# Low response rates and their effects on survey results

## 1 Introduction

### 1.1 Background

The Commonwealth Government Statistical Clearing House (SCH) reviews all collections involving 50 or more businesses that are conducted by or on behalf of any Commonwealth Agency, including the Australian Bureau of Statistics (ABS). The objectives of the SCH are to reduce the load placed on business survey respondents by eliminating duplication, ensuring the design and conduct of business surveys follow good practices and the dissemination of information about the surveys it reviews.

One of the recurring issues that the SCH encounters in reviews is low survey response rates. This problem is particularly prevalent in non ABS surveys, which are primarily voluntary. The SCH tries to encourage it's clients to work towards higher response rates, to reduce non response bias effects on the quality of data outputs.

Non response bias occurs when the response characteristics of businesses that do not respond to a survey are different to the response characteristics of the businesses that do respond. Typically, when a business does not respond to a survey, it has a response created for it through the process of implicit imputation. This process assumes the businesses not responding to the survey have the same response characteristics as the businesses that do respond. When this assumption is not true, a bias in the estimates is introduced.

### 1.2 Aims of the Paper

This paper has two main aims. The first aim is to demonstrate the factors which influence non response bias and the relative magnitude of these influences. Simulated data is used to identify possible influential factors and demonstrate their effects on survey estimates.  To do this, a plausible non response model was postulated, and a number of parameters were varied. These parameters represented factors that might be considered to have a possible effect on the magnitude of the non response bias and were the sampling fraction, population standard deviation, response rate and population distribution. These factors were examined to determine the circumstances under which survey managers should be particularly wary of non response bias effects.

The second aim is to illustrate, using a real life case, how low response rates and any subsequent non response bias can affect the quality of survey results and to clarify why it is important to dedicate time and resources into trying to increase response rates. This is done through a case study of data obtained from the Year 2000 (Y2K) survey conducted by the ABS in November, 1998.

This paper first presents the data sources and methods used in demonstrating and evaluating non response bias effects. The results of this evaluation will then be discussed with some conclusions and recommendations made based on the results of the analysis undertaken.

## 2 Data Sources

### 2.1 Simulated Data

The first data source used for this demonstration was simulated data. Two datasets representing the simulated populations to be sampled from were generated. The response variable generated for each population was a continuous, non-negative variable. The first population followed a normal distribution while the second population followed a positively skewed distribution. The skewed distribution was created by first generating a standard normal distribution. A constant factor of 4 was then applied to all values greater than zero in this distribution. Appropriate factors were then applied to all values to generate the required population mean and standard deviation.

Both populations produced were of size 2000 businesses. The population mean for the normally distributed population with a standard deviation of 3000 was 12416 while for the positively skewed population it was 13780. These populations were used for the demonstrations of sample size and initial response rate effects. For the demonstration of population standard deviation effects, normally distributed and positively skewed populations were produced with population standard deviations of 4000, 3000, 2000 and 1000 with the population means between these population being slightly different.

For simplicity, there is no stratification of businesses in these populations. This is not unrealistic given that non response adjustment is usually undertaken at stratum level. As a result, we can view the population generated as a stratum within a population.

All businesses in the populations are operating and in-scope. So the only businesses in the populations are live, in-scope responding businesses and live, in-scope non responding businesses. A number of samples of varying sizes were selected from these populations using simple random sampling without replacement.

#### 2.1.1 Bias Model for the Simulated Data

An issue in using the simulated data is the need to create a bias in the population response. That is, we need to create differences in the probability of a business responding to the survey, for businesses with differing response characteristics. In this investigation, businesses with a larger response are given a higher probability of responding to the survey than businesses with a lower response. These probabilities are generated based on a model presented by Rancourt, Lee and Sarndal (1992). Refer to Appendix 1, section 1 for the form of the model.

The first step in using this model is to specify the initial population non response rate which is to be obtained. This is the mean probability of non response for all businesses in the population. The constant parameter ($\gamma$) in the model is then solved and substituted in the formula for $\theta_i$. The probability of non response for each business is therefore determined by the values of the responses for all businesses in the population, the population size and the initial non response rate specified.

The initial sample non response rate achieved will be approximately the population non response rate specified. It will not be exactly the same as the population non response rate because the selection process is based on random number generation.

If we take a sample based on the probabilities generated from this model and achieve a 20% response rate, we would expect the responding businesses to have larger responses. Thus the estimates at this time point would be expected to be quite large. After some follow up of non responding businesses, we would expect businesses with smaller responses to respond. If we then calculated the estimate with these additional businesses added in, we would expect the estimate to be smaller. So as more units respond, the estimates tend to get closer to the true population value.

### 2.1.2 Selection Methodology for the Simulated Data

In using the simulated data, the first stage of selection was to take ten simple random samples without replacement from the population generated. For each sample, responding businesses were then selected based on the probabilities of non response generated using the model presented in section 3.2. Another stage of respondent selection was then carried out with the non responding businesses now having another chance to be selected as respondents. This process was repeated a further four times until "final" respondents and non respondents were determined.

### 2.2 Year 2000 (Y2K) data

The second data source to be used in this investigation was obtained from the Y2K survey. The survey was intended to determine the level of business readiness for the Y2K computer problem. A random sample of 7800 businesses, stratified by industry and size, was taken, using the ABS Business Register as the population frame. One of the questions asked of businesses was whether they had taken steps to prepare for the Y2K computer problem. This is the variable of interest in this case study.

This dataset was chosen because the response rate was monitored over a period of intensive follow up action, and it was plausible to determine the numbers of late respondents. The date responses were received was recorded so that businesses could be grouped by the week they responded. This is different to the simulated data where the difference between each group of respondents does not represent a set period of time such as a week, but rather a percentage of the remaining sample businesses which had not responded.

Unlike the simulation studies, in the case of this survey, an outstanding portion of the sample remained non responding. The bias associated with this group is unknown, with a portion of the group being non respondents and another portion being businesses that are no longer operating. So the actual bias measured is only a component of the full bias, and should be taken as indicative of a probable lower limit to the bias.

# 3 Methods

## 3.1 Factors of interest

There are a number of factors of the simulated datasets that we were interested in, including the sampling fraction, the population standard deviation and response rates. Differences in the effects that these factors have on the survey estimates across the two population distributions were also of interest.

## 3.2 Estimation

### 3.2.1 Simulated Data

In the investigation using the simulated data, estimates of totals were produced. Number raised estimation was used with a response for non responding businesses being created using implicit imputation. See Appendix 1, section 3 for the form of the estimator. An estimate of the total was calculated for each group of respondents. For example, if the initial response rate achieved for the survey was 40%, an estimate of the total would be calculated from those responding businesses. At the next stage of selection, more businesses would respond to the survey. An estimate would be calculated based on the responses from all  responding businesses at that time. This process would continue a further four times until a final estimate based on all responding businesses at the final response rate was obtained.

The precision of the estimates at each stage of selection was improved by calculating estimates from each of the ten different samples selected as described in section 2.1.2. The mean of the ten different estimates was then calculated and was used as the final estimate for that time point.

### 3.2.2 Y2K data

In the investigation using the Y2K data, estimates of the proportion of businesses taking action on the Y2K problem were produced for employment size groups and across all businesses. These estimates are calculated at time points which represent one additional week of elapsed time. Number raised estimation for proportions was used with a response for non responding businesses being created using implicit imputation. See Appendix 1, section 2 for the form of the estimator.

## 3.3 Analysis

### 3.3.1 Techniques Used

To evaluate the extent of the non response bias present in both the Y2K and simulated data, estimates of proportions ( for the Y2K data) and totals ( for the simulated data ) were calculated at a number of different time points throughout the collection phase of the survey. These time points represent increases in response rate as time passes and/or as follow up of non responding businesses is carried out.

To indicate the extent of the non response bias and the affect it has on the survey estimates over time, graphs of the estimates against increasing response rate were examined. Another method used to assess the non response bias effects was the calculation of the percentage change in the estimates between two time points, X and Y. This was calculated in the following fashion:

((estX - estY)/(estX)) * 100.

This gives a measure of the difference between the estimates at time points X and Y relative to the estimate at time point X expressed as a percentage.

The third technique used to evaluate the extent of non response bias was a test of the difference between the estimates at different time points. This technique assumes that the difference between the estimates at time points X and Y are normally distributed around the true difference. This technique allows us to determine the statistical significance of the difference between estimates using a specified level of confidence.  In this investigation, the level of significance used is 5%. Refer to Appendix 1, section 4 for the form of the test statistic.

It should be noted that if a very large number of samples was taken, we could have concluded with  certainty whether there was in fact a difference between the estimates at two time points. Due to the limited availability of resources, only ten samples were taken. Therefore, to conclude with some degree of certainty whether there was a significant difference between the estimates at two time points, the test described above was used.

# 4 Results

The results presented in the following sections are in two parts. Firstly, section 4.1 concentrates on the results from the investigation using simulated data. These results will be used to show how the factors investigated influence the non response bias on survey estimates. Section 4.2 presents the results from the case study using the Y2K data. These results are used to demonstrate how non response bias affects survey results.

The tables presented in section 4.1 shows the percentage change between estimates at time points X and Y. These time points represent groups of responding businesses. For example, time point 1 represents businesses that initially respond to the survey. Time point 2 represents the businesses that initially respond to the survey plus those businesses that have responded after some more elapsed time. Similarly, time points 3, 4 and 5 represent groups of respondents after more time has elapsed and more responses are received. Time point 6 represents the final group of respondents and time point 7 represents a response rate of 100%. The difference between points represents a larger response rate where the increase in response rate was not achieved over a set period of time such as a week, but rather a percentage of the remaining sample businesses which had not responded.

Unlike for the results in section 4.1, for the results in section 4.2, the difference between time points represents a set period of time (a week) over which additional responses have been received. Time point 1 represents businesses that initially respond to the survey. Time point 2 represents the businesses that initially respond to the survey plus those businesses that have responded after one additional week. Similarly, time points 3 and 4 represent groups of respondents after more time has elapsed and more responses are received. Time point 5 represents the final group of respondents which is all respondents after the four week collection period. It was found in most cases, that the increase in response rate over a week for the Y2K data was similar to the increase in response rate between time points for the simulated data. As mentioned previously, a response rate of 100% was not achieved for the Y2K case study. Therefore, the true bias could not be determined and the bias measured only a probable lower limit to the true bias.

The bold numbers in the tables show the cases where the difference between the estimates at time points X and Y were statistically significant as based on the test described in section 3.3.1.

In Appendix 2, graphs are provided showing the estimates at different response rates for each of the factors of interest using the simulated data. The horizontal line at the bottom of these graphs represent the true population value. In Appendix 3, graphs are provided showing the estimates at different response rates for each employment size group for the Y2K data.

## 4.1 Simulation Findings

### 4.1.1 Sample Fraction Effects

Tables 1 and 2 below show the percentage change between estimates of total at time points X and Y for different sample sizes. These time points represent  groups of responding businesses. The four sample sizes represented in these tables were drawn from normally distributed and positively skewed populations of size 2000. The initial response rate for each of the samples in the tables was 20% as specified in using the model described in section 2.1.1. These tables also show which differences between time points are statistically significant.

#### 4.1.1.1 Normal Distribution

From the results in table 1, we can see that the percentage change figures for each of the sample size groups are quite similar. Of particular interest is the percentage change between time point 1, which is the initial response rate, and time point 6 which is the final response rate. Also of interest is the percentage change between time point 1 and time point 7, which represents a response rate of 100%.  These figures indicate that the change in the estimates from the initial response rate (time point 1) and the later response rates at time points 6 and 7 were, not surprisingly, of a similar magnitude across sample size groups. From this, we can see that non response bias is not lowered by increasing the sample size.

**Table 1: Simulated estimates of total: Percentage change (%) between estimates of total at time points X and Y with different sample sizes and a normally distributed population**

| | Percentage Change (%) Between Estimates at Time Points X and Y | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample Size | 1,2 | 2,3 | 3,4 | 4,5 | 5,6 | 6,7 | 1,6 | 1,7 |
| 100 | 0.7 | 0.7 | **1.5** | 0.3 | 0 | **2.3** | **3.2** | **5.5** |
| 200 | 0.3 | **1.1** | 0.2 | 0.5 | **0.6** | **2.9** | **2.7** | **5.5** |
| 500 | 0.6 | **0.5** | **0.5** | **0.5** | **0.5** | **2.69** | **2.6** | **5.2** |
| 1000 | **0.8** | **0.5** | **0.7** | **0.4** | **0.4** | **2.7** | **2.8** | **5.5** |

This is further demonstrated by the Graphs 1-4 in Appendix 2, where it can be seen that the estimate obtained at the end of the collection period is quite similar across the different sample size groups. This shows that after a significant amount of time and follow up, the differences in the bias effects across the sample size groups are small.

