ANALYSIS OF SAMPLE SURVEY DATA

Ana Yankovsky
Research Associate
Alberta Cancer Prevention Legacy Fund
Alberta Health Services

Overview

- Explore why we need survey analysis tools;
- Learn about analytics survey procedures in SAS;
- Use procedures with sample survey examples;
- Questions/Discussion;

Why do we need survey data analysis software?

- Regular statistical software analyzes data as if the data were collected using simple random sampling;
- Very few surveys use a simple random sample to collect data because it is not efficient (financially or statistically);
- If you ignore the sampling design both the point estimates and their standard errors will likely be calculated incorrectly.

Complex Sample Design Analysis

- Need to account for weights, stratification, and clustering that was part of the sample design;
  - The sampling weight will affect the calculation of the point estimate;
  - The stratification and/or clustering will affect the calculation of the standard errors.
  - Ignoring the clustering will likely lead to standard errors that are underestimated, possibly leading to results that seem to be statistically significant, when in fact, they are not.

- SURVEY procedures allow for correct estimation of variances/standard errors from complex samples;

- SURVEY procedures cover main analytic techniques:
  - Means/Totals
  - Frequency tables
  - Linear regression
  - Logistic regression
  - Survival models using PH regression approach

Example 1: Big Burn Marketing Survey

- Sampling from an on-line panel – non-random sampling;
- Sample was weighted according to Census 2011;
- Target population parents of children 10 to 15 years old;
- The intend of this survey is to measure the impact of a marketing campaign on the parents' knowledge, believe and behavior towards indoor tanning and assess if they talked to their children about the skin cancer risks that indoor tanning posses;

PROC SURVEYFREQ

PROC SURVEYFREQ <options> ;

BY variables ;

CLUSTER variables ;

REPWEIGHTS variables </ options> ;

STRATA variables </ option> ;

TABLES requests </ options> ;

WEIGHT variable ;

PROC SURVEYFREQ

- For one-way frequency tables
  - Rao-Scott chi-square goodness-of-fit tests, which are adjusted for the sample design.
- For tables computes
  - Estimates and confidence limits for risks (or row proportions), the risk difference, the odds ratio, and relative risks.
- For two-way tables provides
  - Design-adjusted tests of independence, or no association, between the row and column variables.
- Provides a choice of variance estimation methods
  - Taylor series linearization,
  - Balanced repeated replication (BRR),
  - Jackknife.

Example 1: Big Burn Marketing Survey

Test if more parents talked to their child about indoor tanning post marketing intervention;

```
PROC SURVEYFREQ data=ana.allcamp3;
     TABLES post*q7_1/or row nowt chisq
          plots=WTFREQPLOT(scale= percent);
     WEIGHT weight;
RUN;
```

Example1: Big Burn Marketing Survey

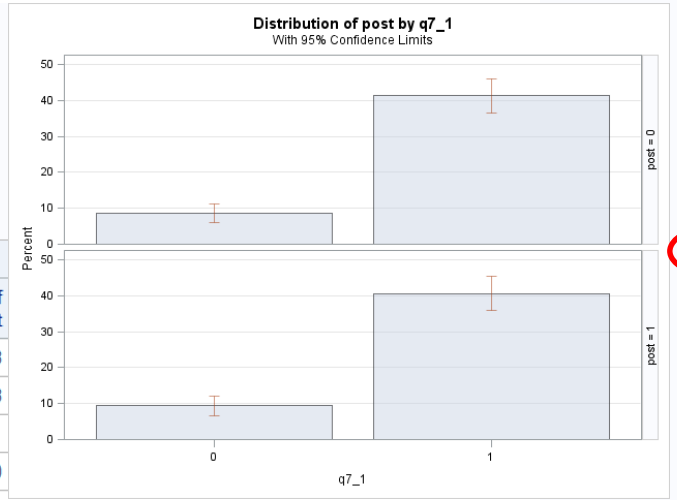
The SAS System

The SURVEYFREQ Procedure

Data Summary	
Number of Observations	500
Sum of Weights	500.000006

Table of post by q7_1

post	q7_1	Frequency	Percent	Std Err of Percent	Row Percent	Std Err of Row Percent
0	0	46	8.6976	1.3120	17.3952	2.5233
	1	204	41.3024	2.4053	82.6048	2.5233
	Total	250	50.0000	2.4126	100.000	
1	0	43	9.3505	1.4277	18.7009	2.6810
	1	207	40.6495	2.3409	81.2991	2.6810
	Total	250	50.0000	2.4126	100.000	
Total	0	89	18.0481	1.8428		
	1	411	81.9519	1.8428		
	Total	500	100.000			



