RC22987 (W0311-080) November 12, 2003
Mathematics

IBM Research Report

Using Surveys to Understand the Present and Predict the Future

Aliza Heching, Ying Tat Leung
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Using Surveys to Understand the Present and Predict the Future

Aliza Heching [*] Ying Tat Leung [†]

November 3, 2003

Abstract

Organizations often conduct surveys to gather information. This information allows the organization to serve its customers more efficiently and to aid in long-term decision-making. We have developed an instrument for using survey responses to analyze current population behavior and to predict future behavior. This instrument is coded using Base SAS, SAS STAT, SAS GRAPH, as well as SAS Macros. Our SAS code uses the new SAS survey statistics procedures (e.g., PROC SURVEYMEANS). More specifically, we have developed the following tools: We generate point and interval estimates of population parameters based upon sample responses. We address methods for improving the robustness of these estimates: we define procedures to identify statistical outliers and we have developed heuristics to modify the weights used in the (weighted) analysis to improve robustness of the survey results. We develop several methods for measuring the impact of different data points on the estimated parameters to help the user verify the analysis results. Finally, we develop methodology that facilitates the use of survey responses to predict future responses.

Key words: survey statistics;

[*] Aliza Heching, IBM T.J. Watson Research Center, Route 134, Yorktown Heights, NY 10598. Ph: (914) 945-3191 Fax: (914) 945-4527 email: ahechi@us.ibm.com

[†] Ying Tat Leung, IBM T.J. Watson Research Center, Route 134, Yorktown Heights, NY 10598.

1 Introduction and Summary

Surveys are often conducted to gather information. The information gathered can be used to understand behaviors or beliefs under a set of conditions or to predict future behaviors or responses. For example, the role of many individuals and organizations is to provide a service to customers. These individuals and organizations may conduct surveys to learn about the underlying conditions that make their services valuable and desirable to others. Better information will allow them to serve their customers more efficiently and effectively and aid in long-term high-level decision making. Further, surveys can be used to capture the public's response to promotional messages sent out by businesses, agencies, governments and institutions. Surveys may also be conducted to test hypotheses and validate or advance theoretical knowledge.

Generally, it is difficult and costly to survey every member of a given population (i.e., to conduct a census). Therefore, those conducting surveys will usually select a subset of the population, a representative sample, from which inferences about the entire population are drawn.

Sample design includes two fundamental elements: (i) a selection process or sampling methodology, which dictates the rules by which members of the population are included in the sample and (ii) an estimation process for computing the statistics of the selected sample that are sample estimates of population values.

A selection process should yield a sample that represents the elements of the population. There are two major categories of sampling methodologies. Model based sampling is sampling based upon broad assumptions about the distribution of the survey variables in the population. Probability sampling assumes that every element of the population has a known non-zero probability of being selected. In general, probability sampling is preferred over model sampling. First, model sampling requires assumptions regarding randomization of the population, while probability sampling bypasses this assumption by introducing randomization into the selection procedures. Second, whereas the results of probability sampling allow for inferences about the population to be made entirely by statistical methods, model sampling depends heavily upon the validity of broad assumptions about the distributions of the survey variables in the population.

The estimation process involves computing statistics from the sample responses that are valid for the entire population. Different statistics are computed depending on whether one is attempting to make a statement about the entire population or about some subset of the population. Further, one must correct for bias and non-sampling errors in the sample data.

Our work, to date, focuses on the second element of sample design, assuming that probability sampling has been used to select the sample elements. More specifically, we assume that the selection process and sample size have been determined. We focus on the elements of analysis that are required once data has been collected and any coding and preprocessing has been completed. We have developed SAS programs that use the survey responses data files as input and perform statistical analysis on any variables of interest. We have developed an instrument for analyzing the survey responses, which includes the following features:

We generate point estimates of population parameters, and confidence bounds for those estimates. We compute an estimate of the population mean, and the variance of this estimate, based upon the sample responses. Some factors that we consider in these calculations include what stratification scheme used when the sample was created. We also consider any weights associated with each respondent, and the impact of those weights on the population parameter estimates. The formulas are adjusted depending upon whether an estimate for the entire population is desired, or just a subpopulation. Finally, we consider the sampling scheme that was used to select the respondents.

Next, we address methods for improving the robustness of the parameter estimates. Here, we consider two different factors that may impact the robustness of these estimates. First, we address the problem of outlier detection and elimination, where an outlier is any observation that falls beyond three standard deviations of the mean for its stratum. Such a definition implicitly assumes that strata were accurately determined to group together respondents with similar characteristics. Second, we consider the impact of weights on the population parameter estimates. More specifically, we address the fact that weights assigned to each respondent are typically estimated and thus not exact. Consequently, we do not want the weights to unduly influence the values of the parameter estimates. We developed a heuristic that analyzes the sample data and weights associated with each data point, and identifies responses that have a significant impact upon the weighted estimate of the population mean. We then adjust these weights, within a user specified range, to limit the sensitivity of the estimate of the mean to the weights assigned to each response.

Finally, we conduct trend analysis on the survey responses and estimates computed from the sample data. Here, our objective is to use survey responses not only as a tool for identifying current population behaviors, but to detect trends in population behaviors so that one can forecast future behaviors. We have considered two different methodologies for trend detection. One approach is appropriate for situations where very few data points are available for trend estimation, and looks for differences in parameter estimates over time. A second approach, in cases where a larger number of data points is available, is to conduct a regression against time. In most cases, parameter estimates are constrained to lie within a given range of values. We have developed a very effective methodology for incorporating these constraints into our trend analysis.

The remainder of this paper is organized as follows. In section 2 we provide a general discussion of various sampling techniques used to collect a representative sample for analysis. Section 3 briefly discusses methods for storing response data. In section 4 we provide detailed analysis of how one uses sample data to estimate population parameters. Section 5 discusses why one may introduce weights into the analysis and how the analysis changes when weights are introduced. In section 6 we introduce the concept of domains and discuss how estimation formulas are revised when one is only analyzing a domain of the population. Section 7 looks at different factors that may influence the accuracy of estimated values. We discuss methods for detecting and eliminating statistical outliers from the sample data collected. We also suggest a heuristic to improve the robustness of weighted estimates.

In section 8 we discuss methods of using survey responses to detect trends in population behaviors and to predict future population behaviors. Finally, we conclude with section 9 where we briefly describe the application of our methodology to a survey administered by IBM's Server Group.

2 Sampling Techniques

Polling organizations will generally survey a subset (i.e., a representative sampling) of the entire population. Inferences about the beliefs or behaviors of the population are then drawn based upon responses from the subset. Sampling may be conducted using either non-probability sampling or probability sampling.

