Sampling eCommerce Data from the Web: Methodological and Practical Issues

Galit Shmueli[1], Wolfgang Jank[1] & Ravi Bapna[2]
[1] Dept. of Decision & Information Technologies
R. H. Smith School of Business
University of Maryland, College Park, MD 20742
[2] Dept. of Operations and Information Management
School of Business
University of Connecticut, Storrs, CT 06269

KEY WORDS: sampling schemes, stratified sampling, cluster sampling, electronic commerce, internet

Abstract:

Empirical research that is based on web-collected data has been rapidly growing thanks to the large amounts of freely available web-data and the technological wonders of web spiders for grabbing data. This is especially true for electronic commerce research, which yields results that can be very influential on the market. Although all studies rely on inferences from the collected data to some population of interest, there has been nearly no attention paid to sampling issues. The methodology of statistical sampling is very relevant in web-data collection. It includes defining observational units and target and sampled populations, determining sources of sampling and non-sampling errors, choosing appropriate sampling designs, and adjusting sample estimators to reduce bias and increase precision. Sampling eCommerce data shares many characteristics with other types of sampling (e.g. surveys), but also has special features that researchers should be aware of and account for. In this paper we discuss web-data, and in particular eCommerce data collection in the context of sampling methodology, and suggest improvements to current practice in this modern sampling setting.

1. Introduction

The seminal question in sampling is "What population does the sample represent", like the title of the paper by [11]. Taking a representative sample from the population of interest is a fundamental of nearly every empirical study. The ability to generalize conclusions from the sample to the population depends on the relationship between the two. The field of sampling deals with the various issues involved in designing and drawing a sample, as well as making inferences from the sample to the population. Classical statistical methods for cross-sectional data e.g., t-tests and regression models, assume that the sample is a "random sample". However, this assumption should not be overlooked since in many situations it is violated. Two examples of violations are a population that is too small to be considered infinite (and thus the probability of sampling one observation affects the sampling probabilities of the other observations), and dependence between observations with regard to the measure of interest. In both cases the sample design must take into account these special features, and the sample statistics are adjusted to be representative of the population.

The wide availability of electronic data on the web has stirred a huge body of empirical research on web content such as electronic commerce. The emergence of web-spiders, which are software programs that are designed to collect data from HTML pages, has greatly enhanced the ability to capture enormous amounts of data almost instantaneously. Although companies are reluctant to share their databases, they sometimes make a large amount of the data freely available through queries. For instance, eBay permits users to browse almost all auctions that have closed in the last 30 days. Similarly, travel websites will reveal airfares for flights in their database through queries.

The issue of sampling, however, appears to be completely disregarded. In the great majority of web-content studies that rely on sampling from the web, the data collection phase focuses on the collection mechanism design and the technical details of how the data will be collected, and ignores other important sampling issues such as defining the population of interest, the goal of the study (exploratory, descriptive, or analytical), and the relative importance and representativeness of sub-populations ([5, p. 47]) What are tolerable error rates? What are the main variables of interest? These, in turn, can affect the sample design directly. For instance, when the goal of the study is exploratory, then it is more important to get broad coverage of the population than reducing sampling error of resulting estimates. However, when the goal is analytical it is more important to assure that the sample will be powerful enough for rejecting null hypotheses.

Another common phenomenon in web-data studies is the inexplicit assumption that the collected data are the entire population of interest or, more commonly, representative of the population of interest. Assuming that the sample consists of the entire population of interest usually stems from the thought that "we sampled all the transactions in the last week". However, in most cases the purpose is to generalize the results to a longer period of time and thus the sampled week is a subset of the population of interest. In fact, in most cases the population of interest includes a longer time frame than that in which the sample was drawn. For instance, [1] used a sample of transactions on eBay that was collected over a period of 3 months to estimate consumer surplus for that entire fiscal year. [4] sampled several million sales of new and used books on Amazon that cover 180 days between 9/2002-3/2003 and 105 days between 4-7/2004. They use this sample to estimate the rate that used books cannibalize on sales of new books (15%), and use it to infer about this rate in the "population" which does not include a time frame. Another example is a study of online price dispersion by [9]. They collected three samples of prices on eight product categories taken in Nov 2000, Nov 2001, and February 2003. They use the three samples to represent three eras of eCommerce: the "blooming, shakeout, and restructuring of eBusiness" ([10]), and the 8 categories to represent the continuum of the global market (based on the price range that they span). In short, a set of collected web-data, no matter how large, is most likely to be partial to the population of interest even if it includes all observations within a limited time period. A recent advancement in web-agent design addresses the issue of long term data collection ([6]). Using this technology in conjunction with good pre-sampling definitions and choices can lead to improved samples.

