

# Sample selection using Proc SURVEYSELECT

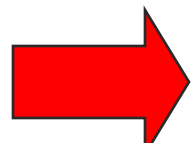
**THE  
POWER  
TO KNOW.**

---

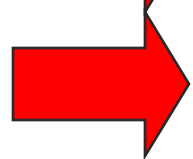
Sylvain Tremblay  
SAS Canada – Education Group

# Stratified random sampling in SAS

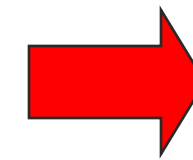
In the good old times...



```
data students;  
  set students;  
  shuffling=ranuni(seed);
```



```
proc sort data=students;  
  by age sex shuffling;
```



```
data sample;  
  retain counter;  
  set students;  
  by age sex;  
  if first.sex  
    then counter=1;  
    else counter=counter+1;  
  if counter<=10;
```



# Stratified random sampling in SAS

## proc surveyselect

```
data=suvprb.HH_frame  
samprate=.004378  
out=suvprb.HH_str_sample  
method=srs  
seed=135102203  
stats;  
  
strata region;  
|  
run;
```

Today!



**ONE procedure, that's it!**

# Agenda

- Sample design
- Selection methods
- Proc SURVEYSELECT syntax
- Demo 1 – Separate sampling
- Demo 2 – Cluster sampling
- Conclusion
- To learn more
- Questions

# Sample Design

## Process for selecting sampling units

- Equal vs non-equal probabilities
- Element vs cluster sampling
- Unstratified vs stratified selection
- Random vs systematic selection
- One phase vs multi-phase sampling