These conclusions are not supported by the results obtained from the test of differences. From the results in table 1 above, it is apparent that as the sample size increases, there tends to be more statistically significant differences between the esti-

estimates at two time points. However, this result is almost certainly due to the increase in sample size leading to smaller standard errors of the estimated differences, which in turn leads to smaller differences being detected as significant. This conclusion is supported by the percentage changes for a sample size of 1000 all being statistically significant while being similar to those for the other sample sizes.

As expected, in taking into account the results obtained from the percentage change figures, graphs and statistical tests, it can be seen that there are only slight, if any, differences in the affects of non response bias on the survey estimates obtained between sample size groups.

### 4.1.1.2 Positively Skewed Distribution

The results and conclusions for the positively skewed population are similar to those obtained for the normally distributed population. From the results in table 2 and Graphs 5-8 in Appendix 2, it can be seen that after a significant amount of follow up, the difference in the bias effects across the sample size groups are small.

**Table 2: Simulated estimates of total: Percentage change (%) between estimates of total at time points X and Y with different sample sizes and a positively skewed population**

| | Percentage Change (%) Between Estimates at Time Points X and Y | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Sample Size | 1,2 | 2,3 | 3,4 | 4,5 | 5,6 | 6,7 | 1,6 | 1,7 |
| 100 | -0.4 | 0.5 | **1.3** | **0.6** | 0.4 | 0.8 | 2.3 | **3.1** |
| 200 | 0.3 | 0.1 | **0.7** | **0.5** | **0.3** | **2** | **1.9** | **3.8** |
| 500 | -0.2 | 0.2 | **0.6** | **0.7** | 0.2 | **2** | **1.5** | **3.4** |
| 1000 | 0.4 | 0.2 | **0.4** | **0.2** | **0.2** | **1.9** | **1.5** | **3.3** |

### 4.1.1.3 Distributional Effects

From the results obtained in tables 1 and 2 and graphs 1-8 in Appendix 2, it can be seen that the population distribution also has an influence on the non response bias effects on the survey estimates. From the percentage change figures, it can be seen that the figures for the positively skewed population are smaller than for the normal population. This is particularly apparent for the differences between time points 1 and 6 and 1 and 7. This shows that the overall change in estimates is not as large for the positively skewed population as for the normally distributed population.

This result may be linked to the bias model presented in section 2.1.1. This model assigns probabilities of non response based on the response of each business, with large businesses more likely to respond. So, for both the positively skewed and normally distributed populations, the large businesses are likely to be the earlier of responding businesses. The remaining businesses in the normal population will still be quite spread out with some businesses having quite small values and other businesses having large values. For the positively skewed population, the remaining businesses will tend to be more grouped.  Most of the businesses that have not responded will have values that are similar. The change in the estimates as more businesses respond is therefore less.

It is expected that if this simulation was run on a population that was negatively skewed, very different results would be found. The non response bias in the survey estimates would be expected to be greater for the negatively skewed population than for both the normal and positively skewed populations.

### 4.1.2 Population Standard Deviation Effects

Tables 3 and 4 below show the percentage change between estimates of total at time points X and Y for populations with different standard deviations. These time points represent  groups of responding businesses. The four standard deviation levels presented in these tables represent samples that were drawn from normally distributed and positively skewed populations of size 2000. The initial response rate for each of the samples in the table was 20% as specified in using the model described in section 2.1.1. These tables also show which differences between time points are statistically significant.

#### 4.1.2.1 Normal Distribution

From the results in table 3, it can be seen that the population standard deviation has a large affect on the non response bias of the survey estimates. The percentage change figures for each of the four standard deviation levels varies markedly, with the larger standard deviation levels having the larger percentage change figures. In particular, the large percentage change between time points 1 and 6 and 1 and 7 indicates that the change in estimates with increasing response rates is large. This would suggest that the non response bias is having a large affect on the survey estimates obtained from the populations with larger standard deviations.

**Table 3: Simulated estimates of total: Percentage change (%) between estimates of total at time points X and Y with different population standard deviations and a normally distributed population**

| | Percentage Change (%) Between Estimates at Time Points X and Y | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Stand. Dev. | 1,2 | 2,3 | 3,4 | 4,5 | 5,6 | 6,7 | 1,6 | 1,7 |
| 1000 | 0 | 0.2 | 0 | **0.1** | 0 | **0.4** | 0.3 | **0.8** |
| 2000 | -0.4 | -0.2 | 0.2 | **0.3** | 0.1 | **1.4** | 0 | **1.5** |
| 3000 | 0.6 | **0.5** | **0.5** | **0.5** | **0.5** | **2.7** | **2.6** | **5.2** |
| 4000 | **1.6** | 0.5 | **1.1** | **0.7** | **0.5** | **5.6** | **4.5** | **9.8** |

This conclusion is supported by the results obtained from the tests of differences. There tend to be more significant differences detected between the estimates at two time points for the populations with larger standard deviations, even though such standard deviations should lead to fewer significant results.

In examining Graphs 9-12 in Appendix 2, it can be seen that the difference between the estimate obtained after the collection period has ended and the estimate at 100% response rate is very different across the different standard deviation groups. This shows that even after considerable time and follow up, the differences in the bias effects across the standard deviation groups is quite large.

From these results, we can conclude that population standard deviation does have an affect on the non response bias effects on survey estimates and that this effect is larger in a population with a more variable characteristic of interest.

There are two reasons for this result. The first is that the potential for non response bias is larger for a population with a larger standard deviation. For a population with a standard deviation 4000, the difference between the largest response and the mean response of the population will be quite large. The corresponding difference for a population with a standard deviation of 1000 will not be as large. Therefore, if the businesses that initially respond to the survey are all large, then the difference in the estimates obtained from the initial respondents and when a response rate of 100% is achieved will also be large. This difference is more marked for a population with a standard deviation 4000 than it will be for a population with a standard deviation of 1000.

The second reason stems from the non response model used in this study. As stated earlier, this model assigns smaller probabilities of non response to businesses with larger responses. If we compared the spread of the probabilities of non response for each of the populations generated, we would find that populations with larger standard deviations have a larger spread of probabilities of non response. Therefore, the businesses with the most extreme large responses in the population with a standard deviation of 4000 will have a smaller probability of non response than those businesses with the most extreme large responses in the population with a standard deviation of 1000.

The difference in the probabilities of non response between the businesses with the largest and smallest responses is smaller for the population with a standard deviation 1000 than it is for the population with a standard deviation of 4000. Therefore, for a population with a standard deviation of 4000, the larger businesses have a higher probability of response relative to the rest of the businesses than for a population with a standard deviation of 1000.

#### 4.1.2.2 Positively Skewed Distribution

The results obtained and conclusions drawn for the positively skewed population are similar to those obtained for the normally distributed population. From the results in table 4 and graphs 13-16 in Appendix 2 it can be seen that the population standard deviation has a large affect on the non response bias of the survey estimates. From these results it can also be seen that the distributional effects here are the same as those outlined in section 4.1.1.3.

**Table 4: Simulated estimates of total: Percentage change (%) between estimates of total at time points X and Y with different population standard deviations and a positively skewed population**

| | Percentage Change (%) Between Estimates at Time Points X and Y | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Stand. Dev. | 1,2 | 2,3 | 3,4 | 4,5 | 5,6 | 6,7 | 1,6 | 1,7 |
| 1000 | 0.1 | **-0.2** | **0.2** | 0 | 0 | **0.2** | 0.2 | **0.3** |
| 2000 | 0.2 | 0.2 | 0.2 | 0 | **0.2** | **0.8** | **0.9** | **1.7** |
| 3000 | -0.2 | 0.2 | **0.6** | **0.7** | 0.2 | **2** | **1.5** | **3.4** |
| 4000 | 0.8 | -0.1 | **0.7** | **0.6** | **0.6** | **3.3** | **2.5** | **5.7** |

### 4.1.3 Initial Response Rate Effects

Tables 5 and 6 below show the percentage change between estimates at time points X and Y for populations with different initial response rates. These time points represent groups of responding businesses. Four different initial response rate levels are presented. The samples were drawn from normally distributed and positively skewed populations of size 2000. These tables also show which differences between time points are statistically significant.

Note in these tables that only the differences between time points 1 and 2, 1 and 3 and 2 and 3 are being examined for the higher initial response rates. Time point 1 represents the businesses that initially respond to the survey. Time point 2 represents the final group of respondents and time point 3 represents a response rate of 100%. Only these time points are examined because the high initial response rates meant that only one attempt at follow up was needed to achieve a 90%+ response rate.

#### 4.1.3.1 Normal Distribution

The results in Table 5 and Graphs 17-20 in Appendix 2 show, not surprisingly, that the affects of the non response bias on survey estimates is larger as the initial response rate decreases. The percentage change figures show that as the initial response rate decreases, the difference between estimates at different time points increases. This is particularly noticeable between time points 1 and 6 and 1 and 7 for initial response rates of 10% and 40% and between time points 1 and 2 and 1 and 3 for initial response rates of 70% and 90%. This indicates that the effects of the non response bias are more marked when the initial response rate is lower.

**Table 5: Simulated estimates of total: Percentage change (%) between estimates of total at time points X and Y with different initial response rates and a normally distributed population**

| | Percentage Change (%) Between Estimates at Time Points X and Y | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Initial Resp. Rate | 1,2* | 2,3* | 3,4 | 4,5 | 5,6 | 6,7 | 1,6 | 1,7 (1,3)* |
| 90% | **1.3** | **0.4** | N/A | N/A | N/A | N/A | N/A | **1.8** |
| 70% | **1.7** | **1.5** | N/A | N/A | N/A | N/A | N/A | **3.2** |
| 40% | **0.7** | **1** | **0.8** | **0.4** | **0.5** | **1.1** | **3.2** | **4.3** |
| 10% | 0.7 | -0.2 | **0.7** | **0.6** | 0.3 | **5** | **2** | **7** |

* For initial response rates of 70% and 90%, time point 2 represents the final group of respondents and time point 3 represents a response rate of 100%

When considering the results obtained from the tests of differences, slightly different conclusions could be drawn. The tests of differences between each pair of estimates for initial response rates of 40% or greater are all significant. For an initial response rate of 10%, not all the tests showed significant differences. From this we might conclude that the non response bias effects with this smaller initial response rate are smaller, which would be surprising result. But this only highlights the care which should be taken in interpreting such statistical tests. There are two possible explanations for this surprising result:

(i) the sample size is much smaller for the lower initial response rates than the sample for the higher initial response rates, leading to larger standard errors of the estimated differences. This is likely to lead to fewer significant results.

(ii) for the 10% initial response rate, after six attempts at follow up, the units which cause the bias - that is the units with the smaller responses - may not have responded. Therefore, the businesses in the sample may all have quite similar response characteristics so the bias between time points 1 and 6 is not very large. This is supported by the fact that the percentage change between time points 6 and 7 and time points 1 and 7 is very large. This indicates that the estimate changes significantly from when the first responses are received (and also from when follow up is concluded) to when a 100% response rate is achieved.

Therefore, even though the tests of differences tend to be more significant with a higher response rate, the percentage change figures and graphs in Appendix 3 show that there are larger non response bias effects with lower response rates. This demonstrates the unsurprising result that the response rate does have an effect on the non response bias effects on survey estimates.

#### 4.1.3.2 Positively Skewed Distribution

The conclusions drawn for the positively skewed population are similar to those obtained for the normally distributed population. As above, the results in Table 6 and Graphs 21-24 in Appendix 2 show that the affects of the non response bias on survey estimates are larger as the initial response rate decreases. From these results it can also be seen that the distributional effects here are the same as those outlined in section 4.1.1.3.

**Table 6: Simulated estimates of total: Percentage change (%) between estimates of total at time points X and Y with different initial response rates and a positively skewed population**

| | Percentage Change (%) Between Estimates at Time Points X and Y | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Initial Resp. Rate | 1,2* | 2,3* | 3,4 | 4,5 | 5,6 | 6,7 | 1,6 | 1,7 (1,3)* |
| 90% | **0.8** | **0.2** | N/A | N/A | N/A | N/A | N/A | **1** |
| 70% | **1.4** | **0.9** | N/A | N/A | N/A | N/A | N/A | **2.2** |
| 40% | **1.2** | **0.6** | **0.7** | **0.5** | **0.3** | **0.5** | **3.2** | **3.7** |
| 10% | 0.7 | 0.6 | **0.6** | **0.5** | **0.4** | **2.9** | **2.8** | **5.6** |

* For initial response rates of 70% and 90%, time point 2 represents the final group of respondents and time point 3 represents a response rate of 100%

## 4.2 Y2K Case Study Findings

### 4.2.1 Size Group Results

Table 7 below shows the percentage change between estimates of the proportion of businesses that are taking action to prepare for the Y2K problem at time points X and Y for different employment size groups.

