PH 538: Biostatistical Methods I

STATA: Lab 1 (Descriptive Statistics)

Dr. Fares Qeadan

fqeadan@salud.unm.edu

Objectives:

In this lab students will learn how to use STATA to describe numerical (quantitative) and categorical (qualitative) variables both numerically and graphically.

Background on the data set:

In this Lab, we will be using the *Diabetes and obesity, cardiovascular risk factors* data set. This data set includes 403 African Americans who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia.

The list of variables in the data set:

	Variable	**Description**
1	id	Subject ID
2	chol	Total Cholesterol
3	stab_glu	Stabilized Glucose
4	hdl	High Density Lipoprotein
5	ratio	Cholesterol/HDL Ratio
6	glyhb	Glycosylated Hemoglobin (A1C)
7	location	County - a factor with levels Buckingham and Louisa
8	age	age in years
9	gender	a factor with levels male and female
10	height	height in inches
11	weight	weight in pounds
12	frame	a factor with levels small, medium and large
13	bp_1s	First Systolic Blood Pressure
14	bp_1d	First Diastolic Blood Pressure
15	waist	waist in inches
16	hip	hip in inches
17	diab	Diabetes status

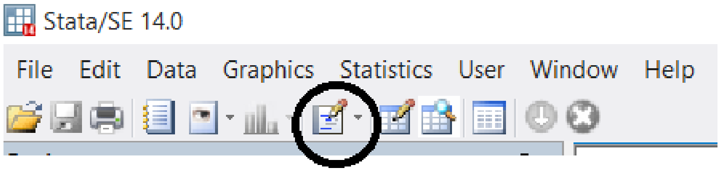
Things to do before starting the analysis of the data

1. Create a do-file:

To reuse your work, you need to save your syntax (STATA commands) into a file. STATA uses do files for this purpose where Do files are simply text files whose names end with .do. There are several ways to create a do file as follows:

a) Type **doedit** in the command line and then a do file editor will pop-up. From the drop down menu of the do file, click on **File** and then select **save as** to save your file under any name and location you like, say **lab1** and save it at the PH538 folder in your computer.

b) Firstly, click the button at the top that looks like a pencil writing in a notebook and then proceed as in option(a) to name and save your do file.



c) From the Menu, click on **Window-> Do-File Editor-> New Do-file Editor** and then proceed as in option(a) to name and save your do file.

Your first command in the do-file should be **clear all** which clears the memory so you don't have to worry about what might have happened before your program was run.

2. Create a log-file:

To record all the commands the do file ran and their results, create a **log file**. There are several ways to create a log file but we will be considering only the way how it's done through the do-file as follows:

```
log using "C:\Users\Fares\Documents\PH538\STATA\lab1\lab1log.log",
text replace
```

Remember to close you log-file after you are done with your work. To do this, end your do-file with the command:

```
log close
```

3. Load the data into STATA:

To load the data you need, use the **use** command as follows:

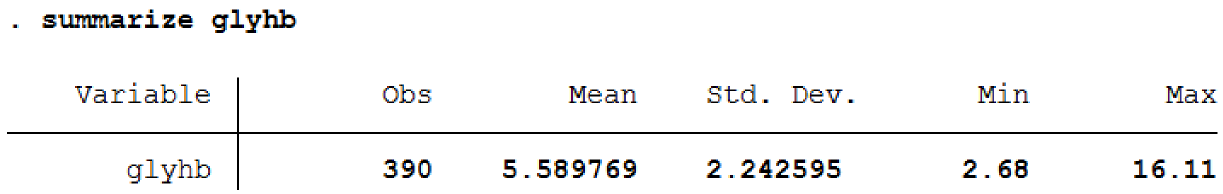
```
use "C:\Users\Fares\Documents\PH538\STATA\lab1\diabetesfall16.dta"
```

Data Analysis

1. Numerical Descriptive Statistics for Numerical (Quantitative) Variables:

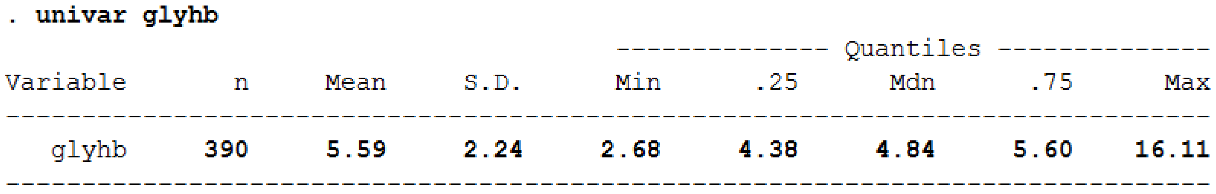
We will describe the Glycosylated Hemoglobin (A1C) variable *[and other variables]* numerically by providing the following sample statistics:

n, Mean, Median, Mode, Standard deviation (or Variance), Q1, Q3, IQR, Min, Max, Range, Mode



```
. summarize glyhb

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       glyhb |       390    5.589769    2.242595       2.68      16.11
```

