MACRO CONSULTING, INC.
Marketing Research / Advanced Analytics

Trade-Off Study Sample Size: How Low Can We go?

The effect of sample size on model error is examined through several commercial data sets, using five trade-off techniques: ACA, ACA/HB, CVA, HB-Reg and CBC/HB. Using the total sample to generate surrogate holdout cards, numerous subsamples are drawn, utilities estimated and model results compared to the total sample model. Latent class analysis is used to model the effect of sample size, number of parameters and number of tasks on model error.

Author: Paul Richard "Dick" McCullough

2001 Sawtooth Software Conference Proceedings,
September 2001,
Victoria, BC, CANADA.

Trade-Off Sample Size:
How Low Can We Go?

Abstract

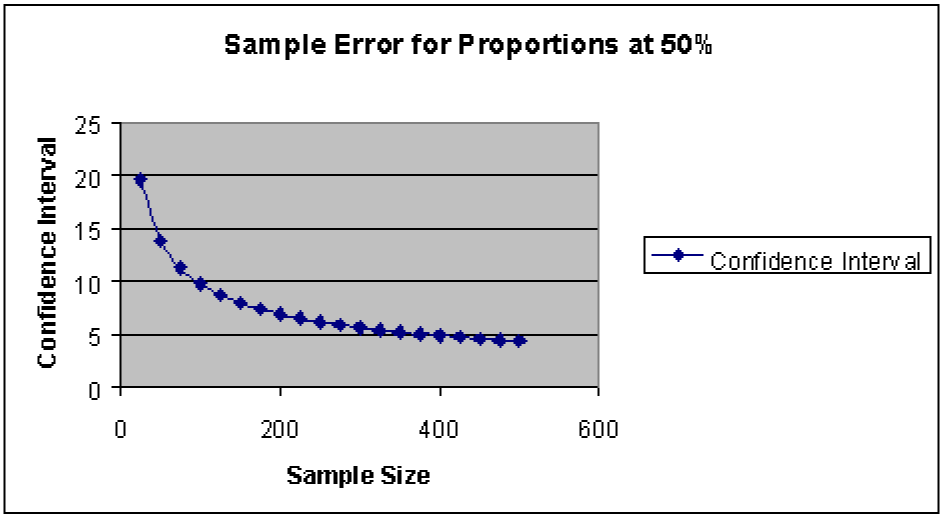
The effect of sample size on model error is examined through several commercial data sets, using five trade-off techniques: ACA, ACA/HB, CVA, HB-Reg and CBC/HB. Using the total sample to generate surrogate holdout cards, numerous subsamples are drawn, utilities estimated and model results compared to the total sample model. Latent class analysis is used to model the effect of sample size, number of parameters and number of tasks on model error.

Introduction

Effect of sample size on study precision is always an issue to commercial market researchers. Sample size is generally the single largest out-of-pocket cost component of a commercial study. Determining the minimum acceptable sample size plays an important role in the design of an efficient commercial study.

For simple statistical measures, such as confidence intervals around proportions estimates, the effect of sample size on error is well known (see Figure 1). For more complex statistical processes, such as conjoint models, the effect of sample size on error is much more difficult to estimate. Even the definition of error is open to several interpretations.

Figure 1.



Many issues face practitioners when determining sample size:

- Research objectives
- Technique
- Number of attributes and levels
- Number of tasks
- Expected heterogeneity
- Value of the information
- Cost and timing
- Measurement error
- Structure and efficiency of experimental design:
    - Fixed designs
    - Blocked designs
    - Individual-level designs

Some of these issues are statistical in nature, such as number of attributes and levels, and some of these issues are managerial in nature, such as value of the information, cost and timing. The commercial researcher needs to address both types of issues when determining sample size.

Objectives

The intent of this paper is to examine a variety of commercial data sets in an empirical way to see if some comments can be made about the effect of sample size on model error. Additionally, the impact of several factors: number of attributes and levels, number of tasks and trade-off technique, on model error will also be investigated.

Method

For each of five trade-off techniques, ACA, ACA/HB, CVA, HB-Reg, and CBC/HB, three commercial data sets were examined (the data sets for ACA, and CVA also served as the data sets for ACA/HB and HBReg, respectively). Sample size for each data set ranged between 431 and 2,400.

Since these data sets were collected from a variety of commercial marketing research firms, there was little control over the number of attributes and levels or the number of tasks. Thus, while there was some variation in these attributes, there was less experimental control than would be desired, particularly with respect to trade-off technique.

Table 1.

	Attr	Lvls	Pars	Tasks	df	SS
CBC/HB						
Data Set 1	4	14	11	8	-3	612
Data Set 2	6	17	12	18	+6	422
Data Set 3	5	25	21	12	-9	444
CVA,HB-Reg						
Data Set 1	6	24	19	30	+11	2,400
Data Set 2	4	9	6	10	+4	431
Data Set 3	6	13	8	16	+8	867
ACA,ACA/HB						
Data Set 1	25	78	54			782
Data Set 2	5	24	20			500
Data Set 3	17	63	47			808

Notice in Table 1 above that the number of parameters and number of tasks are somewhat correlated with trade-off technique. CBC/HB data sets tended to have fewer degrees of freedom (number of tasks minus the number of parameters) than CVA data sets. ACA data sets had a much greater number of parameters than either CBC/HB or CVA data sets. These correlations occur quite naturally in the commercial sector. Historically, choice models have been estimated at the aggregate level while CVA models are estimated at the individual level. By aggregating across respondents, choice study designers could afford to use fewer tasks than necessary for estimating individual level conjoint models. Hierarchical Bayes methods allow for the estimation of individual level choice models without making any additional demands on the study's experimental design. A major benefit of ACA is its ability to accommodate a large number of parameters.

