Analysis of Variance
Illustration – Stata *version 14*

Dear BIOSTATS 640 Spring 2023
Please be aware that this illustration may be out of date.
Thank you - cb

1. Preliminaries

Preliminary – Download the stata data set **hers_640anova.dta**.
Note – This data set is accessible through the internet. Alternatively, you can download it from the course website.

(a) In Stata, input directly from the internet using the command use
use "http://people.umass.edu/biep640w/datasets/hers_640anova.dta", clear

(b) From the course website, right click to download. Afterwards, in Stata, use FILE > OPEN
See, http://people.umass.edu/biep640w/webpages/demonstrations.html

Preliminary – Download the stata command **anovaplot**

```
. * Download "add-on" anova command anovaplot if you don't already have it
. findit anovaplot
```

Preliminary – Download the stata command **anovacontrast**

```
. * Download "add-on" anova command anovacontrast if you don't already have it
. findit anovacontrast
```

2. Illustration: Trial of Estrogen and Progesterone for Prevention of Heart Disease

Source
Hulley et al (1998) Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605-613

Source Data
The actual data set contains information on n=2,763. This was a randomized controlled trial investigation of hormone therapy for the prevention of heart attack and death.

The Data for this Illustration is a subset

(1) This dataset is a subset of size n=612 from a larger data set of Hulley et all (1998) that is described below. I have taken a subsample of size n=612 so that students with "small stata" will be able to reproduce this analysis.

(2) Specifically, I took a random sample of 300 whites, all 218 African-Americans, plus all 94 women of other race ethnicity.

Analysis Question
In this hypothetical data set, does systolic blood pressure (**sbp**) vary by race-ethnicity (**raceth**)?

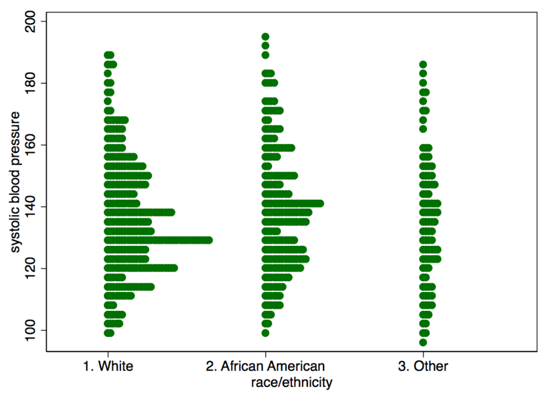
3. ONE WAY Analysis of Variance

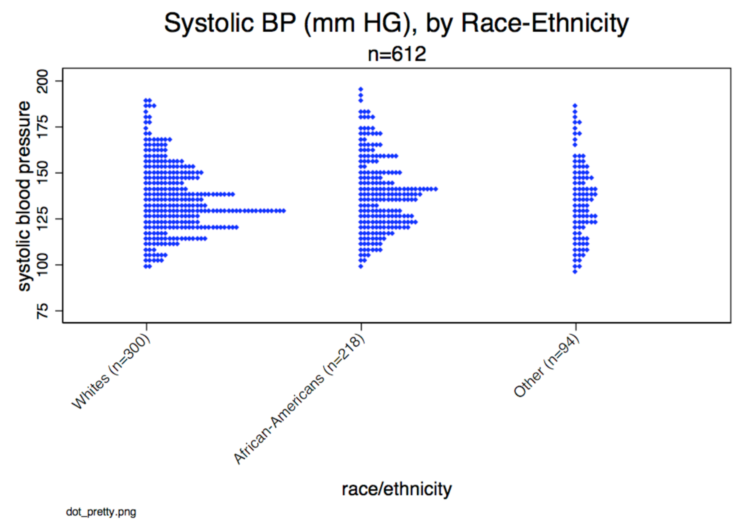
One Way Analysis of Variance Question
In this hypothetical data set, does Y = systolic blood pressure (**sbp**) vary over groups defined by X=race-ethnicity (**raceth**)?

3a. Descriptives – Graphs

```
. **  Set scheme for graphs - user choice
. set scheme s1color

. * get min and max of y=sbp for y-axis tick marks
. tabstat sbp, stat(min max)

    variable |       min       max
-------------+--------------------
         sbp |        95       194
----------------------------------

. * retrieve correspondence between codes and code labels and get sample sizes
. numlabel, add
. tabulate raceth

     race/ethnicity |      Freq.     Percent        Cum.
--------------------+-----------------------------------
           1. White |        300       49.02       49.02
2. African American |        218       35.62       84.64
           3. Other |         94       15.36      100.00
--------------------+-----------------------------------
              Total |        612      100.00


. **  Side-by-side dot plot -  NO FRILLS
. sort raceth
. dotplot sbp, over(raceth)

. ** Side-by-side dot plot - w AESTHETICS
. dotplot sbp, over(raceth) msymbol(d) msize(vsmall) mcolor(blue) ylabel(75(25)200, labsize(small)) xlabel(1 "Whites (n=300)" 2 
"African-Americans (n=218)" 3 "Other (n=94)", labsize(small) angle(45)) title("Systolic BP (mm HG), by Race-Ethnicity") 
subtitle("n=612") caption("dot_pretty.png", size(vsmall))
```