Finally, at the analysis stage, there must be consideration of the sample design in determining if and how estimates should be adjusted to compensate for bias and reduce sampling error. To our knowledge, empirical studies of eCommerce data have completely overlooked these post-sampling issues.

2. Sampling and Non-Sampling Errors

The first steps of planning a sampling scheme involve answering the following questions:

1. What are the observation units of interest? What are the units that we will actually sample?
2. What is the target population which we want to study and what is the actual population from which the sample will be taken?

Answers to these questions are crucial for obtaining correct inferences and are directly related to non-sampling errors. Non-sampling errors result from design issues that are unrelated to the fact that a sample is taken rather than the entire population. There are two types of non-sampling errors: selection bias and measurement bias.

When the target population differs from the population to be sampled this will result in selection bias. This includes under-coverage of the target population, non-response in surveys, and misspecification of the target population. Such errors are just as likely to occur in web-sampled data, but their format can be different. Here are some examples:

- Server problems and internet congestion during "web rush hour" can lead to many unrecorded transactions (such as bids placed by users).
- A website's policy of data refreshing can lead to discrepancies between the target and sampled population: if the data presented on the website are cashed (e.g., Google.com, price comparison engines) then the sampled population might be outdated.
- When using a websites "search" function to retrieve data of interest, choosing the first set of returned records is likely to produce bias in the sample. Search engines do not randomize the results, but rather display them according to some order (e.g., relevance or recency).

It is therefore important to carefully specify the target and sampled populations and compare them before the sample is collected. Skipping this at an early stage can cause complications later on. An example is the study by [12] that investigates price dispersion across internet retailers. The collected data on 22,209 price quotes revealed heterogeneity due to the product condition (which previously has not been accounted for). The researchers then reduced their sample by retaining prices only on products tagged "new", thereby narrowing down the target population. The eliminated data were not random, and in practice a set of retailers were completely eliminated, thereby changing the sample design.

Measurement bias results from the measurement tool: these include interviewer/experimenter effects, non-truthful responses, errors in data recording, and poorly designed questions in questionnaires. Although these types of errors seem less prevalent in web-data, they do exist if in somewhat different format:

- Web-agents that collect data can interfere with the website's traffic causing a slowing down or other extraordinary effects
- Fictitious users or non-truthful postings by users still affect the collected data (e.g. bid shilling)
- Poorly designed websites can lead users to input wrong or even irrelevant information. If the error goes unrecognized, then the sampled population contains errors. If they are recognized, the website might drop those records altogether or record them with missing/wrong values.
- Devices to thwart robots, not allowing collection from some area or limiting the query frequency to, say 5-times-a-day.

Sampling errors result from taking a sample rather than recording the entire population. Although a good sample can provide an accurate representation of the entire population, there are still inaccuracies resulting from including some of the observations and not others. The goal of statistics is to quantify sampling error and thereby allow inference from a sample to the population of interest. Sampling error consists of two main measures: bias and variability. Optimally we prefer unbiased sample estimators that have very low variability. This means that the estimator is accurate, or "on target", and that it produces precise estimates that do not vary much from sample to sample. To construct such estimators, however, requires knowledge of the underlying sampling mechanism. It is not sufficient that web-data samples tend to be very large and that the global population size is huge. It turns out that for different sampling methods the estimators need to be constructed differently to achieve low bias and variability.

Returning to the issue of observation units, determining whether these are transactions, products, users, retailers, or any other entity is directly related to the sampling design. Depending on the website, the data will usually be organized in a way that is most relevant to potential users. However, many of these websites have a relative database where data are linked to multiple observation units. For example, the data on eBay are organized by auction. However, it is possible to collect data on a particular set of bidders. On B&N.com, if we look at book prices, then the natural unit is a book within a certain category. However it is easy to focus on authors as the observation of unit, and collect prices on their books.

Although there might be large amounts of data on the "natural observation unit" of the website, there might be much scarcer data on other units. Returning to the B&N.com example, there might be an abundance of romance contemporary fiction books listed on a certain day. However, if we look at authors who published multiple statistics textbooks, there are likely to be fewer observations. Furthermore, if we are interested in comparing subgroups of this population where a certain subgroup constitutes a very small proportion of the population, then it is likely that this subgroup will not be represented in a random sample.

