Statistical Consulting Report

December 18, 2009

Caitlin Burke
Graduate Student, Forestry and Environmental Resources
Advisor: Dr. Toddi Steelman

By Fikret Isik
Statistical Liaison of College of Natural Resources, NCSU

Background:

An on-line survey was conducted using Survey Monkey. The survey was sent to 3355 people. Researchers received 976 responses.

The objective of the study was to understand (1) what new information fire managers in the south need to do their jobs, (2) how they prefer to get that information, (3) what barriers they perceive to knowledge exchange, and (4) how a consortium of fire science managers and researchers can help them obtain the information they need.

There were 14 main questions in the survey and each question had up to 40 sub-questions. One particular question the scientists were interested in was the differences between states in levels of the response variables. Another question of the interest was the effect of the capacity of the respondents. Some of the states had very few response rates.

Comments:

Survey sample design is important in determining an appropriate statistical analysis method. A complex sample design can include stratification, clustering, multiple stages of selection, and unequal weighting. From the description of the survey, it looks like the survey does not have stratification or subgroups.

Stratified sampling involves selecting samples independently within strata. States can be classified as Strata which are non-overlapping subgroups of the survey population. Having homogeneous sampling units for the characteristics of interest will improve the precision. However, talking to the scientists clarified that 'states' is an independent variable whose association with the response levels is one of the research hypothesis to be answered from statistical analysis.

Most software use equal weights as default. When the sampling units have unequal weights, the weights should be provided to the survey analysis. Examining data suggest that there is no a weight factor.

There are several statistical analysis programs that can handle simple and complex survey data analysis. I will give some examples from the SAS software, which is freely available at NCSU.

Explanatory data analysis

Before testing any hypothesis, it is recommended that explanatory data analysis be performed. Proc FREQ in SAS is a useful procedure to examine the frequency of responses for each state. Such explanatory data analysis helps to eliminate states with very few answers.

```
ods graphics on;
   proc freq data=char;
      tables State / out=FreqCount_State  sparse;
      tables Capacity/ out=FreqCount_Capacity ;
      title 'Frequency of responses for state & capacity';
   run;
ods graphics off;
```

(Note: ODS Graphics are not available with SAS 9.1. You must have SAS 9.2 to generate statistical graphics).

The below table shows that some of the states, such as CA and DC had very few responses and they should be eliminated from further analysis.

State	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AL	30	4.17	30	4.17
AR	16	2.23	46	6.40
CA	1	0.14	47	6.54
DC	3	0.42	50	6.95
FL	145	20.17	195	27.12
GA	43	5.98	238	33.10
IN	1	0.14	239	33.24
KY	5	0.70	244	33.94
LA	21	2.92	265	36.86
MI	1	0.14	266	37.00
MN	1	0.14	267	37.13
MS	37	5.15	304	42.28
MT	1	0.14	305	42.42
NC	281	39.08	586	81.50
OK	7	0.97	593	82.48
PR	1	0.14	594	82.61
SC	54	7.51	648	90.13
TN	15	2.09	663	92.21
TX	24	3.34	687	95.55
VA	32	4.45	719	100.00

The following code keeps states with ‘enough’ respondents without changing the original data.

```
Data char2 ;
set char;
if state in('AL','FL','GA','MS','NC','SC','TX','VA')  ;
run;
```

After running the FREQ procedure again, states with reasonable number of respondents are summarized below.

State	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AL	30	4.64	30	4.64
FL	145	22.45	175	27.09
GA	43	6.66	218	33.75
MS	37	5.73	255	39.47
NC	281	43.50	536	82.97
SC	54	8.36	590	91.33
TX	24	3.72	614	95.05
VA	32	4.95	646	100.00

For the capacity factor, there are only 8 respondents for the 'Other' option in the data. The researchers should consider dropping the 'Other' option when comparing Capacity levels.

Capacity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Academic	38	5.91	38	5.91
Non-profit	33	5.13	71	11.04
Other	8	1.24	79	12.29
Private	212	32.97	291	45.26
Public	352	54.74	643	100.00

The FREQ procedure code above generates the Frequencies, Percent Cumulative Freq and Cumulative Percent for each state and for each capacity. If desired, such summary statistics can be exported to Excel to produce charts.

```
/* Export freq counts to excel */
   PROC EXPORT DATA= FreqCount_capacity
            OUTFILE= "&Folder\FreqCounts.xls"
            DBMS=EXCEL REPLACE;
     SHEET="capacity";
    RUN;
```

Analysis of survey data

The following example shows how you can use PROC SURVEYFREQ to analyze sample survey data.

```
title 'One-way frequency table analysis';
proc surveyfreq data=char2;
   tables  Q4_A ;
   strata  State ;
run;
```

One-way table of response for question Q4_A

An estimation of population total and population percentages for each category of the response are given. The response level 'Very helpful' has a frequency of 387. It is estimated that 60.28% of responses (across the states) in the study fall into this category and the standard error of this estimate is 1.937.

