

# Survey Data Analysis in Stata

Jeff Pitblado

Associate Director, Statistical Software  
StataCorp LP

2009 Canadian Stata Users Group Meeting

---

## Outline

<b>1</b>	<b>Types of data</b>	<b>2</b>
<b>2</b>	<b>Survey data characteristics</b>	<b>4</b>
2.1	Single stage designs . . . . .	4
2.2	Multistage designs . . . . .	9
2.3	Poststratification . . . . .	10
2.4	Strata with a single sampling unit . . . . .	12
2.5	Certainty units . . . . .	15
<b>3</b>	<b>Variance estimation</b>	<b>15</b>
3.1	Linearization . . . . .	16
3.1.1	Total estimator . . . . .	16
3.1.2	Regression models . . . . .	18
3.2	Balanced repeated replication (BRR) . . . . .	20
3.3	Jackknife . . . . .	23
<b>4</b>	<b>Estimation for subpopulations</b>	<b>25</b>
<b>5</b>	<b>Summary</b>	<b>28</b>

---

## Why survey data?

- Collecting data can be expensive and time consuming.
- Consider how you would collect the following data:
  - Smoking habits of teenagers
  - Birth weights for expectant mothers with high blood pressure
- Using stages of clustered sampling can help cut down on the expense and time.

# 1 Types of data

## Simple random sample (*SRS*) data

Observations are "independently" sampled from a data generating process.

- Typical assumption: independent and identically distributed (iid)
- Make inferences about the data generating process
- Sample variability is explained by the statistical model attributed to the data generating process

## Standard data

We'll use this term to distinguish this data from survey data.

## Correlated data

Individuals are assumed not independent.

Cause:

- Observations are taken over time
- Random effects assumptions
- Cluster sampling

Treatment:

- Time-series models
- Longitudinal/panel data models
- `cluster()` option

## Survey data

Individuals are sampled from a fixed population according to a survey design.

Distinguishing characteristics:

- Complex nature under which individuals are sampled
- Make inferences about the fixed population
- Sample variability is attributed to the survey design

## Standard data

- Estimation commands for standard data:
  - `proportion`
  - `regress`
- We'll refer to these as *standard estimation commands*.

## Survey data

- Survey estimation commands are governed by the **svy** prefix.
  - `svy: proportion`
  - `svy: regress`
- **svy** requires that the data is **svyset**.

## 2 Survey data characteristics

### 2.1 Single stage designs

#### Single-stage syntax

```
svyset [psu] [weight] [, strata(varname) fpc(varname) ]
```

- Primary sampling units (PSU)
- Sampling weights – **pweight**
- Strata
- Finite population correction (FPC)

#### Sampling unit

An individual or collection of individuals from the population that can be selected for observation.

- Sampling groups of individuals is synonymous with cluster sampling.
- Cluster sampling usually results in inflated variance estimates compared to *SRS*.

#### ▷ Example

- High schools for sampling from the population of 12th graders.
- Hospitals for sampling from the population of newborns.

◀

#### Sampling weight

The reciprocal of the probability for an individual to be sampled.

- Probabilities are derived from the survey design.
  - Sampling units
  - Strata
- Typically considered to be the number of individuals in the population that a sampled individual represents.
- Reduces bias induced by the sampling design.

▷ Example

If there are 100 hospitals in our population, and we choose 5 of them, the sampling weight is  $20 = 100/5$ . Thus a sampled hospital represents 20 hospitals in the population.

Sampling weights correct for over/under sampling of sections in the population. Many times this over/under sampling is on purpose. ◀

**Strata**

In stratified designs, the population is partitioned into well-defined groups, called strata.

- Sampling units are independently sampled from within each stratum.
- Stratification usually results in smaller variance estimates compared to *SRS*.

▷ Example

- States of the union are typically used as strata in national surveys in the US.
- Demographic information like age group, gender, and ethnicity.

Although there is potential for improving efficiency by reducing sampling variability, it is usually not very practical to stratify on demographic information. ◀

**Finite population correction (FPC)**

An adjustment applied to the variance due to sampling without replacement.

- Sampling without replacement from a finite population reduces sampling variability.

□ Note

- The FPC affects the number of components in the linearized variance estimator for multi-stage designs.
- We can use `svyset` to specify an *SRS* design.

□

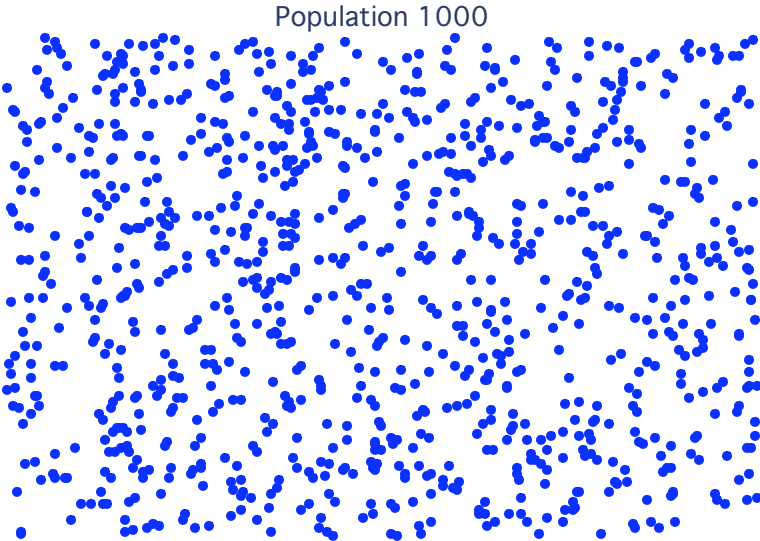
▷ Example: **svyset** for single-stage designs

1. **auto** – specifying an *SRS* design
2. **nmihs** – the National Maternal and Infant Health Survey (1988) dataset came from a stratified design
3. **fpc** – a simulated dataset with variables that identify the characteristics from a stratified and without-replacement clustered design

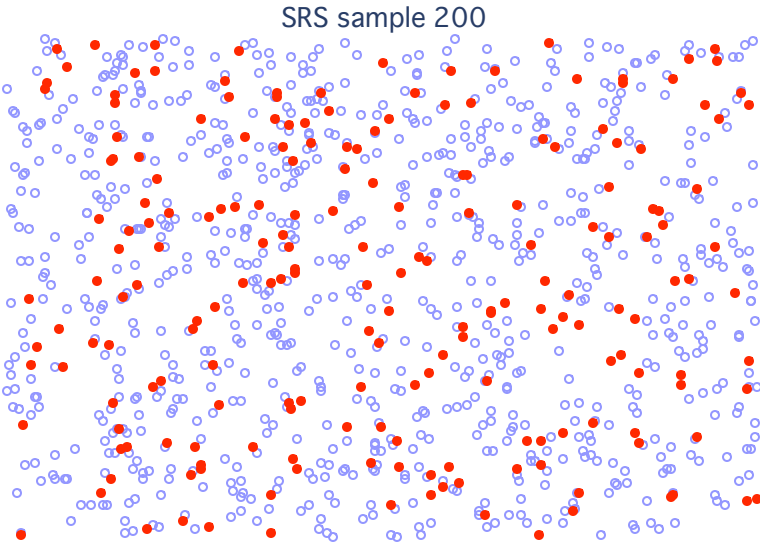
```
*** The auto data that ships with Stata
. sysuse auto
(1978 Automobile Data)
. svyset _n
      pweight: <none>
          VCE: linearized
Single unit: missing
  Strata 1: <one>
      SU 1: <observations>
      FPC 1: <zero>
*** National Maternal and Infant Health Survey
. webuse nmihs
. svyset [pw=finwgt], strata(stratan)
      pweight: finwgt
          VCE: linearized
Single unit: missing
  Strata 1: stratan
      SU 1: <observations>
      FPC 1: <zero>
*** Simulated data
. webuse fpc
. svyset psuid [pw=weight], strata(stratid) fpc(Nh)
      pweight: weight
          VCE: linearized
Single unit: missing
  Strata 1: stratid
      SU 1: psuid
      FPC 1: Nh
```

◀

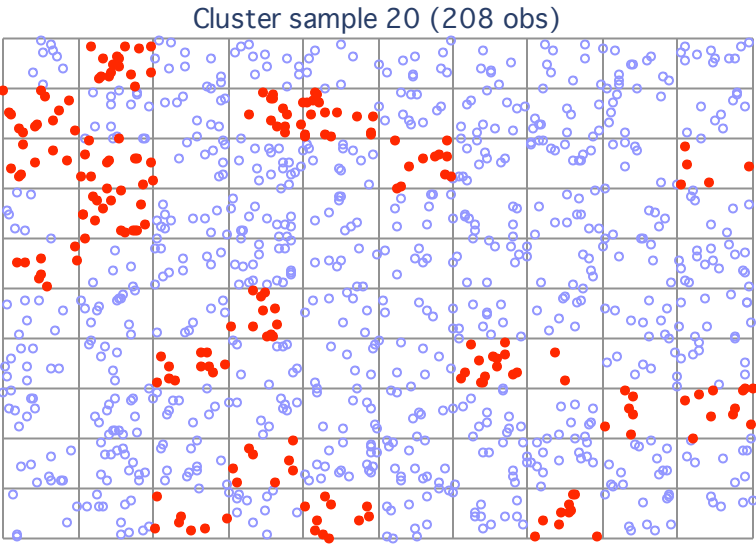
Below is a visual representation of a hypothetical population. Suppose each blue dot represents an individual.



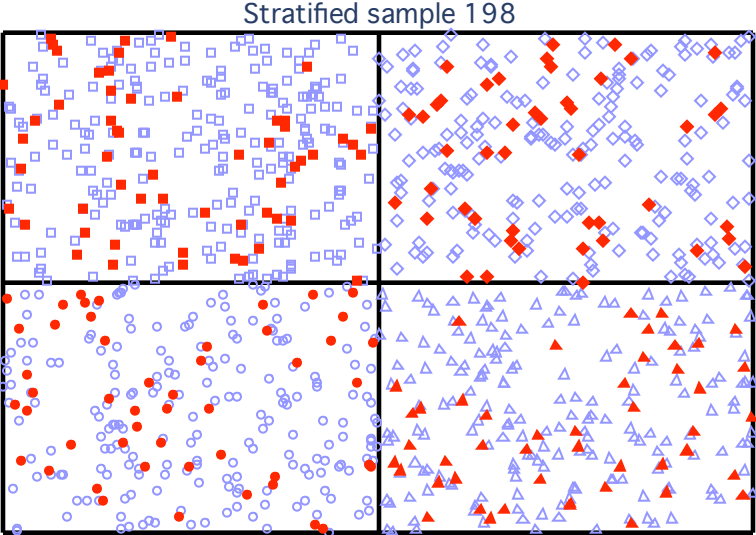
The following shows a 20% simple-random-sample. The solid symbols identify sampled individuals.



Here we partition the population into small blocks, then sample 20% of the blocks. Not all blocks contain the same number of individuals, so the sample size is a random quantity.

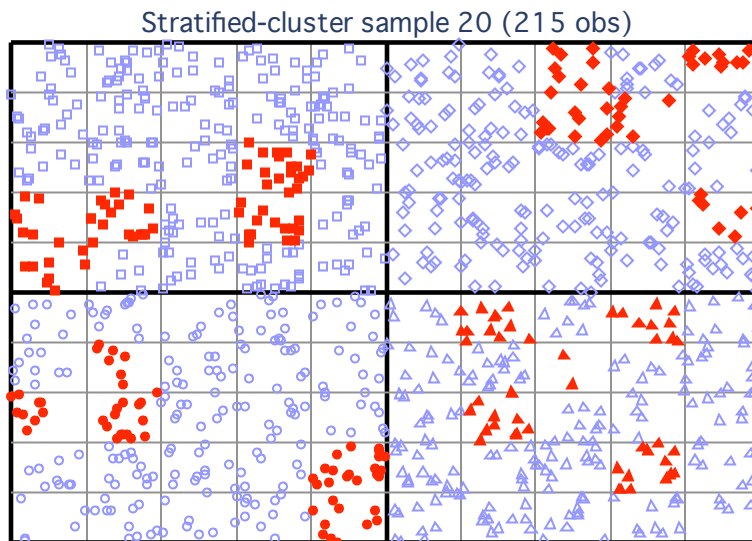


Here we partition the population into four big regions, then perform a 20% sample within each region. The sample size is not exactly 20% of the population size due to unbalanced regions and rounding.





Here we re-establish the smaller blocks within the four regions, then sample 20% of the blocks within each region.



## 2.2 Multistage designs

### Multistage syntax

```
svyset psu [weight] [, strata(varname) fpc(varname)]
  [|| ssu [, strata(varname) fpc(varname)]]
  [|| ssu [, strata(varname) fpc(varname)]] ...
```

- Stages are delimited by “||”
- SSU – secondary/subsequent sampling units
- FPC is required at stage  $s$  for stage  $s + 1$  to play a role in the linearized variance estimator

□ Note

**svyset** will note that it is disregarding subsequent stages when an FPC is not specified for a given stage. □

## 2.3 Poststratification

### Poststratification

A method for adjusting sampling weights, usually to account for underrepresented groups in the population.

- Adjusts weights to sum to the poststratum sizes in the population
- Reduces bias due to nonresponse and underrepresented groups
- Can result in smaller variance estimates

### Syntax

```
svyset ... poststrata(varname) postweight(varname)
```

#### □ Note

Recall that I said it is usually not very practical to stratify on demographic information such as age group, gender, and ethnicity. However we can usually poststratify on these variables using the frequency distribution information available from census data. □

#### ▷ Example: **svyset** for poststratification

A veterinarian has 1300 clients, 450 cats and 850 dogs. He would like to estimate the average annual expenses of his clientele but only has enough time to gather information on 50 randomly selected clients. Thus we have an *SRS* design, the sampling weight is  $26 = 1300/50$ .

