Coding Multiple-Category Categorical Variables

Let's get the 3-group ANOVA to test for fish quality differences between different types of pet stores.

```
ONEWAY fishgood BY storetype
  /STATISTICS DESCRIPTIVES
  /POSTHOC=LSD ALPHA(0.05).
```

ANOVA

rating of fish quality - 1-10 scale

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	36.450	2	18.225	11.538	.003
Within Groups	14.217	9	1.580		
Total	50.667	11			

Descriptives

rating of fish quality - 1-10 scale

	N	Mean	Std. Deviation
chain	5	8.20	.837
coop	4	4.25	1.500
private	3	7.33	1.528
Total	12	6.67	2.146

Multiple Comparisons

Dependent Variable: rating of fish quality - 1-10 scale

LSD

(I) type of store	(J) type of store	Mean Difference (I-J)	Std. Error	Sig.
chain	coop	3.950*	.843	.001
	private	.867	.918	.370
coop	chain	-3.950*	.843	.001
	private	-3.083*	.960	.011
private	chain	-.867	.918	.370
	coop	3.083*	.960	.011

*. The mean difference is significant at the 0.05 level.

There is a significant difference among the three types of stores. The pairwise follow-ups show that chain and privately owned stores have equivalent quality fish, and coops have fish that are lower quality than each of these types of stores. Here are some values to note for later discussion

Groups	Group means		Group mean differences		
			Chain store =1	Coop store =2	Privately owned store =3
Chain store =1	8.20			3.950	.867
Coop store =2	4.25		-3.950		-3.083
Privately owned store =3	7.33		-.867	3.083	

Grand mean (weighted for unequal-n) = 6.667 Group mean average = 6.594

Including Multiple-Category Variables in Regression Models

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.

As with learning to code binary variables for inclusion in multiple regressions, the point here is not that you should use regression instead of ANOVA. You are learning how to incorporate multiple-category qualitative variables into regression analyses, so that you can use multiple regression as a truly all-purpose analysis 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 Multiple-category Predictor

- we need k-1 dummy codes (3 - 1 = 2) for the Storetype variable
- select one of the conditions of Storetype as the "comparison condition"
  - that condition will receive a code of "0" for each dummy code
- the other conditions of Storetype will receive a code of "1" on one dummy code and of "0" on all others

When I created this data set, I chose to code Private stores, the most common type, as "3". That way, each of the other types of stores is compared to this most common type.

```
* original storetype variable  1=chain  2=coop 3=private.
* dummy codes with private store as comparison group.
* first dummy code - chain is the target group.
if (storetype = 1) styp_ch1_co0_p0 = 1.
if (storetype = 2) styp_ch1_co0_p0 = 0.
if (storetype = 3) styp_ch1_co0_p0 = 0.

* second dummy code - coop is the target group.
if (storetype = 1) styp_ch0_co1_p0 = 0.
if (storetype = 2) styp_ch0_co1_p0 = 1.
if (storetype = 3) styp_ch0_co1_p0 = 0.

exe.
```