Q4_A	Frequency	Percent	Std Err of Percent
Not very helpful	56	8.7227	1.1139
Somewhat helpful	199	30.9969	1.8251
Very helpful	387	60.2804	1.9370
Total	642	100.000	
	Frequency Missing = 4		

The following PROC SURVEYFREQ statements request confidence limits for the percentages and a chi-square goodness-of-fit test for the one-way table for question Q4_A:

```
proc surveyfreq data=char2 nosummary;
   tables  Q4_A / cl chisq;
   strata  State  ;
 run;
```

The CHISQ option requests a Rao-Scott chi-square goodness-of-fit test. CL requests the 95% confidence limits.

In the output below, we now have 95% confidence limits of each category. The CL of 'Very helpful' category is 56.47% and 64.08%

.

Q4_A					
Q4_A	**Frequency**	**Percent**	**Std Err of Percent**	**95% Confidence Limits for Percent**	
Not very helpful	56	8.7227	1.1139	6.5354	10.9100
Somewhat helpful	199	30.9969	1.8251	27.4129	34.5809
Very helpful	387	60.2804	1.9370	56.4766	64.0842
Total	642	100.000			
Frequency Missing = 4					

The chi-square (Rao-Scott design-adjusted chi-square test) goodness-of-fit results for the table of Q4_A. **The null hypothesis** for this test is *equal proportions for the levels*. An *F* approximation is also provided. For the table of Q4_A, the *F* value is 128 with a *p*-value of <0.0001, which indicates rejection of the null hypothesis of equal proportions for response levels. In other words, levels (Not very helpful, somewhat helpful, Very helpful) are highly significantly different from each other for the proportions of response levels.

Rao-Scott Chi-Square Test	
Pearson Chi-Square	257.5607
Design Correction	0.9998
Rao-Scott Chi-Square	257.6169
DF	2
Pr > ChiSq	<.0001
F Value	128.8084
Num DF	2
Den DF	1268
Pr > F	<.0001
Sample Size = 642	

The above analysis is for the overall population (across the states). In order to test the null hypothesis that the proportion of the response for a given state are the same, we can generate two-way tables.

```
/* The following PROC SURVEYFREQ statements request confidence limits
   for the percentages and a chi-square goodness-of-fit test for the
Two-way table of Response */

  proc surveyfreq data=Char2 nosummary ;
      tables  State * Q4_A / cl chisq row nototal;
   run;
```

The ROW option in the TABLES statement requests row percentages, which gives the distribution of Response levels within each level of the row variable STATE. The CHISQ option requests a Rao-Scott chi-square test of association between STATE and Q4_A.

The two-way table of STATE by Q4-A. It is estimated that 26.6% of respondents in the study chose 'Very helpful' are from NC. The standard error of this percent is 1.746 and the 95% CL are 23.2% and 30.1%. The **row percent** shows the distribution of categories within NC. About 61.7% of the respondents in NC chose 'Very useful' with a standard error of 2.9226%.

Table of State by Q4_A

State	Q4_A	Frequency	Percent	Std Err of Percent	95% Confidence Limits for Percent		Row Percent	Std Err of Row Percent
AL	Not very helpful	1	0.1558	0.1558	0.0000	0.4616	3.3333	3.2799
	Somewhat helpful	11	1.7134	0.5126	0.7069	2.7199	36.6667	8.8050
	Very helpful	18	2.8037	0.6520	1.5234	4.0841	60.0000	8.9512
NC	Not very helpful	29	4.5171	0.8203	2.9064	6.1279	10.4693	1.8410
	Somewhat helpful	77	11.9938	1.2832	9.4739	14.5136	27.7978	2.6939
	Very helpful	171	26.6355	1.7460	23.2069	30.0641	61.7329	2.9226
VA	Not very helpful	3	0.4673	0.2694	0.0000	0.9962	9.3750	5.1567
	Somewhat helpful	14	2.1807	0.5769	1.0479	3.3135	43.7500	8.7763
	Very helpful	15	2.3364	0.5966	1.1648	3.5081	46.8750	8.8284

Frequency Missing = 4

Table of State by Q4_A

State	Q4_A	95% Confidence Limits for Row Percent	
AL	Not very helpful	0.0000	9.7739
	Somewhat helpful	19.3765	53.9568
	Very helpful	42.4227	77.5773
VA	Not very helpful	0.0000	19.5011
	Somewhat helpful	26.5161	60.9839
	Very helpful	29.5389	64.2111

Frequency Missing = 4

The Rao-Scott chi-square statistic equals 13.42, and the corresponding *F* value is 0.9588 with a *p*-value of 0.4935. This indicates that there is no association between states and the levels of the response for Q4_A. In other words, the levels of the response are not different within each state.

