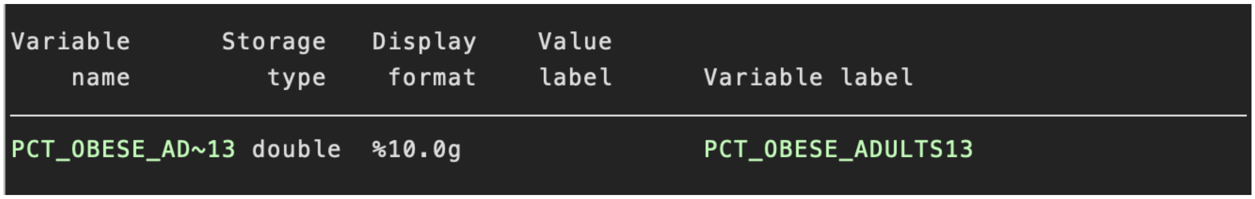
Duc Thanh Nguyen - Boston University

Today, I will be examining data and implementing normality checks in Stata: using 2013 data from the US Census on the county-level obesity rates in the US.

a. Use the command describe to learn about the type of the variable PCT_OBESE_ADULTS13

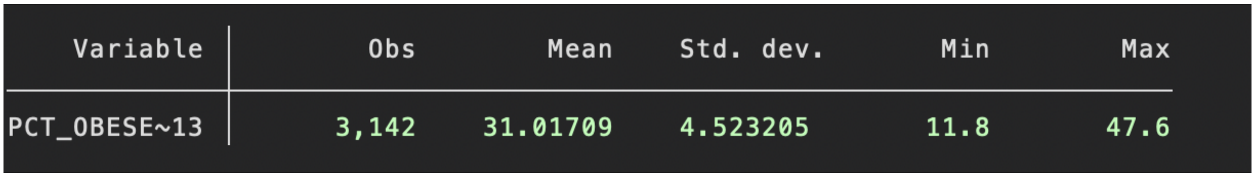
- describe PCT_OBESE_ADULTS13



```
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------
PCT_OBESE_AD~13 double   %10.0g               PCT_OBESE_ADULTS13
```

b. Create a table summary of descriptive statistics for variable PCT_OBESE_ADULTS13 using the summarize command.

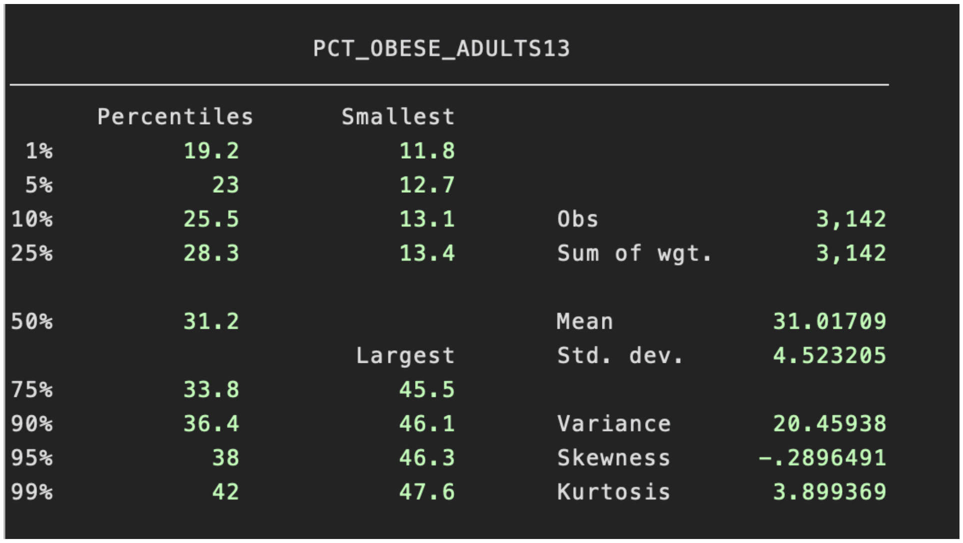
- sum PCT_OBESE_ADULTS13



```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
PCT_OBESE~13 |      3,142    31.01709    4.523205       11.8       47.6
```

c. Utilize the following two commands to further examine the distribution of the obesity rates across the US counties:

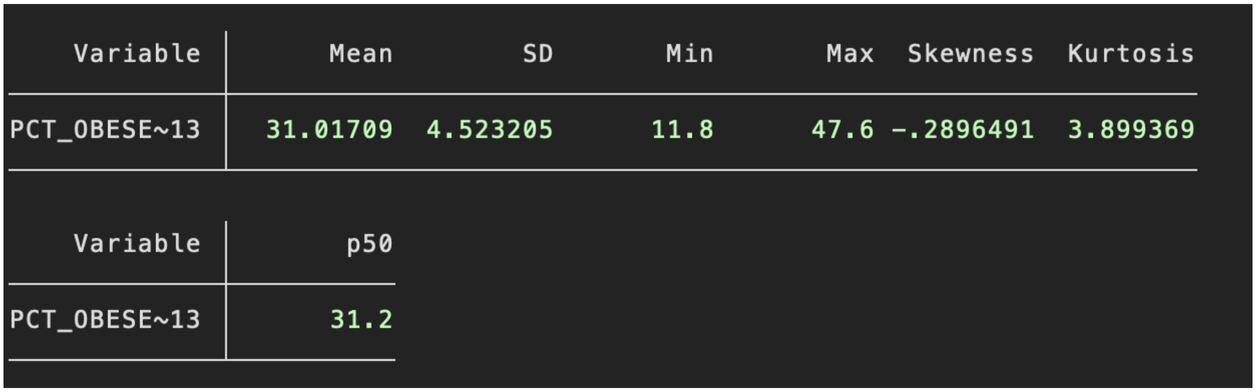
- sum PCT_OBESE_ADULTS13, d



```
                     PCT_OBESE_ADULTS13
-------------------------------------------------------------
      Percentiles      Smallest
 1%         19.2           11.8
 5%           23           12.7
10%         25.5           13.1       Obs               3,142
25%         28.3           13.4       Sum of wgt.       3,142

50%         31.2                      Mean           31.01709
                        Largest       Std. dev.      4.523205
75%         33.8           45.5
90%         36.4           46.1       Variance       20.45938
95%           38           46.3       Skewness      -.2896491
99%           42           47.6       Kurtosis       3.899369
```

- tabstat PCT_OBESE_ADULTS13, stat(mean sd min max sk k med)