In the following we give a short description of the main probability sampling schemes, highlighting issues that are relevant to web-sampling.

3. Simple Random Samples

Probability samples, as opposed to non-probability sample such as convenience sample, are characterized by the feature that each unit in the population has a known, non-zero probability of selection into the sample. The current literature that relies on web data collection generally attempts to take probability samples, but completely ignores sampling issues and considers the data collected as a simple random sample (SRS) without replacement from the target population. An SRS is a sample where each observation has an equal probability of being selected from the population. It is the simplest probability sampling method. The hidden SRS assumption in empirical studies of eCommerce can be seen by the type of analyses that have been performed: t-tests, linear regression models, and parameter estimation, which all rely on sample statistics such as the sample mean, a proportion, etc.

Although SRSs are generally easy to analyze, in web-data sampling they are not always the right design to use. One main reason is that employing an SRS requires identifying and labeling all the records that could potentially be sampled prior to the sampling. In many web-data studies it is hard or impossible to construct such a list. This issue is described in further detail below.

3.1 Constructing the Sampling Frame

A sampling frame is the list of all the observations from which the sample will be drawn, and is in fact the operational definition of the target population. Examples of common sampling frames are phone directories, email lists, and household addresses. A listing on the web is usually labeled using a sequence of digits such as an auction ID, an ISBN, or a product number. To construct a sampling frame we would need to identify each possible listing in our target population and its label before carrying out the sampling. In some cases a list of these labels is possible to obtain or construct. An example is using price comparison engines such as MySimon.com and Shopper.com. [9] compared the completeness of lists of retailers from multiple price comparison websites, and found that BizRate.com returned the most complete list (except for a few cases where not all retailers were covered). However, we performed a search for a popular DVD player on BizRate.com and found that it excludes a major retailer called Amazon.com! Thus, relying on such comparison sites for constructing an exhaustive sampling frame should be carried out cautiously.

In many other cases, especially when the list is potentially very large, enumerating the listings or obtaining their generating mechanism are not disclosed by the website, and therefore it is impractical or impossible to assemble a sampling frame. If querying the website is possible, in many cases the results will only be a subset of the list, and most likely not a random one. Yet in other cases, it is impossible to search the website directly by the observation unit (e.g. obtaining a list of all open auction IDs on eBay.com). Of course, if the company is willing to share its data, then this problem goes away. However, companies tend to be wary of sharing such data with researchers.

Another reason for the absence of a sampling frame in web-data applications is that the target and sampled populations are in many cases assumed to be ongoing (or at least a finite but unknown time interval into the future), and therefore N is not a fixed number at the time of sampling.

If a list does not exist then probability sampling will yield biased results with no information on the bias itself.

3.2 Further SRS Issues

Let us assume that the goal is to estimate a population mean (e.g., the average price of a certain camera model sold online on a certain week) or a population proportion p (e.g., the proportion of transacted auctions on eBay in the last 30 days). If we take an SRS of size n from a very large population, then the sample mean is unbiased for estimating and is optimal in the sense of low variability. Similarly, in a sample of binary valued observations, the sample proportion of 1's, , is unbiased for estimating the population proportion of 1's and is optimal in the sense of low variability.

However, if the population size N is not much larger than the sample, then we must apply the finite population correction (FPC) to obtain unbiased estimators. Thus, even if the sample is very large, the sample mean will still be inaccurate for estimating the population mean. The FPC is given by the ratio (N-n)/(N-1). Although the population size in many web-data applications is very large, in some cases we are actually interested in a subpopulation that is small (see Section 5 on stratified sampling). Furthermore, since the sample sizes in web-collected data tend to be very large, if they approach the population size then the FPC can have a large impact on the results.

Another deviation from the simple estimators given above occurs when the SRS is substituted with an unequal probability sampling scheme: When we are interested in a particular subpopulation that is rarer in the population, it makes more sense to set unequal sampling probabilities and over-sample this particular group. An example is the different rates of sampling used by [4]: they sampled books that sell on Amazon.com at low quantities at 2 hour intervals, whereas books that sell at high quantity were sampled every 6 hours. In such cases unbiased estimators for the population mean and/or proportion require the use of a weighted sample mean/proportion such that observations have weights that are inversely proportional to their selection probabilities. For example, if member of some sub-population are twice as likely to be sampled, then their weights in the estimate must be reduced by ([5, p. 25]).

4. Systematic Sampling

One way to avoid the need for creating a sampling frame before the sampling is carried out is to use systematic sampling. The idea is to choose a random starting point and then to sample every i^th records to achieve a required sample size (i = N/n).

