Melanie Dove, MPH, ScD
UC Davis

mdove@ucdavis.edu

- QSCERT-PC Postdoc, UCD
- Surveys: National Health and Nutrition Examination Survey (NHANES), California Health Interview Survey (CHIS)
- Previously: California Department of Public Health, CDC/NCHS

Katherine Heck, MPH
UC San Francisco

Katherine.Heck@ucsf.edu

- Research analyst, UCSF
- Surveys: California Maternal and Infant Health Assessment (MIHA) survey, Listening to Mothers-CA
- Previously: California Department of Public Health, CDC/NCHS

2018

WESTERN USERS OF SAS SOFTWARE

SACRAMENTO, CA.

Survey data analysis made easy with SAS

Melanie Dove, MPH, ScD
UC Davis

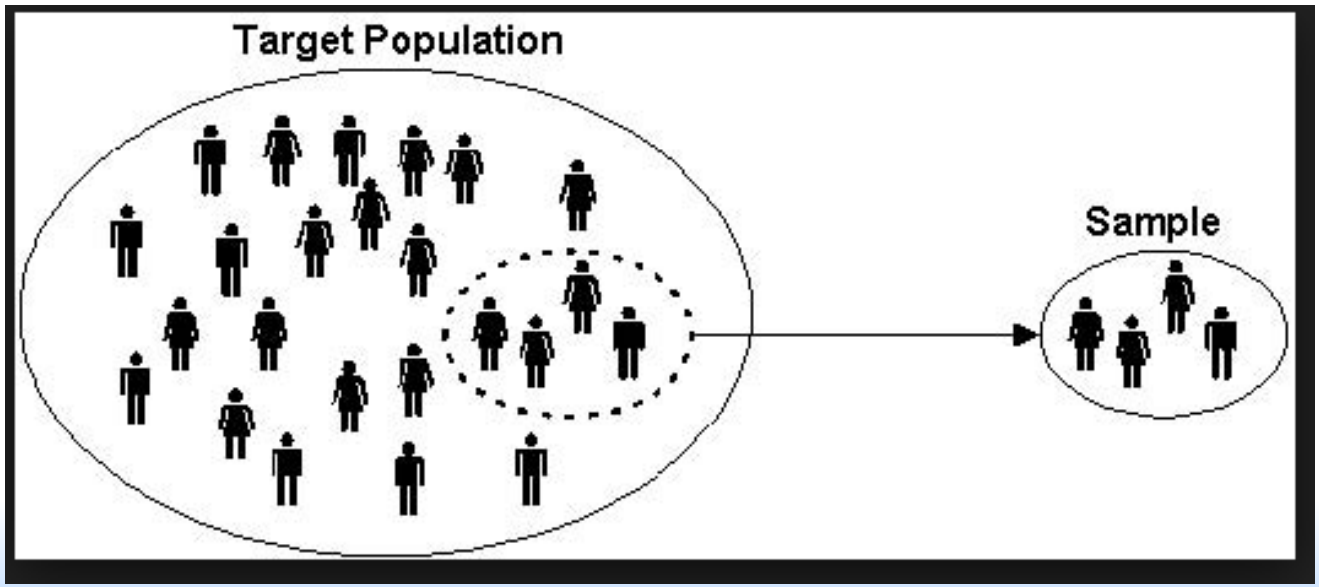
Katherine Heck, MPH
UC San Francisco

Overview

- Background
- Survey design factors (weight and variance)
- How to analyze the data

Surveys

- Representativeness: Using a sample of individuals to represent a population



Survey data

- Different types:
  - Health, economic, marketing, sociology, psychology
- Cross sectional
- Data collection methods
  - In person, phone, mail, online

California Health Interview Survey (CHIS)