# Selection Methods

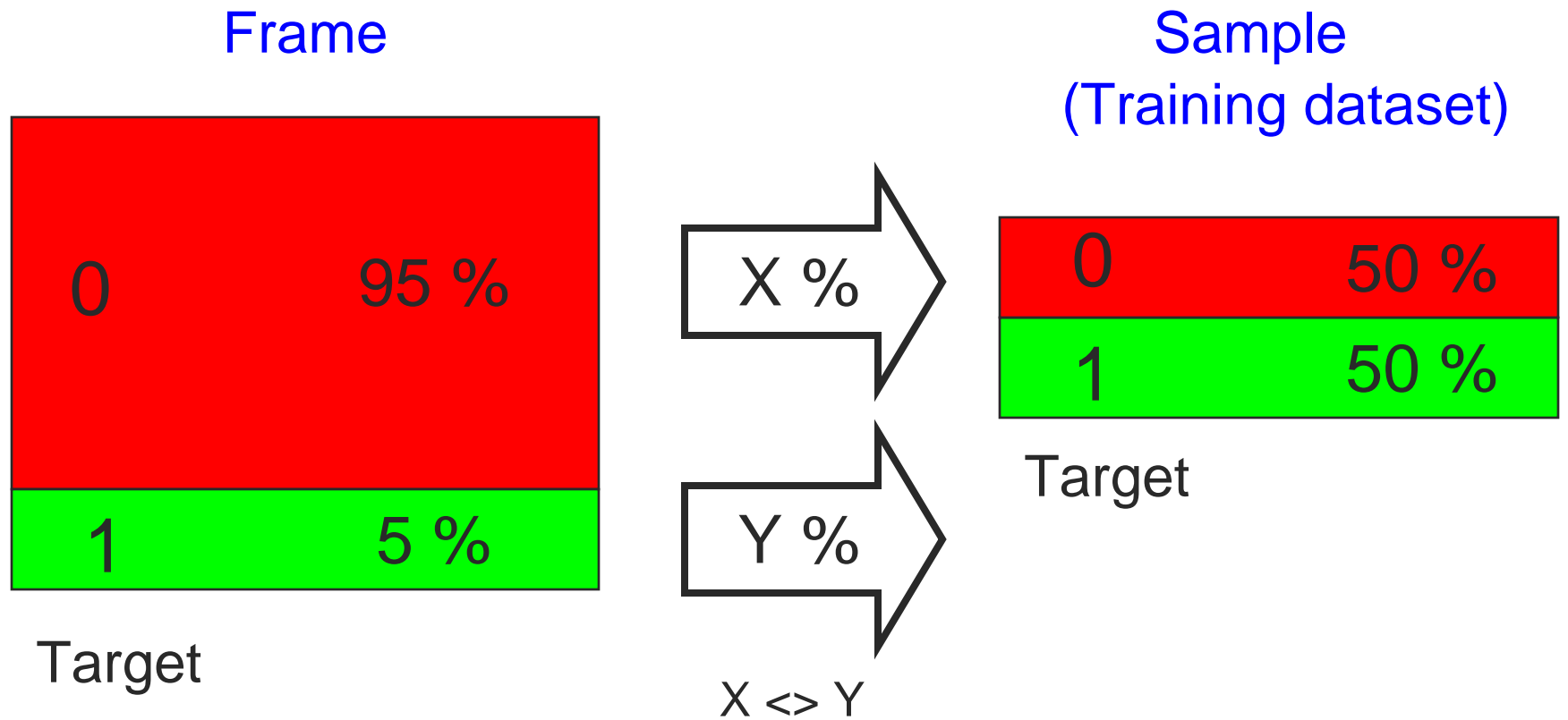
## Supported by Proc SURVEYSELECT

- simple random sampling
- unrestricted random sampling (with replacement)
- systematic
- sequential
- selection probability proportional to size (PPS) with and without replacement
- PPS systematic
- PPS for two units per stratum
- sequential PPS with minimum replacement

# Proc SURVEYSELECT - Basic syntax & options

```
PROC SURVEYSELECT  
  DATA=SAS-data-set  
  OUT=  
  METHOD=  
  SEED=  
  SAMPSIZE=  
  SAMPRATE= ;  
  STRATA variables ;  
  ID variables ;  
RUN;
```

# Demo 1 - Oversampling (separate sampling) to construct a **training** dataset



Generally,  $Y=100\%$

## Demo 2 - Two-Stage Cluster Sampling

Cluster sampling is appropriate in these situations:

- when a sampling frame of the elements of interest is unavailable
- when data collection is expensive because the elements of interest are widely dispersed

## Slide 9

---

**S111**

Cluster sampling can provide efficiency in frame construction and other survey operations. However, it can also result in a loss in precision of your estimates, compared to a nonclustered sample of the same size. To minimize this effect, units within clusters should be as heterogeneous as possible for the characteristics of interest.

SAS Institute Inc., 10/19/2007

# Two-Stage Cluster Sampling

## Examples:

### STAGE 1

### STAGE 2

<b>PSU</b>	<b>SSU</b>
City Block	Household
Household	Family Member
Clinic	Patient
Classroom	Student
Geographic Area	Small Plot
Stream	3-meter Section

# Two-Stage Cluster Sampling

## Data for the demo:

The SAS data set **suvprb.hh\_frame** is a sampling frame that contains 260,396 households in a geographic region of interest. The frame contains these columns:

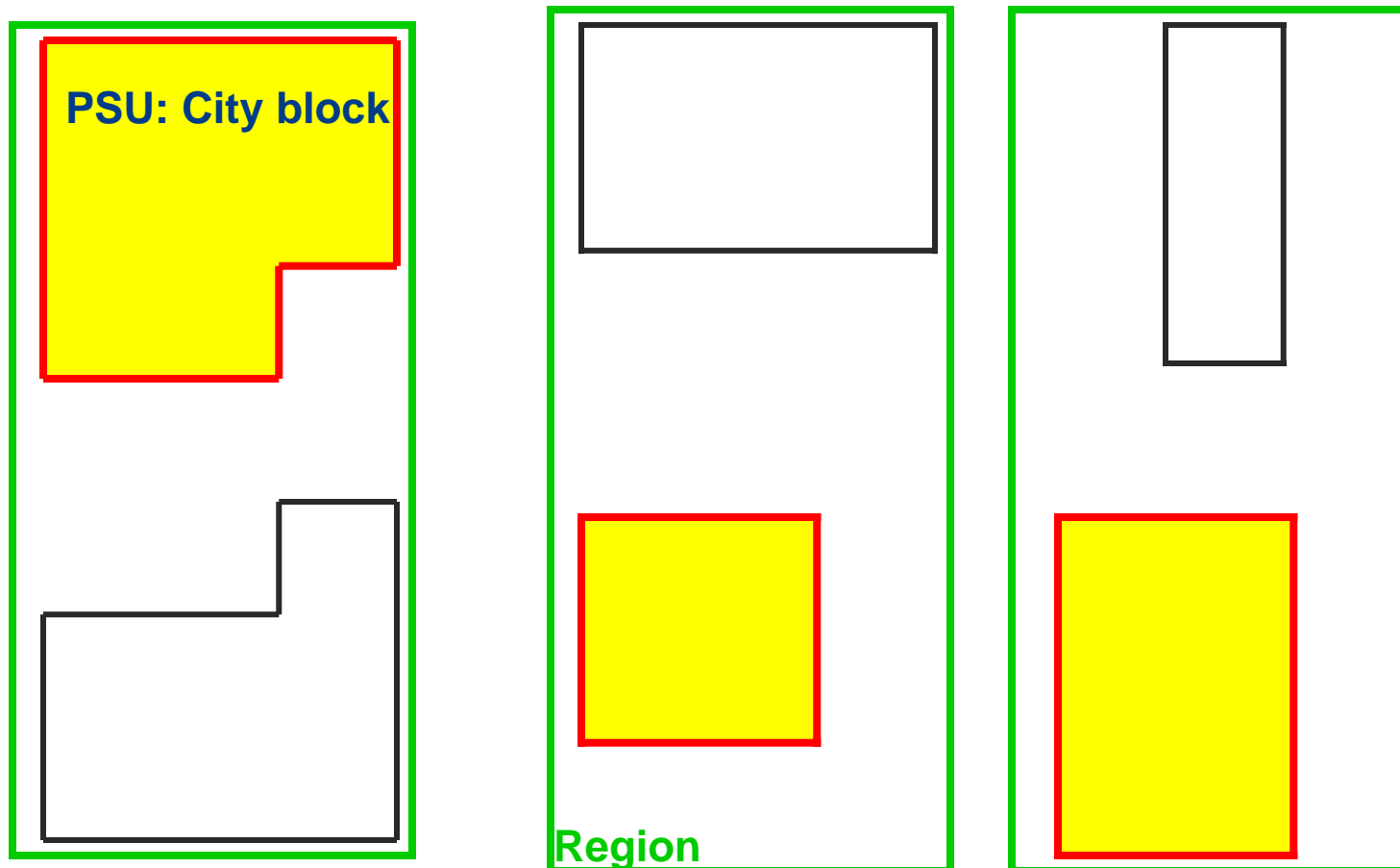
<b>Hhno</b>	household identification number
<b>Block</b>	census block number
<b>Totblock</b>	total number of households in the census block
<b>Lndblock</b>	total land area in the census block
<b>Region</b>	geographic region number



