JOBS

Journal of
Obnoxious Statistics

An International Unica published by TT-Publikaties
Amsterdam
http://www.xs4all.nl/~edithl/jobs.htm
August 2005

Publisher: TT-Publikaties
Chairman of the Board: Svante Öberg

Mailing address: edithl@xs4all.nl

Regular mailing address:
Journal of Obnoxious Statistics
Plantage Doklaan 40
Nl-1018 CN Amsterdam
Netherlands

Telephone:
Chief Editor + 31 20 6223438
Fax: + 31 20 3302597

Chief Editor: Edith D. de Leeuw e.deleeuw@fss.uu.nl
Book Review Editor: Joop Hox j.hox@fss.uu.nl

Associate Editors:
Gunilla Dahlén gunilla.dahlen@scb.se
Lilli Japec lilli.japec@scb.se

Limited Edition
This issue of JOBS was printed in a limited edition of 100 copies, of which the first 50 are numbered and signed by Lars Lyberg personally.

This issue is also available at: http://www.xs4all.nl/~edithl/jobs.htm

This is number: _____

_____ **(Lars E. Lyberg)**

Contents

Journal of Obnoxious Statistics
TT-Publikaties, Amsterdam
ISBN -10: 9080770825
ISBN-13: 9789080770829

Preface

The Journal of Obnoxious Statistics (JOBS) is created specially for Lars Lyberg by his friends and colleagues to commemorate the 20th anniversary of the Journal of Official Statistics (JOS).

It is a *Liber Amicorum* in which we honour and thank Lars Lyberg as friend and as editor. It is also a spoof, which means that the contributions are funny (at least we hope so), but not necessarily scientific, nor reproducible. In fact our main inspiration was the Annals of Irreproducible Results.

We warn the readers explicitly and urge them not to believe a single word they read in JOBS. Also, any resemblances to existing persons, organizations and journals are purely fictional.

We now run out of warnings, but not out of thanks, and want to thank all collaborators to this special issue, who wrote their contributions with intelligence and humor in honour of the greatest editor of all: Lars Lyberg.

Amsterdam, April 2005

Edith de Leeuw
Joop Hox

Introduction to a Very Special Issue

In this special edition of the Journal of Obnoxious Statistics (JOBS), we have collected a wide range of articles that deal with not so important problems that will face official statistics in the coming decade. In unofficial statistics, however, these problems will be a major challenge.

This JOBS issue covers topics that range from new areas of statistics to data analysis. The careful reader will notice that although most of the articles might seem rather doubtful at first they will in fact make perfect sense after reading them many times or late at night. The O'Rourke, Cantwell and Dale, Haraldsen & Kleven articles show how illuminating statistics can be when applied in new areas. They unravel many interesting relationships such as the one between men's first name and IQ level. Harkness, Bates and Biemer illustrate the importance of thinking outside the usual paradigm in their innovative research on questionnaire design and testing. For example, who would have thought that monkeys have such a talent for questionnaire design?

The ultimate objective of survey research is to identify strategies that can increase data quality, reduce survey costs or both. In this issue of JOBS there are many hands on recommendations on how to achieve these goals, for example Campanelli's article on how to train interviewers, and the Beerten & Lynn article on methods for longitudinal surveys. The perhaps most promising method for reducing survey costs is described by Couper. He reports from his CRAPI experiment about a data collection method that can replace the work of interviewers and cut survey costs substantially.

Despite the vast literature that deals with survey errors, there are still some gaps to be filled. In this special issue of JOBS we address these gaps. For example, the underestimation of walking the dog trips in travel surveys (Cohen), the problem of asking a question that accurately estimates household size in reincarnation surveys (Relhom) and the problem of compulsive liars (Andersen, Elvers, Jorner, Münch-Hausen & Vantroen). We also take a closer look at nonresponse errors. Groves looks really deep and sees nonresponse errors in surveys of the dead. Stoop identifies a relationship between dog ownership and the willingness to participate in surveys and Loosveldt pins down the nonresponse problem with a theory about a nonresponse gene. O'Muircheartaigh puts the nonresponse problem into a broader perspective by considering measurement errors in the absence of measurement. Survey errors are not only present in national surveys but also in cross-cultural surveys (Carey and De Leeuw & Japec). The successful Dutch and Swedish cooperation (De Leeuw & Japec) reports on the ultimate solution to the nonresponse problem in cross-cultural surveys. Their method is also theoretically supported in Wretman's article. He shows that we get a better overall accuracy if we leave the nonrespondents alone (and maybe also the respondents).

All respectable journals need articles on quality management. In this special JOBS issue we have five articles on this topic. Scheuren untangles the mystery of different types of data in his article "Plain Old Data, Para Data, and Meta Data: The Three Sopranos of Data". Körner gives a German perspective on quality. Grünewald & Lindén discuss problems that can occur when a statistician have different roles to play. They describe how roles such as survey specialist, professor in statistics, and international statistician might affect and change the view on quality. We also present two examples of quality control; Thorburn looks at coding control and Marker & Morganstein report on using control charts to measure LARS´quality.

In this JOBS issue we also report on some very innovative methods for imputation and variance estimation. Lynn summarizes the limitations that researchers and data users have found with existing imputation methods, as "It's just difficult". He presents the pure imputation method, an extension of the WILD-GUESS method. Fienberg & Straf report on a new imputation method that can replace traditional sample surveys. Researchers that have NO-CLUE, as a standard imputation method, will definitely benefit from these two articles.

"Reinvent a wheel that has always, within predictable limits, worked" is the philosophy used by Mintcoin & Clitsin. They describe the Crossbow procedure for estimating precision in a survey. Rust & Brick propose the Jackknifed Bootstrap Method for variance estimation and they present some interesting results from the Italian Retail Footwear Survey.

Data access is a topic that has been discussed extensively for many years. Singer recognizes the importance of data access for a society to function effectively and gives useful recommendations on how to make confidential data more accessible to research. Tucker's analysis is an example of how access to confidential data can increase our knowledge. He uses linear regression to predict homerun hitting in baseball. The best predictors are steroid use, hot dog and beer consumption. The model can be used to adjust data so that accurate comparisons of players that play in different era can be made.

We end this special issue with a book and software review. Broadsaw has read Donna Dillman´s book "Why Male Surveys Do Not Work: The Total Disaster Method". She recommends it to all male survey researchers. Female researcher however, will still have to do with Don Dillman's old book on the Total Design Method. Addams reviews the software DeSade. He finds the function called *damn,* that automatically removes all nonsignificant data, to be particularly useful and recommends this software package to applied statisticians.

In spite of the substantial effort all contributors have made, I fear that this special edition will not contribute much to the current state of art in official statistics. I am, however, confident that it will contribute to a remarkable improvement of methodology used in unofficial obnoxious statistics. I would like to thank all the contributors for their excellent articles and the large number of referees that have reviewed all submissions to ensure that they are obnoxious enough to qualify for this special edition. I would also like to thank the Chief Editor Edith D. de Leeuw and Review Editor Joop Hox for their patience reading and editing the material and the Associate Editors Gunilla Dahlén and Lilli Japec for their continued moral and editorial support. Last but not least I would like to thank my dear friend Lars Lyberg, who has inspired all of us to put in some extra creativity to make this special edition possible.

Stockholm, April, 2005

Svante Öberg
Director General
Statistics Sweden

The Importance of Being Lars

Diane O'Rourke[1]

This article looks at past research on the consequences of first names and differences between those with various first names. It adds to the literature by reporting the results of a study comparing those named Lars versus other names in Sweden and in the state of Minnesota, U.S.A. Names in Norway were not included in this study, for obvious reasons.

Key words: Names; first names; names; labels; Lars.

1. Introduction

People have had first names since the beginning of time. In fact, surnames or last names were not added until the number of people in a location grew large and some began to travel to other sites. They then began to be known by the location of their birth or the name of their father (e.g., Joseph of Jerusalem, James Anderson). However, first names have continued to play a prominent part of a person's life.

People with different first names can be classified and compared on a number of different characteristics. For example, in 2004, in a large U.S. study of men ages 50 to 60 conducted for the Rush Limbaugh Institute, it was found that those named George had an average IQ of 93, while those named John had an average of 138 (Luntz 2004).

An early study of names was conducted by Johanson and Anderson (1955) who were particularly interested in men named Lars. They studied college men in Sweden and found that those named Lars were rated by their peers as more handsome, smarter, more virile, more liked, and more likely to succeed. However, the results raised some suspicions, the authors were accused of bias and self-interest, and the results were not replicated in subsequent research. Critics blasted the validity of research on the name Lars conducted by men named Lars. Subsequent work on social behaviors was inconclusive (Hefner 1980). More recent works on related topics were questionable (Rumsfeld 2004; Bush 2004).

2. Methods

The research reported here looks at men named Lars versus those with other names. It is based on a study of men ages 50 to 60 in both Sweden and Minnesota, U.S.A., where there are many Swedish Americans and thus a large number of men named Lars. Results are based on data from random-digit-dial samples that screened for men named Lars. The

[1] University of Illinois Survey Research Laboratory.

study compares data from 1,000 men in each country – 500 named Lars and 500 with other names. Each screened respondent was sent a lengthy questionnaire, at their preference either by mail or Internet site. The questionnaire contained questions about their physical characteristics, backgrounds and preferences, including their educational, vocational, and social histories. In addition it included personality and IQ scales. The response rate for the Larses was 78% overall, for the other names it was 36% (see note on "anal tendencies" below).

3. Results

Significant differences were found for most items and in both countries, although somewhat weaker in the Minnesota sample. Larses in both countries were found to have higher IQs and more success educationally and vocationally than men with other names. This may be due to the much higher scoring of Larses on the "anal tendencies" scale inventory. Non-Lars men exhibited higher sociability quotients, although the difference was more pronounced in Minnesota than in Sweden. Despite this, however, Larses in both countries identified more "significant others and partners" than non-Larses. This association was strong in Minnesota and even stronger in Sweden. In terms of physical associations, Larses in both countries were taller and thinner than others, but exhibited less physical fitness. Unexpectedly, Larses were three times as likely to be involved with ice fishing than their counterparts in both countries. However, it also was observed that Lars' fishing was accompanied by increased vodka consumption and female companionship.

4. Discussion and Conclusions

This research found that there are cross-national differences between men named Lars and others. While this is the first academically rigorous study in this area, the authors recommend that further research be conducted to replicate these findings, particularly on younger men.

5. References

Bush, G. (2004). Who is Lars? White House Report # 2223, 2004.

Luntz, F. (2004). Georges and Johns: Is it the Name or What? Journal of Probable Conspiracies, 666 – 911.

Hefner, H. (1980). Associations with Names: The Lars Case Study. Playboy, 45, 1-12.

Johanson, Lars and Anderson, Lars. (1955). The Lars Factor: What's in a Name? Journal of Obnoxious Statistics, 5, 20 – 22.

Rumsfeld, D. (2004). The Lars Conspiracy? CIA internal memo.

Received April 2005

Sports, Reports, News, and Reviews: The Place of JOS in World Events

Patrick J. Cantwell[1]

The author presents relationships between important world events and concurrent developments at the *Journal of Official Statistics* (JOS). Evidence from the fields of sports, politics, and technology is offered to make the case for cause and effect between such events and JOS's average time to publication and level of readership.

Key words: Time to publication; Oslo Accords; Ice hockey; Internet; Sverige.

1. Introduction

It's not unusual for historic world events to be linked. But proving that one incident causes a second is difficult. Yet we present incontrovertible evidence that important episodes in the evolution of the *Journal of Official Statistics* first followed from, and later influenced the course of major world events. Such surprising interrelationships can only be fully appreciated by exploring the causes and effects uncovered below.

2. Examples to Verify the Role of JOS in World Events

The year after Sweden's outstanding performance in the 1958 World Cup football (soccer) tournament–runner-up to champion Brazil–top executives at Statistics Sweden met to discuss an appropriate way to commemorate the event. Their solution was to create a new journal to be called the *Journal of Official Statistics* (JOS). Hoping to reverse the order of the World Cup finalists, they decided to reverse the digits in the year of their near-triumph, and wait until 1985 to start JOS. As a result, for 25 years, the soon-to-be editor of the journal solicited and accepted papers, but was constrained to hold them in waiting (Lyberg 1989). By 1985, when the much anticipated first issue was published, the average time to publication was more than ten years (with a large standard deviation), a bit discouraging to authors.

In the late 1980's and early 1990's, concerned about the lengthy time to publication, the editorial board collected data on the time for reviewing submitted papers, and finalized a declaration of principles to promote expeditious review. Meanwhile, 1993 was

[1] U.S. Census Bureau, Washington, DC 20233-9100, U.S.A.

