# Unit 7

## Stata for Analysis of One, Two and Three$^{+}$ Samples

***version 16***

*"Vive la difference!"*

Statistical analysis often involves the fitting of sophisticated models (multiple predictor linear regression, logistic, survival, mixed models, etc). Among the limitations of these methods are: (1) it is difficult to appreciate the actual data; and (2) their validity rest on assumptions that may or may not hold.

Analyses of data should begin with simple approaches that are as close to the data and as ***"model-free"*** as possible. These have the advantage of being simple, relatively assumption free, and straightforward in their interpretation.

This unit describes the use of Stata for estimation and hypothesis tests of data in one, two and more than two samples.

***Important!*** Be sure that you have already produced your data descriptions (See again, Units 5 – ***Stata for Data Description*** and 6- ***Stata for Graphs***)!

## Table of Contents

# Learning Objectives

When you have finished this unit, you should be able to produce, using Stata:

- Confidence intervals and hypothesis tests for **one continuous variable** under the assumption of normality**;**
- A nonparametric hypothesis test for **one continuous or ordinal variable** in the small study setting where the assumption of normality is not appropriate;
- Confidence intervals and hypothesis tests for **one proportion** under the assumption of a binomial distribution;
- Confidence intervals and hypothesis for **paired continuous variables** under the assumption of normality;
- A nonparametric hypothesis test for **paired continuous or ordinal variables** in the small study setting where the assumption of normality is not appropriate;
- Confidence intervals and hypothesis tests for **two independent variables – continuous** under the assumption of normality;
- Confidence intervals and hypothesis tests for **two independent proportions** under the assumption of independent binomial distributions;
- A nonparametric hypothesis test for **two independent continuous or ordinal variables** in the small sample setting where the assumption of normality is not appropriate;
- A **one way analysis of variance** under the assumption of normality; ***and***
- A nonparametric hypothesis test for the comparison of three or more independent medians in the small sample setting where the assumption of normality is not appropriate.

***How to follow along:***

**These notes utilize 4 data sets.**

**Download from the course website.**
1. **sepsis.dta**
2. **hers_640anova.dta**

**Access using the Stata command sysuse as indicated in the illustrations contained in these notes**
3. **bpwide.dta**
4. **auto.dta**

# Sample Session

***How to follow along:***

**1) Download from the course website the data set sepsis.dta.**
**2) Launch Stata**
**3) From whatever directory is appropriate for you, open sepsis.dta**

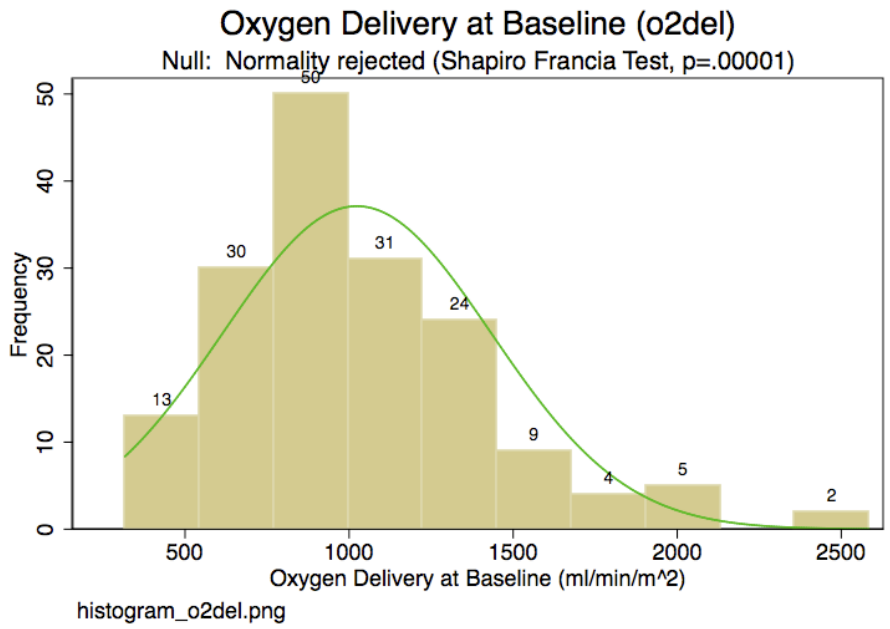
References to data set used:

Dupont WD. Statistical Modeling for Biomedical Researchers, Second Edition. Cambridge University Press, 2008..

Benard GR, Wheeler AP et al (1997) The effects of ibuprofen on the physiology and survival of patients with sepsis. The Ibuprofen in Sepsis Study Group. NEJM 336: 912-8.

```
. *-------------------------------------------------------------------------------
. *  BIOSTATS 690C - Data Management & Applied Data Analysis with Stata
. *
. *   prog:       Carol Bigelow
. *  input:       sepsis.dta
. * output:       none
. *  title:       Illustration of One and Two Plus Sample Inference
. *-------------------------------------------------------------------------------
. * -----------  Preliminaries ------------------------------------------------------
. cd "/Users/cbigelow/Desktop/"
  /Users/cbigelow/Desktop/
 . set more off
 . set scheme s1color


. * -----   Input sepsis.dta to your workspace
From menu at top:  FILE > OPEN > cbigelow/Desktop/sepsis.dta"


. * -----   For this illustration, let's keep just a few variables
. keep temp0 temp7 treat fate apache o2del id
. codebook, compact
Variable   Obs Unique      Mean     Min      Max  Label
-------------------------------------------------------------------------------------------
id         455    455       228       1      455  Patient ID
treat      455      2  .4923077       0        1  Treatment
apache     454     38   15.3304       0       41  Baseline APACHE Score
o2del      168    168  1023.817  316.88  2584.34  Oxygen Delivery at Baseline (ml/min/m^2)
fate       455      2  .3868132       0        1  Mortal Status at 30 Days
temp0      455    122  100.4269   91.58      107  Baseline Temperature (deg. F)
temp7      413    105  99.19448    88.7   104.18  Temperature after 24 hours
-------------------------------------------------------------------------------------------
```

```
. *---------  Command LABEL LIST to review discrete variable value labels ------*
. label list
race:
           0 White
           1 Black
           2 Other
fate:
           0 Alive
           1 Dead
treatmnt:
           0 Placebo
           1 Ibuprofen

. * ___________________  ONE SAMPLE INFERENCE - Continuous Variable  _____________*


. *------------------ Oxygen Delivery at Baseline (o2del) -------------------*
. * Command TABSTAT with option STAT( ) to obtain descriptives*
. tabstat o2del, stat(n mean sd sem med min max) longstub

    variable |         N      mean        sd  se(mean)       p50       min       max
-------------+--------------------------------------------------------------------------
       o2del |       168  1023.817  409.4426  31.58918     947.2    316.88   2584.34
-------------------------------------------------------------------------------------

. * Command SFRANCIA for test of Assumption of Normality (Null: Distribution is Normal)*
. sfrancia o2del

                  Shapiro-Francia W' test for normal data

    Variable |    Obs       W'          V'        z       Prob>z
-------------+-------------------------------------------------
       o2del |    168    0.93575      8.926     4.411    0.00001

Key:
Assumption of the null hypothesis of normality has led to an unlikely result (p-value <
.00001).  The null hypothesis is rejected.  Conclude there is statistically significant
evidence that o2del is not distributed normal.

. * Command HISTOGRAM with option NORMAL for Histogram with overlay Normal*
. histogram o2del, bin(10) frequency addlabels normal title("Oxygen Delivery at Baseline
(o2del)") subtitle("Null:  Normality rejected (Shapiro Francia Test, p=.00001)")
caption("histogram_o2del.png")
(bin=10, start=316.88, width=226.74601)
```