This data set is the Illinois 5% Public Use Micro Sample (PUMS) from the U.S. 2000 Census (<http://www.census.gov>).

# Two-Stage Cluster Sampling

## Stage One: Select Primary Sampling Units (PSU)

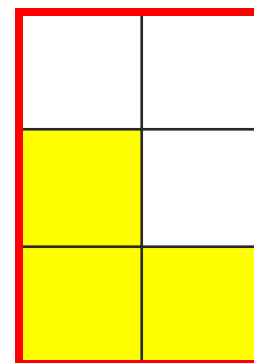
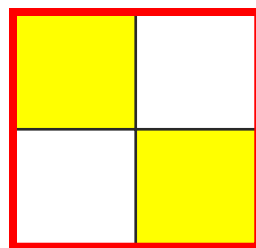
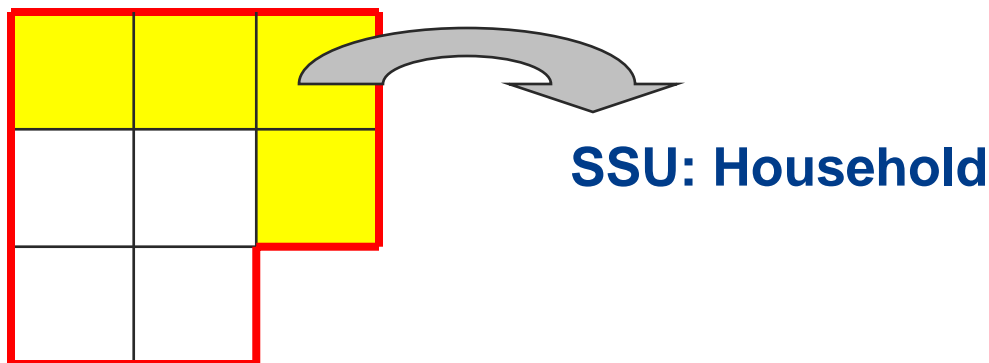


Notes: 1 - PSUs selection: stratified on **REGION**

2 - For household (hh) samples, **selection of PSU**: PPS on # hh

# Two-Stage Cluster Sampling

## Stage Two: Select Secondary Sampling Units (SSU)



Notes: 1 - SSU are randomly selected ONLY from PSUs selected in stage 1

# In conclusion

## Proc SURVEYSELECT

- Can be used for simple or complex **sample designs**
- Covers all the major **selection methods**
- Is an easy way to construct your training and validation datasets for **predictive modeling**
- is your best friend is you need to use re-sampling techniques like **bootstrap**, **jackknife** and **cross validation**

## To learn more

- [SAS Training](#): ‘Design and Analysis of Probability Surveys’
- [SAS Support Website](#) – samples on SURVEYSELECT  
<http://support.sas.com/kb>

# Questions?



THE  
POWER  
TO KNOW®

THANK YOU!

Sylvain Tremblay  
[sylvain.tremblay@sas.com](mailto:sylvain.tremblay@sas.com)