SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

#SASGF

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

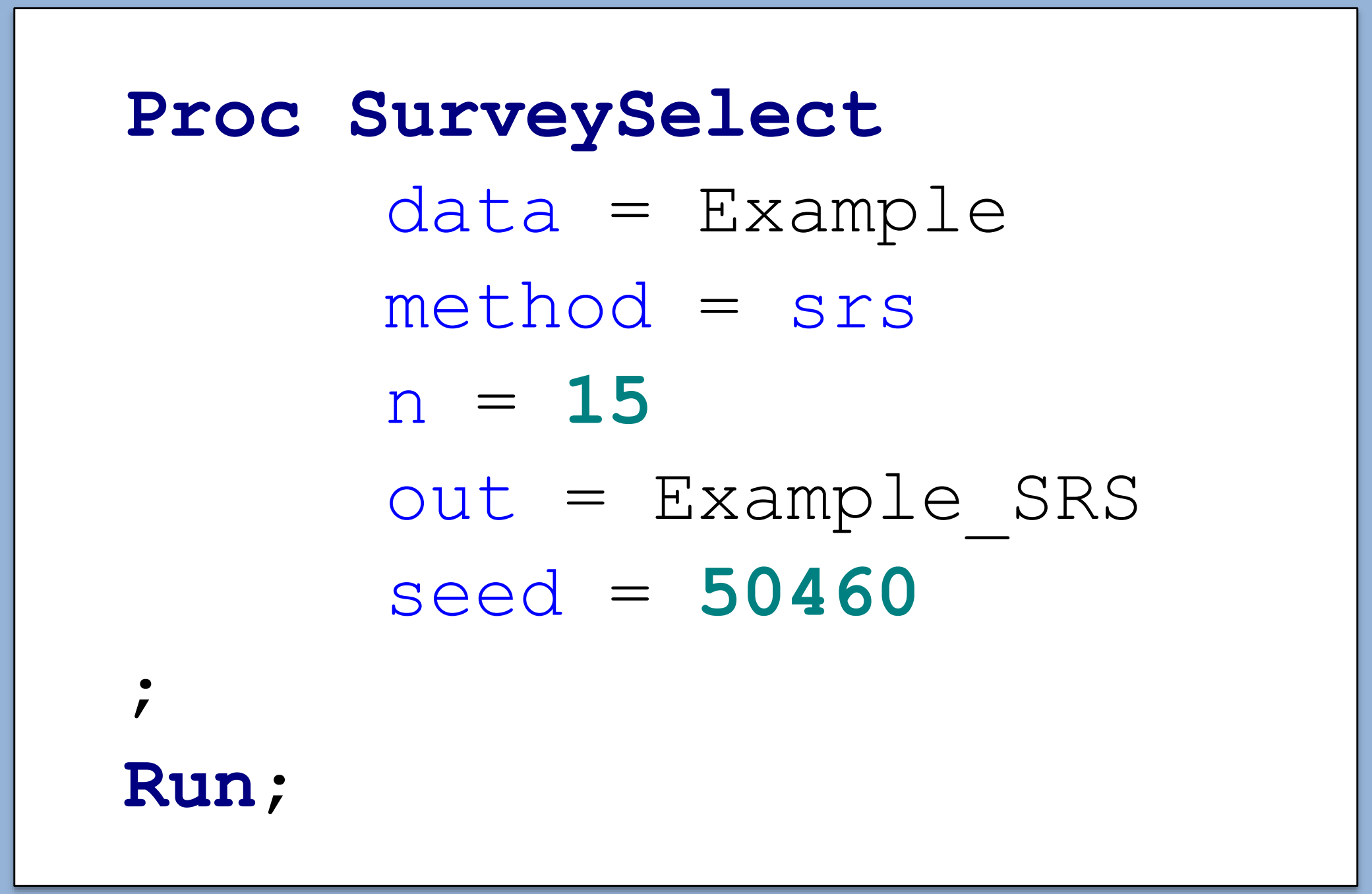
University of Central Florida

Simple Random Sampling

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

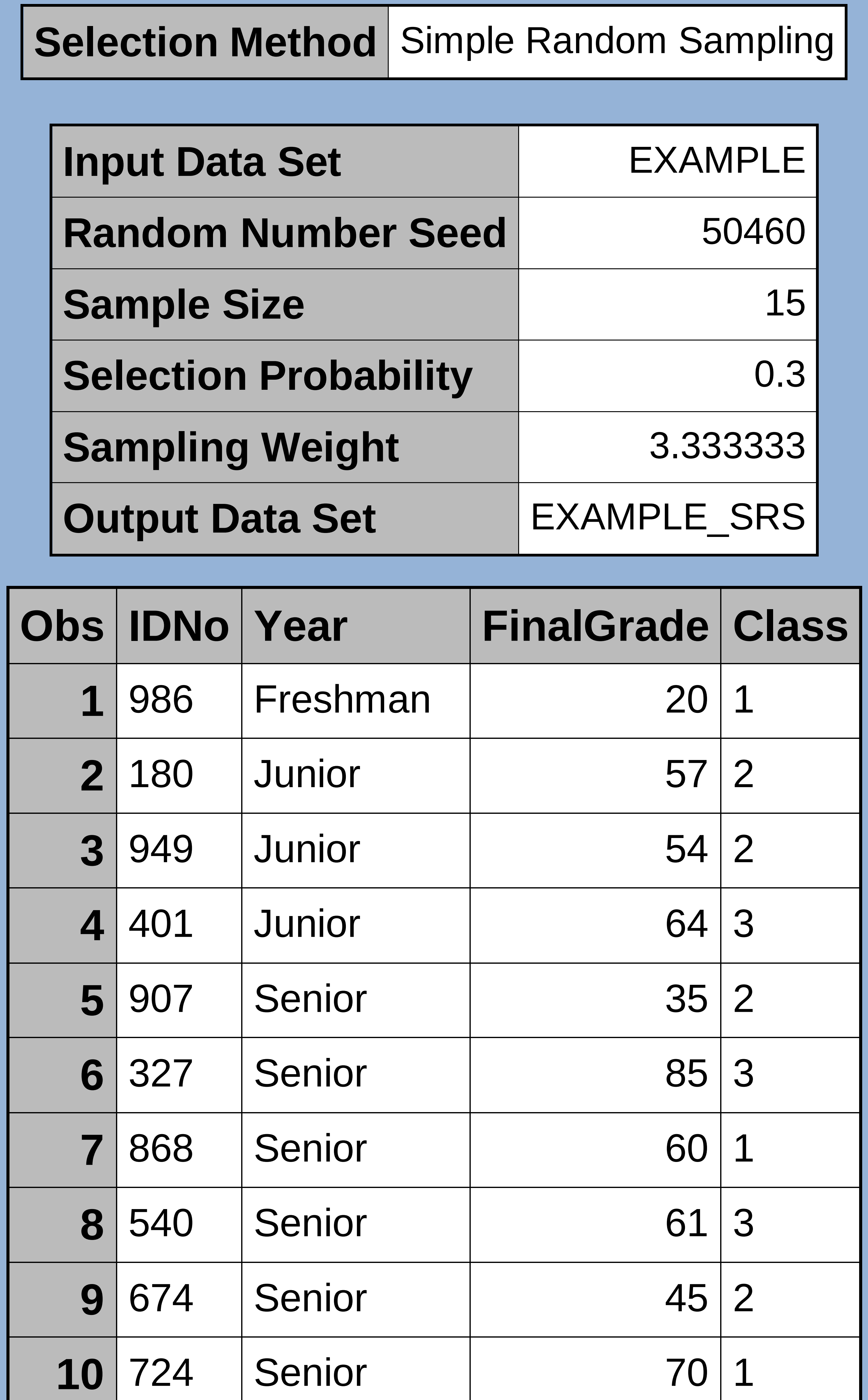
Code

- 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;
```

Results



Selection Method	Simple Random Sampling

Input Data Set	EXAMPLE
Random Number Seed	50460
Sample Size	15
Selection Probability	0.3
Sampling Weight	3.333333
Output Data Set	EXAMPLE_SRS

Obs	IDNo	Year	FinalGrade	Class
1	986	Freshman	20	1
2	180	Junior	57	2
3	949	Junior	54	2