No Frills	W Aesthetics

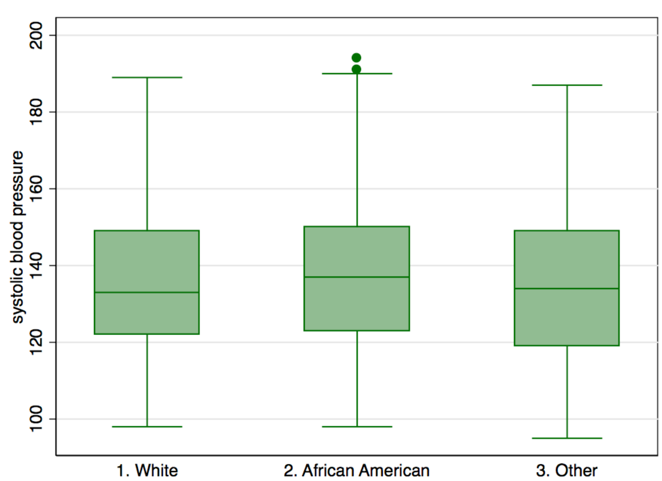


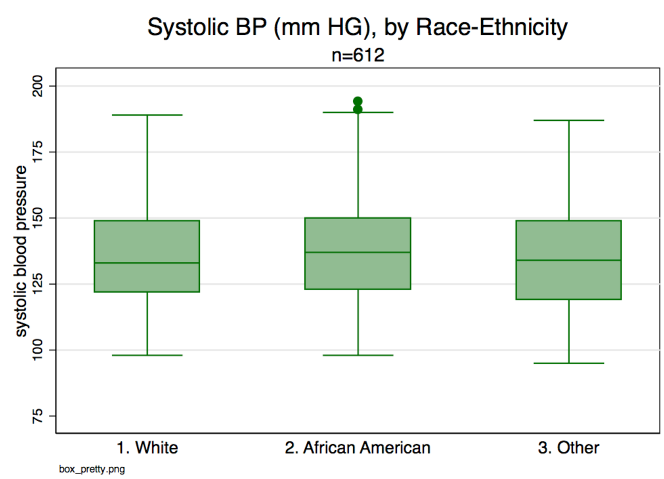


sbp distributions are similar in the 3 groups. It's unlikely that anova will be significant.

```
. ** Side-by-side box plot - NO FRILLS
. graph box sbp, over(raceth)

. ** Side-by-side box plot – w AESTHETICS
. graph box sbp, over(raceth) ylabel(75(25)200, labsize(small)) title("Systolic BP (mm HG), by Race-Ethnicity") subtitle("n=612")
caption("box_pretty.png", size(vsmall))
```

No Frills	W Aesthetics





3b. Descriptives – Numerical

Note - There is more than one way to do these!

```
. *  Descriptives -  Numerical Descriptives of Raw Data
. tabstat sbp, by(raceth) stat(n mean sd sem min q max)


Summary for variables: sbp
     by categories of: raceth (race/ethnicity)

           raceth |         N      mean        sd  se(mean)       min       p25       p50       p75       max
-----------------+--------------------------------------------------------------------------------------------
         1. White |       300  136.0133  18.55138  1.071064        98       122       133       149       189
2. African Ameri |       218  138.2339  19.99252  1.354064        98       123       137       150       194
         3. Other |        94  135.1809  21.25977  2.192778        95       119       134       149       187
-----------------+--------------------------------------------------------------------------------------------
            Total |       612  136.6765  19.50878  .7885956        95       122       135       149       194
--------------------------------------------------------------------------------------------------------------


. tabulate raceth, summarize(sbp)

race/ethnic |  Summary of systolic blood pressure
        ity |        Mean   Std. Dev.       Freq.
------------+------------------------------------
      White |   136.01333   18.551379         300
  African A |   138.23394   19.992518         218
      Other |   135.18085   21.259767          94
------------+------------------------------------
      Total |   136.67647   19.508777         612
```

Numerical descriptives confirm what we saw in the graphs. The means are similar. We see that they range 135.2 mm Hg to 138.2 mm Hg. The standard deviations appear to be similar, too. They range 18.55 mm Hg to 21.26 mm Hg. Test of equality of variances is likely to be non-significant.

```
. *  Descriptives -  Obtain group specific means and associated 95% confidence intervals
. sort raceth
. by raceth: ci means sbp
----------------------------------------------------------------------------------------------------------
-> raceth = White

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         sbp |        300    136.0133    1.071064        133.9056    138.1211

----------------------------------------------------------------------------------------------------------
-> raceth = African American

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         sbp |        218    138.2339    1.354064        135.5651    140.9027

----------------------------------------------------------------------------------------------------------
-> raceth = Other

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         sbp |         94    135.1809    2.192778        130.8264    139.5353
```

No surprise. Lots of overlap of the confidence intervals.

3c. Analysis of Variance Model Estimation

Stata offers at least 2 commands for a one way anova: **oneway** ***or*** **anova**.

KEY:

If you use **oneway**, then the predictor variable is allowed to be **string** or **numeric**
If you use **anova**, then the predictor variable **must be numeric**.

How to convert a string variable into a new, numeric variable.
Use the command **encode** with the option **generate** as follows:

encode *stringvar*, generate(*numericvar*)

Tip – STATA will automatically create variable value labels for your new numericvar.