Acknowledgments: The author thanks the editor of *JOS* for taking a chance on him many years ago, and for retaining him after realizing what a mistake he had made. C'est la vie. The views expressed are those of the author and certainly not those of the U.S. Census Bureau.

the year in which the state of Israel and the Palestine Liberation Organization came together–if only temporarily–to sign the Oslo Accords, formally known as the Declaration of Principles on Interim Self-Government Arrangements (Peres, Abbas and Clinton 1994). Coincidence? Is it also a coincidence that "finalize data" is an anagram of "Intifada zeal"?

In addition to these new principles, in 1994 the editor of JOS began to assess penalties–including instant termination ("sudden death")–to members of his editorial board for late response. Together, these measures helped JOS cut its average time to publication to 58 months and 11 days. The Swedish national ice hockey team, avid readers of JOS, promptly responded with their best performance in the Winter Olympics. After winning three bronze medals in the 1980's, in 1994 they defeated the Canadian Olympic team for the gold medal on penalty shots, after tying the score with less than two minutes remaining in regulation (58 minutes, 11 seconds into the game), and playing through a scoreless period of sudden-death overtime. The coach of the Swedes stated that his team's emotional play was inspired by the hard work of the JOS staff (Forsberg and Salo 1995). It would be hard to disregard the cause and effect at play here.

In the last decade, have changes in technology motivated developments at JOS, or has JOS spurred changes in technology? For many years, the growth of the Internet had been steady but disappointing (Gore to appear). Then, at the turn of the century, JOS made the bold move to post all its issues–including the most recent one–on its website. This epochal event coincided with sudden, astronomical increase in Internet use. Although definitive proof is impossible to document, knowledgeable sources believe that a major part of the increase is due to visits to the JOS website (ibid). The volume there is so high that current attempts to measure it have failed.

3. Conclusion

Have momentous world events been influenced by the activities at JOS? While it is always difficult to prove cause and effect outside a planned experiment, we believe the evidence points overwhelmingly in this direction. Grattis p□ födelsedagen, L.L.

4. References

Forsberg, P. and Salo, T. (1995). The Relationship Between JOS and Ice Hockey in Sverige, Journal of Official Statistics, 11, 5-9.

Gore, A.A., Jr. (to appear). How I Moved the Internet into the 21st Century, Journal of Political Technology, 1, 1-58.

Lyberg, L. (1989). To Publish or Not to Publish: That is the Question. JOS d. Surv. Meth., 4-6, 7-6, 6-4.

Peres, S., Abbas, M., and Clinton, W.J. (1994). The Oslo Accords: What Went Wrong. The New Republic, April 1, 1994.

Received April 2005

The Effects of Steroid Use on Home Run Power

Clyde Tucker[1]

Sabermetricians are in a quandary. The effects of steroid use on homerun hitters today makes the comparison of their records to earlier ones very difficult. Should they all receive an asterisk next to their statistics, or could a method be developed to measure the actual effects of steroid use on power hitting? This paper offers one method, linear regression, for dealing with the problem.

Keywords: Longball, Babe, Say It Ain't So Barry, Jose Canseco (i.e., whining)

1. Introduction

The controversy concerning the comparison of baseball statistics over time actually began back in 1961 when Roger Maris hit 61 homeruns, but not in 154 games. At the time, an asterisk (since removed) was put next to his record (Vincent 2003). The crisis today concerns how to treat the homerun statistics of sluggers in the steroid era. Should their numbers stand alongside those of Aaron, Ruth, and Mays, or should they get the same treatment as Maris' record did in 1961 (Bouton 2005). On the one hand, some of the current generation's homerun hitters may have used a foreign substance, not available to those in an earlier era, to enhance their performance. Yet, at the same time, these same players also have access to much better nutrition and medical care compared to earlier generations of baseball players, and they receive much better physical conditioning. Maybe they should get an asterisk anyway.

In any case, the record keepers could use a way to separate out the effects of each of these factors on homerun hitting. Most sabermetricians, however, have not had much in the way of formal statistical training and must rely on academicians to show them the way. This paper attempts to do just that by applying to the problem a sophisticated regression model involving both continuous and indicator variables.

2. Method

The dependent variable is the average number of homeruns hit per year (to control for number of years played) by the top two- hundred homerun hitters of all time as of January 1, 2005. To be on this list, a player had to have hit 234 homeruns in his career up to that point. At the top of the list is, of course, Hank Aaron with 755 homeruns. Three are at the bottom of the list, tied with 234 homeruns—Craig Biggio, Gary Matthews, and Kevin Mitchell.

[1] U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Washington, DC 20212

Selecting the appropriate independent variables was more problematic. Based on extensive research, the following measures were chosen. Those players who ended their careers before 1980 were coded as "0" on the steroid use variable. For ones playing after 1979, a homerun hitter was coded "1" for having used steroids if the measure of the circumference of his forearms in the year he collected his most homeruns was greater than twice the standard deviation of the league average and "0" otherwise. The only exception was Steve Garvey, who was coded "0," even though his forearms were four times the standard deviation in his best hitting year (See Garvey 1994.). Initially, the number of days on the Disabled List (DL) was considered as a measure of medical care, but DL statistics were not kept before 1960 (apparently players were not allowed to claim they were disabled). Although it took some searching, the average number of days per year a player played with something broken turned out to be the best choice for measuring the state of medical care. To measure the level of training, the average belt size across the years of the player's career was used (Ruth 1940). Surprisingly, the most difficult variable to measure was nutrition. Very few records on nutrition were kept in the early days of baseball. Fortunately, two have been kept since almost the beginning—average number of hot dogs consumed yearly by a player and average yearly consumption of beer. So these two measures served as nutritional surrogates (See Wells and Valenzuela 2004 for more on these important statistics.). Finally, all baseball teams since the early 1920s has administered the same attitudinal questionnaire to each player when they signed with the team, and these scores have been maintained for posterity in Cooperstown at the Baseball Hall of Fame. This measure might serve as an important control variable. As you might expect, Barry Bonds had the lowest score, followed by Reggie Jackson. Jason Giambi had the best score, but who knows what he was on at the time.

3. Results

After coding and reviewing all the variables, no missing data were found. Using the Kitchen Sink approach (pitching all the variables into the equation) with SAS (Bulls 2000), data for all 200 players were analyzed. The resulting equation is given below, and the signs on the coefficients must be reversed to see effects on yearly homerun output. As you can see, the equation is a fairly useful predictor of homerun hitting with 57% of the variance explained. Steroid use (Forearms) is the best predictor, but both Hot Dogs and Beer also explain reasonably large portions of the variance. It might have been expected that they both would have large negative coefficients, but only one does. Could it be possible that their effects counteract each other? Neither Broken Bones nor Attitude has significant effects. The players from long ago must have been a tough lot. I guess attitude doesn't make much difference, except to the press. The effect of Belt Size is almost negligible. Maybe girth is more of a factor in fielding than in hitting.

Table 1. EQUATION: Variable, coefficient, cum R-square and p-value

Variable	Coefficient	Cumulative R-Square	P-value
Intercept	137.81		.0001
Forearms	42.68	.2522	.0001
Broken Bones	-2.44	.2703	.2176
Belt Size	3.57	.3011	.0750
Hot Dogs	-21.70	.4341	.0031
Beer	19.88	.5563	.0059
Attitude	1.66	.4605	.3089

4. Discussion

Could sabermetricians use these results to correct statistics to reflect the era in which the player played? Can we avoid asterisks? Given the amount of variance in homerun hitting power explained by these variables, it certainly is worth a try. For example, if the equation were used, a total of about 65 homeruns would be deducted from Barry Bonds' career numbers. Not only does he have gigantic forearms, but he also drinks more beer than he eats hot dogs.

5. References

Bouton, J. (2005). *Ball Four and Counting*. London: Oxford University Press.

Bulls, D. (2000). *What Has Stepwise Got to Do with It*? Durham, NC: SAS Institute.

Garvey, S. (1994). *Popeye Plays Ball*. Los Angeles: Dodgertown Press.

Ruth, G.H. (1940). *I Can't See the Ground Anymore*. Baltimore, MD: Raven Press.

Vincent, F. (2003). "Former Baseball Commissioner Fay Vincent Speaks Out: Maris, Schmaris," The New York Times, April 27, 2003, OP-ED page.

Wells, D. and Valenzuela, F. (2004). *I Can Pitch Anytime, Anywhere*. Hoboken, NJ: Has Been Press.

Received April 2006

World Perfect Comparability in Mini-Mental State Tests

Janet A Harkness

The article illustrates the cross-cultural calibration undertaken in the Duez-Bastas Yukon Mental Health Scale on the basis of repeat-the-phrase tasks.

Short or Mini-Mental state tests are often used to assess whether a given survey constitutes too great a cognitive burden for respondents. In the cross-national testing context, it is important to ensure that items are measuring the same cognitive ability at the same level of difficulty across populations. The Duez-Bastas Yukon Mental Health Scale (DBY-4 2004) has carefully calibrated cognitive ability test items for 2,112 language versions on a variety of tasks regularly used in mental state tests. Translation procedures used in DBY-4 aim for cultural equivalence plus word-for-word faithfulness backwards and forwards (as in *madam*). We focus here on a repeat-a-phrase task calibrated in the DBY-4 scale for multi-national implementation. In such tasks, the interviewer says a phrase and respondents repeat it, scoring points only if they manage to repeat the phrase exactly.

Hearing proficiency must be ascertained before the test is administered. Interviewers carry disposable hearing aid kits and receive a one-day briefing on fitting these. Nonetheless, the growing cultural diversity among populations may mean that respondents are unfamiliar with the accent of the interviewer and are unable to *exactly* copy her/his version. Alternatively, the interviewer may be unfamiliar with the accent of the respondent and fail to score appropriately. Interviewers with proven mimicry talents can be employed to reduce such difficulties. In some contests, a written test is not available in the language spoken by the respondent. The interviewer is required to translate orally on the spot. Scoring then needs to be adjusted accordingly because answers correct in one language may be wrong in another. Budgets permitting, interviewers carry automatic translation kits or can contact a call center should they be lost for words.

Repeat-a-phrase test items are sophisticated. For example, a phrase frequently used in English repetition tests is "*No ifs, ands or buts."* This resembles another commonly used phrase "No ifs or buts". Respondents must thus be careful not to confuse the phrase they hear with a familiar phrase they do not hear. In addition, in order to decide whether they heard *No if ... nor …* or *No if ...or*; respondents must quickly review the English rules for co-ordination and negation. Finally, since "*No ifs, ands or buts*" contains only function words, it is more difficult to process than "No cats, dogs, or children".

A corresponding German test item "Kein wenn, und oder aber" is an almost perfect word-for-word match for the English (= "No if, and or but"). German does not have a "or/nor" pair, but does have a "one/ no(ne)" pair (= ein/kein). By including an

ein/kein challenge as the first word, a matching item with Kafkaesque-like tasks is achieved. The everyday expression in German is “Without if and but” (= Ohne wenn und aber). The test phrase replaces “without” with “no”. Thus German respondents must decide whether they heard “ein” or “kein”; reproduce the unusual 3-component phrase instead of the normal 2-component phrase, and avoid the idiomatic “without” that ordinary language usage would trigger.

While equivalence of the German and English items is based on careful word-for-word translation, the corresponding item in the Italian test illustrates a functionally equivalence approach to comparability (van Deth 1998). The test task hinges on the value placed on pleasant-sounding speech in the Italian culture. The test asks respondents to repeat a tongue twister with an impressive rolling “r”; “Tigre contro tigre” (= tiger against tiger). Objections that the item is biased against Asian residents are unfounded; evidence for validity and reliability predates recent patterns of immigration. The Italian scoring instructions read: Only accept the exact and full phrase. Do not accept “tiger between tiger”, “leopard against leopard” or any rendering such as *tigel, rion* or *reopard*.

We can immediately sense the functional equivalence balancing Northern functionalism against Southern vitality, colour, and movement. Limitations of space force us to end our illustration here. Details on scoring calibrations for all DBY-4 test items can be downloaded for a commensurate fee from the DBY-4 website (http://www.see-cells-she-sells.com), as can the indispensable 2452-page interviewer training manual.

Literatur

Van Deth, J. (1998). Equivalence in Comparative Political Research in J. v Deth (ed.) Comparative Politics the Problem of Equivalence, Routledge, New York, 1-19.

Duez-Bastas Yukon (2004). Mental Health Scale, Greysell Publications, Amsterdam, Netherlands.

Received April 2005

Expert Review as a Method for Evaluating Survey Questionnaires: Has It "Evolved" Yet?