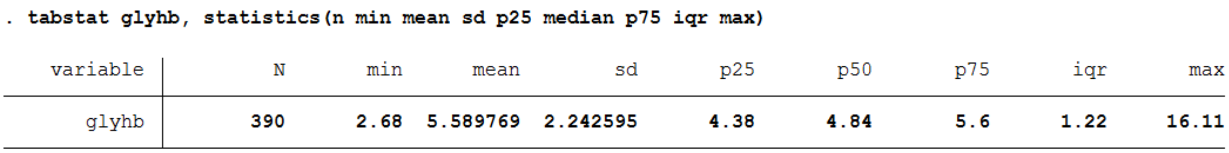
Note that the command `summarize` doesn't provide all summary statistics; it only provides five statistics. So, we should try other commands as follows:



```
. univar glyhb
                                        -------------- Quantiles --------------
Variable       n     Mean     S.D.      Min      .25      Mdn      .75      Max
-------------------------------------------------------------------------------
   glyhb     390     5.59     2.24     2.68     4.38     4.84     5.60    16.11
-------------------------------------------------------------------------------
```

NOTE FROM THE TA: Side note: Univar wouldn't work, so I had to do **findit Univar** and then find the link called "**update to Univar".** Next, I clicked on that link and installed the file. Then **Univar glyhb** worked.

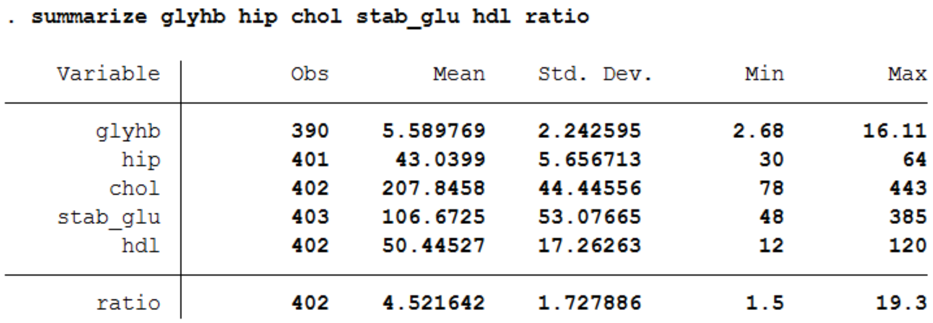
Note that the neither `summarize` nor `univar` provides the mode, Range and IQR statistics. Nonetheless, one could compute the Range and IQR according to Range=Max-Min and IQR=Q3-Q1. So, we should try other commands as follows:



```
. tabstat glyhb, statistics(n min mean sd p25 median p75 iqr max)

    variable |         N       min      mean        sd       p25       p50       p75       iqr       max
-------------+----------------------------------------------------------------------------------------------
       glyhb |       390      2.68  5.589769  2.242595      4.38      4.84       5.6      1.22     16.11
------------------------------------------------------------------------------------------------------------
```

STATA summary statistics commands don't provide the Mode and Range!!!!! See extra credit question in Homework 1.

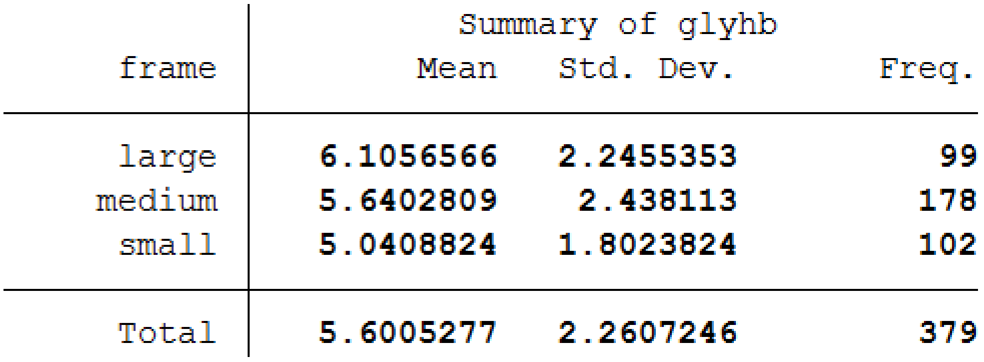
Remark: One could describe more than one variable at a time as follows:



```
. summarize glyhb hip chol stab_glu hdl ratio

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       glyhb |       390    5.589769    2.242595       2.68      16.11
         hip |       401     43.0399    5.656713         30         64
        chol |       402    207.8458    44.44556         78        443
    stab_glu |       403    106.6725    53.07665         48        385
         hdl |       402    50.44527    17.26263         12        120
-------------+--------------------------------------------------------
       ratio |       402    4.521642    1.727886        1.5       19.3
```

