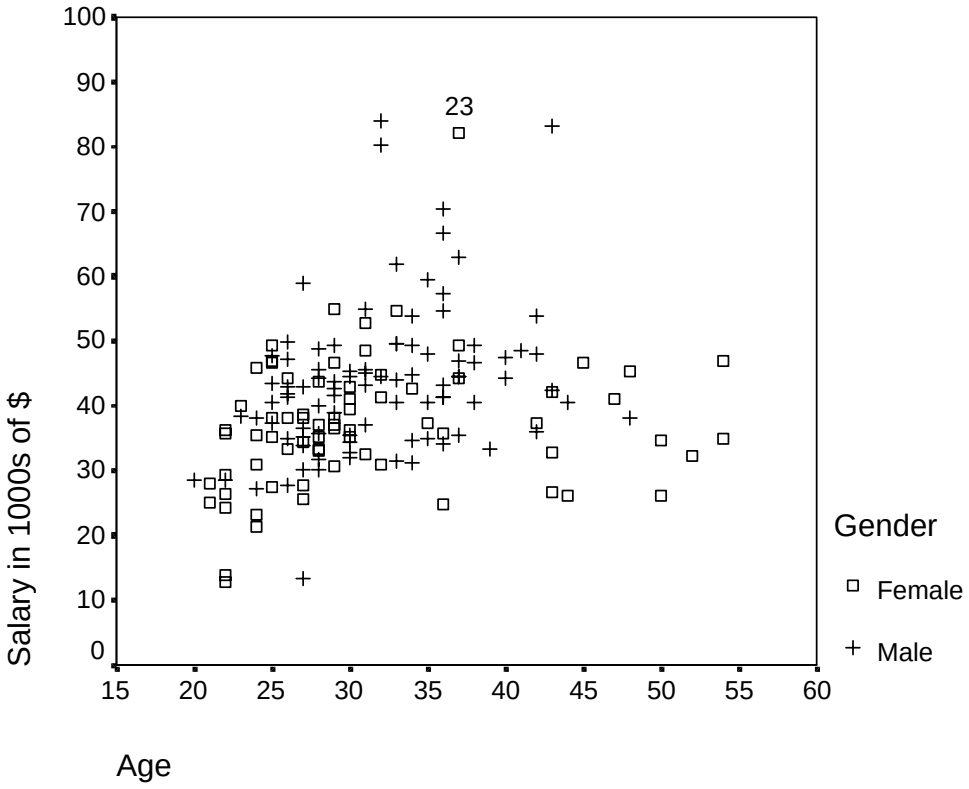
A Framework for Analyses with Numeric and Categorical Dependent Variables

An Exercise in Using GLM

Analyses with Categorical Dependent Variables



Alan Taylor, Department of Psychology, 2001
(Additions 2002-5)

1. When the dependent variable (DV) is an equal-interval numeric variable

Independent variable(s)	Name of procedure	SPSS point-and-click command. *Analyze* ….
a. Groups of different subjects (e.g., *experimental* vs *control*) ["between-subjects analysis"]		
2 groups	independent groups *t*-test	*Compare Means→ Independent-Samples T-Test*
3+ groups	oneway analysis of variance	*Compare Means→One-Way ANOVA* (or *General Linear Model→Univariate*)
More than one grouping factor (e.g., *exptl* vs *control*/ *male* vs *female*)	factorial analysis of variance	*General Linear Model→Univariate*
b. Repeated or matched measures (e.g., *pretest*, *posttest*, *follow-up*) ["within-subjects analysis"]		
2 measures	paired *t*-test	*Compare Means→Paired-Samples T Test*
3+ measures	repeated measures ANOVA	*General Linear Model→ Repeated Measures*
More than one repeated measures factor (e.g., *pre* vs *post*/ *cond1* vs *cond2*)	factorial repeated measures ANOVA	*General Linear Model→Repeated Measures*
c. Both independent groups and repeated measures factors (e.g., *exptl* vs *control*/*pre* vs *post*) ["mixed model analysis"]	mixed factorial ANOVA	*General Linear Model→Repeated Measures*
d. One or more numeric variables (e.g., age, income)	correlation (multiple) regression	*Correlate→Bivariate Regression→Linear*
e. Both idependent groups and numeric variables (includes analysis of covariance)	General linear model	*General Linear Model→Univariate*

