

Techie Tip from Education – provided by Lorne Rothman, Ph.D., SAS Statistical Services Specialist

The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples, allowing for valid inference when it comes time to analyze your survey data.

It may also prove quite useful for the data miner. For example, it provides a simple way to split your data into Training and Validation sets.

```
Proc surveystest noprint
    data= LEARNINGSET
    samprate=.70
    out=LEARNINGSET
    seed=5
    outall;
strata TARGET;
Run;
```

The SAMPRATE=option specifies what proportion of the LEARNINGSET is selected. By default, the procedure outputs just your sample to the OUT= dataset unless you specify the OUTALL option. The OUTALL option returns all of the original data with an extra variable called SELECTED, which takes on a value of 1 for the selected sample and 0 otherwise. In this case 1=Training cases (70%) while 0 = Validation cases (30%).

If you have a categorical target with a rare event, it is often beneficial to stratify on the target (using the STRATA statement as above) to maintain the same proportion of each target level in the training and validation data.

The results of this stratified random sample can be checked using the FREQUENCY procedure.

```
Proc freq data=LEARNINGSET;
Tables TARGET*SELECTED;
Run;
```