describing the data and the distribution shape. (4 points)**

The distribution of the data is right/positively skewed because the right tail is longer than the left tail. It is single-peaked, therefore it has no outliers. Additionally, this peak occurs around 25 BMI.

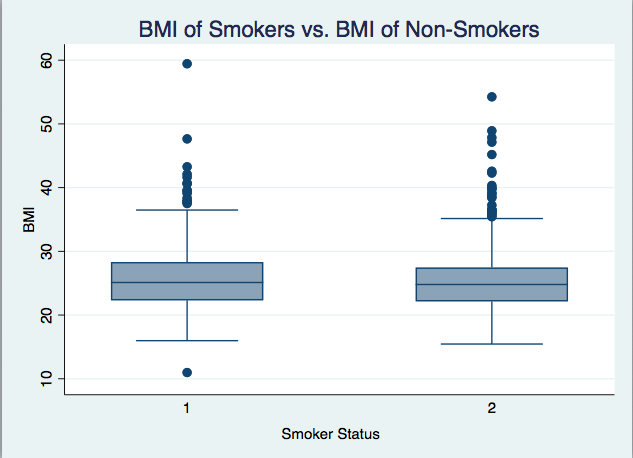


**4. Provide a frequency table for smoking status. Label the output clearly so it can be interpreted. (2 points)**

Found with: tabulate smoker

* Found with: tabulate smoker, plot

**5. To assess how smoking may be associated with BMI, summarize BMI for each strata of smoking status. (Hint: do 3a and 3b above among smokers and then repeat among non- smokers; label the output clearly so it can be interpreted) (4 points)**

* Found with: graph box bmi, over(smoker) ytitle(BMI) title(BMI of Smokers vs. BMI of Non-Smokers) caption(Smoker Status) 

**6. Do there appear to be any differences in BMI by smoking status? Write 2-3 sentences to summarize your findings. (2 points)**

Based on the box graph generated above with smoking status 1 being a current smoker and smoking status 2 being not a current smoker there appears to be a wider range of BMI’s recorded in the smoker group than the non-smoker group. Additionally, in contrast the non-smoker group appears to be more evenly distributed amongst varying BMI’s where the smokers seem more clumped together with the exception of a few outliers.