2. When the dependent variable (DV) is a categorical variable

Independent variable(s)	Name of procedure	SPSS point-and-click command. *Analyze* ….
a. Groups of different subjects (e.g., e*xperimental* vs *control*)		
2 or more groups	crosstabulation	*Descriptive Statistics→Crosstabs*
More than one grouping factor (e.g., *experimental* vs *control*/ *male* vs *female*)		
2 outcome categories	Logistic regression	*Regression→Binary Logistic*
3+ outcome categories Nominal categories Ordinal categories	Log-linear analysis Multinomial logistic regression Ordinal logistic regression	*Loglinear→Model Selection* *Regression→Multinomial logistic* *Regression→Ordinal*
b. Repeated or matched measures (e.g., *pretest*, *posttest*)		
2 measures, 2 outcome categories	McNemar test	*Nonparametric Tests→2 Related Samples→ McNemar*
3+ measures, 2 outcome categories	Cochran's Q test	*Nonparametric Tests→k Related Samples→ Cochran's Q*
2 measures, 3+ ordinal outcome categories	Test of marginal homogeneity	*Nonparametric Tests→2 Related Samples→ Marginal Homogeneity*[*]
c. Both independent groups and repeated measures factors (e.g., *experimental* vs *control*/*pre* and *post*)	Not easily performed in SPSS. Programs like Stata and SAS, which have procedures based on the Generalised Estimating Equations (GEE) model, may be useful.	
d. One or more continuous variables (e.g., age, years of education) **alone or with categorical variables**	(1) The logistic regression methods given in (a) above can include numeric variables as predictors. (2) If the categorical DV has two categories and the IV is one numeric variable, use the Pearson correlation (*Correlate→Bivariate→Pearson*). This is called a *point-biserial* correlation. (3) If the categorical DV is ordinal, and the IV is a numeric variable, use rank correlation (*Correlate→Bivariate→Spearman*). (4) Another approach is to "switch" the IV and DV and to use a oneway ANOVA.	

[*] Not available in the SPSS Gradpack.

An Exercise in Using GLM

Introduction

This exercise gives point-and-click and syntax instructions for a number of analyses given in the framework using the SPSS GLM procedure. The dataset used is *workmot4.sav*, which is *workmot.sav* (the dataset used in Exercise 2 of *Introduction to SPSS for Windows*) but with the scales *perf* and *who* and the recoded version of *salary*, *salrec*, added. For the analyses described in this handout, a continuous version of *salary*, *salcont* (annual salary in 1000s of dollars), cobbled up from the original categorical *salary* variable, and a recoded version of *level*, *levrec* (senior management combined with middle management), have also been added. The dataset is available at *http://online.mq.edu.au/pub/PSYSTAT/download.htm*.

The Analyses

1. Between-subjects (independent groups)

- The dependent variable for these analyses is *salcont*, annual salary in 1000s of dollars.
- The independent variable (factor) for the one-way ANOVA is *levrec* (which shows the level of the respondent in the company: worker [0], junior manager [1], middle or higher management [2]).
- The second factor (used in the factorial example) is *sex* [1: males, 2:female], the gender of the respondent.

(a) One-way ANOVA

Basic Analysis

Analyze➔*General Linear Model*➔*Univariate*
Select *salcont* as *Dependent Variable*
Select *levrec* as *Fixed Factor*
[Optional extras -- see below]
Click on *OK*

Syntax:

```
glm salcont by levrec.
```

Find in the output:

- *p*-value for *levrec*: .000 (report as $p < .001$ or $p < .0005$)
- R^2 for model: .210 (proportion of variance accounted for)

Optional Extras

Note: The point-and-click instructions for extras are additions to the instructions given for the Basic Analysis. They should be performed before clicking the final *OK* in the Basic Analysis.

