#SASGF

VIRTUAL

SAS® GLOBAL FORUM 2021

Survey data analysis made easy with SAS

Melanie Dove, UC Davis

Melanie Dove is an assistant adjunct professor in the Department of Public Health Sciences, Division of Health Policy and Management at UC Davis. She conducts research on the impact tobacco control policies have on health behaviors. She also teaches an introduction to SAS class.

Katherine Heck, UC San Francisco

Katherine Heck is a researcher at the Center for Health Equity, UC San Francisco. She is the data manager and lead analyst for the California Maternal and Infant Health Assessment, an annual stratified sample survey of women giving birth in California.

Outline

Why use survey procedures

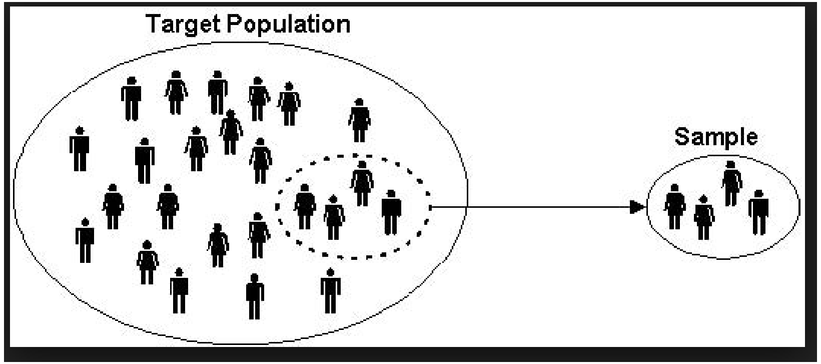
Key survey design features

Examples:

- PROC SURVEYFREQ
- PROC SURVEYMEANS
- PROC SURVEYLOGISTIC
- PROC SURVEYREG

What is a survey?

A sample of individuals to represent a population:



Examples: National Health and Nutrition Examination Survey, National Health Interview Survey

Why do we need to use survey procedures?

To take into account the design of the survey

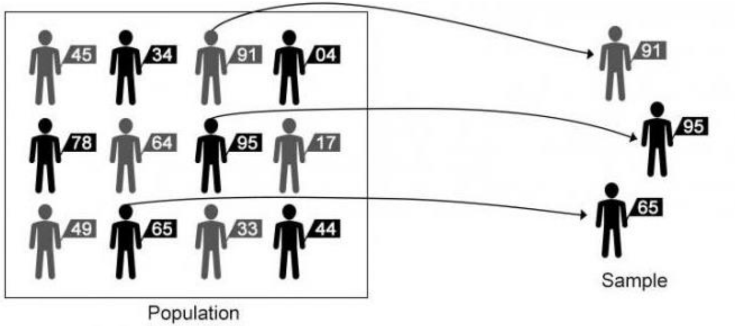
- Sampling
- Weighting

Why do we need to use survey procedures?