**Table 7: Proportion of businesses that are taking action to prepare for the Y2K problem: Percentage change (%) between estimates at time points X and Y for different size groups**

| | Percentage Change (%) Between Estimates at Time Points X and Y | | | | |
|---|---|---|---|---|---|
| Size Group | 1,2 | 2,3 | 3,4 | 4,5 | 1,5 |
| 0-4 | **8.1** | 1 | 3 | 1.2 | **12.7** |
| 5-19 | 1.2 | **3.6** | -0.6 | **1.7** | **5.8** |
| 20-199 | -0.2 | 0.5 | 0.3 | 0.4 | 1 |
| 200+ | 0.4 | -0.3 | 0.3 | -0.1 | 0.3 |
| Aust. Level | **5.7** | 1.7 | 1.8 | **1.3** | **10.2** |

From the results in table 7 and the graphs in Appendix 3, it is clear that the percentage change for the smaller size groups (ie employment 0-4 and 5-19) is larger than for the other sample sizes. Of particular interest is the differences between time points 1 and 5 where the values for the smaller size groups are larger than for the larger size groups. This shows that there was a larger change in the estimates from the initial response rate (time point 1) and the final response rate which was achieved after four weeks (time point 5). This shows that in this case the non response bias for small size groups is greater than for larger size groups.

The conclusions drawn from the tests of the differences between the estimates support those drawn from the percentage change figures. There were statistically significant differences between time points 1 and 2 and 1 and 5 for size group 0-4 and between time points 2 and 3, 4 and 5 and 1 and 5 for size group 5-19. These results demonstrate that the estimate changed significantly over the collection period, showing that a non response bias is present in the early sample of the small size groups.

It is apparent that the bias present in the small size groups has had an effect on the Australian level estimate. There are large percentage change figures between time points 1 and 2 and 1 and 5. The tests for these differences were also statistically significant. These significant results correspond to significant results in both size groups 0-4 or 5-19. From these results it can be concluded that the estimates for the small size groups have been affected by a non response bias. It can also be concluded that this bias has affected the overall estimates.

These results make sense when we consider the variable of interest. We are measuring whether a business has taken action on the Y2K problem. It would seem sensible that most large businesses (ie. employment >20) would be taking some significant action to ready themselves for the Y2K problem. On the other hand, it would seem more likely that small businesses who typically have less resources to call upon, may not have the time or money to take any significant action or are leaving it until later to deal with the problem. Therefore, the potential for there to be a non response bias in the large businesses is less because almost all would respond "yes" to this question.

Also, the proportion of small businesses who will have taken some action to ready themselves for the Y2K problem is lower. You might expect that businesses that have taken some action might be more willing to respond to the survey straight away. This would explain the higher proportions at lower response levels. These businesses might not require much, if any, follow up to respond. Conversely businesses who have not taken any action may be more reluctant to respond to the survey. They might see this issue as less important. They would be more likely to require follow up to respond to the survey. This explains why the proportion of businesses taking action decreased as more responses came in from the businesses who answered "no" to the question.

### 4.3 Simulation vs Case Study Findings

The simulation was designed to be a realistic portrayal of real economic data, where it is known that larger units are generally more likely to respond. The two distributions used in these simulations were also designed to be representations of "common" variables. The normal distribution is known to be representative of many variables, with positively skewed distributions being reflected in many economic variables.

The case study involved a categorical variable, as opposed to the continuous variable used in the simulation exercise. The variables of interest in many social surveys are categorical, and this provides a useful comparison to the continuous variable used in the simulation exercise.

It is reassuring that the demonstrations for both the simulation exercise and Y2K case study resulted in very similar conclusions. Not only was it demonstrated in both cases that as the response rate increased that the accuracy of the estimates increased (a not surprising result), but the general magnitude in the improvements in accuracy were both significant and similar.

## 5 Conclusions

This investigation had two main aims. The first was to demonstrate the effects that low response rates and subsequent non response bias have on survey estimates. The second aim was to determine the circumstances under which non response bias may be minimised.

From the results obtained from the demonstrations using simulated data, it has been shown that there are some determining factors in the non response bias effects on estimates. It was demonstrated for the datasets examined, that the population standard deviation and response rate achieved have a large effect on the magnitude of the non response bias. The population distribution also has some effect while the sample fraction, not surprisingly, has very little effect.

From the results of this demonstration, it is recommended that to reduce the effects of non response bias, it is more important to achieve a higher final response rate than to have a large initial sample size. It follows that the extra resources saved from not taking a large sample, could be used to increase the effort expended in following good survey practices. It is also recommended that particular effort be made to ensure high response rates when it is known or suspected that the characteristic to be estimated is quite variable. This will reduce the affects of the non response bias on the estimates.

From the results obtained from the case study using the Y2K data, it was clear that there was a non response bias present in the estimated proportions of businesses taking action to prepare for the Y2K problem. This bias was particularly prevalent within the smaller sized businesses. The effects that this bias had were apparent not only for the estimates for the smaller sized businesses, but also for the estimated proportions at the Australian level. From these results, it is apparent that as the response rate increased, the non response bias in the estimates decreased, leading to a likely increase in the accuracy of the results.

Overall, we can conclude that non response bias can have significant detrimental effects on the accuracy of survey estimates. These affects can be reduced through higher response rates. From this, it is recommended that steps should be taken to ensure that the response rates achieved in any particular survey are as high as possible. Response rates can be increased through good survey practices such as the use of high quality questionnaires, increased interviewer training, assuring respondents of the confidentiality of the information they provide and dedication of resources and time to following up non respondents.

## 6 References

RANCOURT, E., LEE, H., AND SARNDAL, C. (1992). Bias Corrections For Survey Estimates From Data With Imputed Values For Nonignorable Nonresponse. Proceedings of the Bureau of the Census Annual Research Conference, 523-539.

Appendix 1: Formula

1. Response bias model

$$\theta_i = \exp(-\gamma\ y_i)$$

where : $\theta_i$ is the probability of non response for business i

$\gamma$ is a constant solved subject to $\bar{\theta} = (1/N)\sum \theta_i$

$y_i$ is the response of business i

$\bar{\theta}$ is the mean probability of non response for all businesses in the population

N is the population size

2. Number raised estimator for proportions with implicit imputation

$$\hat{P}_c = \frac{\sum_1^H \frac{N_h}{n_{h,r}} * n_{h,r,c} * (n_{h,r} + n_{h,nr} / n_h)}{\sum_1^H N_h * (n_{h,r} + n_{h,nr} / n_h)}$$

where : $h$ is the stratum identifier

$N_h$ is the stratum population size

$n_h$ is the stratum sample size

$n_{h,r}$ is the number of responding businesses in stratum h

$n_{h,nr}$ is the number of non responding businesses in stratum h

$n_{h,r,c}$ is the number of responding businesses in stratum h with the characteristic of interest.

Note. $(\frac{n_{h,r} + n_{h,nr}}{n_h})$ is the ratio of the number of operating businesses in the sample to the total number of businesses in the sample. This factor, when applied to the population count, estimates the number of operating businesses in the population .

3. Number raised estimator for totals with implicit imputation

$$\overline{\hat{Y}} = \sum \frac{\hat{Y}_k}{10}$$

$$\hat{Y}_k = \frac{N}{n_{r,k}} * \sum y_{i,r,k}$$

where : k is the sample number, $k = 1$ to 10

$\hat{Y}_k$ is the estimate of total from sample k

N is the population size

$n_{r,k}$ is the number of responding businesses in sample k

$\sum y_{i,r,k}$ is the sum of the responses from responding businesses in sample k

4. Test statistic for the estimated difference between the parameter values at two time points.

$$Z = \frac{\hat{D}}{\hat{S}E(\hat{D})}$$

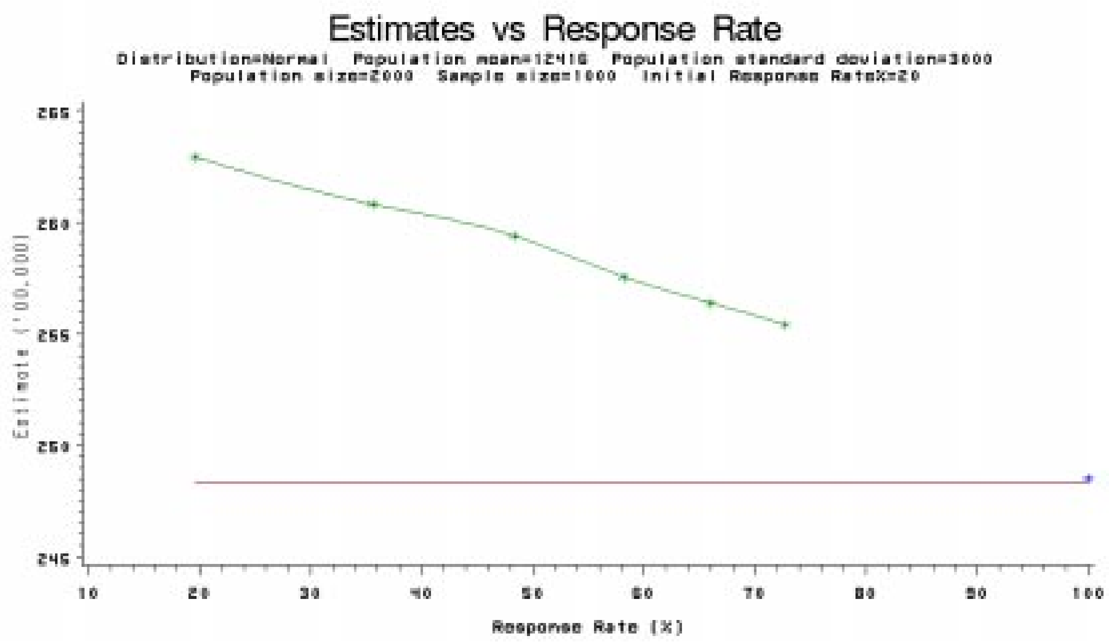
where : Z is the test statistic

$\hat{D}$ is the estimated difference between the parameter values at two time points

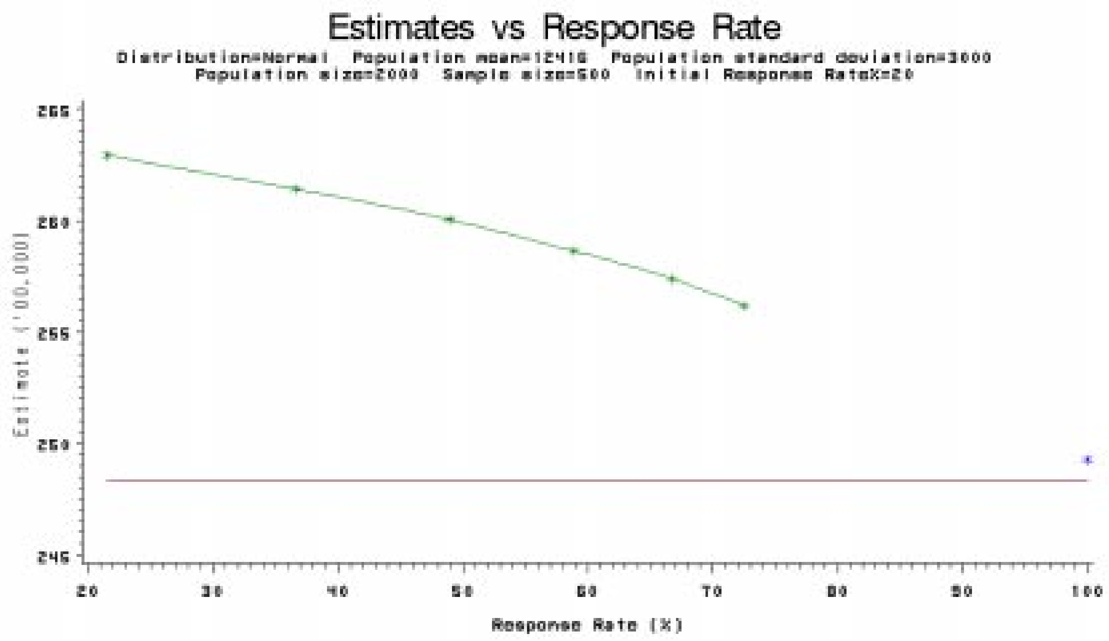
$\hat{S}E(\hat{D})$ is the estimated standard error of the estimated difference between parameter values at two time points

