Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle
Simple Random Sampling

- Each observation the same probability of being selected.
- Observations can only be selected once.

Code

- SRS stands for simple random sample
- \( n \) refers to the sample size
- Seed is used to replicate the sample

```
Proc SurveySelect
  data = Example
  method = srs
  n = 15
  out = Example_SRS
  seed = 50460
; Run;
```
Sampling in SAS ® using PROC SURVEYSELECT
Rachael Becker and Drew Doyle
University of Central Florida

Unrestricted Random Sampling

- Simple random sampling with replacement
- Observations can be selected multiple times

Code

- URS stands for unrestricted random sample
- Outhits creates the column Numberhits
- Numberhits is how frequent the observation occurs

```
Proc SurveySelect
  data = Example
  method = urs
  n = 15
  out = Example_SRS_replacement
  seed = 50460
  outhits
;
Run;
```

Results

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>Unrestricted Random Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Data Set</td>
<td>EXAMPLE</td>
</tr>
<tr>
<td>Random Number Seed</td>
<td>50460</td>
</tr>
<tr>
<td>Sample Size</td>
<td>15</td>
</tr>
<tr>
<td>Expected Number of Hits</td>
<td>0.3</td>
</tr>
<tr>
<td>Sampling Weight</td>
<td>3.333333</td>
</tr>
<tr>
<td>Output Data Set</td>
<td>EXAMPLE_SRS_REPLACEMENT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>IDNo</th>
<th>Year</th>
<th>FinalGrade</th>
<th>Class</th>
<th>NumberHits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>986</td>
<td>Freshman</td>
<td>20</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>464</td>
<td>Junior</td>
<td>52</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>907</td>
<td>Senior</td>
<td>35</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>907</td>
<td>Senior</td>
<td>35</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>907</td>
<td>Senior</td>
<td>35</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>041</td>
<td>Senior</td>
<td>84</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>462</td>
<td>Senior</td>
<td>36</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>724</td>
<td>Senior</td>
<td>70</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>970</td>
<td>Senior</td>
<td>73</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>818</td>
<td>Sophomore</td>
<td>74</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>190</td>
<td>Sophomore</td>
<td>75</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>641</td>
<td>Sophomore</td>
<td>67</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>069</td>
<td>Sophomore</td>
<td>68</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Stratified Random Sampling

- Sampling within subgroups or stratum
- Random sampling without replacement of the subgroups

Code

- Similar code to simple random
- New option STRATA that specifies how to the data should be separated

Sampling in SAS® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

University of Central Florida

Proc SurveySelect
data = Example
method = srs
n = 3
out = Example_Stratification_good
seed = 62493;
strata Year;
Run;

Results

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>Simple Random Sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strata Variable</td>
<td>Year</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input Data Set</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Number Seed</td>
<td>62493</td>
</tr>
<tr>
<td>Stratum Sample Size</td>
<td>3</td>
</tr>
<tr>
<td>Number of Strata</td>
<td>4</td>
</tr>
<tr>
<td>Total Sample Size</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>Year</th>
<th>IDNo</th>
<th>FinalGrade</th>
<th>Class</th>
<th>SelectionProb</th>
<th>SamplingWeight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Freshman</td>
<td>646</td>
<td>50</td>
<td>1</td>
<td>0.60000</td>
<td>1.66667</td>
</tr>
<tr>
<td>2</td>
<td>Freshman</td>
<td>516</td>
<td>56</td>
<td>2</td>
<td>0.60000</td>
<td>1.66667</td>
</tr>
<tr>
<td>3</td>
<td>Freshman</td>
<td>094</td>
<td>61</td>
<td>3</td>
<td>0.60000</td>
<td>1.66667</td>
</tr>
<tr>
<td>4</td>
<td>Junior</td>
<td>949</td>
<td>54</td>
<td>2</td>
<td>0.27273</td>
<td>3.66667</td>
</tr>
<tr>
<td>5</td>
<td>Junior</td>
<td>815</td>
<td>87</td>
<td>3</td>
<td>0.27273</td>
<td>3.66667</td>
</tr>
<tr>
<td>6</td>
<td>Junior</td>
<td>849</td>
<td>88</td>
<td>3</td>
<td>0.27273</td>
<td>3.6667</td>
</tr>
<tr>
<td>7</td>
<td>Senior</td>
<td>868</td>
<td>60</td>
<td>1</td>
<td>0.12000</td>
<td>8.33333</td>
</tr>
<tr>
<td>8</td>
<td>Senior</td>
<td>674</td>
<td>45</td>
<td>2</td>
<td>0.12000</td>
<td>8.33333</td>
</tr>
<tr>
<td>9</td>
<td>Senior</td>
<td>075</td>
<td>86</td>
<td>2</td>
<td>0.12000</td>
<td>8.33333</td>
</tr>
<tr>
<td>10</td>
<td>Sophomore</td>
<td>841</td>
<td>98</td>
<td>2</td>
<td>0.33333</td>
<td>3.00000</td>
</tr>
<tr>
<td>11</td>
<td>Sophomore</td>
<td>013</td>
<td>72</td>
<td>3</td>
<td>0.33333</td>
<td>3.00000</td>
</tr>
<tr>
<td>12</td>
<td>Sophomore</td>
<td>641</td>
<td>67</td>
<td>3</td>
<td>0.33333</td>
<td>3.00000</td>
</tr>
</tbody>
</table>
Cluster Sampling