Now you have two choices for performing your one way anova fit:

(1) oneway *yvariable factor*

(2) anova *yvariable numericfactor*

```
.* Illustration of command ANOVA
.* anova YVARIABLE NUMERICFACTOR
. anova sbp raceth
                           Number of obs =      612     R-squared     =  0.0037
                           Root MSE      = 19.5042     Adj R-squared =  0.0005

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  871.000171     2  435.500085       1.14     0.3190
                         |
                  raceth |  871.000171     2  435.500085       1.14     0.3190
                         |
                Residual |  231670.941   609  380.412054
              -----------+----------------------------------------------------
                   Total |  232541.941   611  380.592375
```

The null hypothesis assumption of equal means does not lead to an unlikely outcome (p=.32); the null hypothesis is NOT rejected.

```
.* TIP - ONEWAY provides Bartlett test of equal variances.
.* oneway YVARIABLE FACTOR
. oneway sbp raceth
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      871.000171      2   435.500085      1.14     0.3190
 Within groups      231670.941    609   380.412054
------------------------------------------------------------------------
    Total           232541.941    611   380.592375

Bartlett's test for equal variances:  chi2(2) =   3.1766  Prob>chi2 = 0.204

.* To obtain summary statistics, use command ONEWAY with option TABULATE
. oneway sbp raceth, tabulate

race/ethnic | Summary of systolic blood pressure
        ity |        Mean   Std. Dev.       Freq.
------------+------------------------------------
   1. White |   136.01333   18.551379         300
  2. Africa |   138.23394   19.992518         218
   3. Other |   135.18085   21.259767          94
------------+------------------------------------
      Total |   136.67647   19.508777         612

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      871.000171      2   435.500085      1.14     0.3190
 Within groups      231670.941    609   380.412054
------------------------------------------------------------------------
    Total           232541.941    611   380.592375

Bartlett's test for equal variances:  chi2(2) =   3.1766  Prob>chi2 = 0.204


.* To obtain reference cell coding regression solution, execute 2 commands
.* First - anova
.* Second - regress

.* Step 1 - Command is anova
. anova sbp raceth

                           Number of obs =       612     R-squared     =  0.0037
                           Root MSE      =   19.5042     Adj R-squared =  0.0005

                  Source | Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  871.000171     2  435.500085       1.14     0.3190
                         |
                  raceth |  871.000171     2  435.500085       1.14     0.3190
                         |
                Residual |  231670.941   609  380.412054
              -----------+----------------------------------------------------
                   Total |  232541.941   611  380.592375
```

```
.* Step 2 - Command is regress
. regress

      Source |       SS           df       MS      Number of obs   =       612
-------------+----------------------------------   F(2, 609)       =      1.14
       Model |  871.000171         2  435.500085   Prob > F        =    0.3190
    Residual |  231670.941       609  380.412054   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0005
       Total |  232541.941       611  380.592375   Root MSE        =    19.504

------------------------------------------------------------------------------------
              sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
           raceth |
African American  |   2.220612   1.735814     1.28   0.201    -1.188296    5.629519
           Other  |  -.8324823   2.305423    -0.36   0.718    -5.360027    3.695063
                  |
            _cons |   136.0133   1.126073   120.79   0.000     133.8019    138.2248
------------------------------------------------------------------------------------
```

Checking with means on page 8 – *they match!*

$$\hat{Y}_{WHITE} = \hat{Y}_{ref} = [\ \hat{b}_0\ ] = 136.01$$

$$\hat{Y}_{AFRICAN\text{-}AMERICAN} = \hat{Y}_2 = [\ \hat{b}_0 + \hat{b}_2\ ] = [136.01 + 2.22] = 138.23$$

$$\hat{Y}_{OTHER} = \hat{Y}_3 = [\ \hat{b}_0 + \hat{b}_3\ ] = [136.01 - 0.83\ ] = 135.18$$

3d. Post-hoc Pairwise Comparison of Groups

Pairwise comparisons of groups is done using the command pwcompare.
Note – You must have fit the model first using anova

```
.* PRELMINARY - Must first fit model using anova
. anova sbp raceth

                   Number of obs =        612     R-squared     =  0.0037
                   Root MSE      =    19.5042     Adj R-squared =  0.0005

          Source |  Partial SS    df       MS           F     Prob > F
      -----------+----------------------------------------------------
           Model |  871.000171     2  435.500085       1.14     0.3190
                 |
          raceth |  871.000171     2  435.500085       1.14     0.3190
                 |
        Residual |  231670.941   609  380.412054
      -----------+----------------------------------------------------
           Total |  232541.941   611  380.592375
```

```
.* Compare pairwise means with NO adjustment for multiple comparisons using PWCOMPARE
.* pwcompare FACTOR
. pwcompare raceth

Pairwise comparisons of marginal linear predictions

Margins      : asbalanced

--------------------------------------------------------------
             |                              Unadjusted
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      raceth |
     2 vs 1  |   .9654558   1.034007     -1.065196    2.996108
     3 vs 1  |   2.228419   1.373318     -.4685942    4.925432
     3 vs 2  |   1.262963   1.433615     -1.552466    4.078392
--------------------------------------------------------------
```

For all pairwise comparisons of groups, the 95% confidence interval includes the null hypothesis value of zero.

```
. * Pairwise comparison of means using Bonferroni adjustment (NOT RECOMMENDED)
. pwcompare raceth, mcompare(bonferroni) sort effects

Pairwise comparisons of marginal linear predictions

Margins      : asbalanced

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
      raceth |            3
---------------------------

-----------------------------------------------------------------------------
             |                            Bonferroni           Bonferroni
             |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
      raceth |
     3 vs 2  |  -3.053094   2.406646    -1.27   0.615    -8.830518    2.724331
     3 vs 1  |  -.8324823   2.305423    -0.36   1.000     -6.36691    4.701945
     2 vs 1  |   2.220612   1.735814     1.28   0.604    -1.946404    6.387628
-----------------------------------------------------------------------------
```

Even with the stringent Bonferroni adjustment, for all pairwise comparisons of groups, no statistically significant differences are found. Not surprising, given what we've already seen.