Notice that the dog clients are (on average) twice as expensive as cat clients. We can use the above frequency distribution of dogs and cats to poststratify on animal type.

```
*** Cat and dog data from Levy and Lemeshow (1999)
. webuse poststrata
. bysort type: sum totexp
```

```
-> type = dog
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	32	49.85844	8.376695	32.78	66.2

```
-> type = cat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	18	21.71111	8.660666	7.14	39.88

Here are the mean estimates with postratification:

```
. svyset [pw=weight], poststrata(type) postweight(postwgt) fpc(fpc)
    pweight: weight
      VCE: linearized
    Poststrata: type
    Postweight: postwgt
    Single unit: missing
      Strata 1: <one>
        SU 1: <observations>
        FPC 1: fpc
. svy: mean totexp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      1          Number of obs   =      50
Number of PSUs  =     50          Population size =    1300
N. of poststrata =      2          Design df      =      49
```

	Mean	Linearized Std. Err.	[95% Conf. Interval]	
totexp	40.11513	1.163498	37.77699	42.45327

Here are the mean estimates without postratification:

```
. svyset _n [pw=weight]
    pweight: weight
      VCE: linearized
    Single unit: missing
      Strata 1: <one>
        SU 1: <observations>
        FPC 1: <zero>
. svy: mean totexp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      1          Number of obs   =      50
Number of PSUs  =     50          Population size =    1300
                          Design df      =      49
```

	Mean	Linearized Std. Err.	[95% Conf. Interval]	
totexp	39.7254	2.265746	35.17221	44.27859

◀

## 2.4 Strata with a single sampling unit

How do we get stuck with strata that have only one sampling unit?

- Missing data can cause entire sampling units to be dropped from the analysis, possibly leaving a single sampling unit in the estimation sample.
- Certainty units
- Bad design

### Big problem for variance estimation

- Consider a sample with only 1 observation
- **svy** reports missing standard error estimates by default

### Finding these lonely sampling units

Use **svydes**:

- Describes the strata and sampling units
- Helps find strata with a single sampling unit

▷ Example: **svydes**

The NHANES2 data has 31 strata, each containing 2 PSUs.

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2
. svydes
Survey: Describing stage 1 sampling units
  pweight: finalwgt
    VCE: linearized
Single unit: missing
Strata 1: strata
  SU 1: psu
  FPC 1: <zero>
```

Stratum	#Units	#Obs	#Obs per Unit		
			min	mean	max
1	2	380	165	190.0	215
2	2	185	67	92.5	118
3	2	348	149	174.0	199
4	2	460	229	230.0	231
5	2	252	105	126.0	147
6	2	298	131	149.0	167
7	2	476	206	238.0	270
8	2	338	158	169.0	180
9	2	244	100	122.0	144
10	2	262	119	131.0	143
11	2	275	120	137.5	155
12	2	314	144	157.0	170
13	2	342	154	171.0	188
14	2	405	200	202.5	205
15	2	380	189	190.0	191
16	2	336	159	168.0	177
17	2	393	180	196.5	213
18	2	359	144	179.5	215
20	2	285	125	142.5	160
21	2	214	102	107.0	112
22	2	301	128	150.5	173
23	2	341	159	170.5	182
24	2	438	205	219.0	233
25	2	256	116	128.0	140
26	2	261	129	130.5	132
27	2	283	139	141.5	144
28	2	299	136	149.5	163
29	2	503	215	251.5	288
30	2	365	166	182.5	199
31	2	308	143	154.0	165
32	2	450	211	225.0	239
31	62	10351	67	167.0	288

Some variables in this dataset have enough missing values to cause us the lonely PSU problem.

```
*** Mean high density lipids (mg/dL)
. svy: mean hresult
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      31      Number of obs   =      8720
Number of PSUs  =      60      Population size = 98725345
                                   Design df      =       29
```

	Linearized		
	Mean	Std. Err.	[95% Conf. Interval]
hresult	49.67141	.	.

Note: missing standard error because of stratum with single sampling unit.

Use **if e(sample)** after estimation commands to restrict **svydes**'s focus on the estimation sample. The **single** option will further restrict output to strata with one sampling unit.

```
*** Restrict to the estimation sample
. svydes if e(sample), single
Survey: Describing strata with a single sampling unit in stage 1
  pweight: finalwgt
    VCE: linearized
Single unit: missing
  Strata 1: strata
    SU 1: psu
    FPC 1: <zero>
```

Stratum	#Units	#Obs	#Obs per Unit		
			min	mean	max
1	1*	114	114	114.0	114
2	1*	98	98	98.0	98

Specifying variable names with **svydes** will result in more information about missing values.