4	401	Junior	64	3
5	907	Senior	35	2
6	327	Senior	85	3
7	868	Senior	60	1
8	540	Senior	61	3
9	674	Senior	45	2
10	724	Senior	70	1

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

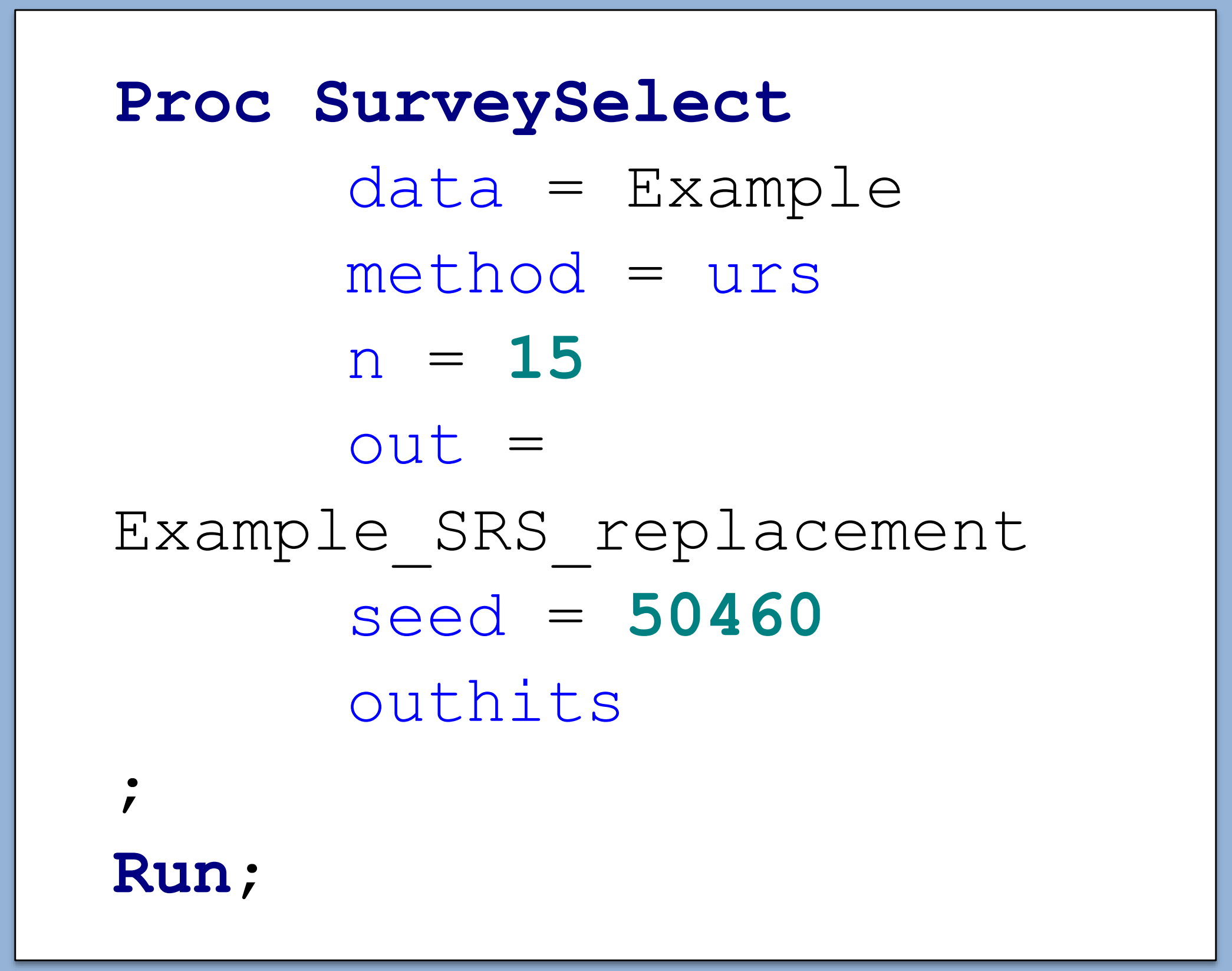
University of Central Florida

Unrestricted Random Sampling

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

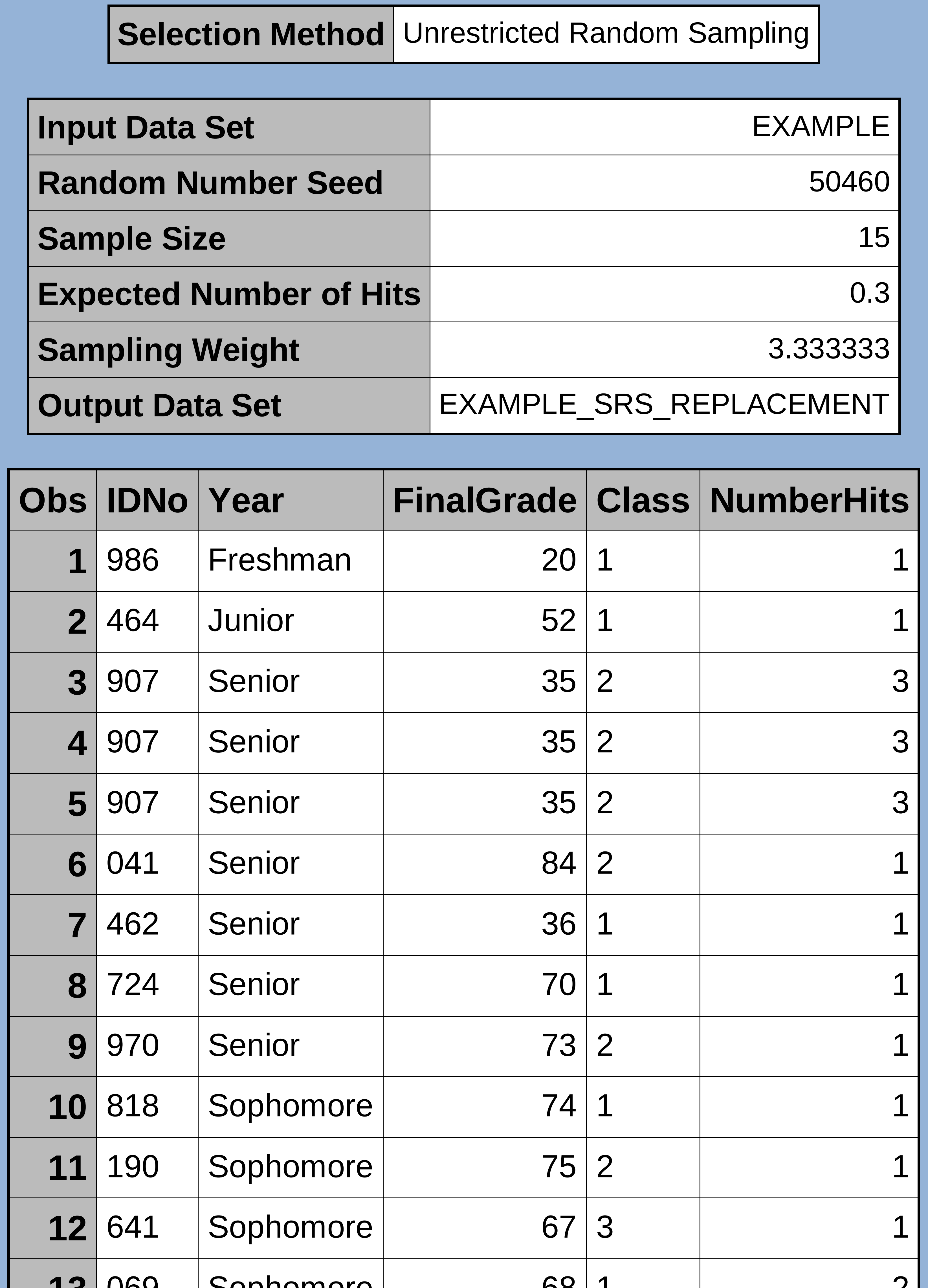
Code

- 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



Selection Method	Unrestricted Random Sampling

Input Data Set	EXAMPLE
Random Number Seed	50460
Sample Size	15
Expected Number of Hits	0.3
Sampling Weight	3.333333
Output Data Set	EXAMPLE_SRS_REPLACEMENT