One could also describe numerical variables within the levels of categorical variables as follows:

```
. tab frame, summarize(glyhb)
```



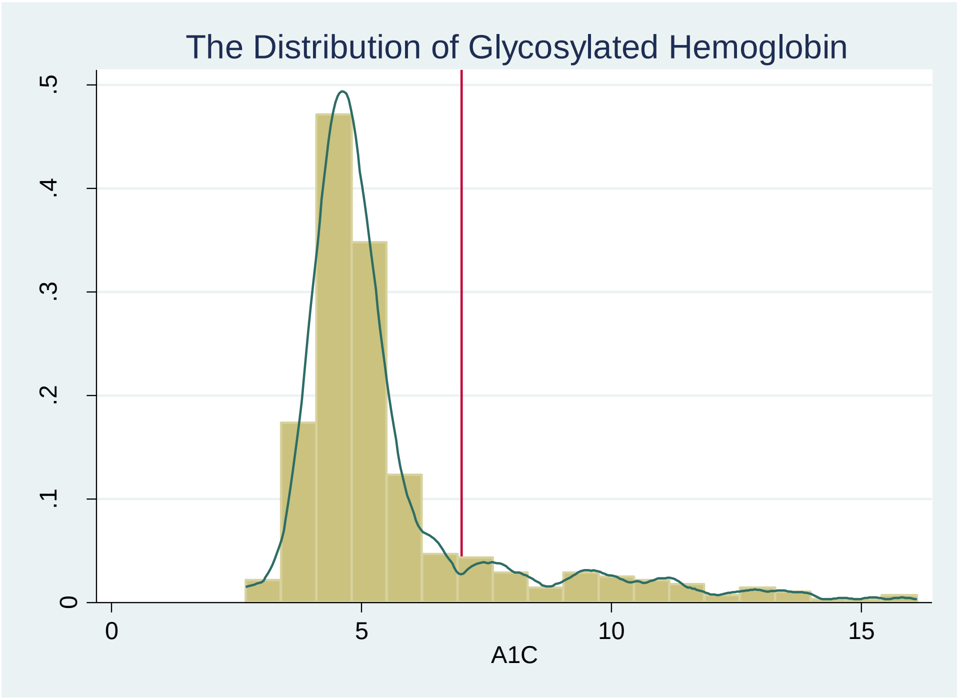
```
            |      Summary of glyhb
      frame |        Mean   Std. Dev.       Freq.
------------+------------------------------------
      large |   6.1056566   2.2455353          99
     medium |   5.6402809    2.438113         178
      small |   5.0408824   1.8023824         102
------------+------------------------------------
      Total |   5.6005277   2.2607246         379
```

2. Graphical Descriptive Statistics for Numerical (Quantitative) Variables:

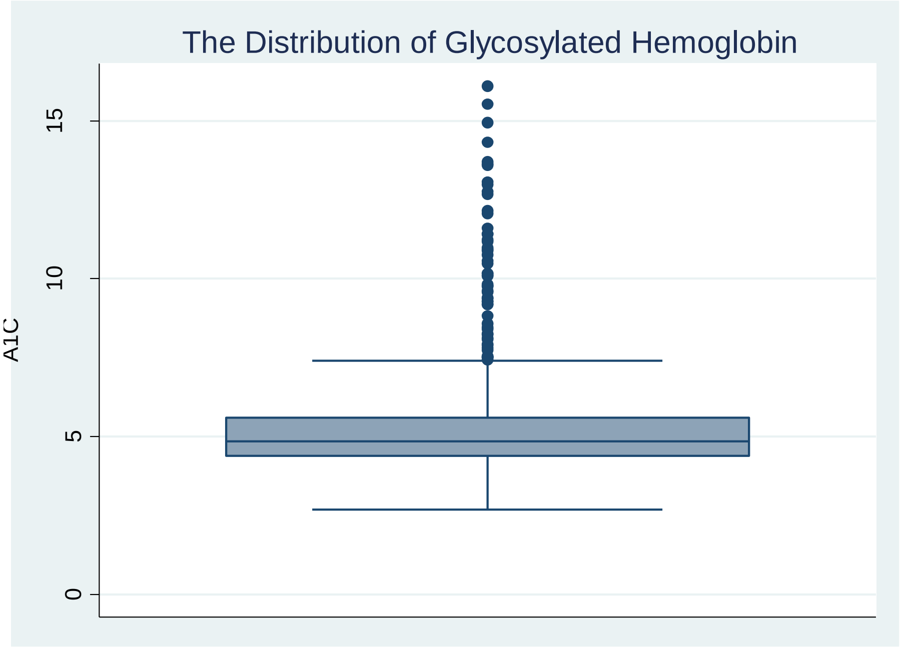
We will describe the Glycosylated Hemoglobin (A1C) variable *[and other variables]* graphically by providing the following presentations:

Histogram, Box-plot, Stem and leaf and Scatter plot.