Nancy Bates[1]

This article explores expert review (ER) as a technique for evaluating and improving survey questionnaires. Expert review has become a commonplace method to pretest and evaluate questionnaires and is especially popular in cases where the survey is running a deficit, behind production, and/or the principle researcher is just plain lazy. We report on a controlled experiment to evaluate the effectiveness of this technique by comparing a questionnaire evaluated and revised by a panel of experts (genus *homo sapiens*) to the same questionnaire evaluated and revised by a panel of randomly selected Howler monkeys (genus *alouattta*).

Key words: Questionnaire design; response rates; honorarium; bananas.

1. Man versus Monkey

In this study we empirically test the merits of "expert review" (ER) as a method for evaluating a self-administered paper and pencil questionnaire. The ER technique is simple, widely used by government statistical agencies, but heretofore, has never been scrutinized under the conditions of a controlled experiment (Presley, E. et. al 2004; DeMothra and Landers; 2004; Forsooth, Rotweiler and Willy; 2004).

2. Methods

The experiment consisted of a mailout/mailback survey involving two split panels. The first panel consisted of a questionnaire designed collaboratively by human methodologists using the ER technique. The second panel represented a questionnaire designed by Howler monkeys. The outcome measure of "success" between the two questionnaires was the final mail response rate (as defined by RR2, AAPOR, 2004).

2.1. The ER panel

A panel of twelve survey methodology "experts" were recruited to participate in the study. Each member was promised a modest honorarium and a trip to Washington, D.C. in exchange for participating in a one-day workshop. The goal of the workshop was to

[1] The author assumes absolutely no responsibility for any part of this article. Except for royalties.

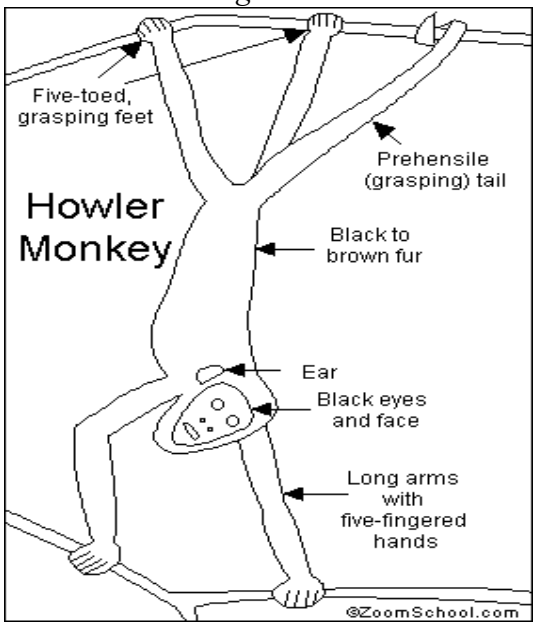
revise and improve a paper and pencil mailout/mailback survey questionnaire. The questionnaire was the National Survey of Really Sensitive Topics (NSRSI) which includes questions about illicit drug use, deviate sexual behavior, pet ownership, tax cheating, patriotism, and trash recycling practices.

In a nutshell, the expert panel spent most of the day quibbling over their honorarium and crummy lunch. When they did find time to critique the questionnaire, they contradicted one another, were petty about who should get credit for pointing out design flaws, and, in the end, could not come to any consensus about how to improve the questionnaire. As a result, the questionnaire for the ER panel was left exactly as it had been to start with.

2.2 The Howler monkey panel

To revise the questionnaire for Panel 2, a group of Howler monkeys and a professional handler were recruited from a local laboratory (see Fig. 1). The paper questionnaire was placed in an empty room equipped with a 2-way mirror. The monkeys were led into the room as the handler observed through the mirror. One monkey cautiously approached the questionnaire and then proceeded to eat some of it. A second monkey then smeared excrement on the cover. The handler interpreted these behaviors to mean simply: reduce the length of the questionnaire and add some nice color and perhaps a graphic or two. The principal research made these revisions and the monkey version was finalized.

Figure 1.



3. Results

The two questionnaires were mailed to a randomly split sample of left-over crap from the 2000 U.S. Decennial Census. Response rate results are presented in Table 1.

Table 1. Final response rates

	Version A – Expert review	*Version B – Howler monkey*
Mail response rate	23.9%	89.2%
N	1,000	1,000
$X^2 = 469.31$, d.f.=1, $p<.001$		

4. Recommendations and Conclusions

Stock up on bananas.

5. References

American Association for Public Opinion Research. 2004. Standard Definitions: How to Finally Dispose of Cases That Refuse to Cooperate. Lenexa: Kansas.

DeMothra, T., and Landers, A. (2004). Do Different Cognitive Interview Techniques Product Different Results (And Who Really Cares Anyway?" In Tried and True Methods for Testing and Evaluating Survey Questionnaires. Presser et al. (eds.) Wiley: New Jersey.

DeMothra, T., and Landers, A. (2003). Senseless Bureau Standard: Pretesting Questionnaires and Related Materials for Surveys and Censuses. http://www.census.gov/important/standards/pretesting.html.

Forsooth, B., Rotweiler, J. and Willy, G. 2004. Q: Does Pretesting Make A Difference? A: No. In Tried and True Methods for Testing and Evaluating Survey Questionnaires. Presser et al. (eds.) Wiley: New Jersey.

Presley, E., Rotweiler, J., Pooper, M., Lassie, J., Martini, E., Martin, J., and Swinger E. (editors). 2004. Tried and True Methods for Testing and Evaluating Survey Questionnaires. Wiley: New Jersey.

Received April 2005

Using IRT for Pretesting Self-Administered Questionnaires: A Test of the LCD Hypothesis

Paul P. Biemer[1]

An important problem in the design of self-administered questionnaires is the construction of survey questions that are readily understood and easy to answer. Many researchers advocate using the "lowest common denominator (LCD)" approach. With LCD, survey questions are designed so that persons of lesser intelligence can answer them accurately, thus ensuring that everyone in the population will readily understand the questions. As described in the article, recruiting unintelligent subjects for pretesting purposes is fraught with difficulties. To address these problems, a method for inducing stupidity in pretest subjects of average intelligence is proposed that we refer to as the "intelligence reduction technique (IRT)." This article briefly discusses IRT and its feasibility for questionnaire design.

Key words: Moron-friendly design; simple (-minded) response variance; induced imbecility; survey design for dummies; inebriation-based question testing.

1. Introduction

A general rule of thumb for writing intelligible survey questions is to assume a 5th grade reading level for survey respondents. However, as Lyberg (2004) illustrates, the reading level of a respondent may have little to do with how well a survey respondent understands a question. More important is the respondent's intelligence level. An intelligent respondent who reads at a low grade level may still be able to respond accurately to a survey question that is written for a higher reading level. On the other hand, a person who reads at a high grade level but is a complete moron is more likely to answer the same survey question inaccurately. This has lead to the use of the lowest common denominator (LCD) approach (described above) in designing questions.

As Lyberg explains, to apply the LCD approach, pretest subjects should be persons who represent the lesser intelligent persons in the population (so-called LCDs). However, in our experience, recruiting LCDs can be quite problematic. Lyberg (2004) recounts one where a general advertisement for pretest subjects produced a group of persons who were too smart to participate in the pretest. Persons who could read and understand the posted ads and show up for pretesting at the right place and at the right

[1] RTI International and University of North Carolina at Chapel Hill.

time scored too high on intelligence tests to qualify as LCDs. Thus, advertising for LCD's is neither effective nor efficient.

In the current study, we attempted to recruit LCDs in the obvious places where they might be found. For example, we contacted high schools and requested lists of the most intellectually-challenged students. The schools were not cooperative and even seemed to resent the request. We approached fashion modeling schools to request a list of there students. However, once they understood our motives, they too were not cooperative.

The most successful approach was to recruiting LCDs from the population of blond females between the ages of 18 and 35; however, only 15% qualified as LCDs. The others were not true blonds. In addition, the no-show rate among LCD subjects was 99%. The most common reasons were forgetfulness, could not find the address of the test site, found the address but could not find the cognitive laboratory and got the time wrong. As before, most of the subjects who did find the cognitive laboratory tested too high on the intelligence test to qualify as LCDs.

2. Solution

These failures led to the idea of inducing imbecility in persons of at least average intelligence. This technique, referred to as the Intelligence Reduction Technique (IRT), eliminates the disadvantages of LCD recruitment described above. Persons can be recruited from the general population without regard to their intelligence-level (or hair color/gender combinations). The only requirement is that they be willing to consume large quantities of alcohol in order to attain LCD-status through inebriation. Through experimentation, we determined that IRT requires a blood alcohol concentration (BAC) of 0.20 to reduce a person of average intelligence to an LCD. Using IRT, the no-show rate for test subjects has been reduced to almost 0. Further, as the next section illustrates, questionnaires developed under IRT satisfy the LCD hypothesis; i.e., questions that can be answered accurately by the IRT induced LCDs can be answered by virtually anyone.

3. Results

To test the IRT approach, we recruited 50 persons of average intelligence and randomly assigned half to the control group and half to the IRT group. We also recruited 25 bonafide LCDs from a local fashion modeling school whose cooperation we were able to obtain. For the IRT group, alcohol was administered to each subject over a one hour period until the BAC for each subject measured 0.20 g/ml using the Alcosenser III breathalyzer. Each person in each group was then asked to complete a self-administered questionnaire containing 20 survey questions. Approximately three days later, all subjects were asked to answer the same survey questions using an interviewer-assisted approach. For the reinterviews, the IRT subjects remained sober for obvious reasons. Discrepancies between the interviews and reinterviews were reconciled to determine the causes of the discrepancies. The results are reported in Table 1.

Table 1. Proportion in LCD and control groups who correctly interpret survey questions that were misinterpreted by the IRT group

	IRT group	
LCD group	Correctly interpret	Not correctly interpret
Correctly interpret	0.96	0.22
Incorrectly interpret	0.04	0.78
Control group		
Correctly interpreted	1.00	0.85
Incorrectly interpret	0.00	0.15

From this table, the high correspondence between interpretations by the IRT and true LCD groups is apparent. Also, note that 100% of the questions that were correctly interpreted by the IRT group were also correctly interpreted by the Control Group while for those that were not correctly interpreted by the IRT group, the rate of correct interpretability by the Control drops to 0.85. These suggest that the LCD hypothesis has some merit.

4. Conclusions

This study provides some evidence that IRT is successful producing a group of LCD test subjects in very cost effective manner. We estimated that the IRT subjects ingested the equivalent of approximately 280 ml of inexpensive whisky on average. The cost of alcohol was only a fraction of the cost of recruiting the 25 LCDs used in the test. We believe that IRT is an indispensable tool for design survey questions that are interpreted correctly by all respondents.

5. References

Lyberg, L. (2004). Moron-Friendly Questionnaire Design. Journal of Obnoxious Statistics, Vol. 29, No. 1, 17-35.

Received April 2005

I.S.I. (Increased Standardised Interviewing): A New Training Method Designed to Increase Interviewer Bias and Respondent Variance

Pamela C. Campanelli[1]

Survey researchers often go to great lengths to standardise questions and interviewers' behaviour. Back in the late 1980's, work from other disciplines suggested that this standardisation assumption should be questioned. For example, anthropologists Mishler (1986) and Suchman & Jordon (1990), among others, have been critical of survey research in this respect. This debate has been kept alive with research on "flexible interviewing" (see, for example, the work of Schober and Conrad, 1997.). This article is about a brand new approach to training interviewers, developed from common interviewer practice, that will provide a new dimension to this debate.

Key words: Interviewers; standardisation; variance; bias.

1. Introduction and Summary

The I.S.I. code of practice in training interviews is contained in the follow 7 simple points.

Specific Points

1. For survey questions on household expenditure, interviewers must be trained to keep householders from going astray and trying to check bills or records in order to answer the survey questions.
2. For the employment question about the number of hours worked, all interviewers must be instructed to interrupt the respondent while they are the process of trying to remember and suggest the answer of 60 hours. This will make all the "lay-about" respondents feel like they have really accomplished something and for those who might have given a higher figure than 60, it eliminates all that "macho showing-off". To minimise variance it is important that ALL interviewers suggest that value of 60 and stay with this answer no matter what the respondent says.
3. Upon encountering respondents with both self-employment and employee income, interviewers must be strictly guided to force respondents to choose only one of these two options. Under no circumstances should income from both

[1] Survey Methods Consultant, Colchester, UK. E-mail: pamc@aspects.net

sources be recorded in the questionnaire, despite respondent protests. Interviewers must learn never to give in to respondent whims.

