GENERALIZED PROGRAM FOR STRATIFICATION & STRATA DETERIORATION ANALYSIS - GPSSD

Bonnie Brown Jacobson, Northeast Utilities Service Company

Most utilities use load research data collected from load research studies of various subgroups of their customer population. It is hoped that the results of these studies will yield accurate profiles of the demand patterns for these subgroups for use in ratemaking, forecasting and load management.

This program is a generalized routine for the calculation of the required sample size needed to satisfy the confidence limits of 90% ± 10% and 95% ± 5% for five separate sampling designs all of which are currently utilized load research designs. Each design is further explored through the calculation of the approximate deterioration of data for each stratum. Summary tables are also generated for quick reference.

The sample sizes required for the appropriate operation of each of the following sample designs are automatically calculated in GPSSD, based on the assumption that the population size is large.

OVERALL SRS:

This design requires a simple random sample to be drawn from the total population without regard to strata boundaries. This would yield data relating to the means that are statistically reliable (for the chosen confidence limits) for the total population. No estimate of reliability can be made for any individual strata before the sample is chosen. The formula used for the sample size is:

$$n = \left(\frac{t\sigma}{e} \right)^2$$

Where: t is the Student's t-value associated with the desired confidence interval.

σ is the standard deviation of the population.

e is the percentage of the mean relating to the fiducial limits.

SRS WITHIN STRATA:

This design requires a separate simple random sample to be drawn from each of the pre-assigned stratum. The design would yield data relating to the mean that are statistically reliable (for the chosen confidence limits) for the individual strata as well as a more rigorous reliability for the total population. The formula for the sample size is:

$$n = \sum_{i=1}^{j} \left(\frac{t\sigma_i}{e_i} \right)^2$$

Where: j is the total number of strata.

t is the Student's t-value associated with the desired confidence interval.

σ_i is the population standard deviation for the ith stratum.

e_i is the percentage of the mean for the ith stratum relating to the fiducial limits.

OPTIMUM ALLOCATION:

This design requires a simple random sample to be drawn within each stratum utilizing the most cost effective method for acquiring data relating to the population mean. Thus, the design yields statistically reliable data (for the chosen confidence limits) relating to the mean of the population. An estimate of reliability is made by the program for the individual stratum, but this reliability is for the most part much more conservative than the overall population mean reliability. The formula for the sample size is:

$$n = \sum_{i=1}^{j} \frac{N_i\sigma_i(\Sigma N_i\sigma_i)}{\frac{N^2e^2}{t^2} + \left(\sum_{i=1}^{j} N_i\sigma_i^2 \right)}$$

Where: j is the total number of strata.

N_i is the population total within the ith stratum.

N is the total population.

σ_i is the population standard deviation for the ith stratum.

t is the Student's t-value associated with the desired confidence interval.

e is the percentage of the population mean relating to the fiducial limits.

SRS WITHIN STRATA-STABLE:

This design is identical to the aforementioned SRS within strata. N is raised to 30 for any stratum with an N of less than 30. This has been determined to be the least number of load research meters that allow a stable result. Thus, the sample yields statistically reliable data overall and per strata.

OPTIMUM ALLOCATION-STABLE:

Similar to the above design, this design is identical to the aforementioned optimum allocation. Again, N is substituted by 30 if it is calculated to be less than 30. This design

yields statistically reliable data overall and within strata (although somewhat more conservative than its SRS counterpart).

PROC SPECIFICATIONS:

GPSSD is invoked by a PROC statement and controlled by the following other statements:

```
PROC GPSSD options;
    MEAN variable;
    STD variable;
    N variable;
    DESC variable;
    SUBDESC variable;
    STRDES variable;
```

THE PROC GPSSD STATEMENT:

These options may appear in the PROC GPSSD statement:

STRATA = n or

STRAT = n or

S = n: specifies the number of strata desired for all sample designs, between 1 and 5. If the STRATA = option is omitted from the PROC statement, the default strata value is 5. If more than 5 strata are required, run the PROC several times.

DATA = data_set: specifies the SAS data set containing the population parameters to be utilized in the sample design. If the DATA = option is omitted from the PROC statement, the most recently created SAS data set is used. See the section concerning the input data set for specific data set format.