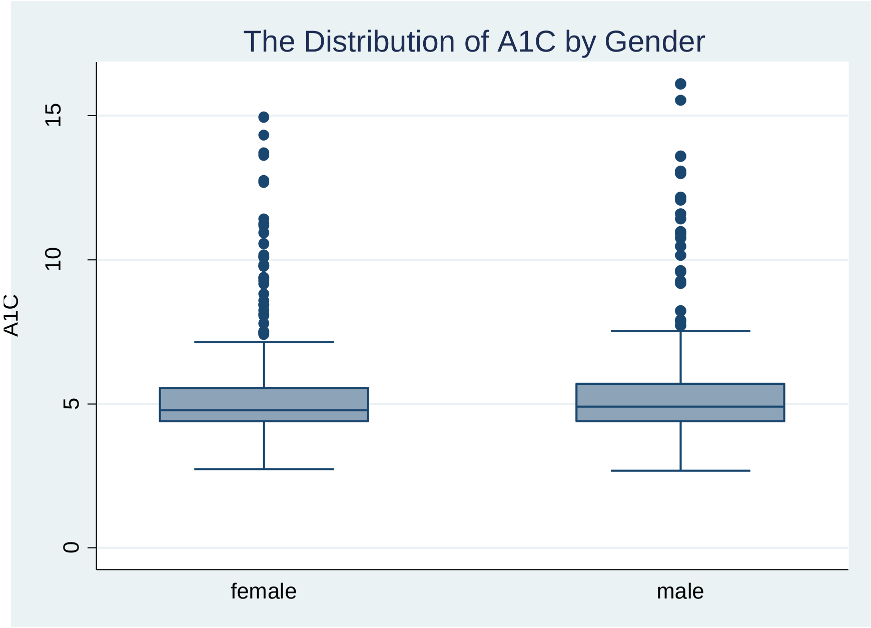
```
. histogram glyhb, kdensity xline(7) xtitle("A1C") title("The Distribution of Glycosylated Hemoglobin")
```

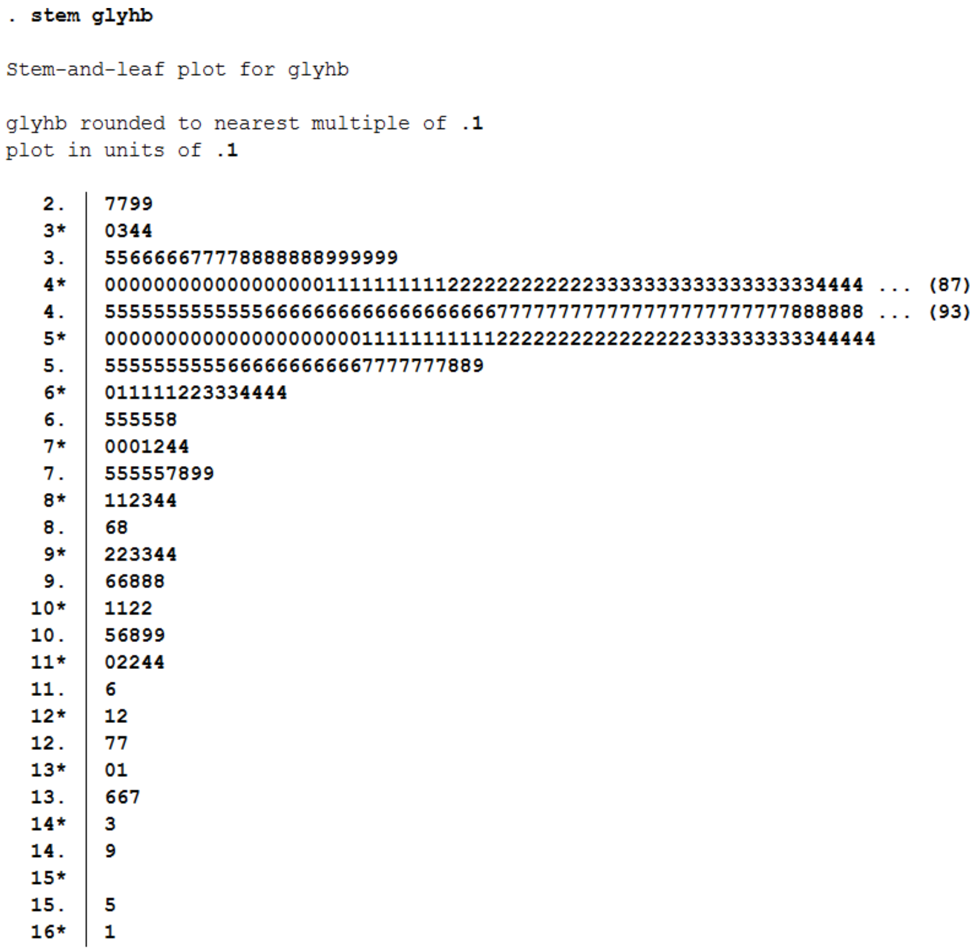


```
. graph box glyhb, ytitle("A1C") title("The Distribution of Glycosylated Hemoglobin")
```



```
. graph box glyhb, over(gender) ytitle("A1C")  title("The Distribution of A1C by Gender")
```

