THE CASE FOR INLIERS

Jeffry N. Savitz, University of Manchester, SavitzConsulting, LLC., USA, jsavitz@savitzresearch.com

ABSTRACT

The Central Limit Theorem is at the foundation of inferential statistical analysis. Advanced over almost three hundred years ago, it dictates the margins of error associated with estimates of many population parameters from random samples. This paper is about Inliers, subsamples of a random sample that are more reliable than the random sample itself and their use in making equally reliable and accurate estimates of population parameters with far fewer data points than required by a random sample. Empirically it was found that over 60% of a random sample consists of Inliers. Their variance is two-thirds that of a random sample, and they estimate population parameters, the average rating of 25 different popular brands within an average of 2.3% of the same values given by a random sample with almost no bias within one dozen key demographic and sixteen psychographic segments. Thus, it is believed Inliers can be used to make equally accurate and reliable estimates of population parameters for virtually any subpopulation. Based on statistics provided by the Council of American Survey Research Organizations on the annual cost of sample acquisition in survey research, it is estimated that the use of Inliers in place of random samples can save the research industry as much as $333M dollars annually or more in the U.S. alone.

Central Limit Theorem, Margins of Error, Individual Variance, Average Variance, Inlier Index, Inliers

INTRODUCTION

Although random samples are used extensively in research, they are not that common in life, business or government. From consumers picking out apples at the supermarket to managers staffing a company to voters at the polls, most people do not select random samples when making choices; instead, they choose the best candidates for the job. Yet random samples pervade research primarily because researchers understand the math behind them. Laplace (1733) advanced an early version of the Central Limit Theorem which was not proved rigorously until the 1930's by Lévy. With this theorem researchers know that, with 95% confidence, all random sample percentage estimates will be within +/- 9.8% of their true values with a random sample of 100; within 4.9% of their true values with a random sample of 400 and within 3.1% of their true values with a random sample size of 1,000.

However, why can't we find samples with lower margins of error than the above for the same sample sizes, the "best candidates" for the job? In fact, we can. And, just as unusually extreme sample points are called "Outliers," this paper will call the sample points which are "closer" to the population parameters "Inliers."

This paper is about Inliers, respondents who have lower margins of error than a random sample of the same sample size. More importantly, an Inlier sample can accurately and reliably be used in survey research with the same margins of error as a random sample but with a far smaller sample size.

The paper covers the following areas: how to find Inliers in a random sample; how many of them are needed in place of a random sample to achieve the same margins of error; how accurate and reliable they are in estimating population parameters and which psychographic and demographic groups have more Inliers than average and how many more they have. Most importantly, the paper speaks as to how Inliers can be used to reduce the cost of the sampling component of all survey research involving random samples because fewer of them are needed to achieve the same margins of error as realized in a random sample.

The original plan called for the development of models that predict who Inliers are based on demographic and psychographic profiles. Knowing who the Inliers are, researchers could then purchase "pure" Inlier samples instead of random samples and at a much lower cost since only two-thirds as many of them are needed to achieve the same margins of error as a random sample. Indeed, the President of the Council of American Survey Research Organizations (CASRO), Diane Bowers, estimates that roughly $1.0B is spent on sample acquisition annually. As such, by using inliers, the savings to the research industry could be as much as $333M or more on an annual basis. Beyond the substantial savings in cost, based on the analysis below, it is believed that the use of Inlier samples has a wide set of applications both in and outside the survey research field.

BACKGROUND

This research divides data points into three classes as follows: Inliers (data points whose values are very close to the average data point), Outliers (data points far from the average data point) and Midliers (data points in between). Most of the literature on the subject of Inliers relates to methods for differentiating between Outliers as this paper defines them and Inliers which are simply defined as all other data points. Lee, Yu, (2014) talk about the value in using both Inliers and Outliers in object tracking. Hawkins et al. (2002), develop a scoring system measuring the "outlyingness" of data using neural networks. There are a few papers that define Inliers as data points within the interior of a dataset which contaminate the dataset just as much as Outliers (Winkler, 1998). However, there is a paucity of papers which speak to the advantageous use of Inliers as this paper has defined them, to replace random samples but with smaller sample sizes, and without sacrificing precision or accuracy. It appears the theory of the use of Inliers as defined herein has yet to be developed.