General Points

4. Make sure interviewers rush the respondent as much as possible. We don't want respondents to think too much about their answers. This also insures that interviewers can do more interviews over a shorter period of time, resulting in overall cost savings.
5. Whenever the respondent is debating between two different answers, interviewers must be trained to interrupt and suggest an answer. This will also save interview time.
6. Another strategy is to train interviewers to talk about their personal lives during the interview. This will produce excellent rapport with the respondent. But interviewers must be made aware that the more they talk about themselves, the more they will need to rush the respondent so as to maintain the overall cost savings.
7. On household surveys where all members of the household are to be interviewed, interviewers should be allowed to collect proxy information for the other household members. But only if the interviewer has a holiday booked for the next week. The collection of proxy information is appropriate for all types of survey questions. If the respondents supplying the proxy information pretend to be frustrated by not knowing the answer for other household members, interviewers must be trained to threaten respondents with being stripped naked and forced to stand wearing a hood over their heads, while the interviewer points a large gun at them and someone takes a picture. If this proves unsuccessful, interviewers must be instructed to simply break off the interview and code it as nonresponse.

The I.S.I. method is a "total design" method, which means that all of the 7 points must be implemented to guarantee the maximum impact of the practice.[1]

2. References

Mishler, E.G. (1986). Research Interviewing: Context and Narrative. Cambridge, MA: Harvard University Press.

Schober, M.F. and Conrad, F.G. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? Public Opinion Quarterly, 61 (4), 576-602.

Suchman, L. and Jordan, B. (1990). Interactional Troubles in Face-to-Face Survey Interviews, Journal of the American Statistical Association, 85, 232-253.

Received April 2005

[1] The core of the I.S.I. method was based on a TRUE STORY of how one new social research interviewer behaved on one UK government survey when the author had been randomly selected as a respondent.

Some Recent Advances in the Methodology of Longitudinal Survey Data Collection

Roeland Beerten[1] *and Peter Lynn*[2]

Over the last few decades there have been many longitudinal surveys across the world, and there is a long tradition of methodological research in this area. However, it is only recently that some truly groundbreaking insights in longitudinal survey design have appeared in scientific publications such as this journal. This article presents an overview of these pioneering new methods and solutions for some of the most difficult problems in the history of survey research, and mankind in general.

Key words: Longitudinal surveys, this journal, mankind.

1. Introduction

In recent years there have been several groundbreaking developments in the design of longitudinal surveys. This article gives an overview of the most important areas of work.

2. Seam Effects

Seam effects occur when retrospective histories collected at each of a number of successive survey waves are combined to create a continuous history. A surfeit of apparent transitions is observed at the "seams" between two waves (see also Malton and Killer, 1991). This arises because respondents may describe the same status differently, or because the information may be coded differently, even when the status being reported at each wave is in fact the same.

Recent research has identified two techniques that are very effective at reducing this seam effect. The first is to ask only about phenomena and characteristics that are unlikely to change and unlikely to be described differently. The new International Longitudinal Survey of Gender and Eye Colour is a shining example. The second technique acknowledges that occasionally it is unavoidable that longitudinal surveys must ask about things that actually change. In this case, the preferred approach is to use dependent interviewing to make it very difficult for the respondent to indicate a change, e.g. "Last time you were <status>; that is still the case isn't it?" In case the respondent insists on answering "no", ingenious researchers have developed the follow-up question, "If your status has changed since last time, I must ask you an additional 36 questions about this change. Are you sure your status has changed?"

[1] Office for Notional Statistics, London, UK.
[2] University of E-sex, Colchester, UK.

3. Accuracy of Measures of Change

Unfortunately, seam effects are only one of the features that are detrimental to survey measurement of change. Other forms of measurement error can occur at each wave and thus reduce the accuracy of measures of change. The recently honed technique of Imputation for Future Years (IFFY) has provided a breakthrough way of reducing to zero the measurement error associated with estimates of change. The IFFY method involves only carrying out one full wave of data collection and then imputing the values for future waves under the scheme $y_{ik}=y_{ij}$, for all $k\neq j$ where y_{ik} is the value of a variable y for respondent i at wave k. Unfortunately, this comes at a price: a) It is necessary to redefine the underlying change as that implied by the imputation scheme; b) Estimates of change are typically very small (a controversial recent paper suggests that the estimates of change are almost entirely determined by imputation errors).

4. Cross-National Comparability

The past decade has seen considerable growth in cross-national longitudinal surveys. It has been suggested that cross-cultural and cross-lingual differences introduce a major source of measurement error to cross-national comparisons, particularly for measures of change. Recent advances in Europe have served to reduce the impact of these factors. This has primarily been achieved by specifying survey designs and survey measures in such a vague way that a huge amount of random error is introduced into the observations. The error component due to cross-cultural differences is thus dwarfed by this new random error and ceases to be a major component of measurement error.

5. Keeping in Touch with Respondents

By their nature most longitudinal surveys must revisit respondents for a second or occasionally even a third or a fourth interview. In these circumstances it is an essential part of the survey process that the survey organisation is aware of the whereabouts of the respondent. This can be a daunting task. Some respondents do not confine their movements within their country's borders, but need to be followed throughout some of the Nature Reserves of the U.S.A., or at international gatherings of their professional peer groups. Apart from the serious difficulties this causes for interviewers it also leads to huge increases in survey costs. As a consequence of these extensive efforts to trace respondents the survey cost/quality balance changes and the survey quality which is achieved is disproportional to the associated cost (see also Biemer and Lyberg 2003).

6. Panel Attrition

Finally, one of the most common problems in panel surveys is panel attrition. This occurs when respondents from the first interview cannot be persuaded to co-operate with the survey, or when they cannot be found; see also Section 4 of this overview and an article by Laurie et al. (1999).

To counter the problem of panel attrition most survey organisation now give incentives to respondents. By giving respondents a small incentive after each visit the interviewer may be able to convince them to continue to co-operate with the survey. There has been a lot of research into the types of incentives that can be used. A new and groundbreaking study from the UK Office for Notional Statistics has found that it is always better to give respondents vouchers for high-street stores rather than cash. Furthermore the study found that vouchers for the Swedish Ikea stores were by far the most effective at retaining respondents. Researchers put this down to the chain's extraordinary ability to retain customers better than a survey could ever hope; as a response the ONS is now considering changing its corporate colours to the familiar blue and yellow colour scheme that was exported so successfully by the Swedish company.

7. Conclusion

The future is bright for longitudinal surveys. Researchers have demonstrated their ability to overcome even the most taxing problems associated with data collection. We believe that in this overview we have identified an important theme in the methodological research that has underpinned recent developments, namely the approach of "avoiding the issue." We believe this to be a promising philosophy that could be further exploited in our quest to be able to claim that surveys can measure all things in all circumstances extremely well.

8. References

Liemer, P. and Byberg, L. (2003). Introduction to Survey Quality. Hoboken, NJ: Wiley.

Kalton, G. and Miller, M. (1991). The Seam Effect with Social Security Income in the Survey of Income and Program Participation. Journal of Official Statistics, 7 (2).

Laurie, H., Smith, R., and Scott, L. (1999). Strategies for Reducing Nonresponse in a Longitudinal Panel Survey." Journal of Official Statistics, 15 (2).

Received April 2005

A New Survey Technology: CRAPI

Mick P. Cooper[1]

This article described the development and initial field tests of CRAPI, or Computer Replacement for All Personal Interviewing. In order to reduce the high costs associated with using human interviewers for face-to-face interviewing, we have been prototyping a robotic device to replace the work of interviewers. Our initial tests have shown much promise, with response rates at least as high as those obtained by recruiting subjects via the Internet. Several bugs still need to be worked out, most notably related to the humanness of the robot interviewers, and the possible introduction of social desirability effects.

Key words: Computerized surveys; automated interviewing; CRAPI.

1. Introduction

One of the perennial problems facing the survey field is the high cost of interviewers, especially in face-to-face surveys. Surveys instruments have been successfully automated over the last several decades, both in computer assisted telephone interviewing and computer assisted personal interviewing (see Cooper and Nicholls 1998). With the addition of voice in audio-CASI, parts of the interview are now being administered without the need for an interviewer, other than to deliver the equipment and provide support. In telephone surveys, automated interviews are increasingly prevalent, both in the form of interviewer-recruited surveys such as outbound interactive voice response (IVR) surveys or telephone audio-CASI, or even fully-automated inbound surveys (touchtone data entry or TDE) (Turner et al. 1998). RTI International has a system called FATI (Fully Automated Telephone Interviewing), and also claims to have a CARI system (Beamer 2002). Political pollsters and market researchers are using a method called "RoboInterviewer" in which a fully automated survey dials respondents, invites them to participate in a telephone survey, and poses the survey questions. But thus far, efforts to reduce the role of interviewers in face-to-face surveys have met with little success. This article reports on the development of a new technology called CRAPI (Computer Replacement for All Personal Interviewing), and the results of the initial field tests.

2. Design of the CRAPI System

The inspiration of the design comes from the brilliant but often-overlooked research of Cooper and his colleagues (e.g., Cooper, Swinger, and Tarantula 2003; 2004). The fact that many of their experimental efforts to introduce human-like features into computer assisted surveys such as audio-CASI, IVR and the Web have failed leads us to believe

[1] University of Mickigan.

that replacing interviewers with robots is likely to have little negative effect on measurement error, but could save large amounts of money. In addition, the failed efforts of Groovy (2003) to replicate earlier findings of training interviewers to avert refusals, suggests that this task could be done just as well by a machine. Given how low survey response rates are these days, could it be any worse with robots?

Our first efforts focused on the development of a prototype with human-like features. We chose as our model a leading survey researcher of the day (see Figure 1), believing that this would add credibility and respect to the robot interviewer's approach.

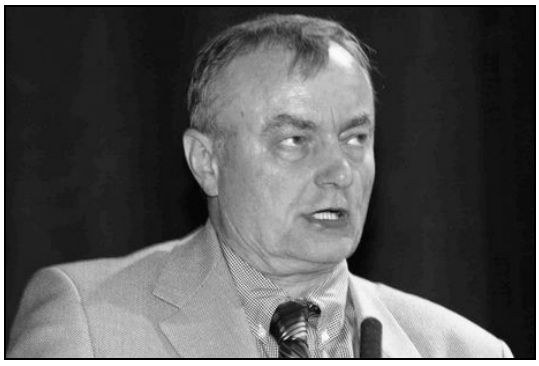


Fig. 1. Model used for human-like CRAPI prototype

Based on usability testing, we decided to contrast the human-like version with a machine-like version, largely devoid of human features or expression. Both versions of the prototype were programmed to comply with the three laws of robotics (see Lyberg 2000). Given Sweden's leading role in technological development, the system was developed and built in kit form by a leading Swedish technology company, Ikeasoft. Of course the system was fully bilingual, capable of conversing in both Swedish and Swenglish.

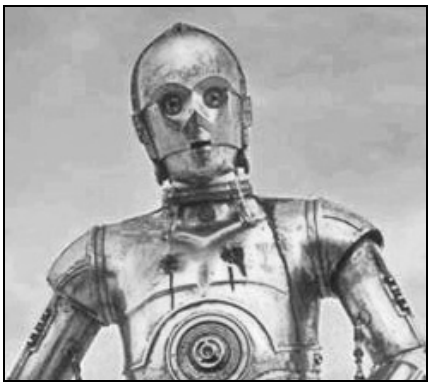


Fig. 2. The human-like CRAPI prototype

Fig. 3. The machine-like CRAPI prototype

3. The Field Experiment

The initial field experiment was carried out in two sites: Lakeland, Florida, and Stockholm, Sweden. These sites were chosen because the model for the human-like

CRAPI system would not seem out of place in either place. We decided to exploit the use of these sites by introducing seasonal variation, with implementation in Florida in the Spring, and Stockholm in mid-Winter. We employed a fully double-crossed design (see Hoax 1995), in which each CRAPI prototype was tested on a random subset of addresses in each site. The initial design called for a sample of 100 addresses in each of the four cells of the design, but because of the poor performance of the human-like interface in early pilot tests (see Cooper 2001), we increased this to 1,000 per cell.

The prototypes were programmed in all the functions of human interviewers, from gaining cooperation at the doorstep to administering the survey questions. An example of the CRAPI prototype conducting a survey interview is shown in Figure 4.

Fig. 4. The machine-like CRAPI prototype handing a show card to a Swedish respondent

4. Results

This article focuses on the results of the cooperation efforts. A later article will focus on measurement error differences between the two styles of CAPI prototype, once the survey responses are translated from the internal machine language into something comprehensible. A summary of the response rates is presented in Table 1.