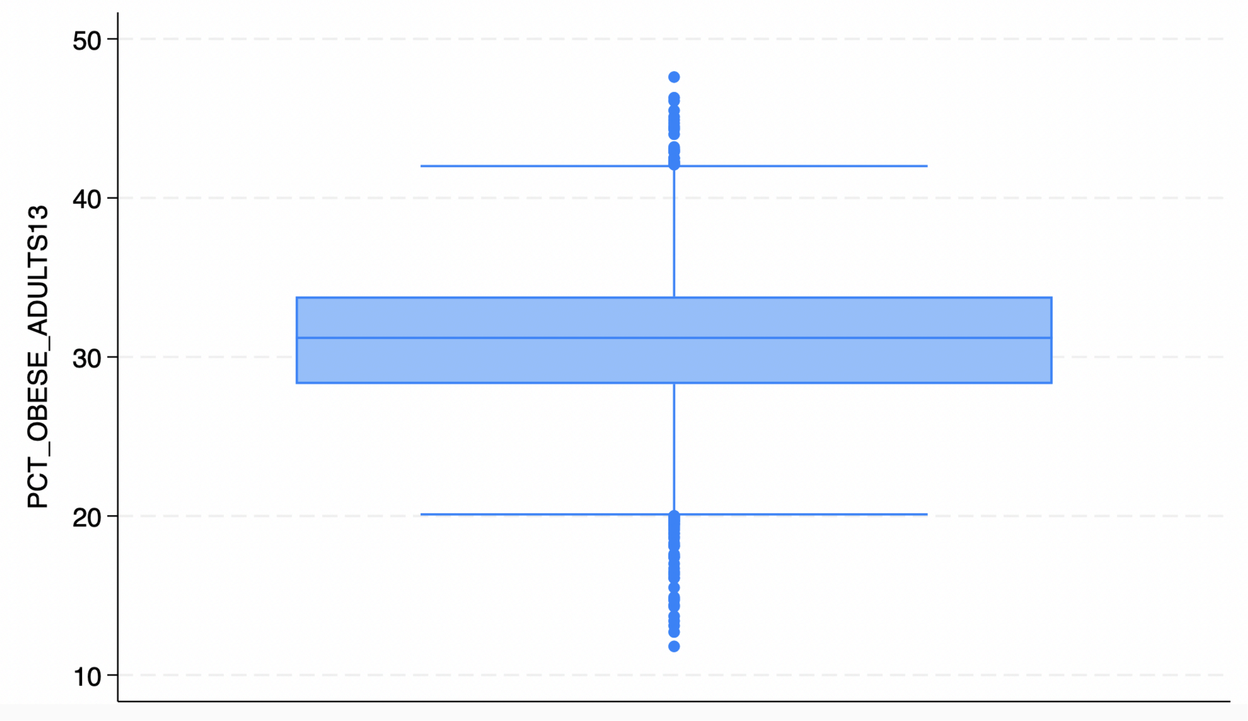


```
    Variable |      Mean        SD       Min       Max  Skewness  Kurtosis
-------------+------------------------------------------------------------
PCT_OBESE~13 |  31.01709  4.523205      11.8      47.6 -.2896491  3.899369
-------------+------------------------------------------------------------

    Variable |       p50
-------------+----------
PCT_OBESE~13 |      31.2
-------------+----------
```

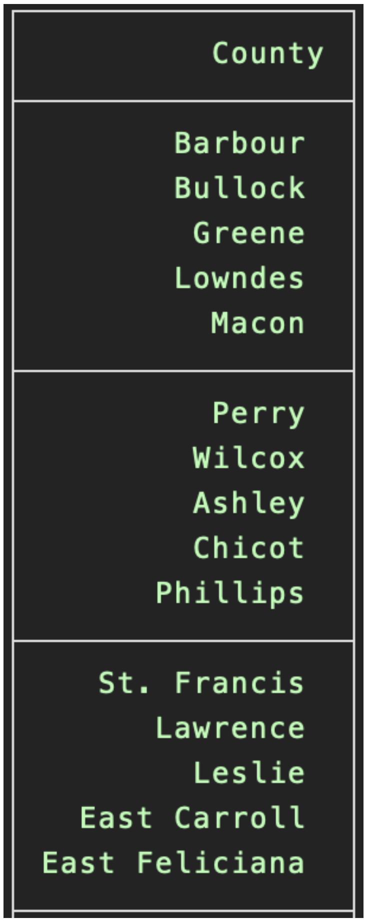
d. Construct a boxplot for the PCT_OBESE_ADULTS13 variable using the graph box command:

graph box PCT_OBESE_ADULTS13



e. The ERS is interested in how many counties have obesity rates above the 99th percentile. From the output in part c above we know that the 99th percentile of the obesity rates across 3142 US counties was 42% in 2013. Using the list command in Stata, we can provide the ERS with the list of counties with obesity rates above the 99th percentile:

list County if PCT_OBESE_ADULTS13 > 42

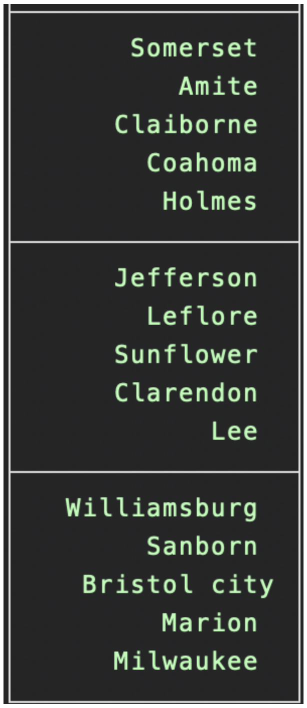


```
          County

         Barbour
         Bullock
          Greene
         Lowndes
           Macon

           Perry
          Wilcox
          Ashley
          Chicot
        Phillips

      St. Francis
        Lawrence
          Leslie
     East Carroll
   East Feliciana
```