Rao-Scott Chi-Square Test	
Pearson Chi-Square	0.1441
Design Correction	1.1394
Rao-Scott Chi-Square	0.1265
DF	1
Pr > ChiSq	0.7221
F Value	0.1265
Num DF	1
Den DF	499
Pr > F	0.7223
Sample Size = 500	

Odds Ratio and Relative Risks (Row1/Row2)			
	Estimate	95% Confidence Limits	
Odds Ratio	0.9155	0.5614	1.4928
Column 1 Relative Risk	0.9302	0.6231	1.3886
Column 2 Relative Risk	1.0161	0.9302	1.1099
Sample Size = 500			

PROC SURVEYLOGISTICS

PROC SURVEYLOGISTIC <options> ;

BY variables ;

CLASS variable <(v-options)> <variable <(v-options)> ...> </ v-options> ;

CLUSTER variables ;

CONTRAST ’label’ effect values <,...effect values> </ options> ;

DOMAIN variables <variablevariable variablevariablevariable ...> ;

FREQ variable ;

MODEL events/trials = <effects < / options>> ;

MODEL variable <(v-options)> = <effects> < / options> ;

OUTPUT <OUT=SAS-data-set> <options> ;

REPWEIGHTS variables < / options> ;

STRATA variables </ option> ;

<label:> TEST equation1 < , ... , equationk> </ options> ;

UNITS independent1 = list1 <... independentk = listk> < / options> ;

WEIGHT variable </ option> ;

Example 1: Big Burn Marketing Survey

Test the model that parents did not talk to their child about indoor tanning because of their *sex*, *low income*, believe that it is *not important to talk about indoor tanning*, that they *allow their children to tan* and that the *parents themselves are tanning beds users*.

```
PROC SURVEYLOGISTIC data=ana.allcamp3;
  CLASS sex low_income not_import alow_tan q10;
  MODEL q7_1 (REF='0') = sex low_income not_import
  alow_tan q10 / link=glogit;
  WEIGHT weight;
RUN;
```

Example 1: Big Burn Marketing Survey

The SAS System

The SURVEYLOGISTIC Procedure

Model Information		
Data Set	ANA.ALLCAMP3	
Response Variable	q7_1	q7_1
Number of Response Levels	2	
Weight Variable	weight	weight
Model	Generalized Logit	
Optimization Technique	Newton-Raphson	
Variance Adjustment	Degrees of Freedom (DF)	

Variance Estimation	
Method	Taylor Series
Variance Adjustment	Degrees of Freedom (DF)

Number of Observations Read	500
Number of Observations Used	500
Sum of Weights Read	500
Sum of Weights Used	500

Response Profile			
Ordered Value	q7_1	Total Frequency	Total Weight
1	0	89	90.24029
2	1	411	409.75972

Logits modeled use q7_1='0' as the reference category.

Class Level Information				
Class	Value	Design Variables		
Sex	1	1		
	2	-1		
low_income	0	1		
	1	-1		
not_import	0	1		
	1	-1		
alow_tan	0	1		
	1	-1		
q10	1	1	0	0
	2	0	1	0
	3	0	0	1
	8	-1	-1	-1

Model Convergence Status
onvergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	474.121	478.735
SC	478.336	512.452
-2 Log L	472.121	462.735

Testing Global Null Hypothesis: BETA=0				
Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	1.34	7	Infty	0.2261
Score	0.55	7	499	0.7996
Wald	33.64	7	499	<.0001

Odds Ratio Estimates				
Effect	q7_1	Point Estimate	95% Confidence Limits	
Sex 1 vs 2	1	0.929	0.553	1.563
low_income 0 vs 1	1	1.369	0.558	3.358
not_import 0 vs 1	1	0.716	0.351	1.458
alow_tan 0 vs 1	1	0.974	0.294	3.227
q10 1 vs 8	1	>999.999	>999.999	>999.999
q10 2 vs 8	1	>999.999	>999.999	>999.999
q10 3 vs 8	1	>999.999	>999.999	>999.999

NOTE: The degrees of freedom in computing the confidence limi

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	47.6	Somers' D	0.131
Percent Discordant	34.5	Gamma	0.160
Percent Tied	17.9	Tau-a	0.038
Pairs	36579	c	0.566

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Sex	1	0.0765	0.7821
low_income	1	0.4729	0.4916
not_import	1	0.8518	0.3560
alow_tan	1	0.0019	0.9649
q10	3	228.8164	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		q7_1	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-2.4933	0.4673	-5.34	<.0001
Sex	1	1	-0.0366	0.1323	-0.28	0.7823
low_income	0	1	0.1570	0.2283	0.69	0.4920
not_import	0	1	-0.1671	0.1811	-0.92	0.3565
alow_tan	0	1	-0.0134	0.3049	-0.04	0.9650
q10	1	1	3.3988	0.4703	7.23	<.0001
q10	2	1	4.0852	0.3390	12.05	<.0001
q10	3	1	4.0183	0.3144	12.78	<.0001

NOTE: The degrees of freedom for the t tests is 499.

