



# The Enigma of Survey Data Analysis: Comparison of SAS Procedures and SUDAAN Procedures

---

By Katherine Baisden

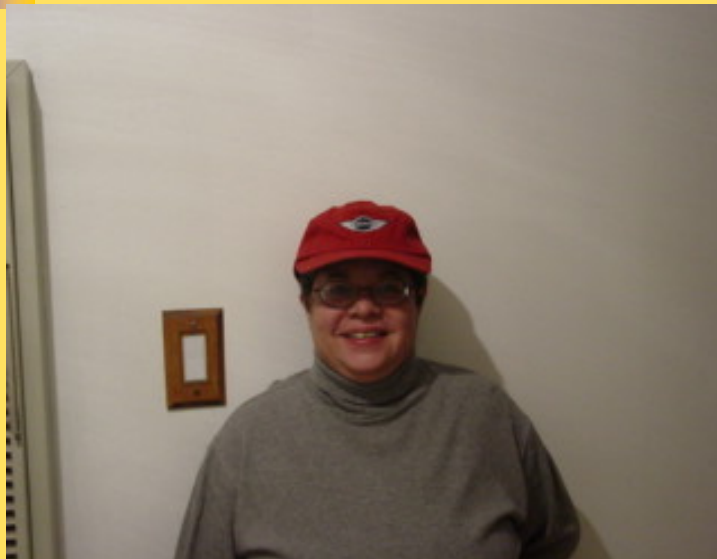
SRI International, Menlo Park

Prepared for Michigan SAS User's Group

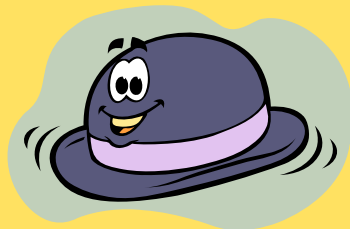
Southfield, MI

November 2007

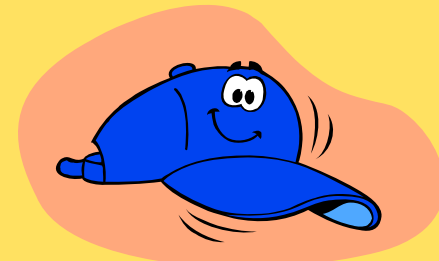
# The Tale of Two Hats



SUDAAN



SAS





# The Taming of the Tigger

---

- Hey Buddy I got you now!



■ Oct 2004



■ Oct 2007



# Purpose of Paper/Presentation

---

This paper  
will compare  
and contrast:

- Proc SURVEYMEANS  
with SUDAAN'S SAS  
CALLABLE PROC  
DESCRIPT
- PROC SURVEYFREQ  
with SUDAAN'S SAS  
CALLABLE PROC  
CROSSTAB



# Purpose of Paper/Presentation

---

- Proc SURVEYREG with SUDAAN'S SAS CALLABLE PROC REGRESS
- PROC SURVEYLOGISTIC with SUDAAN'S SAS CALLABLE PROC RLOGIST

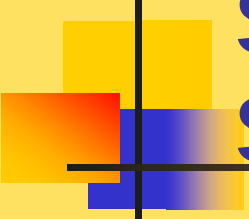


# When do you need to use PROC SURVEY Procedures or SUDAAN?

---

When you are dealing with survey designs that are:

- Stratified
- Clustered
- Have unequal probabilities of selection
- Or any combination of the above



# Why do you need to use PROC SURVEY Procedures or SUDAAN?

---

- If you do not use the Special SAS Survey Procedures or SUDAAN you will obtain **BIASED** estimators of VARIANCE , STANDARD ERRORS and TEST statistics (e.g., chi-square)



# Generalized Rules

---

Type of Procedure	Point Estimate (Means, Percents)	Variances/Test Statistics
SAS Unweighted	Incorrect	Incorrect
SAS Weighted	Correct	Incorrect
SAS Survey Procs	Correct	Correct
SUDAAN	Correct	Correct



# Example of Variance Estimation Differences (T4B)

---


Type of Procedure	Mean	Standard Error of the Mean
Proc Means Weighted	2.89	0.092
Proc Surveymeans	2.89	0.058
Proc Descript SUDAAN	2.89	0.063