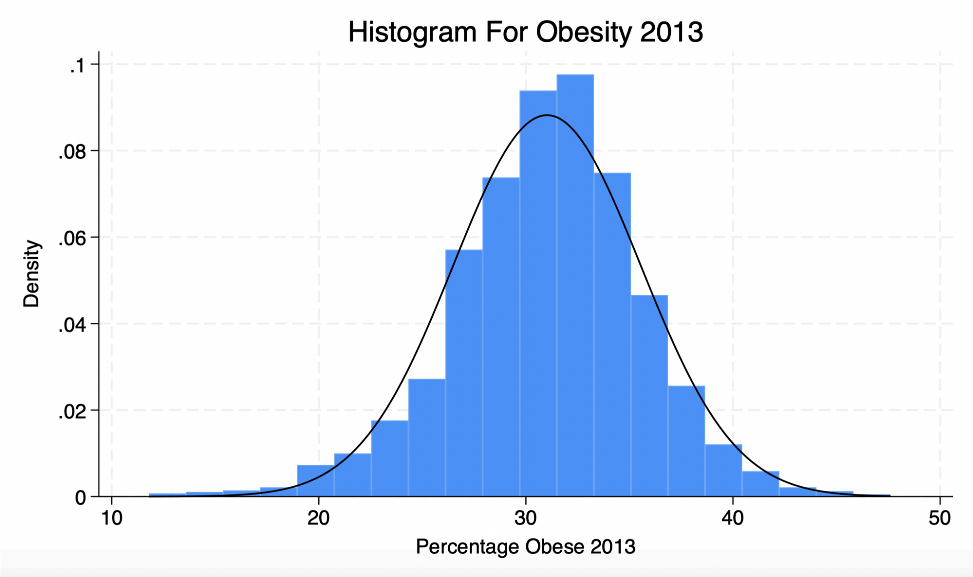
```
      Somerset
         Amite
     Claiborne
       Coahoma
        Holmes

     Jefferson
       Leflore
     Sunflower
     Clarendon
           Lee

  Williamsburg
       Sanborn
  Bristol city
        Marion
     Milwaukee
```

f. Using Stata, create a histogram for the PCT_OBESE_ADULTS13 variable with the superimposed normal curve.