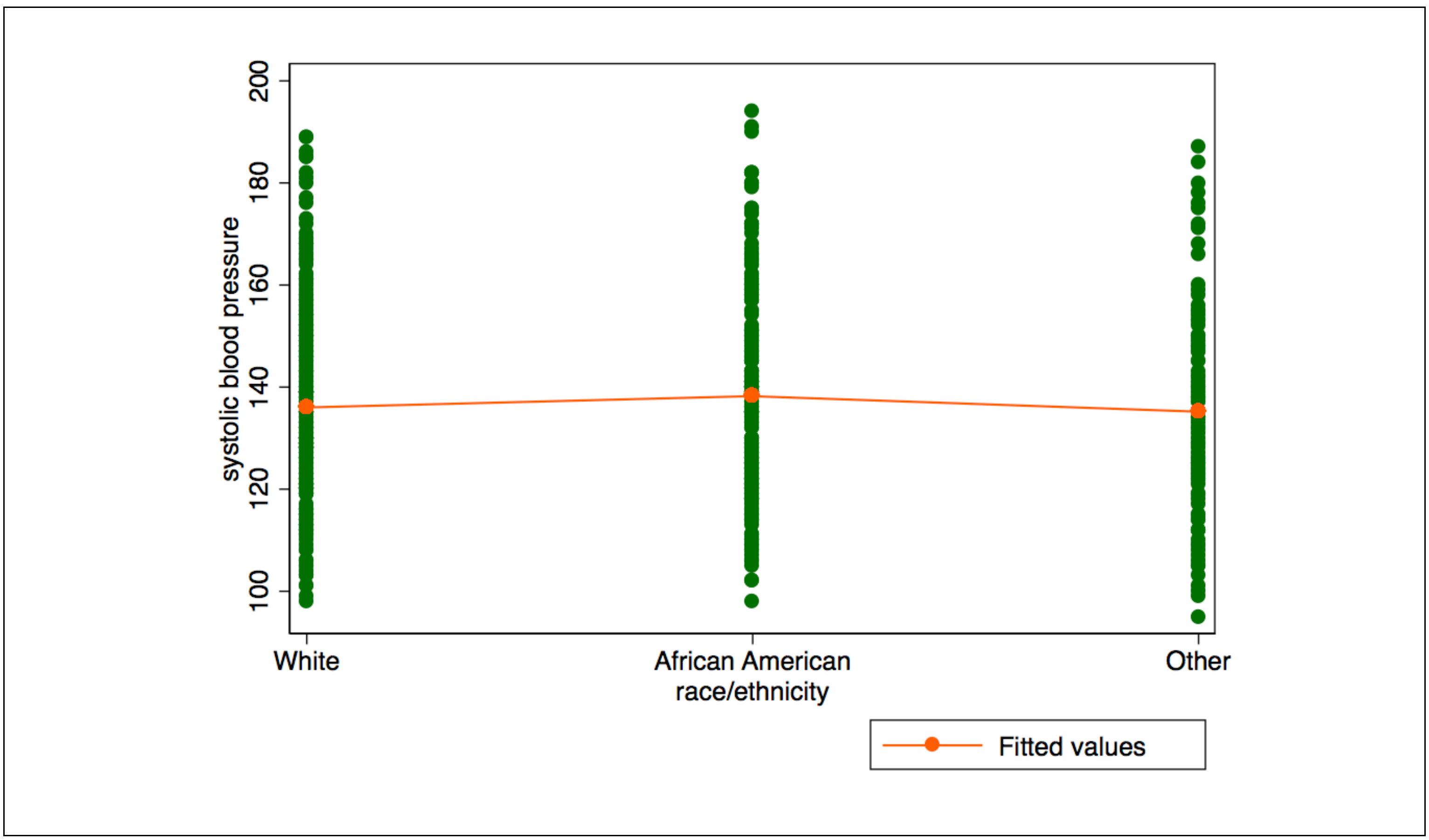
```
. * Pairwise comparison of means using Tukey adjustment for Multiple Comparisons
. * NOTE - This requires equal sample sizes in all groups
. * So, technically, I should not have done this.
. pwcompare raceth, mcompare(tukey)

Pairwise comparisons of marginal linear predictions

Margins      : asbalanced

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
      raceth |            3
---------------------------

-------------------------------------------------------------
             |                                 Tukey
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
      raceth |
     2 vs 1  |   2.220612   1.735814     -1.857604    6.298827
     3 vs 1  |  -.8324823   2.305423      -6.24897    4.584005
     3 vs 2  |  -3.053094   2.406646       -8.7074    2.601212
-------------------------------------------------------------
Note: The tukey method requires balanced data for proper level
      coverage. A factor was found to be unbalanced.
```

Yes, yes, I know. I should not have done the Tukey procedure

3e. Post-Hoc Graphs

PRELIMINARY – You must have fit the model first using **anova**

```
. anova sbp raceth

                    Number of obs =        612     R-squared     =  0.0037
                    Root MSE      =    19.5042     Adj R-squared =  0.0005

              Source | Partial SS    df       MS           F     Prob > F
          -----------+----------------------------------------------------
               Model |  871.000171     2   435.500085      1.14    0.3190
                     |
              raceth |  871.000171     2   435.500085      1.14    0.3190
                     |
            Residual |  231670.941   609   380.412054
          -----------+----------------------------------------------------
               Total |  232541.941   611   380.592375


. * anova plot - NO FRILLS
. anovaplot
```