Obs	IDNo	Year	FinalGrade	Class	NumberHits
1	986	Freshman	20	1	1
2	464	Junior	52	1	1
3	907	Senior	35	2	3
4	907	Senior	35	2	3
5	907	Senior	35	2	3
6	041	Senior	84	2	1
7	462	Senior	36	1	1
8	724	Senior	70	1	1
9	970	Senior	73	2	1
10	818	Sophomore	74	1	1
11	190	Sophomore	75	2	1
12	641	Sophomore	67	3	1
13	069	Sophomore	68	1	2

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

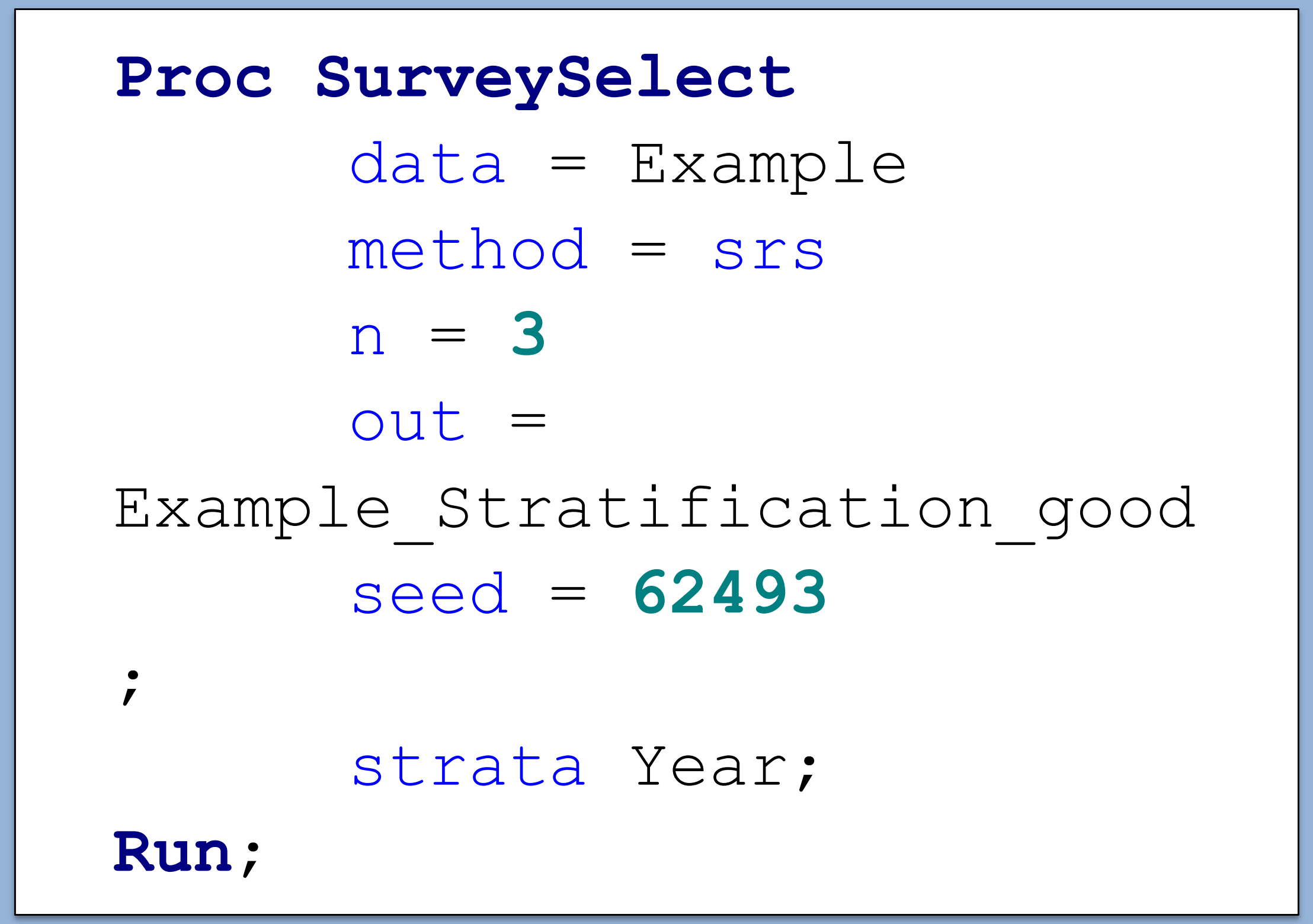
University of Central Florida

Stratified Random Sampling

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

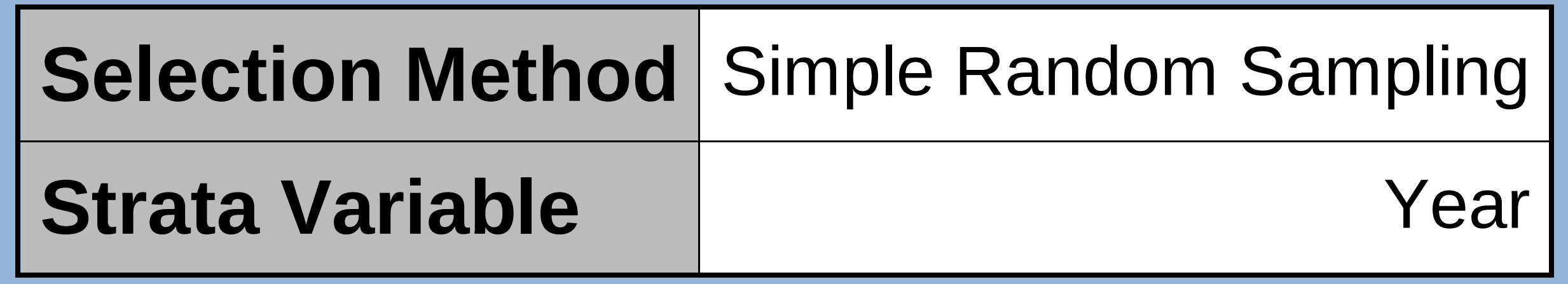
Code

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

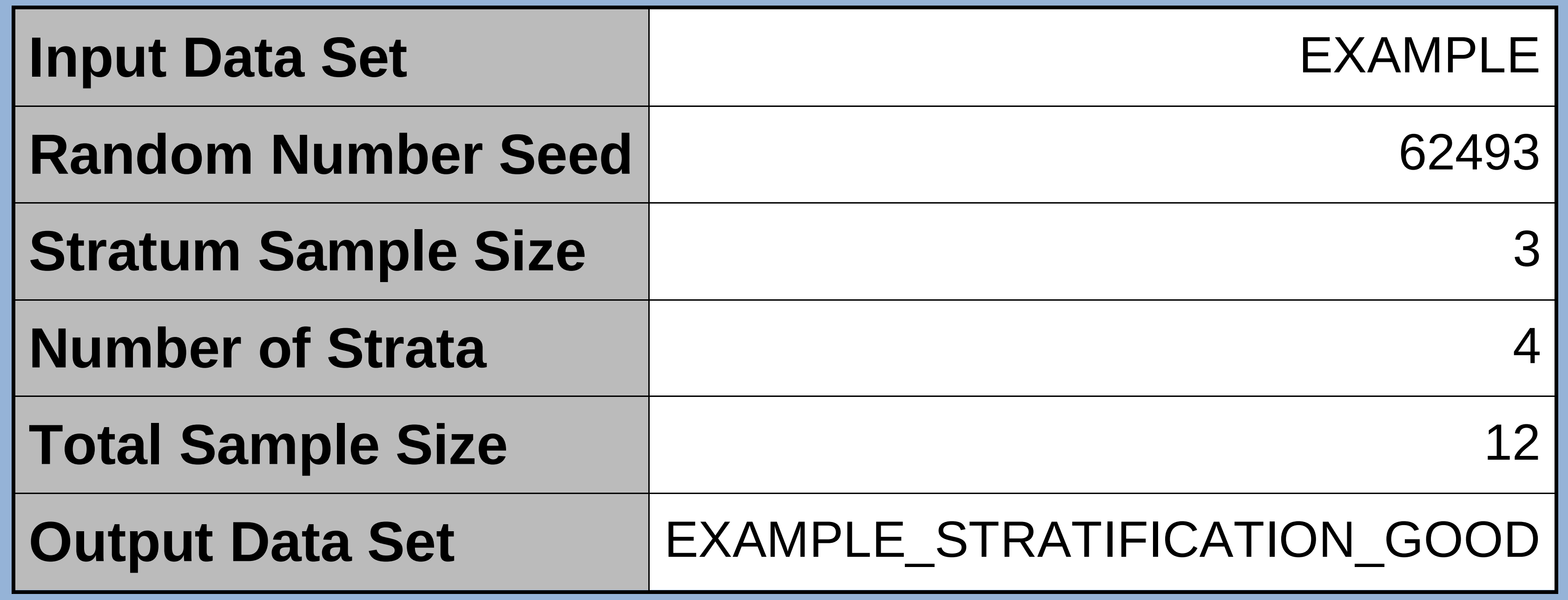