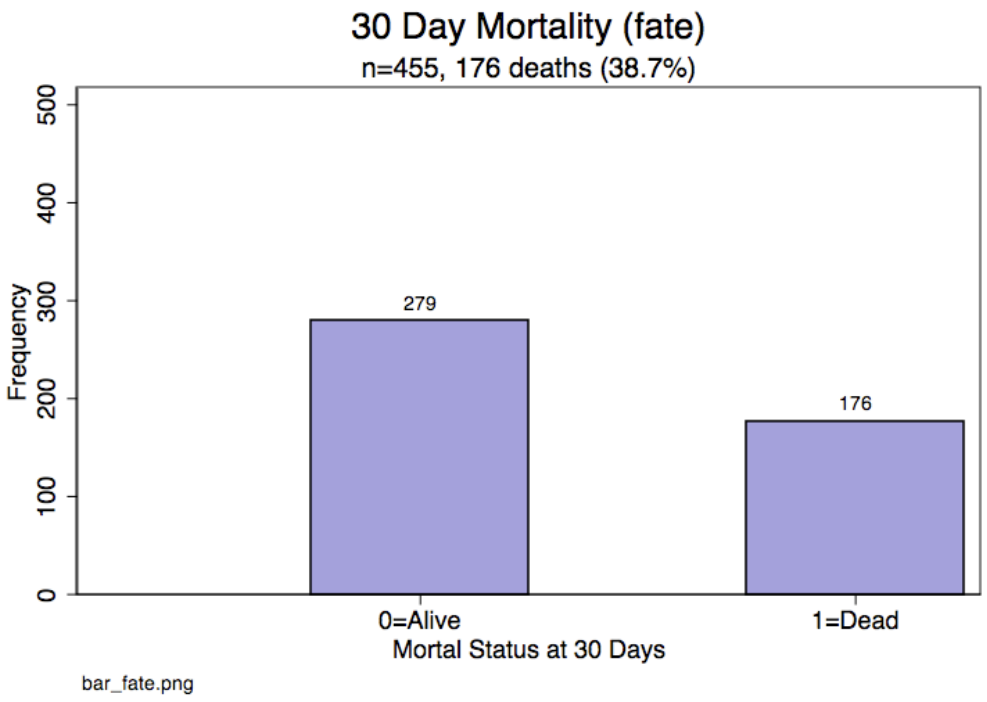

Key:
While the Shapiro-Francia test was statistically significant, the histogram suggests that the departure from normality is not so severe as to warrant a transformation or a non-parametric analysis.

```
. * Command TTEST for One Sample t-test that mean = 950.  *
. ttest o2del=950

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   o2del |     168    1023.817    31.58918    409.4426    961.4515    1086.183
------------------------------------------------------------------------------
    mean = mean(o2del)                                            t =   2.3368
Ho: mean = 950                                 degrees of freedom =      167

   Ha: mean < 950               Ha: mean != 950                 Ha: mean > 950
 Pr(T < t) = 0.9897         Pr(|T| > |t|) = 0.0206          Pr(T > t) = 0.0103

.*  Command CI means with option LEVEL( ) for  confidence interval for mean.
. ci means o2del, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
       o2del |        168    1023.817    31.58918        941.5086    1106.126
```

```
. * ____________________  ONE SAMPLE INFERENCE - Discrete Variable  _______________*
. *------------------ 30 Day Mortality (fate) -------------------*
 * Command TAB1 to obtain one way descriptives*
. tab1 fate
-> tabulation of fate

    Mortal |
  Status at |
    30 Days |      Freq.     Percent        Cum.
------------+-----------------------------------
      Alive |        279       61.32       61.32
       Dead |        176       38.68      100.00
------------+-----------------------------------
      Total |        455      100.00

. * Command FRE to obtain one-way descriptives (my preferred approach – cb)
. fre fate
fate -- Mortal Status at 30 Days
-------------------------------------------------------------
                |      Freq.    Percent      Valid       Cum.
----------------+--------------------------------------------
Valid   0 Alive |        279      61.32      61.32      61.32
        1 Dead  |        176      38.68      38.68     100.00
        Total   |        455     100.00     100.00
-------------------------------------------------------------

.* Command CI means for a confidence Interval estimate of the probability of death
.* normal approximation method
. ci means fate, level(95)

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        fate |        455    .3868132     .022857        .3418946    .4317318

.* Command CI proportions for a confidence interval using exact binomial method
. ci proportions fate, level(95)

                                                      -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        fate |        455    .3868132    .0228319        .3418278    .4332801

.* Command PRTEST for One sample test of proportion = 0.30
.* normal approximation method
. prtest fate=.30, level(95)

One-sample test of proportion                  fate: Number of obs =      455
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        fate |   .3868132   .0228319                      .3420636    .4315628
------------------------------------------------------------------------------
    p = proportion(fate)                                          z =   4.0409
Ho: p = 0.3

     Ha: p < 0.3                 Ha: p != 0.3                 Ha: p > 0.3
 Pr(Z < z) = 1.0000         Pr(|Z| > |z|) = 0.0001          Pr(Z > z) = 0.0000
```

Key:
The 2-sided test is statistically significant (p = .0001). Reject the null.