THE MEAN STATEMENT:

MEAN variable;

The MEAN statement identifies the variable name containing the population mean numbers. This statement is required.

THE STD STATEMENT:

STD variable;

The STD statement identifies the variable name containing the population standard deviation numbers. This statement is required.

THE N STATEMENT:

The N statement identifies the variable name containing the population size numbers. This statement is required.

THE DESC STATEMENT:

DESC variable;

The DESC statement identifies the variable name containing the population description. This statement is required.

THE SUBDES STATEMENT:

SUBDES variable;

The SUBDES statement identifies the variable name containing the subpopulation description. If the SUBDESC statement is omitted, the subpopulation description is equal to the population description.

THE STRDES STATEMENT:

STRDES variable;

The STRDES statement identifies the variable name containing the stratification variable description. This statement is required.

THE INPUT DATA SET:

The input data set must contain both the total population data and each of the population stratum data and optionally, the total subpopulation data and each of the subpopulation stratum data. The input data set must take the following general form:

Line 1: (Total population) mean, standard deviation, size, label, stratification variable label, (optionally) subpopulation label.

Line 2: (Stratum 1) mean, standard deviation, size, label, stratification variable label, (optionally) subpopulation label. Line 2 is repeated (up to 4 times) one line per stratum.

This general population information may optionally be followed by any subpopulation information. Up to 49 subpopulation data groups may follow. The form is the same as the population (i.e., first line is the general subpopulation data followed by lines of stratum data). All results from these groups will be weighted estimates which simulate probable collection data. The order of the information in each line is arbitrary. Any value found missing is treated as a "zero" and the computation of sample sizes proceeds accordingly.

OUTPUT:

The first section of output indicates the required size of the sample for the overall SRS (90% $\pm$ 10%, 95% $\pm$ 5%, the SRS within strata (90% $\pm$ 10%, 95% $\pm$ 5%) and the optimum allocation (90% $\pm$ 10%, 95% $\pm$ 5%). The SRS within stratum and optimum allocation are further explored. Each stratum sample size is listed with the expected standard error of the mean and the standard error as a percentage of the mean for each stratified sample design. This information is also printed for the total sample for each stratified design. With this information, the analyst can determine the likely accuracy of the data to be collected from each design.

The second section repeats this stratified information (since the weight for each stratum with regard to the original stratified population is equal to 1). If, for any of the above stratified designs, the stratum sample size is less than 30, it is raised to 30 and re-evaluated. These analyses yield sample sizes for the SRS within strata-stable (90% $\pm$ 10%, 95% $\pm$ 5%, and the optimum allocation stable (90% $\pm$ 10%, 95% $\pm$ 5%) designs.

If additional subpopulation and/or stratification variable data cards are included in the input data, the sample design process is repeated. Since all evaluation is done in relation to the original strata boundaries, these analyses can at best be only estimates of the accuracy of the data to be collected.

The first section of the optional data analyses is unweighed. It is only intended to give the analyst an idea of the sample size required if the subpopulation or new variable were the original.

The second section of the optional data analyses reflects the potential behavior of the subpopulation and/or new variable within the original population-stratification variable framework. All sample sizes are weighted to simulate the true sample size applicable to the optional situation. The analyst should keep in mind that this information is an estimate of variable accuracy.

Following the optional analyses is a summary table for each variable and population/subpopulation situation. Listed are the required sample sizes for each stratum within each sample design. Also listed is the value of 100 minus t times the standard error (expressed as a percentage of the mean) for the "best" and "worst" strata. This quantity is relative to the mean in that it is a measure of how close to the true mean the stratum should result. The best possible stratum would result in a value of 100.00. Evaluation should be based on this standard.

The final summary table lists the overall standard error times t expressed as a percentage of the mean for each sample design and for each variable and population/subpopulation situation.

ACKNOWLEDGEMENTS

I would like to thank the following people for their help in the development of both the PROC and this paper:
Ms. Karen E. Greeley, Mr. James D. Oleksiw, Mrs. Jean H. Ehle.

REFERENCES

Cochran, W. G., Sampling Techniques Third Edition, c. 1977, John Wiley & Sons, Inc., New York, New York.