Rao-Scott Chi-Square Test

Pearson Chi-Square	13.4237
Design Correction	1.0000
Rao-Scott Chi-Square	13.4237
DF	14
Pr > ChiSq	0.4935
F Value	0.9588
Num DF	14
Den DF	8974
Pr > F	0.4935

Sample Size = 642

Proportional odds model

The association between an independent variable (state in this study) and a categorical response variable can be tested using proportional odds model. The response variable is ordinal with three levels. The SURVEYLOGISTIC procedure fits a common slope cumulative model, which is a parallel lines regression model based on the cumulative probabilities of the Q4_A categories rather than on their individual probabilities. The cumulative model has the form

$$g(Pr(Y \leq d \mid x)) = \alpha_d + x\beta$$

where g is the link function, Y is the response variable, x is the independent variable, $\alpha_1,...\alpha_d$ are intercept parameters and β is the vector of slope parameters. This model is called the cumulative logit model or the proportional odds model.

In order to test a null hypothesis of no differences among states in response levels (Not very helpful, somewhat helpful, Very helpful) the following SAS code can be used.

```
proc surveylogistic data=char2 ;
class  state ;
model Q4_A (ORDER=INTERNAL) = state  ;
contrast 'AL vs NC' state 1 0 0 0 1 ;
run;
```

A partial output with interpretation of important tables for Q4_A is given below.

```
The SURVEYLOGISTIC Procedure

Model Information

Data Set                      WORK.CHAR2
Response Variable             Q4_A
Number of Response Levels     3
Model                         Cumulative Logit
Optimization Technique        Fisher's Scoring
Variance Adjustment           Degrees of Freedom (DF)


Number of Observations Read         646
Number of Observations Used         642
```

The "Response Profile" table (below) lists the three response levels, their ordered values, and their total frequencies for each category. Due to the ORDER=INTERNAL option for the response variable *Q4_A*, the category "Not very helpful" has the Ordered Value 1, the category "Somewhat helpful" has the Ordered Value 2, and the category "Very helpful" has the Ordered Value 3 (this can be changed).

Response Profile

Ordered Value	Q4_A	Total Frequency
1	Not very helpful	56
2	Somewhat helpful	199
3	Very helpful	387

Probabilities modeled are cumulated over the lower Ordered Values.

Class Level Information table (below) shows the parameterization in the regression model for each categorical independent variable. The design variable can be used to test specific hypotheses using the CONTRAST statement in SURVEYLOGISTIC. For example, in order to compare AL and NC, the design variable should be `contrast 'AL vs NC' state 1 0 0 0 1 ;`

Class Level Information

Class	Value	Design Variables						
State	AL	1	0	0	0	0	0	0
	FL	0	1	0	0	0	0	0
	GA	0	0	1	0	0	0	0
	MS	0	0	0	1	0	0	0
	NC	0	0	0	0	1	0	0
	SC	0	0	0	0	0	1	0
	TX	0	0	0	0	0	0	1
	VA	-1	-1	-1	-1	-1	-1	-1

.

The chi-square test for testing the proportional odds assumption is given below. The test is not significant, which indicates that the cumulative logit model adequately fit the data.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
9.4165	7	0.2241

The three chi-square tests show that the independent variable (state) is not significant (H0: Beta=0).

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.1148	7	0.8742
Score	3.3027	7	0.8557
Wald	3.2604	7	0.8599

The Type 3 Analysis of Effects table also shows no significant differences among the states in their response to levels of Q4A.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
State	7	3.2604	0.8599

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	Not very helpful	1	-2.3009	0.1466	246.3708	<.0001
Intercept	Somewhat helpful	1	-0.3633	0.1067	11.6048	0.0007
State	AL	1	-0.1254	0.3227	0.1510	0.6975
State	FL	1	-0.0787	0.1798	0.1915	0.6616
State	GA	1	-0.0750	0.2695	0.0775	0.7807
State	MS	1	-0.1155	0.2914	0.1571	0.6918
State	NC	1	-0.0789	0.1491	0.2801	0.5966
State	SC	1	-0.2171	0.2686	0.6532	0.4190
State	TX	1	0.2880	0.3437	0.7020	0.4021