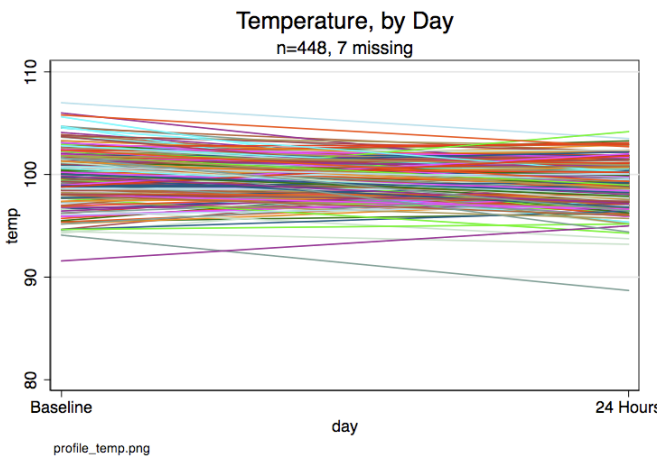
```
. * Command BITEST for one sample test of proportion, exact binomial method
. bitest fate=.30

    Variable |        N   Observed k   Expected k   Assumed p   Observed p
-------------+------------------------------------------------------------
        fate |      455         176        136.5     0.30000     0.38681

  Pr(k >= 176)            = 0.000047  (one-sided test)
  Pr(k <= 176)            = 0.999969  (one-sided test)
  Pr(k <= 98 or k >= 176) = 0.000079  (two-sided test)

Key:
The 2-sided test is statistically significant (p = .00008).  Reject the null.

. * Command HISTOGRAM with option DISCRETE for Discrete Variable Bar Chart
. histogram fate, discrete frequency barwidth(.5) addlabels title("30 Day Mortality (fate)")
   subtitle("n=455, 176 deaths (38.7%)") xlabel(0 "0=Alive" 1 "1=Dead") ylabel(0 (100)500)
   fcolor(lavender) lcolor(black)  note("bar_fate.png")
(start=0, width=1)
```

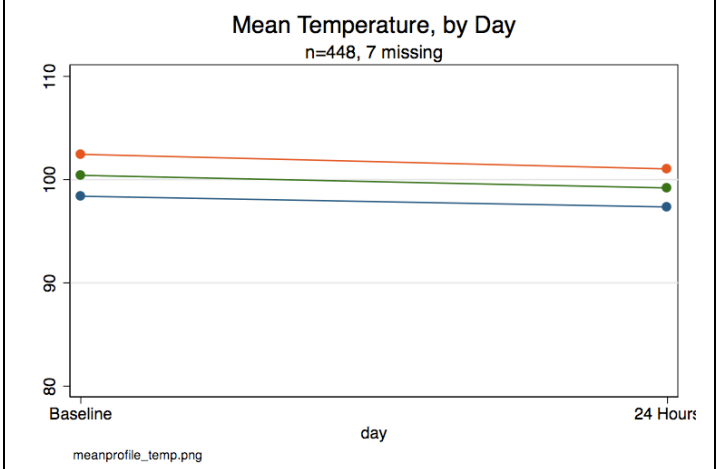


```
. *_____________ Paired Sample Inference _________________*
. *-------------- Repeated Measurement of temperature (temp0 and temp7) -------*
. generate chg_24hrs=temp0-temp7
(42 missing values generated)
. label variable chg_24hrs "Baseline - 24 Hour Change"
. tabstat temp0 temp7 chg_24hrs, col(stat) stat(n mean sd sem med min max) longstub

    variable |         N      mean        sd  se(mean)       p50       min       max
-------------+--------------------------------------------------------------------------
       temp0 |       455  100.4269  2.026105  .0949853     100.7     91.58       107
       temp7 |       413  99.19448  1.842151  .0906463     99.14      88.7    104.18
   chg_24hrs |       413  1.285957  1.988315  .0978386  1.220001 -5.400002  8.299995
-------------------------------------------------------------------------------------

. *------- Command TTEST for paired t test of equality of temp0 and temp 7 ------ *
. ttest temp0=temp7

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   temp0 |     413    100.4804    .0956495    1.943828    100.2924    100.6685
   temp7 |     413    99.19448    .0906463    1.842151    99.01629    99.37267
---------+--------------------------------------------------------------------
    diff |     413    1.285957    .0978386    1.988315    1.093632    1.478282
------------------------------------------------------------------------------
     mean(diff) = mean(temp0 - temp7)                              t =  13.1437
 Ho: mean(diff) = 0                              degrees of freedom =      412

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Key:
The 2-sided test is statistically significant (p << .0001). Assumption of the null hypothesis and its application to the data has led to an unlikely result. → Reject the null hypothesis of equal means.

```
. *------------- individual line plots -----------*
. *------ must re-shape data to long version prior to plot ----*
. * ---- IMPORTANT PRELMINARY:  Use preserve to set aside the original data for later use --*
. preserve
. reshape long temp, i(id) j(day)
(note: j = 0 7)

       ----- some output omitted ---

. *---------- individual profiles of change in temperature (command xtline is good here -----*
. * SUGGESTION: Watch your screen after issuing the command below and see the lines appear!
. xtline temp, i(id) t(day) ylabel(80 (10)110, grid) overlay title("Temperature, by Day")
  subtitle("n=448, 7 missing") xlabel(0 "Baseline" 7 "24 Hours") legend(off)
  note("profile_temp.png")

. *---------- mean profile of change in temperature -----------*
. sort day
. collapse (mean) temp (sd) sdtemp=temp, by(day)
. generate high=temp + sdtemp
. generate low=temp - sdtemp