For each data set, models were estimated using a randomly drawn subset of the total sample, for the sample sizes of 200, 100, 50 and 30. In the cases of ACA and CVA, no new utility estimation was required, since each respondent's utilities are a function of just that respondent. However, for CBC/HB, HB-Reg and ACA/HB, new utility estimations occurred for each draw, since each respondent's utilities are a function of not only that respondent, but also the "total" sample. For each sample size, random draws were replicated up to 30 times. The number of replicates increased as sample size decreased. There were five replicates for n=200, 10 for n=100, 20 for n=50 and 30 for n=30. The intent here was to stabilize the estimates to get a true sense of the accuracy of models at that sample size.

Since it was anticipated that many, if not all, of the commercial data sets to be analyzed in this paper would not contain holdout choice tasks, models derived from reduced samples were compared to models derived from the total sample. That is, in order to evaluate how well a smaller sample size was performing, 10 first choice simulations were run for both the total sample model and each of the reduced sample models, with the total sample model serving to generate surrogate holdout tasks. Thus, MAEs (Mean Absolute Error) were the measure with which models were evaluated (each sub-sample model being compared to the total sample model). 990 models (5 techniques x 3 data sets x 66 sample sizes/replicate combinations) were estimated and evaluated. 9,900 simulations were run (990 models x 10 simulations) as the basis for the MAE estimations.

Additionally, correlations were run, at the aggregate level, between the mean utilities from each of the sub-sample models and the total sample model. Correlation results were reported in the form 100 * (1-rsquared), and called, for the duration of this paper, mean percentage of error (MPE).

It should be noted that there is an indeterminacy inherent in conjoint utility scaling that makes correlation analysis potentially meaningless. Therefore, all utilities were scaled so that the levels within attribute summed to zero (effects coding). This allowed for meaningful correlation analysis to occur.

Sample Bias Analysis

Since each subsample was being compared to a larger sample, of which it was also a part, there was a sample bias inherent in the calculation of error terms.

Several studies using synthetic data were conducted to determine the magnitude of the sample bias and develop correction factors to adjust the raw error terms for sample bias.

Sample Bias Study 1

For each of four different scenarios, random numbers between 1 and 20 were generated 10 times for two data sets of sample size 200. In the first scenario, the first 100 data points were identical for the two data sets and the last 100 were independent of one another. In the second scenario, the first 75 data points were identical for the two data sets and the last 125 were independent of one another. In the third scenario, the first 50 data points were identical for the two data sets and the last 150 were independent of one another. And in the last scenario, the first 25 data points were identical for the two data sets and the last 175 were independent of one another.

The correlation between the two data sets, r, approximately equals the degree of overlap, n/N, between the two data sets (Table 2).

Table 2.

N=200				
n=	100	75	50	25
	0.527451	0.320534	0.176183	0.092247
	0.474558	0.411911	0.255339	0.142685
	0.61104	0.3109	0.226798	0.11125
	0.563223	0.287369	0.223945	0.194286
	0.487692	0.398193	0.368615	0.205507
	0.483789	0.47338	0.229888	-0.09505
	0.524381	0.471472	0.288293	0.250967
	0.368708	0.274371	0.252346	0.169203
	0.446393	0.401521	0.245936	0.109158
	0.453217	0.389331	0.139375	0.184337
r=	0.494045	0.373898	0.240672	0.136459
n/N =	0.5	0.375	0.25	0.125

Sample Bias Study 2

To extend the concept further, a random sample of 200 was generated, a second sample of 100 was created where each member of the second sample was equal to a member of the first sample and a third sample of a random 100 was generated, independent of the first two.

For each of the three samples, the mean was calculated. This process was replicated 13 times and the mean data are reported below (Table 3).

The absolute difference (MAE) between the first two data sets is 0.147308 and the absolute difference between the first and third data sets is 0.218077. By dividing the MAE for the first two data sets by the finite population correction factor (sqrt(1-n/N)), the MAEs become quite similar.

Table 3.

	N=200	n=100	n=100
	11.075	11.18	9.54
	10.275	10.15	11.15
	10.85	11.15	10.62
	10.595	10.51	10.81
	9.99	9.92	10.88
	9.735	10.11	11.19
	10.555	11.3	11.43
	11.44	11.68	10.88
	10.41	10.33	9.37
	10.13	10.55	10.87
	10.34	9.84	11.23
	10.295	10.86	11.46
	10.855	10.88	9.95
	10.50346	10.65077	10.72154
MAE		0.147308	0.218077
MAE/sqrt(1-n/N)=		0.208325	

Sample Bias Study 3

To continue the extension of the concept, a random sample of 200 was generated, a second sample of 100 was created where each member of the second sample was equal to a member of the first sample and a third sample of a random 100 was generated.