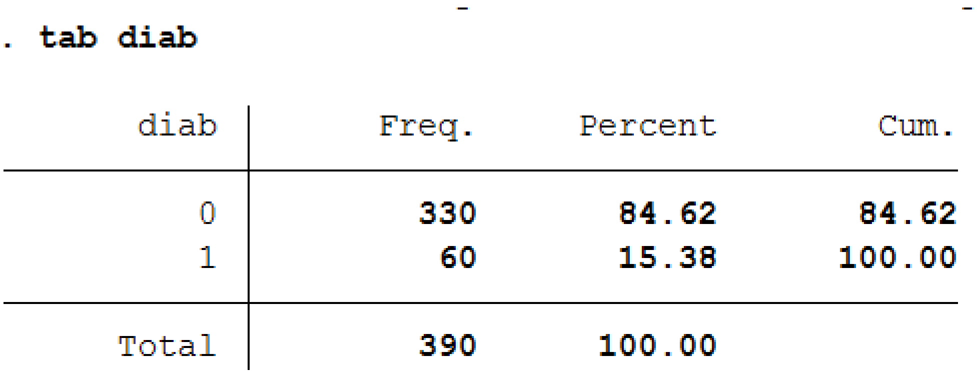




```
. stem glyhb

Stem-and-leaf plot for glyhb

glyhb rounded to nearest multiple of .1
plot in units of .1

   2. | 7799
   3* | 0344
   3. | 556666677778888888999999
   4* | 00000000000000000011111111112222222222223333333333333333334444 ... (87)
   4. | 55555555555556666666666666666666677777777777777777777777888888 ... (93)
   5* | 000000000000000000000111111111112222222222222222333333333344444
   5. | 5555555555666666666677777777889
   6* | 01111223334444
   6. | 555558
   7* | 0001244
   7. | 555557899
   8* | 112344
   8. | 68
   9* | 223344
   9. | 66888
  10* | 1122
  10. | 56899
  11* | 02244
  11. | 6
  12* | 12
  12. | 77
  13* | 01
  13. | 667
  14* | 3
  14. | 9
  15* |
  15. | 5
  16* | 1
```

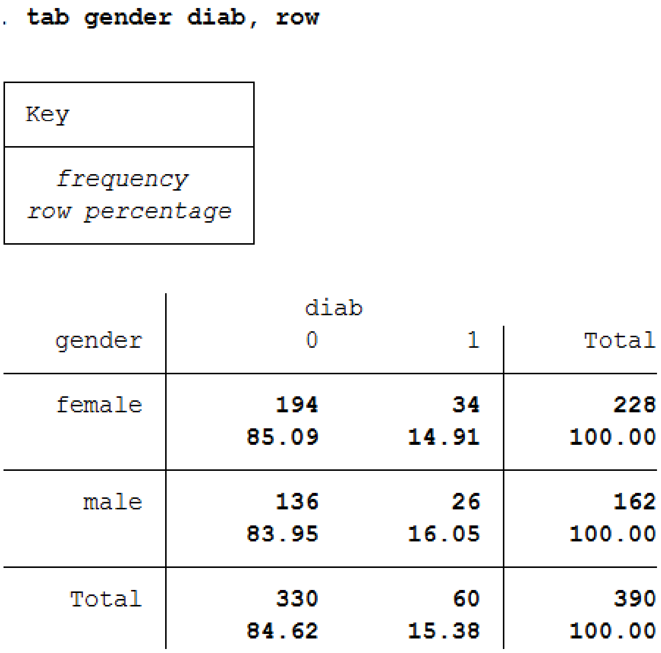
3. **Numerical Descriptive Statistics for Categorical (Qualitative) Variables:**

We will describe the Diabetes status variable *[and other variables]* numerically by providing the frequencies and relative frequencies through contingency tables:



```
. tab diab

       diab |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        330       84.62       84.62
          1 |         60       15.38      100.00
------------+-----------------------------------
      Total |        390      100.00
```