Kish, L., Survey Sampling, c. 1965, John Wiley & Sons, Inc., New York, New York.

For more information, please contact Bonnie B. Jacobson, Consumer Economics, Northeast Utilities Service Company, P. O. Box 270, Hartford, CT 06101 or call (203) 666-6911, Ext. 5030.

INT._COMMERCIAL

SUBSTRATIFICATION VARIABLE: MAX._KW
SUBPOPULATION: INT._COMMERCIAL

	POPULATION	POP MEAN	POP STD	%(90)	%(95)	SAMPLE SIZE WITH T-VALUE = 1.645 STRATA ACCURACY	SAMPLE SIZE WITH T-VALUE = 1.645 OPTIMUM ALLOCATION	SAMPLE SIZE WITH T-VALUE = 1.960 STRATA ACCURACY	SAMPLE SIZE WITH T-VALUE = 1.960 OPTIMUM ALLOCATION
UNSTRATIFIED	9839	169.000	166.000	.100	.050	262		1483	
STRATA #1	4026	65.000	10.000	.100	.050	7	2	37	7
STRATA #2	2528	111.000	18.000	.100	.050	8	2	41	8
STRATA #3	1573	198.000	33.000	.100	.050	8	2	43	9
STRATA #4	1064	349.000	57.000	.100	.050	8	2	41	10
STRATA #5	648	671.000	152.000	.100	.050	14	3	79	17
TOTAL						45	11	241	51
		SAMPLE ACCURACY: T * STANDARD ERROR OF MEAN				7.357	14.943	3.696	8.228
		% OF TOTAL POPULATION MEAN				4.353	8.842	2.187	4.869

	SAMPLE WITH T-VALUE = 1.645						SAMPLE WITH T-VALUE = 1.960					
	SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY		
	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN
STRATA #1	7	6.2	9.565	2	11.6	17.895	37	3.2	4.957	7	7.4	11.397
STRATA #2	8	10.5	9.431	2	20.9	18.863	41	5.5	4.964	8	12.5	11.237
STRATA #3	8	19.2	9.693	2	38.4	19.387	43	9.9	4.982	9	21.6	10.889
STRATA #4	8	33.2	9.499	2	66.3	18.998	41	17.4	4.999	10	35.3	10.123
STRATA #5	14	66.8	9.959	3	144.4	21.514	79	33.5	4.995	17	72.3	10.768
TOTAL SAMPLE	45	7.4	4.353	11	14.9	8.842	241	3.7	2.187	51	8.2	4.869

EVALUATION BASED ON NUMBER OF METERS FROM ORIGINAL STRATIFICATION VARIABLE ANALYSIS (WEIGHTED)

	SAMPLE WITH T-VALUE = 1.645						SAMPLE WITH T-VALUE = 1.960					
	SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY		
	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN
STRATA #1	7	6.2	9.565	2	11.6	17.895	37	3.2	4.957	7	7.4	11.397
STRATA #2	8	10.5	9.431	2	20.9	18.863	41	5.5	4.964	8	12.5	11.237
STRATA #3	8	19.2	9.693	2	38.4	19.387	43	9.9	4.982	9	21.6	10.889
STRATA #4	8	33.2	9.499	2	66.3	18.998	41	17.4	4.999	10	35.3	10.123
STRATA #5	14	66.8	9.959	3	144.4	21.514	79	33.5	4.995	17	72.3	10.768
TOTAL SAMPLE	45	7.4	4.353	11	14.9	8.842	241	3.7	2.187	51	8.2	4.869

EVALUATION WITH MINIMUM 30 METERS/STRATA (WEIGHTED)

	SAMPLE WITH T-VALUE = 1.645						SAMPLE WITH T-VALUE = 1.960					
	SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY		
	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN
STRATA #1	30	3.0	4.621	30	3.0	4.621	37	3.2	4.957	30	3.6	5.505
STRATA #2	30	5.4	4.870	30	5.4	4.870	41	5.5	4.964	30	6.4	5.803
STRATA #3	30	9.9	5.006	30	9.9	5.006	43	9.9	4.982	30	11.8	5.964
STRATA #4	30	17.1	4.905	30	17.1	4.905	41	17.4	4.999	30	20.4	5.844
STRATA #5	30	45.7	6.803	30	45.7	6.803	79	33.5	4.995	30	54.4	8.106
TOTAL SAMPLE	150	4.0	2.354	150	4.0	2.354	241	3.7	2.187	150	4.7	2.804