The squared correlation was calculated for the first two samples and for the first and third samples. This procedure was replicated 11 times. The 11 squared correlations for the first two samples were averaged as were the 11 squared correlations for the first and third samples.

MPEs were caculated for both mean r-squares (Table 4). The MPE for the first two sample is substantially smaller than the MPE for the first and third samples. By dividing the MPE for the first two samples by the square of the finite population correction factor (1-n/N), the MPEs become quite similar.

Note that it is somewhat intuitive that the correction factor for the MPEs is the square of the correction factor for the MAEs. MPE is a measure of squared error whereas MAE is a measure of first power error.

Table 4.

	ns=100	n(R)=100
rsq=	0.603135	0.099661
rsq=	0.648241	0.048967
rsq=	0.357504	0.11173
rsq=	0.30337	0.099186
rsq=	0.790855	0.178414
rsq=	0.883459	0.379786
rsq=	0.829014	0.182635
rsq=	0.477881	0.27063
rsq=	0.798317	0.010961
rsq=	0.425018	0.462108
rsq=	0.785462	0.003547
average rsq=	0.627478	0.167966
MPE=	37.25222	83.2034
MPE/(1-n/N)=	74.50445	

Sample Bias Study 4

Finally, the synthetic data study below involves more closely replicating the study design used in this paper.

<u>Method</u>

The general approach was:

- Generate three data sets
  - Each data set consists of utility weights for three attributes
  - Utility weights for the first and third data sets are randomly drawn integers between 1 and 20
  - Sample size for the first data set is always 200
  - Sample size for the second and third data sets varies across 25, 50 and 100
  - The second and third data sets always are of the same size
  - The second data set consists of the first n cases of the first data set, where n = 25, 50 or 100
- Define either a two, three, four or five product scenario

- Estimate logit-based share of preference models for each of the three data sets, calculating shares at the individual level, then averaging
- Calculate MAEs for each of the second and third data sets, compared to the first, at the aggregate level
- Calculate MPEs (mean percent error = (1- rsq(utils-first data set, utils-other data set))*100) for each of the second and third data sets, compared to the first, at the aggregate level
- Redraw the sample 50 times for each scenario/sample size and make the above calculations
- Calculate mean MAEs and MPEs for each of 50 random draws for each model
- 36 models (3 data sets x 4 market scenarios x 3 sample sizes)

Note: Empirically, the ratio of random sample MAE to overlapping sample MAE equals the scalar that corrects the overlapping sample MAE for sample bias. Similarly for MPE. The issue, then, is to develop a formula for the correction factor that closely resembles the ratio of random sample error/overlapping sample error.

Conclusion

As suggested by Synthetic Data Study 2, the formula (1/(1-percent overlap))^0.5 may represent the desired scalar for correction for MAE. Similarly, as suggested by Synthetic Data Study 3, the formula 1/(1-percent overlap) may represent the desired scalar for correction for MPE:

Table 5.

MAE

Percent Overlap	(1/1-%overlap)^0.5	random/overlap
12.5% (n=25)	1.07	1.17
25% (n=50)	1.15	1.32
50% (n=100)	1.41	1.56

MPE

Percent Overlap	1/1-%overlap	random/overlap
12.5% (n=25)	1.14	1.18
25% (n=50)	1.33	1.84
50% (n=100)	2.00	2.95

Figure 2.