---

# The Mystery of the Design Effect (DE)

$$DE = \text{Variance CSD} / \text{Variance SRS}$$

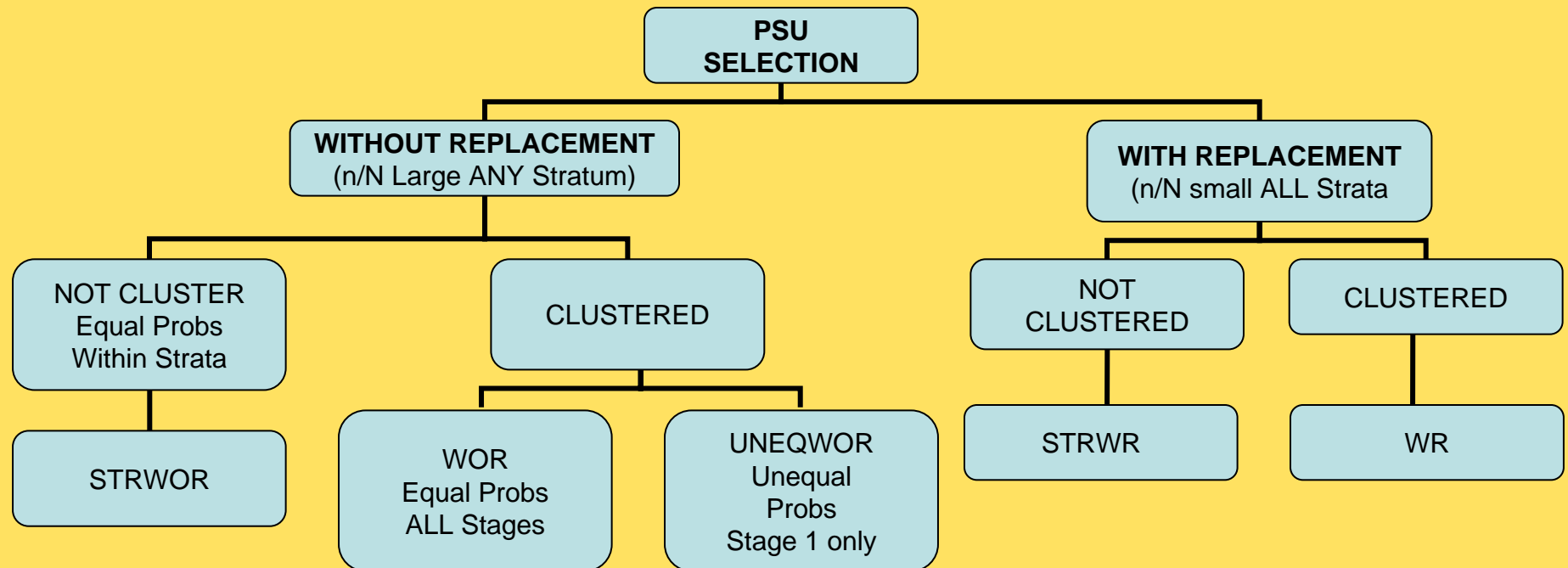
This formula calculates a ratio which represents impact of the CSD on the sample.

DE close to 1  there is no impact

DE exceeds 1+  correlation between respondents in your clusters leading to incorrect variances or test statistics.

# Type of Sampling Design

## Taylor Series Design Options





# Sample Design Options SUDAAN

---

Taylor Linearization Methods		Replication Methods
With Replacement	Without Replacement	
WR STRWR SRS	WOR UNEQWOR STRWOR	Jackknife (Delete -1 or Rep Wgt) BRR



# SAS and SUDAAN Procedures

<b>SAS</b>	<b>SUDAAN</b>	<b>Purpose</b>
	<b>RECORDS</b>	Prints Records
<b>SURVEYFREQ</b>	<b>CROSSTAB</b>	Oneway/Multiway Freqs
	<b>RATIO</b>	Ratio Estimates
<b>SURVEYMEANS</b>	<b>DESCRIPT</b>	Means/SEs
<b>SURVEYREG</b>	<b>REGRESS</b>	Fits Linear Models
<b>SURVEYLOGISTIC</b>	<b>RLOGIST</b>	Fits Logistic Models
	<b>MULTILOG</b>	Log Models Categorical Dependents
	<b>SURVIVAL</b>	Discrete Proportional Hazard Models
<b>SURVEYSELECT</b>		Select a Sample



# Data Set Parameters

---

- Using a data set about CA teachers
- Stratification Vars
  - Emerg (3 levels – High, Med, Low)
  - Distsize (3 levels – Small, Med, Large)
  - Schl\_lvl (3 levels – Elem, Middle, High)