```
*** Specifying variables for more information
. svydes hresult, single
Survey: Describing strata with a single sampling unit in stage 1
  pweight: finalwgt
    VCE: linearized
Single unit: missing
  Strata 1: strata
    SU 1: psu
    FPC 1: <zero>
```

Stratum	#Units included	#Units omitted	#Obs with complete data	#Obs with missing data	#Obs per included Unit		
					min	mean	max
1	1*	1	114	266	114	114.0	114
2	1*	1	98	87	98	98.0	98

◀

## Handling lonely sampling units

1. Drop them from the estimation sample.
2. **svyset** one of the ad-hoc adjustments in the **singleunit()** option.
3. Somehow combine them with other strata.

## 2.5 Certainty units

- Sampling units that are guaranteed to be chosen by the design.
- Certainty units are handled by treating each one as its own stratum with an FPC of 1.

## 3 Variance estimation

Stata has three variance estimation methods for survey data:

- Linearization
- Balanced repeated replication
- The jackknife

### □ Note

- Linearization
  - Stata's **robust** for complex data
  - The default variance estimation method for **svy**.
- Replication methods
  - Motivation
    - \* Linearization can have poor performance in datasets with a small number of sampling units.
    - \* Due to privacy concerns, data providers are reluctant to release strata and sampling unit information in public-use data. Thus some datasets now come packaged with weight variables for use with replication methods.
  - Concept
    - \* Think of a replicate as a copy of the point estimates.
    - \* The idea is to resample the data, computing replicates from each resample, then using the replicates to estimate the variance.

□

## 3.1 Linearization

### Linearization

A method for deriving a variance estimator using a first order Taylor approximation of the point estimator of interest.

- Foundation: Variance of the total estimator

### Syntax

```
svyset ... [vce(linearized) ]
```

- Delta method
- Huber/White/robust/sandwich estimator

#### 3.1.1 Total estimator

##### Total estimator – Stratified two-stage design

- $y_{hijk}$  – observed value from a sampled individual
- Strata:  $h = 1, \dots, L$
- PSU:  $i = 1, \dots, n_h$
- SSU:  $j = 1, \dots, m_{hi}$
- Individual:  $k = 1, \dots, m_{hij}$

$$\begin{aligned}\hat{Y} &= \sum w_{hijk} y_{hijk} \\ \widehat{V}(\hat{Y}) &= \sum_h (1 - f_h) \frac{n_h}{n_h - 1} \sum_i (y_{hi} - \bar{y}_h)^2 + \\ &\quad \sum_h f_h \sum_i (1 - f_{hi}) \frac{m_{hi}}{m_{hi} - 1} \sum_j (y_{hij} - \bar{y}_{hi})^2\end{aligned}$$

- $f_h$  is the sampling fraction for stratum  $h$  in the first stage.
- $f_{hi}$  denotes a sampling fraction in the second stage.
- Remember that the design degrees of freedom is

$$df = N_{\text{PSU}} - N_{\text{strata}}$$



▷ Example: **svy: total**

Let's use our (imaginary) survey data on high school seniors to estimate the number of smokers in the population.

```
. webuse seniors
. svyset
    pweight: sampwgt
      VCE: linearized
Single unit: missing
Strata 1: state
  SU 1: county
  FPC 1: ncounties
Strata 2: <one>
  SU 2: school
  FPC 2: nschools
Strata 3: gender
  SU 3: <observations>
  FPC 3: nseniors
*** Estimate number of seniors who have smoked
. svy: total smoked
(running total on estimation sample)
Survey: Total estimation
Number of strata =      50      Number of obs   =      10559
Number of PSUs  =     100      Population size = 20992929
                                   Design df       =         50
```

	Total	Linearized Std. Err.	[95% Conf. Interval]	
smoked	8347260	331155.1	7682115	9012404

```
*** Use first stage without FPC
. svyset county [pw=sampwgt], strata(state)
    pweight: sampwgt
      VCE: linearized
Single unit: missing
Strata 1: state
  SU 1: county
  FPC 1: <zero>
. svy: total smoked
(running total on estimation sample)
Survey: Total estimation
Number of strata =      50      Number of obs   =      10559
Number of PSUs  =     100      Population size = 20992929
                                   Design df       =         50
```

	Total	Linearized Std. Err.	[95% Conf. Interval]	
smoked	8347260	346853.4	7650584	9043935

◀

### 3.1.2 Regression models

#### Linearized variance for regression models

- Model is fit using estimating equations.
- $\hat{G}()$  is a total estimator, use Taylor expansion to get  $\hat{V}(\hat{\beta})$ .

$$\hat{G}(\beta) = \sum_j w_j s_j \mathbf{x}_j = \mathbf{0}$$

$$\hat{V}(\hat{\beta}) = D \hat{V}\{\hat{G}(\beta)\}|_{\beta=\hat{\beta}} D'$$

#### ML models

- $\hat{G}()$  is the gradient
- $s_j$  is an equation-level score
- $D$  is the inverse negative Hessian matrix at the solution

#### Least squares regression

- $\hat{G}()$  is the normal equations
- $s_j$  is a residual
- $D$  is the inverse of the weighted outer product of the predictors—including the intercept

$$D = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

▷ Example: **svy: logit**

Here is an example of a logistic regression, modeling the incidence of high blood pressure as a function of some demographic variables.

```

*** Second National Health and Nutrition Examination Survey
. webuse nhanes2
. svyset
    pweight: finalwgt
      VCE: linearized
  Single unit: missing
  Strata 1: strata
    SU 1: psu
    FPC 1: <zero>
*** Model high blood pressure on some demographics
. describe highbp height weight age female
variable name      storage  display    value      variable label
                 type     format     label
-----
highbp            byte     %8.0g
height            float    %9.0g      height (cm)
weight            float    %9.0g      weight (kg)
age               byte     %9.0g      age in years
female            byte     %8.0g      1=female, 0=male

. svy: logit highbp height weight age female
(running logit on estimation sample)
Survey: Logistic regression
Number of strata   =          31      Number of obs       =       10351
Number of PSUs    =          62      Population size      =    117157513
Design df         =              31
F( 4, 28)         =       178.69
Prob > F          =         0.0000

```

highbp	Coef.	Linearized Std. Err.	t	P> t	[95% Conf. Interval]	
height	-.0316386	.0058648	-5.39	0.000	-.0435999	-.0196772
weight	.0511574	.0031191	16.40	0.000	.0447959	.057519
age	.0492406	.0023624	20.84	0.000	.0444224	.0540587
female	-.3215716	.0884387	-3.64	0.001	-.5019435	-.1411998
_cons	-2.858968	1.049395	-2.72	0.010	-4.999224	-.7187117

◀

## 3.2 Balanced repeated replication (BRR)

### Balanced repeated replication

For designs with two PSUs in each of  $L$  strata.

- Compute replicates by dropping a PSU from each stratum.
- Find a balanced subset of the  $2^L$  replicates.  $L \leq r < L + 4$
- The replicates are used to estimate the variance.

### Syntax

```
svyset ... vce(brr) [mse]
```

#### □ Note

- The idea is to resample the data, compute replicates from each resample, then use the replicates to estimate the variance.
- Balance here means that stratum specific contributions to the variance cancel out. In other words, no stratum contributes more to the variance than any other.
- We can find a balanced subset by finding a Hadamard matrix of order  $r$ .
- When the dataset contains replicate weight variables, you do not need to worry about Hadamard matrices.

#### □ Note

- These replicate weights are used to produce a copy of the point estimates (replicate). The replicates are then used to estimate the variance.
- **svy brr** can employ replicate weight variables in the dataset, if you **svyset** them. Otherwise, **svy brr** will automatically adjust the sampling weights to produce the replicates; however, a Hadamard matrix must be specified.

## BRR variance formulas

- $\hat{\theta}$  – point estimates
- $\hat{\theta}_{(i)}$  –  $i$ th replicate of the point estimates
- $\bar{\theta}_{(\cdot)}$  – average of the replicates

Default variance formula:

$$\hat{V}(\hat{\theta}) = \frac{1}{r} \sum_{i=1}^r \{\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)}\} \{\hat{\theta}_{(i)} - \bar{\theta}_{(\cdot)}\}'$$

Mean squared error (MSE) formula:

$$\hat{V}(\hat{\theta}) = \frac{1}{r} \sum_{i=1}^r \{\hat{\theta}_{(i)} - \hat{\theta}\} \{\hat{\theta}_{(i)} - \hat{\theta}\}'$$

□ Note

- The default variance formula uses deviations of the replicates from their mean.
- The MSE formula uses deviations of the replicates from the point estimates.
- **BRR \*** is clickable, taking you to a short help file informing you that you used the MSE formula for BRR variance estimation. □

▷ Example: **svy brr: logit**

Let's revisit the previous logistic model fit, but use BRR for variance estimation.