Thus, the sampling frame is constructed "on-the-fly". The main danger in systematic sampling is that the records are ordered in a cyclical pattern which could coincide with the sampling cycle. In a web-data collection environment such a scheme is less appealing, since the cost of sampling the remaining records (between sampled records) is negligent. Furthermore, the list that is compiled at the end of the sample might be temporary, if the website recycles labels. An example is auction IDs in eBay and UPS tracking numbers.

Alternative sampling designs that do not require an apriori sampling frame are stratified and cluster sampling.

5. Stratified Sampling

Stratified sampling means that the population of interest is divided into mutually-exclusive subpopulations and random samples are taken independently from each subpopulation. This is a very popular and useful sampling method which yields more precise estimates compared to a simple random sample, by incorporating some external information about the quantity of interest. It turns out that for web data-collection this method can actually serve a completely different important purpose.

Many procedures for collecting web data rely on querying the website of interest and recording the records that are returned from the query. In most cases, the algorithm that the website uses for ordering the results and the amount of results returned is not disclosed. For example, when querying eBay for all closed auctions on a certain date, it returns only a subset of those auctions! How this subset is chosen and how the records in this subset differ from those that are not returned is hard to assess. The reason for returning a subset is most likely the huge amounts of records that match the query. Using the returned subset could therefore result in selection bias.

A solution to this problem is to use stratified sampling with strata reflecting categories. Since in this scheme queries are category-specific, it greatly reduces the number of resulting records, which should now yield a complete category-level list (at that time point - see the section on Cluster sampling for time-frame issues). On the internet stratification is very natural because of the hierarchical structure of many websites and the relevance of these categories to the research questions. Although website hierarchies are designed for ease of navigation, they can be exploited for the purpose of stratified sampling.

Some examples of hierarchically designed websites are

- eBay's categories (electronics, tickets, collectibles, etc.) and three additional subcategory levels
- Yahoo! Movies
- Barns & Noble's categories (books, DVD&Video, Music) and multiple levels of subcategories (Books> fiction> romance > contemporary romance>. . .)
- Drugstore.com (vitamins & diet> sports nutrition> performance supplements>. . .)

Studies that try to estimate a global mean or rate in many cases tend to use the hierarchical structure even without considering stratified sampling (e.g., [4] sample book sales on Amazon.com by book categories.) If the categories that are used for stratifying are also of interest in themselves (e.g., we want to compare records across these categories, obtain category-level estimates, or treat them differently), then using stratified sampling is advantageous for another important reason: the resulting sample estimates are typically more precise than those from an equivalent simple random sample (SRS). In fact, this is the main reason that stratification is used in survey samples. One challenge in stratified sampling is to determine the sizes of the sub-populations. In some cases estimates based on proxy information can be used. In web-based data, sub-population sizes are usually unknown and tend to change over time. Since many empirical studies use a sample that is a snapshot in time to make inferences about longer time periods, the actual category sizes are unknown until the end of the period of interest. In such cases, if the variability in category sizes is not expected to change drastically from time to time, these "snapshot" sizes can be used as approximations.

A main assumption in stratified sampling is that records are divided into distinct subpopulations and that there is no overlap between strata. However, categories and sub-categories on the web tend not to be mutually exclusive. For instance, the book XXX can be listed in the subcategory Y and subcategory Z. This duplication is more likely to occur at deeper sub-category levels, where the distinction between products/records is smaller. Therefore, the level of hierarchy used for stratification should be the highest possible which creates strata of "manageable" size (i.e., all records are returned by a query of that level).

5.1 Estimating a Mean or Rate from a Stratified Sample

A stratified sample is based on taking an SRS from each stratum. Thus, the stratum-level estimates are unbiased for estimating stratum-level parameters. To obtain unbiased overall population estimates we use a weighted average of the stratum-level estimates, where the weights are proportional to stratum size.

5.2 Sample Size

The main guidelines for determining how many records to sample within each stratum depend on the goal of the analysis: If the goal is to estimate an overall parameter (mean, proportion, etc), it is advantageous to sample more heavily in strata that are (1) bigger, (2) have more variance, and (3) are cheap to sample. However, if the goal is to compare strata, then these subpopulations should be over-sampled (in proportion to their variances). In eCommerce, within-category variances are sometimes the object of study themselves: price dispersion within different categories of products is compared ([9, 12]). In such studies the strata sizes are important in determining whether over-sampling is required, and for adjusting the estimates.