INT._COMMERCIAL

SUBSTRATIFICATION VARIABLE: ANN_KWH
SUBPOPULATION: INT._COMMERCIAL

	POPULATION	POP MEAN	POP STD	%(90)	%(95)	SAMPLE SIZE WITH T-VALUE = 1.645 STRATA ACCURACY	SAMPLE SIZE WITH T-VALUE = 1.645 OPTIMUM ALLOCATION	SAMPLE SIZE WITH T-VALUE = 1.960 STRATA ACCURACY	SAMPLE SIZE WITH T-VALUE = 1.960 OPTIMUM ALLOCATION
UNSTRATIFIED	9836	459537.000	681517.000	.100	.050	596		3380	
STRATA #1	4025	140437.000	82115.000	.100	.050	93	14	526	68
STRATA #2	2526	256701.000	146655.000	.100	.050	89	15	502	76
STRATA #3	1573	552468.000	753941.000	.100	.050	504	48	1573	243
STRATA #4	1064	1027734.000	520079.000	.100	.050	70	23	394	114
STRATA #5	648	2073734.000	1073912.000	.100	.050	73	29	413	143
TOTAL						829	129	3408	644
SAMPLE ACCURACY: T * STANDARD ERROR OF MEAN						16601.449	45566.820	5895.426	22947.754
% OF TOTAL POPULATION MEAN						3.613	9.916	1.283	4.994

	SAMPLE WITH T-VALUE = 1.645						SAMPLE WITH T-VALUE = 1.960					
	SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY		
	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN
STRATA #1	93	14007.1	9.974	14	36101.4	25.706	526	7017.6	4.997	68	19517.5	13.898
STRATA #2	89	25572.2	9.962	15	62289.9	24.266	502	12829.2	4.998	76	32972.1	12.845
STRATA #3	504	55244.4	10.000	48	179012.2	32.402	1573	0.0	0.0	243	94796.1	17.159
STRATA #4	70	102255.4	9.950	23	178390.3	17.358	394	51354.4	4.997	114	95471.3	9.289
STRATA #5	73	206763.2	9.971	29	328046.7	15.819	413	103573.7	4.995	143	176017.8	8.488
TOTAL SAMPLE	829	16601.4	3.613	129	45566.8	9.916	3408	5895.4	1.283	644	22947.8	4.994

EVALUATION BASED ON NUMBER OF METERS FROM ORIGINAL STRATIFICATION VARIABLE ANALYSIS (WEIGHTED)

	SAMPLE WITH T-VALUE = 1.645						SAMPLE WITH T-VALUE = 1.960					
	SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY		
	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN
STRATA #1	7	51055.1	36.354	2	95515.4	68.013	37	26459.3	18.841	7	60831.6	43.316
STRATA #2	8	85293.9	33.227	2	170587.8	66.454	41	44891.2	17.488	8	101626.7	39.590
STRATA #3	8	436488.6	79.369	2	876977.6	158.738	43	225350.8	40.790	9	492574.7	89.159
STRATA #4	8	302475.6	29.431	2	604951.3	58.863	41	159196.4	15.490	10	322348.3	31.365
STRATA #5	14	472139.9	22.768	3	1019938.7	49.184	79	236816.1	11.420	17	510505.3	24.618
TOTAL SAMPLE	45	77827.7	16.936	11	157965.4	34.375	241	39251.6	8.542	51	87051.5	18.943

EVALUATION WITH MINIMUM 30 METERS/STRATA (WEIGHTED)