In non-probability sampling, not all members of the population necessarily have a positive probability of being included in the sample. In contrast, probability sampling is characterized by all members of the population having a known positive probability of being included in the sample.

Non-probability sampling techniques include convenience sampling (select a sample based upon availability for the study), most similar/dissimilar cases sampling (select cases that are judged to be similar or dissimilar, depending upon researcher's objective), typical cases (select cases that are know a priori to be useful and representative), critical cases (select cases that are essential for inclusion in and acceptance of the study), snowball sampling (existing members of the sample identify additional cases), and quota (interviewer selects sample that yields same proportions as population proportions on some specific variables).

While in the case of probability sampling all members of the population have a known positive probability of being included in the sample, these probabilities may not be the same for all members of the population. If all members of the population have the same probability of being selected, this sample design is called an "equal probability sampling." The five most common techniques for probability sampling include:

- Simple Random Sampling: All members of the population have equal probability of being selected. (In this case, if the population size is N and the sample size is n, then a member of the population has probability n/N of being selected as an element of the sample.)

- Systematic Sampling: Each member of the population is listed, a random start is designated, and then members of the population are selected at equal intervals.

- Stratified Sampling: Each member of the population is assigned to a stratum. Simple random sampling is used to select within each stratum. This method of sampling is often used as a means for reducing the variance of the population parameter estimates. One stratifies, for example, so that respondents with similar characteristics are classified into the same stratum.

- Cluster Sampling: Each member of the population is assigned to a cluster. Clusters are selected at random and then all members of the cluster are included in the sample.

- Multi-Stage Sampling: Clusters are selected as in the cluster sampling. Then, sample members are selected from within each cluster, using simple random sampling.

These different methods for probability sampling are used in different situations. For example, cluster sampling is often used if the population is segmented into natural clusters (e.g., schools or households), and stratification may be used to decrease the variances of the sample estimates. Probability sampling is preferred over non-probability sampling in that its results are more valid and credible. On the other hand, it often takes longer and requires more effort to gather a probability sample.

By using probability sampling, one can compute the probability that a given member of the population is included in the sample. We refer to this probability as the “inclusion probability” for that member of the population.

3 Collecting and Storing Sample Responses

Once a sample is identified, survey responses from this subset of the population must be collected. Survey responses may be collected in various ways, including interview studies such as telephone or personal interviews, self administered surveys such as mail surveys or handed/picked-up/dropped off questionnaires, and electronic surveys such as direct entry into a computer, modem, e-mail, fax, disk, or Internet. Another approach may involve collecting "responses" from some records on file (e.g., sample of size of homes in given region, by looking at county records, to make general statements about size of homes in that region).

Once collected, the responses, or data, is preferably stored in a database. The database can either be a fixed-field type (data for each variable in same field location for each respondent) or a free-field type (data for each variable in the same order for each respondent; delimiters, e.g., comma or blank, separate one variable from the next). Preferably, one field or group of fields is used that uniquely identifies each respondent.

4 Estimation

Survey responses are used to make estimates about the beliefs or behaviors of the entire population. By collecting responses from a subset of the population, we want to make statements about beliefs or behaviors of the entire population. We generate point estimates of one or more population parameters. Point estimates estimate a specific value of the population parameter. The methodology used to extrapolate from the sample responses to the entire population differ, depending upon the sampling technique used.

Because simple random sampling can be viewed as a special case of stratified sampling with only a single stratum, we will only discuss results for stratified sampling. Suppose a company is using a survey to estimate the average spending budget for its customers. Let:

π_k =inclusion probability for respondent k, $k = 1, \ldots, n$

y_k =respondent k's spending budget, $k = 1, \ldots, n$

Also, suppose that respondents were stratified into H strata. Let N_h be the size of stratum h, $h = 1, \ldots, H$ and n_h be the size of the sample in stratum h, $h = 1, \ldots, H$. Then a point estimate of the average spending budget for all $N = N_1 + N_2 + \cdots + N_H$ customers, $\tilde{y}_{ST}$, can be derived using the results in Section 5.6 of Sarndal et al. [8] as:

$$\tilde{y}_{ST} = \frac{\sum_{h=1}^{H} \sum_{k \in h} \frac{y_k}{\pi_k}}{\sum_{h=1}^{H} \sum_{k \in h} \frac{1}{\pi_k}} \tag{1}$$

(We use the notation $k \in h$ to denote summing over all the members of the sample that lie in stratum h.) In the special case of simple random sampling in each stratum, the population mean estimator simplifies to:

$$\tilde{y}_{STSI} = \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{\sum_{k \in h} y_k}{n_h} \right). \tag{2}$$

Intuitively, the population mean is equal to a weighted sum of the H strata means (the term inside parenthesis in (2)). More specifically, the sample mean is obtained separately and independently for each stratum and is then multiplied by the weight of the stratum. These products are summed over the H strata to obtain the weighted sample mean. The weight of the stratum is generally the proportion of the population contained in that stratum; in equation (2) this weight is given by $\frac{N_h}{N}$.

Often, the true size of the population is unknown. In this case, we replace N and N_h in equation (2) with their estimators $\hat{N} = \sum_{k=1}^{n} \frac{1}{\pi_k}$ and $\hat{N}_h = \sum_{i \in h} \frac{1}{\pi_i}$, respectively.

However, this point estimate for the population mean is based upon the respondents in sample s only. If a different sample is selected, say s_1, a different point estimate of the population mean will most likely result. Further, most certainly neither of these point estimates is actually the true value of the population mean. Thus, to obtain a more accurate measure of the population mean, a confidence interval is constructed as follows.

First, the variance of the estimator is measured. Let Θ denote the population mean, and $\hat{\Theta}$ denote the estimate of the population mean. We use the statistic of a sample s to approximate $\hat{\Theta}$. Let $\tilde{y}$ be the sample s estimate of $\hat{\Theta}$. The variance of the estimator is defined as $\mathsf{E}(\tilde{y}_s - \mathsf{E}(\hat{\Theta}))^2$, where E is the expected value function. Thus, the variance is a measure of how the point estimate for each sample differs from the expected point estimate based upon all samples. More specifically, let S denote the set of all possible samples. Then,

$$\mathsf{E}(\hat{\Theta}) = \sum_{s \in S} p(s)\tilde{y}_s,$$

where $p(s)$ is the probability that sample s is selected from the set of all possible samples in S, and

$$v(\hat{\Theta}) = \sum_{s \in S} p(s)\{\tilde{y}_s - \mathsf{E}(\hat{\Theta})\}^2. \tag{3}$$

Of course, we cannot compute $v(\hat{\Theta})$ based upon a single sample, as (3) assumes knowledge $\tilde{y}_s$ for all samples $s \in S$. Instead we must estimate $v(\hat{\Theta})$ similar to the way that we estimate $\hat{\Theta}$ by $\tilde{y}_s$.