```

*** Second National Health and Nutrition Examination Survey
. webuse nhanes2brr
. svyset [pw=finalwgt], vce(brr) mse brrweight(brr_*)
    pweight: finalwgt
      VCE: brr
      MSE: on
    brrweight: brr_1 brr_2 brr_3 brr_4 brr_5 brr_6 brr_7 brr_8 brr_9 brr_10
              brr_11 brr_12 brr_13 brr_14 brr_15 brr_16 brr_17 brr_18 brr_19
              brr_20 brr_21 brr_22 brr_23 brr_24 brr_25 brr_26 brr_27 brr_28
              brr_29 brr_30 brr_31 brr_32
    Single unit: missing
      Strata 1: <one>
        SU 1: <observations>
        FPC 1: <zero>
. svy: logit highbp height weight age female
(running logit on estimation sample)
BRR replications (32)
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
Survey: Logistic regression
                                     Number of obs   =      10351
                                     Population size  =    117157513
                                     Replications     =         32
                                     Design df        =         31
                                     F( 4, 28)        =      173.94
                                     Prob > F        =         0.0000

```

highbp	Coef.	BRR * Std. Err.	t	P> t	[95% Conf. Interval]	
height	-.0316386	.0058774	-5.38	0.000	-.0436255	-.0196516
weight	.0511574	.0031267	16.36	0.000	.0447806	.0575343
age	.0492406	.0023449	21.00	0.000	.0444581	.054023
female	-.3215716	.0897343	-3.58	0.001	-.5045859	-.1385574
_cons	-2.858968	1.044318	-2.74	0.010	-4.988868	-.7290671



### 3.3 Jackknife

#### The jackknife

A replication method for variance estimation. Not restricted to a specific survey design.

- Delete-1 jackknife: drop 1 PSU
- Delete- $k$  jackknife: drop  $k$  PSUs within a stratum

#### Syntax

```
svyset ... vce(jackknife) [mse]
```

#### □ Note

- **svy jackknife** can employ replicate weight variables in the dataset, if you **svyset** them. Otherwise, **svy jackknife** will automatically adjust the sampling weights to produce the replicates using the delete-1 jackknife methodology.
- In the delete-1 jackknife, each PSU is represented by a corresponding replicate.
- The delete- $k$  jackknife is only supported if you already have the corresponding replicate weight variables for **svyset**. □

#### Jackknife variance formulas

- $\hat{\theta}_{(h,i)}$  – replicate of the point estimates from stratum  $h$ , PSU  $i$
- $\bar{\theta}_h$  – average of the replicates from stratum  $h$
- $m_h = (n_h - 1)/n_h$  – delete-1 multiplier for stratum  $h$

Default variance formula:

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^L (1 - f_h) m_h \sum_{i=1}^{n_h} \{\hat{\theta}_{(h,i)} - \bar{\theta}_h\} \{\hat{\theta}_{(h,i)} - \bar{\theta}_h\}'$$

Mean squared error (MSE) formula:

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^L (1 - f_h) m_h \sum_{i=1}^{n_h} \{\hat{\theta}_{(h,i)} - \hat{\theta}\} \{\hat{\theta}_{(h,i)} - \hat{\theta}\}'$$

□ Note

- The default variance formula uses deviations of the replicates from their mean.
- The MSE formula uses deviations of the replicates from the point estimates.
- **Jknife \*** is clickable, taking you to a short help file informing you that you used the MSE formula for jackknife variance estimation.
- Make sure to specify the correct multiplier when you **svyset** jackknife replicate weight variables.

□

▷ Example: **svy jackknife: logit**

Here we are again with our now familiar logistic model fit, using the delete-1 jackknife variance estimator.

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2
. svyset
      pweight: finalwgt
      VCE: linearized
  Single unit: missing
  Strata 1: strata
    SU 1: psu
    FPC 1: <zero>

. svy jknife, mse: logit highbp height weight age female
(running logit on estimation sample)
Jackknife replications (62)
-----|-----|-----|-----|-----|-----|----- 50
.....
.....
Survey: Logistic regression
Number of strata   =          31      Number of obs       =      10351
Number of PSUs    =          62      Population size     = 117157513
Replications      =             62
Design df         =             31
F( 4, 28)         =      178.53
Prob > F          =           0.0000
```

highbp	Coef.	Jknife * Std. Err.	t	P> t	[95% Conf. Interval]	
height	-.0316386	.0058674	-5.39	0.000	-.0436052	-.0196719
weight	.0511574	.0031203	16.40	0.000	.0447936	.0575213
age	.0492406	.0023634	20.83	0.000	.0444204	.0540607
female	-.3215716	.088471	-3.63	0.001	-.5020093	-.1411339
_cons	-2.858968	1.049924	-2.72	0.011	-5.000302	-.7176329

◀



### Replicate weight variable

A variable in the dataset that contains sampling weight values that were adjusted for resampling the data using BRR or the jackknife.

- Typically used to protect the privacy of the survey participants.
- Eliminate the need to **svyset** the strata and PSU variables.