Graph of Means

Plots➔Select *levrec* as *Horizontal Axis*➔*Add*➔*Continue*

Post-hoc pairwise comparisons of groups

Post Hoc➔Select *levrec* for *Post Hoc Tests*➔Select one or more post hoc tests (for example Bonferroni)➔*Continue*

Contrasts

Contrasts➔Select *levrec*➔Select *simple* in *Contrast* slot➔Select *first* as *Reference Category*➔Click on *Change*➔*Continue*

Syntax:

```
glm salcont by levrec/
contrast(leverec)=simple(1)/
posthoc=levrec(bonferroni)/
plot=profile(levrec).
```

Find in the output:

- Significance values for Contrast Results of *levrec* = 1 and *levrec* = 2 versus *levrec* = 0
- Significance values of Pairwise Comparisons
- Note that the two contrasts are the same as two of the pairwise comparisons. You wouldn't normally do both. Why are the *p*-values different for the two tests ?

(b) Factorial ANOVA

Basic Analysis

Analyze➔*General Linear Model*➔*Univariate*
Select *salcont* as *Dependent Variable*
Select *levrec* and *sex* as *Fixed Factors(s)*

[Optional extras -- see below]
Click on *OK*

Syntax:

```
glm salcont by levrec sex.
```

Note in output:

- *p*-value for *levrec * sex* interaction: .108 (ns)

Optional Extras

Graph of Means

Plots➔Select *levrec* as *Horizontal* Axis and *sex* as *Separate Lines*➔Add➔*Continue*

Syntax:

```
glm salcont by levrec sex/
plot=profile(levrec*sex).
```

Main Effects Model

Note: The main effects model excludes the interaction term. It should be tested if the interaction is not significant.

Model➔*Custom*➔Select *levrec(F)* and *sex(F)* in the *Factors and Covariates* window
Select *Main effects* in the *Build Term(s)* slot➔click on the *Build Term(s)* arrow➔ *Continue*

Syntax:

```
glm salcont by levrec sex/
plot=profile(levrec*sex)/
design=levrec sex.
```

Find in the output:

- *p*-value for *sex*: .005 (Compare this with the result obtained when the interaction term was in the model. Why is it different?)

2. Within-subjects (repeated measures)

- The dependent variables for these analyses (which define the levels of the within-subject factor *type*) are *perf* and *who*. These variables were created in Exercise 2 of *An Introduction to SPSS for Windows*. They measure the opinions of respondents about the importance of hard work and ability (*perf*) and good luck and who you know (*who*) as ways of getting on in the company.

(a) One-way ANOVA

Basic Analysis

Analyze➔*General Linear Model*➔*Repeated Measures*
Enter the word *type* into the *Within-Subject Factor Name* slot (replacing *factor1*)

Enter *2* into the *Number of Levels* slot
Click on *Define*
Click on *perf* in the variable list➔Click on the arrow pointing to the *Within-Subjects Variables* window
Click on *who* in the variable list➔Click on the arrow pointing to the *Within-Subjects Variables* window
[Optional extras -- see below]
Click on *OK*

Note: As this repeated measures ANOVA has only two levels in the within-subjects factor, it is equivalent to a paired *t*-test. You might like to verify this by carrying out the *t*-test.

Syntax:

```
glm perf who/
wsfactor type 2.
```

Find in the output:

- *p*-value for *type*: .032. What does this tell us?
- The output has both a *Multivariate Tests* table and a *Tests of Within-Subjects Effects* table. In this analysis, the *p*-value is the same in both tables. When there are more than two within-subject variables, this will not be the case.

Optional Extras

Graph of Means

As in 1(a) above.

Post-hoc pairwise comparisons of groups

Not available for within-subject factors.

Contrasts

As in 1(a) above.

(b) Factorial ANOVA

More than one within-subjects factor can be specified (but not in this dataset). After clicking *Define*, you will be asked to specify the variables which represent the value of the dependent variable for each combination of factors. Say in the present dataset we had measures of *perf* and *who* at three times: pretest, posttest and followup. There would then six variables, for example *perf1*, *perf2*, *perf3*, *who1*, *who2* and *who3*. These are the variables which would be specified, in the appropriate order, for GLM, and the factors would be *type* (2 levels) and *time* (3 levels).

3. Mixed Between- and Within-subjects analyses

- The dependent variables for these analyses are as given in 2. above
- The between-subject factor is *levrec* (which shows the level of the respondent in the company: worker, junior manager, middle or higher management).