Table 1: Double-crossed design, sample size and response rate per cell

Interface style	Site			
	Florida		Stockholm	
	Response rate (%)	(*n*)	Response rate (%)	(*n*)
Human-like	2.5	1,000	8.3	1,000
Machine-like	7.9	1,000	9.2	1,000
Total	5.2	2,000	8.75	2,000

The low overall response rate was a disappointment. However, it is on par with many online surveys reported today, and, given the precipitous decline in telephone survey

response rates, will soon be no worse than a typical RDD survey. Furthermore, when compared with typical face-to-face survey response rates in the Netherlands (see de Lewd and de Hair 1998), these rates are quite respectable. Despite the low response rate, we obtained sufficient data points to permit multiple imputation for the missing data, using the techniques of Rittle and Lubin (1990), thereby recreating the entire original sample, and then some.

Interestingly, the human-like CRAPI system produced a significantly ($p<.05$) lower response rate overall than the machine-like system (5.5% versus 8.6%, respectively). In part we attribute this to the model used for the human-like prototype. In addition, the response rate in Florida was significantly lower than that in Stockholm. As can be seen from Table 1, this is largely due to the poor performance of the human-like interface in Florida. Part of this may be due to a technical glitch in the behavior of the prototype. Whenever it was assigned to visit addresses in the vicinity of a baseball stadium, the CRAPI prototype would inexplicably divert from its preprogrammed path, and spend several idle hours watching baseball. Our vendor, Ikeasoft, is working on a solution to this problem. But even if we account for this technical flaw in the prototype, the human-like interface did worse in Florida than in Sweden.

A debriefing was conducted with selected (non)respondents following their interaction with the CRAPI system. Overall, people felt that the human-like prototype was not very human-like. On the other hand, several respondents were unaware that they were interacting with a machine (cf. Nordqvist,1994). Some representative comments made by the debriefing respondents are as follows:

"It was hard to understand the accent – for some reason I kept thinking of my Volvo." (Machine-like interface, Florida)

"It was very sexy, so I invited [the machine] in for a cup of coffee but all it wanted to do was ask me stupid questions." (Human-like interface, Stockholm)

"I've had better conversations talking with my VCR." (Human-like interface, Florida)

"Den var så verklig. Jag förstod inte att det var en maskin." (Translation: "It was so life-like. I did not realize it was a machine." Machine-like interface, Stockholm)

Clearly, the reactions to the CRAPI prototype were mixed. Some persons treated the CRAPI system as human (regardless of the type of interface we designed), while others viewed it as a robot. The effect of this on data quality will be explored later (see Beamer and Lyberg 2003).

5. Conclusions

CRAPI appears to be a promising technology for replacing field interviewers in face-to-face-surveys. While we did not achieve response rates close to what well-trained professional interviewers might obtain, the response rates were higher than those for many Internet surveys, which are the new benchmark for survey quality (see Black and Taylor 1997). In terms of next steps, a clear design improvement is to use a different model to serve as the prototype of the human-like interface. We believe if this was done, we would see a reversal of the finding, with the human-like CRAPI out-performing the machine-like CRAPI system.

Some minor technical errors not reported above were discovered during the test. For example, in some cases overheating caused unpredictable errors in the Florida test, producing nonsensical questions or questions being asked in an unpredictable order. But again, none of the respondents that we debriefed appeared to notice these anomalies, and answered all the questions as if they made sense. Similarly, severe cold in the Stockholm site slowed down the CRAPI system considerably in a number of cases, but apparently most of those debriefed did not appear to notice. When prompted they said they thought this was the normal pace of a conversation.

6. References

Beamer, P.P. (2002). CARI: Computer-Assisted Robotic Interviewing. Paper presented at the Joint Sadistical Meeting (JSM) of the Acronym Society of America (ASA), August, New York, New York (NYNY).

Beamer, P.P. and Lyberg, L.E. (2004).The Improtance Survye of Data Qaulity. New Yrok: Wiley-Coyote.

Black G. and Taylor, H. (1997). Our Method can Beat up Your Method Any Day: Why the Internet will Replace All Other Modes of Data Collection. Rochester, NY: Harry Interactive, Technical Paper.

Cooper, M.P<.05. (2001). Paradata and Parachutes: Important Tools in Pilot Testing. Journal of Obnoxious Statistics, 17 (4): 24.95 or less.

Cooper, M.P. and Nicholls IV, W.L. (1998). The History and Development of Acronyms: CASIC, CAI, CATI, CAPI, ACASI, T-ACASI, IVR, TDE, VRE, ASR, and More. In Cooper et al. (eds.), A Collection of Computer Assisted Survey Information. New York: Wiley-Coyote, pp. 1-∞.

Cooper, M.P., Swinger, E., and Tarantula, R. (2002). In Search of the Null Finding: Making Machines Human-Like and Humans Machine-Like, and Failing at Both. Journal of Obnoxious Statistics, 18 (2), 457-459.

Cooper, M.P., Swinger, E., and Tarantula, R. (2004). Does Anything Matter? An IVR Experiment. Journal of Obnoxious Statistics, 20 (3), 47-32.

De Lewd, E.D. and de Hair, W.X.Y.Z. (1997). Somebody Has to Be on the Bottom. A Meta-Analysis of Response Rates in the Netherlands Versus the Rest of the World. Journal of Obnoxious Statistics, 13 (1), 25% - 36%.

Groovy, R.M. (2003). Training Interviewers to Become Tailors – A Bad Fit? Private Opinions Quarterly, 69 (3), 42-44.

Hoax, J.J. (1995). Double-Crossing: A New Approach to Experimental Design. Journal of Obnoxious Statistics, 11 (22), 12-21.

Nordqvist, S. (1995). Tomtemaskinen. Stocholm: Bokförlaget Opal AB

Lyberg, L.E. (2000). I, Robot. Private Opinions Quarterly, 66 (1), 451-1234.

Rittle, R., and Lubin, D. (1990). Multiply Imputing Everything. Journal of the American Sadistical Association, 198 (11), 1-93.

Turner, T., Koo, G., Rogers, F., Lindberg, C., Peck, J.Q., and Sommerstein, F.L. (1998). I Say Tomayto, You Say Tomato: Telephone Audio-CASI, IVR, and the Invention of Acronyms. Pseudo-Science, 280 (May 8): 867-873.

Received April 2005

The Walking-the-Dog Bias in Household Travel Surveys: Problem and Solution

Michael P. Cohen[1]

It is generally acknowledged that household travel surveys underestimate the number of walking trips. The walking-the-dog-trip is particularly prone to not being reported by respondents because they do not think of it as a trip in the same sense as, say, a journey-to-work trip. This article proposes a novel method for picking up more walking-the-dog trips: include the dog as a household member.

Key words: Canine; sidewalks.

1. Introduction

Estimating the number and characteristics of walking trips is important for public policy. The need for sidewalks, crosswalks, street lights, and other transportation infrastructure is measured, at least in part, by the amount of walking and when it occurs. There is, unfortunately, substantial evidence that walking trips are underestimated by household travel surveys (Litman, 2004, p. 2).

2. Importance of Walking Trips

Walking trips are a distant second to personal vehicle trips in frequency in the United States (Table 1 and Fig. 1). Similar results are reported in other developed nations (Litman, 2004, pp. 2-3).

But the frequency data provide a misleading picture of the social, economic, and public health importance of walking. As Litman (2004, p.2) puts it:

> "But consider another perspective. Would you rather lose your ability to drive or your ability to walk? Being able to drive, although useful, is less essential than the ability to walk. With a little planning, a physically-able non-driver can engage in most common activities, but being unable to walk affects nearly every aspect of life, creating barriers to employment, recreation and social activities."

As Litman (2004, p.2) continues putting it:

> "Homo sapiens are walking animals. Walking is a fundamental activity for physical and mental health, providing physical exercise and relaxation. It is a social and recreational activity. Environments that are conducive to walking are

[1] U.S. Bureau of Transportation Statistics, 400 Seventh Street SW #4432, Washington DC 20590 USA. E-mail: Michael.Cohen@bts.gov

The views in this article are those of the author. Official support by the U.S. Department of Transportation or any other entity is not implied nor should it be, even remotely, inferred.

Acknowledgments: The author acknowledges the enumerable annoying comments and suggestions of the referees and especially of the Chief Editor.

conducive to people. Walking is also a critical component of a transportation system, providing connections between homes and transit, parking lots and destinations, and within airports. Often, the best way to improve another form of transportation is to improve walkability."

Table 1. Distribution of trips by mode of transportation, in percent

	Percent	**SE**
Personal vehicle (POV)	86.6	0.18
POV-single occupant	37.6	0.25
POV-multiple occupants	48.9	0.28
Transit	1.5	0.06
School bus	1.7	0.05
Walk	8.6	0.13
Other	1.7	0.07
Total	100.0	

NOTE: SE = standard error. SOURCE: The 2001 National Household Travel Survey, daily trip file, U.S. Department of Transportation.

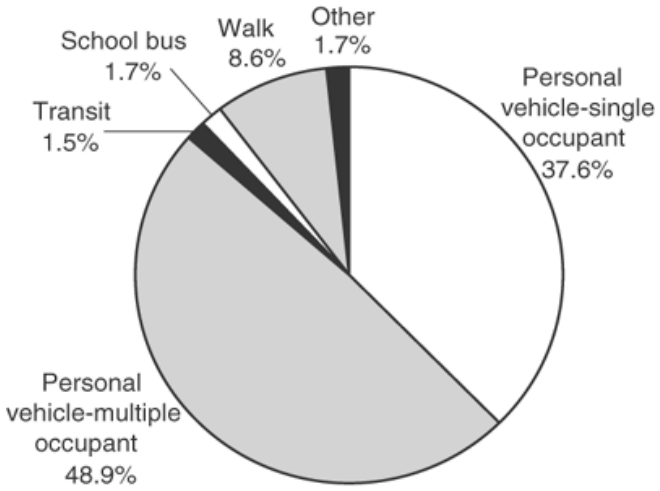


Fig. 1. Distribution of trips by mode of transportation, in percent
SOURCE: The 2001 National Household Travel Survey, daily trip file, U.S. Department of Transportation.

3. Walking-the-Dog-Trips

Among walking trips, walking-the-dog trips is clearly a major subcategory. Most dogs need to be walked at least once a day. Giles-Corti (2001, pp. 62-64) did a logistic regression analysis of "walking as recommended for good health" on individual, social, and physical environmental determinants. Dog ownership was a significant determinant (p=0.002).

4. Making the Dog a Member of the Household

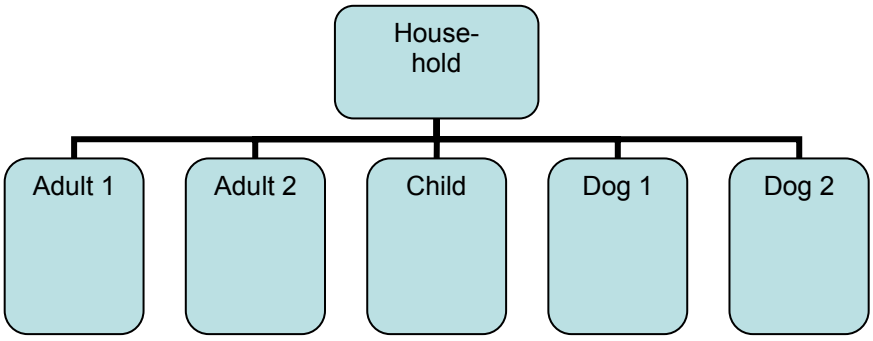


Fig. 2. A diagram of a household showing the household members, both human and canine.

Walking-the-dog trips are likely to be undercounted because respondents do not consider such trips to be "serious" in the same way that, say, journey-to-work trips are. We propose a radical but compelling solution to this problem: make the dogs members of the household. Under this system, household membership would consist of human household members (HHMs) and canine household members (CHMs) (Fig. 2). An HHM could proxy-report on trips for a CHM, just as adult HHMs now do for child HHMs. By prompting the proxy reporter for the CHM about who accompanied the CHM on his (or her) trips, we uncover missed trips by HHMs.

Extensive field testing of the proposal is warranted but has not yet been undertaken. Such testing would typically be done before publication of a journal article. But this new method is of such revolutionary importance that waiting for such testing was deemed inadvisable.

5. Concluding Remarks

This article proposed a new method for household travel surveys in which dogs are included as household members. This method is expected to lead to a quantum jump in reporting of walking-the-dog trips. This improvement will have crucial public policy implications.

6. References

Bose, J., Giesbrecht, L., and Sharp, J. (2003). Highlights of the 2001 National Household Travel Survey. Washington: U.S. Department of Transportation.

Giles-Corti, B. (2001). Walk This Way for Health. In Australia: Walking the 21st Century. Perth: Western Australia Department of Transport.

Litman, T.A. (2004). Economic Value of Walkability. Victoria: Victoria Transport Policy Institute.

Received April 2005

The REC-Survey "Remember the Households You Lived In" Question: A Research Note

Retep HP Relhom (AMUZ) and Nelletheb Llennepp (RSI) [1]

Household size is a key demographic variable, either as a single indicator of household complexity or a control variable for Kish-Grids in determining respondents in a given household or for controlling the completeness of household composition indicators. To determine who counts as a household member varies greatly among cultures and across time. The newly designed household size question of the RE-inCarnation (REC) survey may serve as a generic model for all cultures across all times.

Key words: Household size; demographics; cross-temporal; cavalry.

The unique and important contributions of the RE-inCarnation (REC) survey has created numerous requests for the actual question wording used to determine the size of household. The household size measure aimed to provide an inclusive and thorough determination of the household size of respondent accommodations during their lives here on earth. Due to the survey's cross-temporal and cross-cultural nature, the study's demographic questions had to cover a very wide variety of possible situations. The development of this critical question went through several stages of rigorous methodological testing and reformulation. In the following we present the final, condensed but comprehensive version of the 'number of household members' source questionnaire item:[2]

"How many people, including yourself, were living in the same household as you. Please include people with whom you shared the same accommodation such as a house, apartment or flat in a building, tepee, hut, yurt or other complex architectural agglomeration, such as castles, fortresses, citadels, bastions, palaces, etc. in a common living arrangement. A shared accommodation is defined as living in the same household. A common living arrangement includes people in the household who shared in the expenses of the household, had meals together, or shared a room. Sharing expenses includes those who benefited from the household expenses, such as children, a guest who never left, other persons with no income who lived in the household during this time or people who had income but did not share. People include related or family members and unrelated or non-family members. For those living in architectural agglomerations, please do not include unrelated or non-family members such as palace guards, household

[1] Excerpted from Llennep & Relhom: *From Here to Next*. Sunk-Cost-Publications. Googless/Queen Maud Land, 2005. The RE-inCarnation project was funded by the Strüby-Foundation, Ingenbohl.

[2] Measurement properties: Reliability -.8 and Validity approx. +≡; Standard Shrewdness Index .007.

cavalry, courtiers, or other non-family or unrelated household member household staff. Please do not include unrelated or non-family members who share a room for one evening only. Please do not include reincarnated persons unless they fit the above criteria (reincarnation as domestic animal should not be included under any circumstances, but see hondenstukCHM…). Shared living arrangements in public institutions such as circus maximus, boarding school, poorhouse, dungeon, asylum, monastery, convent, or the like, should be described in detail but estimating the number of people that lived there should not be attempted."

Further information about the REC survey can be found at *The Encyclopedia of Ingnorance* (2005 B.C. p. 6000-6010).

Literature

Go D. K. Nows it al. (2005 b.c.). The Encyclopedia of Ignorance. ET Publications/Jericho.

Relhom, Retep HP and Llennep, Nelletheb (2005). From Here to Next. Sunk Cost Publications. Googless/Queen Maud Land.

Received April 2005

Nonresponse Error in Surveys of the Dead[1]

Robert M. Groves[2],
Eesi Elpoepdaed[3]

Much of survey methodological research on nonresponse has led to erroneous conclusions because of use of sampling frames that include only living persons. This article builds on a speculation in earlier work of Lyberg, tests outcomes forecasted by that speculation, and shows that nonresponse issues are quite different among the so-called nonliving.

Key words: Nonresponse; latent life analysis; Lybergism.

1. Introduction

Nonresponse error is one of the most pressing issues facing statistical surveys (Lyberg 1902). Although it has its proponents, as in Godambe's (1987) defense of the crucial role of nonresponse error in the unbiasedness properties of the ratio mean in unrestricted random sampling, most research has concluded it's a pain in the ass.

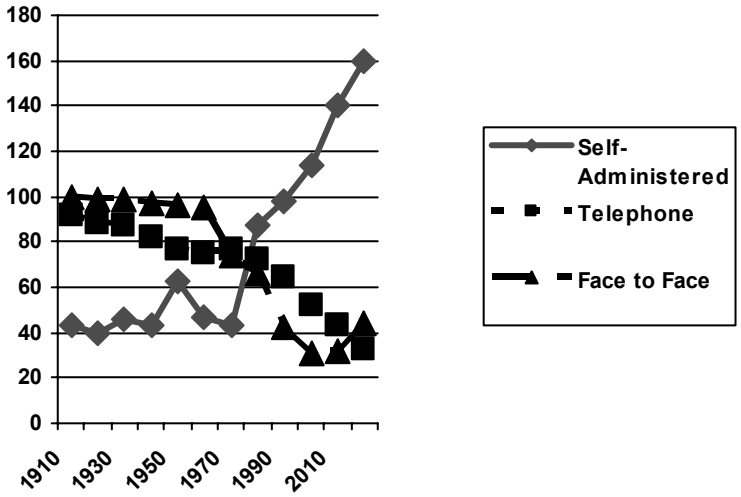


Fig. 1. Response rate by mood of data collection by year

[1] Funding for the research on persons prior to birth from the Right to Life Foundation; for that involving persons after death, from the Ted Williams Family Fund.
[2] University of Bath.
[3] Delecarlia Technical Institute. Dr. Elpoepdaed, a specialist in participant observation, was active at the beginning of this study, but mysteriously disappeared and has not been found as of this writing.

Figure 1 presents response rates (AAPOR 73.127, subpart Ж) over recent years. Both telephone and face to face surveys show declines in response rates over time, with large drops occurring around 1990, the start of the International Workshop on Household Survey Nonresponse. Self-administered surveys, in contrast, show gains starting in 1978 with the discovery of the Total Design Method; with rates climbing to far above 100%.

In an early 1619 manuscript in survey methodology, the Lyberg Genome Decay Theory was forwarded as a potential conceptual framework to describe causes of declining participation rates over time and over the lifecourse. Under the theory, $\xi(\varsigma)$ is not convex bounded within Martingale confined posteriors, instead when using AC current $\phi\left(\int \Im \sum \psi_{\delta}\right) = 6$ and $\varphi_{\alpha}\left(\Phi\Xi\right) = \sum\sum\sum\sum \eta_{i}\kappa_{j}\nu_{\kappa\varsigma}$ when the laptop is using batteries (consistent with the findings of Dalenius (1955)). In a handwritten note in the margins of the galleys of this early manuscript, Lyberg writes "This could easily been extended to explaining nonresponse in surveys" (first discovered by Andreenkov 1989). This paper applies the Genome Decay Theory to survey nonresponse over the life course.

2. Research Methods

A stratified multistage sample of the Swedish population register records from 1900 to 2012 was drawn and lifetime measurement of response propensity for all survey requests for each sample person completed. In addition, unobtrusive measures of autonomic thoughts were recorded continuously for all sampled persons. All human subject protections were compliant with those established by Project Metropolit. Following the implications of Genome Decay Theory of Survey Participation, we included measurements both before the birth date listed on the register and after the death date listed on the register. Consistent with the theory, we expected propensity for life to be causally related to response propensity in surveys. Both qualitative and quantitative investigations were conducted. The qualitative work involved ethnographers matched on propensity to be alive with those of the sample persons. No substitution was permitted, except the use of cyclamates instead of sugar.

3. Results

We first examine the descriptive estimates of response propensity after correcting for the censoring problems of restricting observations to those between 1900 and 2012. Figure 2 contains the essential support for the Lybergian assertion: response propensities are quite high pre-birth and post-death. This reveals a previously undiscovered solution to the nonresponse problem – the extension outside the conservative lifespan restrictions common to current sampling frames. While such an approach might offer attractive solutions to the nonresponse *rate* challenge, it may not, however, offer relief to nonresponse error concerns.

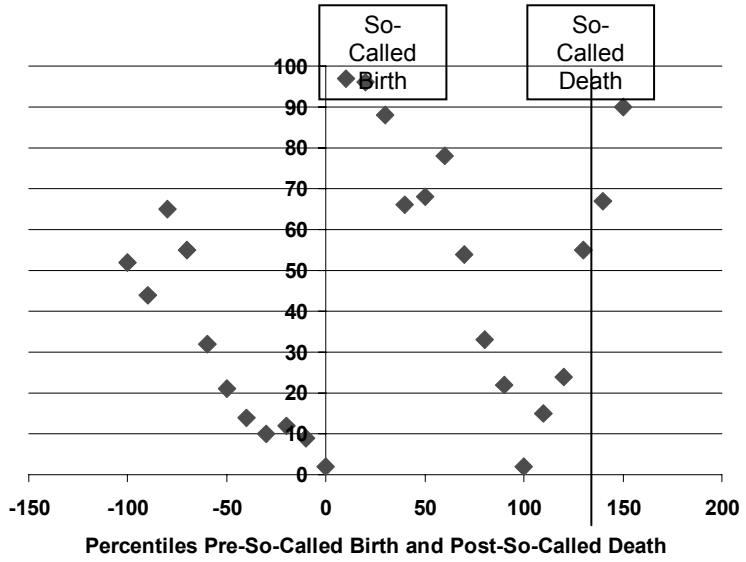


Fig 2: Pre-So-Called Birth and Post-So called Death in Percentiles

To address this, we use two totally separate analytic techniques to test for nonresponse error implications of the extended sampling frame – first a LISRELMXMVII model and second Monte Carlo simulation. The Lisrel model involving 247 terms and 1,419 equations showed no nonresponse error. For the simulation we used 1,213 replications using Gibbs sampling from an infinite number of priors and variance adjusted posteriors, under the assumption of MCAR. We found no nonresponse error, confirming the findings of the Lisrel models.

Finally, we report on a set of qualitative studies with retrospective self-reports. A common report of ethnographers in learning about reactions to survey requests among the pre-births was that they encountered difficulties in gaining well-elaborated answers.

4. Summary and Conclusions

This article is the first to test and prove the simple Lybergian assertion in the 1619 manuscript. We have learned that unpublished research findings (Lyberg, personal communication) show that negative incentives are effective for those prebirth, and postdeath sample persons. That is, that target population achieves higher response rates when they provide money to the investigator. Thus, the 1,619 marginal notes of Lyberg, as with much of his work, offers greater insights than originally thought. Lyberg notes that the entire business model of Statistics Sweden is undergoing renewal, with refocusing of studies of the nonliving.

Received April 2005

Editorial Note: After the disappearance of dr. Elpoepdaed, we now fear for the fate of Dr. Groves. Despite repeated contact attempts we did not receive a second version of the manuscript. So the interesting references are lost forever!

Beware of the Dog
A Review of Canine Nonresponse Bias

Ineke Stoop[1]

This article presents the results of a review of the literature on the effects of dog ownership on nonresponse rates. The (presumed) presence of a dog at sample addresses in face-to-face surveys has an impact on noncontact rates, and may be related to survey cooperation in general. Additional evidence indicates that dog ownership correlates with expenditure patterns and political preferences and may therefore be a determinant of a substantial amount of nonresponse bias. Independent evidence on dogs should be routinely collected in face-to-face surveys to allow nonresponse adjustment.

Key words: nonresponse bias; noncontact rate; survey cooperation; dogs

1. Introduction

In their seminal study on nonresponse Groves and Couper (1998) identify several factors related to survey participation by households, namely socio-environmental influences, socio-structural household characteristics and the socio-psychological make-up of the sample person. One major factor is systematically ignored, namely the presence of pets, and dogs in particular (see also Cohen this issue). Morton-Williams (1993, p. 33) identifies dogs as a disturbing influence in face-to-face surveys in which the selection of respondents is to a certain extent left to the discretion of the interviewer: "*Bias can also arise from the opportunity that the interviewer has to exercise some choice in which dwellings or which people to approach: the flat with the large dog, the house with filthy windows and decaying furniture in the garden or blaring forth loud music and the dwelling with an entryphone can be avoided*." In addition, several authors have shown the serious impact of dogs on contactability in probability samples. Finally, reluctance to cooperate may also vary between dog owners and non-dog owners. These findings, combined with empirical evidence that dog ownership not only correlates with the environment of the dwelling and household characteristics but may also reflect personality characteristics of the respondent, provides ample reason to collect independent evidence on dog ownership as a major determinant of survey nonresponse.

