14 Chapter Statistical Techniques -- III

Chair: Sylvia Kay Fisher, Bureau of Labor Statistics

Thierry Delbecque ◆ Sid Laxson ◆ Nathalie Millot

Jai Choi

Richard Wendt ◆ Irene Hall ◆ Patricia Price-Green
V. Ramana Dhara ◆ Wendy E. Kaye

14

Chapter

Statistical Analysis of Textual Information

Thierry Delbecque, Sid Laxson, and Nathalie Millot, Infoware, Inc.

Abstract

Open-ended questions are a very important information source in marketing research, quality management, psychology, and survey analysis. It is critical that analysts have data mining tools that allow them to extract valuable information.

The techniques utilized to analyze complex textual information must incorporate advanced information technology in data management, linguistics, and statistics.

STATlab exploratory data analysis integrates a Natural Language Processor (NLP) module that allows users to easily analyze textual information. This module has been especially designed to meet the unique requirements for survey analysis. STATlab NLP capabilities include:

❒ text reduction, using lemmatization;

❒ filtering of the key-words;

❒ systematic counting of the significant terms; and

❒ systematic recoding and customizable recoding of the presence of terms, by creating numerical variables.

Illustrations of the use of these techniques, with classical data analysis methods such as correspondence analysis and clustering, will be demonstrated. ■

14
Chapter

The Impact of Ratio Weighting

Jai Choi, National Center for Health Statistics

Abstract

A sample is weighted to be consistent with the population it represents. The weighting procedure is an attempt to make the sample as close to certain population characteristics as possible.

A population is made up to demonstrate the impact of ratio weighting when an attempt is made to align the sample to the selected characteristics of the population. For example, as a result of multi-stage weighting, in some surveys one sample count could represent from 100 to 120,000 -- depending on the ratios being used. The ratios vary greatly due to the nature of the characteristics. The degree to which the final weighted data approximate the true population affects the size of the variance.

The Impact of Ratio Weighting

Jai Choi, National Center for Health Statistics

Introduction

Sample units are often weighted a number of times in an attempt to make the sample compatible with the characteristics of population. For instance, the National Health Interview Survey data obtained by the National Center for Health Statistics were weighted at least five times: basic weight, nonresponses, cell counts of residential areas in Nonself-representing (NSR) PSU's, alignment of the age-sex-race cells to those of population, and expansion of the two-week reference period to 13 weeks.

Such weighting may be extended to more steps to reflect other features of population and/or correct sample biases.

The following diagram shows that a random sample S is taken from the population U:

U ______________________> S__________________________>E

A sample taken by
Simple random
Stratified
Cluster
Probability proportional
To population size
Equal
Unequal
Single stage
Multistage stages
With replacement
Without replacement

Weighted by

Basic weight: W1

Non-response adjust: W2
1st ratio adjust: W3
2nd ratio adjust: W4
Recall adjust : W5

The sample S is then weighted five times to estimate the population. We want to have the final estimate E close to population in every aspect. But final number of estimates and cell counts are not close to those of the population after these weightings. For instance, the cells of age-sex-race table have changed after each weighting. Although we want cell estimates close to those of the population, the final results are quite different, as seen in the next section.

The weighting may reduce the difference and result in reducing variance and/or bias. It depends on each particular situation in weighting. The weighting may be repeated until the difference is minimized between the population U and final estimate E not only in number, but also in cell distributions.

In the next section, the weights are estimated five times for a sample of 12 persons. The changes of cell distribution are shown after each weighting. The impacts of weighting are discussed in the following section.

Example

A population is created for illustration, from which a sample of 12 persons is taken to show the five steps of weighting in Table 1. The population of 1,600 is divided into four strata, the first stratum of 300, the second of 300, the third stratum of 600, and the fourth stratum of 400.

Table 1.--Weighting of Doctor Visits

Str	PSU	n	Basic weight	Nonresponse		1-Ratio		2-Ratio		V	2WV	52WV
				R	W2	R1	W3	R2	W4			
(1)	(2)	(3)	(4)	(5)		(6)		(7)			(8)	
I	A300	1	100	1	100	1	100	.8	80	2	160	4,160
300		2	100	1	100	1	100	1	100	0	0	0
		3	100	1	100	1	100	1	100	0	0	0
II	B300	4	100	3/2	150	1	150	1	150	1	150	3,900
		5	100	n-resp								
300		6	100	3/2	150	1	150	1	150	0	0	0
III	A200	7	150	1	150	1	150	1	150	0	0	0
		8	150	1	150	1	150	1	150	3	450	11,700
600*	B200	9	150	2	300	.9	270	.8	216	1	216	5,616
		10	150	n-resp								
VI	A100	11	200	1	200	1	200	1	200	2	400	10,400
400*	B100	12	200	1	200	1	200	1	200	0	0	0
1600	**1,200**	**12**	**1,600**		**1,600**		**1,570**		**1,496**	**9**	**1,376**	**35,776**

*NSR-stratum R=ratio str=strata sp=sample
V=2 weeks Doctor visits
2WV=weighted visits for 2 weeks recall
52WV=weighted visits for 52 weeks.