Basic Analysis

Analyze➔*General Linear Model*➔*Repeated Measures*
Type *type* into the *Within-Subject Factor Name* slot (replacing *factor1*)
Enter *2* into the *Number of Levels* slot
Click on *Define*
Click on *perf* in the variable list➔Click on the arrow pointing to the *Within-Subjects Variables* window
Click on *who* in the variable list➔Click on the arrow pointing to the *Within-Subjects Variables* window
Select *levrec* for the *Between-Subjects Factor*
[Optional extras -- see below]
Click on *OK*

Syntax:

```
glm perf who by levrec/
wsfactor type 2.
```

Find in the output:

- *p*-value for the *type***levrec* interaction: .571 (ns) [This is a test of whether any differences between *perf* and *who* differ for different levels of *levrec*.]
- *p*-value for *Tests of Between-Subjects Effects*: .431 (ns) [This is a test of whether there is a difference between levels of *levrec* for the average of *perf* and *who*.]

Optional Extras

Graph of Means

As in 1(b) above. It is usual to specify the within-subject factor on the horizontal axis and the between-subject factor as separate lines.

Post-hoc pairwise comparisons of groups

Available for the between-subject factor (*levrec*).

4. Numeric Variables as Predictors ("regression")

- The dependent variable for these analyses is *salcont*, annual salary in 1000s of dollars.
- The independent variable (predictor) in the bivariate regression is *age* in years.
- The second independent variable in the multiple regression is *sex* (coded 1: male and 2: female). This is included to show that a categorical variable with two categories can be entered into GLM as either a covariate or a factor. The same results are obtained in both

cases. A series of (0,1) variables may be used as dummy coding to represent a categorical variable with more than two categories.

(a) Bivariate regression

Analyze➔General Linear Model➔Univariate
Select *salcont* as *Dependent Variable*
Select *age* as *Covariate*
[Optional extras – see below]
Click on *OK*

Syntax:

```
glm salcont with age.
```

Note in output:

- *p*-value for *age*: .001
- R^2 for model: .065 (proportion of variance accounted for)

Optional Extras

Regression Coefficients

Options➔Parameter estimates➔Continue

Syntax:

```
glm salcont with age/
print=parameter.
```

Note in output:

- Regression coefficient for *age*: .412. This implies that for each extra year of age, salary increases by approximately $400.

(a) Multiple regression

Analyze➔General Linear Model➔Univariate
Select *salcont* as *Dependent Variable*
Select *age* as *Covariate*
Select *sex* as *Covariate*
[Optional extras -- see below]
Click on *OK*

Syntax:

```
glm salcont with age sex.
```

Find in the output:

- *p*-value for *age* adjusted for *sex*: .001 (i.e., it hasn't changed much).
- *p*-value for *sex* adjusted for *age*: .000 (report as <.001 or .0005)
- R^2 for both variables combined: .143

Optional Extras

Regression coefficients

As above.

- Regression coefficient for *sex*: -6.482. A one-unit increase in *sex* (i.e., going from male (=1) to female (=2) leads to a $6500 decrease in salary.
- Regression coefficient for *age*: .389. Adjusting for *sex* has decreased the effect of *age* on salary, but not by much.

Testing the interaction of *age* and *sex*

Note: This asks whether the relationship between *age* and *salcont* is different for males and females (OR, whether the difference between males and females is different at different ages).

Model➔*Custom*➔Select *age(C)* and *sex(C)* in the *Factors and Covariates* window
Select *Main effects* in the *Build Term(s)* slot➔click on the *Build Term(s)* arrow
Select *age(C)* and *sex(C)* in the *Factors and Covariates* window
Select *Interaction* in the *Build Term(s)* slot➔click on the *Build Term(s)* arrow➔*Continue*
Options➔*Parameter estimates*➔*Continue*

Syntax:

```
glm salcont with age sex/
print=parameter/
design=age sex age*sex.
```

Find in the output:

- *p*-value for interaction of *age*sex*: .082 (ns, but approaching it)
- Regression coefficient for *age*sex*: -.432. This implies that the slope of the relationship between *age* and *salcont* is .432 less steep for females than it is for males. In other words, that salary does not increase with age as much for women as it does for men.
- Regression coefficient for age: 1.103. This is the effect of age for *sex* = 0. Not very helpful here, since *sex* is coded (1,2). It would be a good idea to recode *sex* to (0,1) and then the coefficient would show the effect of age for whichever sex was coded 0.
- Regression coefficient for *sex*: 7.247. This is the effect of age for *age* = 0. Not very helpful here, because no respondents were aged 0 (or even near it!). It would be a good idea to centre *age* (by subtracting the mean of *age* from each value of *age*) and then the coefficient would show the effect of gender for the mean age.
- You may want to try the suggestions above. The mean of *age* is 31.92. The syntax is as follows:

```
temporary.
compute age = age – 31.92.  [this is called centring]
compute sex = sex -1.       [the code is now male: 0, female: 1]
glm salcont with age sex/
print=parameter/
design=age sex age*sex.
```

Specifying sequential sums of squares

By default GLM uses Type III sums of squares, which are equivalent to *unique* in *manova* (each term is adjusted for every other term). If you would like to use sequential sums of squares in GLM, click on *Model* when specifying the analysis, and choose Type I in the *Sum of Squares* slot at the bottom of the display.

Syntax:

```
glm salcont with age sex/
  method=sstype(1).
```

- The result is the same as before for *sex*, but not for *age*? Why?

5. Both Categorical and Numeric Variables

For this sort of analysis, both *Fixed Factor(s)* and *Covariate(s)* are specified. The analyses are pretty much as would be expected, with a few exceptions:

- GLM automatically assumes that you want to test every interaction between variables specified as *Fixed Factor(s)* but that you don't want to test any interactions between variables specified as *Covariates(s)* or between variables specified as *Fixed Factor(s)* and those specified as *Covariates(s)*. If you want to test interactions between covariates or between covariates and factors, you need to click on *Model* then *Custom* and specify the terms you do want in the model.
- Plots cannot be requested for numeric predictors. For example, GLM won't allow you to produce a scatterplot with *salcont* on the *y*-axis and *age* on the *x*-axis.

As a final exercise, run the following analyses using either syntax or point-and-click:

	Model 1	Model 2	Model 3
Dependent variable	*salcont*	*salcont*	*salcont*
Factors	*levrec, sex*	*levrec, sex*	*levrec, sex*
Covariate	*age*	*age*	*age*
Interactions	*levrec*sex*	None	*leverec*age, sex*age*

Analyses With Categorical Dependent Variables

Introduction

This exercise gives SPSS point-and-click and syntax instructions for a number of analyses in which the dependent variable is categorical. The dataset used is *workmot4.sav*, which is *workmot.sav* (the dataset used in Exercise 2 of *Introduction to SPSS for Windows*) but with the scales *perf* and *who* and the recoded version of *salary*, *salrec*, added. A recoded version of *level*, *levrec* (senior management combined with middle management), has also been added. The dataset is available at *http://online.mq.edu.au/pub/PSYSTAT/download.htm*. It it also in the Share/Psych directory on the Student Server (see *http://online.mq.edu.au/pub/PSYSTAT/share.htm*).

The Analyses

1. Between-group comparisons of subjects (independent groups)

- The dependent variable for these analyses is *salrec*, annual salary in categorical form: 2: up to $30K, 3: >$30K-$40K, 4: >$40K-$50K, 5: >$50K.
- The independent variable (factor) for the two-group analysis is *sex*, 1: male, 2: female. The second factor (used in the three-way table example) is *levrec*, further recoded so that all management subjects are in one group.

(a) Two-way table

Basic Analysis

Analyze➔*Descriptive Statistics*➔*Crosstabs*
Select *sex* as *Row* variable
Select *salrec* as *Column* variable
Cells➔Select *Observed* in *Counts* in (probably already selected) and *Row* in *Percentages* ➔*Continue*
Statistics➔*Chi-square*➔*Continue*
[Optional extras -- see below]
Click on *OK*

Syntax:

crosstabs sex by salrec/cells=count row/statistics=chisq .