Let:

$\hat{N}$ $= \sum_{k \in s}(\frac{1}{\pi_k})$, where $\hat{N}$ is an estimate of the size of the population,

Δ_{kl} $= \pi_{kl} - \pi_k \pi_l$. (where π_{kl} is the probability that both k and l are included in the sample),

$\hat{\Delta}_{kl}$ $= \frac{\Delta_{kl}}{\pi_{kl}}$,

and denote the estimate of $v(\hat{\Theta})$ by $\hat{v}(\tilde{y}_{STSI}, \pi)$, in the case of stratified sampling with simple random sampling within each stratum.

Based on results in Sections 2.8 and 5.6 of Sarndal et al. [8], we compute $\hat{v}(\tilde{y}_{STSI}, \pi)$ as follows:

$$\begin{aligned} \hat{v}(\tilde{y}_{STSI}, \pi) &= \frac{1}{\hat{N}^2} \sum_{h=1}^{H} (\sum_{k \in h} \sum_{l \in h} (\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l})(\frac{y_k - \tilde{y}_s}{\pi_k})(\frac{y_l - \tilde{y}_s}{\pi_l})) & (4) \\ &= (\frac{1}{\hat{N}})^2 \sum_{h=1}^{H} (\sum_{k \in h} \sum_{l \in h} \ddot{\Delta}_{kl} \ddot{y}_k \ddot{y}_l), & (5) \end{aligned}$$

where

H =number of strata

y_k =y-value for k^{th} observation in sample

$\ddot{y}_k = \frac{y_k - \tilde{y}_s}{\pi_k}$

$$\ddot{\Delta}_{kl} = \begin{cases} 1 - \frac{\pi_k \pi_l}{\pi_{kl}} = \frac{-1+\frac{n_h}{N_h}}{n_h - 1} & \text{if } k \neq l \\ 1 - \pi_k = 1 - \frac{n+h}{N_h} & \text{if } k = l \end{cases}$$

$\hat{N} = \sum_{k \in s} \frac{1}{\pi_k}$

Note that for stratum h, $\pi_k = \frac{n_h}{N_h}$ and $\pi_{kl} = \frac{n_h(n_h-1)}{N_h(N_h-1)}$.

Having computed an estimator of the variance of the estimate of the population mean, we construct a confidence interval on the true population mean, Θ, in the traditional manner:

$$\tilde{y}_{STSI} - z_{1-\frac{\alpha}{2}} \hat{v}(\tilde{y}_{STSI}, \pi)^{\frac{1}{2}} \leq \Theta \leq \tilde{y}_{STSI} + z_{1-\frac{\alpha}{2}} \hat{v}(\tilde{y}_{STSI}, \pi)^{\frac{1}{2}} \tag{6}$$

where $z_{1-\frac{\alpha}{2}}$ is the constant exceeded with probability $\alpha/2$ by the $N(0,1)$ distribution and $\tilde{y}_{STSI}$ and $\hat{v}(\tilde{y}_{STSI}, \pi)$ are the estimated mean and estimated variance of the stratified sample estimator of the population mean, respectively.

5 Weighted Analysis

Often, survey responses are weighted. The weights are computed so as to assign greater "importance" to responses of certain respondents with given characteristics. The weights are in addition to and separate from the sampling probabilities. Typically, the weights are values estimated by the individual analyzing the survey, or by an executive who is interested in the survey results. For example, a company may weight the responses of its customers according to each customer's relative size (e.g., the number of employees) or according to each customer's prior purchase volume. The company may wish to place greater emphasis, for example, on customers who traditionally have larger purchase volume. The weights are then incorporated in the estimation formulas. For the sake of brevity, we shall not provide formulas here; the formulas for weighted analysis are a special case of the formulas for weighted domain analysis (described in the following section), where the domain is the entire population.

6 Domain Analysis

One may consider whether an estimate of the entire population's behavior is desired, or just a sub-population. For example, a company that desires to survey its customers to gain insight into what customer demand will be for different products offered by the company, may only wish to survey the subset of its customer population that intends to make purchases in the near future. The customers who intend to make purchases in the near future represent a subset, or a domain, of the entire population of customers. A domain is a subset of the population for which separate estimates are planned in the sample design. All estimation formulas must be revised to reflect the fact that only a domain of the total population is being analyzed.

For stratified sampling, the weighted point estimate for the domain mean and its variance can be derived (albeit, after significant algebraic manipulation) using results in Sarndal [8] Sections 5.7 and 10.3, and are given as follows:

$$\tilde{y}_w = \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in h_d} w_k y_k}{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in h_d} w_k} \tag{7}$$

where

w_k =weight assigned to respondent k, $k = 1, \ldots, n$

N_h =size of population in stratum h, $h = 1, \ldots, H$

n_h =sample size for stratum h, $h = 1, \ldots, H$

y_k =value of response for respondent k, $k = 1, \ldots, n$

h_d =subset of stratum h that is in the domain, $h = 1, \ldots, H$

Similarly, the formula for the variance of this estimator must be revised as follows:

$$\hat{v}(\tilde{y}_w) = \frac{\sum_{h=1}^{H} \frac{N_h^2}{n_h^2} \sum_{k \in h_d} \sum_{l \in h_d} \ddot{\Delta}_{kl} \; w_k w_l (y_k - \tilde{y}_w)(y_l - \tilde{y}_w)}{[\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in h_d} w_k]^2}. \tag{8}$$

7 Improving Robustness of Estimated Parameters

In this section we address methods for improving the robustness of population parameters estimated using sample responses, with respect to two different factors that may impact the accuracy of estimates. We refer to the first factor as statistical outliers. By statistical outliers, we mean observations that fall “statistically outside” of the remaining observations in the sample. The second factor that we consider is the impact of the weight assigned to each observation on the overall parameter estimate. More specifically, we address the fact that assigned weights are typically estimated and thus inexact. Consequently, the weights should not disproportionately influence the value of the parameter estimates.

7.1 Eliminating Statistical Outliers

We first address the problem of identifying and eliminating statistical outliers. By “outliers” we mean observations that are statistically significantly different from the remaining observed values. To illustrate, if 100 members of a population are sampled, we will check if, for example, observed value for member 1 is an outlier by comparing its observed value to the remaining 99 observed values. In the case that stratified sampling was used to obtain the sample responses, one would search for outliers separately within each stratum. The reason for taking this approach is that populations are generally stratified so that observations within any stratum are similar.