. graph twoway (connected temp day) (connected high day) (connected low day), ylabel(80 (10)110, grid)
xlabel(0 "Baseline" 7 "24 Hours") legend(off) note("meanprofile_temp.png") title("Mean Temperature, by
Day") subtitle("n=448, 7 missing")
```

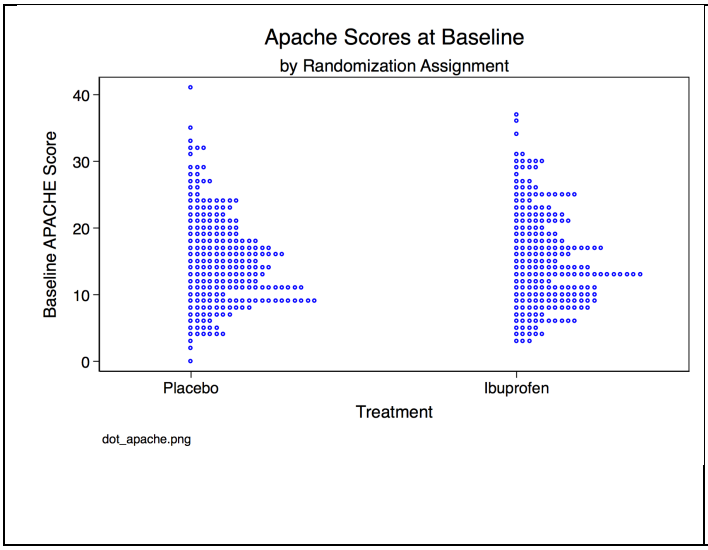


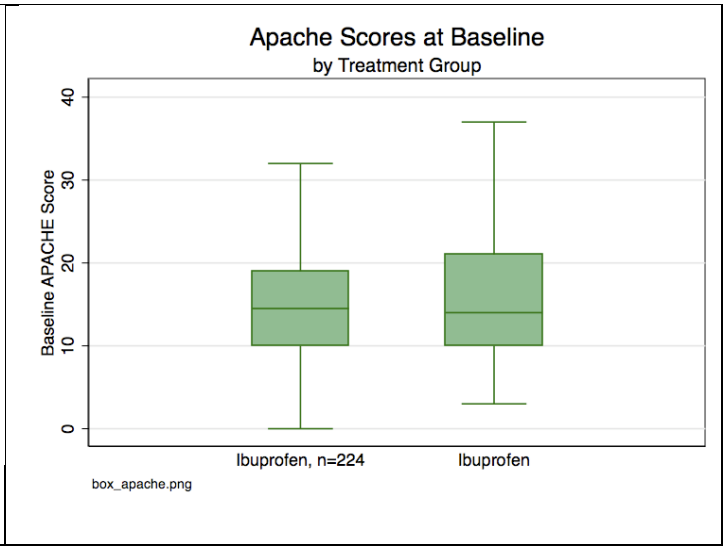



```
. *_________________  Two Independent Samples Inference _________________*
. * Continuous outcome (apache) in independent groups (treat)
. *---- IMPORTANT:  Issue the command restore to recover back the original data ------*
. clear
. restore
. tab1 treat
-> tabulation of treat

  Treatment |      Freq.     Percent        Cum.
------------+-----------------------------------
    Placebo |        231       50.77       50.77
  Ibuprofen |        224       49.23      100.00
------------+-----------------------------------
      Total |        455      100.00



.* Descriptives of outcome (apache) by group (treat)
. sort treat
. tabstat apache, by(treat) col(stat) stat(n mean sd sem med min max) longstub

treat         variable |         N      mean        sd  se(mean)       p50       min       max
-----------------------+--------------------------------------------------------------------------
Placebo         apache |       230  15.18696  6.922831   .456478      14.5         0        41
Ibuprofen       apache |       224  15.47768  7.261882  .4852049        14         3        37
-----------------------+--------------------------------------------------------------------------
Total           apache |       454   15.3304  7.085794  .3325528        14         0        41
--------------------------------------------------------------------------------------------------



. *------ test of equality of variances -----*
. sort treat
. sdtest apache, by(treat)
Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Placebo |     230    15.18696     .456478    6.922831    14.28752    16.08639
Ibuprofe |     224    15.47768    .4852049    7.261882    14.52151    16.43385
---------+--------------------------------------------------------------------
combined |     454     15.3304    .3325528    7.085794    14.67686    15.98393
------------------------------------------------------------------------------
    ratio = sd(Placebo) / sd(Ibuprofe)                            f =   0.9088
Ho: ratio = 1                                 degrees of freedom = 229, 223

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.2362         2*Pr(F < f) = 0.4724          Pr(F > f) = 0.7638
```

Key:
Assumption of the null hypothesis and its application to the data has NOT led to an unlikely result (p-value = .47). The null hypothesis of equal variances is NOT rejected. Conclude these data provide NO statistically significant evidence that the variances are different.

```
. *------- Command TTEST with option by( ) for 2 sample t test for independent groups -----*
. ttest apache, by(treat)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Placebo |     230    15.18696     .456478    6.922831    14.28752    16.08639
Ibuprofe |     224    15.47768    .4852049    7.261882    14.52151    16.43385
---------+--------------------------------------------------------------------
combined |     454     15.3304    .3325528    7.085794    14.67686    15.98393
---------+--------------------------------------------------------------------
    diff |            -.290722    .6657587               -1.599088    1.017644
------------------------------------------------------------------------------
    diff = mean(Placebo) - mean(Ibuprofe)                         t =  -0.4367
Ho: diff = 0                                     degrees of freedom =      452

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.3313         Pr(|T| > |t|) = 0.6626          Pr(T > t) = 0.6687

Key:
Assumption of the null hypothesis has NOT let to an unlikely result (p-value = .66).  The null
hypothesis of equal means is NOT rejected.  Conclude these data provide NO statistically significant
evidence that the means are different.

. *------- Command DOTPLOT with option over( ) for side by side dot plot ---------*
. dotplot apache, over(treat) nx(50) msymbol(oh) msize(small) mcolor(blue) title("Apache Scores at
Baseline") subtitle("by Randomization Assignment") note("dot_apache.png")