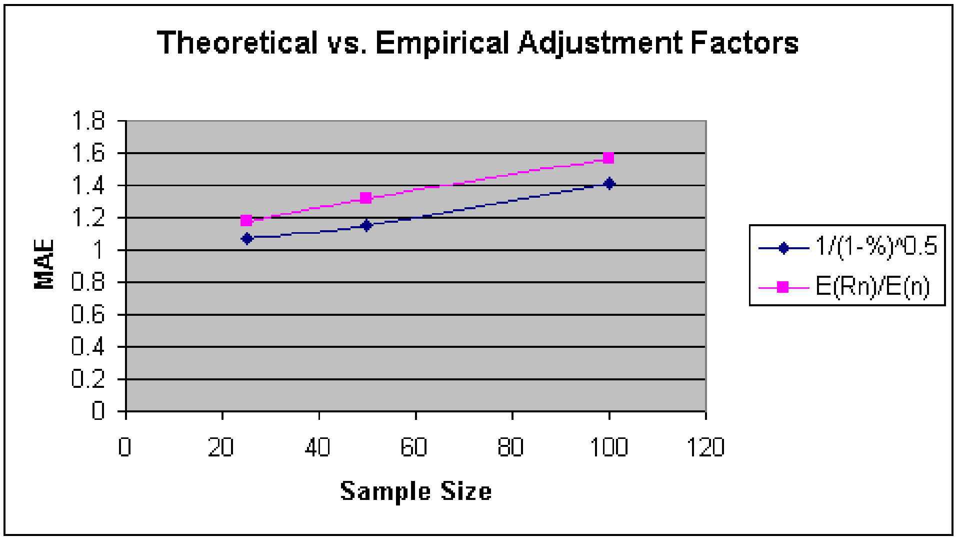
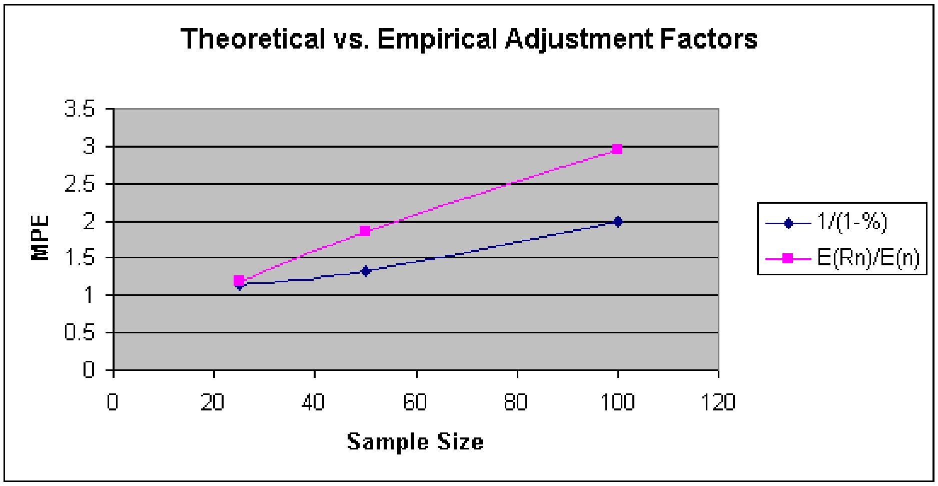


Figure 3.



Additional conclusions:

- There is a definite bias due to overlapping sample, both in MAE and MPE.
- This bias appears to be independent of the number of products in the simulations (see Tables 6 and 7).
- The bias is directly related to the percent of the first data set duplicated in the second.
- The amount of bias is different for MAE and MPE.

Table 6.

MAE					
	two products	**three products**	**four products**	**five products**	**mean**
n=25	0.077167327	0.065818796	0.04373055	0.03523201	0.055487
n=R25	0.081864359	0.078921973	0.05091588	0.04456126	0.064066
n=50	0.041639879	0.046603952	0.0292088	0.02401724	0.035367
n=R50	0.057030973	0.057865728	0.03926258	0.03140596	0.046391
n=100	0.02421646	0.024658317	0.01847943	0.01383198	0.020297
n=R100	0.033042464	0.040345804	0.02954819	0.02281538	0.031438

Table 7.

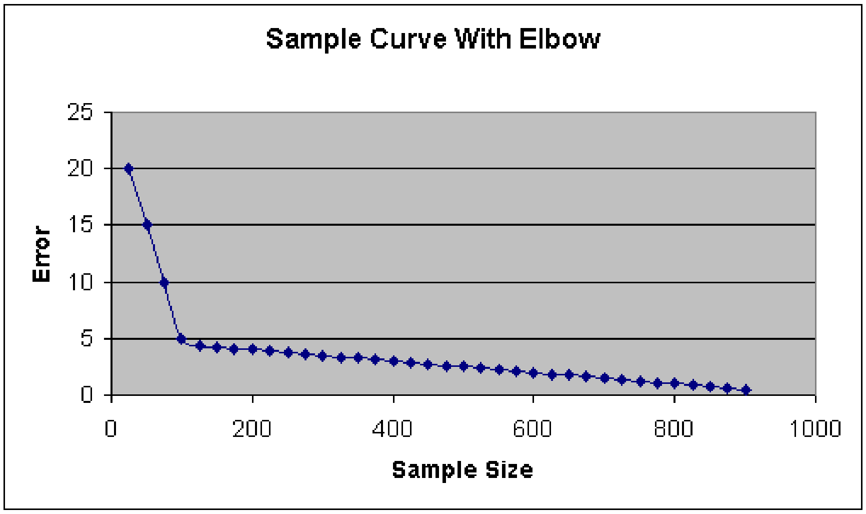
MPE					
	two products	**three products**	**four products**	**five products**	**mean**
n=25	0.70718724	0.687751783	0.85695737	0.664759341	0.729164
n=R25	0.785403871	0.870813277	0.869094024	0.88440592	0.852429
n=50	0.242856551	0.312908934	0.292542572	0.246179851	0.273622
n=R50	0.437063715	0.554906027	0.453530099	0.552845437	0.499586
n=100	0.094198823	0.096766941	0.123103025	0.099623936	0.103423
n=R100	0.281835972	0.335639163	0.490892078	0.296887426	0.351314

Sample Size Study Results

Referring to the error curve for proportions once again (Figure 1), a natural point to search for in the error curve would be an elbow. An elbow would be a point on the curve where any increase in sample size would result in a declining gain in precision and any decrease in sample size would result in an increasing loss in precision. This elbow, if it exists, would identify the maximally efficient sample size.

Visually, and intuitively, an elbow would appear as noted in Figure 4.

Figure 4.