- Division of data into mutually exclusive groups
- Data is usually related in a certain manner (e.g. geography)
- Used for convenience and monetary benefits

Code

- Again the code looks similar to simple random
- SAMPLINGUNIT is how define the variable that the data was clustered by

```
proc surveyselect
  data = Example2
  method = srs
  sampsize = 5
  out = Example_Clustering
  seed = 7162010;
  samplingunit IDNo;
run;
```
Systematic Random Sampling

- Selection of every k\(^{th}\) observation
- Formula:
  \[ K = \frac{N}{n} \]
  \[ K_{th} = \frac{\text{Total # in the Population}}{\text{# of Observation in the Sample}} \]

Code

- Note the change in METHOD
- SYS stands for systematic random sampling

```
Proc SurveySelect
  data = Example3
  method = sys
  n = 15
  out = Example_Systematic
  seed = 31636
; Run;
```
Sampling in SAS® using PROC SURVEYSELECT
Rachael Becker and Drew Doyle
University of Central Florida

Sequential Random Sampling

- Takes population size of strata into account
- Sequential vs. Stratified in SAS: Sequential calculates the appropriate sizes of the stratum on its own, Stratified does not

Code

- Note the change in the METHOD
- SEQ stands for sequential random sampling
- Addition of SORT, CONTROL, and STRATA options

```
Proc SurveySelect
  data = Example3
  method = seq
  n = 1
  out = Example_Sequential
  seed = 31636
  sort = nest
  control Name;
strata NoSib;
Run;
```

Results

Selection Method: Sequential Random Sampling With Equal Probability
Strata Variable: NoSib
Control Variable: Name

Input Data Set: EXAMPLE3
Random Number Seed: 31636
Stratum Sample Size: 1
Number of Strata: 8
Total Sample Size: 8
Output Data Set: EXAMPLE_SEQUENTIAL

<table>
<thead>
<tr>
<th>Obs</th>
<th>NoSib</th>
<th>Name</th>
<th>SelectionProb</th>
<th>SamplingWeight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>Kristen</td>
<td>0.16667</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Marilyn</td>
<td>0.12500</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Grant</td>
<td>0.07692</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Tracy</td>
<td>0.10000</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>Richard</td>
<td>0.20000</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>Dennis</td>
<td>0.25000</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>Carlton</td>
<td>0.33333</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>John</td>
<td>1.00000</td>
<td>1</td>
</tr>
</tbody>
</table>
Conclusion

• PROC SURVEYSELECT helps apply useful sampling techniques
• PROC SURVEY SELECT has many more options than what were described in this presentation

References


Contact Information

Rachael Becker:
Email: Leahcarbecker@knights.ucf.edu
Drew Doyle:
Email: Drewdoyle@knights.ucf.edu

Acknowledgments

We would like to thank Dr. Mark Johnson for sparking our interest in sampling methods. We would also like to say a special thank you to Kelcey Ellis, without her this paper would not exist.