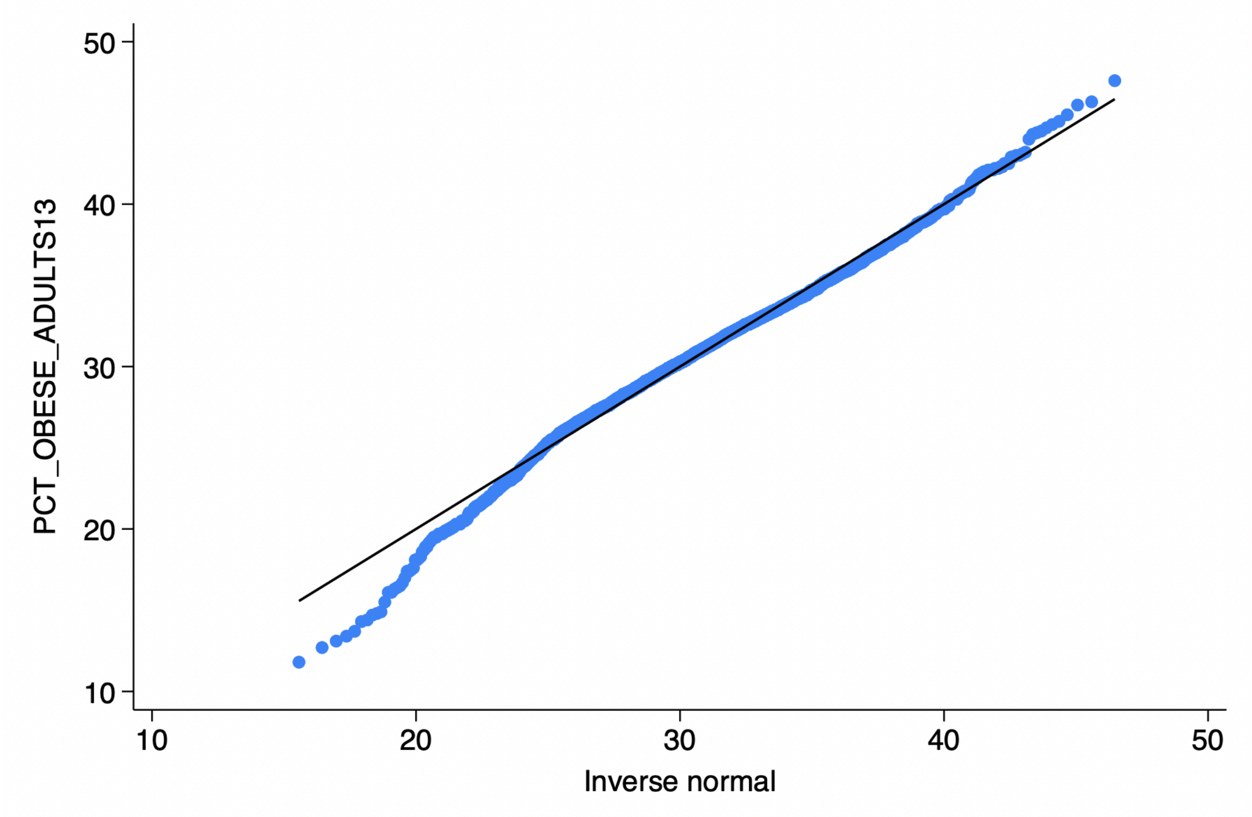
- histogram PCT_OBESE_ADULTS13, bin(20) normal xtitle("Percentage Obese 2013") title("Histogram For Obesity 2013")



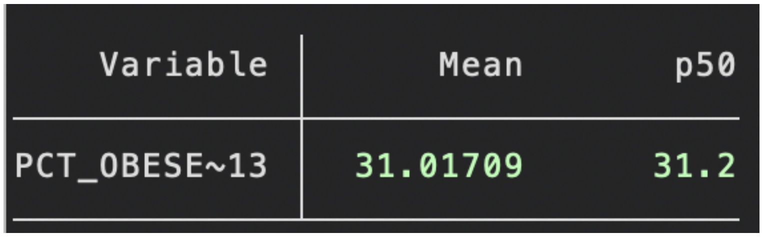
I can qualitatively say that the distribution of obesity percentages in 2013 seems to approximate a normal distribution fairly well. The shape is roughly bell-shaped, though there's a slight skew to the left, indicating that while most of the data points are concentrated in the middle (around the 30% obesity range), there are fewer areas with much lower or much higher percentages.

The smooth curve suggests the data might follow a normal distribution, but the leftward skew hints that there may be a small bias toward lower obesity percentages. Overall, the distribution doesn't look perfectly normal, but it is fairly close.

g.



h.
tabstat PCT_OBESE_ADULTS13, stat(mean med)



```
    Variable |      Mean       p50
-------------+--------------------
PCT_OBESE~13 |  31.01709      31.2
----------------------------------
```

Since the normal probability density function is symmetric, we would expect that for a variable with a distribution close to normal, the sample mean and median should be nearly the same. This is because in a perfectly normal distribution, the mean and median coincide at the center of the distribution, reflecting symmetry. If the distribution is close to normal, the sample mean and median should not differ significantly, as both would be located near the peak of the bell curve. However, any noticeable differences between them might indicate slight skewness or departures from normality. In the case of the obesity distribution from 2013, the slight leftward skew observed in the histogram might cause a small difference between the sample mean and median.