

Techno-conseil de formation – fourni par Lorne Rothman, Ph. D., spécialiste des services statistiques SAS

La procédure SURVEYSELECT offre une variété de méthodes pour sélectionner des échantillons aléatoires basés sur la probabilité, permettant une inférence valide au moment d'analyser vos données de sondage.

Elle peut aussi se révéler fort utile pour l'explorateur de données (*data miner*). Par exemple, elle offre une façon simple de séparer vos données dans les ensembles Formation et Validation.

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

SAMPRATE=option indique quelle proportion de LEARNINGSET est choisie. Par défaut, la procédure ne donne en sortie que votre échantillon dans l'ensemble de données OUT=, à moins que vous ne spécifiez l'option OUTALL. L'option OUTALL retourne toutes les données d'origine avec une variable supplémentaire appelée SELECTED, qui prend la valeur 1 si l'échantillon est sélectionné, et la valeur 0 autrement. Dans le cas présent, 1 = cas de formation (70 %) et 0 = cas de validation (30 %).

Si vous avez une cible catégorique avec un événement rare, la stratification est souvent profitable selon la cible (au moyen de l'énoncé STRATA mentionné ci-dessus) pour garder la même proportion de chaque niveau cible dans les données de formation et de validation.

Les résultats de cet échantillon aléatoire stratifié peuvent être vérifiés au moyen de la procédure FREQUENCY.

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