In performing the search for statistical outliers, it is difficult to determine with certainty whether or not an observation is an outlier. Statisticians have devised several methods for detecting outliers. We adopt one method as follows (Dixon [3]). First, some assumption is made about the distribution of the data from which the sample was collected. Then, one will quantify how far the suspected outlier is from the other values. This may be computed in different ways, including computing the difference between the value of the suspected outlier and the mean of all points, the difference between the value of the suspected outlier and the mean of the remaining values, or the difference between the value of the suspected outlier and the next closest value.

Next, this distance value is standardized by dividing by some measure of scatter, such as the standard deviation of all values, the standard deviation of the remaining values, or the range of the data (i.e., maximum observed value minus minimum observed value).

Finally, the probability associated with the answer to the following question is computed: If all the observed values were sampled from the assumed distribution for the population, what is the probability of randomly obtaining an outlier so far from the other values? If the probability is small, it is concluded that the deviation of the outlier from the other values is statistically significant.

If non-stratified sampling is employed, we search for statistical outliers once for the entire sample. Different methods may be used to detect statistical outliers. The choice of method depends upon the number of observations in the sample (or stratum). The following rules may be followed:

(i) If the sample size is smaller than a first predetermined number this sample is deemed to be too small to conduct a valid search for outliers. We suggest setting this predetermined number equal to 4.

(ii) If the sample size is between the first predetermined number and a second predetermined number, the so-called Dixon-type test may be used to detect outliers. See Dixong [3] or Barnett and Lewis [1] We suggest setting this second predetermined number equal to 10.

(iii) If the sample size is greater than the second predetermined number, a generalized Extreme Studentized Deviate (g-ESD) procedure may be sued to identify the outliers.

See Rosner [7].

7.1.1 Eliminating Outliers When Sample Size is Less Than Ten

In applying the Dixon-type test, observations are first ordered in increasing order of value. Then, a search is performed for (i) a single outlier on the right, (ii) a single outlier on the left, (iii) two outliers on the right and (iv) two outliers on the left. (The terms “right” and “left” mean the following: If all observations are ordered in increasing order, then the observations with lowest value will be on the left side, and the values of the observations will increase as one goes from left to right. So an outlier on the left means to test if the observation with the lowest value is an outlier; an outlier on the right means to test if the observation with the highest value is an outlier.)

In performing the search, the Dixon-type test requires some guess to be made of the number and location (right or left) of suspected outliers in the data. One option is to first use a boxplot technique to locate suspected outliers and then use the Dixon-type test to test these outliers. Several boxplot techniques are known. The standard boxplot (see e.g., Newbold [6]) has a higher chance of detecting false outliers than typical tests. (A “false outlier” means that the test determines that the observation is an outlier when, in fact, it is not.) A known variation of the standard boxplot slightly modifies the standard test to ensure that a random normal sample has a prespecified probability of containing no outliers. (For description of this modified boxplot rule see Hoaglin and Iglewicz [4].) Either type of boxplot technique may be used to identify the location of potential outliers, and then a Dixon-type test may be used to test if these observations are, in fact, outliers.

7.1.2 The Boxplot Rule

The boxplot rule works as follows. First, order the observations in increasing order of magnitude and label these ordered observations $x_1,\ x_2,\ \ldots, x_{n_h}$. Second, define the lower quartile Q_1 as $Q_1 = x_{[f]}$, where the f^{th} observation is defined as $f = \frac{\lfloor \frac{n+1}{2} \rfloor + 1}{2}$ where $\lfloor \rfloor$

denotes the "floor," i.e., the largest integer whose value is less than or equal to $\frac{n+1}{2}$, where n is the number of observations. If f involves a fraction, Q_1 is the average of $x_{[f]}$ and $x_{[f+1]}$. Third, define the upper quartile, Q_3, in a similar manner. Specifically, count f observations from the top. Thus, $Q_3 = x_{(n+1-f)}$. Fourth, define the inter-quartile range as $R_f = Q_3 - Q_1$. Finally, potential outliers are those observations that either (a) lie above $Q_3 + 1.5R_f$ or (b) lie below $Q_1 - 1.5R_f$. Once the locations of the potential outliers have been identified, the Dixon-type test may be applied to determine if these observations are, in fact, outliers.

7.1.3 The Dixon-Type Test

The Dixon-type test works as follows.

- Order all observations in increasing order and them $x_1,\ x_2,\ \ldots, x_{n_h}$.
- Specify a significance level α_i.
- $x_{(n)}$ is an outlier on the right if

  $$r_{11} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}} > \lambda_{11}, \tag{9}$$

  where values for λ_{11} are computed using, e.g., the table in Dixon [3] for the specified value of α_i and value of n. For an outlier on the left, the negative is taken of all observations and the Dixon-type test is performed as above.
- For two outliers on the right, $x_{(n)}$ and $x_{(n-1)}$ are outliers if

  $$r_{21} = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}} > \lambda_{21}, \tag{10}$$

  where values for λ_{21} are computed using, e.g., the table in Dixon [3] for the specified value of α_i and value of n. For two outliers on the left, the negative is taken of all observations and the Dixon-type test is performed as above.

7.1.4 Eliminating Outliers When Sample Size is Greater than Ten

When the sample size is greater than ten, a generalized ESD ("g-ESD") procedure may be used to identify the outliers. g-ESD includes the following steps. (See Rosner [7].)

- Specify m, the maximum potential number of outliers.
- Specify α_i, the significance level.
- Compute
$$R_1 = \max_{i \leq n}(\frac{|x_i - \overline{x}|}{s}), \tag{11}$$
where $s^2 = (\sum(x_i - \overline{x})^2)/(n-1)$. Find and remove the observation that maximizes this term.
- Compute R_2 in the same way, except that for R_2 the sample size $n-1$ is used by removing the observation identified by R_1.
- Compute $R_3, \ldots, R_m$ in a similar manner (with sample sizes $n-2$, $n-3, \ldots, n-m+1$).
- Assume λ_i is known (below we discuss how values for λ_i are determined), outlier identification works as follows. If for all i $R_i \leq \lambda_i$, then no outliers exist. If for some i $R_i > \lambda_i$ then let $l = \max\{i : R_i > \lambda_i\}$ and declare $x^{(0)},\ x^{(1)}, \ldots, x^{(l-1)}$ as outliers. (Here, $x^{(0)},\ x^{(1)}, \ldots, x^{(l-1)}$ are the observations selected in iterations 1 through l of this algorithm.)

The parameter λ_i may be determined as follows.

(i) The table in Rosner [7] lists values for λ_i for different values of α_i and n.