T4B – number of classes taught

T40 – Gender

WGTD – weight variable

T6 – Leave prep program for employment (Y/N)

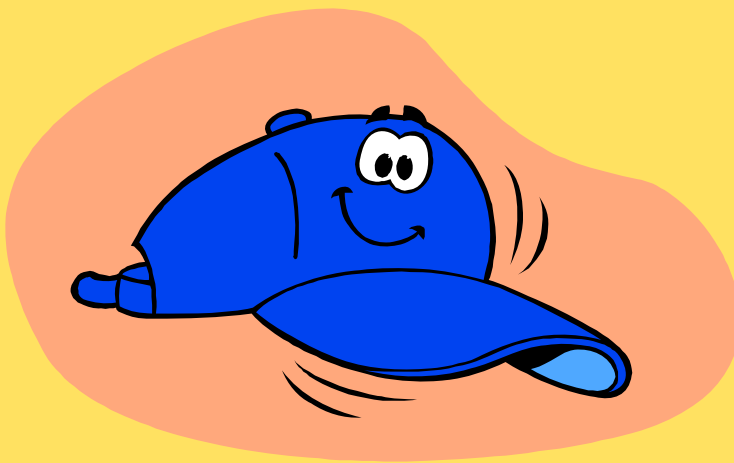
T41 – Age

T36 - Number of years teaching fulltime

# Code Example

## PROC SURVEYMEANS

---

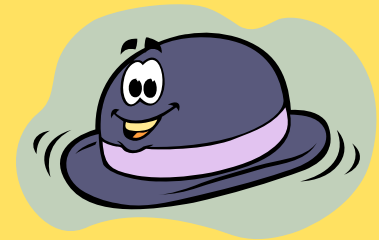


- Proc Surveymeans;
- Var T4B;
- Strata Emerg  
Distsize Schl\_lvl;
- Weight Wgtd;
- Domain T40;
- Run;



# Code Example

## PROC DESCRIPT (SUDAAN)



- Proc Descript data=one  
filetype=SAS  
design=strwr;
- Nest emerg distsize  
schl\_lvl;
- Weight Wgtd;
- Var T4B;
- Subgroup T40;
- Class T40 / nofreq;
- Setenv labwidth=28 colspce=1  
colwidth=10 decwidth=4;
- Print nsum="Sample Size"  
Wsum="Population size" Mean  
semean="S.E." Deffmean="Design  
effect" / style=nchs nsumfmt=f6.0  
wsumfmt=f10.0 Deffmeanfmt=F6.2  
semeanfmt=F7.4;
- Rtitle "Mean of T4B";
- Run;

## SUDAAN Means Output Example

Number of observations read: 441

Weighted count: 289707

Denominator degrees of freedom: 414

Variance Estimation Method: Taylor Series (STRWR)

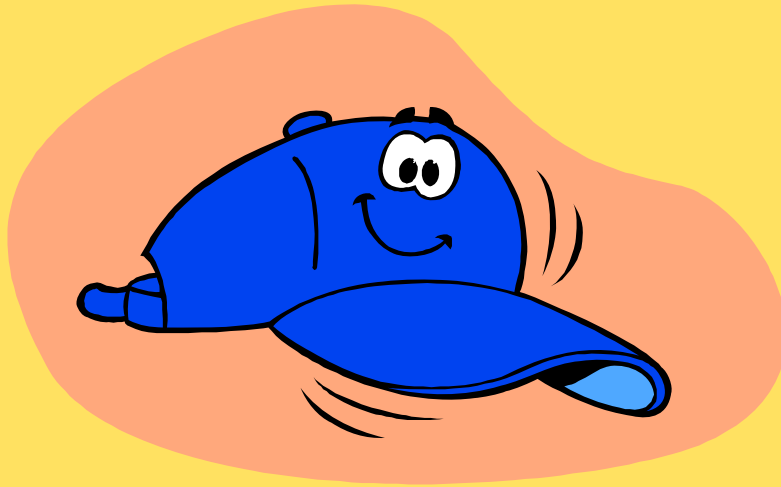
Mean of T4b by T40 by: Variable, T40:GENDER.