```
Proc SurveySelect
      data = Example
      method = srs
      n = 3
      out =
Example_Stratification_good
      seed = 62493
;
      strata Year;
Run;
```

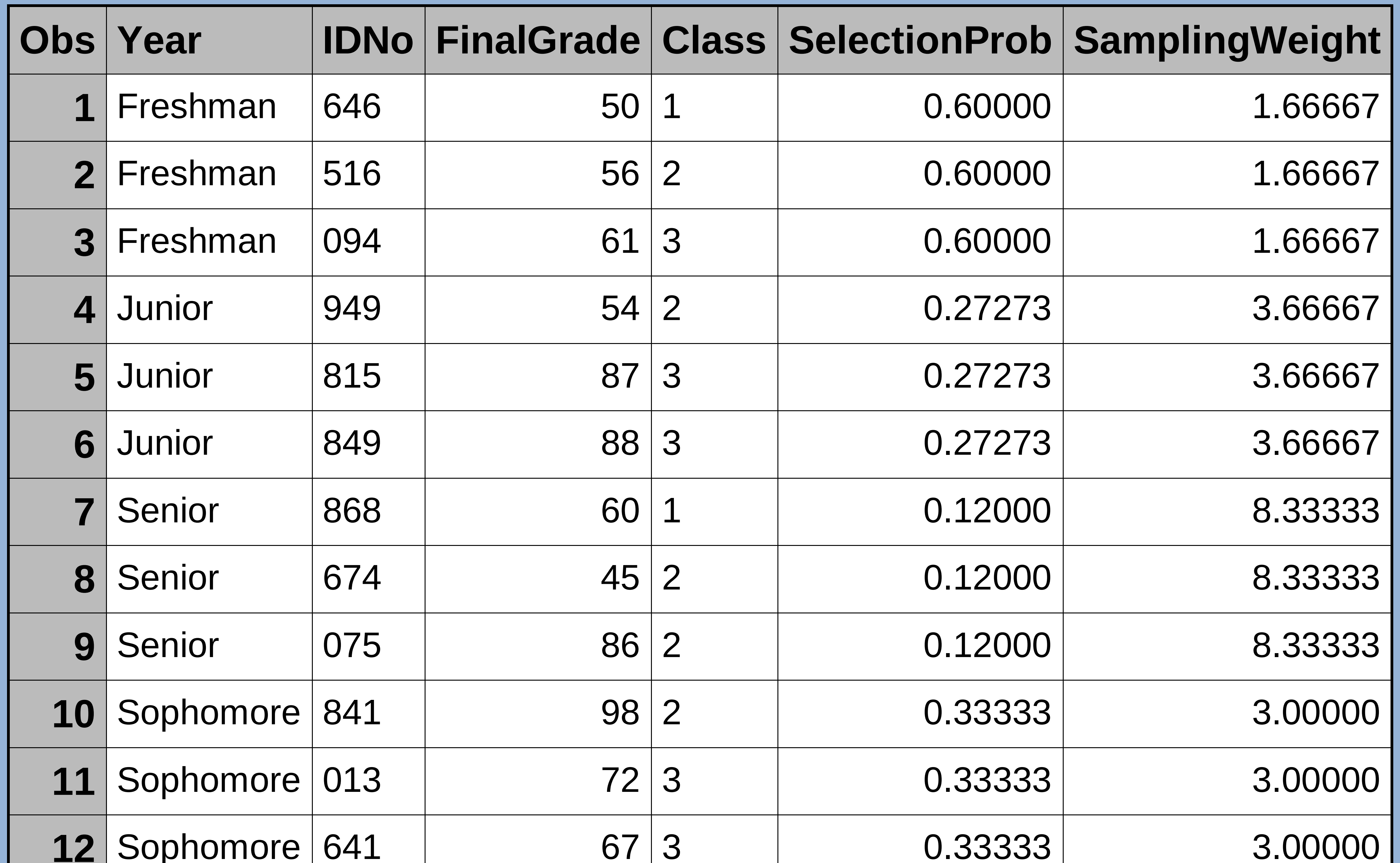
Results



Selection Method	Simple Random Sampling
Strata Variable	Year



Input Data Set	EXAMPLE
Random Number Seed	62493
Stratum Sample Size	3
Number of Strata	4
Total Sample Size	12
Output Data Set	EXAMPLE_STRATIFICATION_GOOD



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

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

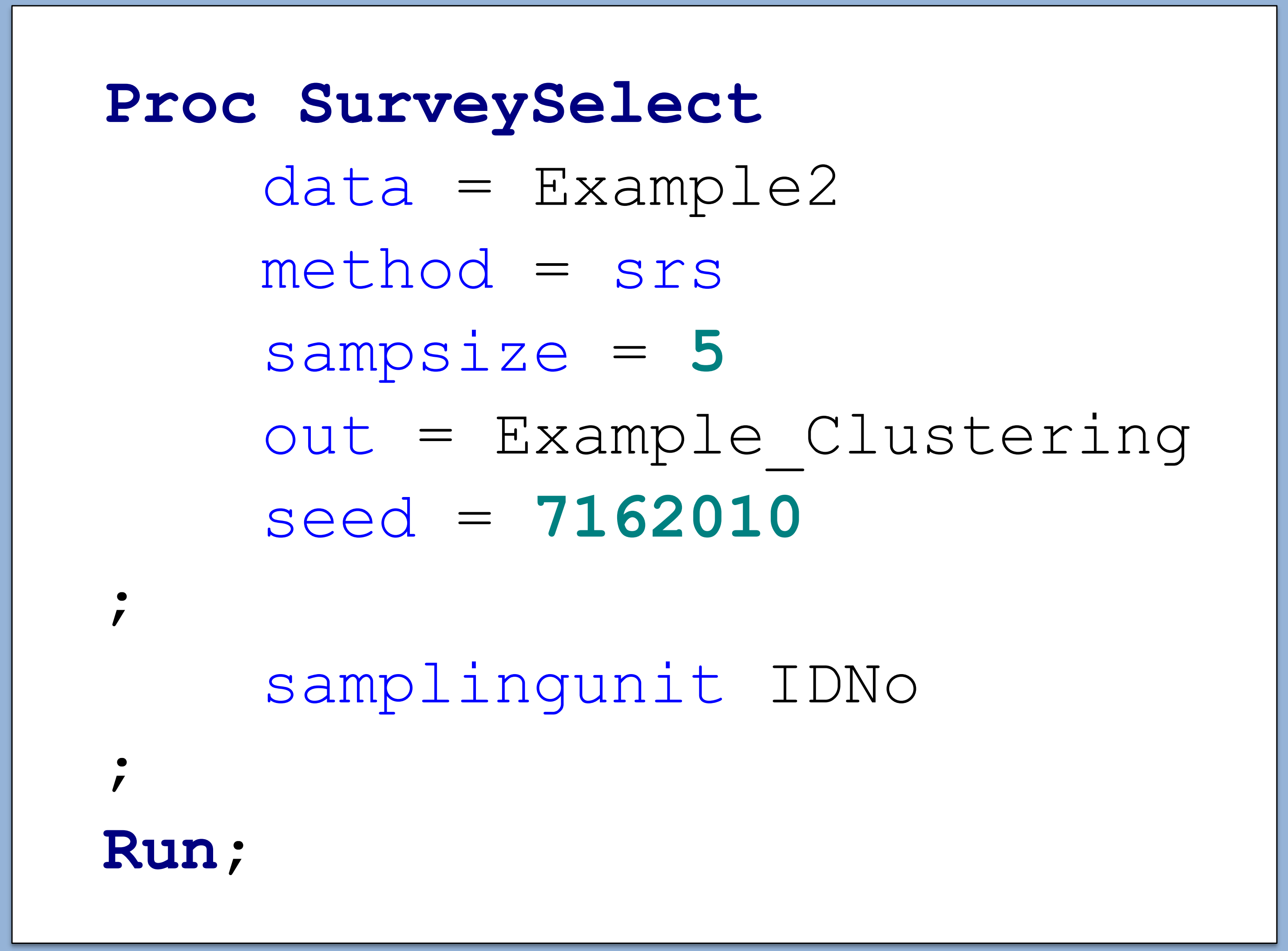
University of Central Florida

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

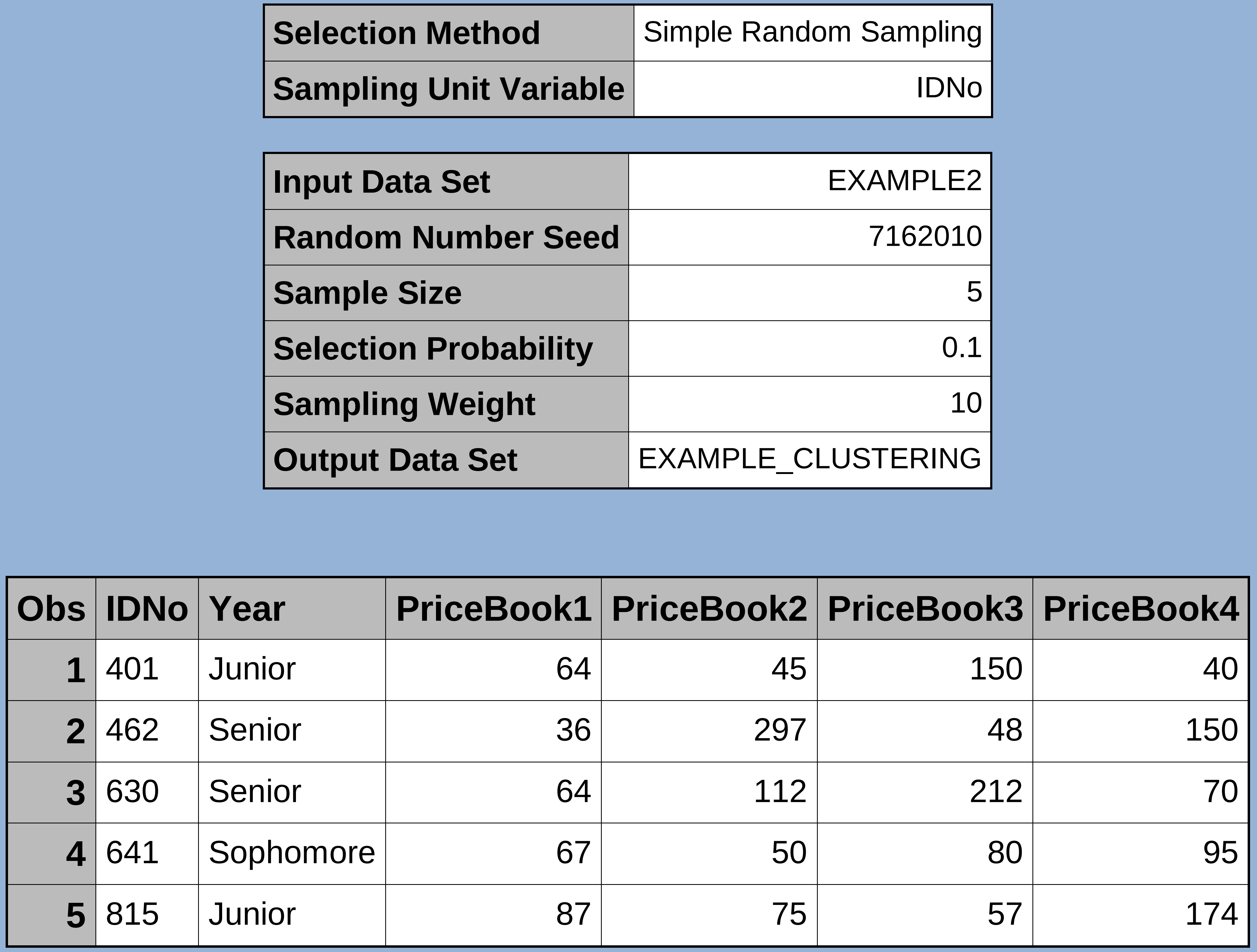
Code

- 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;
```