In web-sampling the cost of sampling is usually negligent. However, some categories of records can be harder to sample in the sense that they require specialized programming or even manual intervention per record. An example is adult-only merchandize auctions on eBay, which require extra password logins. Such "sampling costs" can be integrated into the sampling scheme such that smaller samples are taken from these categories (for the purpose of estimating a global parameter).

6. Cluster Sampling

Cluster sampling is a widely used method aimed at reducing the cost of sampling. The basic idea is to divide the population of interest into groups whose members are "close", and in that sense sampling all members is easier, cheaper, or faster. A random set of clusters is then selected and all cluster members are drawn into the sample. An example is survey that randomly selects households, and then surveys each of the household members.

Cluster sampling occurs naturally in web-collection data when the target population consists of a longer time interval than the time frame in which the sample is collected. eCommerce data typically change faster than offline data. Therefore inference from a temporal snapshot to a longer time frame is not always feasible without taking longitudinal sampling. Such a setting can be seen as a cluster sampling design, where the clusters are the different time points of which a random sample is taken. Then, at each of the selected time points either the entire set of observations is recorded, or a sample of them is taken (see section on multistage sampling). Even if we assume that during the population time frame there is no change in the measured phenomena, cluster sampling can emerge in another format: concurrency. An SRS assumes that the sampling units are independent. When sampling eCommerce records there exists an effect of concurrency that in most cases affects the relationship between records. For instance, a set of auctions for the same item that have some time overlap are most likely to interact and affect the price in each of the auctions. Similarly, a set of movie ratings for similar movies is most likely to be dependent for two reasons: the same person might be rating multiple movies, and in addition ratings are affected by comparisons between movies showing at the same time. From a sampling point of view, we might treat such subsets of concurrent records as clusters. Cluster sampling is typically used to simplify and reduce the cost of sampling when the cost of sampling an additional observation within a cluster is very low, but it usually has the price of reduced precision of the estimates. A typical example is the sampling of households when the unit of interest is actually a single person. The household is then selected at random, and all its inhabitants are surveyed. Like systematic sampling, cluster sampling does not require an exhaustive list of the observational units. Instead it requires a list of the clusters.

The reduction in precision which usually accompanies cluster sampling stems from a high intra-cluster correlation: records within a cluster tend to be more similar than records across clusters (e.g., education level within the same household). To reduce this effect, the number of clusters to be sampled should be large, and the cluster sizes small. In a web-context this means that if the goal is to get a representative picture of the entire population of records (auctions, movies, etc) we want to make sure that we have a large enough number of records that do not overlap in time. In the opposite case where the intra-cluster correlation is very low, we prefer to sample a small set of very large samples.

The key point is that internet-sampling actually employs cluster sampling, and therefore the sampling should actively be designed and recognized as such. One implication is that clusters (i.e. time-intervals) should be identified before data collection and then a cluster-scheme sampling should be employed accordingly. The simplest scheme is a one-stage cluster sampling, where a random set of clusters is chosen and all records in those clusters are recorded. This procedure guarantees that each record has an equal chance of being selected (self-weighting). In an online auction context this would mean, for example, specifying a set of time-intervals where auctions within an interval are most likely to interact (e.g., auctions for a specific DVD model that took place between Dec. 1-7, 2004).

6.1 Estimating a Mean or Proportion from a Cluster Sample

When the goal is to estimate a population mean or proportion, an unbiased estimate from the cluster sample is based on averaging the cluster-based means/proportions using a weighting scheme that produces self-weighting. The weights are selected as to achieve required probabilities of selection. For instance, setting the weights proportional to the relative cluster size yields equal probabilities of selection for each record.

In both cases, the standard error of the estimate differs in proportion to the intra-cluster correlation. When clusters are of average size M and have intra-cluster correlation (defined as the correlation between each pair of elements within the same cluster, and ranges between [-1/(M-1) , 0]), the variance of a cluster-estimate is larger than that of an SRS-estimate by a factor of 1+(M-1) This implies that the sample size required from cluster sampling is usually much larger than that from a simple random sample to achieve the same accuracy. The only exception is when items within a cluster are negatively correlated (and thus the cluster sample estimates are more accurate than those from an SRS.)