. *------- Command GRAPH BOX with option over( ) for by side box and whisker plot ---------*
. graph box apache, nooutsides over(treat,relabel(0 "Placebo, n=230" 1 "Ibuprofen, n=224"))
outergap(150) title("Apache Scores at Baseline") subtitle("by Treatment Group") note("box_apache.png")
```



# Introduction to "Immediate" Commands in Stata

**In a Stata "immediate" command, a data set is NOT used. Instead, the command includes input of the statistics needed to perform the analysis.**

- An "immediate" command instructs STATA to perform a calculation using values of statistics provided in the command.
- A Stata "immediate" command always ends in a **"i"**

**Example**

| | |
|---|---|
| **. * cii n xbar s**<br>**. cii 74 21.29 5.78, level(90)**<br>**n xbar s** | This produces a 90% confidence interval estimate of the mean of a variable for which we do not have the actual data but for which we do know that<br><br>n = 74<br>xbar = 21.29<br>s = 5.78 |

## Immediate Commands in Stata

**They end in "i"**
**and you provide the data inputs….**

**Examples: cii, ttesti, sdtesti, bitesti.**

# 1. One Sample Inference

***How to follow along;***
**. clear**
**. sysuse auto**

## 1.1 Continuous Outcome – Nonparametric Test of Median: The Signed Rank Test

| Command | Example |
|---|---|
| **One Sample Signed Rank Test of Median**<br>**signrank** ***variable=nullmedian,***<br>**The option level( ) is NOT allowed** | **.signrank mpg=20**<br>**This produces a one sample Wilcoxon Signed Rank test of the median of mpg is = 20** |

## 1.2 Continuous Outcome – Tests of Assumption of Normality

**Review -** Many statistical methods (especially linear regression) assume that the distribution of a variable (for example the dependent or Y-variable) is normal. Thus, it is useful to test this assumption. Stata offers two tests of normality: **Shapiro-Wilks** and **Shapiro-Francia**. Each is a test of the null hypothesis that the data are distributed normal.

**What to look for -**

| | Data are Normal | Data are NOT Normal |
|---|---|---|
| Null hypothesis ("normality") | NOT rejected | rejected |
| p-value* | large | small |

*** Note – In Stata the p-value appears the value listed under "Prob > z"**

Violations of the assumption of normality, if modest, are sometimes not a serious problem:

- Estimation and hypothesis tests of regression parameters are fairly robust to modest violations of normality;
- ***When to worry***: Predictions are sensitive to violations of normality
- ***Beware***: Sometimes the cure for violations of normality is worse than the problem.

| Command | Example |
|---|---|
| **Shapiro-Wilk Test**<br>swilk *variable* | .swilk mpg<br>The null hypothesis is normality. Thus, the assumption of normality is reasonable when the test returns a p-value that is NOT statistically significant. |
| **Shapiro-Francia Test**<br>sfrancia *variable* | .sfrancia mpg<br>The null hypothesis is again normality. Thus, the assumption of normality is reasonable when the test returns a p-value that is NOT statistically significant. |
| **Skewness-Kurtosis Test**<br>sktest *variable* | .sktest mpg<br>The null hypothesis is again normality. Thus, the assumption of normality is reasonable when the test returns a p-value that is NOT statistically significant. |

## 1.3 Continuous Outcome: Mean of a Normal Distribution

| Command | Example |
|---|---|
| **Confidence Interval for Mean**<br>ci means *variable,* level*(#)* | .ci means mpg, level(90)<br>This produces a 90% confidence interval estimate of the mean of the variable mpg |
| **Confidence Interval for Mean,** *"immediate"*<br>cii means *n xbar s,* level*(#)* | .cii means 74 21.2973 5.785503, level(90)<br>This produces a 90% confidence interval estimate of the mean of an UNNAMED variable for which n=74, xbar=21.2973 and the sample s=5.785503 |
| **t-test for Mean**<br>ttest *variable=nullmean,* level*(#)* | .ttest mpg=20, level(90)<br>This produces a one sample t-test of the null hypothesis that the mean of mpg is μ = 20 for the . The output includes a 90% confidence interval |
| **t-test for Mean,** *"immediate"*<br>ttesti *n xbar sigma nullmean,* level*(#)* | .ttesti 74 21.2973 5.785503 20, level(90)<br>This produces a one sample t-test of the null hypothesis that the mean of an UNNAMED variable is μ = 20 in the setting where n=74, xbar=21.2973 and the sample s=5.785503 |

## 1.4 Continuous Outcome – Variance of a Normal Distribution

| Command | Example |
|---|---|
| **One Sample Test of Variance**<br>**sdtest** *variable=nullsigma,*<br>**NOTE! You supply the null standard deviation, NOT the null variance**<br><br>**One Sample Test of Variance,** ***"immediate"***<br>**sdtesti** *variable n . sigma nullsigma*<br><br>**Notes - (1) The period that you type is in place of the sample mean. You could supply this if you have it, but it is not necessary for the test of variance. (2) You specify the null standard deviation, NOT the null variance.** | **.sdtest mpg=5**<br>**This produces a one sample test of the null hypothesis that the variance of mpg is $5^2 = 25$**<br><br>**.sdtesti 74 . 5.78 6**<br><br>**Note the period -**<br>**Take care to provide the period in place of the sample mean. Otherwise you will get an uninterpretable error message!** |

## 1.5 Discrete Outcome – Binomial Proportion

| Command | Example |
|---|---|
| **Exact Confidence Interval for Binomial π**<br>**ci proportions *variable*, level(#)**<br>**This produces Clopper-Pearson "exact" confidence interval** | **.ci proportions foreign, level(90)**<br>**This produces an exact 90% confidence interval estimate of the binomial parameter π for the variable foreign** |
| **Confidence Interval for Binomial π, *"immediate"***<br>**cii proportions *n observedproportion*, level(#)** | **.cii proportions 74 .2973, level(90)**<br>**This produces an exact 90% confidence interval estimate of the binomial parameter π for an UNNAMED variable** |
| **Exact test for Binomial π**<br>**bitest *variable=nullpi***<br>**The option level( ) is NOT allowed** | **.bitest foreign=.28**<br>**This produces an exact test of significance of the null hypothesis that the binomial parameter π = .28 for the variable foreign** |
| **Exact test for Binomial π, *"immediate"***<br>**bitesti *n #successes nullpi***<br>**The option level( ) is NOT allowed** | **.bitesti 74 22 .28**<br>**This produces an exact test of significance of the null hypothesis that the binomial parameter π = .28 for an UNNAMED variable in the setting where N=74, # successes = 22 and the null hypothesis that π = .28** |
| **Normal Approximation test for Binomial π**<br>**prtest *variable=nullpi*, level(#)**<br><br>**Normal Approximation test for Binomial π, *"immediate"***<br>**prtesti *n #successes nullpi*, count level(#)** | **.prtest foreign=.28, level(95)**<br>**This produces a normal aproximation test of significance of the null hypothesis that the binomial parameter π = .28 for the variable foreign. The output includes a 95% confidence interval estimate of π.**<br><br>**.prtesti 74 22 .28, count level(95)**<br>**This produces a normal approximation test of significance of the null hypothesis that the binomial parameter π = .28 for an UNNAMED variable in the setting where N=74, # successes = 22 and the null hypothesis that π = .28. The output includes a 95% confidence interval estimate of π.** |

# 2. Paired Sample Inference

***How to follow along:***
**. clear**
**. sysuse bpwide**

## 2.1 Nonparametric Tests of Paired Medians

**Tip** – Two tests are provided here.

(1) **signrank** - Use for paired outcomes measured on an ordinal scale.
(2) **signtest** - Use for paired outcomes measured on a nominal scale.

| Command | Example |
|---|---|
| ***Ordinal data …***<br><br>**Paired Data Wilcoxon Signed Rank Test of Equal Medians**<br>**signrank** ***var1=var2***<br>**The option level( ) is NOT allowed** | **.signrank bp_before=bp_after**<br>**This produces a paired data Wilcoxon Signed Rank test of equality of medians** |
| ***Nominal data …***<br><br>**Paired Data Sign Test of Equal Medians**<br>**signtest** ***var1=var2***<br>**The option level( ) is NOT allowed** | |

## 2.2 Continuous Outcome – Paired Means Under Normality

| Command | Example |
|---|---|
| **Paired t-test for Mean**<br>**ttest** *var1==var2,* **level***(#)*<br><br>**Tip – Note the requirement of TWO equal signs.** | **.ttest bp_before==bp_after, level(99)**<br>**This produces a paired t-test of the null hypothesis that the mean of bp_before equals the mean of bp_after. The output includes three 99% confidence intervals: (1) for bp_before (2) bp_after (3) difference** |

## 2.3 Continuous Outcome – Paired Variances Under Normality

| Command | Example |
|---|---|
| **Paired Data Test of Equal Variances**<br>**sdtest** *var1=var2*<br>**NOTE –This will produce an Unpaired comparison of the variances using Levene's test, thus disregarding the paired-ness of the data. Stata does have a test of equality of variances for paired data. The command is sdpair and must be installed from the internet** | **.sdtest bp_before=bp_after**<br>**This tests the equality of variances of bp_before and bp_after, as if the data were UNpaired** |

# 3. Two Independent Samples Inference

***How to follow along.***
**. clear**
**. sysuse auto**

## 3.1 Nonparametric Test of Two Independent Medians: Rank Sum Test

**Tip -** The nonparametric test of equality of two independent medians goes by multiple names. All are referring to the same thing:

- Mann Whitney
- Wilcoxon Rank Sum
- Rank Sum

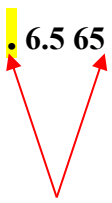
The Stata command to use is the same one: **ranksum**

| Command | Example |
|---|---|
| **2 Sample Rank Sum Test for Equality of Medians**<br>**sort *groupvariable***<br>**ranksum *variable*, by(*groupvariable*)**<br>**The option level( ) is NOT allowed** | **.sort foreign**<br>**. ranksum mpg, by(foreign)**<br>**This produces a Wilcoxon Rank Sum test of the equality of medians of the variable mpg, across the two groups of the variable foreign.** |

## 3.2 Continuous Outcome – Comparison of Two Normal Means

| Command | Example |
|---|---|
| *Assuming Equal Variances …*<br><br>**2 Sample t-test for Equality of Means**<br>sort *groupvariable*<br>ttest *variable,* by(*groupvariable*) level*(#)* | .sort foreign<br>. ttest mpg, by(foreign) level(99)<br>This produces a two-sample t-test of the equality of means of the variable mpg, across the two groups of the variable foreign. The output includes a 99% confidence interval. Variances are assumed equal. |
| *Assuming UNequal Variances …*<br><br>**2 Sample t-test for Equality of Means**<br>sort *groupvariable*<br>ttest *variable,* by(*groupvariable*) unequal level*(#)* | .sort foreign<br>. ttest mpg, by(foreign) unequal level(99)<br>This produces a two sample t-test of the equality of means of the variable mpg, across the two groups of the variable foreign. The output includes a 99% confidence interval. Variances are assumed UNequal. |

## 3.3 Continuous Outcome: Comparison of Two Independent Variances

| Command | Example |
|---|---|
| **2 Sample Test for Equality of 2 Variances**<br>sdtest *variable,* by(*groupvariable*)<br>The option level( ) is NOT allowed<br><br>**2 Sample Test for Equality of Variance,** *immediate*<br>sdtesti *n1* . *sigma1 n2* . *sigma2*<br>The option level( ) is NOT allowed | . sdtest mpg, by(foreign)<br>This produces a test of the equality of variances of the variable mpg, across the two groups of the variable foreign.<br><br>. sdtesti 75 . 6.5 65 . 7.5<br>  <br>Again, take care to provide two periods, this time as placeholders for the two sample mean values. |

## 3.4 Discrete Outcome: Comparison of Two Binomial Proportions

**Recall** - The normal approximation two sample test of equality of independent proportions and the chi square test of association in a 2x2 table are equivalent.

### (a) Two Sample Normal Approximation Test of Equality of Independent Proportions

| Command | Example |
|---|---|
| **Normal Approximation Test for Equality of Two Indpendent Binomial π**<br>**sort** *groupingvar*<br>**prtest** *0/1variable,* **by(***groupingvar***) level***(#)* | **.sort sex**<br>**. prtest cure, by(sex) level(95)**<br>**This produces a normal approximation test of significance of the null hypothesis equality of probability of cure in the two groups defined by sex. The output includes a 95% confidence interval estimate of the difference in the two binomial proportions π.** |
| ***"immediate with n's and observed proportions"***<br>**Normal Approximation test for Binomial π,**<br>**prtesti** *n1 proportion1 n2 proportion2* | **.prtesti 30 .4 45 .67**<br>**In the 1st group: n = 30 % event = .40**<br>**In the 2nd group: n=45 % event = .67** |
| ***"immediate with all counts"***<br>**Normal Approximation test for Binomial π,**<br>**prtesti** *n1 eventcount1 n2 eventcount2,* **count** | **.prtesti 30 12 45 30, count** |

## (b) Two Sample Chi Square Test of Association for a 2x2 Table

| Command | Example |
|---|---|
| **Chi Square Test of Null: Zero Association**<br>tabulate *rowvar colvar,* chi2 *OR*<br>tab *rowvar colvar,* chi2 | . tab drug died, chi2 |
| **Chi Square Test, *immediate***<br>tabi *#11 #12 ..\#21 #22… ,* chi2 | . tabi 1 19\8 6\8 6, chi nolog |
| **All possible Two Way Tests of Null: Zero Association**<br>tab2 *var1 var2 var3 …..,* exact *OR*<br>tab2 *var1 var2 var3 …..,* chi2<br>Use the command tab2 to obtain tests of associations for all pairwise combinations of discrete variables. | |

## 3.5 Fisher's Exact Test of Association for a 2x2 Table

| Command | Example |
|---|---|
| **Fisher's Exact Test of Null: Zero Association**<br>tabulate *rowvar colvar,* exact nolog *OR*<br>tab *rowvar colvar,* exact nolog | . tab drug died, exact nolog<br>Tip! The option nolog suppresses the printing of the enumeration log for Fisher's exact test. |
| **Fisher's Exact Test as an "immediate" command**<br>tabi *#11 #12 ..\#21 #22 ...,* exact | . tabi 1 19\8 6\8 6, exact nolog |

# 4. ONE WAY Analysis of Variance

***How to follow along (beginning Section 4.2):***

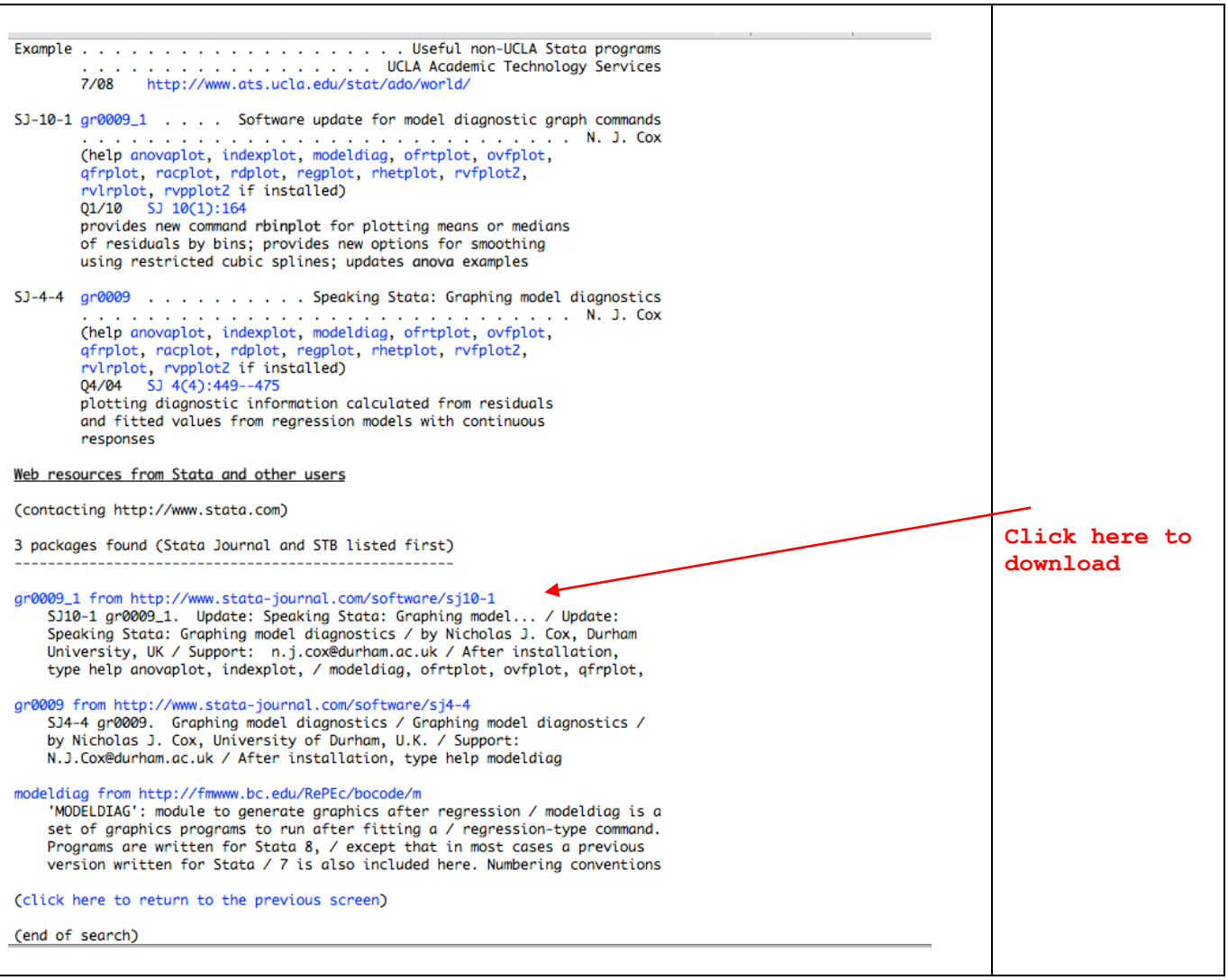
**1) Download from the course website the data set hers_640anova.dta.**
**2) Launch Stata**
**3) From whatever directory is appropriate for you, open hers_640anova.dta**

## 4.1 Nonparametric Test of K Medians – The Kruskal Wallis Test

| Command | Example |
|---|---|
| **K Sample Kruskal Wallis Test of Null: Equality of Medians**<br>**sort *groupvariable***<br>**kwallis *variable*, by(*groupvariable*)**<br>**The option level( ) is NOT allowed** | .<br>**.sort foreign**<br>**. kwallis mpg, by(foreign)**<br>**This produces a kruskal wallis test of the equality of medians of the variable mpg, across the K groups of the variable foreign. When the number of groups K=2, the results are identical to those obtained with the ranksum command.** |

## 4.2 Preliminary: Download ANOVAPLOT

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



## 4.3 Assessment of Normality

| Command | Example |
|---|---|
| **Shapiro-Wilk Test**<br>swilk *variable* | .swilk sbp<br>The null hypothesis is normality. Thus, the assumption of normality is reasonable when the test returns a p-value that is NOT statistically significant. |
| **Shapiro-Francia Test**<br>sfrancia *variable* | .sfrancia sbp<br>The null hypothesis is again normality. Thus, the assumption of normality is reasonable when the test returns a p-value that is NOT statistically significant. |
| **Skewness-Kurtosis Test**<br>sktest *variable* | .sktest sbp<br>The null hypothesis is again normality. Thus, the assumption of normality is reasonable when the test returns a p-value that is NOT statistically significant. |

**Example -**

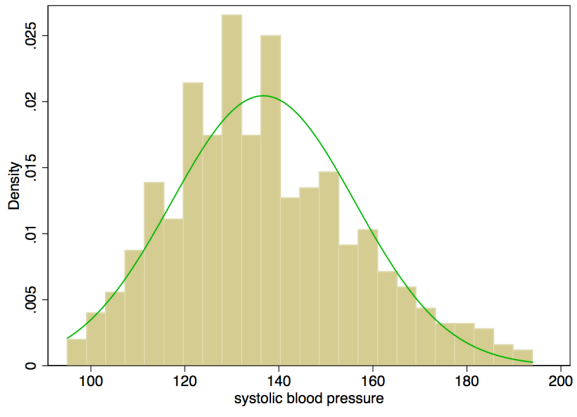
```
. *  Shapiro-Wilk Test (NULL: Distribution is normal)
. swilk sbp
```

Shapiro-Wilk W test for normal data

| Variable | Obs | W | V | z | Prob>z |
|---|---|---|---|---|---|
| sbp | 612 | 0.98098 | 7.681 | 4.945 | 0.00000 |

*KEY: 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 of overlay normal - PLAIN
. histogram sbp, normal
(bin=24, start=95, width=4.125)
```