Literature on the evidence that dogs may be a serious impediment to contacting households is presented in Section 2. It turns out that a distinction needs to be made between two general types of dogs, namely security dogs and dogs predominant in crime-

[1] Social and Cultural Planning Office of the Netherlands www.scp.nl, Postbus 16164 2500 BD The Hague, The Netherlands. Email: i.stoop@scp.nl.

infested areas. Section 3 presents conflicting evidence on the relationship between willingness to participate in a survey and dog ownership. Section 4 emphasises the fact that this is serious since having a dog may correlate with consumption patterns and political preferences. This section gives a number of recommendations for further study.

2. Noncontacts and Type of Dogs

The presence of dogs in a household is generally seen as a serious impediment to contacting respondents. Many authors have drawn attention to the detrimental effect of dog ownership on accessibility, including Cornish (2002, p. 6) "*A listed dwelling might be too hard to find, access is difficult because of security and dogs, or people are never home when the interviewer tries to make contact.*", Bates (2003, p.3) "… *category 4 (PV - barrier) covers personal visit situations where interviewers could not access the sample household because of environmental barriers (drugs, crime, dogs) or physical barriers (buzzed entry, locked gate).”*, LTSA (2003) "*Here, 'non-response' includes refusals, households where no contact could be made after four attempts, households where no person spoke sufficient English to participate in the survey, and dwellings which were inaccessible because of security features or guard dogs*", and the New Zealand Ministry of Justice (2003, note 30) "*This means that the interviewer felt unsafe entering the property (for example, there were dogs or it appeared to be a gang house) or could not gain access (for example, because of a security fence).”*

A striking feature in this review (see also Morton-Williams, op. cit.) is the distinction between dogs that indicate a concern for security and dogs that are presumably regarded as an indicator of lower class (blaring music, decaying garden furniture) or crime (gangs, drugs). In studying the impact of drugs on survey nonresponse this distinction should be taken into account. It should also be noted that the concern about dogs is more widespread than this overview might suggest. Viragh (2000), for instance, presented a list of abilities and skills needed by the interviewer, one of which was a lack of fear of dogs.

3. (Temporary) Refusal and 'Beware of the Dog' Signs

Irrespective of the type and size of the dog, the verdict of the literature on contactability is unequivocal: dogs are an impediment to contacting respondents. Studies of survey cooperation and dogs produce conflicting results, however. More than 25 years ago the Dutch Interview-groep (1978, p. 9) carried out an in-depth study of nonresponse. One of their findings was that converted refusers were somewhat less affluent (possibly related to age) and more often had a dog (38% compared to 27% in other groups). The researchers concluded that the proportion of the household budget spent on dog food might be underestimated in surveys. Joye et al. (2004) had the presence of 'Beware of the Dog' signs recorded in the Swiss fieldwork for the European Social Survey 2002/2003. These signs were present at 1.8% of the sample households. In line with expectations, the sign was found almost twice as often at 'no contact' households though, contrary to the Dutch findings, it was found substantially less frequently at the dwellings of final refusals. One explanation might be that the Swiss researchers focused more on the 'Beware of the Dog' sign than the actual presence of a dog. This could be misleading: not

every dog owner might advertise its presence with a sign, while dogless people might put up a sign anyway, in order to deter unwelcome interviewers.

4. Discussion and Recommendations

This overview demonstrates a uniform negative effect of dog ownership on contacting respondents. More research is needed on the relationship between dog ownership and survey cooperation. These initial findings might turn out to be confined to face-to-face surveys, while the correlation with survey cooperation might apply irrespective of the mode. The results of this study are especially alarming as Dekker (1992) has demonstrated convincingly that dog owners are situated further to the political right than cat owners. One reason for faulty electoral predictions might therefore be the lower response rates of dog owners and the resultant nonresponse bias. For this reason it is strongly recommended that the presence of 'Beware of the Dog' signs and the actual presence of dogs, categorised by size (large/small) and type (guard/attack dog) be recorded and keyed during survey fieldwork.

5. References

Bates, N. (2003). Contact Histories in Personal Visit Surveys: The Survey of Income and Program Participation (SIPP) Methods Panel. In: Proceedings of the Annual Meetings of the American Statistical Association (ASA), AAPOR Conference.

Cornish, J. (2002). Response Problems in Surveys. Improving Response and Minimising the Load. Paper prepared for the UNSD Regional Seminar on 'Good Practices in the Organisation and Management of Statistical Systems' for ASEAN countries, Yangon Myanmar, December.

Dekker, P. (1992). Zijn hondebeziters rechtser dan kattebezitters? In: Paul Dekker and Marjanne Konings-van der Snoek (eds.) Sociale en Culturele Kennis. Rijswijk: Sociaal en Cultureel Planbureau. [In Dutch]

Groves, R.M. and Couper, M.P. (1998). Nonresponse in Household Interview Surveys. New York: Wiley.

Inter/view-groep (1978) Rondje Non-response van de Inter/view-groep. Amsterdam: Inter/view B.V. [In Dutch]

Joye, D., Schöbi, N., and Bergman, M.M.(2004). Response Propensity and Acquiescence Effects on Data Quality. Paper presented at the RC33 Sixth International Conference on Social Science Methodology, Amsterdam, August.

LTSA (2003). Public Attitudes to Road Safety – 2003, Appendix A: Sample details. Wellington, New Zealand, Land Transport Safety Authority.

Ministry of Justice (2003). The Nature and Design of the New Zealand National Survey of Crime Victims 2001. Wellington, New Zealand, Ministry of Justice.

Morton-Williams, J. (1993). Interviewer Approaches. Aldershot: Dartmouth Publishing.

Viragh, E. (2000). Issues of the Training of the Interviewers and Their Role in Treating Nonresponse. Paper presented at the 11th International Workshop on Household Survey Nonresponse, Budapest, September.

Received April 2005

About Interdisciplinarity and a Genetic Approach to Nonresponse in Surveys

Geert Loosveldt[1]

Survey methodology is characterised by a high degree of interdisciplinarity. It is argued that the increasing self-sufficiency of survey methodology should not curb interdisciplinarity. This is illustrated based on the possibilities offered by a genetic approach to survey nonresponse. Moreover, the underlying assumption is that non-response is a disorder caused by a defect in human genetic material. It is explained that this approach may bring a refreshing research agenda and that the genetically manipulated respondent would be the answer to many problems in survey research.

Key words: Interdisciplinarity; genetics; nonresponse.

1. Introduction

Interdisciplinarity is an important characteristic of survey methodology. It could be argued that a different discipline plays an important role for every component of survey research. Sampling theory, relying on the theory of probability, is used for the sample design and to determine design effects. An economic cost-benefit analysis is made for evaluating aspects of the survey design such as sampling design and mode of data collection. Cognitive psychologists made important contributions to understanding how respondents interpret questions and reply to them (Schwarz and Sudman 1996). This led to guidelines regarding the phrasing of questions and the order in which questions were to be placed. Conversation analysis has accentuated the insight into the interaction between the interviewer and respondents and served to clarify the interviewer's task in a structured interview (Maynard et al. 2002). Social psychology principles were particularly inspiring for identifying ways to obtain cooperation in an interview (Cialdini 1984). Even ordinary sociologists, using vague concepts to put survey participation into perspective (e.g., utilitarian individualism, confidence in institutions, political powerlessness, see Loosveldt 1999) form part of the interdisciplinary club of survey methodologists.

The integration of the contribution made by various disciplines has ensured that survey methodology has become an independent discipline. Some excellent recent manuals (e.g., Biemer and Lyberg 2003; Groves et al. 2004) and the launch of a European Association for Survey Research clearly illustrate this fact. The self-sufficiency of survey methodology has undoubtedly been a contributory factor in the quality improvement of survey data. However, there is a real threat that further interdisciplinarity may be curbed, resulting in survey methodological problems not being tackled adequately. The possibility of a genetic approach survey nonresponse seems to a case in point.

[1] Departement of Sociology, University of Leuven , E. Van Evenstraat 2B, 3000 Leuven, Belgium. Email: Geert.Loosveldt@soc.kuleuven.ac.be

2. On the Feasibility of a Genetic Approach to Survey Nonresponse

It is remarkable that survey methodologists with their interdisciplinary orientation have not explored the genetics. As a matter of fact, 'it's all in our genes' and genetics offer a fundamental insight into important aspects of the development of human behaviour. Certain deviations and defects in our development may be traced to an underlying cause.

When we approach survey nonresponse from the angle of genetics, we come up with some refreshing ideas. Nonresponse is a disorder caused by a shortage or defect in the human genetic material. It is a genetically determined inability, or lack of talent, to react positively to a request to cooperate with a survey interview. Viewed from this perspective, we can assume the existence of a gene responsible for our tendency to cooperate with survey research. This is a completely new fundamental hypothesis in nonresponse research. Apparent from this approach is that the tendency to non-cooperation is hereditary and not infectious. It opens up an entirely new research agenda. An urgent investigation is needed into whether the nonresponse of sons and daughters is related to the nonresponse of fathers and mothers. A close inspection is required of whether respondents can be carriers without developing the defect and which factors stimulate or inhibit its development. It cannot be excluded that increasing environmental pollution is a contributory factor in the negative consequences of this genetic defect coming to the fore more readily. More specifically, the growing hole in the ozone layer and the associated greenhouse effect spring to mind. This may explain the increasing non-response in industrialised countries. If the results of genetic research confirm the existence of a nonresponse gene, it is clear that the nonresponse problem cannot be resolved by further incentives and additional training for interviewers. A solution must be found in genetic manipulation. A confirmation of the existence of the nonresponse gene would open up many possibilities for more genetic research into respondent related survey errors. Plausible options are response tendencies, such as giving socially desirable replies, acquiescence and the refusal to reply to threatening questions. In the event that this promising research obtains positive results, the genetically manipulated respondent clearly offers the ultimate solution to many survey errors. Our recommendation is to allocate high priority to genetic research into survey errors on the research agenda.

3. References

Biemer, P. and Lyberg, L. (2003). Introduction to Survey Quality. New York: Wiley.

Cialdini, R. (1984). Influence: The New Psychology of Modern Persuasion. New York: Quill.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R., (2004). Survey Methodology. New York: Wiley.

Loosveldt, G. and Carton, A. (1999). Utilitarian Individualism and Panel Nonresponse. International Journal of Public Opinion Research., 14, 428- 438.

Maynard, Houtkoop-Steenstra, Schaeffer, Van der Zouwen (2002). Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview. New York: Wiley

Schwarz, N. and Sudman, S. (1996). Answering Questions. Methodology for Cognitive and Communicative Processes in Survey Research. San Francisco: Jossey-Bass.

Received April 2005

Survey Practices in 8 European Countries in 2004

Siobhan Carey[1]

This article updates the current position on survey practice in the eight European countries originally described in deHeer (2000) and identifies change in the period 1998 to 2004 in the capacity of European countries to undertake high quality household surveys.

Key words: Survey practice; survey methodology; European statistics.

1. Background

As part of the methodological review of the International Adult Literacy Survey led by ONS, the team involved conducted a review of the capacity of a number of European countries to undertake complex household based surveys to a high standard deHeer (2000). In particular, the paper sought to identify the extent to which unnecessary variation in survey practice existed and to what extent this could be avoided. The report concluded that there seemed to be a lot of variation in the way surveys were conducted and that most of this variation was not necessary. There was therefore considerable scope to improve comparability of international surveys through increased harmonisation of survey practice.

This article looks at the same countries as the original review and updates the information for 2004. Interviews were conducted with the same organisations (both public and private in all countries) and broadly covered the same topics as the original study. The article highlights aspects of the survey process and specific countries where improvements have been made but also identifies some aspects and countries that have apparent reductions in survey quality since the original research in 1998.

2. Sample Design and Sample Procedure

Since 1998 improvements have been made in the procedures for sampling in France, Greece and Portugal. In France improved procedures with regard to the updating of the Census file have been implemented and is now available to both public and private organisations free on the Internet. This file also contains all the information relating to household members and so can be used as a sampling frame for individual characteristics such as ethnicity or long standing illness. In both Portugal and Greece the possibility for private organisations to draw high quality probability samples is much improved since

[1] Office for National Statistics, London, UK.

1998 with the implementation of population registers which can be used for sampling purposes.
Only one country of those surveyed has had a reduction in capability to produce high quality samples. Until 2002 Sweden both public and private organisations in Sweden could use the national population register to carry out probability sampling. Following recommendations from Statistics Sweden the national population register was withdrawn and now both public and private surveys in Sweden use quota sampling.