4. ONE WAY Analysis of Variance Diagnostics

4a. Homogeneity of Variances

```
. * BARTLETT's Test is provided in output from command oneway
. * Caution:  This test is sensitive to the assumption of normality
. oneway sbp raceth

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      871.000171      2   435.500085      1.14     0.3190
 Within groups      231670.941    609   380.412054
------------------------------------------------------------------------
    Total           232541.941    611   380.592375

Bartlett's test for equal variances:  chi2(2) =   3.1766  Prob>chi2 = 0.204
```

The null hypothesis of equal variances is not rejected (Bartlett test p-value=.20)

```
. * LEVENE and BROWN-FORSYTHE tests are obtained using the command robvar
. * These are good choices when assumption of normality is in question.
. *  W_0  =  Levene test
. * W_50  =  Forsythe-Browne modification of Levene test (mean is replaced by median)
. * W_10  =  Fosythe-Browne modification of Levene test (mean is replaced by 10% trim)
. * robar(YVAR), by(FACTOR)
. robvar sbp, by(raceth)

race/ethnic | Summary of systolic blood pressure
        ity |        Mean   Std. Dev.       Freq.
------------+-----------------------------------
      White |   136.01333   18.551379         300
  African A |   138.23394   19.992518         218
      Other |   135.18085   21.259767          94
------------+-----------------------------------
      Total |   136.67647   19.508777         612

W0  =  1.4143305   df(2, 609)     Pr > F = 0.24388559   <----  Levene

W50 =  1.4701779   df(2, 609)     Pr > F = 0.23069929   <----  Brown-Forsythe with median

W10 =  1.4741613   df(2, 609)     Pr > F = 0.22978655   <----  Brown-Forsythe with 10% trimmed mean
```

The null hypothesis of equal variances is not rejected by Levene's test either (p-value=.24)

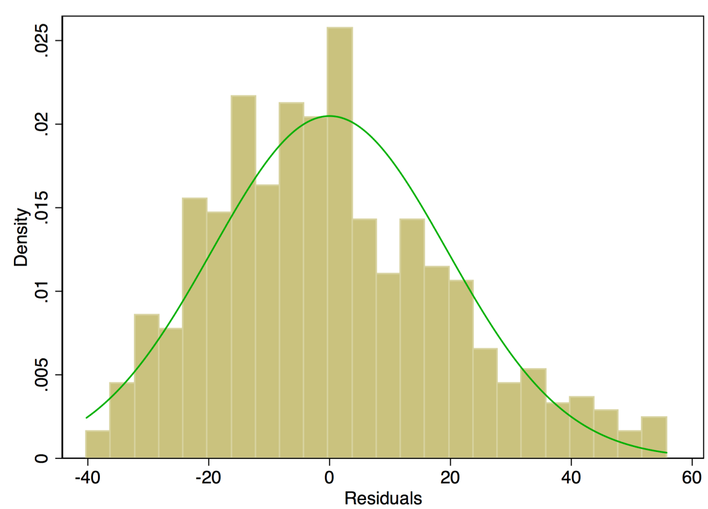
4b. Normality of Residuals

*NOTE – The following requires that you fit your ONE WAY anova model using the command **anova**. THe following will not work if you fit your model using the command **oneway**.*

```
. anova sbp raceth

                         Number of obs =        612    R-squared     =  0.0037
                         Root MSE      =    19.5042    Adj R-squared =  0.0005

                  Source | Partial SS         df         MS        F    Prob>F
              -----------+----------------------------------------------------
                   Model |  871.00017          2   435.50009      1.14  0.3190
                         |
                  raceth |  871.00017          2   435.50009      1.14  0.3190
                         |
                Residual |  231670.94        609   380.41205  
              -----------+----------------------------------------------------
                   Total |  232541.94        611   380.59238  

. predict ehat, residuals
. * Shapiro-Wilk Test (NULL:  Distribution is normal)
. swilk ehat

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
        ehat |        612    0.98066      7.810     4.986    0.00000
```

Boo hoo. There is statistically significant evidence of departure from normality. Null is rejected (p < .00001). Let's do a graphical look to see if we're really in trouble.

```
. histogram ehat, normal
```



This departure from normality is not so severe as to warrant transforming the data. For purposes of this illustration, we'll proceed under the assumption of normality.