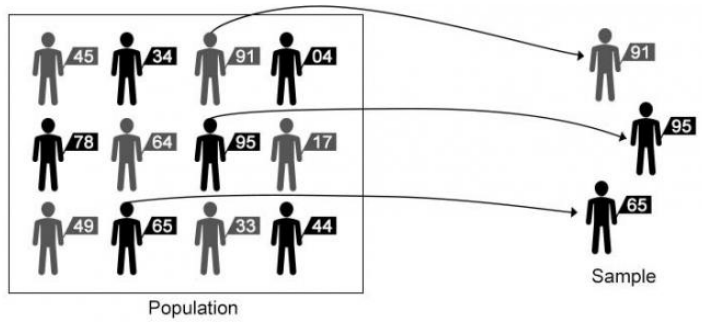
Health survey that represents California

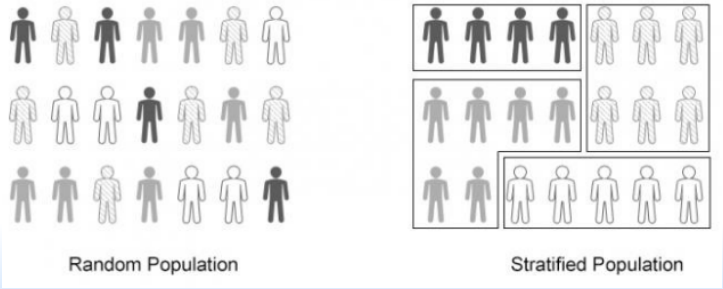
California's population: 39,809,693 (1/1/2018)

State of California, Department of Finance, E-1 Population Estimates for Cities, Counties and the State with Annual Percent Change — January 1, 2017 and 2018. Sacramento, California, May 2018.

Sampling

- Convenience
- Simple random
- Stratified

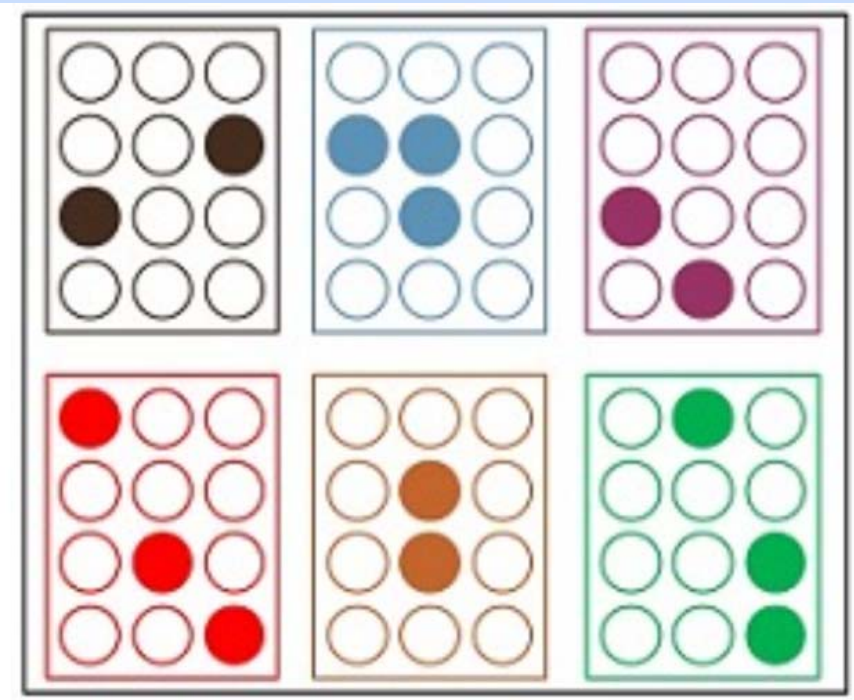




Sampling

- Cluster
  - within specified groups or geographic areas
  - sometimes called primary sampling units (PSUs)

- Stratification
  - select a specified number of individuals from a particular population group
  - can be used for oversampling



Stratified

Cluster

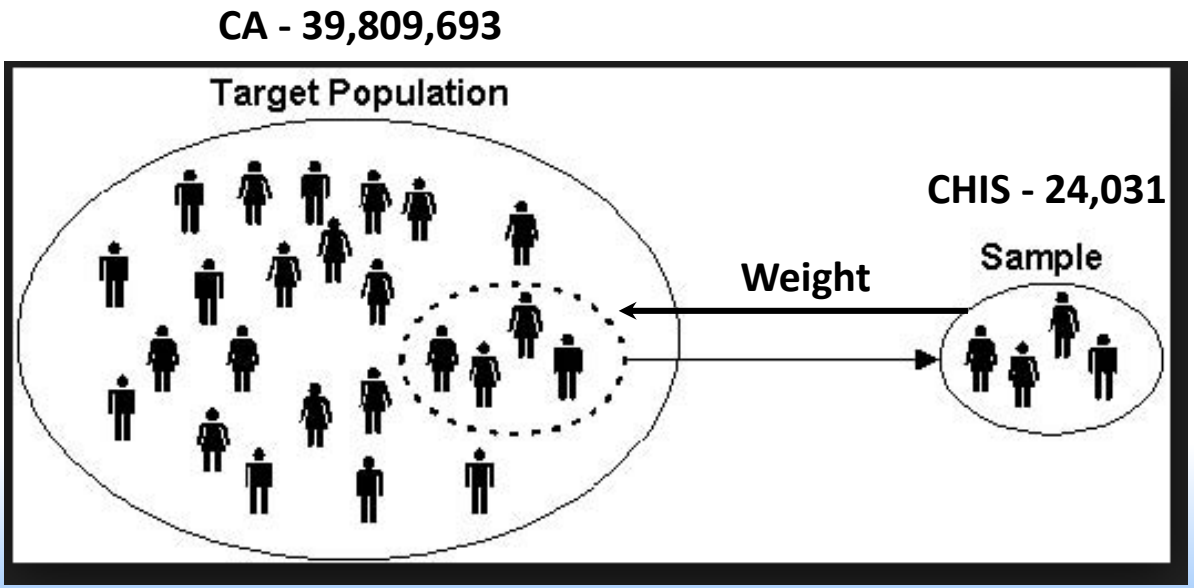
Variance

- Individuals within clusters are similar
  - Overestimate variance – significance

- Need to account for the sample design if any stratification, clustering, or weighting was used

Weighting

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



Weights

- **Single** weight variable

  -or-

- **Replicate** weights, a series of weight variables which must be used in combination to correctly weight the sample

SAS survey procedures

SAS survey procedures:

- **Proc Surveyfreq**: Frequencies, crosstabs
- **Proc Surveymeans**: Means, medians
- **Proc Surveyreg**: Linear regression
- **Proc Surveylogistic**: Logistic regression
- **Proc Surveyphreg**: Cox proportional hazards model
- **Proc Surveyselect**: Sample selection

Procedures can produce standard errors and confidence intervals

Results with and without survey procedures: confidence intervals

Example: CHIS, 2016 adult survey

Weighted percent and confidence interval

* Ever diagnosed with asthma, age 30-34

- Proc Freq results: 13.89% (13.85%-13.93%)
- Proc Surveyfreq results: 13.89% (9.97%-17.80%)

Survey components and syntax

- Stratification: STRATA statement
- Clustering: CLUSTER statement
- Weighting: WEIGHT statement (and REPWEIGHT if using replicate weights)
- Subpopulation analyses: DOMAIN statement or “flag” variables
  - *Do not use “where” to subset data*

Survey procedure examples

Proc Surveyfreq - stratum/cluster

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

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

Proc Surveyfreq - stratum/cluster

Missing data

```
proc surveyfreq nomcar data=dataset total=c.sampfrac;
strata  stratum;
cluster PSU;
weight  weightvar;
tables  agegrp * disease / row col cl;
format  agegrp agegrpf.;
run;
```

Finite pop correction

Row %

Col %

Confidence limits

Proc Surveyfreq - replicate weights

Variance estimation method

```
proc surveyfreq data=dataset varmethod=jackknife;
weight  weightvar;
repweight  wtvar1-wtvar80/JKCOEFS=1;
tables  agegrp * disease / row cl ;
format  agegrp agegrpf.;
run;
```

Two weighting statements

Libname statement

```
libname CHIS ‘C:\HOW\Heck’;

data adult;
     set chis.adult;
run;
```

CHIS age variable

VARNAME: SRAGE_P1 QNAME16:

QNAME14: QNAME11:

LABEL: SELF-REPORTED AGE (PUF 1 YR RECODE)

VALUE:			FREQ	%
	18	18-25 YEARS	1998	9.49
	26	26-29 YEARS	804	3.82
	30	30-34 YEARS	1087	5.16
	35	35-39 YEARS	1058	5.02
	40	40-44 YEARS	1156	5.49
	45	45-49 YEARS	1286	6.11
	50	50-54 YEARS	1753	8.33
	55	55-59 YEARS	1935	9.19
	60	60-64 YEARS	2207	10.48
	65	65-69 YEARS	2442	11.60
	70	70-74 YEARS	1833	8.71
	75	75-79 YEARS	1425	6.77
	80	80-84 YEARS	1093	5.19
	85	85+ YEARS	978	4.64

UNIVERSE: ALL ADULTS

Proc Surveyfreq - age

```
proc surveyfreq data=adult varmethod= ?????;
weight   ????? ;
repweight   ????? /JKCOEFS=1;
tables   ?????   ;
run;
```

Proc Surveyfreq - age

```
proc surveyfreq data=adult varmethod=jackknife;
weight  rakedw0;
repweight  rakedw1-rakedw80/JKCOEFS=1;
tables  srage_p1 / cl  ;
run;
```

Proc Surveyfreq - Results

The SAS System

The SURVEYFREQ Procedure

Data Summary	
Number of Observations	21055
Sum of Weights	29390199.7

Variance Estimation	
Method	Jackknife
Replicate Weights	ADULT
Number of Replicates	80

SELF-REPORTED AGE (PUF 1 YR RECODE)							
SRAGE_P1	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent	95% Confidence Limits for Percent	
18	1998	4335525	93184	14.7516	0.3171	14.1206	15.3826
26	804	2077862	93184	7.0699	0.3171	6.4390	7.7009
30	1087	3061032	69368	10.4151	0.2360	9.9454	10.8848
35	1058	2287010	69368	7.7815	0.2360	7.3118	8.2512
40	1156	2637708	69623	8.9748	0.2369	8.5034	9.4462
45	1286	2439452	69623	8.3002	0.2369	7.8288	8.7716
50	1753	2540550	48634	8.6442	0.1655	8.3149	8.9735
55	1935	2120498	100679	7.2150	0.3426	6.5333	7.8967
60	2207	2640511	123347	8.9843	0.4197	8.1491	9.8195
65	2442	1893775	67029	6.4436	0.2281	5.9897	6.8974
70	1833	1218959	58178	4.1475	0.1980	3.7536	4.5414
75	1425	916757	76999	3.1193	0.2620	2.5979	3.6406
80	1093	691306	51784	2.3522	0.1762	2.0015	2.7028
85	978	529254	36193	1.8008	0.1231	1.5557	2.0459
Total	21055	29390200	1.71143E-6	100.000			

CHIS: Asthma variable

VARNAME: AB17 QNAME16: QA15_B2

QNAME14: QA13_B2 QNAME11: QA11_B2

LABEL: DOCTOR EVER TOLD HAVE ASTHMA

VALUE:			FREQ	%
	1	YES	3257	15.47
	2	NO	17798	84.53

UNIVERSE: ALL ADULTS

Proc Surveyfreq syntax

```
proc surveyfreq data=adult varmethod=jackknife;
weight   ?????;
repweight   ????? / JKCOEFS=1 ;
tables   ????? *   ?????  / row cl nototal ;
run;
```

Category (age)

Outcome (asthma)

No row/col totals

Proc Surveyfreq syntax

```
proc surveyfreq data=c.adult varmethod=jackknife;
weight  rakedw0;
repweight  rakedw1-rakedw80 / JKCOEFS=1 ;
tables  srage_p1 * ab17 / row cl nototal ;
run;
```

Category (age)

Outcome (asthma)

No row/col totals

Proc Surveyfreq output

The SAS System

The SURVEYFREQ Procedure

Data Summary	
Number of Observations	21055
Sum of Weights	29390199.7

Variance Estimation	
Method	Jackknife
Replicate Weights	ADULT
Number of Replicates	80