Results



Selection Method	Simple Random Sampling
Sampling Unit Variable	IDNo

Input Data Set	EXAMPLE2
Random Number Seed	7162010
Sample Size	5
Selection Probability	0.1
Sampling Weight	10
Output Data Set	EXAMPLE_CLUSTERING

Obs	IDNo	Year	PriceBook1	PriceBook2	PriceBook3	PriceBook4
1	401	Junior	64	45	150	40
2	462	Senior	36	297	48	150
3	630	Senior	64	112	212	70
4	641	Sophomore	67	50	80	95
5	815	Junior	87	75	57	174

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

University of Central Florida

Systematic Random Sampling

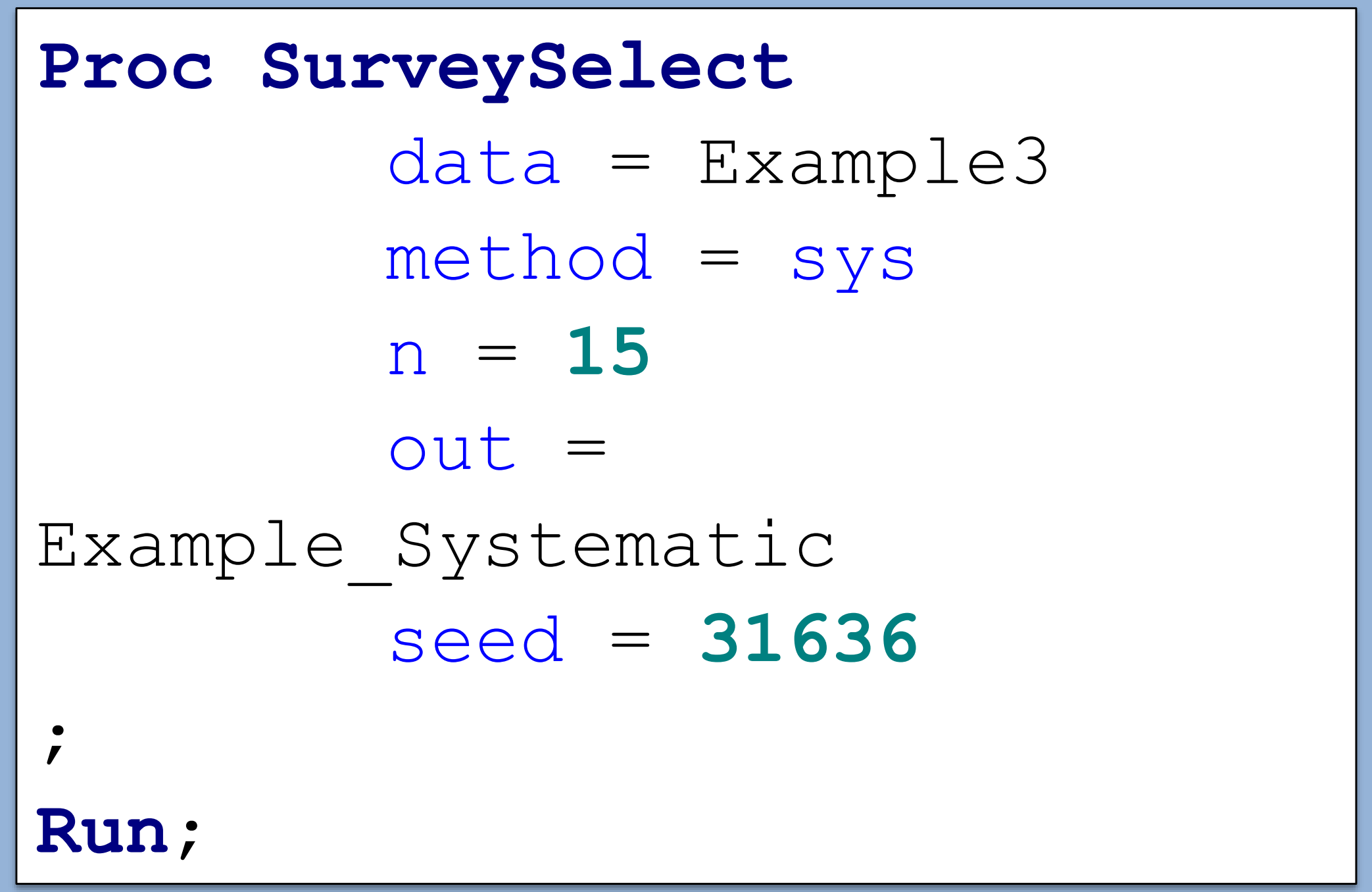
- Selection of every kth observation
- Formula:

$$K = \frac{N}{n}$$

$$Kth = \frac{Total\ \#\ in\ the\ Population}{\#\ of\ Observation\ in\ the\ Sample}$$