Example 2: Weight loss data

- Fictional data to show how other survey analysis procedures work;
- Sample is stratified by region of residence of the participant and is weighted according to Census 2011;
- Data reflects a longitudinal survey design, intended to measure the weight loss after 6 months among those participating in a diet/exercise program.

PROC SURVEYMEANS

```
PROC SURVEYMEANS <options> <statistic-keywords> ;
    BY variables ;
    CLASS variables ;
    CLUSTER variables ;
    DOMAIN variables <variable variable variable...> < / option> ;
    RATIO <'label'> variables / variables ;
    REPWEIGHTS variables < / options> ;
    STRATA variables < / option> ;
    VAR variables ;
    WEIGHT variable ;
```

Example 2: Weight loss data

```
PROC SURVEYMEANS data=example
  /*sum df clm fraction*/;
  VAR w_loss_6mon_kg;
  STRATA Strata/ list;
  DOMAIN sex;
  WEIGHT weight;
RUN;
```

Example 2: Weight loss data

The SAS System

The SURVEYMEANS Procedure

Data Summary	
Number of Strata	13
Number of Observations	340
Number of Observations Used	332
Number of Obs with Nonpositive Weights	8
Sum of Weights	327.324263

Stratum Information

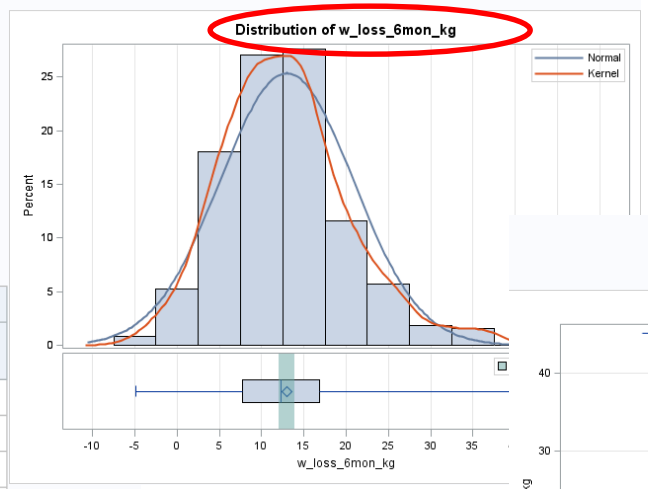
Stratum Index	Strata	N Obs	Variable	Label	N
1	1	21	w_loss_6mon_kg	w_loss_6mon_kg	21
2	10	30	w_loss_6mon_kg	w_loss_6mon_kg	30
3	11	29	w_loss_6mon_kg	w_loss_6mon_kg	29
4	12	21	w_loss_6mon_kg	w_loss_6mon_kg	21
5	13	25	w_loss_6mon_kg	w_loss_6mon_kg	25
6	2	27	w_loss_6mon_kg	w_loss_6mon_kg	27
7	3	25	w_loss_6mon_kg	w_loss_6mon_kg	25
8	4	26	w_loss_6mon_kg	w_loss_6mon_kg	26
9	5	26	w_loss_6mon_kg	w_loss_6mon_kg	26
10	6	26	w_loss_6mon_kg	w_loss_6mon_kg	26
11	7	29	w_loss_6mon_kg	w_loss_6mon_kg	29
12	8	27	w_loss_6mon_kg	w_loss_6mon_kg	27
13	9	20	w_loss_6mon_kg	w_loss_6mon_kg	20

Statistics

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
w_loss_6mon_kg	w_loss_6mon_kg	332	12.955811	0.482975	12.0055910	13.9060302

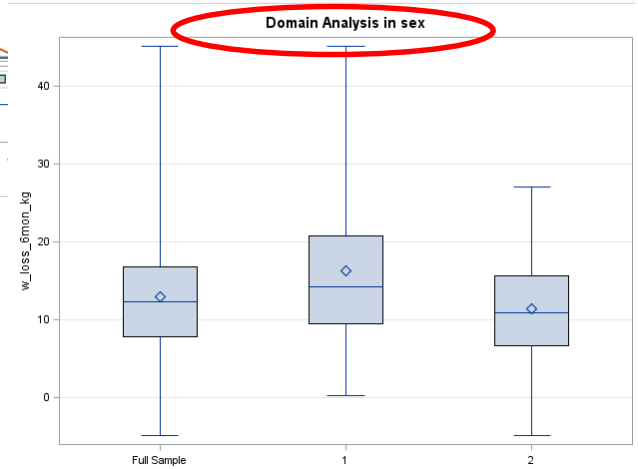
The SAS System

The SURVEYMEANS Procedure



The SAS System

The SURVEYMEANS Procedure



The SAS System

The SURVEYMEANS Procedure

Domain Statistics in sex

sex	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
1	w_loss_6mon_kg	w_loss_6mon_kg	114	16.167076	1.048999	14.1032465	18.2309060
2	w_loss_6mon_kg	w_loss_6mon_kg	218	11.299256	0.462134	10.3900409	12.2084719