The first two are self-representing PSU's, while the last four are the nonself-representing (NSR) ones. No sampling is involved in the first two PSU's got the first stage. Two PSU's are selected out of the three PSU's, each with 300 persons, in the stratum III by equal probability, and two PSU's taken from the four PSU's of equal size (100) in the stratum VI.

The weights given to these six PSU's are 1, 1, 3/2, 3/2, 4/2 and 4/2 in the respective stratum.

A sample of 12 persons is selected from the six sample PSU's by simple random sample, 3 from each of the first two PSU's, 2 from each of the third and fourth PSU's, and 1 from each of the last two PSU's.

The weights are 300/3, 300/3, 200/2, 200/2, and 100/1, 100/1 for the selection of persons from the respective sample PSU.

The basic weights are shown in the 4th column, and they are the multiplications of the two weights arising from the selection of PSU's and persons. The basic weights are 100, 100, 150, 150, 200, and 200 for the respective PSU.

The nonresponses are adjusted within the PSU, and a nonresponse ratio, used to adjust the missing numbers, is the sample persons divided by the number of respondents within the PSU. The fifth and sixth column shows the 5th and 10th samples did not responded, and adjusted accordingly.

The living areas in NSR-PSU are divided into three cells of city, urban and rural places. Six sample persons, 7, 8, 9, 10, 11 and 12, are from the NSR-PSU's in the strata III and VI, shown in Table 2; hence, they are the subject of the first stage ratio estimation. The cell ratios of population to the sample estimates are 1.0, 0.9, and 1.1, and used for the first stage ratio estimation. Since the ratio is 0.9 for the second cell, the 9th sample requires the adjustment, while no adjustment is needed for the remaining 5 sample persons in the first cell as seen in column 6, Table 1, for the lst Ratio estimation.

Table 2.--1st Ratio in NSR-PSU's

Cell	1 City	2 Urban	3 Rural
Population	510	90	100
Estimation	510	100	90
1st ratio (sample no.)	1(7,8,10,11,12)	0.9(9)	1.1(0)

Each of the 12 persons belongs to one of the 8 age-sex-race cells. Table 3 consists of the three tables of eight cells for population, estimation, and the ratios. The ratios in the last table are population divided by estimates, and shown in column 7 in Table 1. They are used for the second ratio estimation.

Table 3.--2nd Ratio of Population, Estimation

Cell	Age	Male-White	Male-Black	Female-White	Female-Black
Population	1-49 yr	350	40	350	60
	50+ yr	350	60	350	40
Estimation	1-49 yr	350	50	350	50
	50+ yr	350	50	350	50
Ratio	1-49 yr	1(5,8,10)	0.8(1)	1(2,3,7)	1.2(-)
	50+yr	1(4,11)	1.2(-)	1(6,12)	0.8(9)

Table 4 shows the weighting process of one doctor visit of the 9th sample person. The weight of this person is changed five times, W1 through W5.

The W1 is 150, the sample persons selected from two stages, the first stage of selection of two PSU's out of three PSU's of equal size. Two persons were sampled from each sampled PSU. The basic weight is the product of these two weights, i.e., 150 = (3/2) x (200/2).

The W2 is the number adjusted for the nonresponse. Since one of the two sampled persons in the same PSU did not respond, the weight of the respondent is doubled (300 = 2 x 150) to cover the nonrespondent.

Table 4.--The Changes of Weight for One Visit of the 9th Sample Person

W1(basic)	W2(n-resp)	W3(1st ratio)	W4(2nd ratio)	W5(52 wks)
150	300	270	216	5,616
1/(p1 P2)	W1 x 2/1	W2 x 0.9*	W3 x 0.8	W4 x 26

The W3 is 270 from the first ratio weighting (270 = 0.9 x 300). As this sample person lives in urban area, her first stage ratio is 0.9 as shown in Table 2.

The W4 is 216 by the second ratio weighting (216 = 270 x 0.8). Since this sample belongs to the cell (2,4), the black female of 50+ years, her second ratio is 0.8 for her age-sex-race. She represents 216 people for her stratum, PSU, residential area, and age-sex-race class.