	SAMPLE WITH T-VALUE = 1.645						SAMPLE WITH T-VALUE = 1.960					
	SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY			SAMPLE ACCURACY		
	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN	SAMPLE SIZE	T * STD ERR OF MEAN	% OF POP MEAN
STRATA #1	30	24662.0	17.561	30	24662.0	17.561	37	26459.3	18.841	30	29384.5	20.924
STRATA #2	30	44045.6	17.158	30	44045.6	17.158	41	44891.2	17.488	30	52479.8	20.444
STRATA #3	30	226434.6	40.986	30	226434.6	40.986	43	225350.8	40.790	30	269794.3	48.834
STRATA #4	30	156197.7	15.198	30	156197.7	15.198	41	159196.4	15.490	30	186107.9	18.109
STRATA #5	30	322532.9	15.553	30	322532.9	15.553	79	236816.1	11.420	30	384294.5	18.532
TOTAL SAMPLE	150	42163.7	9.175	150	42163.7	9.175	241	39251.6	8.542	150	50237.5	10.932

POPULATION: INT._COMMERCIAL

DETERIORATION OF STRATA

SUBSTRATIFICATION VARIABLE: MAX._KW
SUBPOPULATION: INT._COMMERCIAL

STRATA = 5

#METERS FOR:	STRATA 1	2	3	4	5	TOTALS	100% - |T * S.E.| (AS % OF MEAN) BEST STRATA	WORST STRATA
90% +- 10% SRS	7	8	8	8	14	45	90.57	90.04
95% +- 5% SRS	37	41	43	41	79	241	95.04	95.00
90% +- 10% OPT	2	2	2	2	3	11	82.10	78.49
95% +- 5% OPT	7	8	9	10	17	51	89.88	88.60
90% +- 10% OVERALL SRS						262		
95% +- 5% OVERALL SRS						1483		
90% +- 10% SRS-STABLE	30	30	30	30	30	150	95.38	93.20
95% +- 5% SRS-STABLE	37	41	43	41	79	241	95.04	95.00
90% +- 10% OPT-STABLE	30	30	30	30	30	150	95.38	93.20
95% +- 5% OPT-STABLE	30	30	30	30	30	150	94.49	91.89

POPULATION: INT._COMMERCIAL

DETERIORATION OF STRATA

SUBSTRATIFICATION VARIABLE: ANN_KWH
SUBPOPULATION: INT._COMMERCIAL

STRATA = 5

#METERS FOR:	STRATA 1	2	3	4	5	TOTALS	100% - |T * S.E.| (AS % OF MEAN) BEST STRATA	WORST STRATA
90% +- 10% SRS	7	8	8	8	14	45	77.23	20.63
95% +- 5% SRS	37	41	43	41	79	241	88.58	59.21
90% +- 10% OPT	2	2	2	2	3	11	50.82	31.99
95% +- 5% OPT	7	8	9	10	17	51	75.38	10.84
90% +- 10% OVERALL SRS						262		
95% +- 5% OVERALL SRS						1483		
90% +- 10% SRS-STABLE	30	30	30	30	30	150	84.80	59.01
95% +- 5% SRS-STABLE	37	41	43	41	79	241	88.58	59.21
90% +- 10% OPT-STABLE	30	30	30	30	30	150	84.80	59.01
95% +- 5% OPT-STABLE	30	30	30	30	30	150	81.89	51.17

POPULATION: INT._COMMERCIAL

OVERALL SUMMARY TABLE

T * S.E. AS % OF MEAN

(BASED ON ORIGINAL METER ALLOCATIONS)

DESIGN	MAX._KW	ANN_KWH				
90% +- 10% SRS	4.353	16.936	0.0	0.0	0.0	0.0
95% +- 5% SRS	2.187	8.542	0.0	0.0	0.0	0.0
90% +- 10% OPT	8.842	34.375	0.0	0.0	0.0	0.0
95% +- 5% OPT	4.869	18.943	0.0	0.0	0.0	0.0
90% +- 10% SRS STABLE	2.354	9.175	0.0	0.0	0.0	0.0
95% +- 5% SRS STABLE	2.187	8.542	0.0	0.0	0.0	0.0
90% +- 10% OPT STABLE	2.354	9.175	0.0	0.0	0.0	0.0
95% +- 5% OPT STABLE	2.804	10.932	0.0	0.0	0.0	0.0
90% +- 10% OVERALL SRS	9.982	15.072	0.0	0.0	0.0	0.0
95% +- 5% OVERALL SRS	4.999	7.548	0.0	0.0	0.0	0.0