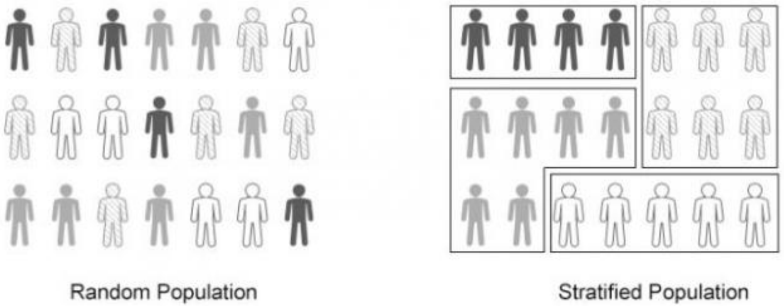
Sampling

Sampling

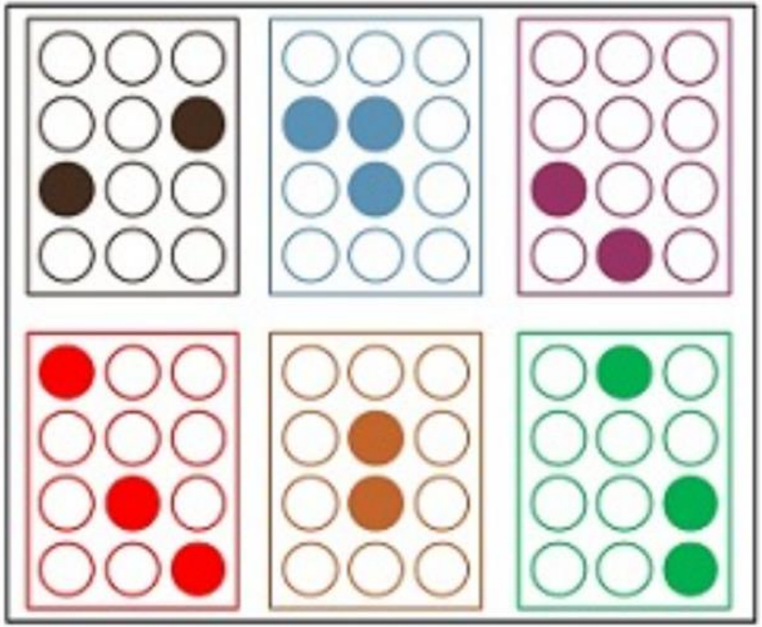
Simple random



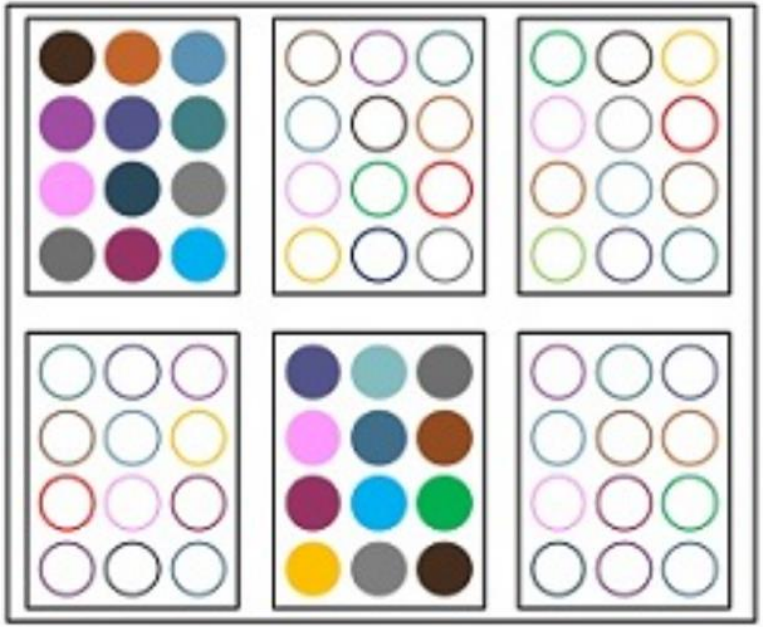
Stratified



Sampling



Stratified

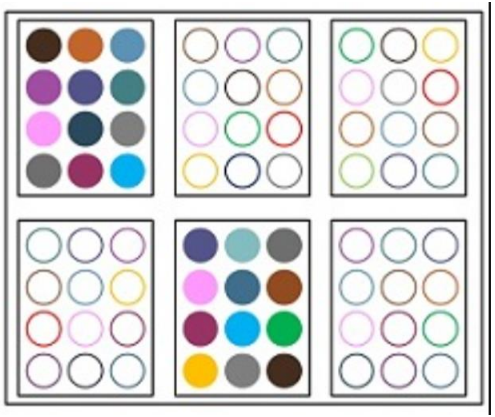


Cluster

Sampling

Individuals within clusters are similar

- Overestimate variance – significance

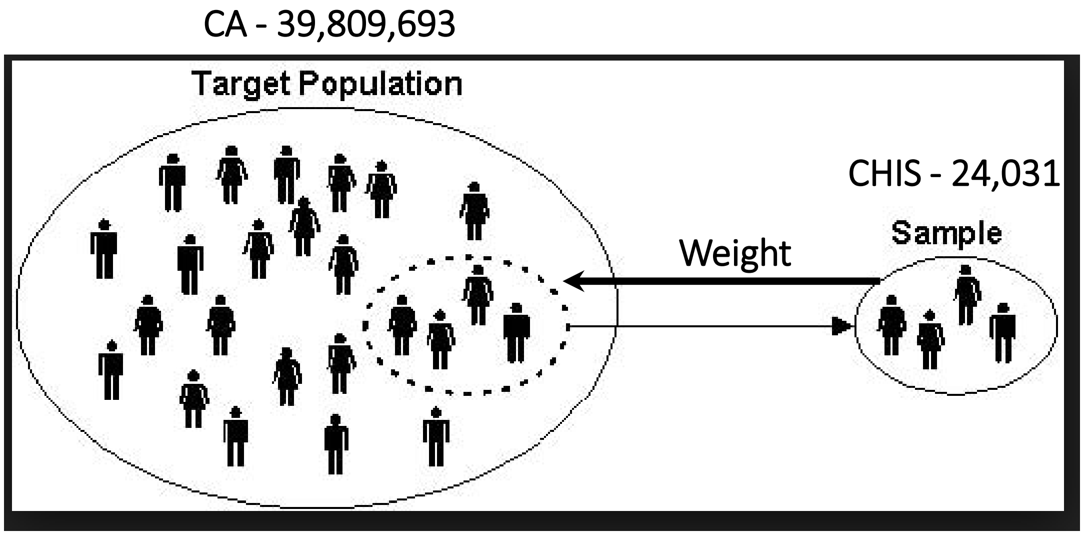


Weighting

Weight: a value indicating the number of people the respondent represents

Weighting

Weight: a value indicating the number of people the respondent represents



Weighting

Corrects for:

- differing probability of sampling in different clusters or strata
- nonresponse