7. Multistage Sampling

This type of sampling is a complex form of cluster sampling which does not require a sampling frame, can reduce the cost of sampling all the items within a cluster, and is useful when the population is so large that it is easier to sample items by going through a hierarchical selection process. Multistage sampling applies cluster sampling hierarchically with a final random sampling stage. A simple two-stage design first chooses a random set of clusters and then within each cluster takes a random sample of observations. A three-stage design would mean randomly choosing a set of clusters, and then within each cluster randomly choose a set of sub-clusters, and finally take an SRS within each chose sub-cluster. Many eCommerce studies that look at changes over time in fact employ a two-stage sampling design: the target population covers a time interval of which a sample of time points is taken. Within each time point a sample is collected (e.g., the price dispersion study by [9], which collects data on Nov 2000, Nov, 2001, and February 2003, and in those times samples a set of products (which are representative of a larger population of products).

Advanced surveys can have even 6 or more stages. In addition, they might pre-select some clusters instead of drawing a random sample of clusters. Preselection is also usually common in eCommerce data collection, where a set of categories is preselected and they are to represent a wider population of categories. As long as the selection at the next stages is random the bias can be adjusted for. Determining the probability of selection in multistage samples and estimating sampling variability is much more complicated. The two popular methods are linearization and resampling, which can be executed through software packages such as SAS.

8. Further Issues and Conclusions

Web-data and their collection open the door to exciting new research. However, issues related to the sampling schemes that are employed should be carefully evaluated before performing the actual sampling. Designing a good sampling scheme is necessary for reducing both sampling and non-sampling errors. Data grabbers should carefully consider what the target population is, what they are actually sampling, and where the two do not overlap. Populations and sub-populations should be spelled out and used for creating improved sampling schemes and inference. Although simple random sampling currently dominates the field, its application is flawed due to the lack of sampling frames and therefore inferences might be incorrect with relation to the population of interest. In order to apply SRS in a valid way, methods for constructing sampling frames must be developed.

The hierarchical nature of websites and the relevance of these categories to many research problems lead us to favor stratified sampling for many applications. On the other hand, the concurrency of web records, which is completely ignored in the SRS setting, can be approached from a cluster sampling approach. In short, although simplicity is a blessing, there are several details that must be acknowledges and accounted for in order to reach valid generalizations from web-collected data. Finally, if a complex sampling design is used then resampling methods can be employed to quantify sampling error.

Another challenging issue is the dynamic nature of eCommerce (and other) websites. Aside from the technical burden of adjusting the web-spiders to changes in the website, it also means that the definition of target population should either be restricted in time according to the times when the website changed, or more global, but taking into account this new source of heterogeneity.

References

[1] Bapna, R., Jank, W., and Shmueli, G. (2004), "Consumer Surplus in Online Auctions", *Working Paper*, University of Connecticut.

[2] Chevalier, J.A. and Mayzlin, D. (2003), "The effect of word of mouth on sales: online book reviews", *Yale SOM Working Paper* No's. ES-28 & MK-15.

[3] Dellarocas, C. N. (2001), Analyzing the Economic Efficiency of eBay-like Online Reputation Reporting Mechanisms, *MIT Sloan Working Paper*, No. 4181-01.

[4] Ghose, A., Smith, M. D. and Telang, R. (2004), "Internet Exchanges for Used Books: An Empirical Analysis of Welfare Implications and Policy Issues", Working paper, http://ssrn.com/abstract=584401.

[5] Henry, G. T. (1990) Practical Sampling, SAGE Publications, Applied Social Research Methods, vol 21.

[6] Kauffman, R. J., March, S. M., and Wood, C. (2000), "Mapping Out Design Aspects for Data-Collecting Agents", International Journal of Intelligent Systems in Accounting, Finance, and Management, vol 9, no 4, pp. 217-236.

[7] Levy, P. S. and Lemeshow, S. (1999), Sampling of Populations, Methods and Applications, Wiley-Interscience, 3rd edition.

[8] Lohr, S. L. (1999), Sampling: Design and Analysis, Duxbury Press.

[9] Pan, X., Ratchford, B. T., and Shankar, V. (2003), "The Evolution of Price Dispersion in Internet Retail Markets", Advances in Applied Microeconomics, volume 12, pp. 85-105.

[10] Pan, X., Ratchford, B. T., and Shankar, V. (2004), "Price Dispersion on the Internet: A Review and Directions for Future Research," Journal of Interactive Marketing, Forthcoming.

[11] Rao, C.R. (1985) "Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent?", In *A Celebration of Statistics: The ISI Centenary Volume*, A.C. Atkinson and S.E. Fienberg, eds., Springer-Verlag, pp. 543-569

[12] Venkatanesan, R., Mehta, K., and Bapna, R. (2004), "Do market characteristics impact the relationship between Internet retailer characteristics and prices?"