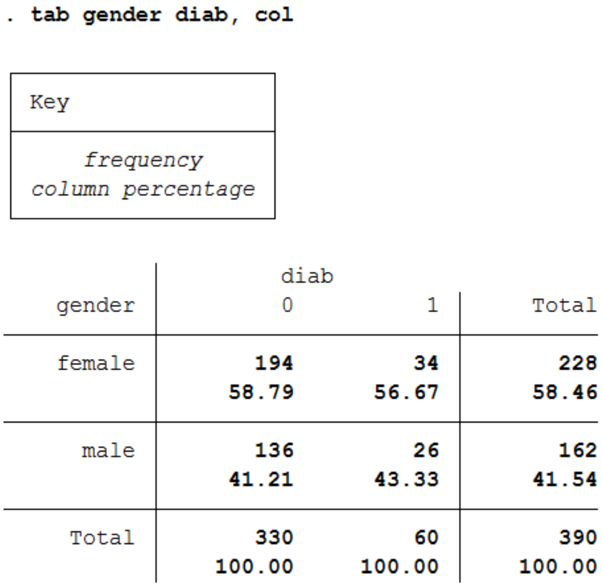
Note that we could also find the sample proportion of diabetes by gender as follows:



```
. tab gender diab, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |         diab
    gender |         0          1 |     Total
-----------+----------------------+----------
    female |       194         34 |       228 
           |     85.09      14.91 |    100.00 
-----------+----------------------+----------
      male |       136         26 |       162 
           |     83.95      16.05 |    100.00 
-----------+----------------------+----------
     Total |       330         60 |       390 
           |     84.62      15.38 |    100.00 
```

How to read the above Table? Here is a correct statement: 14.91% of females were found to have diabetes

Remark: Note that there are three different percentages one could obtain, the total one, the row one and the column one and each one of them has a different denominator and hence different interpretation.



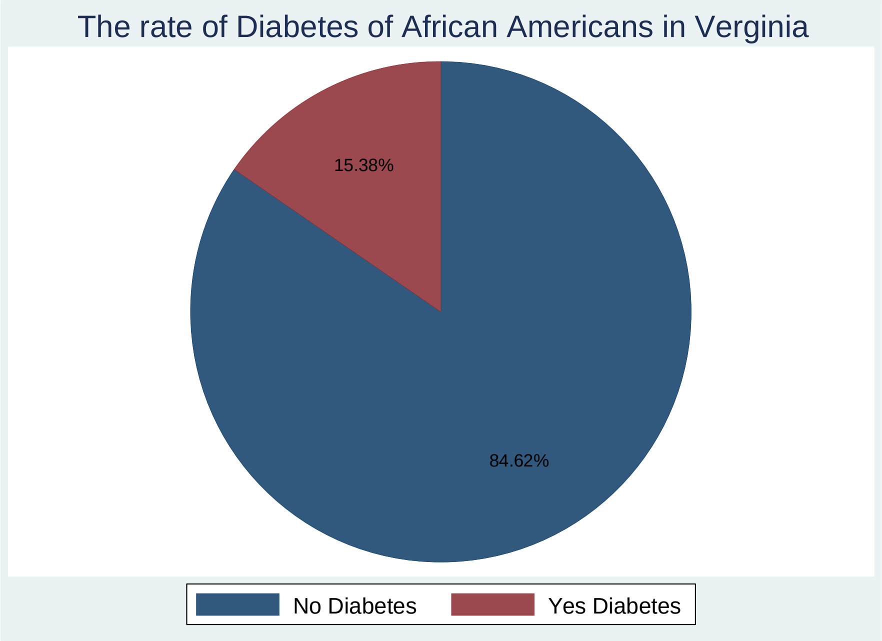
```
. tab gender diab, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |         diab
    gender |         0          1 |     Total
-----------+----------------------+----------
    female |       194         34 |       228 
           |     58.79      56.67 |     58.46 
-----------+----------------------+----------
      male |       136         26 |       162 
           |     41.21      43.33 |     41.54 
-----------+----------------------+----------
     Total |       330         60 |       390 
           |    100.00     100.00 |    100.00 
```

How to read the above Table? Here is a correct statement: 56.67% of subjects with diabetes were female.

