

STATA FOR ONE-WAY ANALYSIS OF VARIANCE

GPA BY SEAT LOCATION EXAMPLE

There are 384 students in the dataset. Y = GPA and there is one categorical variable, “seat” which is a response to the question “Where do you typically sit in a classroom – in the front, middle or back?” We want to know if population mean GPA differs for students who typically sit in the 3 classroom locations.

For one-way ANOVA, you can use the special command “oneway” or you can use the more general command “anova.” Both are illustrated below. If the categorical variable is a “string” variable (i.e. not numerical), before you use the “anova” command, you need to “encode” it to give turn it into a numerical variable. In this example, the “seat” variable was recorded as strings, like “3_Back.” The following command turns it into an integer variable, in what’s called “long” format:

```
encode seat, generate(location)
```

Let’s see what happened, by listing the first few rows of data:

```
list seat gpa location in 1/3
```

```
+-----+
|      seat      gpa   location |
+-----+
1. | 2_Middle      2.6   2_Middle |
2. | 2_Middle      2.7   2_Middle |
3. | 1_Front        3    1_Front  |
+-----+
```

Notice that Stata kept the labels when it created the variable “location” so it’s hard to see that it’s numerical. We can list the variables without the labels attached, which makes it more clear:

```
list seat gpa location in 1/3, nolabel
```

```
+-----+
|      seat      gpa   location |
+-----+
1. | 2_Middle      2.6          2 |
2. | 2_Middle      2.7          2 |
3. | 1_Front        3          1 |
+-----+
```

We can use either “seat” or “location” with the oneway command, but we can use only “location” with the anova command. Here are the results (using “oneway gpa location” would be identical):

oneway gpa seat					
Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	3.99726251	2	1.99863125	6.69	0.0014
Within groups	113.777936	381	.298629753		
Total	117.775198	383	.307507046		
Bartlett's test for equal variances: chi2(2) = 0.8641 Prob>chi2 = 0.649					

Notice the difference between the names given for the “Sources” for the oneway command (above) and the anova command (below).

anova gpa location

Number of obs = 384 R-squared = 0.0339
 Root MSE = .54647 Adj R-squared = 0.0289

Source	Partial SS	df	MS	F	Prob > F
Model	3.99726251	2	1.99863125	6.69	0.0014
location	3.99726251	2	1.99863125	6.69	0.0014
Residual	113.777936	381	.298629753		
Total	117.775198	383	.307507046		

Next, we would like to use the Tukey method to find out which population means are significantly different. Stata doesn't do this automatically, but we can install the "prcomp" software:

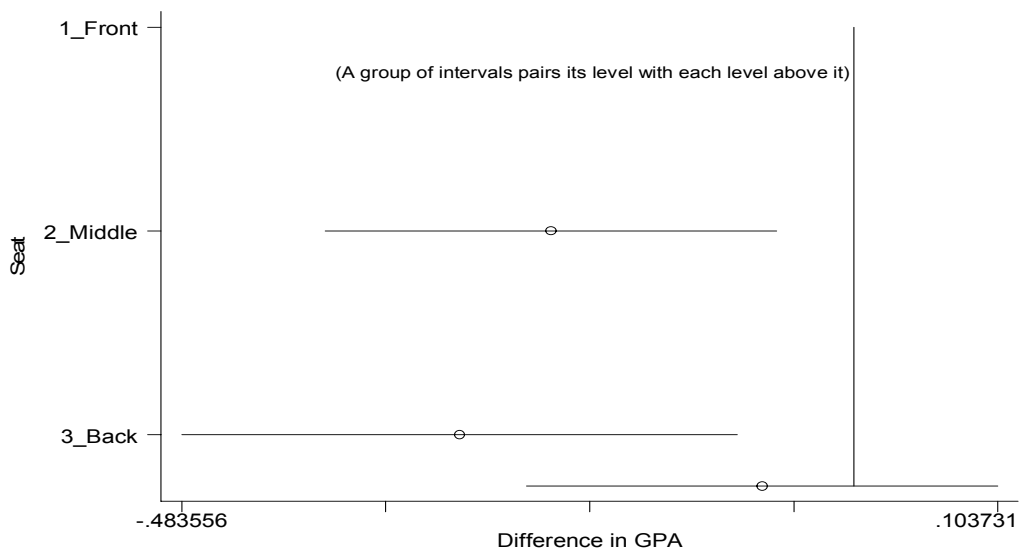
```
findit prcomp
prcomp gpa location, tukey order(m) graph xlin(0)
```

Pairwise Comparisons of Means
 Response variable (Y): gpa GPA
 Group variable (X): location Seat

Group variable (X): location		Response variable (Y): gpa		
Level	Label	n	Mean	S.E.
1	1_Front	88	3.202955	.0585445
2	2_Middle	218	2.985275	.0377666
3	3_Back	78	2.919359	.0577982

Simultaneous confidence level: 95% (Tukey wsd method)
 Homogeneous error SD = .5464703, degrees of freedom = 381

Level (X)	Mean (Y)	Level (X)	Mean (Y)	Diff Mean	95% Confidence Limits	
2_Middle	2.985275	1_Front	3.202955	-.2176793	-.3800729	-.0552858
3_Back	2.919359	1_Front	3.202955	-.2835956	-.4835555	-.0836356
		2_Middle	2.985275	-.0659162	-.2355639	.1037314



Stata will also provide output similar to regression output, with the following command:

```
anova gpa location, regress
```

Source	SS	df	MS	Number of obs = 384		
Model	3.99726251	2	1.99863125	F(2, 381) = 6.69		
Residual	113.777936	381	.298629753	Prob > F = 0.0014		
Total	117.775198	383	.307507046	R-squared = 0.0339		
				Adj R-squared = 0.0289		
				Root MSE = .54647		

gpa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	2.919359	.0618756	47.18	0.000	2.797699	3.041019
location						
1	.2835956	.084983	3.34	0.001	.1165012	.4506899
2	.0659162	.0721003	0.91	0.361	-.075848	.2076805
3	(dropped)					

Notice that it has made the 3rd category (“back”) the reference, and the coefficients for locations 1 and 2 are the additional terms needed to get the average GPA for those two locations. The test of interest is the overall F test, $F^* = 6.69$. That’s the test of whether anything other than the constant is needed.

CELL MEANS PLOT

Especially with two factor ANOVA and higher, it’s useful to plot the “cell means.” Again you need to install software:

```
findit anovaplot
anovaplot, scatter(ms(i))
```

This command follows the anova command, so Stata knows what variables you want to plot. Including “scatter(ms(i))” eliminates the actual data, and just plots the means. This will be more important for two-factor ANOVA when looking for interactions. The right hand side shows an example of a cell means plot with interactions, to be discussed next.

