

Stata Program Notes
Biostatistics: A Guide to Design, Analysis, and Discovery Second Edition
Chapter 12: Analysis of Variance

Program Note 12.1 - One-Way ANOVA and Multiple Comparisons

The Stata command **oneway** can be used to conduct a one-way analysis of variance (ANOVA). After the **oneway** command is specified, the dependent variable should be the first variable listed followed by the independent variable.

As an example, we analyze the data presented in Table 12.1 to examine if the age distribution of patients who had undergone surgery was similar to members of two control groups. The variable **group** consists of values 1 to 3 where 1 is the surgery group, 2 is the first control group, and 3 is the second control group. The variable **age** contains the ages of the patients in the study.

Stata Command:

```
* Call in the data directly from the website
use http://www.biostat-edu.com/files/Table12-1.dta, clear
list in 1/10
```

Stata Output:

```
+-----+
| group  age |
+-----+
1. |      1   32 |
2. |      1   28 |
3. |      1   22 |
4. |      1   25 |
5. |      1   20 |
+-----+
6. |      1   20 |
7. |      1   28 |
8. |      1   28 |
9. |      1   20 |
10. |      1   29 |
+-----+
```

Below are the Stata commands used to add labels to the values of the variable **group**. Additionally, there are commands used to create boxplots allowing for the visualization of the age distribution across the different groups under study.

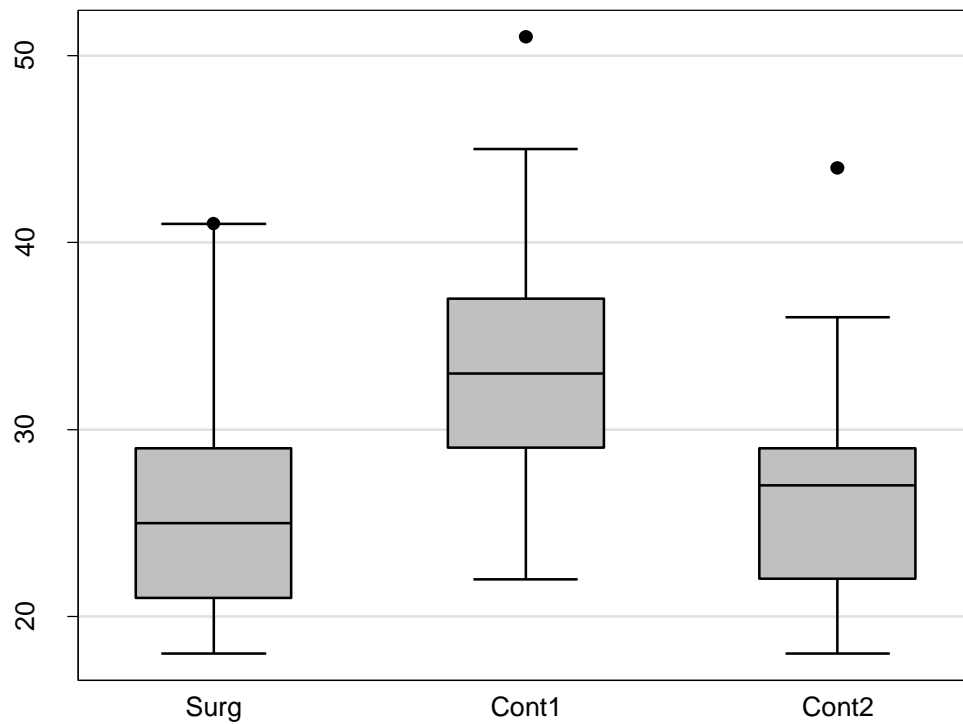
Stata Command:

```
label define grouplab 1 "Surg" 2 "Cont1" 3 "Cont2"
label values group grouplab
```

Stata Command:

```
graph box age, over(group)      /// box plot for each group
  scheme(s1color)              /// change plot scheme
  ytitle(Age (years))         /// y-axis title
  yscale(titlegap(4))         /// creates title gaps
  box(1, bfcolor(gray) lc(black)  /// change color fill of boxes
  marker(1, mcolor(black)))
```

Stata Output: (converting from Word to PDF may not show the y-axis title)



The Stata commands below illustrate the use of the **oneway** command.