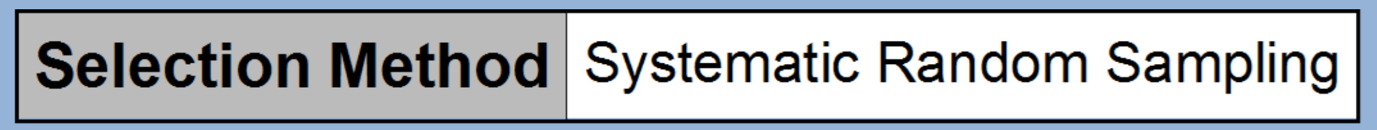
Code

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

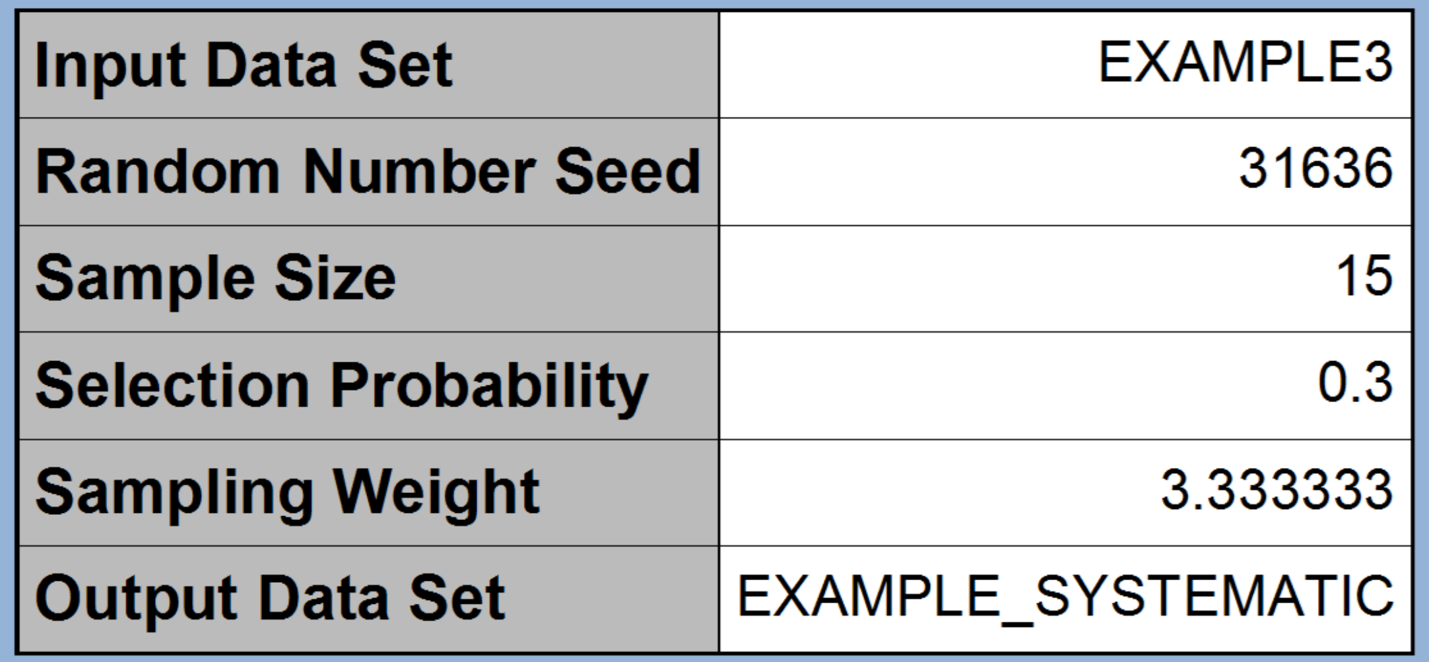


```
Proc SurveySelect
        data = Example3
        method = sys
        n = 15
        out =
Example_Systematic
        seed = 31636
;
Run;
```

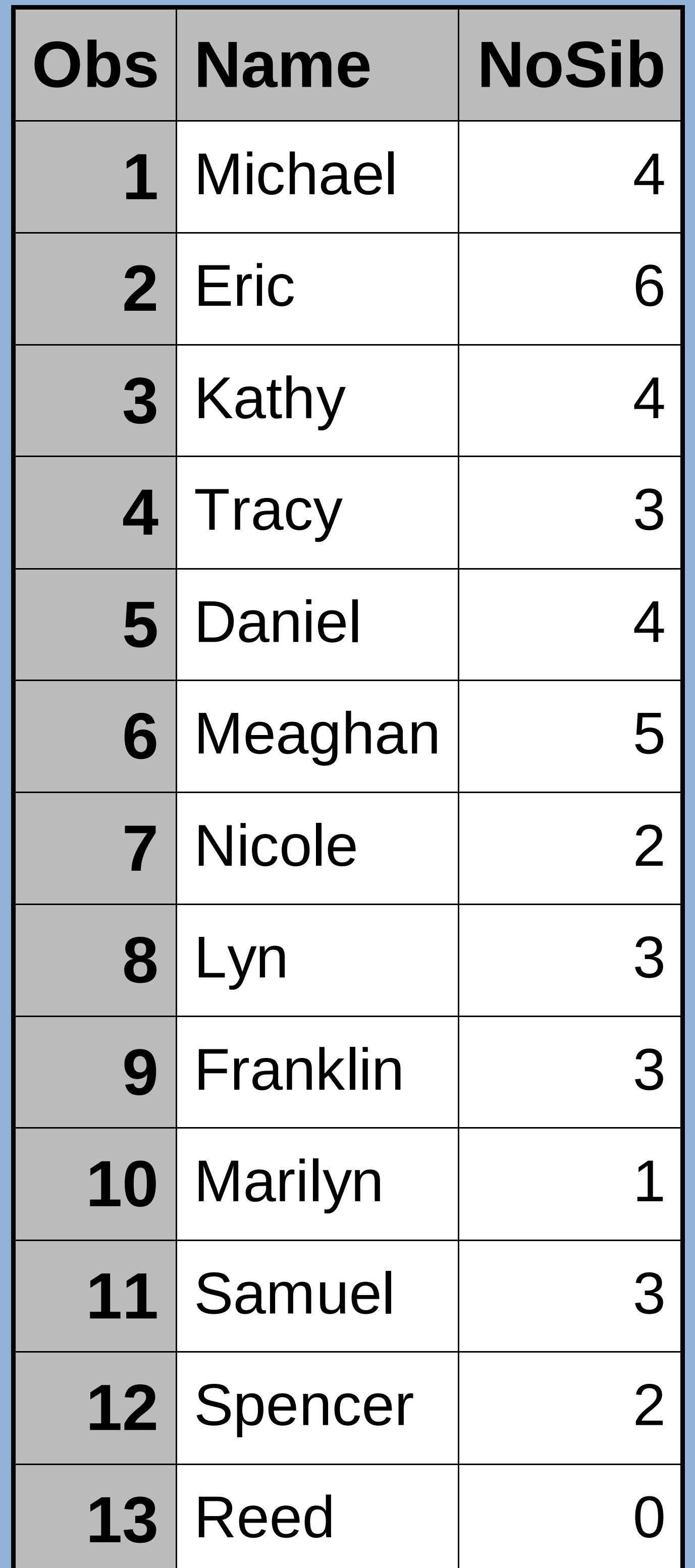
Results



Selection Method	Systematic Random Sampling



Input Data Set	EXAMPLE3
Random Number Seed	31636
Sample Size	15
Selection Probability	0.3
Sampling Weight	3.333333
Output Data Set	EXAMPLE_SYSTEMATIC



