SPSS®

white paper

Missing data: the hidden problem

Draw more valid conclusions with
SPSS Missing Data Analysis

Just about everyone doing analysis has some missing data, especially survey researchers, market researchers, database analysts, researchers and social scientists. Missing data are questions without answers or variables without observations. Even a small percent of missing data can cause serious problems with your analysis leading you to draw wrong conclusions. This white paper presents a case study demonstrating how missing data can affect your analysis and the decisions you make based on your results. It uses SPSS Missing Value Analysis to overcome a missing data problem to make better decisions.

Missing data occurs for many reasons

In surveys, missing data can be caused by many things. Respondents may refuse to answer a question because of privacy issues. Or, the person taking the survey does not understand the question. Perhaps, the respondent would have answered, but the answer he or she might have given was not one of the options presented. Or, perhaps there wasn't enough time to complete the questionnaire or the respondent just lost interest. Every survey question without an answer is a missing data point.

Missing data can be treacherous because it is difficult to identify the problem

Besides survey data, research data are also prone to missing data. Missed observations may occur due to human error. For example, a researcher may forget to take a measurement such as the patient's pulse. Or, a test tube can be broken which eliminates a measurement.

Databases also have missing data. Whenever there is a mismatch of variables between databases, there are missing occurrences. For example, a database analyst is analyzing sales databases consolidated from three regions. If the central region did not record one variable such as the educational background of sales representatives, then this variable would have missing occurrences when the three databases are merged.

Missing data can be serious

Missing data can be treacherous because it is difficult to identify the problem. You can't predict when missing data are problematic because sometimes your results are affected and sometimes they are not. Also, it is not always obvious when missing data will cause a problem. Each question or variable may only have a small number of missing responses, but in combination, the missing data could be numerous. Only thorough analysis on your missing data can determine whether missing data are problematic. Until now, this analysis has been time consuming and error prone.

Missing data can cause serious problems. First, most statistical procedures automatically eliminate cases with missing data. This means that in the end, you may not have enough data to perform the analysis. For example, you could not run a factor analysis on just a few cases. Second, the analysis might run but the results may not be statistically significant because of the small amount of input data. Third, your results may be misleading if the cases you analyze are not a random sample of all cases.

Not enough data

Most statistical procedures usually eliminate entire cases whenever they encounter missing data in any variable included in the analysis. For example, a regression analysis to predict home ownership based on age and educational background would ignore all cases where either of these variables had a missing response (Figure 1). So, although each individual variable may only have a small percent of missing data, when examined in combination, the total number of cases in the analysis is reduced drastically.

Missing data can also lead to misleading results by introducing bias

Case	Age	Gender	Home	Education	Occupation
1	.	Female	No	16	Non-professional
2	22	Male	No	.	Non-professional
3	39	Male	.	20	Professional
4	.	Female	Yes	.	Professional
5	40	.	Yes	16	Non-professional
6	22	Female	No	16	.
7	35	Male	Yes	18	Professional
8	39	Male	Yes	20	Professional

Figure 1. In this data set, data are missing for some respondents as shown by the ".". Any case with either missing data in age, home or education is ignored in the analysis. So, only respondents 5-8 would be included in your analysis which would drastically reduce the amount of data to be analyzed and increase the risk of a costly error.

Misleading results

Missing data can also lead to misleading results by introducing bias. Whenever segments of your target population do not respond, they become under represented in your data. In this situation, you end up not analyzing what you intended to measure. For example, suppose you surveyed a group of customers, but many people refused to answer the question about their age. If you calculate the average age based on the data you have, you would conclude that the average age of your customers is 39 (Figure 2). However, some segments of customers may be under repre-

Case	Age	Gender
1	.	Female
2	.	Male
3	39	Male
4	.	Female
5	42	Male
6	.	Female
7	37	Male
8	39	Male

Figure 2. Missing data can impact your results. Here, the average age is 39 when respondents with missing data are ignored.

Case	Age	Gender
1	21	Female
2	22	Male
3	39	Male
4	20	Female
5	42	Male
6	18	Female
7	37	Male
8	39	Male

Figure 3. With complete data, the mean age is 29 - a difference in generation. Missing data can seriously impact the conclusions you draw from your data.

sented, so this conclusion could be incorrect. If every customer reported their age, you might get different results. For example, if those who did not respond are younger, the actual average age of your customer base is 29 (Figure 3).

For this question, the "youth" segment of your customers is "under represented" and your conclusion would have been incorrect.

Case study

A consumer goods company's primary source of customer information is a few survey questions on the warranty card returned by customers. The warranty card survey collects data on age, occupation, gender, marital status, family size and income. The marketing department wants to analyze these data to better understand the demographics of their customer base to more effectively target promotions. First, the analysis is performed ignoring missing data, and then rerun taking missing data into account to easily compare the impact missing data had on the results.

SPSS Missing Value Analysis determines whether the missing data are problematic and may affect the results

Traditional analysis

To determine the demographics of our largest customer segment, examine how many customers comprise different groups defined by combinations of the demographic characteristics: gender, occupation, marital status and income. By looking for the group with the largest percent (Figure 4), it is discovered that most of the customers are married women: 38.4 %. Within this subgroup, there is an equal split among professional and non-professional women, so it is not necessary to further define customers by occupation. Using the mean and range of income for this group, these customers earn between $18,000-$42,000 a year. In summary, the largest customer segment can be defined as married women earning $18,000-$42,000 a year.

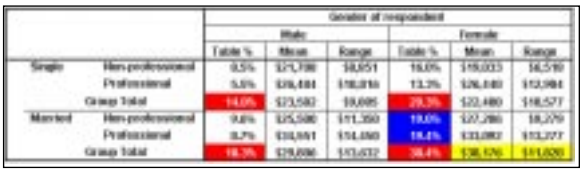

Figure 4. By examining customers by demographic characteristics, the largest segment, 38.4%, is married women. Within married women, there is about an equal split between professional and non-professional occupations.

But is this a valid conclusion? This analysis did not account for missing data - questions on the warranty card that some people did not answer. Further analysis with SPSS Missing Value Analysis determines whether the missing data are problematic and may affect the results.

Explore missing data

The missing data analysis begins by investigating the extent of missing data. A summary table (Figure 5) gives an overview of the responses for each question. The question with the highest rate of missing data is income. In fact, 34% of the customers returning the warranty card did not answer this question. Such a large percent warrants further investigation.

Figure 5. 34% of respondents did not answer the question about income.

Although the other questions have only a small percent of missing data, these missing data might still have an impact on the analysis. There might be combinations of questions which customers did not answer, or only certain types of customers did not answer the questions. Either case could cause under-representation of certain groups in the data. Investigating which questions were not answered together, and counting the number of people who responded to each combination give a good indication of where to focus the analysis of missing data. Thirty-one customers (Figure 6) did not report occupation and income. But, 257 people (Figure 6), or almost one third of the customers, left the question about income blank. Missing income data appears to be a potential problem.

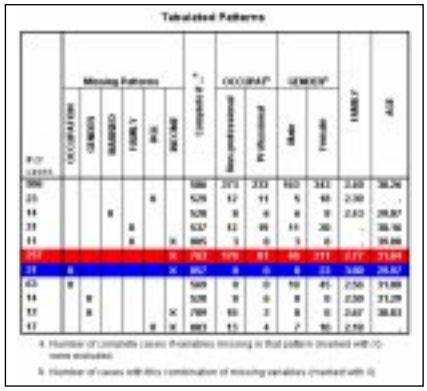

Figure 6. The most common missing data pattern is "income" with 257 of the 596 respondents not answering. 31 respondents did not answer either the "income" or the "occupation" questions together.

Since gender is a key demographic characteristic of the customer base, it is important to know how males and females do not respond to the income question. The diagnostic report (Figure 7) shows that many more women than men did not answer the question on income. This could be problematic since women are our primary target market.

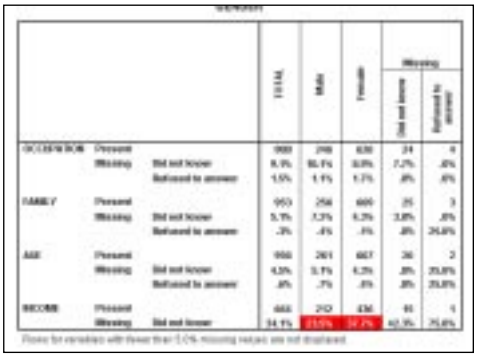

Figure 7. Many more women, 37.7%, than men, 23.5%, did not answer the question about income.

A better approach to missing value imputation is based upon the statistical method of maximum likelihood

Take action

To compensate for the under-representation of women, the missing data for these respondents can be replaced with statistical estimates of what they would have answered. One popular naïve method is mean substitution. That is, take the average income for those who answered the question and plug it into every case with a missing value for income. Then, if listwise deletion is employed, you will have more data to analyze than if the missing values had been left in place. However, mean substitution in general cannot be recommended. It is easy to see that if the mean is substituted in more than a handful of cases, then surely this adversely affects the estimated variance or standard deviation of the variable in question. Beyond that, estimated covariances and correlations involving that variable are also adversely affected. Therefore, any subsequent analysis such as regression or factor analysis is suspect.

A better approach to missing value imputation is based upon the statistical method of maximum likelihood. Statisticians employ maximum likelihood methods as a general approach to develop estimators with desirable properties. Applied in the context of missing values, the researcher assumes a model for the distribution of the data in the absence of missing data, and a model for the missing-data mechanism. For example, the researcher

might assume that the data are multivariate normal, and that the missing data mechanism is Missing Completely at Random (the pattern of missingness is random and independent of the data values of any of the variables). These assumptions should be assessed so far as is possible, but making them provides a way to getting "good" imputed values.

The SPSS Missing Values Analysis module provides two methods for maximum likelihood estimation and imputation. First, the EM (Expectation-Maximization) algorithm is an iterative algorithm that can provide estimates of statistical quantities such as correlations, or imputed values for missing values, in the presence of a general pattern of missingness. Second, Regression imputation relies on the fact that the EM approach is mathematically very similar to using regression to fill in the missing values using predicted values from a regression of a given variable on other variables in the analysis. Imputing regression predictions in this fashion can underrepresent the variance of the variable in question, so one might "jitter" the predicted values by adding a random component to the values. Either or both of these methods can be tried on your data using SPSS Missing Values Analysis. You can inspect the results, and in general you expect them to perform similarly. In general, either of these methods are superior to naïve approaches such as listwise deletion, pairwise deletion, or mean substitution.

The SPSS Missing Values Analysis module provides two methods for maximum likelihood estimation and imputation

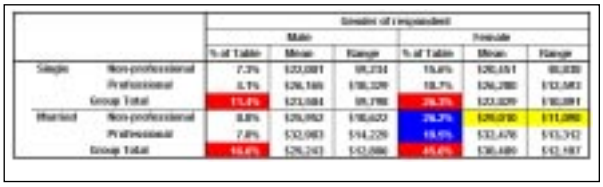

Figure 8. The largest customer segment, 45.6%, is married women. Of this group, there are more non-professional, 26.2%, than professional, 19.5%, occupations.

Compare results

Rerun the initial report used to identify the demographic characteristics of the largest customer segment using the new "completed" data. Look (Figure 8) at the subgroup with the largest percent. Married women are still the biggest segment of our customer base. In fact, they comprise 46% of the total which is up from the 38% we previously thought. Look at the break down of occupations for this group. Now there is a difference among the women's occupations. Non-professional married women makeup 26% of our customers, and 19% are married professional women. Clearly, most of the customers are non-professional married women, so this is our largest customer segment. Using the mean and range of income, they earn between $17,000-$39,000.

Better results translate to better decisions

Compare this conclusion to the customer description prior to analyzing the missing data. The new result is more focused and precise by narrowing the income range and including occupation as another dimension.

Incomplete data

- Married women
- $18,000-$42,000 income range
- Any occupation

Complete data

- Married women
- $17,000-$39,000 income range
- Non-professional occupation

While these difference may appear subtle, they translate into significant cost savings. Consider the case where the marketing department purchases a list for a direct mail promotion using these demographics about the target market.

The cost per piece is $1.00. Basing the list selection criteria on the incomplete data, more names are mailed to who are not in the primary target market. This results in a poorer response rate. Using the complete data, a more precise list is purchased. When the list is more precise, there is a higher return and fewer wasted promotion dollars. Marketing dollars are not wasted promoting the product to the less likely target market: professional women. Therefore, the direct mail campaign based on complete data is more profitable.

Incomplete data	**Complete data**
100,000 names purchased	100,000 names purchased
Total cost for mailing: $100,000	Total cost for mailing: $100,000
Response rate: 2%	Response rate: 4%
Total number of responses: 2000	Total number of responses: 4000
Closed sales: 1000	Closed sales: 2000
Average revenue per sale: $100	Average revenue per sale: $100
Total revenue: $100,000	Total revenue: $200,000
Profit: $0	Profit: $100,000

SPSS Missing Value Analysis

In this case study, missing data did in fact affect the analysis and results. By thoroughly analyzing the missing data and imputing the missing data, a more valid conclusion was reached. SPSS Missing Value Analysis provides the tools needed to diagnose missing data and take action.

About SPSS

SPSS Inc. is a multinational software products company that delivers statistical product and service solutions for survey research, marketing and sales analysis, quality improvement, scientific research, government reporting and education. Primary product lines include: SPSS for a variety of business solutions, SYSTAT and BMDP for scientific analysis, and QI Analyst for manufacturing and quality improvement applications. More than 2 million people worldwide use SPSS products.

Chicago-based SPSS has sales and support offices and distributors worldwide. In 1995, SPSS completed the best year in its 28-year history with total revenues of $63 million.

SPSS software operates on most models of all major computers. It is widely used on personal computers running Microsoft® Windows® and Windows 95. Versions for the Power Macintosh® and many UNIX® platforms are also available. In addition, many products are offered in Catalan, French, German, Italian, Japanese, Spanish and traditional Chinese.

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at **http://www.spss.com**

SPSS Inc. United States and Canada	+1.312.329.2400 Toll-free: +1.800.543.2185
SPSS Federal Systems (U.S.)	+1.703.527.6777
SPSS Argentina srl.	+541.816.4086
SPSS Asia Pacific Pte. Ltd.	+65.392.2738
SPSS Australasia Pty. Ltd.	+61.2.9954.5660 Toll-free: +1800.024.836
SPSS Bay Area	+1.415.453.6700
SPSS Belgium	+32.162.389.82
SPSS Benelux	+31.183.636711
SPSS Central and Eastern Europe	+44.(0)1483.719200
SPSS France SARL	+33.1.4699.9670
SPSS Germany	+49.89.4890740
SPSS Hellas SA	+30.1.7251925
SPSS Hispano-portuguesa S.L.	+34.1.447.37.00
SPSS Ireland	+353.1.66.13788
SPSS Israel Ltd.	+972.9.9526700
SPSS Italia srl	+39.51.252573
SPSS Japan Inc.	+81.3.5466.5511
SPSS Korea	+82.2.552.9415
SPSS Latin America	+1.312.494.3226
SPSS Malaysia Sdn Bhd	+603.704.5877
SPSS Mexico Sa de CV	+52.5.575.3091
SPSS Newton	+1.617.965.6755
SPSS Scandinavia AB	+46.8.102610
SPSS Schweiz AG	+41.1.266.90.30
SPSS Singapore Pte.	+65.2991238
SPSS Taiwan	+886.2.5771100
SPSS UK Ltd.	+44.1483.719200

Printed in the U.S.A MISSINGP-1097M

Real Stats. Real Easy.℠