Find in the output:

- The percentage of males and females in each category of *salrec*. What do they tell us about the salaries of females compared to those of males?
- Pearson chi-squared and *p*-value. Report as $\chi^2(3) = 20.2$, $p < .0005$. Sex and salary are clearly related, i.e., not independent.
- The number of cells with expected cell frequency of less than 5. Rules of thumb say that if more than a few cells have expected frequencies less than 5, the results may be questionable. (The number of cases which are expected in a cell [say the cell which is in row *R* and column *C*] if the two variables [*sex* and *salrec* in this case] are independent [unrelated] is equal to (row *R* total x column *C* total)/total number of cases. It is the discrepancy between the observed and expected numbers which are the basis of the chi-squared test. See *http://online.mq.edu.au/pub/PSYSTAT/chi2.htm* for further details.)

Optional Extras

Note: The point-and-click instructions for extras are additions to the instructions given for the Basic Analysis. They should be performed before clicking the final *OK* in the Basic Analysis.

Expected frequencies and adjusted standardised residuals

Cells➔*Expected* in *Counts* and *Adj. standardized* in *Residuals*➔*Continue*

Syntax:

crosstabs sex by salrec/cells=count row expected asresid/statistics=chisq.

Find in the output:

- The expected frequencies, calculated as described above. Compare the expected and observed frequency in each cell.
- The adjusted standardised residuals show, for each cell, the magnitude of the difference between observed and expected frequencies. Residuals with an absolute value greater than two suggest a significant difference for that cell.

(b) Three-way table

The first analysis showed that female employees earned less than male employees. One reason for this could be that (1) females are more likely to be "workers" than males, and that (2) "workers" receive lower salaries. Two two-way tables, *sex* by *levrec* and *levrec* by *salrec*, show that (1) and (2) are both true (you can run *crosstabs* to obtain the appropriate tables and verify that this is the case).

To evaluate the above explanation, we'll obtain a separate table of *sex* by *salrec* for each of two levels of *levrec*, "worker" and "management" (we'll use *recode* to combine the two management groups). If our explanation fully accounts for the difference between male and female salaries, there should be no relationship between *sex* and *salrec* at the separate levels

of *levrec* (because our hypothesis is that differences in *levrec* produce the differences between males and females). In this sort of analysis, we're literally "holding *levrec* constant", or "adjusting for *levrec*".

Before doing the analysis, recode *levrec* so that 2 becomes 1:

Transform➔*Recode*➔*Into Same Variables*
Select *levrec* as the *Variable*
Old and New Values➔Enter *2* as *Old Value, 1* as *New Value*➔*Add*➔*Continue*➔*OK*

The Analysis

Analyze➔*Descriptive Statistics*➔*Crosstabs*
Select *sex* as *Row* variable
Select *salrec* as *Column* variable
Select *levrec* as *Layer 1 of 1*
Cells➔*Observed* in *Counts* (probably already selected) and *Row* in *Percentages*➔*Continue*
Statistics➔*Chi-square*➔*Continue*
Click on *OK*

Syntax:

```
recode levrec (2=1).
crosstabs sex by salrec by levrec/cells=count row/statistics=chisq.
```

Note in the output:

- The percentages of male and female "workers" in each salary category. Do they (and the significant chi-squared for this table) support the suggestion that differences in *levrec* produce the differences between males and females?
- What do the results for the combined management group suggest?

Note: Logistic regression, which can be used for multifactorial analyses with categorical dependent variables, is beyond the scope of this course.

2. Within-subject (repeated measures)

- The dependent variables for two of the following analyses are *satjob* and *satorg*, which measure employees' satisfaction with their jobs and the organisation they work for. The original variables are based on five-point rating scales ranging from 1 (Very dissatisfied) to 5 (Very satisfied). For the analysis which handles two measures and two outcome categories, new variables with the values 1 (Very satisfied) and 0 (the rest) will be created. For the analyses of three or more measures with two outcomes, recoded versions of *d6a* to *d6h* [1=Extremely or Very important, 0=the rest] are used. As will be recalled from Exercise 2 in *Introduction to SPSS for Window*, these variables are the employees' ratings of the importance of various factors for getting ahead in the orgranisation on a five-point scale ranging from 1 (Not important at all) to 6 (Extremely important).