To formally identify an elbow, one would need to set the third derivative of the error function to zero. It is easy to demonstrate that, for the proportions error curve, the third derivative of the error function cannot be zero. Therefore, for a proportions error curve, an elbow does not exist.

Below in Figure 5 and in Figure 7, the error curves for both the MAE and MPE error terms have been plotted for the aggregate data, that is, for all five techniques averaged together. In Figures 6 and 8, the error curves for each trade-off technique has been plotted separately.

The MAE curves are all similar in shape to one another as are the MPE curves.

Visually, the MAE curves appear to be proportionate to 1/sqrt(n) and the MPE curves appear to proportionate to 1/n. By regressing the log of the error against the log of sample size it can be confirmed that the aggregate MAE is indeed proportionate to 1/sqrt(n) and the aggregate MPE proportionate to 1/n (coefficients of –0.443 and –0.811, respectively).

The third derivative of both 1/sqrt(n) and 1/n can never equal zero. Therefore, neither of these error curves can have an elbow.

Figure 5.

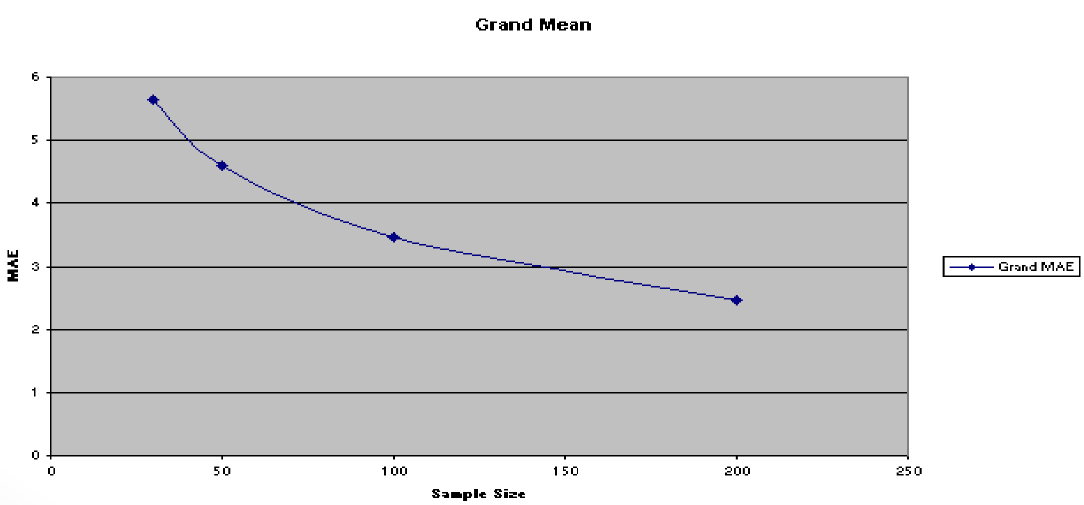


Figure 6.

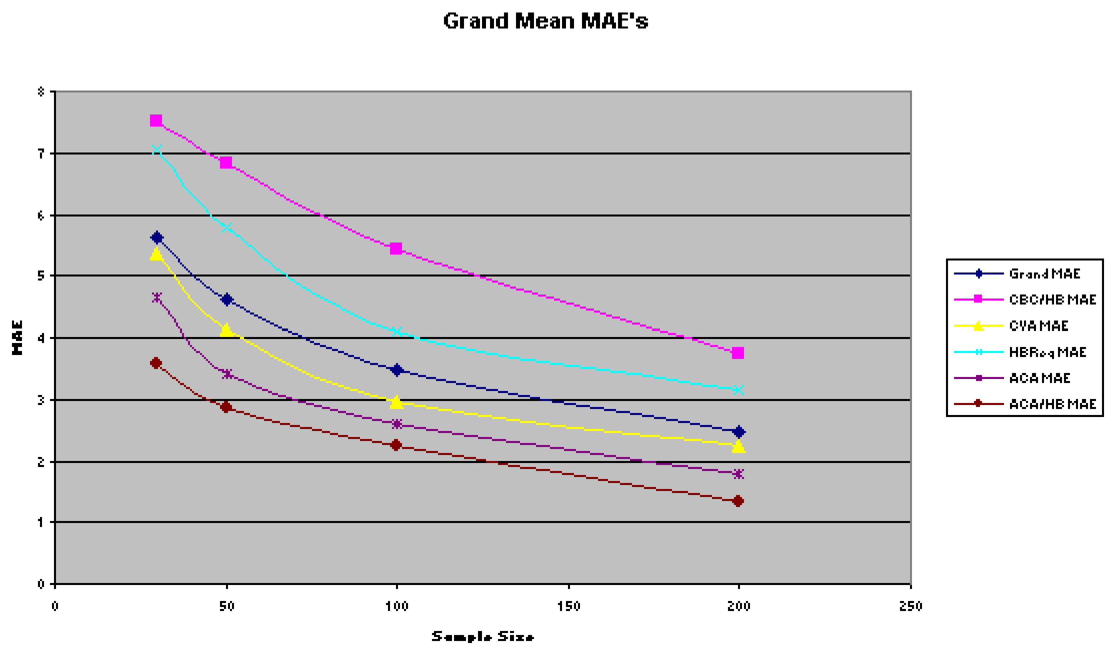


Figure 7.