### Syntax

```
svyset ... brrweight (varlist)
svyset ... jkrweight (varlist [, ... multiplier(#)])
```

## 4 Estimation for subpopulations

### Focus on a subset of the population

- Subpopulation variance estimation:
  - Assumes the same survey design for subsequent data collection.
  - The **subpop ()** option.
- Restricted-sample variance estimation:
  - Assumes the identified subset for subsequent data collection.
  - Ignores the fact that the sample size is a random quantity.
  - The **if** and **in** restrictions.

### □ Note

- As I mentioned earlier on, variability is governed by the survey design, so our variance estimates assume the design is fixed. The **subpop ()** option assumes this too.
- If we discourage you from using **if** and **in**, why does **svy** allow them?
  - You might want to restrict your sample because of known defects in some of the variables.
  - Researchers can use **if** and **in** to conduct simulation studies by simulating survey samples from a population dataset without having to use **preserve** and **restore**.

- We can illustrate the difference between these estimators with an *SRS* design.

□

### Total from *SRS* data

- Data is  $y_1, \dots, y_n$  and  $S$  is the subset of observations.

$$\delta_j(S) = \begin{cases} 1, & \text{if } j \in S \\ 0, & \text{otherwise} \end{cases}$$

- Subpopulation (or restricted-sample) total:

$$\hat{Y}_S = \sum_{j=1}^n \delta_j(S) w_j y_j$$

- Sampling weight and subpopulation size:

$$w_j = \frac{N}{n}, \quad N_S = \sum_{j=1}^n \delta_j(S) w_j = \frac{N}{n} n_S$$

### Variance of a subpopulation total

Sample  $n$  without replacement from a population comprised of the  $N_S$  subpopulation values with  $N - N_S$  additional zeroes.

$$\hat{V}(\hat{Y}_S) = \left(1 - \frac{n}{N}\right) \frac{n}{n-1} \sum_{j=1}^n \left\{ \delta_j(S) y_j - \frac{1}{n} \hat{Y}_S \right\}^2$$

### Variance of a restricted-sample total

Sample  $n_S$  without replacement from the subpopulation of  $N_S$  values.

$$\tilde{V}(\hat{Y}_S) = \left(1 - \frac{n_S}{\hat{N}_S}\right) \frac{n_S}{n_S-1} \sum_{j=1}^n \delta_j(S) \left\{ y_j - \frac{1}{n_S} \hat{Y}_S \right\}^2$$

▷ Example: `svy, subpop()`

Suppose we want to estimate the mean birth weight for mothers with high blood pressure. The `highbp` variable (in the `nmihs` data) is an indicator for mothers with high blood pressure.

In the reported results, the subpopulation information is provided in the header. Notice that although the restricted sample results reproduce the same mean, the standard errors differ.

```

*** National Maternal and Infant Health Survey
. webuse nmihs
. svyset [pw=finwgt], strata(stratan)
    pweight: finwgt
      VCE: linearized
Single unit: missing
Strata 1: stratan
  SU 1: <observations>
  FPC 1: <zero>
*** Focus: birthweight, mothers with high blood pressure
. describe birthwgt highbp

```

variable name	storage type	display format	value label	variable label
birthwgt	int	%8.0g		Birthweight in grams
highbp	byte	%8.0g	hibp	High blood pressure: 1=yes,0=no

```

. label list hibp
hibp:
    0 norm BP
    1 hi BP
*** Subpopulation estimation
. svy, subpop(highbp): mean birthwgt
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      6          Number of obs   =    9953
Number of PSUs  =   9953          Population size = 3898922
                                          Subpop. no. obs =     595
                                          Subpop. size   = 186196.7
                                          Design df     =    9947

```

	Mean	Linearized Std. Err.	[95% Conf. Interval]	
birthwgt	3202.483	33.29493	3137.218	3267.748

```

*** Restricted sample estimation
. svy: mean birthwgt if highbp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      6          Number of obs   =     595
Number of PSUs  =     595          Population size = 186197
                                          Design df     =     589

```

	Mean	Linearized Std. Err.	[95% Conf. Interval]	
birthwgt	3202.483	28.7201	3146.077	3258.89

## 5 Summary

1. Use **svyset** to specify the survey design for your data.
2. Use **svydes** to find strata with a single PSU.
3. Choose your variance estimation method; you can **svyset** it.
4. Use the **svy** prefix with estimation commands.
5. Use **subpop()** instead of **if** and **in**.

## References

- [1] Levy, P. and S. Lemeshow. 1999. *Sampling of Populations*. 3rd ed. New York: Wiley.
- [2] StataCorp. 2009. *Survey Data Reference Manual: Release 11*. College Station, TX: StataCorp LP.