PROC SURVEYREG

```
PROC SURVEYREG  <options> ;
   BY variables ;
   CLASS variables ;
   CLUSTER variables ;
   CONTRAST ’label’ effect values < ... effect values> </ options> ;
   DOMAIN variables <variablevariable variablevariablevariable ... > ;
   EFFECT name = effect-type ( variables </ options> ) ;
   ESTIMATE <'label'> estimate-specification < / options> ;
   LSMEANS <model-effects> </ options> ;
   LSMESTIMATE model-effect lsmestimate-specification < / options> ;
   MODEL dependent = <effects> </ options> ;
   OUTPUT <keyword <=variable-name> ... keyword <=variable-name>> </
   option> ;
   REPWEIGHTS variables < / options> ;
   SLICE model-effect </ options> ;
   STORE <OUT=>item-store-name</LABEL='label'> ;
   STRATA variables </ option> ;
   TEST <model-effects> </ options> ;
   WEIGHT variable ;
```

Example 2: Weight loss data

```
PROC SURVEYREG data=example;
   STRATA Strata / list;
   CLASS sex tobacco_use diabetes;
   MODEL w_loss_6mon_kg = sex tobacco_use
                          diabetes / covB;
   WEIGHT Weight;
RUN;
```

Example 2: Weight loss data

The SAS System

The SURVEYREG Procedure

Regression Analysis for Dependent Variable w_loss_6mon_kg

Data Summary	
Number of Observations	332
Sum of Weights	327.32426
Weighted Mean of w_loss_6mon_kg	12.95581
Weighted Sum of w_loss_6mon_kg	4240.8

Design Summary	
Number of Strata	13

Fit Statistics	
R-Square	0.1559
Root MSE	7.2766
Denominator DF	319

Class Level Information			
CLASS Variable	Label	Levels	Values
sex	sex	2	1 2
tobacco_use	tobacco_use	2	0 1
diabetes	diabetes	2	0 1

Tests of Model Effects			
Effect	Num DF	F Value	Pr > F
Model	3	14.77	<.0001
Intercept	1	108.98	<.0001
sex	1	18.11	<.0001
tobacco_use	1	28.70	<.0001
diabetes	1	0.34	0.5585

Stratum Information		
Stratum Index	Strata	N Obs
1	1	21
2	10	30
3	11	29
4	12	21
5	13	25
6	2	27
7	3	25
8	4	26
9	5	26
10	6	26
11	7	29
12	8	27
13	9	20

Covariance of Estimated Regression Coefficients

	Intercept	sex 1	sex 2	tobacco_use 0	tobacco_use 1	diabetes 0	diabetes 1
Intercept	5.601523823	-1.113423746	0	-2.171143261	0	-3.491379847	0
sex 1	-1.113423746	1.3274567726	0	0.3300077848	0	0.6302834032	0
sex 2	0	0	0	0	0	0	0
tobacco_use 0	-2.171143261	0.3300077848	0	2.0102337875	0	0.3103225691	0
tobacco_use 1	0	0	0	0	0	0	0
diabetes 0	-3.491379847	0.6302834032	0	0.3103225691	0	3.2972947971	0
diabetes 1	0	0	0	0	0	0	0

Other Sample Survey Analysis Procedures

PROC SURVEYPHREG - procedure performs regression analysis based on the Cox proportional hazards model for sample survey data;

PROC SURVEYSELECT - procedure provides a variety of methods for selecting probability-based random samples;

References:

1. Statistical Computing Seminar. Applied Survey Data Analysis using SAS 9.4. UCLA: Statistical Consulting Group. From http://statistics.ats.ucla.edu/stat/sas/seminars/SAS_survey/ (accessed March 21, 2016).

2. An, A. & Watts, D. New SAS Procedures for Analysis of Sample Survey Data. SAS Institute. From http://www.ats.ucla.edu/stat/sas/libra.ry/svy_survey.pdf (accessed March 30, 2016).

3. Survey Analysis. SAS Support. From https://support.sas.com/rnd/app/stat/procedures/SurveyAnalysis.html (accessed March 30, 2016)

4. Berglund, P. A. Analysis of Survey Data Using the SAS SURVEY Procedures: A Primer. From: http://www.isr.umich.edu/src/smp/asda/Berglund%20Survey%20Data%20Analysis%20WIILL%20June%202014.pdf (accessed March 21, 2016)

Questions?