RESEARCH METHODS

ANALYTICAL PLAN

Seven hundred nineteen (719) usable interviews were conducted with a random sample drawn from the Toluna online panel, asking participants to rate 25 popular brands. The sample was divided into two groups of 360 and 359, where each group was almost perfectly balanced against U.S. demographics for gender and age. One group was to be used as a training sample to develop theories and the other as a testing sample.

The plan was to derive an Inlier Index for the 360 respondents from the training sample, an index which measures how close the respondents' ratings are to the average ratings across all 25 brands. A model would then be built predicting the Inlier Index from respondent demographics and psychographics using multivariate analysis. A second model was planned differentiating between the Inliers and the Midliers using other multivariate techniques.

These models were then going to be used in conjunction with the demographics and psychographics of the testing sample to determine if these variables could be used to reliably and accurately predict the Inlier Index and discriminate between Inliers and Midliers in the testing sample.

THE INTERVIEW

After screening for gender and age, each respondent was asked to rate the 25 popular brands on a scale of one-to-five, poor to excellent. Brands were grouped by category to provide participants with a more user-friendly survey. The categories and the brands within the categories were randomized to reduce positioning and fatigue bias.

Next, participants were queried on how much they agreed or disagreed with 103 randomly presented psychographic statements selected from the GfK battery of 400 psychographics (Gfk MRI, 2014). Examples include “I always know broadly how much is in my bank account at any one time,” and, “I'd rather travel by myself or with just a small group of people.”

The interview continued with an extensive battery of personal and family demographics including: level of education, indication of Hispanic origin, race, employment status, employment level, marital status, family size, presence of children, presence of children 0-6, 7-12, 13-18 and total household annual income.

INITIAL COMPUTATIONS – THE FULL SAMPLE

Before dividing the respondents into the two balanced group of 360 and 359, it was decided to explore patterns in the entire sample. The average rating was computed across all 719 respondents for each of the brands. Since the goal is to find respondents whose individual ratings are closer to the average rating for all brands, a measure of this “closeness” was needed. To this end, for each respondent-brand pair, the squared deviation of the individual’s brand rating from the average rating for each brand was computed. Next, for each respondent, the Individual’s Variance was developed as the sum of these squared deviations across all brands for each individual divided by 25 for the number of brands. The Average Variance was also computed across all 719 respondents. Finally, an Inlier Index was derived for each participant which is the ratio of the Average Variance to the Individual’s Variance. This is detailed algebraically below.

Let x_{ij} = the rating of respondent “i” on brand “j” , i=1 to 719, j = 1 to 25

$\overline{x}_j = \sum_i x_{ij}/719$, the average brand rating

$V_i = \sum_j (x_{ij} - \bar{x}_j)^2/25$, the variance in brand ratings relative to the average brand

$\overline{V} = \sum_i (V_i/719)$,the average variance in brand ratings relative to the average brand

$I_i = \overline{V}/V_i$, the Inlier Index for respondent i

THE INLIER INDEX

Borrowing from methods of multivariate analysis, when using principal component analysis (PCA), a component with an eigenvalue more than 1 explains more variance than the average original variable used to develop the components. Thus, those components with eigenvalues above 1 are preferred to the original variables in doing analytics. Similarly, because the Inlier Index for each respondent is inversely proportional to his/her Individual Variance, like an eigenvalue in a PCA, it reflects the amount of variance the respondent explains relative to the Average Variance for the entire random sample.

An individual with an Inlier Index of 1 is “average” relative to the other respondents in terms of the variance he or she explains. Those with Inlier Indices above 1 explain more of the variance in the data compared to the “average” respondent.

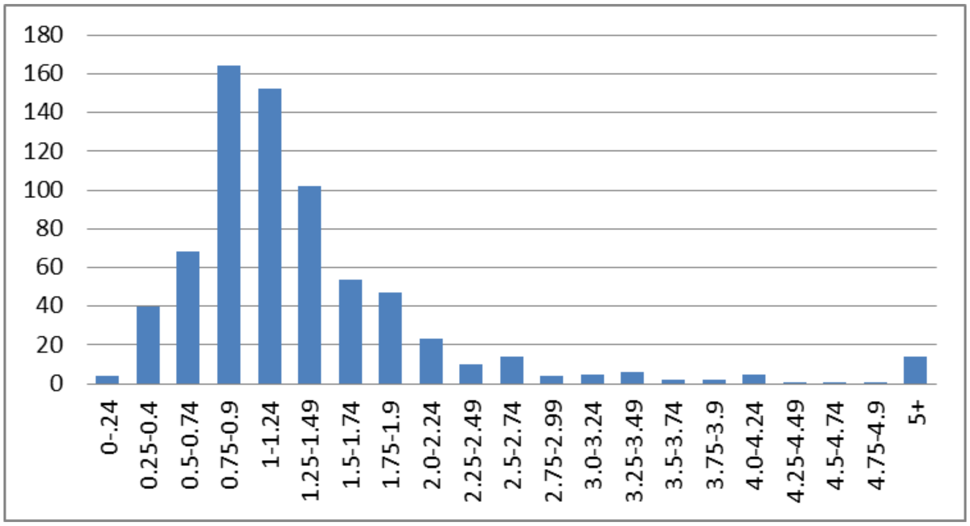
RESULTS AND DISCUSSION

INLIER INDEX REVISTED

The Individual Variances range from 0.03 to 7.50 with an average of 1.00. As such, the corresponding Individual Inlier Indices vary from 0.13 up to 31.31. Only 1.9% of the Inlier Indices are over 5 and 7.7% of the data points over 2.5.

Since the inlier index relates to the average squared difference between the ratings and their averages it follows the Chi squared distribution as illustrated in the diagram below. Using a Chi-squared goodness of fit test we cannot reject the hypothesis that the distribution of inlier indices actually follows the chi squared distribution at the 5% level of significance.

Figure 1 – Distribution of Inlier Indices



The term “Inlier” herein is used to mean those respondents whose Individual Variance is at or above the Average Variance for the entire sample. This means their Inlier Indices are 1 or more. In total 446 out of 719 or 62.0% of the respondents are Inliers and their Inlier indices range from 1.00 to 31.31 with an average of 1.98. The Inlier Indices for the Midliers, the complimentary set to the Inliers have Indices of 0.03 up to 0.99 with an average 0.74.

Importantly, the average variance of the inlier as a group is 0.67 compraed to the average variance for the entire random sample of 1.00. This implies that 1 Inlier is equivalent to 1/0.67 = 1.5 randomly selected respondents. Thus, if researchers could find a pure sample of Inliers for use in a survey instead of a random sample, only two thirds as many respondents would be needed, and at two thirds the cost.

AVERAGE BRAND RATINGS: RANDOM SAMPLE VS. INLIERS

Figure 2 below shows that the “pure” Inlier sample does an excellent job of predicting the average ratings of all 25 brands relative to the Random Sample. Indeed, the average absolute difference between these pairs of ratings is only 0.9 points out of an average Random Sample rating of 3.96 or 2.3%. This ranges from a high for McDonald’s with a difference of 0.16 on an average rating of 3.79 or 4.09% to a low for Lysol Disinfectant with a difference of 0.03 on an average random sample rating of 4.34 or 0.63%.

Figure 2 – Average Ratings Inliers vs Random Sample

Brand	Coke Soft Drinks	Minute Maid	Birdseye Frozen	Wrigley's Gum	Jiffy Peanut Butter	Campbell's Soup	Tide Detergent	Lysol Disinfectant	Levi's Jeans
Avg. Inlier	4.22	4.30	4.22	4.03	4.23	4.34	4.39	4.37	4.34
Avg. Random Sample	4.07	4.21	4.12	3.89	4.14	4.27	4.29	4.34	4.27
Difference	0.15	0.08	0.10	0.14	0.09	0.08	0.09	0.03	0.08
% Difference	3.59	1.99	2.36	3.53	2.14	1.76	2.17	0.63	1.81
Brand	Timex Watches	Ford Cars	Crest Toothpaste	Bayer Aspiring	Viagra Medicine	Verizon Mobile	Apple iPhone	Sony TV	Microsoft Software
Avg. Inlier	4.03	3.93	4.48	4.11	3.24	3.80	3.99	4.13	4.33
Avg. Random Sample	3.95	3.81	4.36	4.05	3.15	3.74	3.90	4.08	4.26
Difference	0.07	0.12	0.11	0.06	0.09	0.06	0.09	0.05	0.06
% Difference	1.85	3.16	2.56	1.52	2.87	1.48	2.25	1.13	1.47
Brand	7-Eleven Convenienc	JCPenny Department	McDonalds	TGI Friday's	Mastercard	American Airlines	New York Yankees		
Avg. Inlier	3.64	3.89	3.94	3.85	4.14	3.69	3.64		
Avg. Random Sample	3.59	3.79	3.79	3.77	4.06	3.64	3.54		
Difference	0.05	0.10	0.16	0.09	0.08	0.05	0.10		
% Difference	1.26	2.58	4.09	2.31	1.90	1.28	2.72		