Key SAS Survey Design Features

- Stratification: **STRATA** statement

- Clustering: **CLUSTER** statement

- Weighting: **WEIGHT** statement

Key SAS Survey Design Features

Subpopulation analyses:

DOMAIN statement or “flag” variables

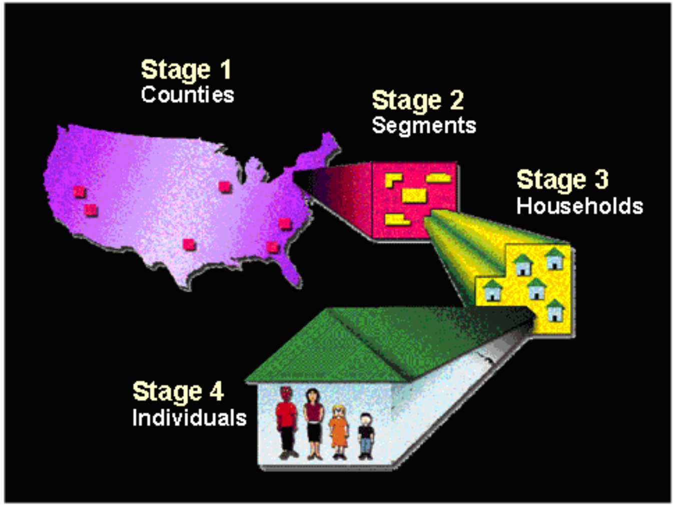
- *Do not use “where”, “by”, or “if” to subset data*

Examples- NHANES and CHIS

National Health and Nutrition Examination Survey (**NHANES**)

- stratified, cluster design
- in person survey
- one weight per person

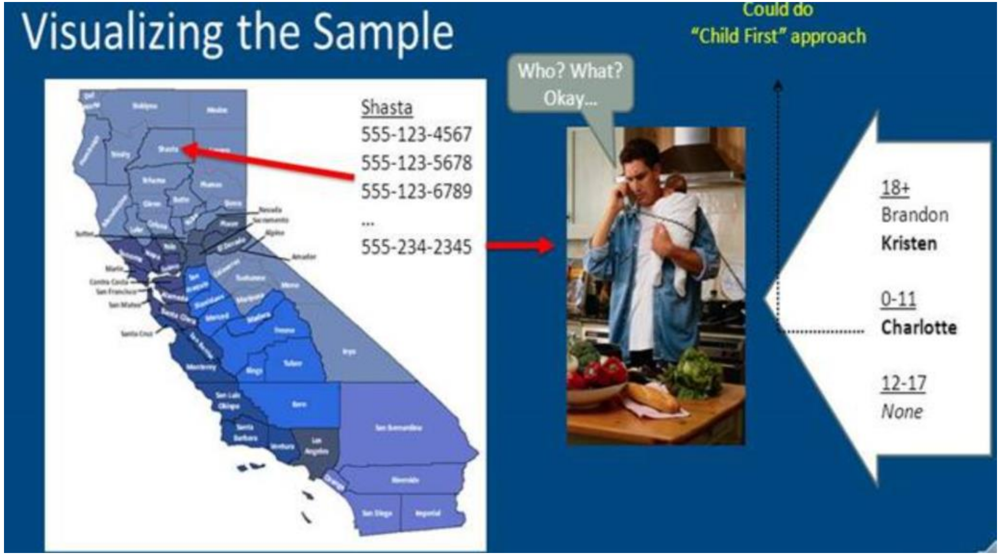
Four Stages of NHANES Sampling Procedure



Examples – NHANES and CHIS

California Health Interview Survey (**CHIS**)

- stratified random sample
- telephone survey
- replicate weights
  - 80 weights per person



Visualizing the Sample

Could do "Child First" approach

Who? What? Okay...

Shasta
555-123-4567
555-123-5678
555-123-6789
...
555-234-2345

18+
Brandon
Kristen

0-11
Charlotte

12-17
None

SAS code!

```
0 0 1 1 1 1 1 0 1 1 1 1 0 1 1 1 0 0 0 0 0
0 1 1 1 1 0 1 0 0 0 1 1 0 1 1 0 0 0 1 0 1
1 0 1 1 1 0 1 1 0 0 0 1 1 0 1 1 0 0 0 1 0
0 0 0 1 0 1 0 0 0 1 0 0 1 1 0 0 0 1 0 0 1
1 1 1 1 1 0 0 0 1 0 0 1 1 0 0 1 0 0 1 1 0
0 1 1 1 1 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1 1
1 1 1 0 0 0 1 1 0 1 0 1 1 1 0 0 0 1 0 0 0
1 0 0 0 0 1 1 1 0 0 1 0 0 0 1 0 1 0 0 0 1
1 1 1 0 0 0 1 1 0 0 1 0 0 0 0 1 0 0 0 0 0
0 0 0 0 1 1 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1
1 1 1 1 1 0 1 0 0 1 0 0 0 0 0 1 1 0 1 1 1
0 0 1 0 1 1 1 1 0 1 1 0 0 0 0 0 1 0 0 1 0
0 0 1 0 1 1 0 1 1 1 0 0 0 0 1 1 0 1 1 1 1
1 1 0 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 1 0 0
1 0 1 1 0 1 1 0 1 0 0 0 1 0 1 0 0 1 1 0 0
0 1 1 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1
1 1 0 1 1 1 0 1 0 0 0 1 0 1 1 1 0 1 1 1 1
1 1 0 0 0 1 1 0 0 1 1 1 0 1 1 0 0 1 1 1 1
1 1 1 0 1 0 0 1 0 1 1 1 1 0 0 1 0 1 1 0 1
0 0 0 0 0 0 1 1 0 0 1 0 1 1 0 0 0 1 1 1 1
1 1 0 1 1 1 0 1 0 0 1 0 1 1 0 1 1 1 0 1 0
0 1 1 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1
```