Appendix 2: Graphs using simulated data

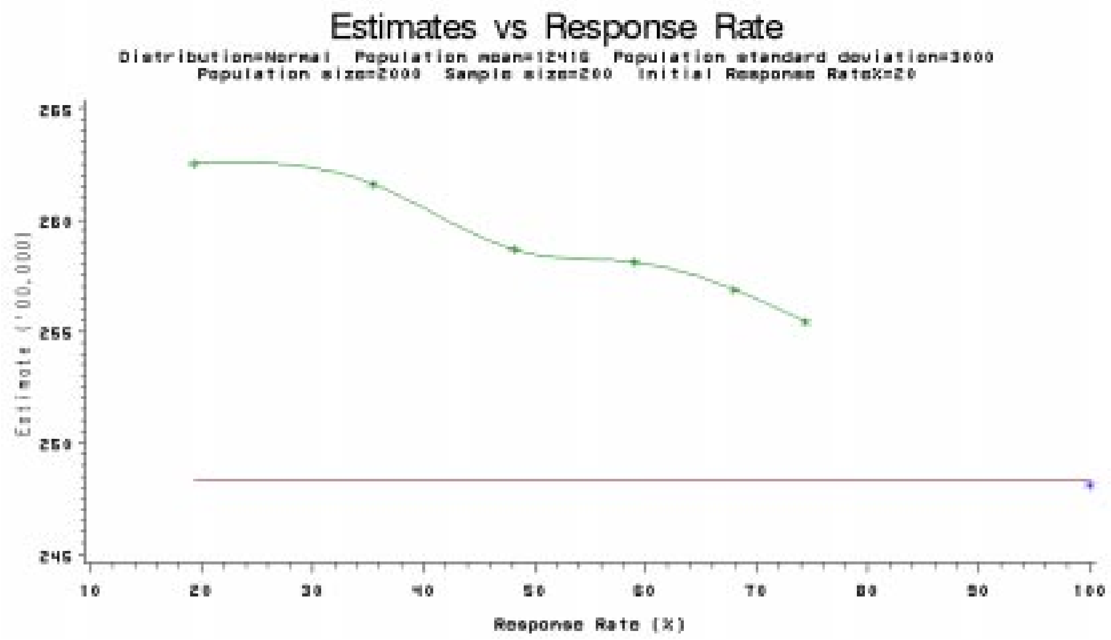
Graph 1: Estimates vs Response Rate (%) for a sample size of 1000 businesses drawn from a normally distributed population



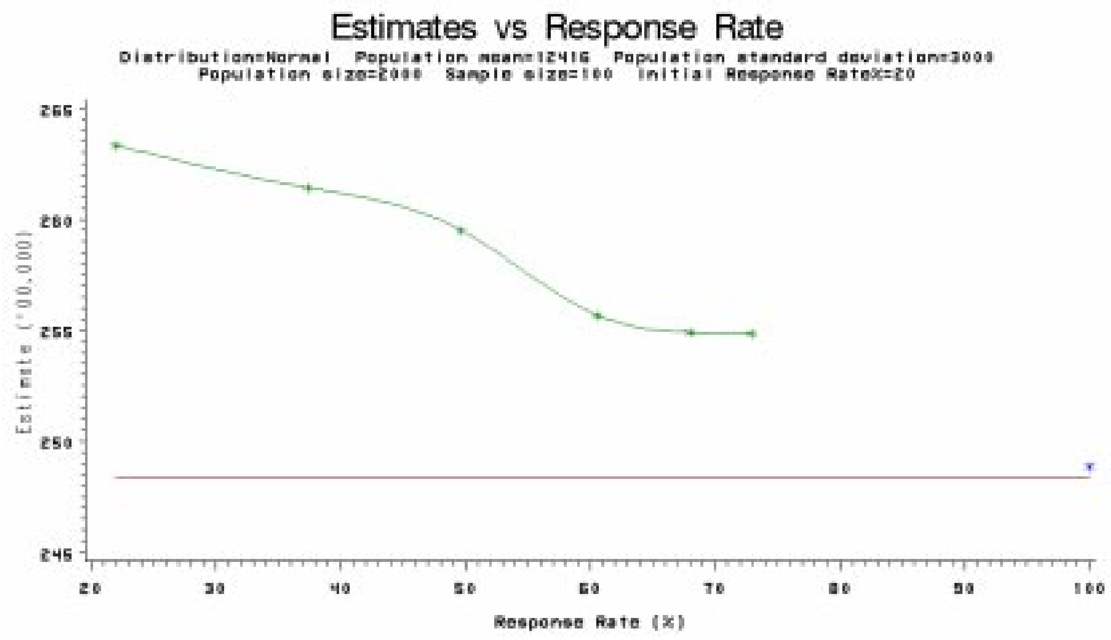
Graph 2: Estimates vs Response Rate (%) for a sample size of 500 businesses drawn from a normally distributed population



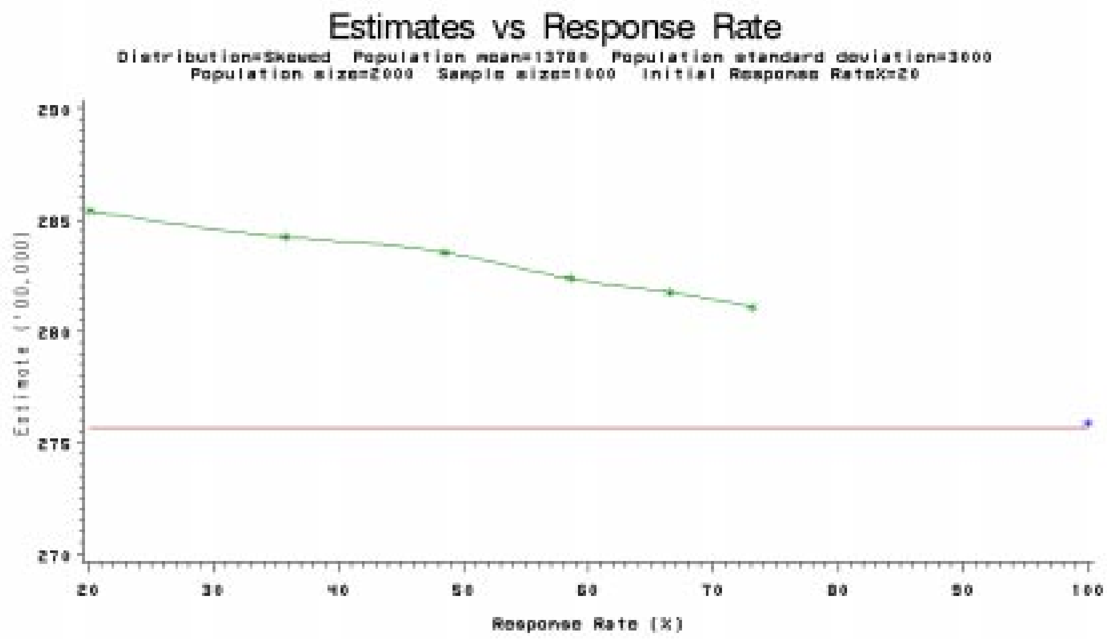
Graph 3: Estimates vs Response Rate (%) for a sample size of 200 businesses drawn from a normally distributed population



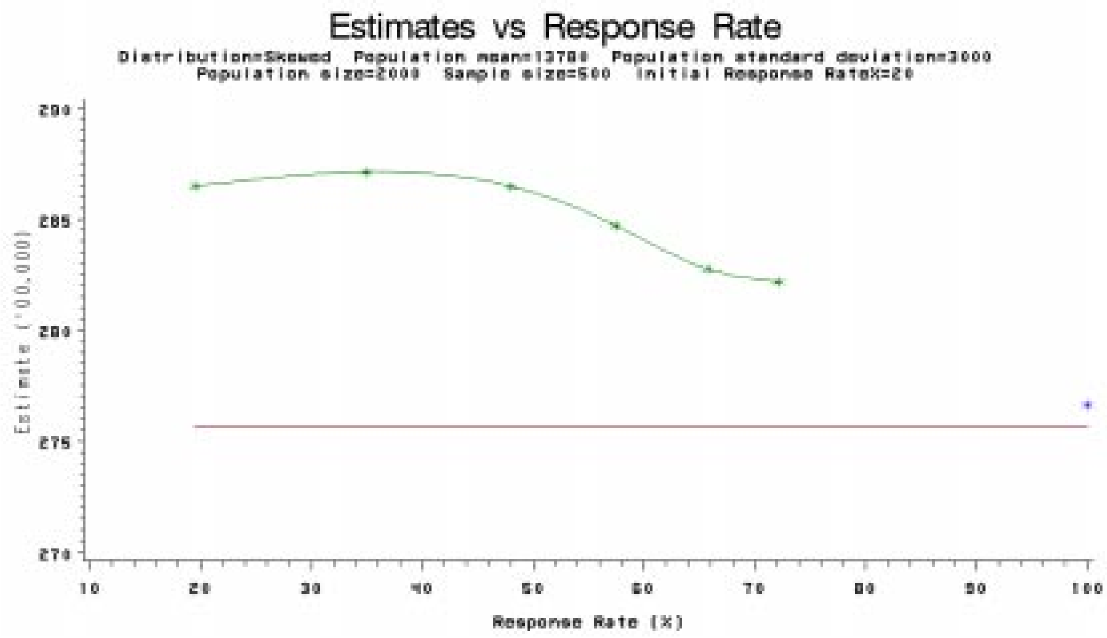
Graph 4: Estimates vs Response Rate (%) for a sample size of 100 businesses drawn from a normally distributed population



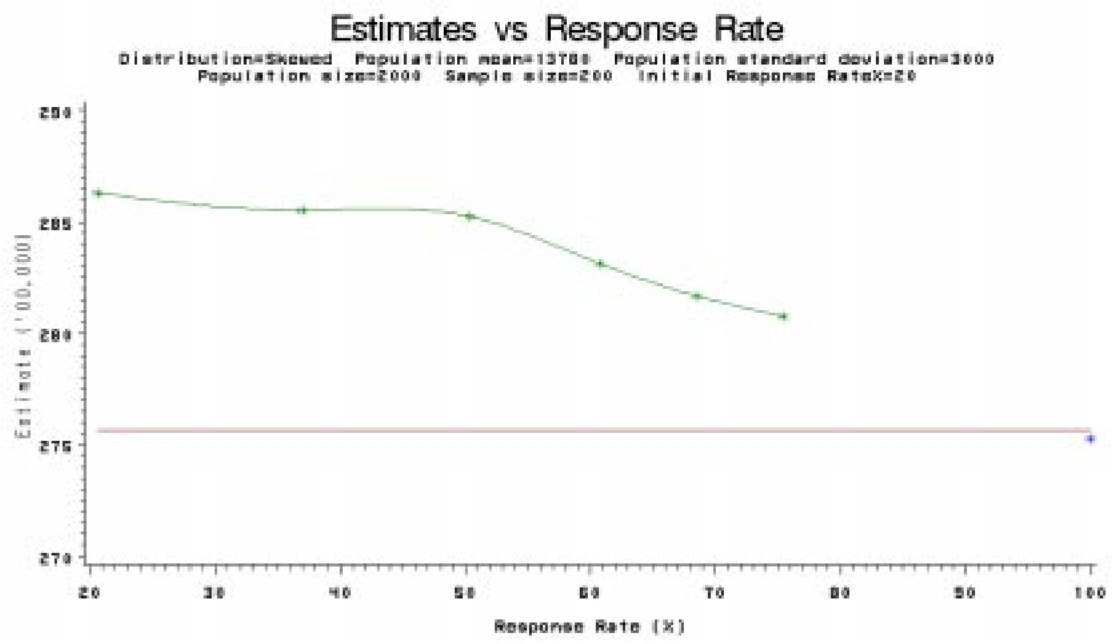
Graph 5: Estimates vs Response Rate (%) for a sample size of 1000 businesses drawn from a positively skewed population



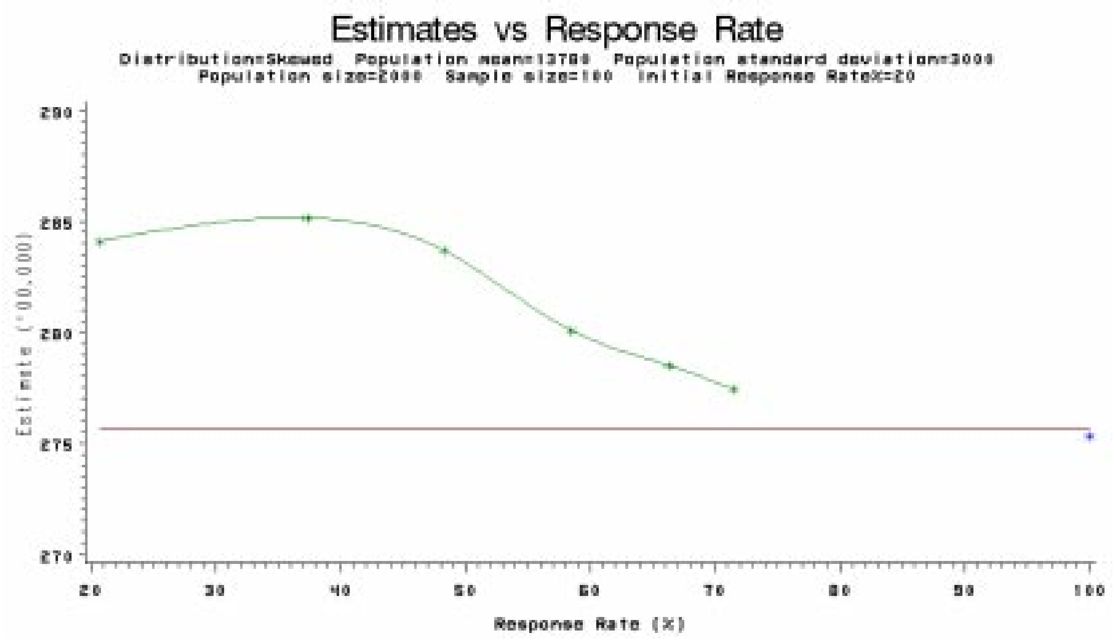
Graph 6: Estimates vs Response Rate (%) for a sample size of 500 businesses drawn from a positively skewed population