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT fishgood
  /METHOD=ENTER styp_ch1_co0_p0  styp_ch0_co1_p0.
```

Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.848[a]	.719	1.257

a. Predictors: (Constant), styp_ch0_co1_p0, styp_ch1_co0_p0

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36.450	2	18.225	11.538	.003[b]
	Residual	14.217	9	1.580		
	Total	50.667	11			

a. Dependent Variable: rating of fish quality - 1-10 scale

b. Predictors: (Constant), styp_ch0_co1_p0, styp_ch1_co0_p0

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.333	.726		10.106	.000
	styp_ch1_co0_p0	.867	.918	.208	.944	.370
	styp_ch0_co1_p0	-3.083	.960	-.707	-3.212	.011

a. Dependent Variable: rating of fish quality - 1-10 scale

Using the dummy codes in a regression produces the same model fit and F-test as the ANOVA

- the F and and p-values are the same as from the ANOVA (as are the SS, MS and df)!!
- computing eta-squared from the ANOVA, we get SS_{BG} / SS_{Tot} = 36.450 / 50.667 = .7196, the R^2!!

The regression weights "tell the same story" as the pairwise comparisons from the ANOVA

- The constant tells us the mean of the comparison group coded "0" in both dummy codes (private = 7.33)
- Each dummy code regression weight tells us about the mean difference between the comparison group and the target group of that dummy code
- The t-test of each regression weight tests if those two groups have significantly different means
  - styp_ch1_co0_p0 = .867 tells us that the chain stores have a mean .867 higher than the private stores ➔ 7.333 + .867 = 8.20
  - styp_ch0_co1_p0 = -3.083 tells us that the coop stores have mean 3.083 less than the private stores ➔ 7.333 + (-3.083) = 4.25

Using this set of dummy codes, we do not get a direct test of the mean difference between the private and the coop stores.

-1 **& 1 Effect Coding for a Multiple-category Predictor**

- we need k-1 effect codes (3 - 1 = 2) for the Storetype variable
- select one of the conditions of Storetype as the "comparison condition"
  - that condition will receive a code of "-1" for each effect code
- the other conditions of Storetype will receive a code of "1" on one effect code and of "0" on all others

When I created this data set, I chose to code Private stores, the most common type, as “3”. That way, each of the other types of stores is compared to this most common type.

```
* original storetype variable  1=chain  2=coop 3=private.
* effect codes with private store as the comp group.
* first effect code - chain is the target group.
if (storetype = 1) styp_ch1_co0_pn1 = 1.
if (storetype = 2) styp_ch1_co0_pn1 = 0.
if (storetype = 3) styp_ch1_co0_pn1 = -1.

* second effect code - coop is the target group.
if (storetype = 1) styp_ch0_co1_pn1 = 0.
if (storetype = 2) styp_ch0_co1_pn1 = 1.
if (storetype = 3) styp_ch0_co1_pn1 = -1.

exe.
```

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT fishgood
  /METHOD=ENTER styp_ch1_co0_pn1    styp_ch0_co1_pn1.
```

Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.848[a]	.719	1.257

a. Predictors: (Constant), styp_ch0_co1_pn1, styp_ch1_co0_pn1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36.450	2	18.225	11.538	.003[b]
	Residual	14.217	9	1.580		
	Total	50.667	11			

a. Dependent Variable: rating of fish quality - 1-10 scale

b. Predictors: (Constant), styp_ch0_co1_pn1, styp_ch1_co0_pn1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.594	.371		17.785	.000
	styp_ch1_co0_pn1	1.606	.493	.625	3.258	.010
	styp_ch0_co1_pn1	-2.344	.519	-.866	-4.519	.001

a. Dependent Variable: rating of fish quality - 1-10 scale

Using the effect codes in a regression produces the same model fit and F-test as the ANOVA& dummy code regression.

- the F and and p-values are the same as from the ANOVA and regression (as are the SS, MS and df)!!
- computing eta-squared from the ANOVA, we get SS_{BG} / SS_{Tot} = 36.450 / 50.667 = .7196, the R^2!!

The regression weights “tell the same story” as the pairwise comparisons from the ANOVA

- The constant tells us the expected grand mean (6.594). Notice: that the “expected grand mean” is the mean of the group means, not the weighted grand mean from the ANOVA Descriptives table above.
- Each effect code regression weight tells us about the difference between the expected grand mean and the mean of the target group of that effect code
- The t-test of each regression weight tests if the target group of each effect code has a mean that is significantly different from the expected grand mean.
  - styp_ch1_co0_pn1 = 1.606 tells us that the chain stores have a mean 1.606 higher than the expected grand mean ➔ 6.594 + 1.606 = 8.20
  - styp_ch0_co1_pn1 = -2.344 tells us that the coop stores have mean 2.344 less than the expected grand mean ➔ 6.594 + (-2.344) = 4.25