(ii) For combinations of α_i and n not in the table, the following formula can be used:
$$\lambda_i = \frac{t_{n-i-1,p}(n-i)}{\sqrt{(n-i-1+t^2_{n-i-1,p})(n-i+1)}}, \tag{12}$$
where $i \in [i, m]$, $t_{\nu,p}$ is $100p$ percentage points from the t distribution with ν degrees of freedom, and $p = 1 - [\frac{\alpha}{\alpha(n-i+1)}]$ and parameters m and α.

In summary, the steps of generalized ESD are:

(1) Specify m, the maximum number of possible outliers.

(2) Compute $R - 1,\ R_2, \ldots, R_m$ where $R_i = \max_{s \in S}\{\frac{(x_i - \overline{x})}{s_i}\}$, where s_i is the standard deviation of observations considered when computing R_i.

(3) Compute $\lambda_1, \lambda_2, \ldots, \lambda_m$.

(4) Determine the maximum $i = 1, \ldots, m$ for which $R_i > \lambda_i$. Let $k = \max\{i : R_i > \lambda_i\}$. Observations $X_1, \ldots, x_i$ are outliers.

7.2 Improving Robustness of Weighted Estimates

As discussed in section 5 sample responses are often weighted so that some responses have a greater influence on the value of the final estimator than others. These weights often represent some estimated value rather than an exact known quantity. In our case, the weights in some cases represented an estimate of the buying power of the customer. It was not clear in all cases that the value of the assumed weight was exact or that it was accurately computed. In many cases, the weights represent educated guesses by an analyst and should only be taken as approximations to the relative orders of magnitude. When such uncertainty exists surrounding the accuracy or appropriateness of the weights themselves, the weights can negatively impact the accuracy of the overall population parameter estimates. We describe here a heuristic for reducing or eliminating the potential inaccuracies introduced by these estimated weights on the population parameter estimates. In this heuristic we analyze the sample response data and the weights associated with each response. We then identify responses that have a significant impact on the weighted estimate of the population parameter. The corresponding weights are then adjusted, within a user prespecified range, to ensure that the response no longer has such a significant impact on the overall estimate. The goal is to have the population parameter estimates reflect the weights, but the exact combined value of any single response and its associated weight should not overly influence the final parameter estimate.

To perform the heuristic, we first define some concepts.

First, we define mathematically what we mean by "overly influence the final parameter estimate" by developing a measure of the impact of any weighted observation on the overall parameter estimate as follows. Suppose that we are computing an estimate of the population mean. We define the impact of observation k, D_k, as:

$$D_k = |\frac{\overline{y} - \overline{y}_{-k}}{\overline{y}}|, \tag{13}$$

where $\overline{y}$ is the estimated domain mean (or population mean, a special case of domain mean where the domain is the entire population) based upon the current set of weights and observed values, and $\overline{y}_{-k}$ is the corresponding mean estimate based upon the current set of weights and observed values excluding observation k. In other words, we consider the case if we were to choose one fewer respondent (observation k) from that stratum.

Second, specify a tolerance level for the impact of any observation on the final parameter estimate. We denote this tolerance level by T.

We say that weight w_k, the weight for observation k, satisfies the *Tolerance Condition* if $D_k < T$. Note that because any single weight affects the value of the mean, changing the weight assigned to observation k will change D_j for $j \neq k$ and therefore may cause some other observation to violate its tolerance limit. It is possible that no set of positive weights exists that simultaneously satisfies the Tolerance Conditions for all observations. Similarly, it is possible that there are multiple sets of positive weights that simultaneously satisfy all Tolerance Conditions. In the latter case, we select for each observation a weight that is as close to the original weight (assigned by the analyst) as possible (not precluding the possibility of using the original weight itself).

The heuristic works as follows:

Step 0: The Tolerance Condition is given. Define a maximum number of iterations, m, after which point the algorithm terminates. Let i denote the counter which counts the current iteration. Set $i = 1$.

Step 1: Mark all observations as feasible.

Step 2: For each feasible observation, check whether the Tolerance Condition is satisfied. If the Tolerance Condition is satisfied for all observations goto Step 6. Else, goto Step 3.

Step 3: Find the observation that violates the Tolerance Condition and for which D is largest. Denote this observation by k_0. Compute a revised weight $w\prime_{k_0}$ such that

$$D_{k_0} = T.$$

Thus, we are revising the weight for observation k_0 such that the Tolerance Condition is exactly satisfied.

Step 4: Step 3 may yield more than one revised weight. (Below we provide an example where this is the case.) If all revised weights are nonpositive, retain the current weight for observation k_0 and mark observation k_0 as infeasible. Otherwise, select the revised positive weight whose value is closest to the value of the original weight, w_{k_0}, and mark all observations as feasible.

Step 5: Increment counter i by 1. If $i = m$ or if all observations are marked infeasible, goto Step 6. Else, goto Step 2.

Step 6: Exit heuristic.

We now compute the revised weight $w\prime_{k_0}$ for the case where the sample was selected using stratified sampling and an estimate of the domain mean is being computed. First, we assume that each stratum contains more than one member and that observation k_0 is in the domain of interest. We are interested in adjusting the weights of the observations in the domain. In this case,

$$\overline{y} = \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in h} w_k y_k}{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in h} w_k},$$

where

H = number of strata

N_h = population size of population in stratum h

n_h = size of sample in stratum h

y_k = value of variable for observation k

w_k = weight assigned to y_k

and

$$\overline{y}_{-k_0} = \frac{\sum_{h=1,h\neq h_0}^{H}[\frac{N_h}{n_h}\sum_{k\in h_d} w_k y_k] + \frac{N_{h_0}}{(n_{h_0}-1)}\sum_{k\in h_{0d},k\neq k_0} w_k y_k}{\sum_{h=1,h\neq h_0}^{H}[\frac{N_h}{n_h}\sum_{k\in h_d} w_k] + \frac{N_{h_0}}{(n_{h_0}-1)}\sum_{k\in h_{0d},k\neq k_0} w_k}$$

where h_0 denotes the stratum containing observation k_0. The revised weight for observation k_0, $w\prime_{k_0}$, is computed to satisfy the Tolerance Condition, i.e.,

$$\frac{\overline{y} - \overline{y}_{-k_0}}{\overline{y}} = T,$$

resulting in the following:

$$w\prime_{k_0} = \frac{C \cdot (\sum_{h\neq h_0}(\frac{N_h}{n_h}\sum_{k\in h_d} w_k) + C \cdot \frac{N_{h_0}}{n_{h_0}}\sum_{k\in h_d,k\neq k_0} w_k - \sum_{h\neq h_0}(\frac{N_h}{n_h}\sum_{k\in h_d} w_k y_k) - \frac{N_{h_0}}{n_{h_0}}\sum_{k\in h_d,k\neq k_0} w_k y_k}{\frac{N_{h_0}}{n_{h_0}}(y_{k_0} - C)}$$

where $C = \frac{(1\pm T)y_{-k_0}}{1-T^2}$. Thus, two values arise for $w\prime_{k_0}$, one by selecting the "+" in the formula for C and one by selecting the "−" in the formula for C. Denote the former by $w_{k_0}^{+}$ and the latter by $w_{k_0}^{-}$. Next, we must determine which of these values, if any, will be used to replace the existing weight w_{k_0}. Three situations may arise: (i) Both $w_{k_0}^{+}$ and $w_{k_0}^{-}$ are positive. In this case select the revised weight whose value is closest to w_{k_0}; (ii) Only one revised weight, $w_{k_0}^{+}$ or $w_{k_0}^{-}$ is positive. In this case select the revised weight with the positive value to replace w_{k_0}; (iii) Both revised weights are negative. In this case retain the current weight w_{k_0}.