5. TWO WAY Analysis of Variance

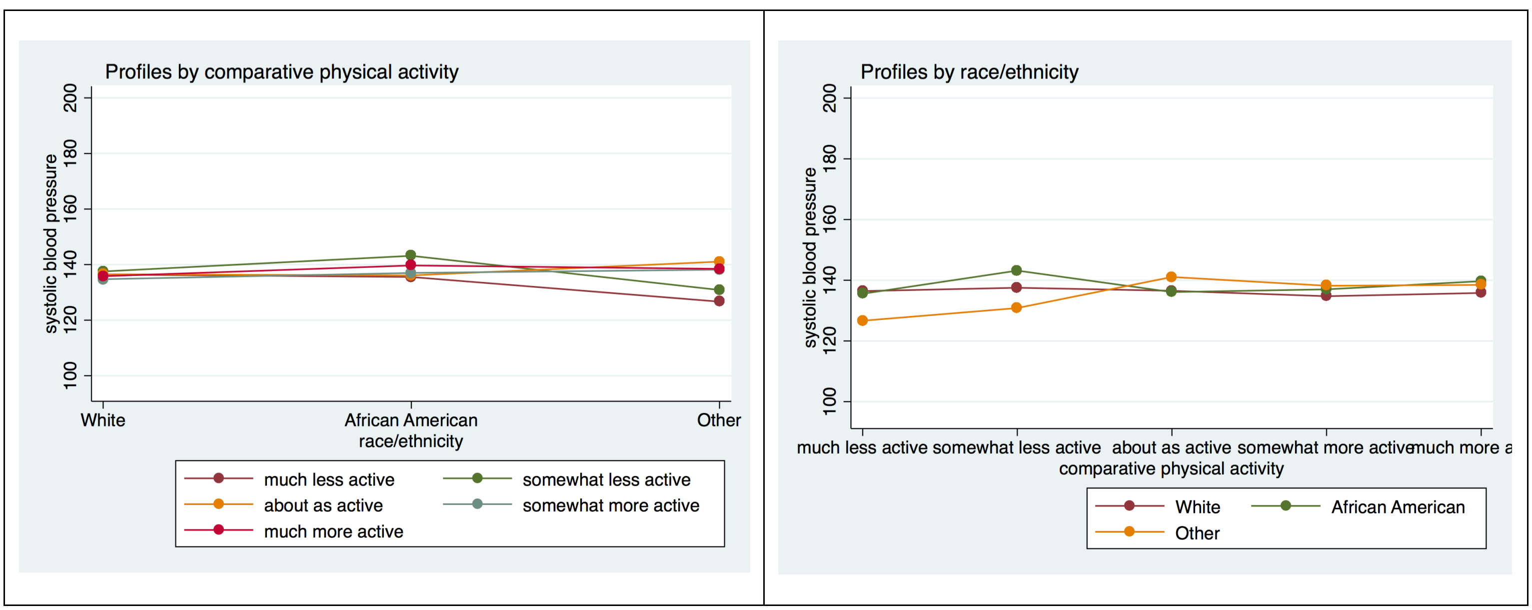
Two Way Analysis of Variance Question
In this hypothetical data set, does Y = systolic blood pressure (**sbp**) vary over groups defined by X_1=race-ethnicity (**raceth**) and X_2=physical activity (**physact**)?

5a. Descriptives – Graphs

We will use the command **anovaplot** here.

Our goal at this point is just a visualization of the means to see if there is interaction present. However, the stata command **anovaplot** requires that we have first executed the command **anova**. So we will do this "**quietly:**" which means that the output will not be shown.

```
. * Obtain plot of mean Y=sbp over groups in 2 ways to assess interaction
. quietly: anova sbp raceth physact raceth#physact

.* Plot of means of Y=sbp with raceth, SEPARATELY for groups defined by physact (left panel)
. anovaplot raceth physact, scatter(msym(none))

. * Plot of means of Y=sbp with physact, SEPARATELY for groups defined by raceth (right panel)
. anovaplot physact raceth, scatter(msym(none))
```



Okay, maybe this was not the best choice of example. These mean profiles look pretty similar.

5b. Descriptives – Numeric

```
. table physact raceth, contents(n sbp mean sbp sd sbp)

----------------------------------------------------------------------
comparative physical |                  race/ethnicity
activity             |          White  African American          Other
---------------------+------------------------------------------------
     much less active |             14                33             18
                     |       136.4286          135.5455       126.6667
                     |       19.63793          18.88467       22.68454
                     |
somewhat less active |             55                51             21
                     |       137.5273          143.1765       130.8571
                     |       18.71012          21.08621        19.1371
                     |
      about as active |             99                67             26
                     |       136.5253          136.1045       141.0769
                     |       17.62208          20.27921       21.07875
                     |
somewhat more active |            105                44             16
                     |       134.7333               137       138.1875
                     |       18.94337          19.55196       19.73565
                     |
     much more active |             27                23             13
                     |       135.8148          139.6956       138.4615
                     |       20.47895           18.5948       22.59283
----------------------------------------------------------------------
```

Pretty tidy summary! A quick scan, however, suggests that the means are not likely to be assessed as statistically significantly different. Not surprising given the graphs we obtained earlier.

5c. Analysis of Variance Model Estimation

```
. anova sbp raceth physact raceth#physact

                         Number of obs =        612    R-squared     =  0.0258
                         Root MSE      =    19.4804    Adj R-squared =  0.0029

                  Source | Partial SS         df         MS        F    Prob>F
          ---------------+----------------------------------------------------
                   Model |  5988.6916         14   427.76368      1.13  0.3300
                         |
                  raceth |  762.56725          2   381.28362      1.00  0.3668
                 physact |  1149.1915          4   287.29788      0.76  0.5535
          raceth#physact |  3563.4604          8   445.43255      1.17  0.3125
                         |
                Residual |  226553.25        597   379.48618
          ---------------+----------------------------------------------------
                   Total |  232541.94        611   380.59238
```

Oh dear. How boring. Nothing significant here.

