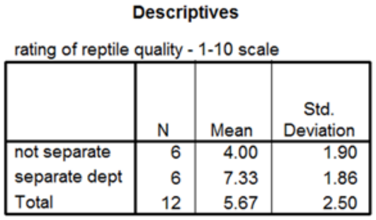
Coding Binary Categorical Variables

Let’s get the 2-group ANOVA to test for reptile quality differences between stores that do and do not have separate reptile departments.

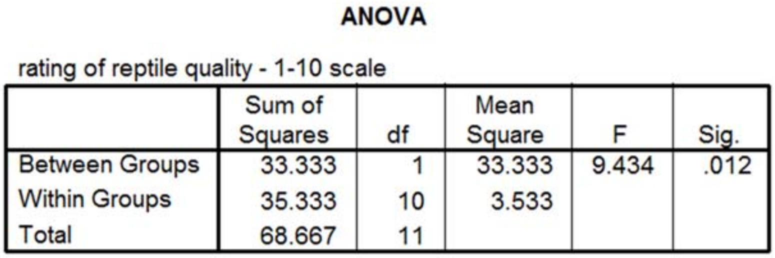
```
 oneway    reptgood by reptdept (1,2).
```



Descriptives

rating of reptile quality - 1-10 scale

	N	Mean	Std. Deviation
not separate	6	4.00	1.90
separate dept	6	7.33	1.86
Total	12	5.67	2.50



ANOVA

rating of reptile quality - 1-10 scale

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	33.333	1	33.333	9.434	.012
Within Groups	35.333	10	3.533		
Total	68.667	11			

1 & 2 Unit Coding of Binary Predictors

We know we can put binary predictors into a regression model. How do those results compare with the ANOVA?

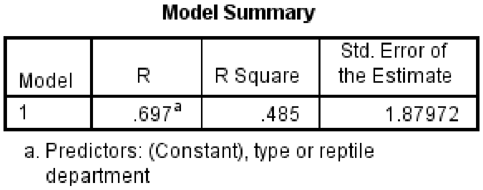
```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT reptgood
  /METHOD=ENTER reptdept.
```

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33.333	1	33.333	9.43	.012[a]
	Residual	35.333	10	3.533		
	Total	68.667	11			

a. Predictors: (Constant), type or reptile department

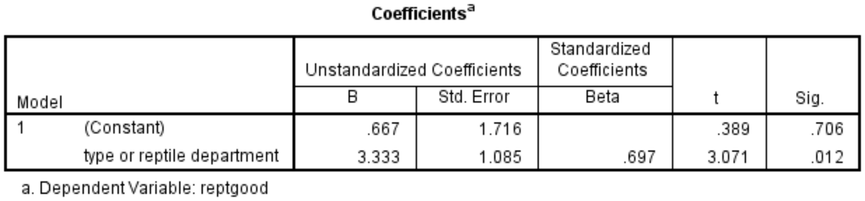
b. Dependent Variable: rating of reptile quality - 1-10 scale



Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.697[a]	.485	1.87972

a. Predictors: (Constant), type or reptile department



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.667	1.716		.389	.706
	type or reptile department	3.333	1.085	.697	3.071	.012

a. Dependent Variable: reptgood

You should notice -- we get exactly the same information from the two analyses

- the F and and p-values are the same from the ANOVA and regression (as are the SS, MS and df)!!
- computing eta-squared from the ANOVA, we get SS_{BG} / SS_{Tot} = 33 .33 / 68.67 = .4854, the R^2!!
- the constant tells the mean of the group coded = 0 ➔ but there isn’t one in this 1 & 2 unit coding!
- the regression weight tells the mean differences between the mean of the roup coded=1 and the group coded=2 ➔ 4.00 + 3.33 = 7.33

As we talked about before, when working with several of the models we’ll be learning soon, it can make things much easier if we have “sensible zeros”. Also, sometimes we will want to “point” a regression model at a particular group of a binary or categorical variable, by setting that group to “0”.

Time to learn about re-coding categorical predictors

Re-coding (or coding) the conditions of a binary variable is simply re-valuing the codes given to specific conditions – a kind of additive linear transformation. The result is to improve the interpretability of the regression weights and their significance tests for binary.

As you know, multiple-category variables cannot be included in a regression or multiple regression model, because the "values" of the variable don't reflect interval quantitative differences among the groups. However, we can re-code multiple category variables so they can be included in multiple regression models.

Keep in mind -- the point here is not that you should use regression instead of ANOVA. This is just the first in learning how to incorporate qualitative variables into regression analyses, so that you can use multiple regression as a truly all-purpose analytic tool!!

A reminder about naming variables…

As with the re-centering of quantitative variables, you may often end up having multiple "versions" of a coded categorical in your data set. So, variable names become increasingly important. Most statistical packages have some capacity to represent the meanings of the condition values for categorical variables (e.g., "Values" in SPSS). However, if you are transferring data across platforms or software packages, often these sorts of ancillary information get dropped! For example, if you export your SPSS .sav data set as an xls file, the Values (and Type, Label, Missing, etc) information is dropped, and stays dropped if you later transfer that xls file back into an SPSS data file!
So, if becomes important to use variable names that carry key details about the variable – like which condition has what value.

0 & 1 Dummy Coding for a Binary Predictor

When making a dummy code for a binary predictor, one of the group of the binary variable is selected at the "comparison group", and receives a code of "0", the other is the “target group” and is coded "1". Syntax to do this a couple of ways is shown below.

Some procedures in SPSS and other packages will create the dummy codes for you. It is important to know what condition is set as the “target=1” and “comparison=0”. In SPSS, the highest coded conditions is set as the “comparison= 0” condition.

The original variable coded: 1=not separate departments (mean=4.00) 2=separate reptile department (mean=7.33)

Using a “Recode” command

```
recode reptdept (1=1) (2=0) into repdep_1not_0sep.
```

🡐 “separate dept” is the comparison=0 group

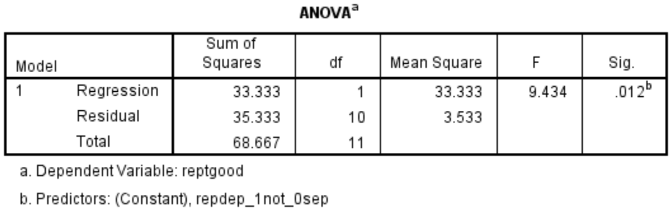
Using a set of “If” command

```
If (reptdept =1) repdep_1not_0sep = 1.
If (reptdept =2) repdep_1not_0sep = 0.
```

🡐 “separate dept” is the comparison=0 group

Regression results using 0 & 1 Dummy Coding for a Binary Predictor

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT reptgood
  /METHOD=ENTER repdep_1not_0sep.
```

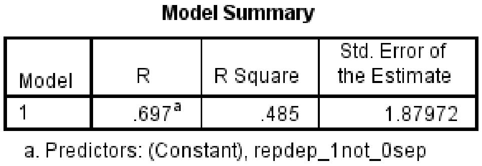


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33.333	1	33.333	9.434	.012[b]
	Residual	35.333	10	3.533		
	Total	68.667	11			

a. Dependent Variable: reptgood

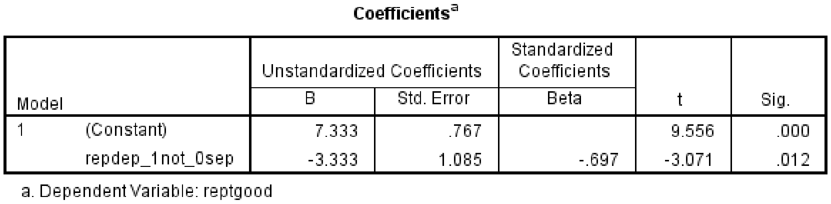
b. Predictors: (Constant), repdep_1not_0sep



Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.697[a]	.485	1.87972

a. Predictors: (Constant), repdep_1not_0sep



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.333	.767		9.556	.000
	repdep_1not_0sep	-3.333	1.085	-.697	-3.071	.012

a. Dependent Variable: reptgood

Using the dummy codes in a regression produces the same model fit and F-test as the ANOVA and the unit-coding regressions above. However, the regressions weight and constant are different..

- The constant tells the mean of the comparison group coded = 0 ➔ mean of the separate department stores = 7.33
- The regression weight tells the mean differences between the comparison group mean and the target group mean ➔ 7.33 + (-3.33) = 4.00

1 * -1 Effect Coding for a Binary Predictor

When we make an effect code for a binary predictor, one of the values of the binary variable is selected at the "comparison group", and receives a code of "-1", the other group is the “target group” and is coded "1". Syntax to do this a couple of ways is shown below.

The original variable coded: 1=not separate departments (mean=4.00) 2=separate reptile department (mean=7.33)

Also, the grand mean was 5.67

Using “Recode” command

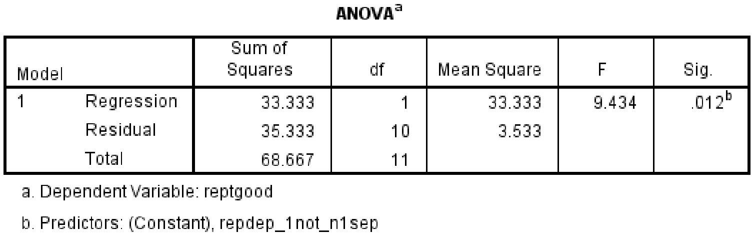
```
recode reptdept (1=1) (2=-1) into repdep_1not_n1sep.          🡐 “separate dept” is the comparison = -1 group
```

Using sets of “If” command

```
If (reptdept =1) repdep_1not_n1sep = 1.                       🡐 “separate dept” is the comparison = -1 group
If (reptdept =2) repdep_1not_n1sep = -1.
```

Regression results using -1 & 1 Effect Coding for a Binary Predictor

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT reptgood
  /METHOD=ENTER repdep_1not_n1sep.
```

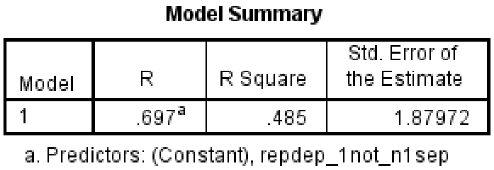


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33.333	1	33.333	9.434	.012[b]
	Residual	35.333	10	3.533		
	Total	68.667	11			

a. Dependent Variable: reptgood

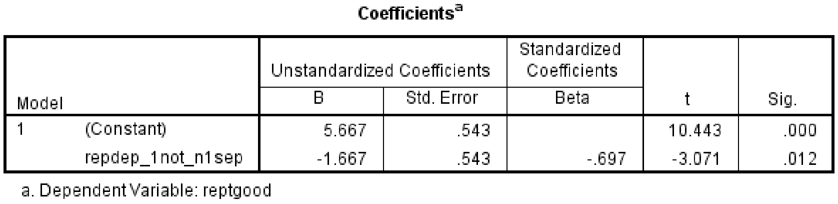
b. Predictors: (Constant), repdep_1not_n1sep



Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.697[a]	.485	1.87972

a. Predictors: (Constant), repdep_1not_n1sep



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.667	.543		10.443	.000
	repdep_1not_n1sep	-1.667	.543	-.697	-3.071	.012

a. Dependent Variable: reptgood

Using the effect codes in a regression produces the same model fit and F-test as the ANOVA, unit-coding, and the dummy coding regressions above. However, the regressions weight and constant are different..

- The constant tells the grand mean mean (the midpoint of -1 & 1 is 0) 🡒 5.667
- The regression weight tells the mean differences between the grand mean and the target group mean
  🡒 5.667 + (-1.67) = 4.00

.5 * -.5 Effect Coding for a Binary Predictor

Another variation of effect coding is to use weights of .5 & -.5. When we make an effect code for a binary predictor, one of the values of the binary variable is selected at the "comparison group", and receives a code of "-.5", the other group is the "target group" and is coded ".5". Syntax to do this a couple of ways is shown below.

The advantage of using these weights is that there is a 1-unit difference between them, and so, the regression weight will tell the mean difference between the comparison group and the target group (instead of telling the difference between the grand mean and the target group mean).

The original variable coded: 1=not separate departments (mean=4.00)   2=separate reptile department (mean=7.33)

Also, the grand mean was 5.67

Using "Recode" command

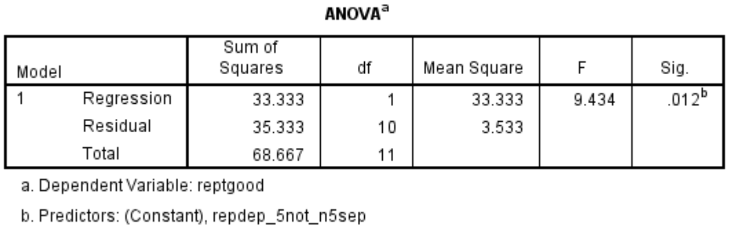
recode reptdept (1=.5) (2=-.5) into repdep_5not_n5sep.   ← "separate dept" is the comparison = -.5 group

Using sets of "If" command

If (reptdept =1) repdep_5not_n5sep = .5.   ← "separate dept" is the comparison = -.5 group
If (reptdept =2) repdep_5not_n5sep = -.5.

Regression results using -.5 & .5 Effect Coding for a Binary Predictor

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT reptgood
  /METHOD=ENTER repdep_5not_n5sep.
```

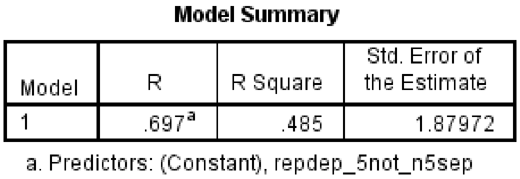


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33.333	1	33.333	9.434	.012[b]
	Residual	35.333	10	3.533		
	Total	68.667	11			

a. Dependent Variable: reptgood

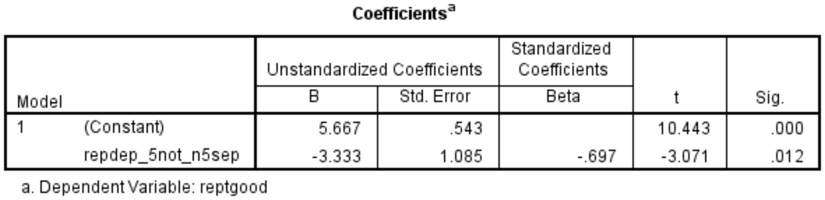
b. Predictors: (Constant), repdep_5not_n5sep



Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.697[a]	.485	1.87972

a. Predictors: (Constant), repdep_5not_n5sep



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.667	.543		10.443	.000
	repdep_5not_n5sep	-3.333	1.085	-.697	-3.071	.012

a. Dependent Variable: reptgood

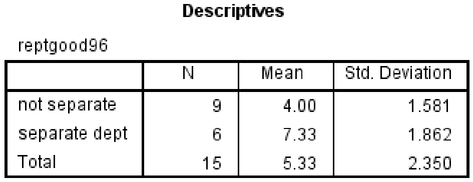
Using the effect codes in a regression produces the same model fit and F-test as the ANOVA, unit-coding, the dummy coding, and the 1 & -1 effect coding regressions above. However, the regressions weight and constant are different..

- The constant tells the grand mean mean (the midpoint of -.5 & .5 is 0) → 5.667
- The regression weight tells the difference between the comparison group mean and the target group mean
  → 7.333 + (-3.333) = 4.00

About using Effect Codes to Compare Groups with Unequal-n

Here are results comparing stores “with separate” and “not separate” reptile departments from a sample with uequal-n.

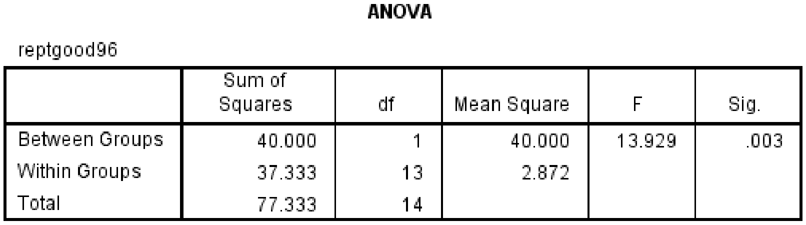
```
ONEWAY reptgood96 BY reptdept
  /STATISTICS DESCRIPTIVES
```



Descriptives

reptgood96

	N	Mean	Std. Deviation
not separate	9	4.00	1.581
separate dept	6	7.33	1.862
Total	15	5.33	2.350



ANOVA

reptgood96

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	40.000	1	40.000	13.929	.003
Within Groups	37.333	13	2.872		
Total	77.333	14			

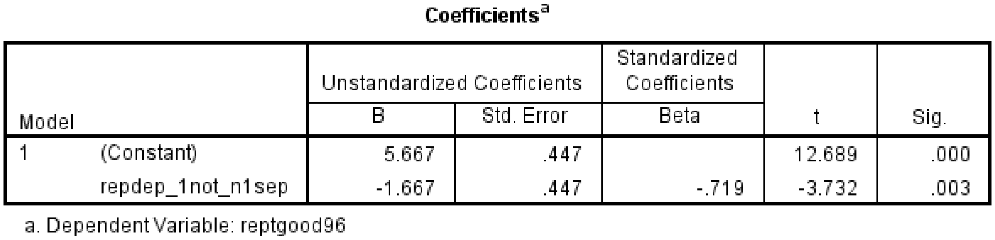
The sample grand mean is 5.33, which is the weighted mean of 4.00 for 9 “not separate” stores and 7.33 for 6 “with separate” reptile departments ➔ ((9 * 4.00) + (6 * 7.33)) / 16 = (36 + 44) / 15 = 5.333.

However, the estimate of the population grand mean would be the unweighted mean of the group means (or the midpoint between the group means) ➔ (4.00 + 7.333) / 2 = 5.667

Why? The estimate of the population mean does not assume that the relative sample sizes of the groups represents the relative population sizes of the groups.

When we apply effect coding to an unequal-n groups comparison, the constant will represent the expected population grand mean (the midpoint or unweighted average of the group means) not the sample grand mean.

Here’s the model obtained using 1 & -1 effects coding



Coefficients[a]

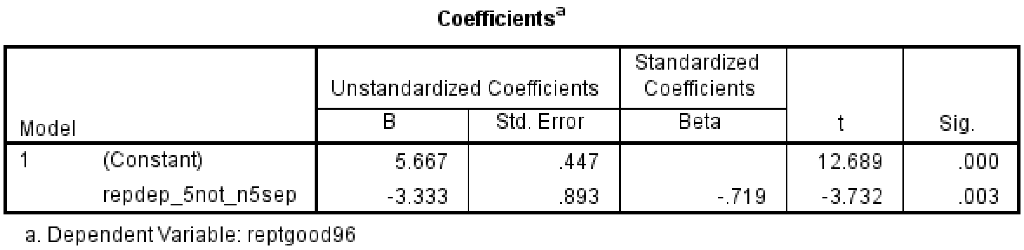
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.667	.447		12.689	.000
	repdep_1not_n1sep	-1.667	.447	-.719	-3.732	.003

a. Dependent Variable: reptgood96

The constant tells the estimate of the population grand mean mean (the midpoint of -1 & 1 is 0) ➔ 5.667

The regression weight tells the difference between the population estimate of the grand mean and the mean of the target group (separate reptile departments = 1) ➔ 5.667 + (-1.667) = 4.00

Equivalent results are obtained if we use -.5 & .5 effects coding.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.667	.447		12.689	.000
	repdep_5not_n5sep	-3.333	.893	-.719	-3.732	.003

a. Dependent Variable: reptgood96

The constant tells the estimate of the population grand mean mean (the midpoint of -.5 & .5 is 0) ➔ 5.667

The regression weight tells the difference between the comparison group mean and the target group mean

➔ 7.333 + (-3.333) = 4.00

Using SPSS GLM with Binary Predictors

In addition to regression, SPSS also offers a GLM procedure that can be used to build models from combinations of quantitative and categorical variables. GLM (UNINOVA) will “do several things for us”, including create coded categorical variables & interactions, as well as perform various kinds of pairwise comparisons.

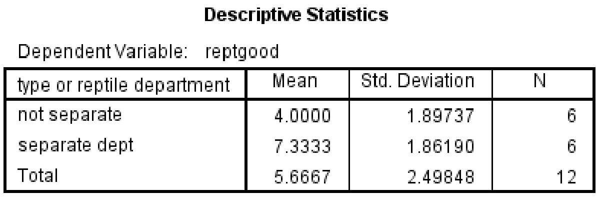
```
UNIANOVA reptgood BY reptdept                          ←  format is “DV”  by “categorical variable IV”
  /METHOD=SSTYPE(3)                                    ← SS to test each effect controlling for all others
  /EMMEANS=TABLES(reptdept) COMPARE(reptdept)          ← gets specific pairwise group comparisons
  /PRINT=DESCRIPTIVE PARAMETER                         ← gets sample descriptives and regression model weights
  /DESIGN=reptdept.                                    ← defines predictors to include in model
```

GLM will code the categorical reptdept variable and include it in the model. GLM actually uses two different codings of categorical variables and presents results for each.

The original coding of reptdept was → 1=not separate department   2=separate reptile department

For the ANOVA & F-test results, categorical variables are effect coded using .5 & -.5 code values. The highest valued group is coded as the comparison group = -.5. The results will be the same as when we recoded using:
recode reptdept (1=.5) (2=-.5) into repdep_5not_n5sep.

For the “parameter estimates,” categorical variables are dummy coded using 1 & 0 values. The highest valued group is coded as the comparison group = 0. The results will be the same as when we recoded using:
recode reptdept (1=1) (2=0) into repdep_1not_0sep.

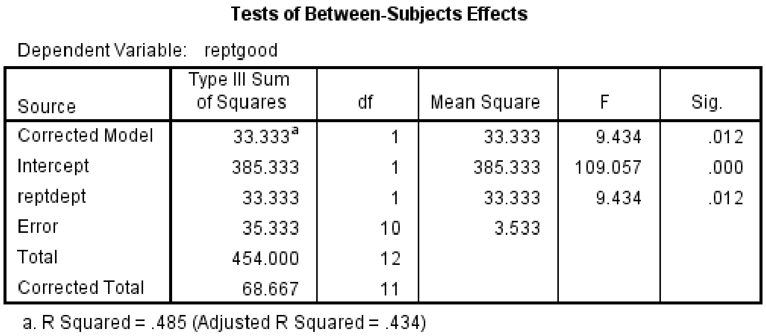


Descriptive Statistics

Dependent Variable: reptgood

type or reptile department	Mean	Std. Deviation	N
not separate	4.0000	1.89737	6
separate dept	7.3333	1.86190	6
Total	5.6667	2.49848	12

“Descriptives” are the sample uinvariate statistics.



Tests of Between-Subjects Effects

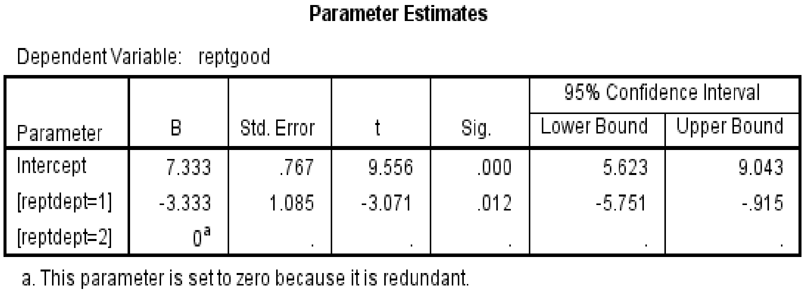
Dependent Variable: reptgood

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	33.333[a]	1	33.333	9.434	.012
Intercept	385.333	1	385.333	109.057	.000
reptdept	33.333	1	33.333	9.434	.012
Error	35.333	10	3.533		
Total	454.000	12			
Corrected Total	68.667	11			

a. R Squared = .485 (Adjusted R Squared = .434)

For this simple model, the ANOVA table provides the same information as the related table in the multiple regression analysis.

However, with more complex models, the GLM ANOVA table will give an F-test for each predictor/effect in the model.



Parameter Estimates

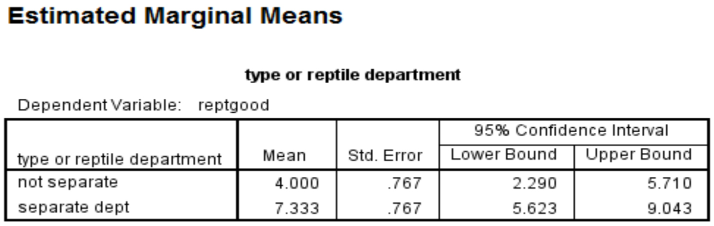
Dependent Variable: reptgood

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	7.333	.767	9.556	.000	5.623	9.043
[reptdept=1]	-3.333	1.085	-3.071	.012	-5.751	-.915
[reptdept=2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

For these parameter estimates, categorical variables are dummy coded, with the highest valued condition coded as the comparison group = 0.

With this dummy coding:
Constant → mean of group coded 0 (separate dept)
Reptdept=1 → difference between comparison group and the target group

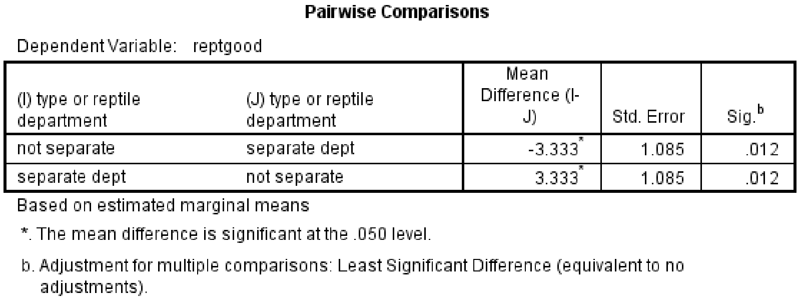


Estimated Marginal Means

type or reptile department

Dependent Variable: reptgood

type or reptile department	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
not separate	4.000	.767	2.290	5.710
separate dept	7.333	.767	5.623	9.043



Pairwise Comparisons

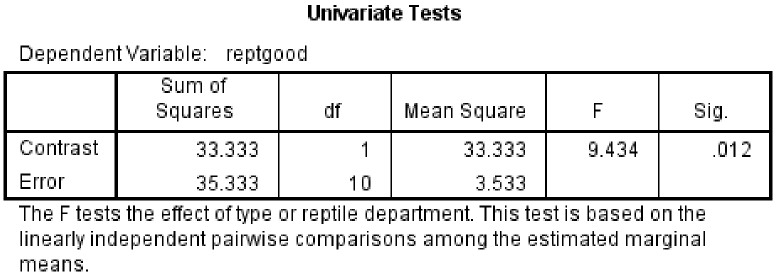
Dependent Variable: reptgood

(I) type or reptile department	(J) type or reptile department	Mean Difference (I-J)	Std. Error	Sig.[b]
not separate	separate dept	-3.333[*]	1.085	.012
separate dept	not separate	3.333[*]	1.085	.012

Based on estimated marginal means

*. The mean difference is significant at the .050 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).



Univariate Tests

Dependent Variable: reptgood

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	33.333	1	33.333	9.434	.012
Error	35.333	10	3.533		

The F tests the effect of type or reptile department. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.

For a very simple design like this, the estimated means, pairwise comparisons and Univariate tests will match the information given in the ANOVA and parameter estimates.

For more complex models, the estimated means will provide important follow-up analyses of categorical variable effects that are “controlled for” other effects in the model.

The pairwise comparison show the mean difference between the groups, and provide a significance test of that difference.

Notice that the mean difference, Std, Error and p-values match those from the “Parameter Estimates” table above

In this simple model, this ANOVA comparison of the two group means is redundant with the pairwise comparison of them just above.

In more complex models, these will provide usefully different pieces of information.