The W5 are the estimated number of visits for 52 weeks or one year. She visited the doctor's office once during the past two weeks, and her one visit became 5,616 visits (= 26 x 216) for 52 weeks as shown in the column 8.

Each sample in Table 1 is weighted the same way. The 9 visits from 12 sample persons became 35,776 visits after the sample visits were weighted five times.

The eight cells in age-sex-race table are the basic weights, W1 and they have been changed three times through the three weighting processes as seen in Table 5. After these weightings, the last row of W4 is quite different from that of of the population. This difference is mainly due to sampling, non-responses of the samples 5 and 10, and the first ratio adjust of the sample 9, and the second ratio adjust of the samples 1 and 9.

Similarly, the 3 residential cells of population in NSR-PSU's differ from those of the last estimates in Table 6. This difference is also due to the sampling, empty cell, and first and second ratio adjustments of the ninth sample.

Table 5.--Eight Cells of the Age-Sex-Race Table

Cell	1	2	3	4	5	6	7	8
Pop	350	40	350	60	350	60	350	40
W1	400	100	350	0	250	0	300	150
W2	300	100	350	0	300	0	350	150
W3	270	100	350	0	300	0	350	150
W4	216	80	350	0	150	0	350	150

Table 6.--Three Cells of Residential Areas

Cell	1 City	2 Urban	3 Rural
Population	300	90	100
Estimate given	300	100	90
W1	300(8,10)	150(9)	0(-)
W2	300(8,10)	150(9)	0(-)
W3	270(8,10)	150(9)	0(-)
W4	216(8,10)	150(9)	0(-)

Remarks in Weighting

In the process of ratio weighting, we observed that each step of weighting may reduce or increase the differences between the estimates and population. This may also increase relative bias and/or variance, depending on the specific situations in sampling and weighting. Each step may have contributed to the estimation, as discussed on the following page.

❐ The Basic Weight (W1)

There are many ways to select a sample. For instance, if population were structured in three stages, and the sample taken by pps design, the variance would be minimized. Thus, the basic weight is decided by sampling design.

If a sample is randomly selected, the basic weighting may reduce the relative variance, while a non-random design might increase the variance and bias dramatically.

There may be empty cells when the sample persons are distributed over the cells in a large table. This may happen more likely for a small sample.

❐ Nonresponse Adjusted Weight (W2)

Nonresponse may be adjusted at a proper stage or stages. If nonresponses arise randomly and the nonresponse rate is low, the ratio adjustment may be valid especially for a large sample, and bias and/or variance reduced.

On the other hand, if the nonresponse rate is more than 30 percent, the ratio estimate may cause severe biases even for a large sample.

Alternative methods may be used in order to reduce bias in the presence of high rates of nonresponse. Other methods such as regression and Bayesian methods are often useful for nonresponse estimation. But such methods usually bring problems later at the stage of data analysis.

❐ The First Stage Ratio Adjusted Weight (W3)

We often do not have enough sample persons in sparsely populated area or among specific subpopulations, such as African Americans or older people. Consequently, the small number of a sample may not reflect the characters of population. Thus, we may use the ratio between a population and its estimation.

If the previous weights were already biased, this process may further increase biases.

❐ The Second Stage Ratio Adjusted Weight (W4)

The weights from the previous adjustment may not reflect the age-sex-race cells of the population. We may multiply the ratio, population to its estimate, to the previous weights. This is done for each person in the age-sex-race cell. But the resulting cells may differ from those of the population due to the empty cells, small sample, and the previous weighting. Although this process reduces the difference between the population and estimate in the age-sex-race cells, it may make the difference greater for the cells of living areas, which one may like to avoid.

❒ The Recall Adjusted Weight (W5)

The number of doctors' visits in the past two weeks is only 1/26 of one year; hence, we multiply 26 to make the previous weights to be the visits for one year.

The resulting number of visits per year may mislead readers for a calendar year, as two weeks could be extended to the future or past 52 weeks from the point of an interview. In this case, the visits may be counted to a different year, depending on the date of an interview.

If the nonresponse was already biased, the recalls adjust may further increase the bias.

Comments

The ratio method does not create new estimates for empty cells in a age-sex-race table. Unless we use the estimates for empty cells, no improvement can be made. However, we may put one in an empty cell for estimation or we may incease sample size if it is possible.

The high rates of nonresponses may cause bigger biases, especially when the units in the PSU are different.

The ratios may be unstable for a small sample. Since a small sample may leave more empty cells, large biases may be introduced, and nonresponse may cause more problems.