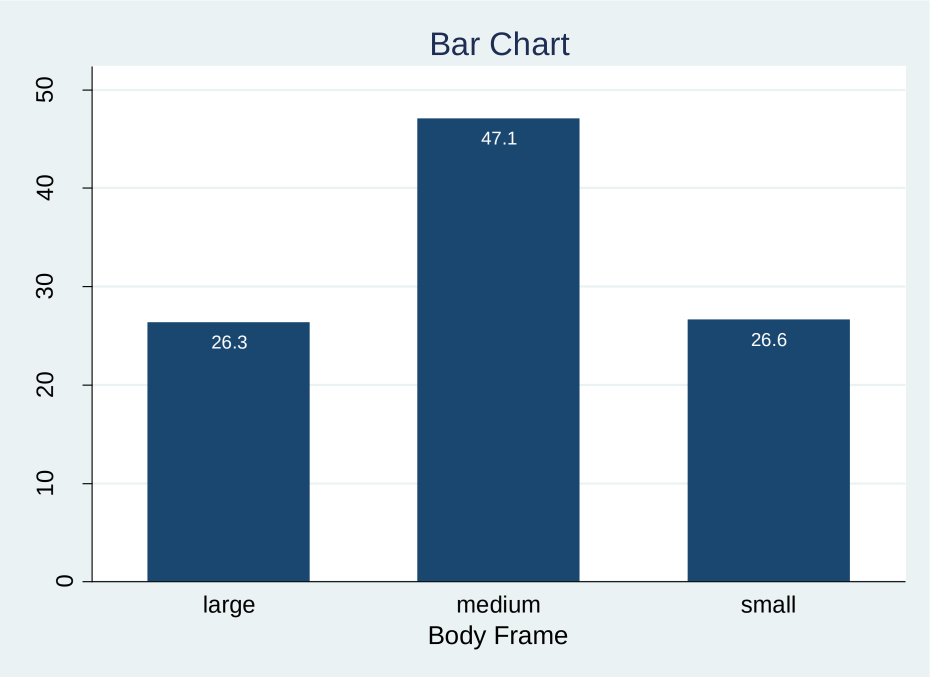
4. Graphical Descriptive Statistics for Categorical (Qualitative) Variables:

We will describe the Diabetes status variable *[and other variables]* graphically by providing the pie and bar charts:

```
. graph pie, over(diab) plabel(_all percent) legend(label(1 "No Diabetes") label(2 "Yes Diabetes")) title("The rate of Diabetes of African Americans in Verginia")
```

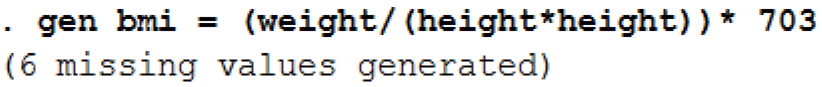


```
. graph bar, over(frame) b1title("Body Frame") title("Bar Chart") blabel(bar, position(inside) format(%9.1f) color(white))
```



Data Management:

1. Please create a BMI variable from the given weight and height variables?



```
. gen bmi = (weight/(height*height))* 703
(6 missing values generated)
```

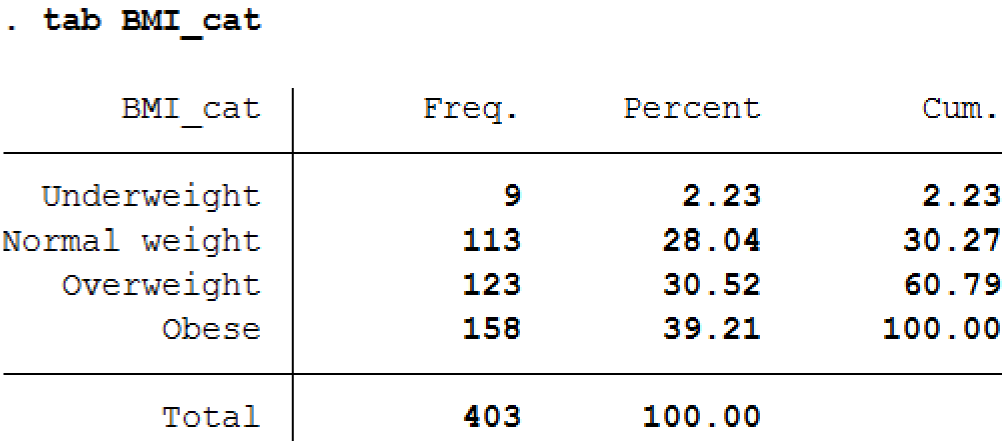
2. Please create a BMI categorical variable from the BMI numeric one? Note that, in public health, BMI for adults is often divided into four categories:
    1. Underweight if BMI<18.5
    2. normal weight if BMI is within [18.5, 25)
    3. overweight if BMI is within [25, 30)
    4. obese if BMI ≥ 30

```
gen BMI_cat=1 if bmi<18.5 & age>=18

replace BMI_cat=2 if bmi>=18.5 & bmi<25 &age>=18

replace BMI_cat=3 if bmi>=25 &bmi<30 &age>=18

replace BMI_cat=4 if bmi>=30 &age>=18
```