Stata Command:

```
oneway age group
```

Stata Output:

Source	Analysis of Variance			F	Prob > F
	SS	df	MS		
Between groups	842.740065	2	421.370033	10.29	0.0001

Within groups	2660.95111	65	40.9377094

Total	3503.69118	67	52.2938982
Bartlett's test for equal variances: chi2(2) = 0.1871 Prob>chi2 = 0.911			

The **tabulate** option produces a table of means, standard deviations, and frequencies for each level of the independent variable.

```
Stata Command:
oneway age group, tabulate
```

```
Stata Output:
```

Summary of Age			
Group	Mean	Std. Dev.	Freq.
Surg	26.08	6.1571097	25
Cont1	33.8	6.7082039	25
Cont2	27.222222	6.2830824	18
Total	29.220588	7.231452	68

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	842.740065	2	421.370033	10.29	0.0001
Within groups	2660.95111	65	40.9377094		
Total	3503.69118	67	52.2938982		

```
Bartlett's test for equal variances: chi2(2) = 0.1871 Prob>chi2 = 0.911
```

Stata reports the results of three different multiple comparison procedures. The procedures are **Bonferroni**, **Scheffe**, and **Sidak**. All of these procedures focus on the family error rate. In the Stata commands below, we request all three multiple comparison procedures.

```
Stata Command:
oneway age group, bonferroni scheffe sidak
```

```
Stata Output:
```

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	842.740065	2	421.370033	10.29	0.0001
Within groups	2660.95111	65	40.9377094		
Total	3503.69118	67	52.2938982		

Bartlett's test for equal variances: $\chi^2(2) = 0.1871$ Prob> $\chi^2 = 0.911$

Comparison of Age by Group
(Bonferroni)

Row Mean-		
Col Mean	Surg	Cont1
-----+-----		
Cont1	7.72	
	0.000	
Cont2	1.14222	-6.57778
	1.000	0.004

Comparison of Age by Group
(Scheffe)

Row Mean-		
Col Mean	Surg	Cont1
-----+-----		
Cont1	7.72	
	0.000	
Cont2	1.14222	-6.57778
	0.847	0.006

Comparison of Age by Group
(Sidak)

Row Mean-		
Col Mean	Surg	Cont1
-----+-----		
Cont1	7.72	
	0.000	
Cont2	1.14222	-6.57778
	0.918	0.004

Program Note 12.2 - ANOVA Continued

1. For Randomized Block with k Replicates per Cell

The **anova** command can be used to conduct a one or two-way analysis of variance (ANOVA). The dependent variable immediately follows the **anova** command and the independent variables are given after the dependent variable.

As an example, we show the Stata commands used to analyze the data presented in Table 12.4. First we present the Stata commands used to add labels to the values of the variables **program** and **site**.

Stata Command:

```
label define proglab 1 "diet" 2 "exer" 3 "both"  
label values program proglab  
label define sitelab 1 "office" 2 "factory"  
label values site sitelab
```

Next, the commands used to provide the results in Table 12.6 are presented.

Stata Command:

```
anova reduct program site
```

Stata Output:

```
Number of obs =      30      R-squared      = 0.3808  
Root MSE      = 4.27725      Adj R-squared = 0.3093
```

Source	Partial SS	df	MS	F	Prob > F
Model	292.5	3	97.5	5.33	0.0054
program	274.866667	2	137.4333333	7.51	0.0027
site	17.63333333	1	17.63333333	0.96	0.3353
Residual	475.666667	26	18.2948718		
Total	768.166667	29	26.4885057		

2. For Balanced Two-Way ANOVA with Interaction

The **anova** command can also be used here to analyze the data shown in Table 12.7. The only difference from the previous analysis is that we include an interaction term. The interaction of the two independent variables is specified by the * (an asterisk) between the two variables.

First we present the Stata commands used to add labels to the values of the variable **method**.

Stata Command:

```
label define methodlab 1 "lecture" 2 "discuss"  
label values method methodlab
```

Next, we present the commands used to provide the results at the bottom of Example 12.8 from the textbook.

Stata Command:

```
tabulate text method, summarize(increase) means
```

Stata Output:

```
Means of increase
```

text	method		Total
	lecture	discuss	
1	23.5	31.333333	27.416667
2	17.333333	27	22.166667
3	23.166667	35.833333	29.5
Total	21.333333	31.388889	26.361111

Finally, we present the commands used to provide the results from including an interaction term in Table 12.9.

Stata Command:

```
anova increase text method text*method
```

Stata Output:

```
Number of obs = 36      R-squared = 0.2936  
Root MSE = 10.165     Adj R-squared = 0.1759
```

Source	Partial SS	df	MS	F	Prob > F
Model	1288.47222	5	257.694444	2.49	0.0529
text	342.722222	2	171.361111	1.66	0.2074
method	910.027778	1	910.027778	8.81	0.0058
text*method	35.7222222	2	17.8611111	0.17	0.8421
Residual	3099.83333	30	103.327778		
Total	4388.30556	35	125.380159		

Program Note 12.3 - General Linear Models Procedure

Stata uses the **glm** command to fit generalized linear models. The **glm** command conducts a linear regression analysis by default. The **regress** command is also used to conduct linear regression analysis, and more information on the **regress** command is presented in the program notes for Chapter 13.