The order of weighting also has influence on the final outcome of a table. If the order of weighting were changed in the previous example, or the age-sex-race adjusted first and then residential area in NSR-PSU's, the result would be quite different. One may do the most important adjustment at the last stage.

The above example illustrates the difficulty to estimate population by ratio weighting to satisfy all of its aspects. In order to reduce the difference between the population and estimate in the age-sex-race as well as in residential areas, we may repeat steps from the first ratio W3 to the second ratio W4, leaving W1, W2, and W5 out, and stop when the difference between population and final estimates is minimum for both tables. Each time a new ratio table is created from the ratios between the population and new estimate of W3 or W4.

The ratio estimation may work better if no cells were empty, response rates high, sample size reasonably large, and cell members homogeneous. ■

Fitting Square Text Into Round Computer Holes -- An Approach to Standardizing Textual Responses Using Computer-Assisted Data Entry

Richard D. Wendt, Irene Hall, Patricia Price-Green,
V. Ramana Dhara, and Wendy E. Kaye,
Agency for Toxic Substances and Disease Registry

14
Chapter

Abstract

Textual responses in data collection efforts present major programming and analysis challenges. One type of text response that poses a problem in the data collection efforts of the Agency for Toxic Substances and Disease Registry's (ATSDR) Hazardous Substances Emergency Events Surveillance is chemical names. Many different chemicals are used and released into the environment in the United States each year. Additionally, varied chemical names, trade names, and mixtures add to the difficulty of establishing a uniform naming convention. Existing naming conventions or coding schemes (for example, Chemical Abstract Service Registry Numbers) are often too complicated for use by data entry personnel. Additionally, for many chemicals no codes are available. Analysis efforts involve not only the identification but also the classification of chemicals released during emergency events. Standardizing the names of chemicals is necessary to automate the analysis process.

To solve these problems, ATSDR has created a semiautomated chemical selection system that combines chemical names previously supplied by the users and chemical names supplied by ATSDR into a single database. This data entry system incorporates chemical names and chemical category assignments from previous data collection years. Users scroll through a window containing the list of chemical names and select the substance of interest. When a user selects a chemical name, the computer stores the associated chemical in the appropriate data field. A search function allows the user to locate chemicals by typing the first few letters of the desired name.

This feedback system minimizes the use of different names for the same chemical by basing chemical name selections on chemical and substance names in previous event reports and names supplied by ATSDR. This increases chemical name standardization. Since users are more likely to select names from the menu, this method reduces the workload of ATSDR staff and increases the consistency of chemical categorization.

Fitting Square Text Into Round Computer Holes -- An Approach to Standardizing Textual Responses Using Computer-Assisted Data Entry

Richard D. Wendt, Irene Hall, Patricia Price-Green,
V. Ramana Dhara, and Wendy E. Kaye,
Agency for Toxic Substances and Disease Registry

Introduction

Since 1990, the Agency for Toxic Substances and Disease Registry (ATSDR) has maintained the Hazardous Substances Emergency Events Surveillance (HSEES) system. This epidemiologic surveillance system currently tracks hazardous substance releases in 14 states. HSEES allows health officials to evaluate both the nature and extent of hazardous releases (both threatened and actual) and their effects on public health. The HSEES system is an active surveillance system. Participating state health departments use a variety of reporting sources (for example, individuals, state environmental protection agencies, newspapers, police, fire departments, and hospitals) to collect HSEES information on a data collection form. Information from the data collection form is entered into a computerized data entry system that is a simulation of the paper data collection form. Participating state health departments transmit this information toATSDR quarterly.

Although most data are categorical in nature and easy to code uniformly, there are some textual responses that require special treatment, including descriptions of the type of industry, responses indicating "other," and chemical names. In this regard, standardizing chemical names presents the greatest challenge.

HSEES defines hazardous substances emergency events as uncontrolled or illegal releases or threatened releases of substances or their hazardous by-products. From 1990 through 1992, reportable substances included the 200 chemicals identified as most hazardous at Superfund sites. Also included were insecticides, herbicides, chlorine, hydrochloric acid, sodium hydroxide, nitric acid, phosphoric acid, acrylic acid, and hydrofluoric acid (*Federal Register*, 1988). Since 1993, all hazardous substances (except petroleum products) have been included in the HSEES definition.

Events are reported if the substance(s) must be removed, cleaned up, or neutralized according to Federal, state, or local law. Additionally, a potential release is reported if it involves one of the designated substances and if it results in an action (for example, an evacuation) to protect public health (Hall et al., 1994). Presently, ATSDR maintains a database of over 11,000 hazardous substance spills and over 13,000 chemical data records.