Obs	Name	NoSib
1	Michael	4
2	Eric	6
3	Kathy	4
4	Tracy	3
5	Daniel	4
6	Meaghan	5
7	Nicole	2
8	Lyn	3
9	Franklin	3
10	Marilyn	1
11	Samuel	3
12	Spencer	2
13	Reed	0

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

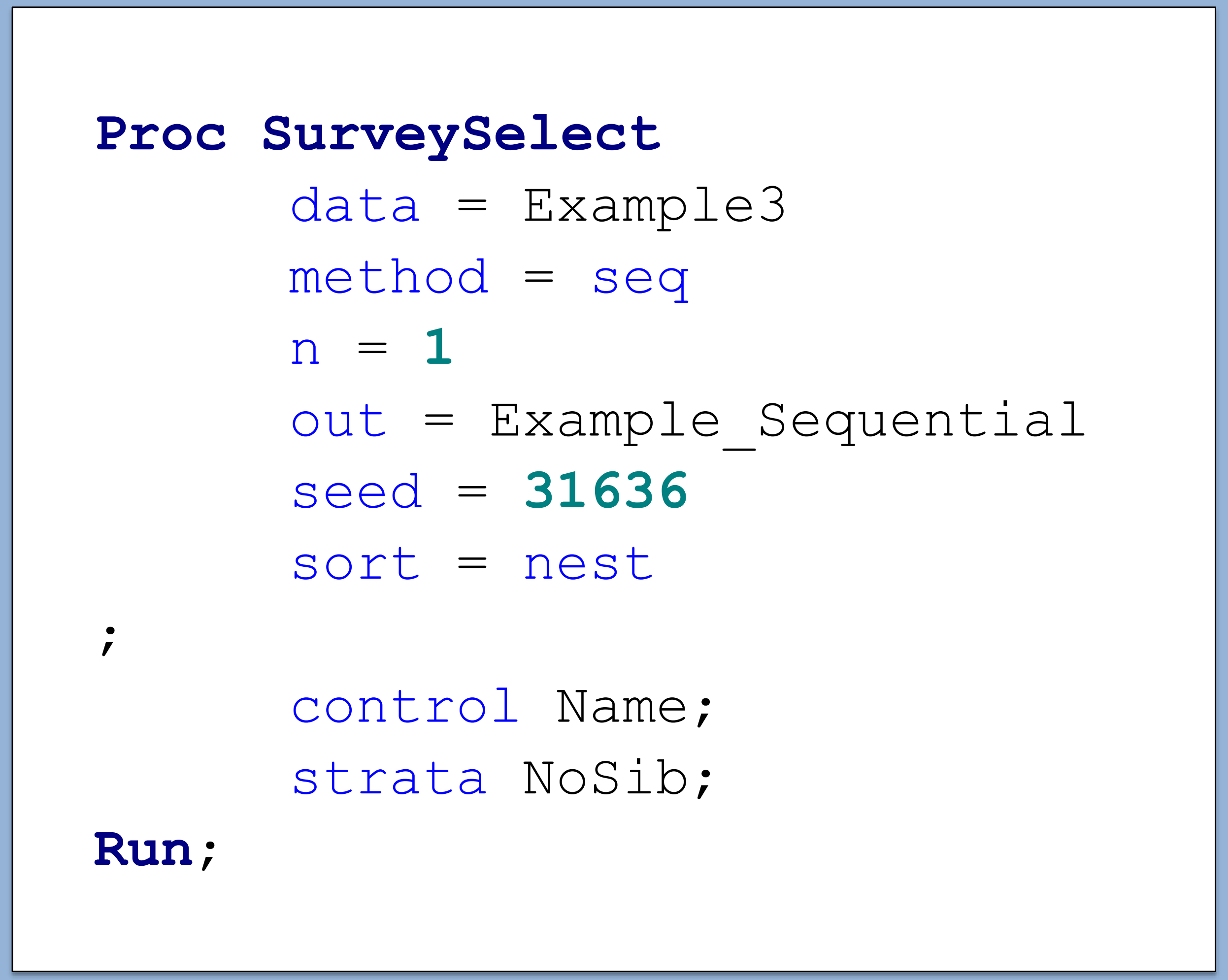
University of Central Florida

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

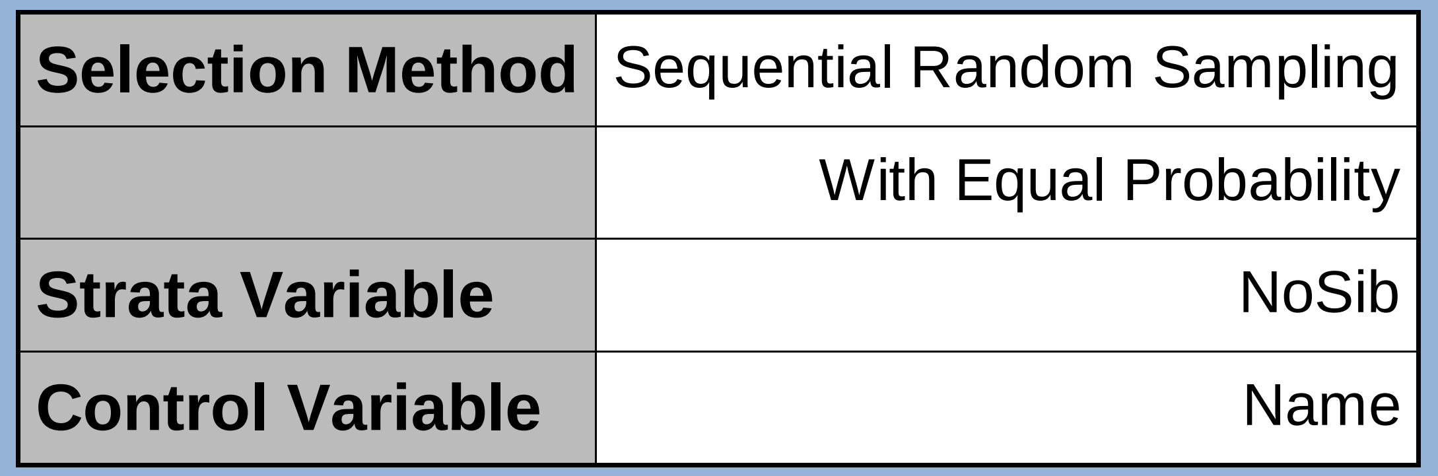
Code

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

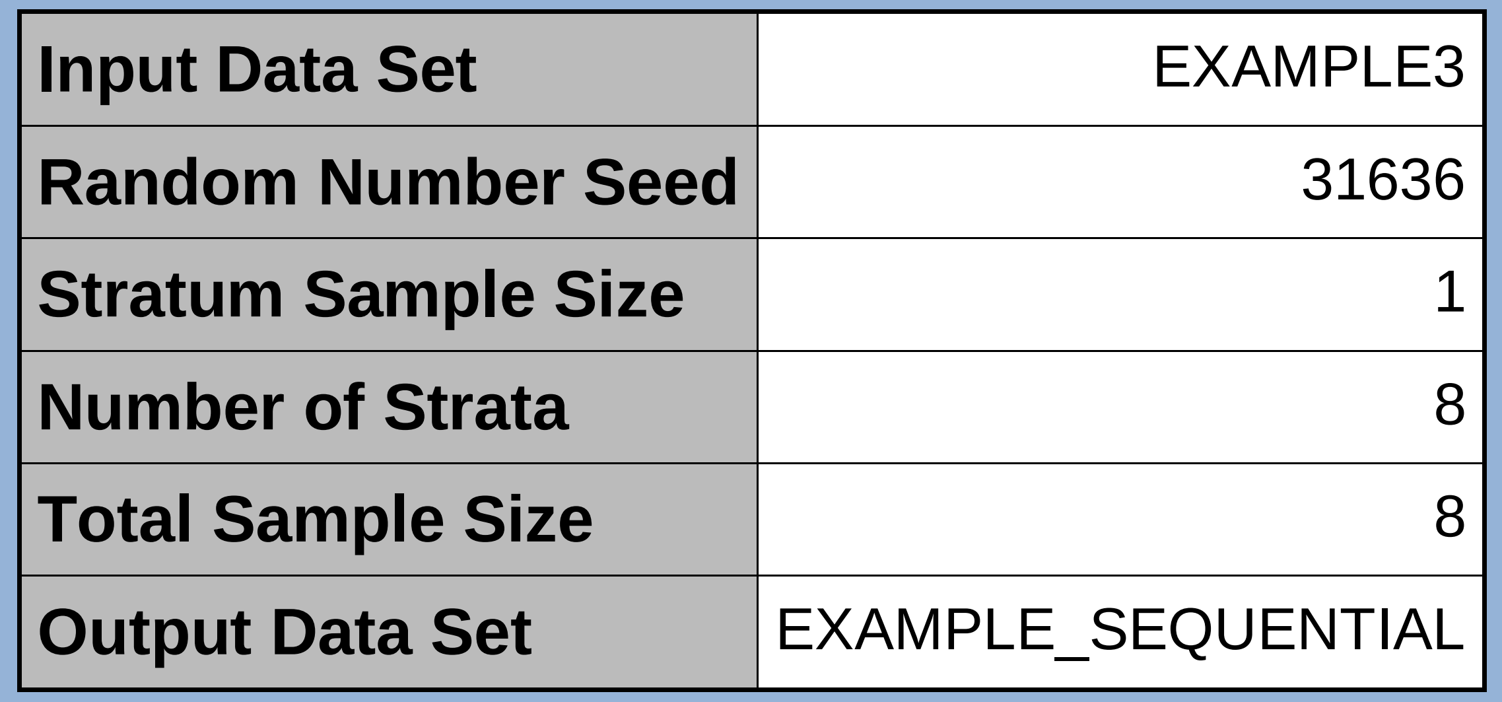


```
Proc SurveySelect
      data = Example3
      method = seq
      n = 1
      out = Example_Sequential
      seed = 31636
      sort = nest
;
      control Name;
      strata NoSib;
Run;
```

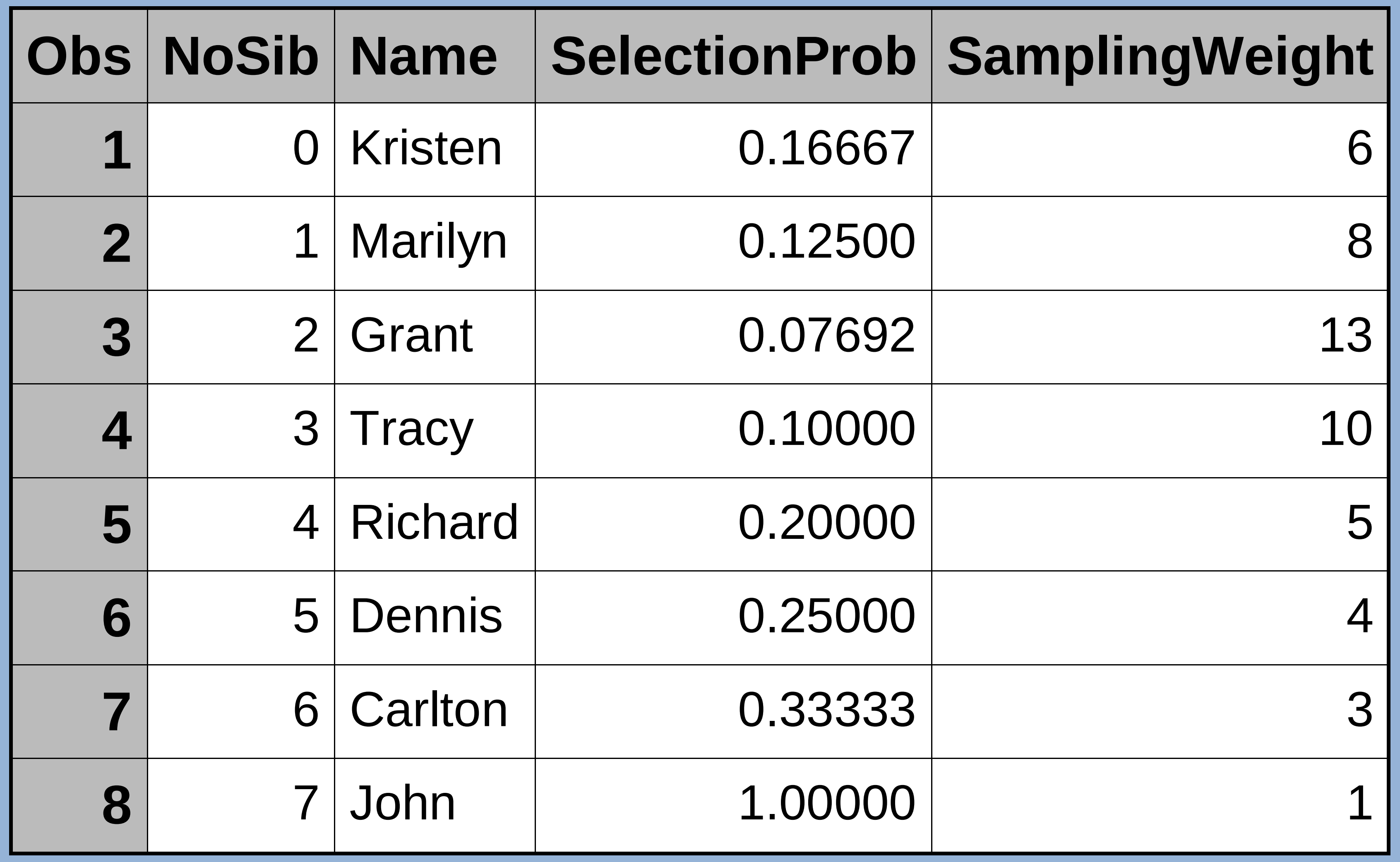
Results



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



Obs	NoSib	Name	SelectionProb	SamplingWeight
1	0	Kristen	0.16667	6
2	1	Marilyn	0.12500	8
3	2	Grant	0.07692	13
4	3	Tracy	0.10000	10
5	4	Richard	0.20000	5
6	5	Dennis	0.25000	4
7	6	Carlton	0.33333	3
8	7	John	1.00000	1

Sampling in SAS ® using PROC SURVEYSELECT

Rachael Becker and Drew Doyle

University of Central Florida

Conclusion

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

References

- 24555 - Using PROC SURVEYSELECT for single-stage cluster sampling. (n.d.). Retrieved July 17, 2015, from http://support.sas.com/kb/24/555.html