5d. Post-hoc Comparison of Groups

```
. anova sbp raceth physact raceth#physact


. * Table of Fitted Means - NICELY LAID OUT
. predict yhat, xb
. label var yhat "TWO WAY Anova Fit"
. tabulate raceth physact, summarize(yhat) means

                        Means of TWO WAY Anova Fit

race/ethni |           comparative physical activity
      city | much less  somewhat   about as   somewhat   much more |     Total
-----------+-------------------------------------------------------+----------
     White | 136.42857  137.52727  136.52525  134.73334  135.81482 | 136.01333
 African A | 135.54546  143.17647  136.10448        137  139.69565 | 138.23394
     Other | 126.66666  130.85715  141.07692   138.1875  138.46153 | 135.18085
-----------+-------------------------------------------------------+----------
     Total | 133.27692  138.69291  136.99479  135.67273  137.77778 | 136.67647
```

No surprise here. Whether you read across the rows or down the columns, either way, we're not seeing much by way of changes in mean sbp. See again the pics on page 15.

```
. * Table of Fitted Means - not so nicely laid out, but with lots more information.
. margins raceth#physact, asbalanced post

Adjusted predictions                              Number of obs     =        612

Expression   : Linear prediction, predict()

------------------------------------------------------------------------------------------------------
                                        |            Delta-method
                                        |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------------------------+-------------------------------------------------------------
                         raceth#physact |
              White#much less active    |   136.4286   5.206357    26.20   0.000     126.2036    146.6536
          White#somewhat less active    |   137.5273   2.626737    52.36   0.000     132.3685     142.686
               White#about as active    |   136.5253   1.957854    69.73   0.000     132.6801    140.3704
          White#somewhat more active    |   134.7333   1.901093    70.87   0.000     130.9997     138.467
              White#much more active    |   135.8148   3.749006    36.23   0.000      128.452    143.1777
   African American#much less active    |   135.5455   3.391103    39.97   0.000     128.8855    142.2054
African American#somewhat less active   |   143.1765   2.727802    52.49   0.000     137.8192    148.5337
    African American#about as active    |   136.1045    2.37991    57.19   0.000     131.4305    140.7785
African American#somewhat more active   |        137   2.936782    46.65   0.000     131.2323    142.7677
   African American#much more active    |   139.6957   4.061945    34.39   0.000     131.7182    147.6731
              Other#much less active    |   126.6667   4.591575    27.59   0.000     117.6491    135.6843
          Other#somewhat less active    |   130.8571   4.250973    30.78   0.000     122.5085    139.2058
               Other#about as active    |   141.0769   3.820422    36.93   0.000     133.5738      148.58
          Other#somewhat more active    |   138.1875   4.870101    28.37   0.000     128.6229    147.7521
              Other#much more active    |   138.4615   5.402892    25.63   0.000     127.8506    149.0725
------------------------------------------------------------------------------------------------------
```

```
. * Post-hoc Tests of Simple Main Effects
. * PRELIMINARY - Must have fit model w anova command first
. anova sbp raceth physact raceth#physact

                     Number of obs =        612    R-squared     =  0.0258
                     Root MSE      =    19.4804    Adj R-squared =  0.0029

              Source | Partial SS         df         MS        F    Prob>F
      ---------------+----------------------------------------------------
               Model |  5988.6916         14   427.76368      1.13  0.3300
                     |
              raceth |  762.56725          2   381.28362      1.00  0.3668
             physact |  1149.1915          4   287.29788      0.76  0.5535
      raceth#physact |  3563.4604          8   445.43255      1.17  0.3125
                     |
            Residual |  226553.25        597   379.48618
      ---------------+----------------------------------------------------
               Total |  232541.94        611   380.59238

. * Test of simple main effect of predictor = physact, separately at each raceth
. sme physact raceth

Test of physact at raceth(1): F(4/597)  = .21576196
Test of physact at raceth(2): F(4/597)  = 1.2545496
Test of physact at raceth(3): F(4/597)  = 1.9011494


Critical value of F for alpha = .05 using ...
--------------------------------------------------
Dunn's procedure              = 2.9791402
Marascuilo & Levin            = 3.2193716
per family error rate         = 3.0491547
simultaneous test procedure   = 4.8816539
```

At each level of raceth, the F-statistic for testing equality of sbp over groups defined by physact is NOT in the critical region. Thus, we have no statistically significant evidence of variations in sbp by physact.

```
. * Test of simple main effect of predictor = raceth, separately at each physact
. sme raceth physact

Test of raceth at physact(1): F(2/597)  = 1.4432739
Test of raceth at physact(2): F(2/597)  = 3.1481138
Test of raceth at physact(3): F(2/597)  = .66958053
Test of raceth at physact(4): F(2/597)  = .35753139
Test of raceth at physact(5): F(2/597)  = .25655016


Critical value of F for alpha = .05 using ...
--------------------------------------------------
Dunn's procedure              = 4.003167
Marascuilo & Levin            = 4.414349
per family error rate         = 4.6408772
simultaneous test procedure   = 9.7633078
```

At each level of physact, the F-statistic for testing equality of sbp over groups defined by raceth is NOT in the critical region. Thus, we have no statistically significant evidence of variations in sbp by raceeth

```
. * Pairwise comparison of groups using command anovacontrast (requires installation)
. * CAUTION!!!!! If interaction is significant, this analysis may not be appropriate.

. * Preliminary:  Use command fre to see again the codings for raceth
. fre raceth

raceth -- race/ethnicity
------------------------------------------------------------------------
                           |      Freq.    Percent      Valid       Cum.
---------------------------+--------------------------------------------
Valid   1 White            |        300      49.02      49.02      49.02
        2 African American |        218      35.62      35.62      84.64
        3 Other            |         94      15.36      15.36     100.00
        Total              |        612     100.00     100.00
------------------------------------------------------------------------

. *  Pairwise comparison of 2 groups:  (1=White) v (2=African American)
. *  KEY:  Entries in option values(  ) must total 0 and must make sense
. *  (1=White) → 1,   (2=African American) → -1 and  (3=Other)  → 0
. quietly: anova sbp raceth physact raceth#physact
. anovacontrast raceth, values(1 -1 0)


Contrast variable: raceth (1 -1 0) Dep Var: sbp
source           SS          df      MS          Contrast =     -2.10
---------+--------------------------------       N of obs =       612
contrast | 397.018529        1    397.0185       F        =      1.05
error    |  226553.25      597    379.4862       Prob > F =    0.3068
---------+--------------------------------       t        =      1.02
```

Overall, there is no statistically significant difference in sbp, Whites v African-Americans (p=.31)

```
.  * Another example
. fre physact

physact -- comparative physical activity
----------------------------------------------------------------------------
                               |      Freq.    Percent      Valid       Cum.
-------------------------------+--------------------------------------------
Valid   1 much less active     |         65      10.62      10.62      10.62
        2 somewhat less active |        127      20.75      20.75      31.37
        3 about as active      |        192      31.37      31.37      62.75
        4 somewhat more active |        165      26.96      26.96      89.71
        5 much more active     |         63      10.29      10.29     100.00
        Total                  |        612     100.00     100.00
----------------------------------------------------------------------------

. *  Compare all "less" active versus all "more" active
. quietly: anova sbp raceth physact raceth#physact

. anovacontrast physact, values(-1 -1 0 1 1)


Contrast variable: physact (-1 -1 0 1 1) Dep Var: sbp
source           SS          df      MS          Contrast =      4.56
---------+--------------------------------       N of obs =       612
contrast | 378.775904        1    378.7759       F        =      1.00
error    |  226553.25      597    379.4862       Prob > F =    0.3182
---------+--------------------------------       t        =      1.00
```

Overall, there is no statistically significant difference in sbp, Less active v More active (p=.32)

6. TWO WAY Analysis of Variance Diagnostics

6a. Homogeneity of Variances

Spoiler – we need to trick Stata.

We will use the command **robvar** to obtain our test of equality of variance. However, STATA will not accommodate groups defined by 2 factors (in our case raceth and physact). So we will use the command **egen** with the option **group** to create a new variable that is just one collection of groups. In this case, I've named my new variable ***allgroups***.

```
. egen allgroups=group(raceth physact)
. robvar sbp, by(allgroups)
```

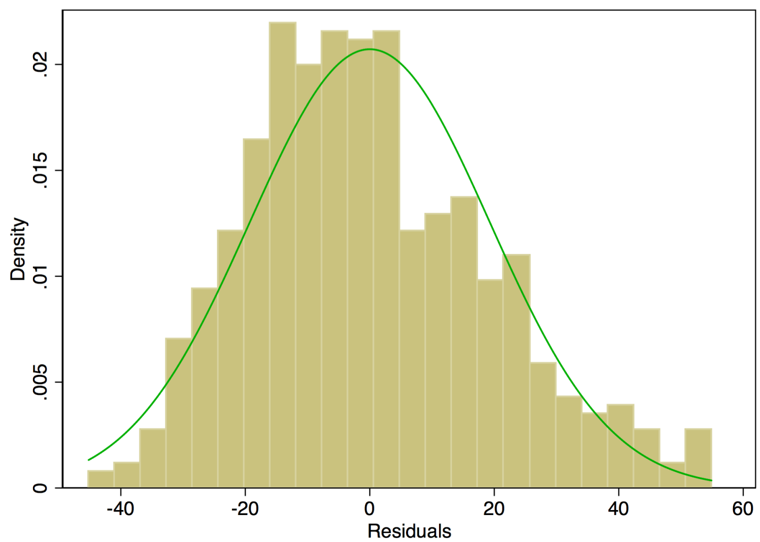
```
group(racet | Summary of systolic blood pressure
 h physact) |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   136.42857   19.637931          14
          2 |   137.52727   18.710123          55
          3 |   136.52525   17.622078          99
          4 |   134.73333    18.94337         105
          5 |   135.81481   20.478952          27
          6 |   135.54545   18.884668          33
          7 |   143.17647    21.08621          51
          8 |   136.10448   20.279213          67
          9 |         137   19.551958          44
         10 |   139.69565   18.594806          23
         11 |   126.66667   22.684537          18
         12 |   130.85714   19.137099          21
         13 |   141.07692   21.078753          26
         14 |    138.1875   19.735649          16
         15 |   138.46154   22.592828          13
------------+------------------------------------
      Total |   136.67647   19.508777         612

W0  =  0.27440108   df(14, 597)     Pr > F = 0.99621084

W50 =  0.24684554   df(14, 597)     Pr > F = 0.99786588

W10 =  0.27125653   df(14, 597)     Pr > F = 0.99643754
```

NOT Significant. We have no statistically significant evidence of a departure from equality of variances

6b. Normality of Residuals

Same as before(see again Section 4.b)

```
. quietly: anova sbp raceth physact raceth#physact

. predict ehat, residuals
. * Shapiro-Wilk Test (NULL:  Distribution is normal)
. swilk ehat


                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
        ehat |        612    0.98058      7.840     4.995    0.00000


. histogram ehat, normal
```



This departure from normality is not so horrible.