With so many chemical names recorded in one database, the problems associated with standardizing chemical or substance names are very large. In addition, varied chemical names, trade names, and mixtures add to the difficulty of establishing a uniform naming convention. Existing naming conventions or coding schemes such as Chemical Abstract Service (CAS) Registry numbers, Department of Transportation (DOT) numbers, Chemical Hazards Risk Information System (CHRIS), or United Nations (UN) numbers are often too complicated for use by data entry personnel. Additionally, for many chemicals no codes exist. As an example, trichloroethane can be listed four ways depending on the reporting source; as TCA, as 1,1,1 trichloroethane, as 1,1,1-trichloroethane, or as trichloroethane. All of these responses are correct names, but the text fields represent completely different answers to a computerized statistical analysis system. To the nonchemist, the choices between proper naming conventions is overwhelming.

Data Entry Method

Originally, the HSEES staff addressed this problem by creating a chemical pick-list that consisted of the 36 most reported chemical names and 2 "Other" fields (one for pesticides and another for all remaining chemicals). While this approach reduced the number of user-defined names, it was not completely successful. The problem of unique chemical naming conventions still remained a major data analysis concern.

This problem hampered efforts to analyze events by chemical substance and persisted for three reasons. First, almost one- third of the spills that were reported to the HSEES system were chemical mixtures. Second, many spills consisted of pesticides and herbicides (which may be made from complex mixtures and compounds). Third, and most significantly, the system itself was being used by ATSDR and the state health departments for two different purposes.

ATSDR staff must classify chemical names and substance names into a standardized format for analysis as part of their data processing procedures. This assists them in disseminating the public health consequences of the release events. The data entry personnel at the state health departments use chemical names as descriptions of events for other state agencies, such as emergency responders. For emergency responders, there are major differences between a pure ammonia release and a 1-percent ammonia solution release. Both spills involve the same chemical, but the level of protective measures used, the issuance of evacuation orders, the use of in-place sheltering, and the level and extent of clean-up are very different.

To address these problems, ATSDR has created a semiautomated chemical selection system that combines chemical names previously supplied by the users and chemical names supplied by ATSDR. The system incorporates data from previous years into a database file that contains a set of selected chemical names.

When users reach the chemical information data entry screen, the chemical name database file is opened. The chemical and substance names in this database are then displayed in a browse screen format. Users scroll, page, or search through the window containing the list of chemical names and select the substance of interest. At this point the computer program enters the chemical name in the appropriate data entry field. A search function allows the user to locate chemical records by typing the first few letters of the desired name.

Presently, most chemicals names in the list are selected by ATSDR staff for correct syntax, but the variety of these chemicals and substances are based on all previously reported releases. Data entry personnel are not allowed to modify these predetermined chemical names, but may still select "Other" and edit that name. As an added incentive for selecting predefined names, all CAS, DOT, CHRIS, and UN chemical codes are automatically entered and are saved to the chemical name database file. These codes are then automatically retrieved each time the user selects a predetermined chemical name.

Conclusion

This pseudo-feedback system should reduce the addition of different names for the same chemical because many selections are based on previous event reports and chemical names supplied byATSDR. It should also increase chemical name standardization. Finally, because users are more likely to select names from the menu, this method should reduce the workload ofATSDR staff and increase consistency of chemical categorization.

The full potential of this approach should be seen when it is used as a full-fledged feedback system. By incorporating all user defined chemical names into the system and translating this feedback into standardized chemical names during quarterly data processing atATSDR, the addition of new chemical names by data entry personnel can be reduced to a minimum. While this approach will require an intense amount of programming at first, as time passes the maintenance effort for the translation program will be greatly reduced. As seen from examination of previous data submissions, many hazardous substance releases are repetitive. The problem has been that the chemical name choices do not adequately reflect the idiosyncratic naming habits of each user By tailoring the data entry system to each user's response, the use of "Other" as a free-form text input selection should be reduced.

The main source of nonstandard chemical names will then come from either truly unique hazardous releases or data input from new states as they are added to the surveillance system. HSEES staff are currently evaluating the need for implementing a fully functioning feedback system and will decide on its development once sufficient data has been collected.

References

Federal Register (1988). Hazardous Substances Priority List, 53, 41280-41285.

Hall, I. H., Dhara, R. V., Kaye, W. E., and Price-Green, P. (1994). Surveillance of Emergency Events Involving Hazardous Substances -- United States, 1990-1992. *Morbidity and Mortality Weekly Review*, 43 (suppl S-2), 1-6. ■