PSYCHOGRAPHIC ANALYSIS

A Principal Component Analysis (PCA) was conducted of the 103 psychographics which led to 17 components with eigenvalues above 1. Each component was trimmed dropping any variable (psychographic statement) with a loading below 0.5 which led us to 16 components. These components include 79 of the 103 variables and explained 59.4% of the variance in the psychographic data.

The components were then named with the groups as follows: Image Compulsive Snob, Environmentalist, Overly Busy Individual, Family Oriented Traditionalist, Online Purchaser, Monetarily Conscious Individual, Loyal Foodie, Convenience Oriented Mobile User, Politically Engaged Individual, Technologically Adaptive Individual, Ad Resistant Individual, Meticulously Compliant Car Owner, Thrifty Shopper, Clothes Utilitarian, Loner/Semi-Loner and Meticulously Compliant Patient.

A shown in figure 3 below, there was only 1 of the 16 components which showed any significant correlation with the Inlier Index, Monetarily Conscious Individual whose correlation is significant but small (r=0.083). Apparently, to at least a small degree, people with lower awareness of their monetary situation carry a higher Inlier Index. Assuming that people with lower awareness are less cautious with their funds than bigger spenders, this would mean that Inliers as a group may be slightly more likely to spend more money.

Furthermore, in terms of any relationships between the Inlier Indicator and the psychographic groups, there are three statistically significant correlations that surface (all low, r<0.09). Again, to a small degree, Inliers are more likely to be influenced by advertisements, less apt to be thrifty shoppers and more likely to be social. This means that there are essentially no significant differences in mindsets between the Inliers and the full random sample from which they are extracted!

Figure 3 – Principal Component Analysis

Principal Component Description	Eigenvalue	Highest Loading	Correlation with Inlier Index	P-value with Inlier Index	Correlation with Inlier Indicator	P-value with Inlier Indicator
Snobby image-compulsive	20.8	.787	.033	.37	.036	.33
Environmentalist	6.5	.777	.010	.78	.017	.63
Overly Busy	3.6	.639	.049	.19	.019	.62
Traditional Family Orientation	3.4	.702	.041	.27	.052	.16
Online Purchaser	3.2	.663	.048	.12	.025	.51
Monetary Awareness	3.2	.614	.083	.02**	.060	.11
Loyal Foodie	3.2	.591	.047	.21	.011	.77
Mobile Convenience user	2.5	.580	.007	.86	.064	.09
Politically Engaged	2.2	.685	.006	.88	.013	.74
Technologically Adaptive	2.2	.752	.026	.49	.023	.54
Uninfluenced by Ads	2.0	.633	.013	.73	.145	.00**
Meticulously Compliant Car Owner	2.0	.554	.020	.59	.014	.70
Thrifty Shopper	1.8	.630	.039	.30	.075	.05**
Clothes Utilitarian	1.7	.649	.015	.68	.050	.18
Loner/Semi-Loner	1.6	.530	.033	.38	.080	.03**

Meticulously Compliant Patient	1.6	.706	.043	.25	.006	.85

DEMOGRAPHIC ANALYSIS

Similar to the psychographic analysis above, a Chi-squared analysis was completed to determine if there is a relationship between the Inlier Indicator and different demographic groups. Figure 4 below shows the Chi-squared values and the p-values associated with the Chi-squared tests for the various demographic groups. This implies there are virtually no significant differences between Inliers as a group and the full random sample from which they are taken!

No significant findings surface in the figure. Evidently, few if any subgroups stand out as having more than an average percentage of Inliers within a given demographic implying the use of Inliers to estimate population parameters should be equally accurate across most all demographic groups. At least, there is no real evidence to the contrary.