Table of SRAGE_P1 by AB17

SRAGE_P1	AB17	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	
18	1	403	810412	120772	2.7574	0.4109	1.9397	3.5752	18.6924	2.5869	13.5444	23.8404
	2	1595	3525114	102455	11.9942	0.3486	11.3004	12.6879	81.3076	2.5869	76.1596	86.4556
26	1	133	273548	56650	0.9307	0.1928	0.5472	1.3143	13.1649	2.6180	7.9549	18.3749
	2	671	1804314	93983	6.1392	0.3198	5.5028	6.7755	86.8351	2.6180	81.6251	92.0451
30	1	175	425103	58322	1.4464	0.1984	1.0515	1.8413	13.8876	1.9684	9.9703	17.8049
	2	912	2635929	95833	8.9687	0.3261	8.3198	9.6176	86.1124	1.9684	82.1951	90.0297
35	1	147	329958	58962	1.1227	0.2006	0.7234	1.5219	14.4275	2.5158	9.4209	19.4340
	2	911	1957052	80282	6.6589	0.2732	6.1153	7.2025	85.5725	2.5158	80.5660	90.5791
40	1	172	258127	49060	0.8783	0.1669	0.5461	1.2105	9.7860	1.8594	6.0857	13.4863
	2	984	2379581	82090	8.0965	0.2793	7.5407	8.6524	90.2140	1.8594	86.5137	93.9143
45	1	174	331148	50948	1.1267	0.1734	0.7817	1.4717	13.5747	2.0332	9.5285	17.6208
	2	1112	2108304	75993	7.1735	0.2586	6.6589	7.6881	86.4253	2.0332	82.3792	90.4715
50	1	280	304216	41613	1.0351	0.1416	0.7533	1.3169	11.9744	1.6157	8.7591	15.1898
	2	1473	2236333	58487	7.6091	0.1990	7.2131	8.0051	88.0256	1.6157	84.8102	91.2409

Proc Surveyfreq with chi-square

```
proc surveyfreq data=c.adult varmethod=jackknife;
weight  weightvar;
repweight  wtvar1-wtvar80 / JKCOEFS=1 ;
tables  srsex * ab29 / row cl nototal chisq ;
run;
```

Gender

Hypertension

Chi-square

Proc Surveyfreq output

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	14.0611	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	

Proc Surveymeans example

CHIS 2016, number of times walked for leisure, past 7 days, by family type

```
proc surveymeans data=c.adult varmethod=JACKKNIFE;
weight     rakedw0;
repweight  rakedw1-rakedw80 / JKCOEFS=1 ;
var        AD41W ;
domain     FAMT4 ;
run;
```

AD41W = how often walked

Domain = group(s) of interest
FAMT4 = family structure

Results

The SAS System

The SURVEYMEANS Procedure

Domain Statistics in FAMT4							
FAMILY TYPE - 4 LVLS	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
Single, no kids	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	10214	2.765439	0.113541	2.53948603	2.99139231
Married, no kids	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	6544	2.957396	0.144112	2.67060365	3.24418841
Married, has kids	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	3182	2.753278	0.162631	2.42963262	3.07692377
Single, has kids	AD41W	# TIMES WALKED AT LEAST 10 MIN FOR LEISURE PAST 7 DAYS	1115	2.710728	0.542839	1.63044466	3.79101177

Proc Surveylogistic example

Usual source of care by uninsured, adults 18-64, CHIS 2016

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

Proc Surveylogistic results

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

Resources to analyze CHIS data

Analyze CHIS Data website:

http://healthpolicy.ucla.edu/chis/analyze/Pages/default.aspx

Webinar:

http://www.authorstream.com/Presentation/mattjans-1668262-chis-data-analysis-webinar-recording/

Thank you!

Questions?

Contact Information

Name: Melanie Dove

Company: UC Davis

City/State: Sacramento, CA

Phone: 916-734-8364

Email: mdove@ucdavis.edu

Contact Information

Name: Katherine Heck

Company: UCSF

City/State: San Francisco, CA

Phone: 530-219-8895

Email: Katherine.Heck@ucsf.edu