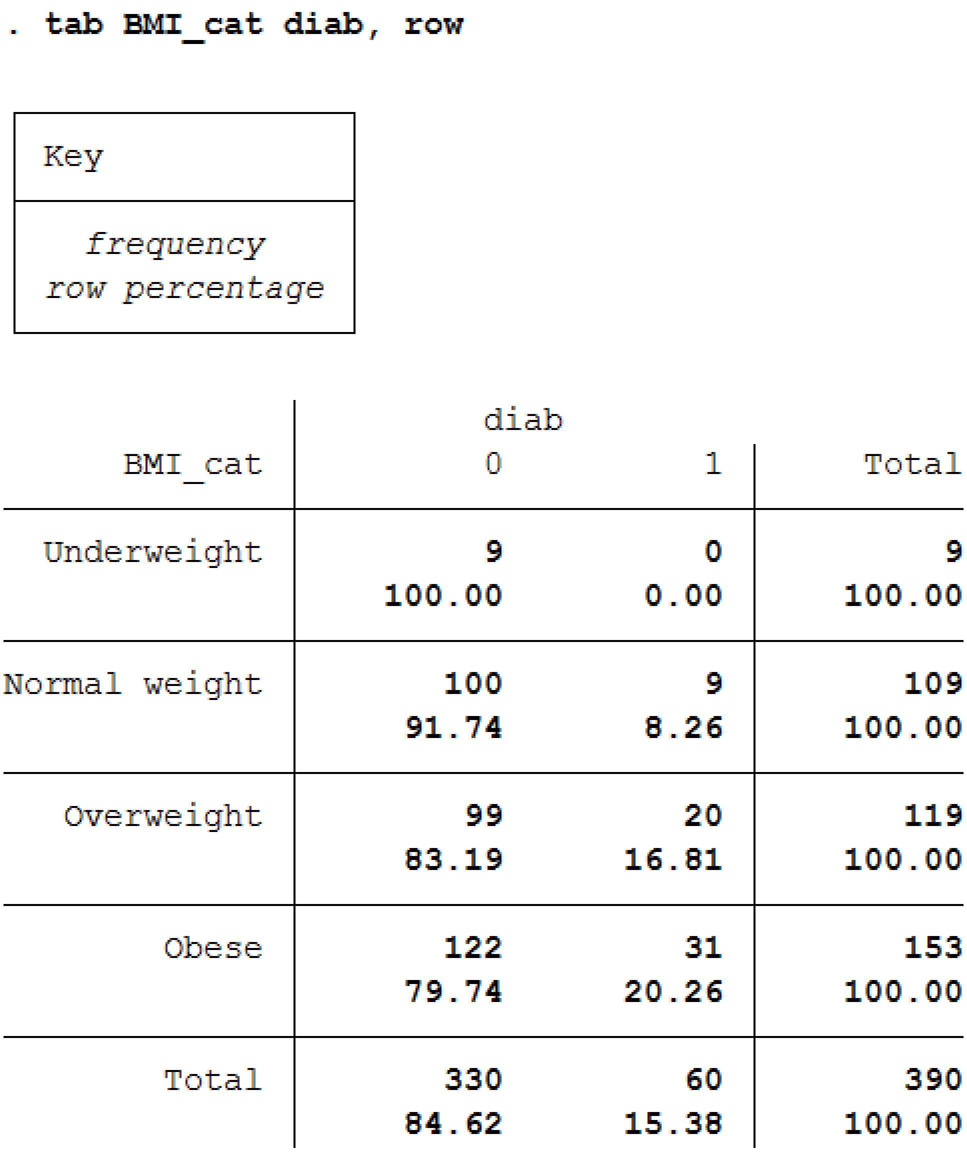
label define BMI_label 1 "Underweight" 2 "Normal weight" 3 "Overweight" 4 "Obese"

label values BMI_cat BMI_label

3. Get the contingency table for BMI categories and cross tab it with diabetes status?

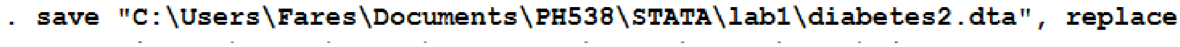


```
. tab BMI_cat

      BMI_cat |      Freq.     Percent        Cum.
--------------+-----------------------------------
  Underweight |          9        2.23        2.23
Normal weight |        113       28.04       30.27
   Overweight |        123       30.52       60.79
        Obese |        158       39.21      100.00
--------------+-----------------------------------
        Total |        403      100.00
```



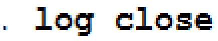
```
. tab BMI_cat diab, row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
+-------------------+

               |         diab
       BMI_cat |         0          1 |     Total
---------------+----------------------+----------
   Underweight |         9          0 |         9 
               |    100.00       0.00 |    100.00 
---------------+----------------------+----------
 Normal weight |       100          9 |       109 
               |     91.74       8.26 |    100.00 
---------------+----------------------+----------
    Overweight |        99         20 |       119 
               |     83.19      16.81 |    100.00 
---------------+----------------------+----------
         Obese |       122         31 |       153 
               |     79.74      20.26 |    100.00 
---------------+----------------------+----------
         Total |       330         60 |       390 
               |     84.62      15.38 |    100.00 
```

Finally, save the final data set:



```
. save "C:\Users\Fares\Documents\PH538\STATA\lab1\diabetes2.dta", replace
```

And close the log:



```
. log close
```