(a) Two measures, two outcome variables (McNemar's test)

Before doing the analysis, use *recode* to create dichotomous versions of *satjob* and *satorg*:

Transform➔*Recode*➔*Into Different Variables*
Select *satjob* and *satorg* as the *Numeric Variables*
Highlight *satjob* and enter *satjobr* as *Output Variable* name. Click on *Change*
Highlight *satorg* and enter *satorgr* as *Output Variable* name. Click on *Change*
Click on *Old and New Values*
Click on *Range*. Enter *1* in the left slot, *4* in the right slot. Enter *0* as *New Value*. Click on *Add*
Click on *Value*. Enter *5*. Enter *1* as *New Value*. Click on *Add*
Click on *Continue*
Click on *OK*.

The Analysis

Analyze➔*Nonparametric Tests*➔2 *Related Samples*
Select *satjobr* and *satorgr* as *Test Pairs*
Select *McNemar* as *Test Type*
Deselect other *Test Types*
Click on *OK*

Note: What we are testing here is whether the proportion of respondents who were "Very satisfied" with their jobs was significantly different from the proportion who were "Very satisfied" with the organisation. If we obtain a one-way *frequencies* distribution for *satjobr* and *satorgr*, we can see that 19.6% were very satisfied with their job while 8.6% were very satisfied with the organisation. The McNemar test will tell us whether the difference between these percentages is statistically significant.

Syntax:

```
recode satjob satorg (1 thru 4=0)(5=1) into satjobr satorgr.

npar tests mcnemar=satjobr satorgr.
```

Find in the output:

- *p*-value: .000. Report as $p < .0005$.
- The table shows that 127 + 11 respondents had the same rating (0 or 1) on both items. It also shows that while only two subjects had 0 for *satjobr* and 1 for *satorgr*, 21 had 1 for *satjobr* and 0 for *satorgr*. It is this difference which reflects the greater proportion who were very satisfied with their job as against the proportion who were very satisifed with the organisation.

(b) Three+ measures, two outcome categories (Cochran's Q test)

Cochran's Q test is like McNemar's test but allows us to test the difference between three or more dichotomous measures.

Before doing the analysis, use *recode* to create dichotomous versions of *d6a* to *d6h*:

Transform➔*Recode*➔*Into Same Variables*
Select *d6a* to *d6h* as the *Variables*
Click on *Old and New Values*
Click on *Range*. Enter *1* in the left slot, *4* in the right slot. Enter *0* as *New Value*. Click on *Add*
[Click on *Range*.] Enter *5* in the left slot, *6* in the right slot. Enter *1* as *New Value*. Click on *Add*
Click on *Continue*
Click on *OK*.

The Analysis

Analyze➔*Nonparametric Tests*➔*K Related Samples*
Select *d6a* to *d6h* as *Test Variables*
Select *Cochran's Q* as *Test Type*
Deselect other *Test Types*
Click on *OK*

Syntax:

```
recode d6a to d6h (1 thru 4=0)(5,6=1).

npar tests cochran=d6a to d6h.
```

Find in the output:

- *p*-value: .000. Report as $p < .0005$.
- The output says "0 is treated as a success". This is an arbitrary choice. It makes no difference to the result whether 0 or 1 is treated as the success.

Hint: Having found a significant overall difference between the proportions of respondents who rated each of the eight factors "Very important" or "Extremely important", you may want to compare each pair of items to see which are different from each other. You could use the commands

```
npar tests mcnemar=d6a to d6h.
```

or the point-and-click equivalent. This would carry out all 8!/([8-2)! 2!) = 28 comparisons. Because of the large number of tests, you might want to Bonferroni-adjust the significance level to .05/28 = .00179.

(c) Two measures, three+ outcome categories (test of marginal homogeneity)

The test of marginal homogeneity is like McNemar's test but allows us to test differences between outcomes with more than two (ordinal) categories. We'll use it to test whether the distribution of subjects' responses on the unrecoded versions of *satjob* and *satorg* are different.

Note: It wouldn't be unreasonable to treat *satjob* and *satorg* as numeric variables, in which case the ratings could be compared using a paired *t*-test. For the purposes of this exercise, we're assuming that we want to treat the variables as categorical.

The Analysis

Analyze➔Nonparametric Tests➔2 Related Samples
Deselect *satjobr* and *satorgr* if necessary
Select *satjob* and *satorg* as *Test Pairs*
Select *Marginal Homogeneity* as *Test Type*
Deselect other *Test Types*
Click on *OK*

Syntax:

```
npar tests mh=satjob satorg.
```

Find in the output:

- *p*-value: .000. Report as $p < .0005$.