Variable T40:GENDER

Sample	Population Size	Size	Mean	S.E.	Design Effect
<b>T4b: TOTAL NUMBER OF CLASSES TAUGHT</b>					
Total	485	271,054	2.8982	0.0641	0.47
(1) FEMALE	361	201,576	2.4026	0.0961	0.91
(2) MALE	124	69,478	4.3360	0.2167	1.78

# Code Example

## PROC SURVEYFREQ



- **Proc Surveyfreq;**
- **Strata emerg distsize  
schl\_lvl;**
- **Tables T40\*T6 / chisq  
wchisq row col chisq1;**
- **Weight WGTd;**
- **TITLE 'Crosstab of T40  
by T6';**
- **Run;**

# Code Example



## PROC CROSSTAB (SUDAAN)

- Proc Crosstab data=one  
filetype=SAS  
design=strwr;
- Nest emerg distsize  
schl\_lvl;
- Weight WGTD;
- Subgroup T40 T6;
- Class T40 T6 / nofreq;
- Tables T40\*T6;
- Setenv colwidth=9 decwidth=2  
colspce=2;
- Print nsum wsum colper rowper  
totper /wsumfmt=f9.0  
nsumfmt=f9.0 cmhtest=all  
tests=all cmhfmt=f8.2  
Cmhdfmt=f8.0  
cmhpvalfmt=f8.4  
chisqfmt=f11.2;
- Rtitle "Crosstab of T40 by T6";
- Run;

# SUDAAN Crosstab Output Example

Number of observations read : 510      Weighted count : 285680  
 Denominator degrees of freedom : 483

Variance Estimation Method: Taylor Series (STRWR)

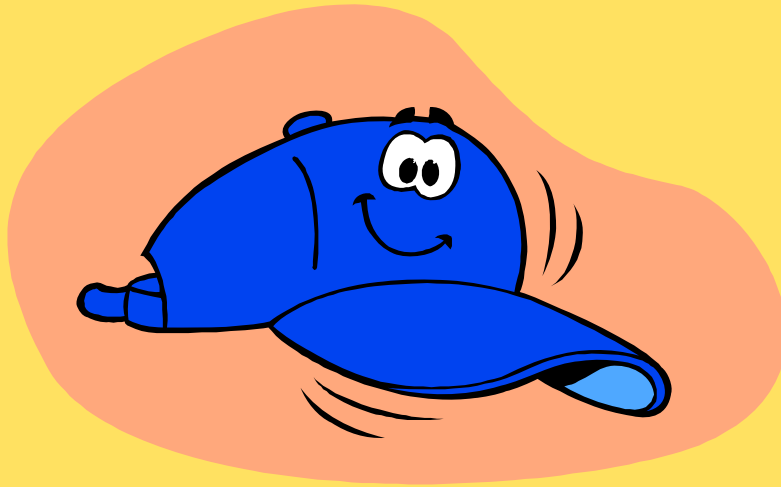
Crosstab of T40 by T6

by: T40:GENDER, T6:LEAVE MA OR PREP PGM FOR FT PAID POSITION.

T40:GENDER		T6:LEAVE MA OR PREP PGM FOR FT PAID POSITION			Pearson Chisq: 0.41 P-value of Pearson: .52 DF 1.00
		Total	1 (YES)	2 (NO)	
Total	Sample Size	452	25	427	Cochran-Mantel Haenszel Chisq: 0.41 P-value of CMH 0.52 DF 1.00
	Weighted Size	248022	12340	235683	
	Col Percent	100.00	100.00	100.00	
	Row Percent	100.00	4.98	95.02	
	Tot Percent	100.00	4.98	95.02	
(1) FEMALE	Sample Size	336	16	320	
	Weighted Size	185441	10006	175435	
	Col Percent	74.77	83.73	74.44	
	Row Percent	100.00	5.60	94.60	
	Tot Percent	74.77	4.23	70.73	
(2) MALE	Sample Size	116	9	107	
	Weighted Size	62581	2333	60248	
	Col Percent	25.23	18.91	25.56	
	Row Percent	100.00	3.73	96.27	
	Tot Percent	25.23	0.94	24.29	

# Code Example

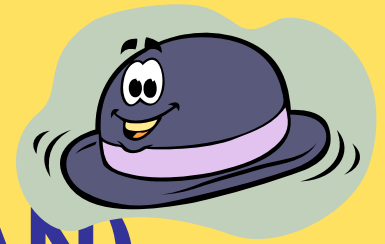
## PROC SURVEYREG



- **Proc Surveyreg;**
- **Strata emerg distsize  
schl\_lvl / list;**
- **Class T40;**
- **Model T36=T40 T41 / Anova  
Deff Adjrsq Solution;**
- **Weight WGTD;**
- **TITLE 'Surveyreg T36=T40  
T41';**
- **Run;**