PROC SURVEYFREQ -syntax

Taylor is the default

```
proc surveyfreq data=dataset varmethod=taylor;
strata   stratum;
cluster PSU;
weight   weightvar;
tables   agegrp;
run;
```

```
proc freq data=dataset;
tables agegrp;
run;
```

PROC SURVEYFREQ – NHANES example

```
proc surveyfreq data=NHANES;
strata   sdmvstra;
cluster  sdmvpsu;
weight   WTINT2YR;
tables   ridreth3/ cl;
run;
```

Confidence limits

```
proc freq data=NHANES;
table ridreth3/binomial (level=1) cl;
run;
```

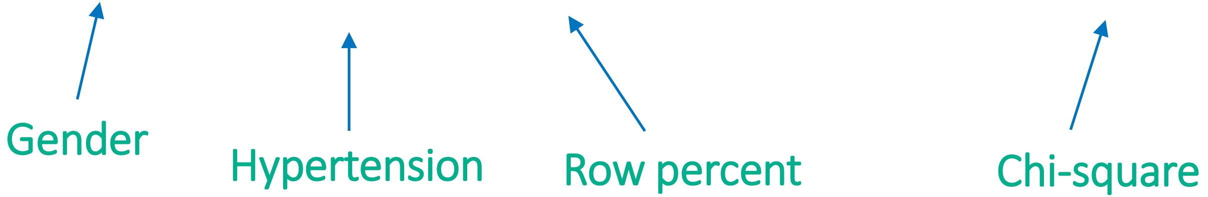
Results with and without adjusting for survey factors

Race/ethnicity	Without survey procedures			With survey procedures		
	n	Percent	95% CI	n	Percent	95% CI
Mexican American	1730	17.0	16.3, 17.7	1730	11.1	6.8, 15.3
Other Hispanic	960	9.4	8.9, 10.0	960	6.0	3.8, 8.3
NH White	3674	36.1	35.2, 37.0	3674	62.2	54.5, 70.0
NH Black	2267	22.3	21.5, 23.1	2267	12.1	8.4, 15.8
NH Asian	1074	10.6	10.0, 11.2	1074	5.2	3.9, 6.5
other	470	4.6	4.2, 5.0	470	3.4	2.4, 4.4

PROC SURVEYFREQ- CHIS example

Does hypertension differ by gender?

```
proc surveyfreq data=CHIS varmethod=jackknife;
weight   rakedw0;
repweight  rakedw1-rakedw80 / JKCOEFS=1 ;
tables   srsex * ab29 / row cl nototal chisq ;
run;
```

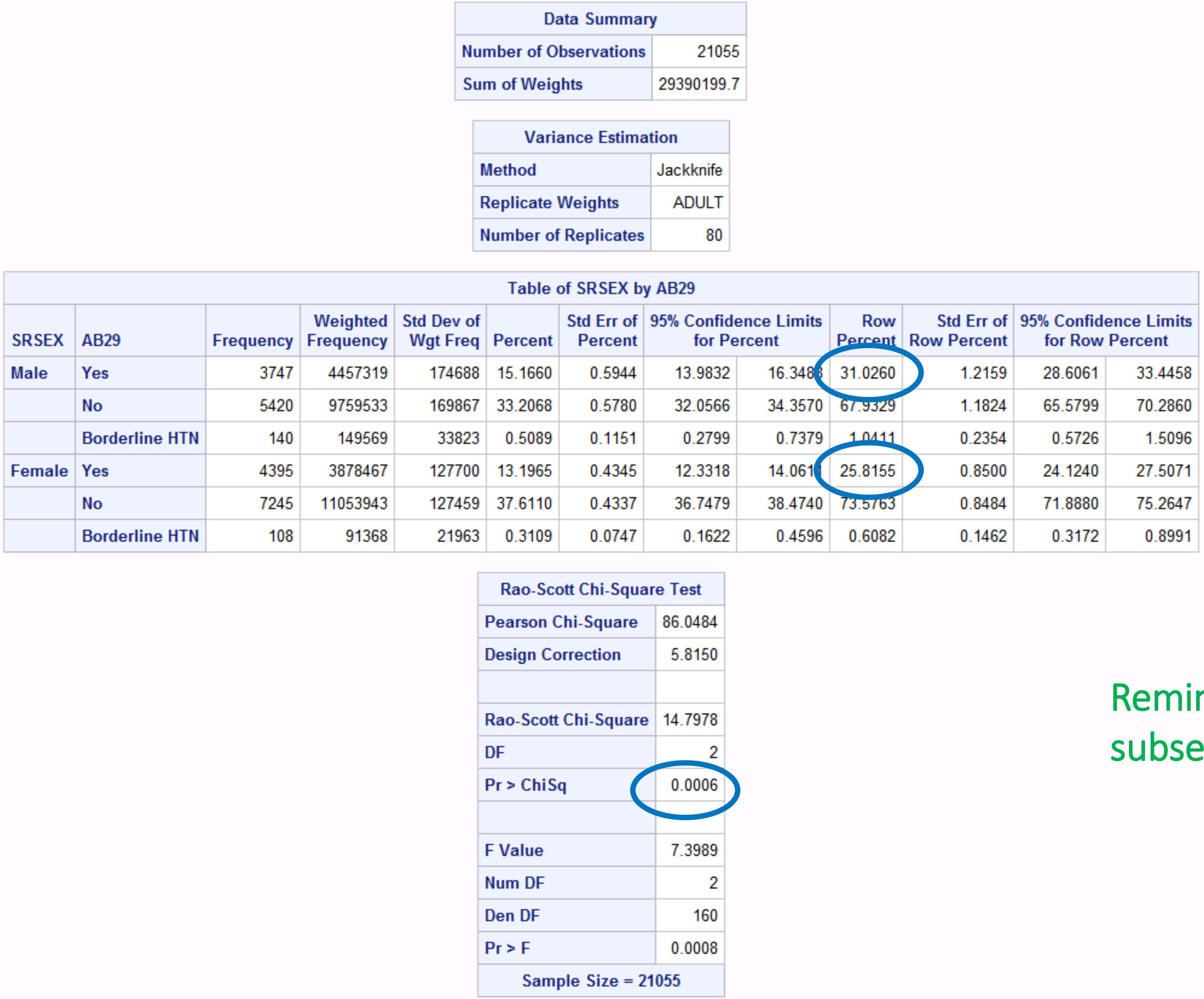


Gender

Hypertension

Row percent

Chi-square



Data Summary	
Number of Observations	21055
Sum of Weights	29390199.7

Variance Estimation	
Method	Jackknife
Replicate Weights	ADULT
Number of Replicates	80

Table of SRSEX by AB29

SRSEX	AB29	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	95% Confidence Limits for Percent		Row Percent	Std Err of Row Percent	95% Confidence Limits for Row Percent	
Male	Yes	3747	4457319	174688	15.1660	0.5944	13.9832	16.3488	31.0260	1.2159	28.6061	33.4458
	No	5420	9759533	169867	33.2068	0.5780	32.0566	34.3570	67.9329	1.1824	65.5799	70.2860
	Borderline HTN	140	149569	33823	0.5089	0.1151	0.2799	0.7379	1.0411	0.2354	0.5726	1.5096
Female	Yes	4395	3878467	127700	13.1965	0.4345	12.3318	14.0610	25.8155	0.8500	24.1240	27.5071
	No	7245	11053943	127459	37.6110	0.4337	36.7479	38.4740	73.5763	0.8484	71.8880	75.2647
	Borderline HTN	108	91368	21963	0.3109	0.0747	0.1622	0.4596	0.6082	0.1462	0.3172	0.8991

Rao-Scott Chi-Square Test	
Pearson Chi-Square	86.0484
Design Correction	5.8150
Rao-Scott Chi-Square	14.7978
DF	2
Pr > ChiSq	0.0006
F Value	7.3989
Num DF	2
Den DF	160
Pr > F	0.0008
Sample Size = 21055	

Reminder: Don’t subset the data

PROC SURVEYMEANS- CHIS example

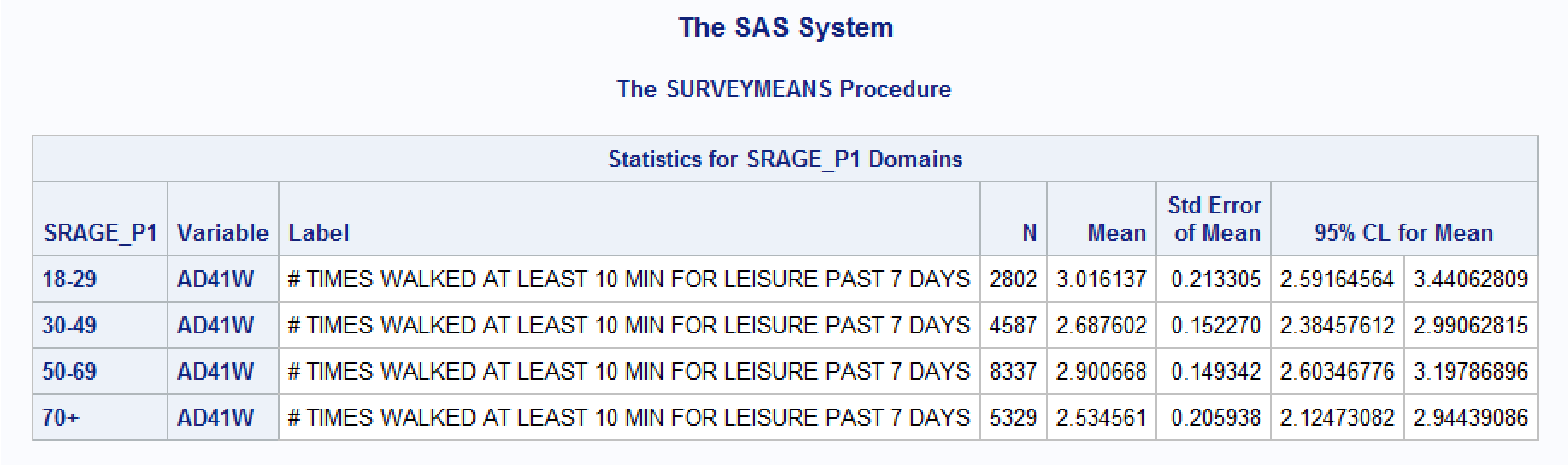
Does the frequency of walking for leisure differ by age?

```
proc surveymeans data=CHIS varmethod=JACKKNIFE;
weight     rakedw0;
repweight rakedw1-rakedw80 / JKCOEFS=1  ;
var        AD41W  ;
domain     SRAGE_P1  ;
run;
```

AD41W = how often walked

Domain = group(s) of interest
SRAGE_P1 = age

Results



The SAS System

The SURVEYMEANS Procedure

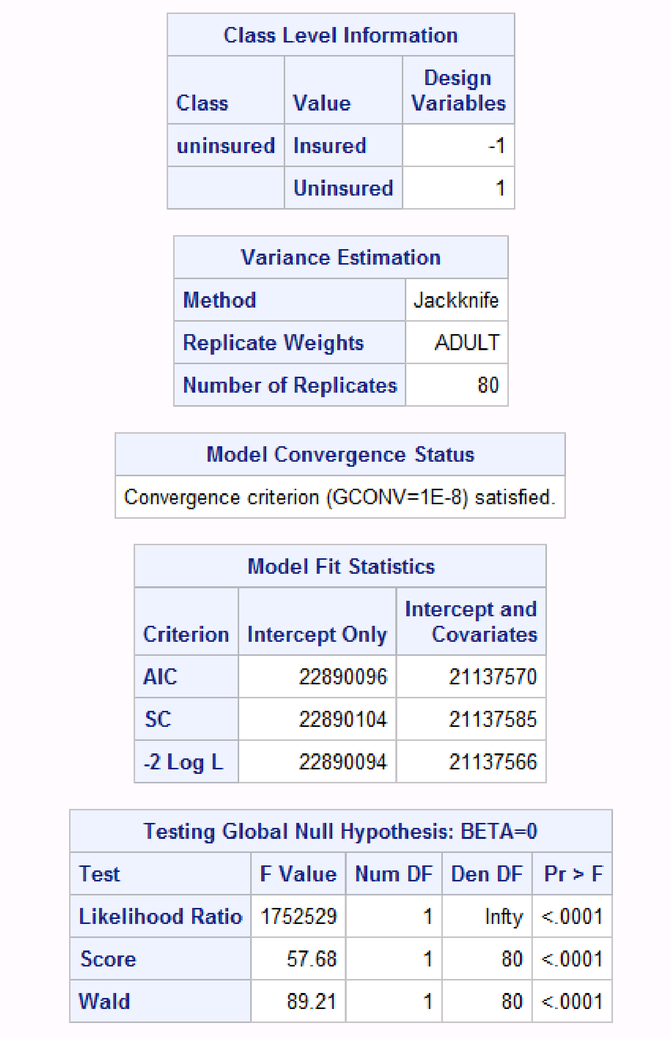
Statistics for SRAGE_P1 Domains

SRAGE_P1	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
18-29	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	2802	3.016137	0.213305	2.59164564	3.44062809
30-49	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	4587	2.687602	0.152270	2.38457612	2.99062815
50-69	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	8337	2.900668	0.149342	2.60346776	3.19786896
70+	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	5329	2.534561	0.205938	2.12473082	2.94439086

PROC SURVEYLOGISTIC – CHIS example

Are you more likely to not have a usual source of care if you are uninsured?

```
proc surveylogistic data=CHIS varmethod=JACKKNIFE;
weight    rakedw0;
repweight rakedw1-rakedw80/JKCOEFS=1;
class     uninsured (ref='Insured’)/ param=ref;
model     nousual (descending) = uninsured ;
format    uninsured unins.;
run;
```



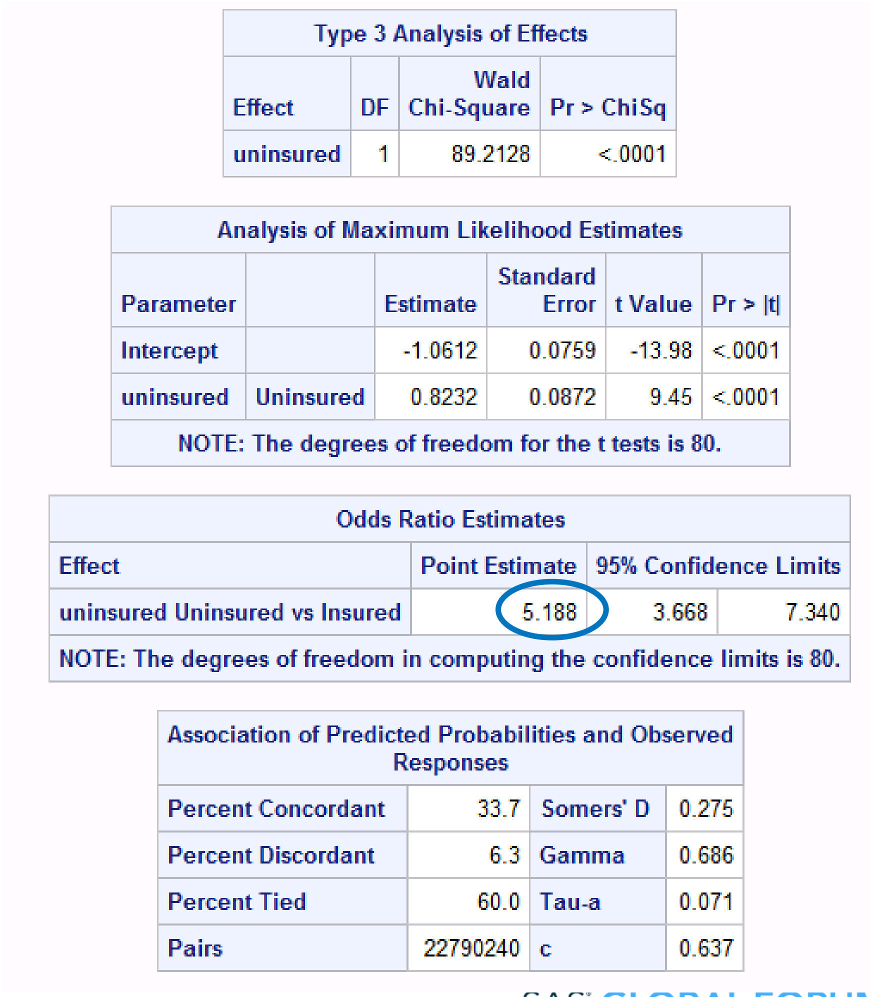
Class Level Information		
Class	Value	Design Variables
uninsured	Insured	-1
	Uninsured	1

Variance Estimation	
Method	Jackknife
Replicate Weights	ADULT
Number of Replicates	80

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	22890096	21137570
SC	22890104	21137585
-2 Log L	22890094	21137566

Testing Global Null Hypothesis: BETA=0				
Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	1752529	1	Infty	<.0001
Score	57.68	1	80	<.0001
Wald	89.21	1	80	<.0001



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
uninsured	1	89.2128	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-1.0612	0.0759	-13.98	<.0001
uninsured	Uninsured	0.8232	0.0872	9.45	<.0001
NOTE: The degrees of freedom for the t tests is 80.					

Odds Ratio Estimates			
Effect	Point Estimate	95% Confidence Limits	
uninsured Uninsured vs Insured	5.188	3.668	7.340
NOTE: The degrees of freedom in computing the confidence limits is 80.			

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	33.7	Somers' D	0.275
Percent Discordant	6.3	Gamma	0.686
Percent Tied	60.0	Tau-a	0.071
Pairs	22790240	c	0.637