Figure 4 – Inlier Memberships within Different Demographic Groups

Demographic	**Chi-squared Value**	**P-value**
Gender (Males/Females)	1.58	0.208
Age (18-24, 25-34, 35-44, 45-54, 55-64, 65-74)	2.82	0.729
Level of Education (College, No College)	0.322	0.570
Race (White, Black, Asian, American Indian, Other)	3.805	0.433
Hispanic (Yes/No)	0.058	0.810
Marriage (Single never married, single previously married, married)	5.60	0.347
Family Size (1,…)	(t = .498)	0.619
Presence of Kids (Yes/No)	0.04	0.850
Kids (Aged 0-6)	0.013	0.910
Kids (Aged 7-12)	0.032	0.859
Kids (Aged 13-18)	0.236	0.627

Importantly, no psychographic or demographic groups stand out as being particularly inlier dense. As such, the plan to develop psychographic models using a training sample and apply them to a testing sample would not be effective and was abandoned. There is no obvious way to target any subpopulation to identify and acquire a pure sample of inliers and take advantage of their lower variance and consequent lower cost. Additional research will be needed to target Inliers perhaps using a combination of psycho-socioeconomic variables.

CONCLUSIONS AND IMPLICATIONS

Inliers are special subsamples of a random sample that can be used in smaller numbers than random samples but with the same margins of error, because the variance of the inliers is two thirds that of the overlying random sample. Furthermore, because Inliers appear to be fairly uniform across all demographic and psychographic subgroups, there should be no decline in the accuracy in using them in any subgroup analysis.

According to CASRO, companies in the U.S. spend roughly $1.0B on sample acquisition annually. Thus, with a variance of only two thirds that of a random sample, by using samples of pure Inliers, researchers could reduce this cost by as much as 33% leading to a potential savings to the industry of $333M annually.

There are many applications of the use of Inliers with survey as well as other data. In the marketing arena alone, Inliers can be used to reduce sample sizes in applications such as: brand and advertising tracking, advertising testing, image and positioning, packaging and channel research, sales analysis and retail store customer counts, customer engagement and satisfaction, survival analysis, new product testing and prediction models, price elasticity, product line optimization, website design and analytics, social media effectiveness, facial recognition and eye tracking marketing, analysis of churn, RFM (reach, frequency and monetary value of customers) and other database and Big Data applications.

In addition to marketing and advertising applications, there may be other applications for the use of Inliers. For example, if one could use Inliers to analyze the S&P 500 it might be possible to study two thirds of these stocks, the “S&P 333”, and get the same level of accuracy and reliability in portfolio predictions. This might even enable Wall Street to reduce the number of analysts it uses to do portfolio analysis.

In summary, Inliers could mean a real boon to the research industry and business in general, cutting the cost of samples down by a substantial margin. Further research is being conducted to learn more about the details of Inliers and their applications to marketing and business.

However, because there are no individual psychographic or demographic groups that index higher on having inliers than others, the ability to obtain a “pure” inlier sample remains a mystery at this time.

However, using structural equation modeling and/or partial least squares and/or neural networks, we may be able to find other latent variables besides individual demographics and psychographic which identify what makes people Inliers versus Midliers. Armed with this information, researchers may then be able to purchase “pure” or at least highly dense samples of Inliers reducing the cost of sample acquisition by as much as one=third.

REFERENCES

Lee, J., & Wonpil, Y. (2014). Concurrent Tracking of Inliers and Outliers. *ArXiv E-prints*. Retrieved October 5, 2015 from http://arxiv.org/pdf/1409.3913v1

Gfk MRI (2014) *Gfk MRI Psychographic Sourcebook*. Retreived July 20, 2015 from www.gfkmri.com/assets/PDF/GfKMRIPsychographicSourcebook.pdf

Hair, J., Ringle, C., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *The Journal of Marketing Theory and Practice, 19*, 139-152.

Hair, J., Sarstedt, M., Ringle, C., & Mena, J. (2011). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science, 40*(3), 414-433.

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier Detection Using Replicator Neural Networks. *Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science, 2454*, 170-180.

Winkler, W. (1998). Problems with inliers. Retrieved October 5, 2015, from http://www.census.gov/srd/papers/pdf/rr9805.pdf