*This departure from normality is not so severe as to warrant transforming the data. We'll proceed.*

## 4.4. One Way Analysis of Variance Model Estimation

***There are lots of ways to do a one way anova in Stata. Three commands are* oneway, anova. *and* regress**

**Tips if you want to use ONEWAY:**
- If you use **oneway**, then the predictor variable is allowed to be **string** or **numeric**

**Tips if you want to use ANOVA:**
- If you use **anova**, then the predictor variable **must be numeric**.
- Thus, you may need to convert a string variable into a new, numeric variable.

> **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.

**Tips if you want to use REGRESS:**
- If you use **regress**, then you may need to use **design variables** to represent your grouping variable.

**Preliminary look at the codings of the predictor variable 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

------------------------------------------------------------------------
```

## ONEWAY

| Command | Example |
|---|---|
| **sort** ***groupvariable***<br>**oneway** ***yariable groupvariable,*** **tabulate level***(#)*<br><br>**Nice features**:<br>1) Use option **tabulate** to get descriptives<br>2) Use option **level ( )** to set confidence level<br>3) Output includes Bartletts test of equal variances. | **.sort raceth**<br>**. oneway sbp raceth, level(99)**<br>**This produces a one way anova of the equality of the means. Here, the output includes a 99% confidence interval** |

### Example -

```
. 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
```

*Key: Bartlett's test of equality of variance is not statistically significant (p-value = .204). Whew. We don't have to worry about having violated the assumption of equal variances*

# ANOVA

| Command | Example |
|---|---|
| **sort** ***numericgroupvariable***<br>**anova** ***variable numericgroupvariable***<br><br>**Notes**:<br>1) The option **tabulate** is NOT available<br>2) The option **level ( )** is NOT available | **. sort raceth**<br>**. anova sbp raceth**<br>**This produces a one way anova of the equality of the means of the variable sbp, across the k=3 groups of the variable raceth.** |

**Example -**

```
. 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
```

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

# REGRESS

**Note – There are a variety of ways of using the regress command to do a one-way analysis of variance.**

| Command | Example |
|---|---|
| **1st solution: Two steps.**<br>Step 1: Issue the command **anova**<br>Step 2: Issue a post anova command **regress**<br>NOTE: The first group is the referent.<br><br>**anova** ***yvariable numericgroupvariable***<br>**regress** | **. sort raceth**<br>**. anova sbp raceth**<br>**. regress** |

**Example –**

```
.* 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
------------------------------------------------------------------------------
```

**From the coefficients table you can obtain the group means. These match those on page 29.**

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

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

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

| Command | Example |
|---|---|
| **2st solution: Regress with i.groupingvariable**<br><br>**regress *yvariable i.numericgroupvariable***<br><br>or (I prefer the following actually)<br><br>**xi: regress *yvariable i.numericgroupvariable*** | **. regress sbp i.raceth**<br><br>**. xi: regress sbp i.raceth** |

**Example –**

```
. regress sbp i.raceth

      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
------------------------------------------------------------------------------


. xi: regress sbp i.raceth
i.raceth          _Iraceth_1-3        (naturally coded; _Iraceth_1 omitted)

      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]
-------------+----------------------------------------------------------------
  _Iraceth_2 |   2.220612   1.735814     1.28   0.201    -1.188296    5.629519
  _Iraceth_3 |  -.8324823   2.305423    -0.36   0.718    -5.360027    3.695063
       _cons |   136.0133   1.126073   120.79   0.000     133.8019    138.2248
------------------------------------------------------------------------------
```

## 4.5. Test of Equality of Variances

| Command | Example |
|---|---|
| **Solution 1 – Obtain Bartletts test from output of a oneway**<br>**oneway** ***yariable groupvariable***<br><br>Caution: Bartlett's test is sensitive to the assumption of normality. | **.sort raceth**<br>**. oneway sbp raceth** |
| **Solution 2 – robvar**<br>**sort** ***groupvariable***<br>**robvar** ***yvariable,*** **by*****(groupvariable)***<br><br>You will get:<br>W_0 = Levene test<br><br>W_50 = Forsythe-Browne modification of Levene test (mean is replaced by median)<br><br>W_10 = Forsythe-Browne modification of Levene test (mean is replaced by 10% trim) | **.sort raceth**<br>**. robvar sbp, by(raceth)** |

**Example -**

```
. 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
```

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

```
. 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 any of these tests***

## 4.6. Post-hoc Comparisons of Groups

***As with all post-hoc commands in Stata, 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
```

| Command | Example |
|---|---|
| **Pairwise comparison of means – NO ADJUSTMENT for multiple comparisons**<br>**pwcompare *groupvariable*** | **. pwcompare raceth** |
| *** Not recommended!**<br>**Pairwise comparison of means – BONFERRONI ADJUSTMENT**<br>**pwcompare *groupvariable*, mcompare(bonferroni) sort effects** | **. pwcompare sraceth, mcompare(bonferroni) sort effects** |
| *** Must have equal group sample sizes!**<br>**Pairwise comparison of means – TUKEY ADJUSTMENT**<br>**pwcompare *groupvariable*, mcompare(tukey)** | **. pwcompare sraceth, mcompare(tukey)** |

**Example –**

```
. pwcompare raceth

Pairwise comparisons of marginal linear predictions

Margins      : asbalanced

------------------------------------------------------------------------------
                             |                                  Unadjusted
                             |   Contrast   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                      raceth |
African American vs White    |   2.220612   1.735814     -1.188296    5.629519
           Other vs White    |  -.8324823   2.305423     -5.360027    3.695063
Other vs African American    |  -3.053094   2.406646     -7.779427    1.673239
------------------------------------------------------------------------------
```

***Key:***
**African American vs White**
**2.220612 = Mean sbp (African American) - Mean sbp (Whites)**
**= 138.23394 - 136.01333**

```
. 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
------------------------------------------------------------------------------
```

## 4.7. Post-hoc Graphs

***Again, as with all post-hoc commands in Stata, you must have fit the model first using the command*** **anova**

| Command | Example |
|---|---|
| **anovaplot**<br><br>This command produces a side-by-side dot plot with an overlay line that connects the means | **. anovaplot** |

```
. * Example -
. anovaplot
```