# Code Example

## PROC REGRESS (SUDAAN)



- Proc Regress data=one  
filetype=SAS  
design=strwr;
- Nest emerg distsize  
schl\_lvl;
- Weight WGTD;
- **Subgroup T40 ;**
- **Class T40 / nofreq;**
- Test Satadjchi adjwaldf;
- Model T36=T40 T41;
- Setenv colwidth=9 decwidth=2  
colspce=2;
- Print Beta Sebeta defft T\_beta  
P\_beta Df Satadjaf Satadchi  
Adjwaldf satadchip Adjwaldp /  
risk=all deffmt=F6.2  
Satadjdffmt=F6.3  
Satadchifmt=F7.2  
Adjwaldffmt=F7.2  
Betafmt=F8.4 Sebetafmt=F8.4  
Betafmt=F7.4 Deffmt=F6.2  
Satadchpfmt=F7.4  
Adjwaldpfmt=F7.4;
- Rtitle "Sudaan Reg Proc  
T36=T40 T41";
- Run;

# Comparison of Information in Regression

<b>Statistical Concept</b>	<b>SAS PROC SURVEYREG</b>	<b>SUDAAN PROC REGRESS</b>
<b>R Square</b>	<b>Adjusted R- Square</b>	<b>Multiple R-Square</b>
<b>F Tests</b>	<b>Test of Model Effects Table Model</b>	<b>Contrast Table Model Minus Intercept</b>
<b>Betas and SEs</b>	<b>Estimated Reg Coefficient Table</b>	<b>Independent Variables Effects Table</b>

# Results of Regression Part I

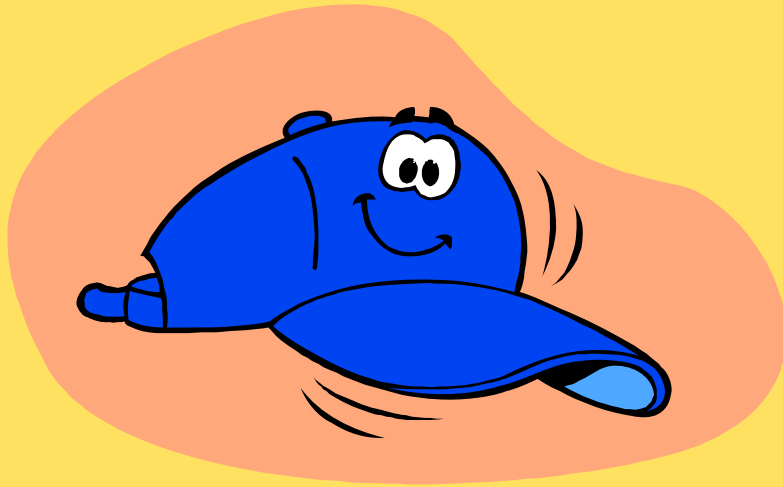
	<b>SAS PROC SURVEYREG</b>	<b>SUDAAN PROC REGRESS</b>
<b>R Square</b>	<b>.5181</b>	<b>.5199</b>
<b>F-Tests</b>		
<b>Model</b>	<b>120.28 (prob &lt;.0001)</b>	<b>121.24 (prob=.000)</b>
<b>T40</b>	<b>0.13 (prob=.7140)</b>	<b>0.14 (prob=.7125)</b>
<b>T41</b>	<b>240.34 (prob &lt;.0001)</b>	<b>242.76 (prob=0.000)</b>

# Results of Regression Part II

	<b>SAS PROC SURVEYREG</b>	<b>SUDAAN PROC REGRESS</b>
<b>Betas and Standard Errors</b>		
<b>Model</b>	<b>52.0321874 (se 2.936)</b>	<b>52.0322 (2.9219)</b>
<b>T40 (Males)</b>	<b>-0.42355 (se 1.155)</b>	<b>-0.4236 (se 1.1491)</b>
<b>T40 (Females)</b>	<b>0.0000 (se 0.000)</b>	<b>0.0000 (se 0.0000)</b>
<b>T41</b>	<b>-0.6496 (se .04190)</b>	<b>-0.6497 (se .0417)</b>