- An, A. and Watts, D. (2000). New SAS Procedures for Analysis of Sample Survey Data. SUGI 23, Retrieved from http://www2.sas.com/proceedings/sugi23/Stats/p247.pdf
- Diseker, R. and Permanente, K. (2004). Simplified Matched Case-Control Sampling using PROC SURVEYSELECT. SUGI 209-29, Retrieved from http://www2.sas.com/proceedings/sugi29/209-29.pdf
- Frerichs, R.R. Rapid Surveys (unpublished), 2008, Retrieved from http://www.ph.ucla.edu/epi/rapidsurveys/RScourse/RSbook_ch3.pdf
- Hadden, L. (2005). PROC SURVEYSELECT: A Simply Serpentine Solution for Complex Sample Designs. NESUGI 18, Retrieved from http://www.nesug.org/proceedings/nesug05/an/an5.pdf
- Putnam, D. (2011). PROC SURVEY…Says!: Selecting and Analyzing Stratified Samples. SESUGI ST-05, Retrieved from http://analytics.ncsu.edu/sesug/2011/ST05.Putnam.pdf
- Suhr, D. (2009). Selecting a Stratified Sample with PROC SURVEYSELECT. SUGI 058-2009, Retrieved from http://support.sas.com/resources/papers/proceedings09/058-2009.pdf
- Thompson, Steven K. *Sampling*. 3rd ed. Hoboken, N.J.: John Wiley & Sons, 2012. Print.
- Unknown Author, Retrieved from, http://www.math.wpi.edu/saspdf/stat/chap63.pdf

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.

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF