Survey Data Analysis in Stata

Jeff Pitblado
Associate Director, Statistical Software
StataCorp LP

2009 Canadian Stata Users Group Meeting

Outline

1 Types of data 2

2 Survey data characteristics 4
   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

3 Variance estimation 15
   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

4 Estimation for subpopulations 25

5 Summary 28
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 $20 = \frac{100}{5}$. 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

```plaintext
*** 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.

Cluster sample 20 (208 obs)

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.

Stratified sample 198
Here we re-establish the smaller blocks within the four regions, then sample 20% of the blocks within each region.

Stratified-cluster sample 20 (215 obs)

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 $\frac{1300}{50}$.

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.

```plaintext
** 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

<table>
<thead>
<tr>
<th></th>
<th>Linearized</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Err.</td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>totexp</td>
<td>40.11513</td>
<td>1.163498</td>
<td>37.77699</td>
</tr>
</tbody>
</table>
```

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

<table>
<thead>
<tr>
<th></th>
<th>Linearized</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Err.</td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>totexp</td>
<td>39.7254</td>
<td>2.265746</td>
<td>35.17221</td>
</tr>
</tbody>
</table>
```
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.

```plaintext
*** 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>
#Obs per Unit

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

### Mean high density lipids (mg/dL)
```
. svy: mean hdresult
(running mean on estimation sample)
```

Survey: Mean estimation
---
Number of strata = 31
Number of PSUs = 60
Population size = 98725345
Design df = 29

<table>
<thead>
<tr>
<th>Linearized</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>hdresult</td>
<td>49.67141</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

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>

<table>
<thead>
<tr>
<th>#Obs per Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

### Specifying variable names with `svydes`
```
. svydes hdresult, 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>

<table>
<thead>
<tr>
<th>#Obs per included Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratum</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>
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

\texttt{svyset ... [vce(linearized)]}

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

3.1.1 Total estimator

Total estimator – Stratified two-stage design

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

\[
\hat{Y} = \sum w_{hihk} y_{hihk}
\]

\[
\hat{V}(\hat{Y}) = \sum_h (1 - f_h)\frac{n_h}{n_h - 1} \sum_i (y_{hi} - \bar{y}_h)^2 + \sum_h f_h \sum_i (1 - f_{hi})\frac{m_{hi}}{m_{hi} - 1} \sum_j (y_{hij} - \bar{y}_{hi})^2
\]

- \( f_h \) 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

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>smoked</td>
<td>8347260</td>
<td>331155.1</td>
<td>7682115 - 9012404</td>
</tr>
</tbody>
</table>
```

*** 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

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>smoked</td>
<td>8347260</td>
<td>346853.4</td>
<td>7650584 - 9043935</td>
</tr>
</tbody>
</table>
```
3.1.2 Regression models

Linearized variance for regression models

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

\[
\hat{G}(\beta) = \sum_j w_j s_j x_j = 0
\]

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

ML models

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

Least squares regression

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

\[
D = (X'WX)^{-1}
\]
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.

```stata
*** 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
.storage display value
variable name type format label variable label
highbp byte %8.0g 1 if BP > 140/90, 0 otherwise
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
.svys 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
```

| Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|--------|-----------|-------|------|----------------------|
| height | -.0316386 | .0058648 | -5.39 | 0.000 | -.0435999 to -.196772 |
| weight | .0511574  | .0031191  | 16.40 | 0.000 | .0447959 to .057519  |
| age    | .0492406  | .0023624  | 20.84 | 0.000 | .0444224 to .0540587 |
| female | -3.215716 | .0884387  | -3.64 | 0.001 | -5.019435 to -.1411998 |
| _cons  | -2.858968 | 1.049395 | -2.72 | 0.010 | -4.999224 to -.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
- \( \bar{\theta}() \) – average of the replicates

Default variance formula:

\[
\hat{V}(\hat{\theta}) = \frac{1}{r} \sum_{i=1}^{r} \{ \hat{\theta}(i) - \bar{\theta}() \} \{ \hat{\theta}(i) - \bar{\theta}() \}'
\]

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.

```stata
*** 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)

+----------+------------------+
|          | Coef. Std. Err.  t  P>|t| [95% Conf. Interval] |
+----------+------------------+
| highbp   |                  |
| height   | -0.0316386 0.0058774 -5.38 0.000 -.0436255 -.0196516 |
| weight   | 0.0511574 0.0031267 16.36 0.000 0.0447806 0.0575343 |
| age      | 0.0492406 0.0023449 21.00 0.000 0.0444581 0.054023 |
| female   | -0.3215716 0.0897343 -3.58 0.001 -0.5045859 -0.1385574 |
| _cons    | -2.858968 1.044318 -2.74 0.010 -4.988868 -.7290671 |
+----------+------------------+
```

Survey: Logistic regression
Number of obs = 10351
Population size = 117157513
Replications = 32
Design df = 31
F( 4, 28) = 173.94
Prob > F = 0.0000

22
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.

```stata
*** 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

|         | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|--------|-----------|-------|-----|---------------------|
| highbp  |        |           |       |     |                     |
| 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 |
```

24
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`.  

25
• We can illustrate the difference between these estimators with an SRS design.

Total from SRS data

• Data is $y_1, \ldots, y_n$ 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.

$$\hat{V}(\hat{Y}_S) = \left(1 - \frac{n_S}{N_S}\right) \frac{n_S}{n_S - 1} \sum_{j=1}^{n} \delta_j(S) \left(y_j - \frac{1}{n_S} \hat{Y}_S\right)^2$$
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
storage  display    value
variable name type format 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
  Linearized
  Mean    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
  Linearized
  Mean    Std. Err.    [95% Conf. Interval]
  birthwgt  3202.483   28.72010    3146.077    3258.890
```
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