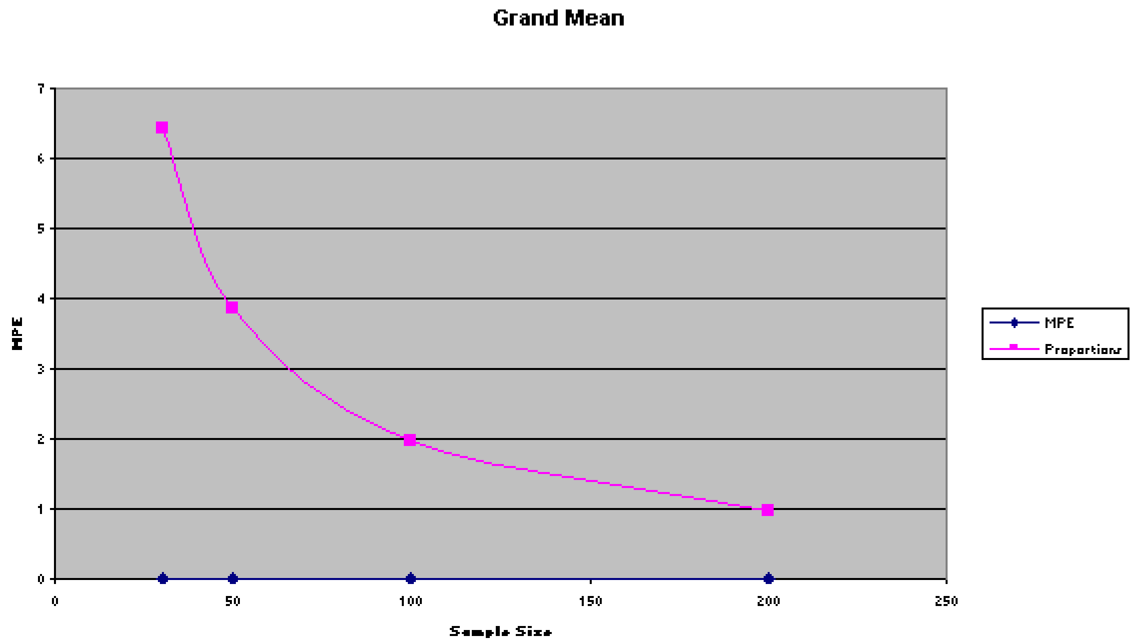
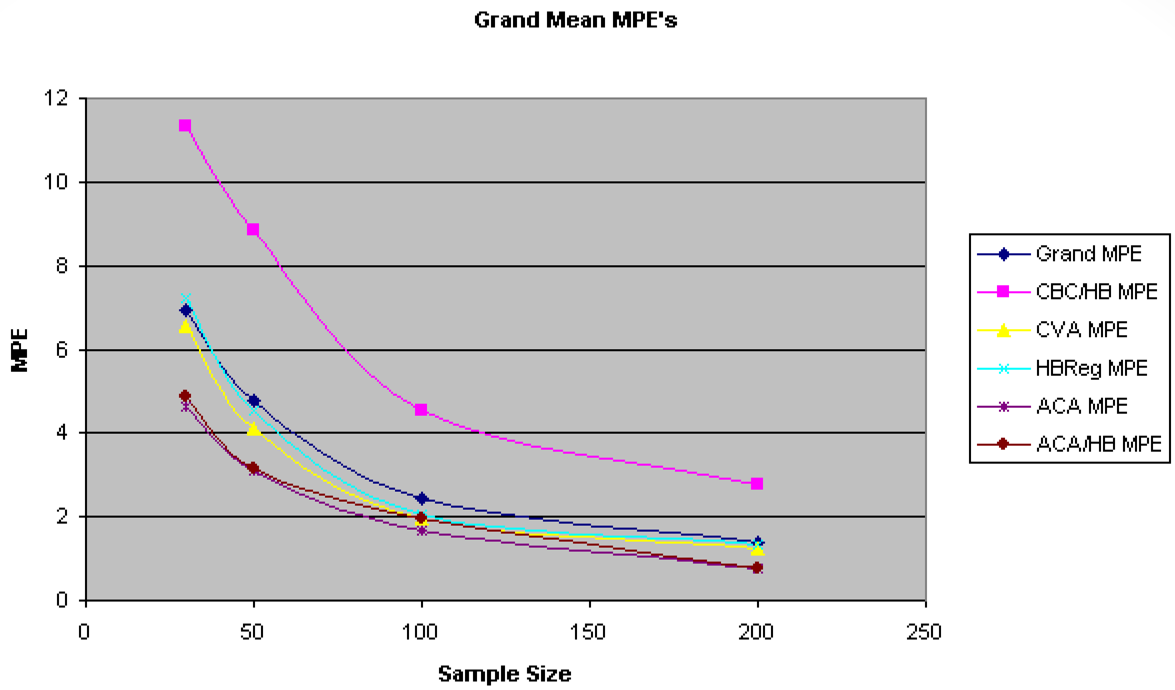


Figure 8.