-.5 **& .5 Effect Coding for a Multiple-category Predictor**

- we need k-1 effect codes (3 - 1 = 2) for the Storetype variable
- select one of the conditions of Storetype as the "comparison condition"
  - that condition will receive a code of "-,5" for each effect code
- the other conditions of Storetype will receive a code of ".5" on one effect code and of "0" on all others

When I created this data set, I chose to code Private stores, the most common type, as "3". That way, each of the other types of stores is compared to this most common type. This is the type of effect coding that SPSS GLM performs.

```
* original storetype variable 1=chain  2=coop 3=private.
* effect codes with private store as the comp group.
* first effect code - chain is the target group.
if (storetype = 1) styp_ch5_co0_pn5 = .5.
if (storetype = 2) styp_ch5_co0_pn5 = 0.
if (storetype = 3) styp_ch5_co0_pn5 = -.5.

* second effect code - coop is the target group.
if (storetype = 1) styp_ch0_co5_pn5 = 0.
if (storetype = 2) styp_ch0_co5_pn5 = .5.
if (storetype = 3) styp_ch0_co5_pn5 = -.5.

exe
```

```
REGRESSION
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT fishgood
  /METHOD=ENTER styp_ch5_co0_pn5  styp_ch0_co5_pn5.
```

Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.848[a]	.719	1.257

a. Predictors: (Constant), styp_ch0_co5_pn5, styp_ch5_co0_pn5

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36.450	2	18.225	11.538	.003[b]
	Residual	14.217	9	1.580		
	Total	50.667	11			

a. Dependent Variable: rating of fish quality - 1-10 scale

b. Predictors: (Constant), styp_ch0_co5_pn5, styp_ch5_co0_pn5

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.594	.371		17.785	.000
	styp_ch5_co0_pn5	3.211	.985	.625	3.258	.010
	styp_ch0_co5_pn5	-4.689	1.038	-.866	-4.519	.001

a. Dependent Variable: rating of fish quality - 1-10 scale

Using these effect codes in a regression produces the same model fit & F-test as the ANOVA & other coded regressions.

- the F and and p-values are the same as from the ANOVA and regression (as are the SS, MS and df)!!
- computing eta-squared from the ANOVA, we get SS_{BG} / SS_{Tot} = 36.450 / 50.667 = .7196, the R^2!!

The regression weights of -.5 & .5 effect coding must be considered carefully!!!

- The constant tells us the expected grand mean (6.594). Notice: that the "expected grand mean" is the mean of the group means, not the weighted grand mean from the ANOVA Descriptives table above.
- However, the effect code regression weights get a bit tricky…
  - Notice that each regression weight is **twice** the difference between the expected grand mean and the mean of the target group of that regression weight
  - styp_ch5_co0_pn5 ➔ 3.211 = 2 * (8.20 – 6.594)
  - styp_ch0_co5_pn5 ➔ -4.680= 2 * (4.25 – 6.594)

Using SPSS GLM with Multiple-Category 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 fishgood BY storetype                          ← format is "DV" by "categorical variable IV"
  /METHOD=SSTYPE(3)                                     ← SS to test each effect controlling for all others
  /EMMEANS=TABLES(storetype) COMPARE(storetype)         ← gets specific pairwise group comparisons
  /PRINT=DESCRIPTIVE PARAMETER                          ← gets sample descriptives and regression model weights
  /DESIGN=storetype.                                    ← defines predictors to include in model
```

GLM will code the multiple-category storetype variable and include it in the model. GLM actually uses two different codings of multiple-category variables and presents results for each.

The original coding of storetype t was → 1=chain 2=coop 3=private.

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:

```
* effect codes with private store as the comp group.
    * first effect code - chain is the target group.
      if (storetype = 1) styp_ch5_co0_pn5 = .5.
      if (storetype = 2) styp_ch5_co0_pn5 = 0.
     if (storetype = 3) styp_ch5_co0_pn5 = -.5.

  * second effect code - coop is the target group.
      if (storetype = 1) styp_ch0_co5_pn5 = 0.
      if (storetype = 2) styp_ch0_co5_pn5 = .5
```

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:

```
* dummy codes with private store as comparison group.
    * first dummy code - chain is the target group.
      if (storetype = 1) styp_ch1_co0_p0 = 1.
      if (storetype = 2) styp_ch1_co0_p0 = 0.
      if (storetype = 3) styp_ch1_co0_p0 = 0.

  * second dummy code - coop is the target group.
      if (storetype = 1) styp_ch0_co1_p0 = 0.
      if (storetype = 2) styp_ch0_co1_p0 = 1.
      if (storetype = 3) styp_ch0_co1_p0 = 0.
```

This odd-appearing combination of codings gives us parallel, but usefully different information from the ANOVA and Parameter Estimates portions of the output, although the really useful aspects of this won't be demonstrable until we get to more complex design with larger mixes of predictor variables (more to come…).

Descriptive Statistics

Dependent Variable: rating of fish quality - 1-10 scale

type of store	Mean	Std. Deviation	N
chain	8.20	.837	5
coop	4.25	1.500	4
private	7.33	1.528	3
Total	6.67	2.146	12

"Descriptives" are the sample uinvariate statistics.

Tests of Between-Subjects Effects

Dependent Variable: rating of fish quality - 1-10 scale

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	36.450[a]	2	18.225	11.538	.003
Intercept	499.634	1	499.634	316.298	.000
storetype	36.450	2	18.225	11.538	.003
Error	14.217	9	1.580		
Total	584.000	12			
Corrected Total	50.667	11			

a. R Squared = .719 (Adjusted R Squared = .657)

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: rating of fish quality - 1-10 scale

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	7.333	.726	10.106	.000	5.692	8.975
[storetype=1.00]	.867	.918	.944	.370	-1.210	2.943
[storetype=2.00]	-3.083	.960	-3.212	.011	-5.255	-.912
[storetype=3.00]	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 (private stores = 3) coded as the comparison group = 0.

- The constant tells us the mean of the comparison group coded "0" in both dummy codes (private = 7.33)
- Each dummy code regression weight tells us about the mean difference between the comparison group and the target group of that dummy code
- The t-test of each regression weight tests if those two groups have significantly different means

    - Storetype=1 (like styp_ch1_co0_p0) = .867 tells us that the chain stores have a mean .867 higher than the private stores ➔ 7.333 + .867 = 8.20

    - Storetype=2 (like styp_ch0_co1_p0) = -3.083 tells us that the coop stores have mean 3.083 less than the private stores ➔ 7.333 + (-3.083) = 4.25

Using this set of dummy codes, we do not get a direct test of the mean difference between the private and the coop stores.

Estimates

Dependent Variable: rating of fish quali

type of store	Mean	Std. Error
chain	8.200	.562
coop	4.250	.628
private	7.333	.726

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.

Univariate Tests

Dependent Variable: rating of fish quality - 1-10 scale

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	36.450	2	18.225	11.538	.003
Error	14.217	9	1.580		

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

The ANOVA tells us there is a difference among the 3 groups. But, like any k-group ANOVA, doesn't tell us between which groups the mean differences are significant and not significant.

Pairwise Comparisons

Dependent Variable: rating of fish quality - 1-10 scale

(I) type of store	(J) type of store	Mean Difference (I-J)	Std. Error	Sig.[b]
chain	coop	3.950*	.843	.001
	private	.867	.918	.370
coop	chain	-3.950*	.843	.001
	private	-3.083*	.960	.011
private	chain	-.867	.918	.370
	coop	3.083*	.960	.011

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).

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