Hint: Having found a significant overall difference between the way responses are distributed on the two variables, you may want to do follow-up comparisons between the two variables to track down where the differences lie. You could use *recode* to create dichotomous versions of both variables and compare them using McNemar's test. For example:

```
temporary.
recode satjob satorg (1 thru 3=0)(4,5=1).
npar tests mcnemar=satjob satorg.
```

or the point-and-click equivalent.

3. Mixed Between- and Within-subjects analyses

SPSS does not provide easily-performed tests for mixed analyses involving categorical dependent variables and both within-and between-group factors. However, you can improvise. For example, say you're interested in whether the difference between *satjobr* and *satorgr* (performed in 2(a) above) occurs for both males and females. You could simply perform the analysis separately for males and females. One way of doing this would be to use the *split file* command. This command is equivalent to using a series of temporary *select if* commands, one for each category of the between-subject factor.

The Analysis

[splitting the file]

Data➔Split File➔Compare groups
Select *sex* for *Groups Based on*
Make sure that *Sort the file by grouping variables* is selected
Click on *OK*

[performing the analysis]

Analyze➔*Nonparametric Tests*➔2 *Related Samples*
Select *satjobr* and *satorgr* as *Test Pairs*
Select *McNemar* as *Test Type*
Deselect other *Test Types*
Click on *OK*

[turning *split file* off]

Data➔*Split File*➔*Analyze all cases, do not compare groups*
Click on *OK*

Syntax:

```
sort cases by sex.
split file by sex.
npar tests mcnemar=satjobr satorgr.
split file off.
```

Find in the output:

- *p*-values for males and females of .013 and .004 respectively. It appears from these results, and the tables showing the distribution of responses, that the same pattern of responses occurs for males and females.

Hint: Beware of instances where one of the sub-groups contains a small number of subjects. A non-significant result may be the result of too little power. You need to look at the pattern of results as well as at the significance tests. For instance, if one of the *p*-values in the above test had not been significant, the distribution of results in the accompanying table would nevertheless have made it inappropriate to conclude that there was clearly no difference for one group or the other.

4. Numeric Variables as Predictors

If your dependent variable is categorical, and the predictor is a numeric variable, you could use logistic regression for an analysis. This procedure is beyond the scope of this course. However, you can still assess the association between the two variables by reversing the analysis and treating the categorical variable as the predictor and the continuous variable as the outcome variable. You can then perform an analysis like that in 1(a) of *An Exercise in Using GLM*. A nice way of carrying out such analyses is to use the *means* procedure and to take advantage of the options which allow tests of linear and non-linear relationships.

For example, say we wanted to know if there was any association between the categorical variable *satjob* and the numeric variable *who* (created in Exercise 2 of *Introduction to SPSS for Windows*). In other words, do those who think that who you know and good luck are important factors in getting ahead tend to be more, or less, satisfied with the organisation for which they work?

The Analysis

Analyze➔*Compare Means*➔*Means*
Select *who* in *Dependent List*
Select *satorg* in *Independent List*
Options➔Click on *Anova table and eta* and *Test for linearity*➔*Continue*
Click on *OK*

Syntax:

```
means who by satorg/statistics=anova linearity.
```

Note in the output:

- *p*-value for the overall one-way analysis of variance: .000. Report as $p < .0005$. This result is what you would get with GLM.
- *p*-value for linear part of the relationship: .005. The linear part of the relationship is that part which can be summarised with a straight line.
- *p*-value for non-linear (i.e., curved) part of the relationship: .606. Clearly only a straight line is needed to summarise the relationship between the two variables.
- The measure of (linear) association *R*: -.435. This is negative, which shows that the more strongly someone believes that it's who you know and luck that get you ahead in the organisation, the less satisfied they are with the organisation (reassuring). This value of *R* is what you would get if you used the command *correlations who satorg*.
- The measure of (linear + non-linear) association *eta*: .446. This is a correlation-like quantity which takes account of the non-linear as well as linear parts of the relationship between the variables. Because the non-linear part is not significant in this case, *eta* is not much greater than *R*.

Alan Taylor
Department of Psychology 2001
Additions and slight modifications 2002-2005