Using the aggregate MAE and MPE curves as surrogate formulae, tables of error terms as a function of sample size have been constructed below. Given that no elbow exists for these curves, it is left to the researcher, just as it is with proportions curves, to determine the level of error that is acceptable.

There is substantial increase in precision (or decrease in error) when increasing sample from 30 to 50, both for MAE and MPE. There is also substantial increase in precision in terms of both MAE and MPE when increasing sample size from 50 to 75. However, the amount of increased precision may become less relevant to many commercial studies when increasing sample size beyond 75 or 100.

Table 8.

Estimated MAE by Sample Size

Sample Size	MAE
30	5.8
50	4.6
75	3.9
100	3.5
125	3.2
150	3.0
175	2.7
200	2.5

Table 9.

Estimated MPE by Sample Size

Sample Size	MPE
30	6.4
50	3.9
75	2.7
100	2.0
125	1.6
150	1.4
175	1.2
200	1.0

A careful review of Figures 6 and 8 will reveal a pattern of error terms which might suggest that certain trade-off techniques generate lower or higher model error terms than others. This conclusion, at least based on the data presented here, would be false. Each error term is based on total sample utilities computed with a given trade-off technique. Thus, for example, the CVA MPE at a sample size of 100 is determined by taking the CVA-generated mean utilities from the five replicates of the 100 subsample and correlating them with the CVA-generated mean utilities for the total sample. Similarly, for HB-Reg, the subsample mean utilities are correlated with the total sample mean HB-Reg utilities. Even though the underlying data are exactly the same, MPEs for the CVA subsamples are based on one set of "holdouts" (total sample CVA-based utilities) while the MPEs for the HB-Reg subsamples are based on an entirely separate and different set of "holdouts" (total sample HB-Reg-based utilities). Because the reference points for calculating error are not the same, conclusions contrasting the efficiency of the different trade-off techniques cannot be made.

To illustrate how different the total sample models can be, MAEs were calculated comparing the total sample CVA-based models with the total sample HB-Reg-based models for three data sets.

	MAE
Data set 1	7.7
Data set 2	6.5
Data set 3	6.7

These MAEs are larger than most of the MAEs calculated using much smaller sample sizes. Thus, while we cannot compare error terms as calculated here, we can conclude that different trade-off techniques can generate substantially different results.

Having said the above, it is still interesting to note that both the ACA and ACA/HB utilities and models showed remarkable stability at low sample sizes despite the burden of a very large number of parameters to estimate; much larger number of parameters than any of the other techniques.

Latent Class Models

The above analysis is based upon a data set of 1,950 data points, 975 data points for each error term, MAE and MPE. Excluding ACA data, there were 585 data points for each error term.

Latent Class models were run on these data to explore the impact on model error of sample size, number of attributes and levels (expressed as number of parameters) and number of tasks. ACA data were excluded from the latent class modeling because of the fundamentally different nature of ACA to CVA and CBC.

A variety of model forms were explored, beginning with the simplest, such as error regressed against sample size. The models that yielded the best fit were of the form:

MAE = *k*(sqrt(Pc/(na*Tb)))*

and

MPE = *k*Pc/(na*Tb)*

Where P is the number of parameters, n is sample size, T is number of tasks and k, c, a and b are coefficients estimated by the model.

The k coefficient in the MAE model was not significantly different from 1 and therefore effectively drops out of the equation.

For both the MAE and MPE models, latent class regressions were run for solutions with up to 12 classes. In both cases, the two class solution proved to have the optimal BIC number.

Also in both models, sample size (n) and number of tasks (T) were class independent while number of parameters was class dependent. In both models, all three independent variables were highly significant.

It is interesting to note that the most effective covariate attribute was, for the MAE model, trade-off technique (CBC/HB, CVA, HB-Reg). In that model, CBC/HB data points and HB-Reg data points tended to be members of the same class while CVA data points tended to be classified in the other class.

For the MPE model, the most effective covariate was data type (CBC, CVA), which would, by definition group CVA data points and HB-Reg data points together, leaving CBC/HB data points in the other class.

Table 10.

MAE 2-Latent Class Model Output

Latent Variable(s)	**(gamma)**					
	Class1	Class2	Wald	p-value		
Intercept	-0.0395	0.0395	0.0119	0.91		
Covariates	Class1	Class2	Wald	p-value		
Technique						
CBC/HB	0.9122	-0.9122	7.9449	0.019		
CVA	-1.8192	1.8192				
HB-Reg	0.907	-0.907				
Dependent Variable	**(beta)**					
	Class1	Class2	Wald	p-value	Wald(=)	p-value
logAdjVal						
	1.5988	1.2358	905.8524	2.00E-197	3.77E+01	8.30E-10
Predictors						
logn						
	-0.4166	-0.4166	481.2585	1.10E-106	0.00E+00	.
logT						
	-0.2255	-0.2255	41.0711	1.50E-10	0.00E+00	.
logP						
	0.1471	0.3588	62.4217	2.80E-14	14.2287	0.00016
	Class1	Class2				
Class Size	0.5751	0.4249				

Table 11.