If stratum h_0 contains only one chosen sample, we can develop a similar set of formulae to replace those described. (This situation is, however, unlikely in practice.) As well, an analogous set of formulae can be developed for the case where observation k_0 is not in the domain in question. The latter case is generally less interesting in practice.

When observation k_0 is not in the domain of interest, removing k_0 from the sample still affects the value of $\overline{y}_{-k_0}$ through the value of n_{h_0}. An analogous formula for $w\prime_{k_0}$ can be derived.

We note that the user may wish to consider different values for T (e.g., increase T from $\underline{T}$ to $\overline{T}$ in some specified increments) and then chart (i) impact of the value of T on the revised estimate of the population parameter and (ii) the number of weights that are revised by the heuristic. The user can then select a value for $\underline{T} \leq T \leq \overline{T}$ with which he feels most comfortable.

8 Trend Analysis

Often, one is interested in studying the estimated population means over time, to detect the presence of some trend in these values. In this case, it is the objective to use the survey responses not only as a tool for identifying current beliefs and behaviors of the population, but also to detect trends in population behaviors so that future beliefs, behaviors, actions, and/or attitudes of the population may be predicted.

The steps used to perform this trend analysis are as follows. Here, we use computed estimates of the population parameter (e.g., mean) to compute the trends. We perform trend analysis by looking at responses to a single question over time. Thus, if a survey is administered multiple times, and the same question is repeated each time the survey is administered, one can perform trend analysis on the response to this question. The general methodology involves first computing the estimated mean each time the survey was administered (as discussed in Section 4) and then testing for trends in these means using the methodology outlined below. This trend analysis may be performed where there are at least three data points available (i.e., the same questioned was repeated in three different survey instruments). This trend analysis involves conducting a (weighted) regression over time. This regression analysis may then be used to *forecast* future estimated parameter values, thus forming the basis of predicting population behavior. We propose using weighted regression with the reciprocal of the standard deviation of the mean estimator as weights to take in account the relative precision of the different survey repetitions.

Sometimes, parameter estimates are constrained to lie within a given range of values. For example, if the question of interest was "What percentage of your overall spending budget will be allocated to hardware purchases?" the response is clearly constrained to lie within the range of $[0, 100]$. Thus, one must ensure that the forecasted value lies within the allowable range as well. In more complex situations, we may have multiple questions for which we wish to forecast population parameters, and the *sum* of these population parameters are constrained to lie within some given range. For example, a company may ask a customer to specify the percentage of its overall spending budget that will be allocated to each of a list of expenditure classes. Clearly, the sum of these responses (for each respondent) is constrained to equal 100. Again, we have developed methodology to allow for such constrained forecasting. The methodology is described in the steps below.

Step 1: For each repetition of the survey, compute (as above) the estimated domain mean, weighted or unweighted, and the variance of this estimator. Eliminate outliers and execute the robustness improvement algorithm as described above. The following variables are applicable:

t= the index of repetition of survey (e.g., the survey was conducted T times, where $t = 1, \ldots, T$ and T is the total number of times the survey was repeated)

y= the response variable of interest

$\tilde{y}_t$= the estimated domain mean for survey repetition t ($\tilde{y}_t$ is computed as described in section 4.)

Step 2: Perform a regression to forecast $\tilde{y}_{T+j},\ (j \geq 1)$ as follows. Two situations may apply: (I) $\tilde{y}_{T+j},\ (j \geq 1)$ is unconstrained, i.e., its value is not constrained to lie within a given range or (II) $\tilde{y}_{T+j},\ (j \geq 1)$ is constrained.

For the case where $\tilde{y}_T + j,\ (j \geq 1)$ is unconstrained, regression of the estimated population mean against time is given by $\tilde{y}_t = a + bt$. Thus, if we have T repetitions of the survey, we can estimate the T means and variances $\tilde{y}_1, \ldots, \tilde{y}_T$ and $v(\tilde{y}_1), \ldots, v(\tilde{y}_T)$. Using the data set $91, \tilde{y}_1), \ldots, (T, \tilde{y}_T)$ and the corresponding set of weights $1/\sqrt{v(\tilde{y}_1)}, \ldots, 1/\sqrt{v(\tilde{y}_T)}$, standard weighted regression techniques may be used to solve for the least squares estimates of a and b, denoted by $\hat{a}$ and $\hat{b}$, respectively. Thus, the following regression formula is produced:

$$\overline{y}_t = \hat{a} + \hat{b}t \tag{14}$$

For the case where $\overline{y}_{T+j}$ $(i \geq 1)$ is constrained, regression of the estimated population mean against time may be performed according to the following example. Suppose that a company is trying to forecast the percentages of its customer's IT spending budget that will be used to purchase hardware ("H") and the percentage of the budget that will be used to purchase software ("S") at time $T+1$. (Assume that hardware and software form a collectively exhaustive set of all possible ways in which a customer can allocate its IT spending budget.) Let:

$\overline{y}^H_{T+1}$ =estimated percentage of hardware capacity used for business intelligence in period $T+1$

$\overline{y}^S_{T+1}$ =estimated percentage of hardware capacity used for network computing in period $T+1$

Clearly, $\overline{y}^H_1 + \overline{y}^S_1 = 100\%$. In view of this constraint, regression equations may be formed as follows:

$$\begin{aligned}
\overline{y}^i_1 &= a^i + b^i, \quad i = H, S \\
\overline{y}^i_2 &= a^i + 2b^i, \quad i = HI, S \\
&. \\
&. \\
&. \\
\overline{y}^i_T &= a^i + Tb^i, \quad i = H, S
\end{aligned}$$

where $\overline{y}_t^H,\ t = 1, \ldots, T$ estimated percentage of IT spending budget used for hardware purchases in period t and $\overline{y}_t^S,\ t = 1, \ldots, T$ estimated percentage of IT spending budget used for software purchases in period t. Further, we define the following set of constraints:

$$\begin{aligned}
\overline{y}_1^H + \overline{y}_1^S &= 100 \\
\overline{y}_2^H + \overline{y}_2^S &= 100 \\
&. \\
&. \\
&. \\
\overline{y}_T^H + \overline{y}_T^S &= 100
\end{aligned}$$