# Code Example

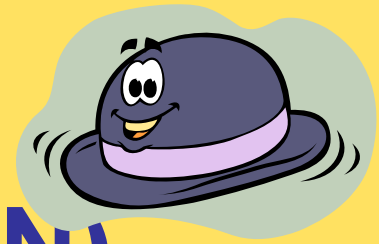
## PROC SURVEYLOGISTIC



- **Proc Surveylogistic;**
- **Strata emerg distsize  
schl\_lvl / list;**
- **Model  
T42AB(Event='1')=T40  
T41 / Stb Rsq;**
- **Weight WGTD;**
- **TITLE 'Surveylogistic  
T42AB=T40 T41';**
- **Run;**

# Code Example

## PROC RLOGIST (SUDAAN)



- Proc Rlogist data=one  
filetype=SAS  
design=strwr;
- Nest emerg distsize  
schl\_lvl;
- Weight WGTD;
- **Subgroup T40 ;**
- **Class T40 / nofreq;**
- Test Satadjchi adjwaldf;
- Model T42AB=T40 T41;
- Setenv colwidth=9 decwidth=2  
colspce=2;
- Print Beta Sebeta deft T\_beta  
P\_beta Df Satadjaf Satadchi  
Adjwaldf satadchip Adjwaldp /  
risk=all deffmt=F6.2  
Satadjdfmt=F6.3  
Satadchifmt=F7.2  
Adjwaldffmt=F7.2  
Betafmt=F8.4 Sebetafmt=F8.4  
Betafmt=F7.4 Deftfmt=F6.2  
Satadchpfmt=F7.4  
Adjwaldpfmt=F7.4;
- Rtitle "Sudaan Reg Proc  
T42AB=T40 T41";
- Run;

# Comparison of Information in Logistic Regression

<b>Statistical Concept</b>	<b>SAS PROC SURVEYLOGISTIC</b>	<b>SUDAAN PROC RLOGIST</b>
<b>Dependent Variable Information</b>	<b>Response Profile</b>	<b>Sample and Population Counts for Response Var</b>
<b>Betas and SEs</b>	<b>Analysis of Maximum Likelihood Estimates Table</b>	<b>Independent Variables and Effects Table</b>

# Results of Logistic Regression

	<b>SAS Proc SURVEYLOGISTIC</b>	<b>SUDAAN PROC RLOGIST</b>
<b>Betas and Standard Errors</b>		
<b>Intercept</b>	<b>-0.3926 (se 1.0329)</b>	<b>0.842 (se .924474)</b>
<b>T41</b>	<b>-0.00049 (se 0.0149)</b>	<b>-0.00 (se .014905)</b>
<b>T40 (Females)</b>	<b>0.6172 (se .3843)</b>	<b>-0.617 (se .3929)</b>



# Limitations of Each Package

---

## **SAS**

- Does not offer other methods such as BRR
- Limited number of procedures and options
- Does not easily provide actual design effects for all PROC SURVEY procedures

## **SUDDAN**

- Documentation difficult to understand
- Tech Support is not as fully staffed or responsive
- Additional \$ for licensing and training is required
- More labor intensive and more difficult to code
- Output is paper intensive



# New Features

---

- **SAS**

- Inclusion of Proc Surveyfreq
- Inclusion of Proc Surveylogistic

- **Sudaan**

- Inclusion of Class Statement
- Being able to work with binary responses (0,1)



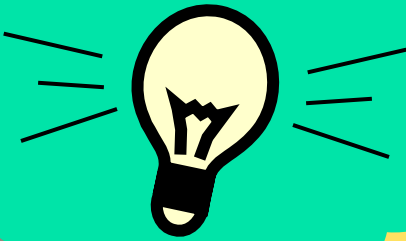
# Conclusion

---

- At this point in time, we as programmers must wear more than **ONE** hat to correctly analyze complex sample designs
- Neither program is sufficient to provide both ease and variety of choice
- We must work with the **manual** transmission mode of SUDAAN in conjunction with the **automatic** transmission mode of SAS

# CONCLUSION

Two Hats are  
Necessary





# The future of CSD

---

- Maybe a future SAS Programmer will not have wear two hats – only the future knows!



■ Oct 2004



■ June 2007



# Contact Information

---

- Katherine Baisden
- 6187 Lodi Lane
- Saline, MI 48176
- Email: [katherine.baisden@sri.com](mailto:katherine.baisden@sri.com)
- Phone: 734-316-2910