MPE 2-Latent Class Model Output

Latent Variable(s)	**(gamma)**					
	Class1	Class2	Wald	p-value		
Intercept	1.0976	-1.0976	4.7383	0.03		
Covariates	Class1	Class2	Wald	p-value		
DataType						
CBC	1.2901	-1.2901	6.6741	0.0098		
CVA	-1.2901	1.2901				
Dependent Variable	**(beta)**					
	Class1	Class2	Wald	p-value	Wald(=)	p-value
logAdjVal						
	0.7849	2.9455	575.3608	1.20E-125	252.1947	8.60E-57
Predictors						
logP						
	2.0587	0.1556	499.858	2.90E-109	186.1446	2.20E-42
logT						
	-0.7422	-0.7422	79.3514	5.20E-19	0	.
logn						
	-0.9422	-0.9422	467.1816	1.30E-103	0	.
	Class1	Class2				
Class Size	0.6005	0.3995				

Conclusions

Minimum sample size must be determined by the individual researcher, just as is the case with simple proportions tests. There is no obvious “elbow” in the error curve which would dictate a natural minimum sample size.

However, using the aggregate error tables as a guide, sample sizes of approximately 75 to 100 appear to be sufficient to provide reasonably accurate models. Larger sample sizes do not provide a substantial improvement in model error. If fact, sample sizes as low as 30 provided larger but not unreasonable error terms, suggesting that, in some instances, small sample sizes may be appropriate.

These data do not suggest that sample size needs to be larger for any trade-off technique relative to the others. Specifically, HB methods do not appear to require greater sample size than traditional methods.

In addition to sample size, both the number of tasks and the number of parameters being estimated play a significant role in the size of model error. An obvious conclusion from this finding is that when circumstances dictate the use of small sample sizes, the negative effects on model precision can be somewhat offset by either increasing the number of tasks and/or decreasing the number of parameters estimated.

These results appear consistent for both error terms calculated for this study: MAE and MPE.

Discussion

There are many aspects of this study which could be improved in future research. The inclusion of more data points would provide better estimates of the shape of the error curve. More replicates at lower samples sizes would provide more stability. MSE (Mean Squared Error) could be included as an additional error term that may prove to be more sensitive than MAE.

The most serious limitation to this paper is the absence of objective standards, that is, holdout cards. Ideally, holdout cards and also attributes and levels would be identical across trade-off techniques. This would require custom designed studies for the purpose of sample size research. An alternative to funding fieldwork for a non-commercial study would be to construct synthetic data sets based on the means and covariances of existing, commercial data sets. If the synthetic data sets were constructed, the sample bias problem would be eliminated, a variety of sample sizes could be independently drawn and attribute collinearities, which commonly exist in commercial data sets, would be maintained.

There are other factors that may affect model error. The number of tasks may have a non-linear relationship to model error. Increasing the number of tasks increases the amount of information available to estimate the model. Excessive number of tasks, however, may increase respondent fatigue to the point of offsetting the theoretical gain in information. Many aspects of measurement error, such as method of data collection (online vs telephone vs mall intercept), use of physical or visual exhibits, interview length, level of respondent interest, etc. may all play a role in model error that could affect the ultimate decision regarding sample size.

The ultimate question that remains unanswered is, what is the mathematics behind model error? If a formula could be developed, as exists for proportions, researchers could input various study parameters, such as number of tasks, number of parameters, sample size, etc. and chart the error term by sample size. They could then make an informed decision, weighing both the technical and managerial aspects, and select the sample size most appropriate for that situation.

References

Johnson, Rich (1996), “Getting the Most From CBC – Part 1,” Sawtooth Software, Inc., Sequim, WA.

Johnson, Richard M. and Bryan K. Orme (1996), “How Many Questions Should You Ask In Choice-Based Conjoint Studies?” 1996 Advanced Research Techniques Forum Proceedings, American Marketing Association.

Moskowitz, Howard (2000), “Stability of the Mean Utilities In Hybrid Conjoint Measurement,” 2000 Advanced Research Techniques Forum Poster Presentation, American Marketing Association.

Orme, Bryan (1998), “Sample Size Issues for Conjoint Analysis Studies,” Sawtooth Software, Inc., Sequim, WA.

The author wishes to thank Rich Johnson for his invaluable suggestions and guidance during the preparation of this paper. The author also thanks POPULUS and The Analytic Helpline, Inc. for generously sharing commercial data sets used in the analysis.

2001 Sawtooth Software Conference Proceedings, September 2001, Victoria, BC, CANADA.

MACRO
CONSULTING, INC.

We are an independent ***marketing research consulting firm*** *dedicated to helping you make the most informed, insightful marketing decisions possible. We specialize in technology, consumer, and new product research, and are well recognized for our* ***State-of-the-Art Research*** *techniques.*

Ultimately, we provide more than just technical expertise. We focus on developing ***pragmatic solutions*** *that will have a positive impact on the profitability of our clients.*

CONTACT US:

Telephone: 650-823-3042

General Inquiries:
info@macroinc.com

Advanced Analysis Inquiries:
analysis@macroinc.com

richard@macroinc.com

www.macroinc.com