Now, the task is to determine values for $a^H,\ b^H,\ a^S$, and b^S that satisfy the constraints $\overline{y}_t^H + \overline{y}_t^S = 100\ \ \forall t$. More generally, suppose that there are p variables of interest (i.e., not only variables H and S) for which we have:

$$y_t^i = a^i + b^i t, \quad \text{where } i = 1, \ldots, p \tag{15}$$

$$\sum_{i=1}^{p} y_t^i = k, \quad \text{where } t = 1, \ldots, T. \tag{16}$$

One must find the estimates for parameters a^i and b^i in equation (15), subject to the constraints in equation (16). Suppose for $t = 1$ we have

$$y_1^i = a^i + b^i, \quad \forall i = 1, \ldots, p \tag{17}$$

$$\sum_{i=1}^{p} y_1^i = k \quad \rightarrow \quad \sum_{i=1}^{p} (a^i + b^i) = k \tag{18}$$

For $t = 2$:

$$y_2^i = a^i + 2b^i, \quad \forall i = 1, \ldots, p \tag{19}$$

$$\sum_{i=1}^{p} y_2^i = k \quad \rightarrow \quad \sum_{i=1}^{p} (a^i + b^i + b^i) = k \tag{20}$$

Incorporating (18) into (20), we have:

$$k + \sum_{i=1}^{p} b_i = k \tag{21}$$

Here, equation (21) implies that $\sum_{i=1}^{p} b_i = 0$.

For $t = 3$:

$$y_3^i = a^i + 3b^i, \quad \forall i = 1, \ldots, p \tag{22}$$

$$\sum_{i=1}^{p} y_3^i = k \quad \rightarrow \quad \sum_{i=1}^{p} (a^i + b^i + b^i + b^i) = k \tag{23}$$

Incorporating (20) into (23), produces:

$$k + \sum_{i=1}^{p} b^i = k, \tag{24}$$

again implying that

$$\sum_{i=1}^{p} b^i = 0. \tag{25}$$

Equation (25) is equivalent to

$$b^p = -\sum_{i=1}^{p-1} b^i \tag{26}$$

If we define I^i as an indicator variable, with $I^i = 1$ if the data point is variable i, then the y_t^i expressions can be replaced with the following single expression:

$$\begin{aligned} y_t &= \sum_{i=1}^{p} I^i (a^i + b^i t) \\ &= \sum_{i=1}^{p} a^i I^i + \sum_{i=1}^{p} I^i b^i t \\ &= \sum_{i=1}^{p} a^i I^i + \sum_{i=1}^{p-1} I^i b^i t + (-\sum_{i=1}^{p-1} b^i) I^p t \\ &= \sum_{i=1}^{p} a^i I^i + \sum_{i=1}^{p-1} b^i (I^i t - I^p t) \end{aligned}$$

Equation (27) represents the regression equation, with independent variables I^i and $(I^i t - I^p t)$. The dependent variable is y_t, which is the estimated population mean at time t. The regression equation can be rewritten as:

$$y_t = a_0' + \sum_{i=2}^{p} a^i + \sum_{i=1}^{p-1} b^{i\prime} (I^i t - I^p t) I^i,$$

with least square estimates for the parameters given by

$$\begin{aligned} a^1 &= a^{0\prime} \\ a^i &= a^{0\prime} + a^{i\prime}, \text{ where } i = 2, \ldots, p \\ b^i &= b^{i\prime}, \text{ where } i = 1, \ldots, p-1 \\ b^p &= -\sum_{i=1}^{p-1} b^i \end{aligned}$$

Thus, the following regression equation is produced:

$$y_{it} = a^i + b^i t \text{ where } i = 1, \ldots, p \tag{27}$$

and a^i and b^i defined as above.

Regression equations (14) and (27) may be used in two ways: to detect trends and to forecast future values.

First, the trend of y_t^i (the estimate of the population mean) is determined over time. This may be done using standard hypothesis testing in linear regression as follows. We wish to test $H_0 : \; b^i = 0$. If this null hypothesis is rejected, then it can be said that the variable t (time) does impact y_t^i. In this case, if $b^i > 0$, then y_t^i is increasing over time, i.e., one can say that the population mean is increasing over time. If $b^i < 0$, then y_t^i is decreasing over time. If the null hypothesis is not rejected, then y_t^i is not changing over time.

Second, the regression equations can be used for prediction, to predict future population behavior. If the predicted values are unconstrained, regression equation (14) is used, with the period to be predicted substituted for t in equation (14); if the predicted values are constrained, regression equation (27) is used.

9 Practical Applicaton

In this section we briefly describe the motivating factor which stimulated our interest in this area of survey data analysis. This work was initiated as a result of a desire to maximize the utilization and therefore the value of surveys conducted by a group in the IBM Server Group. This group conducts two types of surveys. The first survey, known as the IT Trends Survey (the "Trends Survey") is conducted on a quarterly basis. This survey targets all IBM server customers and asks questions on topics including current and planned IT spending and current and planned uses for IBM servers. One of the goals of this survey is to help IBM focus future development efforts by understanding why customers currently value IBM technology and what future needs customers may have. The second survey, known as the Win-Loss Survey, is conducted on an annual basis. This survey targets all companies that have purchased a server in the past year, regardless of whether they purchased from IBM or from a competitor. The goal of this survey is to understand why IBM wins some business (i.e., why some companies did buy from IBM) and why IBM loses some business (i.e., why some potential customers decide to purchase from a competitor). We only analyzed responses to the Trends Survey; the Win-Loss Survey had not yet been implemented by the time we began our analysis.

We applied our methodology to data collected from administration of the Trends Survey in the second quarter of 1999. In particular, the group was primarily interested in the responses to two strategic questions included in the survey: (i) "Do you plan to increase your hardware capacity in the upcoming year?" and for those who responded to this question in the affirmative, (ii) "What percentage of your increased capacity do you intend to use for each of the following application areas?" This second question was followed by the following list of ten application areas: (a) Enterprise Resource Planning ("ERP"), (b) Transaction Processing, (c) e-Commerce, (d) Web Serving, (e) Business Intelligence ("BI"), (f) e-Mail, Groupware, Collaborative Computing, (g) Scientific and Technical, (h) Infrastructure, (i) Supply Chain Management ("SCM"), and (j) Customer Relationship Management ("CRM"). The responses to these questions are critical, as they help to shape IBM's technology investment strategy for the upcoming years.