PROC SURVEYREG – NHANES example

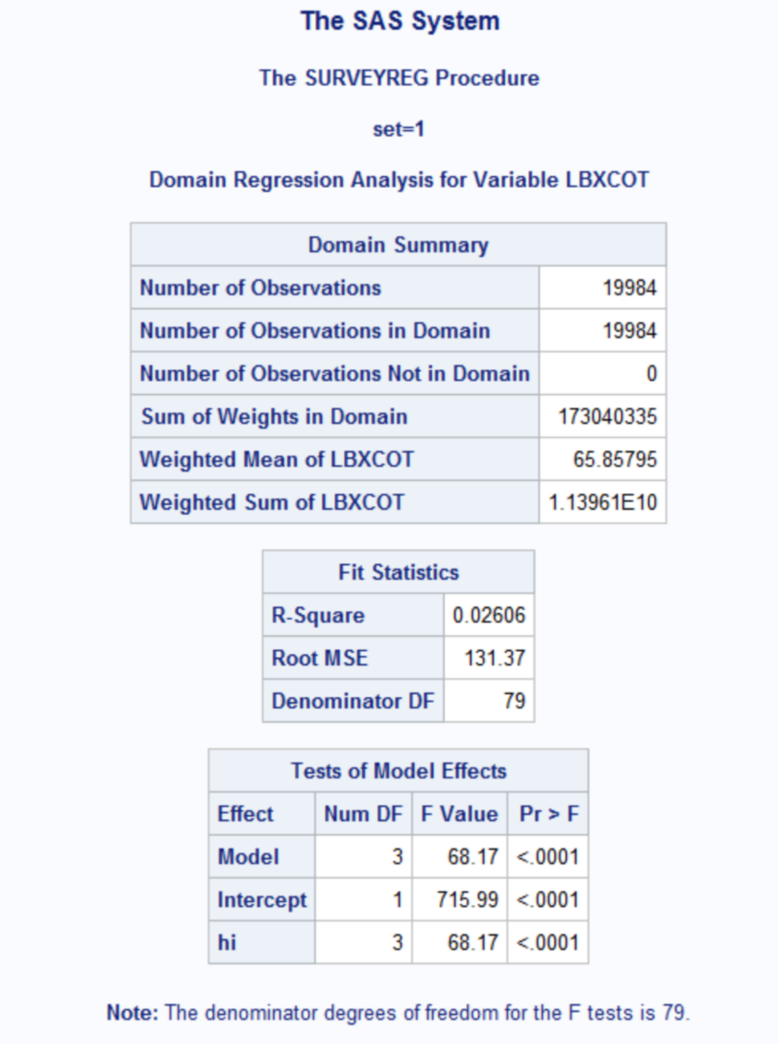
Does cotinine differ by health insurance status?

```
PROC SURVEYREG DATA = NHANES;
STRATUM sdmvstra;
CLUSTER sdmvpsu;
WEIGHT mec10yr;
DOMAIN set;
CLASS hi;
MODEL lbxcot=hi / solution clparm;
run;
```

hi = health insurance

Requests parameter estimates

Confidence limits



The SAS System

The SURVEYREG Procedure

set=1

Domain Regression Analysis for Variable LBXCOT

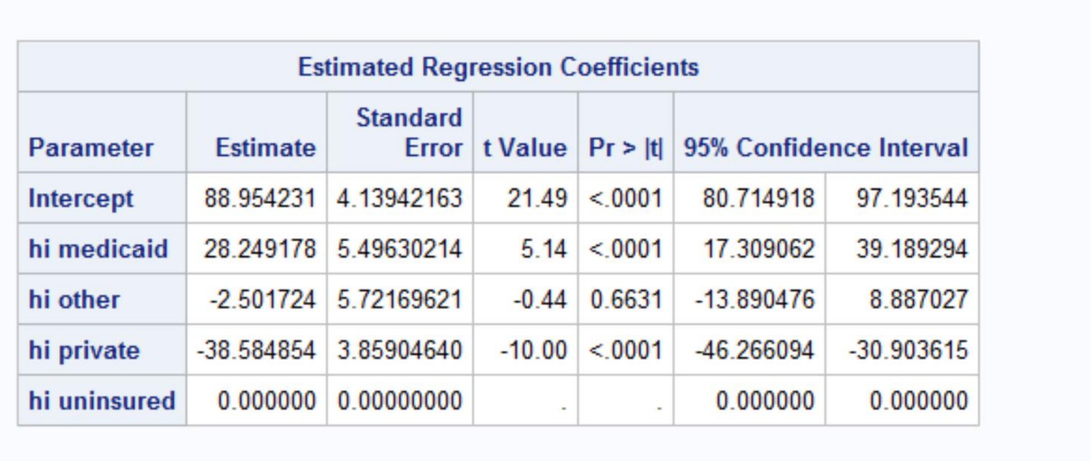
Domain Summary	
Number of Observations	19984
Number of Observations in Domain	19984
Number of Observations Not in Domain	0
Sum of Weights in Domain	173040335
Weighted Mean of LBXCOT	65.85795
Weighted Sum of LBXCOT	1.13961E10

Fit Statistics	
R-Square	0.02606
Root MSE	131.37
Denominator DF	79

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	3	68.17	<.0001
Intercept	1	715.99	<.0001
hi	3	68.17	<.0001

Note: The denominator degrees of freedom for the F tests is 79.



Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Interval	
Intercept	88.954231	4.13942163	21.49	<.0001	80.714918	97.193544
hi medicaid	28.249178	5.49630214	5.14	<.0001	17.309062	39.189294
hi other	-2.501724	5.72169621	-0.44	0.6631	-13.890476	8.887027
hi private	-38.584854	3.85904640	-10.00	<.0001	-46.266094	-30.903615
hi uninsured	0.000000	0.00000000	.	.	0.000000	0.000000

Conclusion

- Survey is a sample of the population
- Adjust for the survey design features in SAS
- Examples using CHIS and NHANES data

Thank you!

Contact Information:

mdove@ucdavis.edu

katherine.heck@ucsf.edu