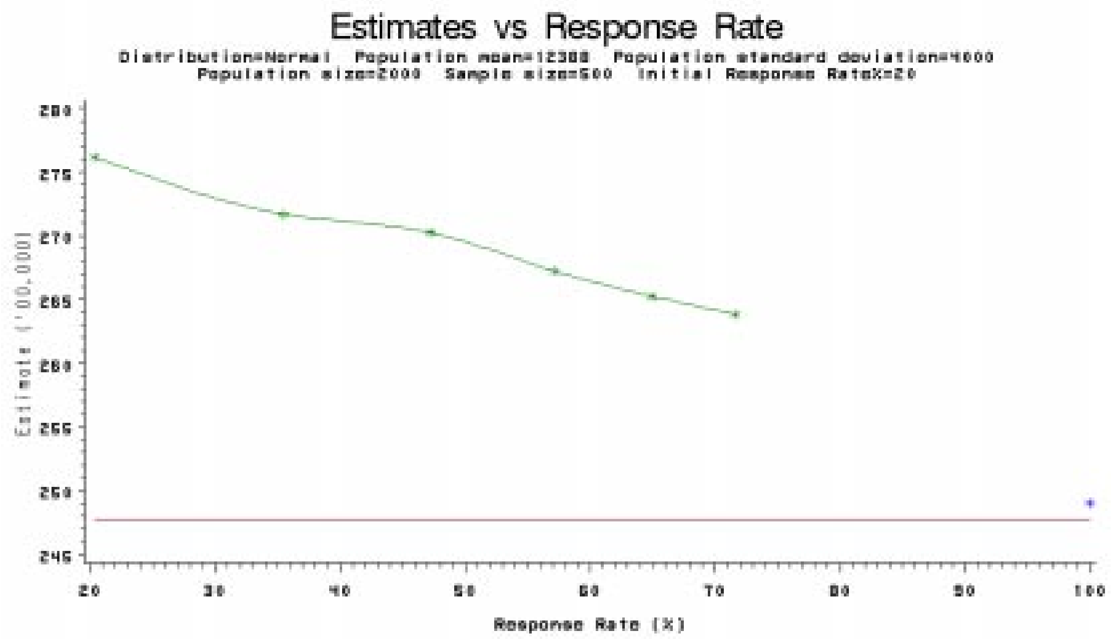
Graph 7: Estimates vs Response Rate (%) for a sample size of 200 businesses drawn from a positively skewed population



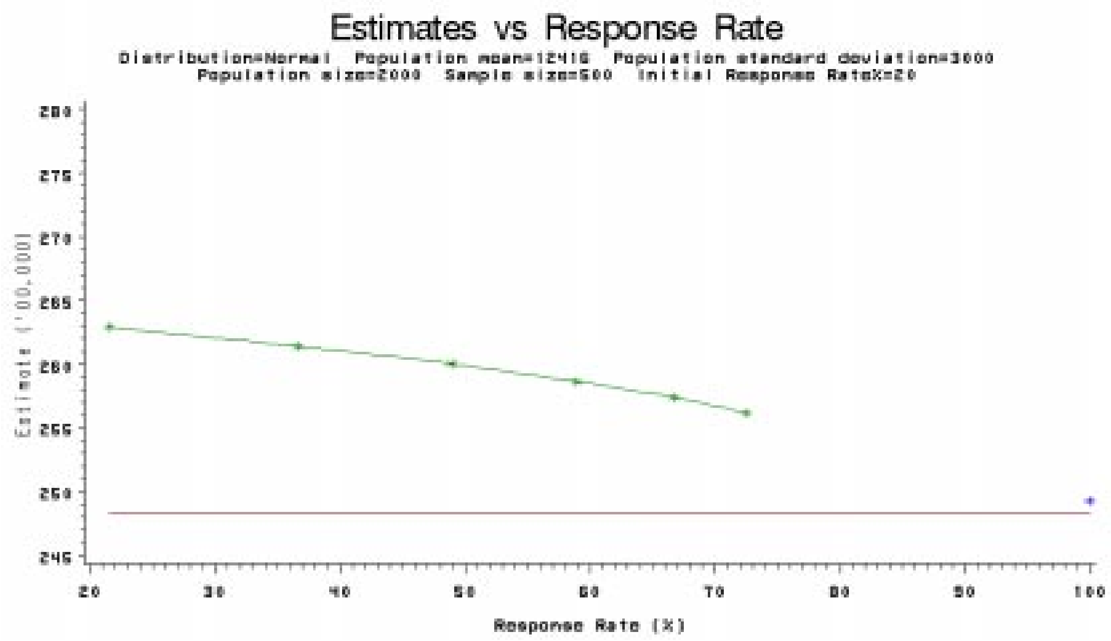
Graph 8: Estimates vs Response Rate (%) for a sample size of 100 businesses drawn from a positively skewed population



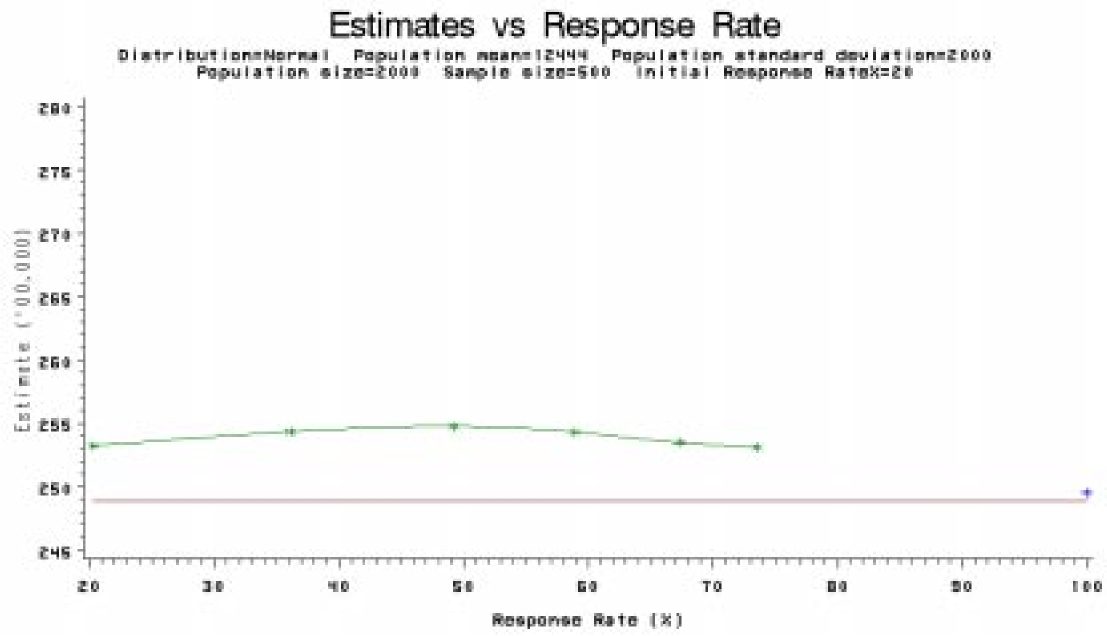
Graph 9: Estimates vs Response Rate (%) for a normally distributed population with a standard deviation of 4000



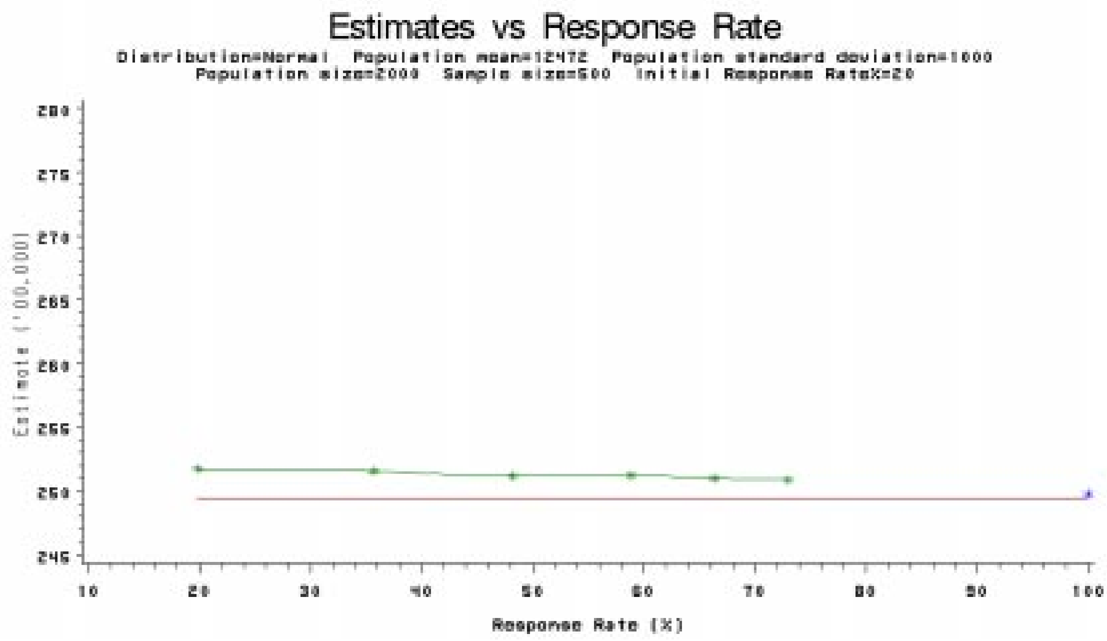
Graph 10: Estimates vs Response Rate (%) for a normally distributed population with a standard deviation of 3000



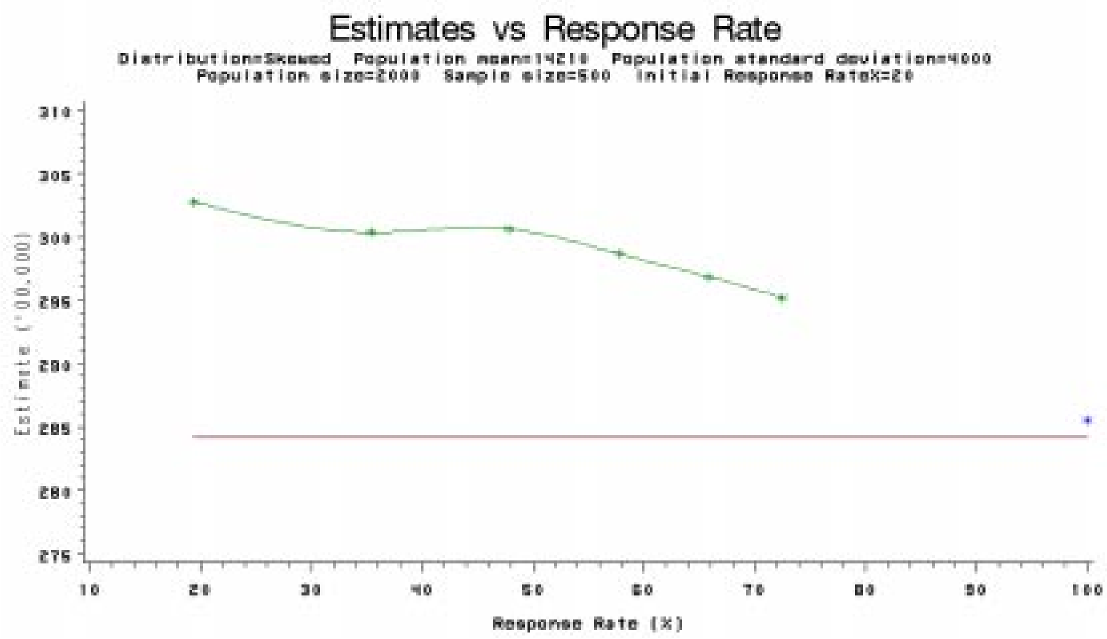
Graph 11: Estimates vs Response Rate (%) for a normally distributed population with a standard deviation of 2000



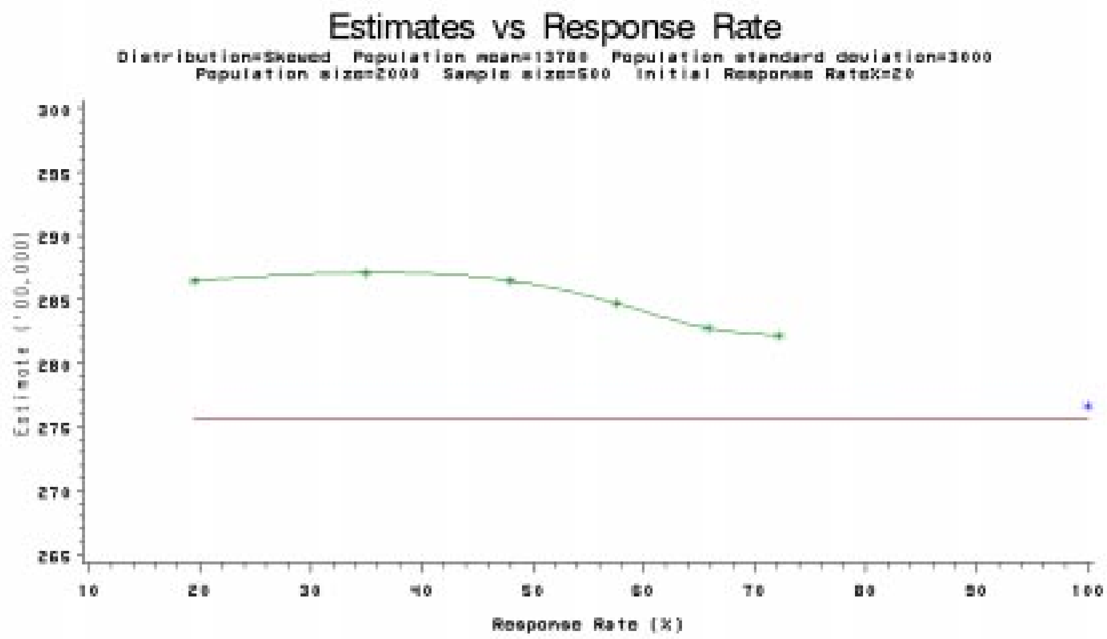
Graph 12: Estimates vs Response Rate (%) for a normally distributed population with a standard deviation of 1000



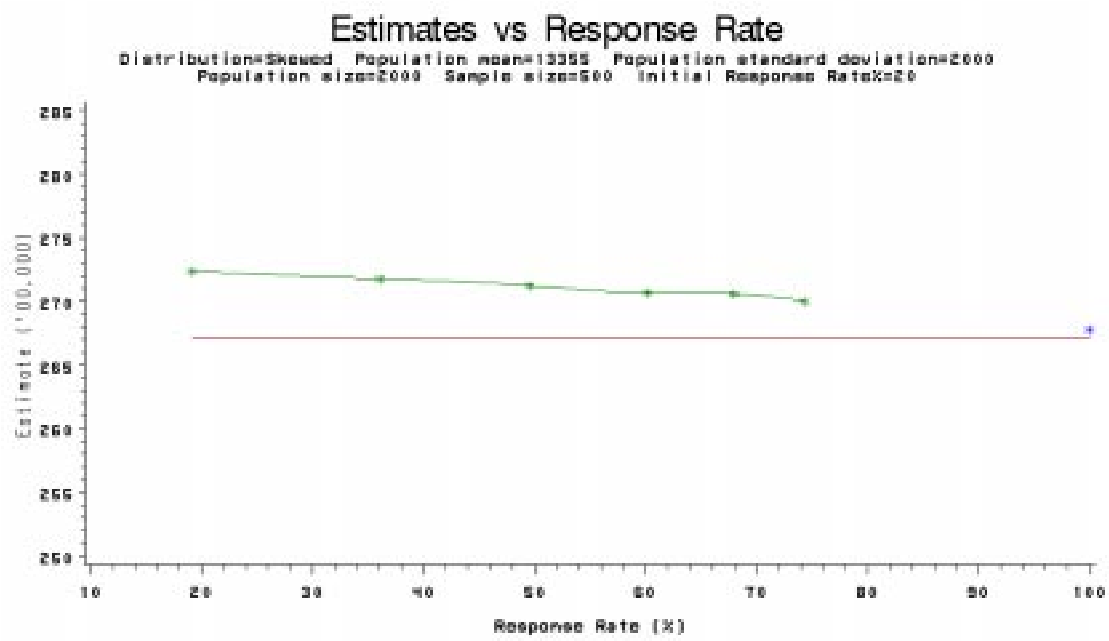
Graph 13: Estimates vs Response Rate (%) for a positively skewed  population with a standard deviation of 4000



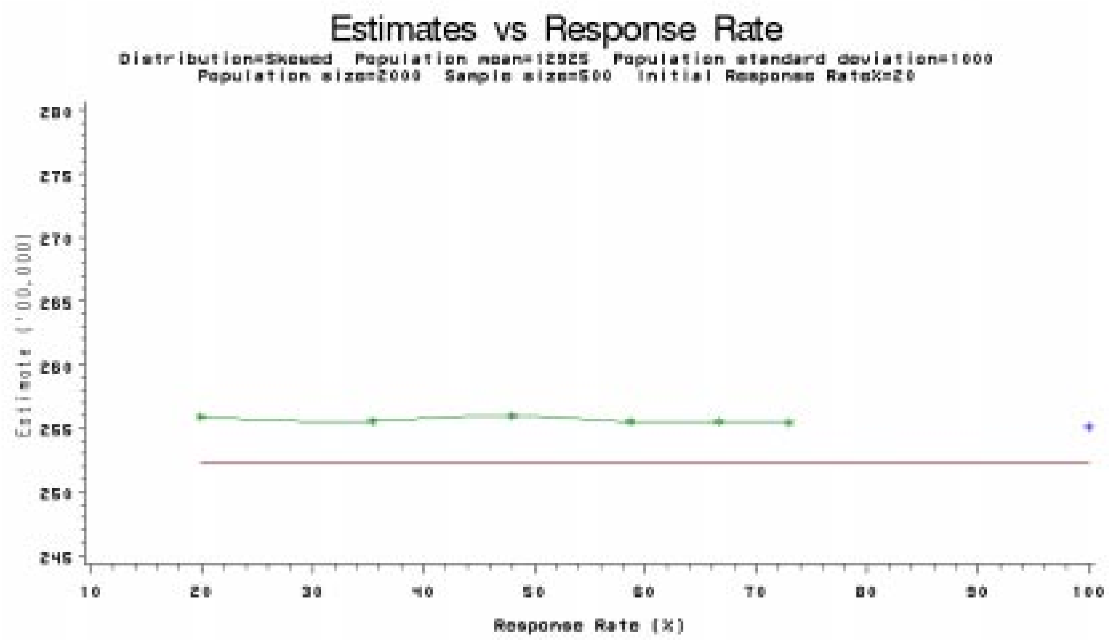
Graph 14: Estimates vs Response Rate (%) for a positively skewed  population with a standard deviation of 3000



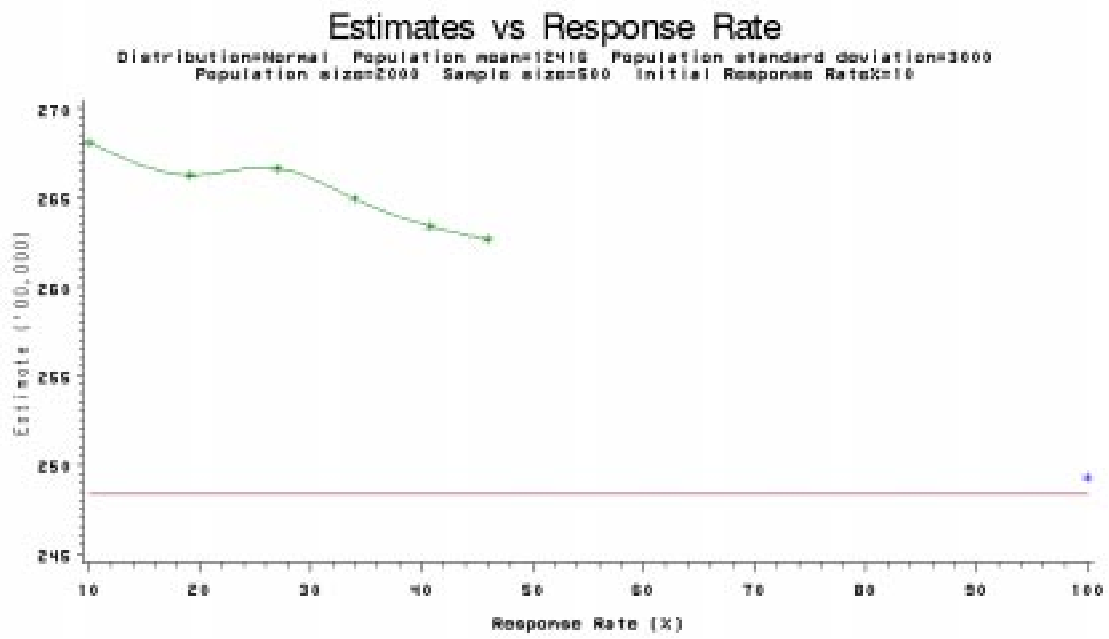
Graph 15: Estimates vs Response Rate (%) for a positively skewed population with a standard deviation of 2000



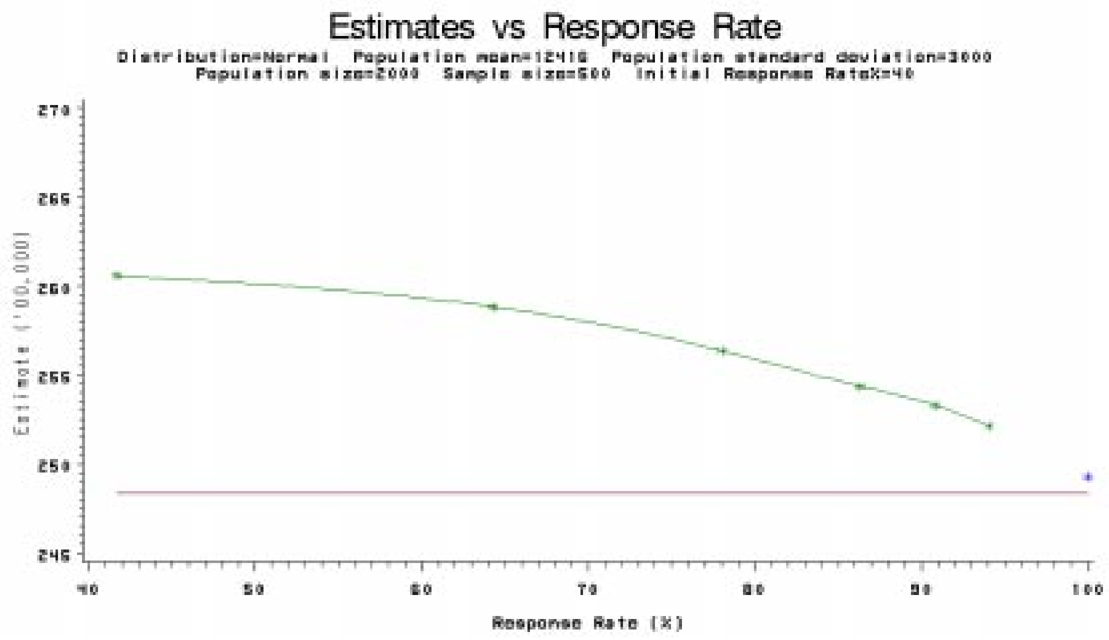
Graph16: Estimates vs Response Rate (%) for a positively skewed population with a standard deviation of 1000



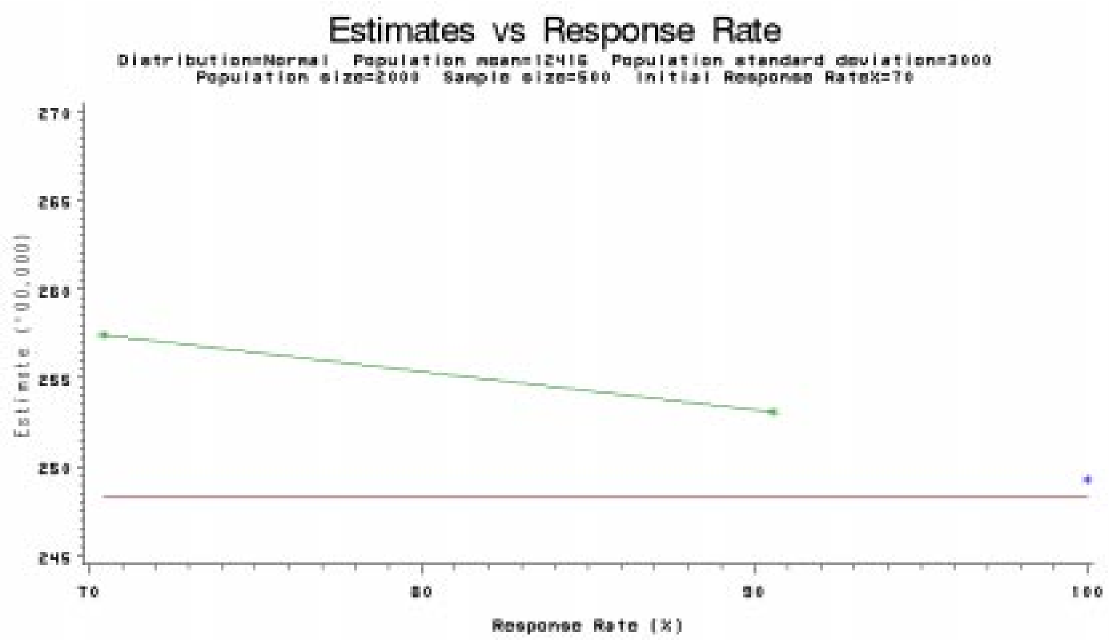
Graph 17: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 10% drawn from a normally distributed population



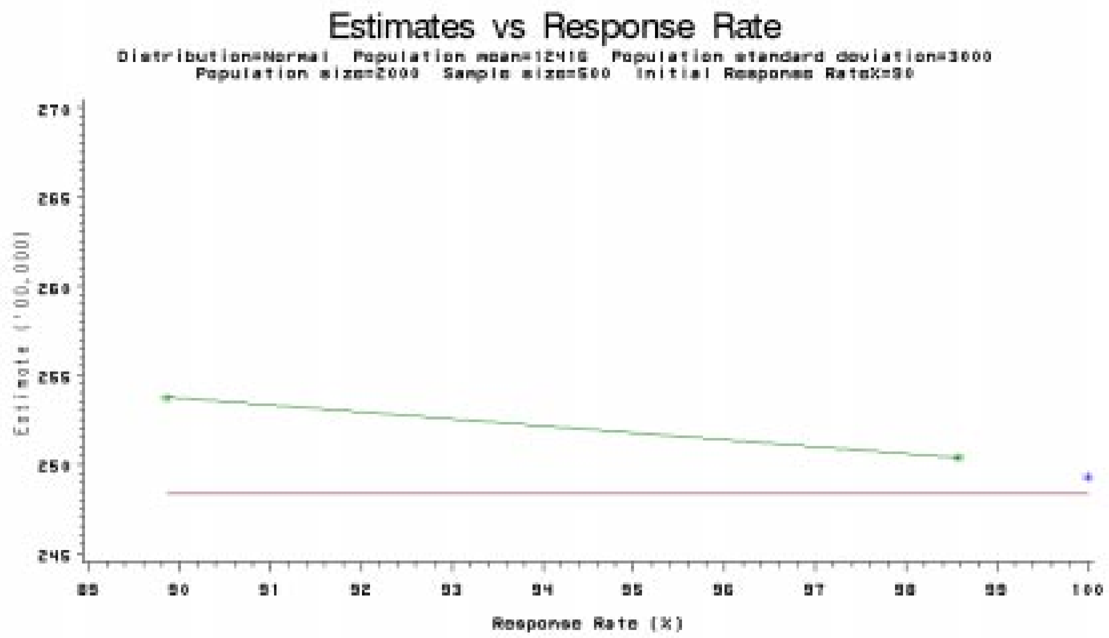
Graph 18: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 40% drawn from a normally distributed population



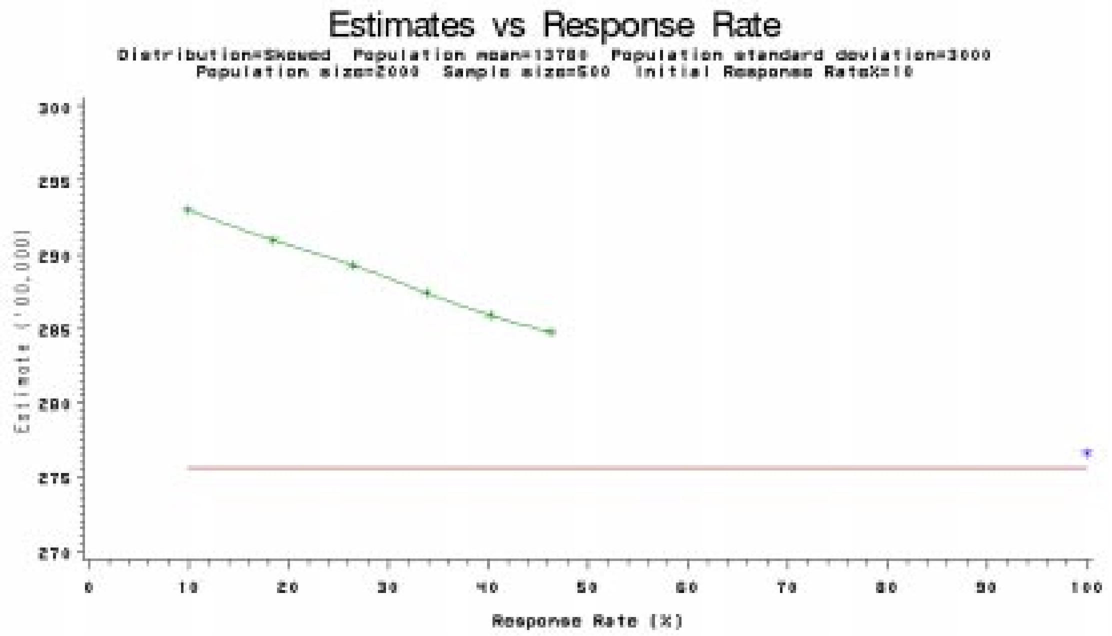
Graph 19: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 70% drawn from a normally distributed population



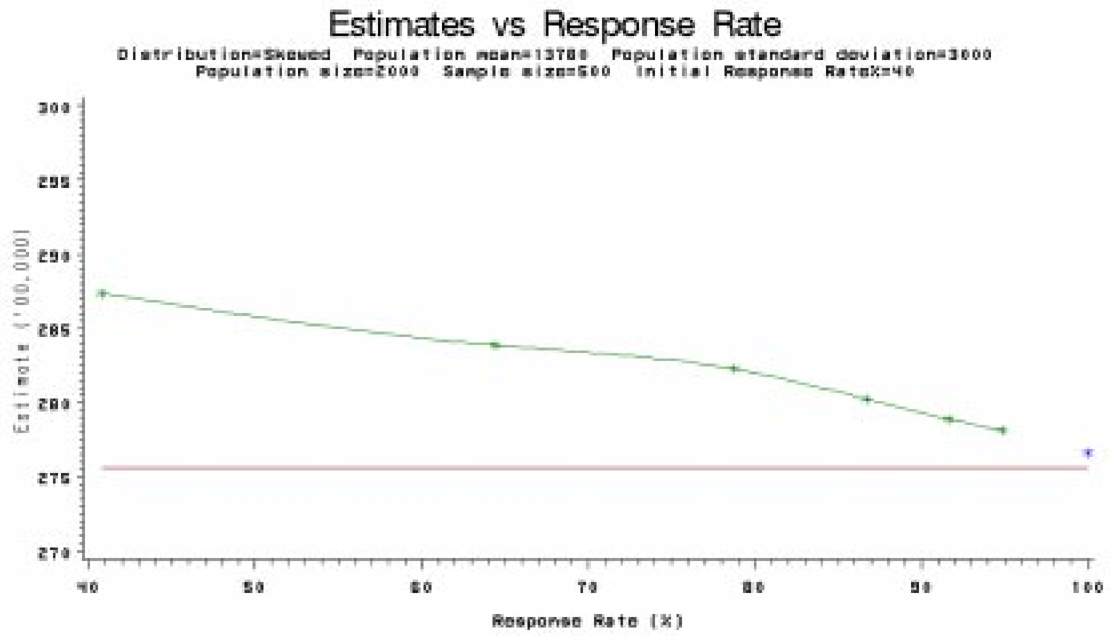
Graph 20: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 90% drawn from a normally distributed population



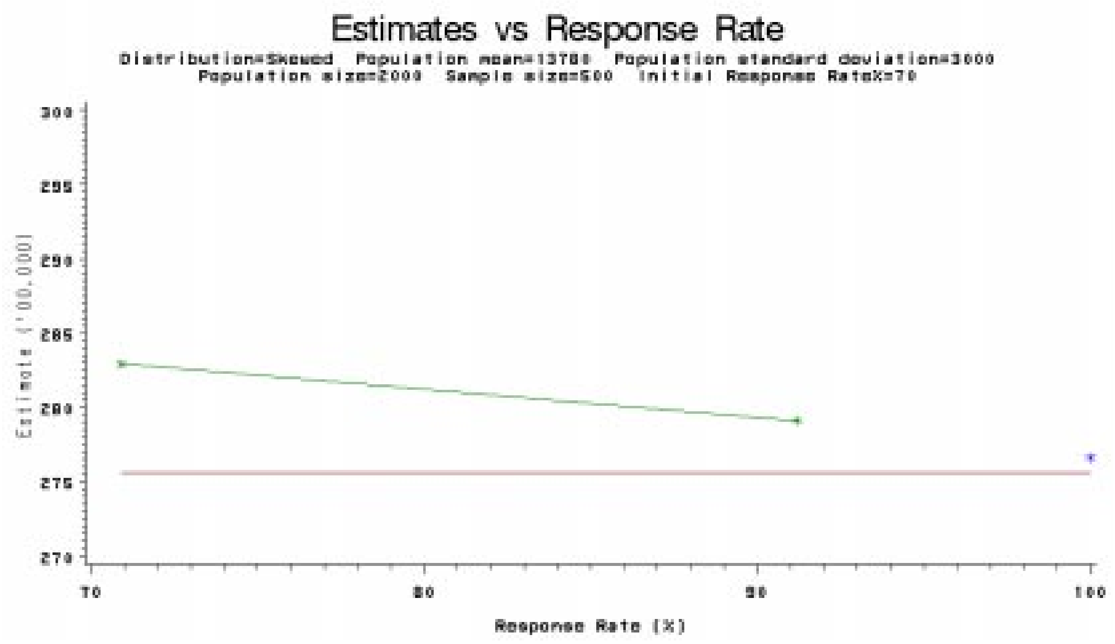
Graph 21: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 10% drawn from a positively skewed population



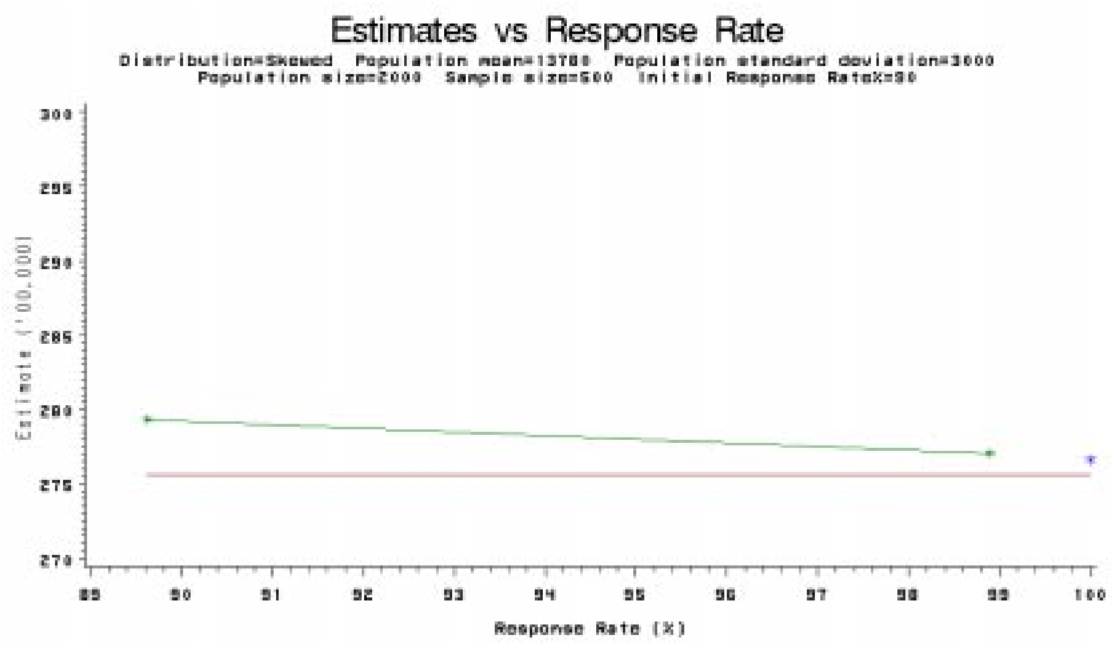
Graph 22: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 40% drawn from a positively skewed population



Graph 23: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 70% drawn from a positively skewed population

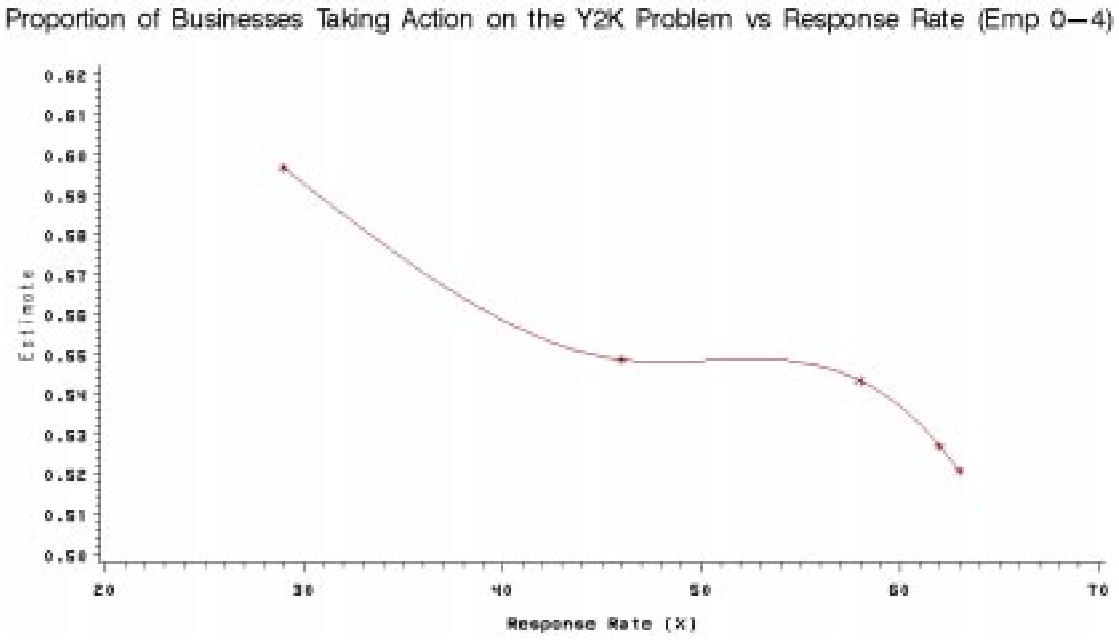


Graph 24: Estimates vs Response Rate (%) for a sample of 500 businesses with an initial response rate of 90% drawn from a positively skewed population

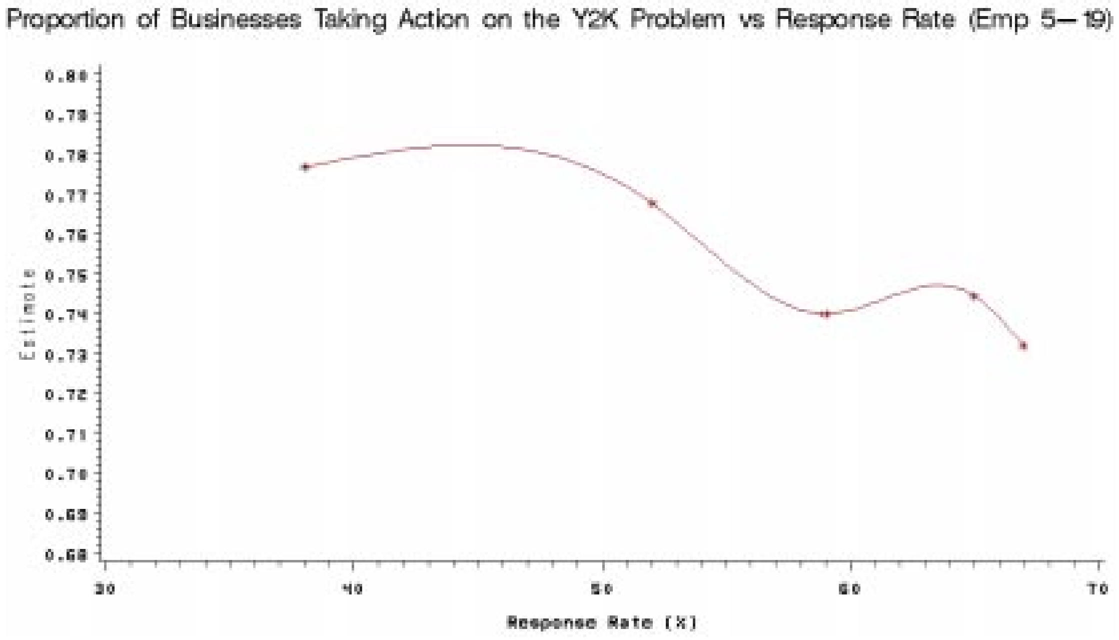


Appendix 3: Y2K Graphs

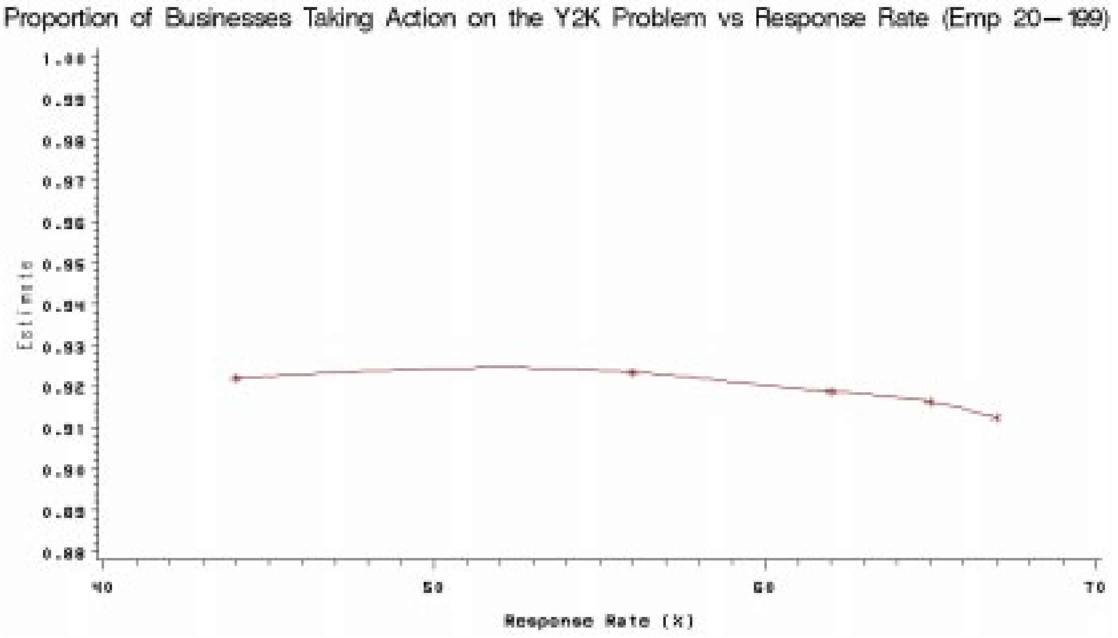
Graph 1: Estimates vs Response Rate for Employment Size Group 0-4



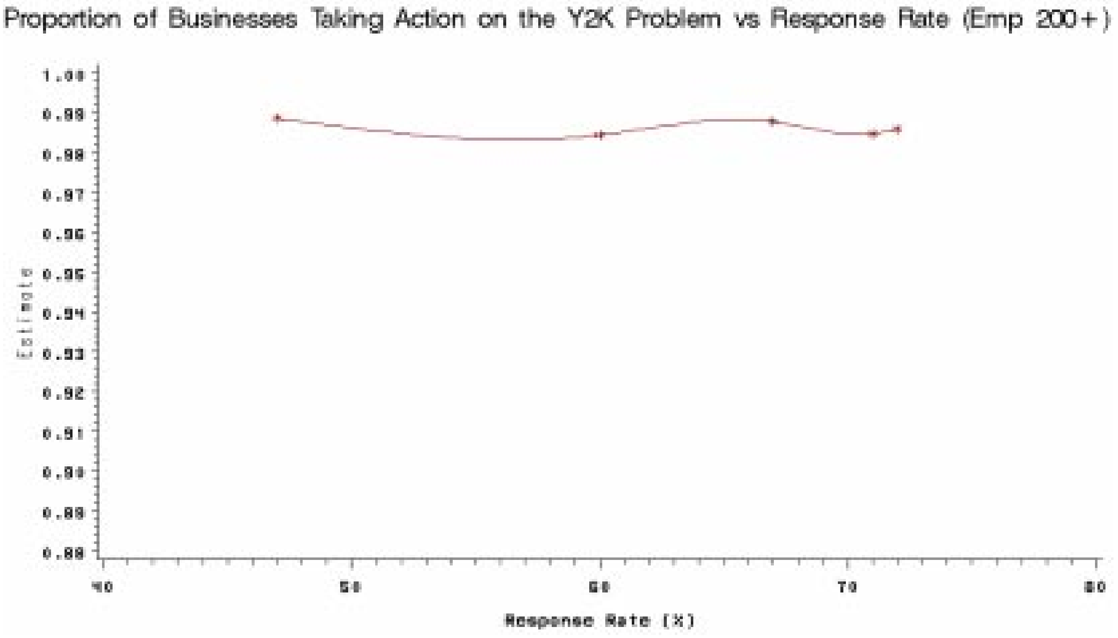
Graph 2: Estimates vs Response Rate for Employment Size Group 5-19



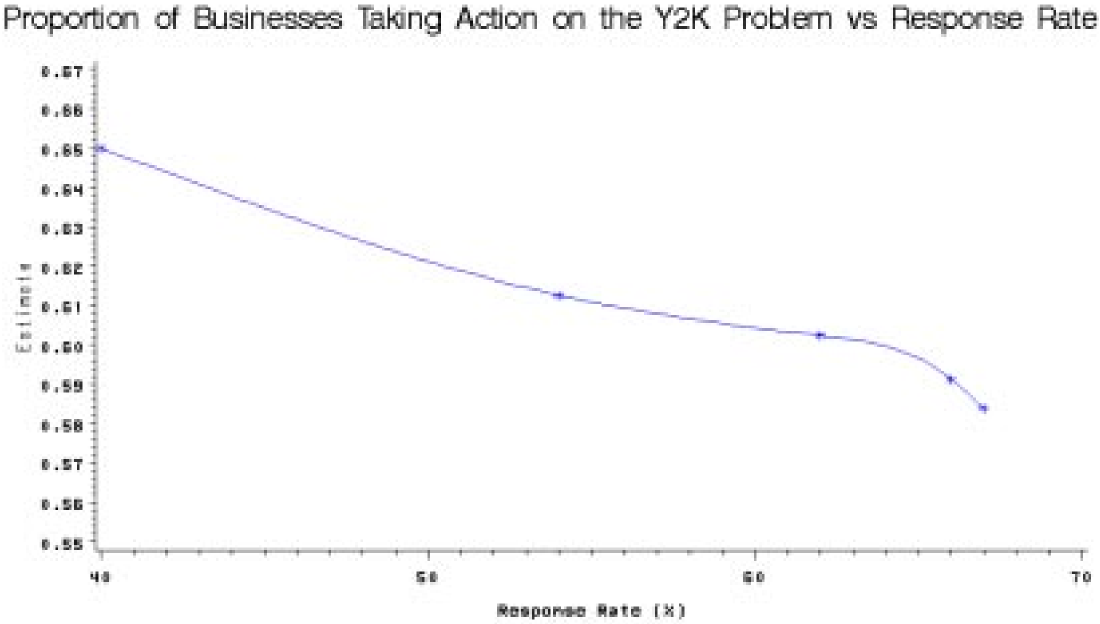
Graph 3: Estimates vs Response Rate for Employment Size Group 20-199


Proportion of Businesses Taking Action on the Y2K Problem vs Response Rate (Emp 20–199)

Graph 4: Estimates vs Response Rate for Employment Size Group 200+


Proportion of Businesses Taking Action on the Y2K Problem vs Response Rate (Emp 200+)

Graph 5: Estimates vs Response Rate for Australian Level


Proportion of Businesses Taking Action on the Y2K Problem vs Response Rate
Estimate
Response Rate (%)