When the survey was administered, the customer population was stratified according to the following characteristics: server brand, size of customer (as measured by number of employees), amount of computing power (measured in mips), customer geographical location, and the cost of the multi-user system or server purchased in the most recent year. The total possible number of strata is computed as follows: server brand could take on one of five values, customer size was grouped into three classes, amount of computing power was grouped into three classes, geographical location were classified into three regions, and cost of system purchased in the most recent year was classified into four categories. Thus, the total number of possible strata is $5 \cdot 3 \cdot 3 \cdot 4 = 180$ strata. However, given the actual combinations that could occur in reality, the actual number of strata was 53. (Some combinations of values for these five variables are not, in practice, feasible.)

A target number of respondents was specified for each stratum. The target sample size per stratum differed for each stratum. The Server Group was interested in collecting a larger sample of large customers with historically greater purchase volumes. Within each stratum, customers were randomly selected and the survey was administered via telephone interview.

The Trend Survey was conducted via telephone interview and the responses were stored in a fixed-field type database within the SAS system. A total of 1582 responses were collected for this survey. Of those responses, 589 respondents answered that they plan to increase capacity in the upcoming year. Only those 589 were asked the subsequent question regarding expected use of the increased capacity.

Since this question is only asked of those customers who responded that they intend to increase computing capacity, domain analysis is the appropriate analysis methodology. The domain of interest is the subset of respondents who responded that they intend to increase capacity.

The IBM Server Group assigned weights to each survey response. The responses were weighted according to the size of each of the responding customers (i.e., the number of employees). The intent was to place greater emphasis on responses provided by larger customers. We note that one may assign different weights for responses to different questions for the same respondent (customer). Thus, weights can be question specific. For example, in our analysis responses to some questions were weighted by the size of the customer while other questions were weighted by factors such as past revenues or past purchase dollars of the respondents. The weights are then incorporated in the estimation formulas.

First, for each of these application areas we compute the estimated population mean percentage of capacity that will be allocated to that application area. Notice that each estimate cannot be made in isolation because this percentage represents a fraction of total increased capacity. Thus, we constrain the estimates so that the sum of all these estimates must equal one (or 100%). To estimate population mean percentage of capacity that will be allocated to each application area we apply formula (7) with the added constraint that the sum of all of these estimates must equal to 100. The domain is the set of all customers who responded that they intend to increase computing capacity. We then use formula (8) to compute the variance of this estimated population mean. Finally, equation (6) is used to obtain confidence bounds on the estimated population mean.

We continue our analysis of these questions by now considering methods for improving the robustness of our estimates of capacity allocation to each of the application areas. Consider the changes in expected capacity allocation to different application areas that were caused by assigning weights to each respondent. IBM may be making significant development decisions based upon the responses to this survey. Thus, we want to be certain that, since the weighting factors are often estimated (and somewhat arbitrary) values, the calculated survey results are not too sensitive to exact values of the weighting factors. Toward this end, we will employ the heuristic described in section 7.2 to adjust the weights assigned to each response. We then compute the adjusted weighted estimate of mean capacity allocation to each application area.

Finally, we conducted a trend analysis as described in section 8. Trend analysis requires a collection of survey responses over time. However, the Trends Survey was new at the time of our analysis, so multiple collections of data did not exist. Instead, we conducted a trend analysis on data collected from another survey administered by the IBM Server Group - the Large Systems Panel ("LCIP") survey. LCIP is a predecessor to the Trends Survey and

Application Area in LCIP Survey	Application Area in Trends Survey
Business Intelligence	BI
Network Computing	e-Commerce, Web serving, e-Mail, Groupware, etc.
Traditional Transaction Processing or Batch	Transaction Processing, Infrastructure
Enterprise Resource Planning	ERP, SCM, CRM
Other	Scientific and Technical

Table 1: Mapping of application areas between LCIP and Trend Surveys

was conducted to collect information on mainframe computer customers. The LCIP survey contained questions similar to those that we analyzed in the Trends Survey, namely, what was the intended use for the increased capacity. The list of possible answers provided by LCIP included: business intelligence, network computing, traditional transaction processing or batch applications, enterprise resource planning, and other applications. This list was not the same as the Trends Survey, but a rough mapping exists, as shown in Table 1. The five areas provided by LCIP can be seen as aggregates of those provided by the Trends Survey.

We used the methodology described in section 8 to analyze four waves of the LCIP survey conducted over a period of 2 years. There was some overlap in the sample sets across waves so they are not completely independent. As a result the regression analysis is only approximate. The analysis was performed for illustrative purposes, as four data points is too small a set to obtain a reliable estimate of a model with two parameters.

10 Conclusion

We have demonstrated how survey information can be used to draw conclusions about current population behaviors. We have discussed how the conclusions drawn can be made more robust both by eliminating statistical outliers from the data as well as by tempering the impact of estimated weights on the prediction of overall population behavior. Finally, we have shown how survey data can be used to detect trends over time in population behavior as well as to predict future population behavior. The methodology was used to analyze responses to surveys conducted by the IBM Server Group. The analysis indicates that the value of surveys can be enhanced by applying the techniques described in this report to obtain more robust estimates of population actions, based upon the sample responses collected.

Acknowledgment

Joseph Caruso of the IBM Server Group provided the context and data for the practical application mentioned in this article, and inspired some of our ideas through our many discussions. Ramesh Inaganti wrote the SAS code to implement the methodology described.

References

[1] Barnett, V. and T. Lewis (1994), Outliers in Statistical Data, New York, John Wiley and Sons.

[2] Cochran, W.G. (1977), Sampling Techniques, New York, John Wiley and Sons.

[3] Dixon, W.J. (1951), “Ratios Involving Extreme Values,” *Annals of Mathematical Statistics*, 22, 68-78.

[4] Hoaglin, D. and B. Iglewicz (1987), “Fine Tuning Some Resistant Rules for Outlier Labeling,” *Journal of the American Statistical Association*, 82, 1147-1149.

[5] Kish, L. (1995), Survey Sampling, New York, John Wiley and Sons.

[6] Newbold, P. (1995), Statistics for Business and Economics, New Jersey, Prentice Hall.

[7] Rosner, B., (1983), “Percentage points for a Generalized ESD Many-Outlier Procedure,” *Technometrics*, 25, 165-172.

[8] Sarndal, C., B. Swensson, and J. Wretman (1992), Model Assisted Survey Sampling, New York, Springer-Verlag.