3. Response, Nonresponse and Fieldwork Procedures

The original review identified room for improvement in a number of countries. It concluded that in general the use of the battery of minimum measures must be improved and that this seemed to be possible. In particular, the contact strategy must be improved in Germany, Italy, the Netherlands and for the nonpublic organisations in Greece and Portugal. The professionalism of the interviewer corps must be improved in Germany, Italy, the Netherlands and Portugal.

In 2004 it was apparent that some countries had managed to make astounding improvements in their fieldwork operations while others had disimproved. In both Portugal and the Netherlands a major programme of systematic improvements to address the causes of nonresponse have yielded dividends. In the Netherlands, where response to surveys has traditionally been low, many surveys conducted by the NSI are achieving response rates of 95%-97%. In Portugal, response rates to the Labour Force Survey are currently 98% and to the newly launched survey of Wealth and Assets, which has a high respondent burden of approximately 15 hours interview time, the response rate in the first year was 105%.

In the UK however, response rates continue to slide and response to the LFS has now reached 28% and is expected to decline further. This is partly attributable to a campaign launched by the Prime Minister who announced he would be “tough on response, tough on the causes of response” rather than the intended “nonresponse”. For once the population gladly heeded the PMs exhortations and refused to answer the door. A sharp decline in response was also observed in Sweden where the LFS achieves its (reduced) set target of 35% response. The decline in response is at least partly attributable to the decision by the Chief Methodologist in Statistics Sweden to reduce survey development costs by using the BLAISE questionnaires produced by ONS since he judged that all Swedes spoke perfectly good English. An error resulted in the Swedish interviewers being issued with the Welsh language version of the ONS questionnaire. In defence of the decision Statistics Sweden said that as a result, for the first time they had reliable estimates of the number of Swedes who spoke Welsh.

4. Data Processing, Weighting and Analysis

The 1998 review concluded that only NSIs seemed to have sufficient experience and expertise to carry out complex data processing, weighting and analyses to the minimum standards set. In 2004 many of the NSIs, in particular the ONS in the UK, had lost the ability to carry out even the most simple computations accurately following a major

investment by Government to modernise their systems. The new system was delivered without any functionality for some of the most basic tasks and the ONS now employs large numbers of administrative staff that manually collate survey results using five bar gates.

5. Summary and Overview

While some countries had made good progress in improving the quality of their surveys others seem to have disimproved markedly, most notably Sweden and the UK. In both these countries the review identified several areas where performance by the NSI was now less than that available through even the cheapest private survey organisations. A crude multivariate analysis identified the most significant factor in declining quality as attendance at international conferences. The variable was inversely related to improvements in survey quality so that countries that never attended made the most gains while those countries most frequently represented showed the biggest decline in quality.

6. References

De Heer, W. (2000). Survey Practice in Europe in Measuring Adult Literacy. London: ONS.

Received April 2005

Equating Response Rates in Cross Cultural Surveys: A Swedish – Dutch Example

Edith de Leeuw and Lilli Japec[1]

There is a growing literature indicating that response rates differ between countries. This threatens the validity of cross-cultural and cross-national surveys. We provide an overview of past research and suggested remedies to equate response rates. We will show that these old treatments are unsatisfactory and present a new revolutionary solution and cure-all.

Key words: Cross national surveys; international comparability; nonresponse trends; miracles.

1. Introduction

For the past 16 years international experts have been studying nonresponse in such exotic places as Oslo, Leuven, and Maastricht. Despite (or possibly because of) their efforts, nonresponse has been increasing over time. Still, some nonresponse experts are better in nonresponse than others, and there are relatively large differences between countries in response rates and in trends over time (de Leeuw and de Heer 2002). The potential risk for differential nonresponse error between countries is considerable and threatens the quality of international and cross-national research as the international leadership group on millennium quality points out (Lyberg et al. 2000). This article analyses causes of differential nonresponse, summarizes suggested remedies to equate response rates and presents a new and revolutionary remedy. As an illustration we use Holland and Sweden: two European countries that have much in common, but differ highly in response.

2. Why Response Rates Differ

Despite heroic attempts at data crunching, De Leeuw and De Heer (2002) could not identify clear factors that explain the differences in response rates between countries. One potential factor named is interviewer training and supervision, as interviewers do differ in their behaviour and attitudes across countries (Hox and friends 2002). Campanelli (this issue) shows how interviewer training can indeed be used to increase differences.

---

[1] The authors are executive secretaries at Lasse's Angels Inc. The views expressed are those of the authors and do not necessarily reflect those of our director.

Acknowledgment: We thank our present and former colleagues at Lasse' Angels Inc, and especially Birgit Glimenius, Gunilla Dahlén, Nancy Bates, and Pat Brick Dean, for their support and suggestions.

Fieldwork procedures and survey practices have been named as factors (see Carey this issue).

Others point to cultural differences; for instance Sayers (1975, p. 46-47) points out that English people won't fill up questionnaires, and that as a Nation the British are not questionnaire conscious. This in contrast to the Scottish, who as early as 1788 were able to obtain a 100% response on a mail survey for the Statistical Account of Scotland (Hacking 1990). Although prior beliefs (cf. Cialdini et al. 1991-1993) suggest otherwise, U.S. citizens are not great respondents when compared to the Swedes (Couper this issue). Like the French, the Dutch are prone to say 'no', be it to a European constitution or to surveys. It seems unlikely that response rates can be raised to the high Swedish norms, but remedies against nonresponse have been suggested and tried, as is reviewed in the next section.

3. Suggested Remedies

Numerous attempts have been made in order to increase response rates, but nevertheless response over the years has been decreasing. This has been documented over 16 years by the members of the International workshop on household survey nonresponse, guided by Lars Lyberg. When the reducers failed, the adjusters took over and suggested new standard definitions (APE/OR 2004) and new estimation techniques (O'Muircheartaig this issue).

Unfortunately all attempts have been in vain. In accordance with the old philosophy 'lay back and enjoy it & think of England', we suggest to go with the flow. Instead of trying to raise the international response rates to the uncanny heights of Sweden, it would be more cost-effective to reduce Swedish response rates and aim at a maximization of nonresponse worldwide. Carey (this issue) points out how changed survey practices may lower Swedish response. As a consequence Statistics Sweden has switched the goal of their strategic research programme from improvement to detoriation of response rates, and developed CBMs for response rate reduction (Japec et al. 1997, op cit Biemer and Lyberg 2003, page 369). As a by-product, surveys costs have gone down dramatically in Sweden. This is a big step forward in equating response rates for cross-cultural studies and especially equating survey response rates in Sweden and Holland. Hopefully the Scottish will follow this excellent example. Still, some researchers lack the flexibility needed for this dramatic improvement in survey methodology. For instance, a major obstacle at Statistics Sweden that stands in the way of further detoriations of response rates is Lars Lyberg, who continues to insist on reducing respondent burden, sending advance letters, training interviewers, conducting nonresponse follow-ups; in short on increasing response rates! For a concise summary, see Biemer and Lyberg, Chapter 3.

4. Recommendations and Conclusions

Fire Lars Lyberg!!!!!!

5. References

APE/OR 2004. *Standard Definitions: How to Finally Dispose of Cases That Refuse to Cooperate.* Lenexa: Kansas.

Cialdini, R.B. and various others. International workshops on household survey nonresponse, 1991-1993.

Hacking, I. (1990). The taming of chance. Cambridge: Cambridge University Press.

Hox, J, de Leeuw, E., Couper, M., Groves, B., deHeer, W., Kuusela, V., Lehtonen, R., Loosveldt, G., Lundqvist, P., Japec, L., Martin, J., Beertens, R., Michaud, S., Knighton, T., Mohler, P., Sturgis, P., Campanelli, P., Vehovar, V., Zaletel, M., Belak, E. (2002) The Influence of Interviewers' Attitude and Behaviour on Household Survey Nonresponse: An International Comparison. In: R.M. Groves, D.A. Dillman, J.L.Eltinge, R.J.A. Little. Survey Nonresponse. New York: Wiley.

de Leeuw, E.D. and de Heer, W. (2002).Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In: R.M. Groves, D.A. Dillman, J.L.Eltinge, R.J.A. Little. Survey Nonresponse. New York: Wiley.

Lyberg et al. (2000) No Leg to Stand on: Summary Report from the Leadership Group (LEG) on Millennium Quality.

Sayers, D.L. (1975). Gaudy Night. New English Library.

Received April 2005

Plain Old Data, Para Data, and Meta Data: The Three Sopranos of Data

Fritz Scheuren[1]

Here we give some historical insights into one aspect of the foundational mathematical work of Euclid, as retold by Aesop. In particular, we confine attention simply to the theorem of the three little pigs. Included also is a long overdue reinterpretation of the misunderstood figure of the wolf. When put in her proper statistical context, the wolf's role in quality becomes clear.

Key words: The three houses of quality; fitness for eating use; constancy of purpose or constancy in the reduction of portliness.

1. Menu

The present article has been written to put back into its proper mathematical statistical form what has come down to us as the story of the three little pigs. We would argue, in fact, that this revision is long overdue (See Kyle 2005 for more background).

While quite familiar, the three little pig story is very sparse on essential details and completely omits the most important points. For example, the names of the three little pigs are missing from all versions that we have. Almost all we know is that they are supposed to be "little." And that, in fact, can hardly be true since most quality pigs (or problems as they are also called) are big.

Rubin (1976) seems to have alluded to the story of the three little pigs when he wrote about missing data. Indeed the pigs' last name is Data and it has been long missing. What Rubin's paper, despite its seminal nature, does not do is to give the pigs their complete or multiple names. This is done here, for the first time, along with how we were able to impute them (See Section 2).

Moreover the role of the wolf has been completely misunderstood, seriously hampering our use of this example in quality theory. Most of the time the wolf is cast as the villain – when, in fact, she is the real hero. Arguably, the idea of constancy of purpose that Deming (1986) made a centerpiece of his quality message may have come about as a result of the wolf's role in the traditional story.

---

[1] NORC University of Chicago.

Acknowledgments: Some of the ideas here and many of the references in this paper come from the contributions of Patrick Baier, Susan Hinkins, Robin Lee, Lars Lyberg, Susan Kyle, Michael Kwanisai, Ali Mushtaq, and Joe Walker. Thanks are due too to the editor for her forbearance in all matters pertaining to this article. The author, of course, is responsible for all misattributions and other gaps in the retelling of this important quality story.

It is believed by the author that in the original versions of the three little pigs, no longer extant, the wolf was sometimes praised for his "constancy of portliness" or, when expressed more fully, the wolf's constant desire to reduce portliness. We are unclear as to whether Deming may have independently come up with the idea of "constancy of purpose." But, in any case, "constancy of portliness" appears to predate "constancy of purpose" in the quality literature by thousands of years.

For more on this last point, see the recent updates on the three-pig theorem. One source is the June 2005 issue of *Quality Progress*, which features several articles on "Lean" – formerly lean production (Womack et al. 1991).

2. Three Houses of Quality

In the story, to escape the wolf, each of the pigs builds a different kind of house. Incidentally, this is where the idea of the houses of quality comes from (e.g., Angst and Newman 1973). Again, alas, without full attribution.

As you will remember, one of the pigs builds a house of straw, the second of wood and the third of brick (actually this was of Mike Brick and hence very sound indeed).

We do not know much more about these pigs than this. In most versions, the pigs triumph over the wolf, but in the versions I like best, the pigs or the problems are dealt with by being eliminated or at least reduced by the wolf -- succumbing to the wolf's constancy in the reduction of portliness.

Now, you will ask, how can you impute the pigs' names if this is all you know? Well I employ ideas taken from both new and continuing survey practice:

Despite the importance of paradata (e.g., Lyberg and Cooper 2005), the profession has, so far, given it little permanency in the Data family, particularly in survey Data files (even though advocated by Scheuren 2005). Hence the pig that built a straw (bamboo?) hut must have been called Para Data, otherwise how would the concept have originated?

Meta Data must be the pig that built his quality house of wood. Why? Well, we know that this pig had to cut down trees to build his house. Hence, since he was in a hurry for fear of the wolf, the wood must have still been green. The concept of metadata is still new (e.g., Dippo and Sundgren 2000), so QED.

By a process of elimination we are led to the revelation that the pig that built the house of brick must have been Plain Data -- better know by his full name Plain Old Data. In the survey field he was the oldest too and that is another tip-off.

These are the names anyway that are given in the paper's title and I will stick with them for now, although a cousin in the Data family gave me some information that I have put into the paper's subtitle which is the "The Three Sopranos of Data." Now this information comes from a cell phone conversation with poor reception and one that ended abruptly, so I am unable to say whether this Data family detail about the three pigs' occupations can be taken to mean that the pigs, before their encounter with the wolf, sang as they worked (liked the mythical seven dwarfs) or, as in the TV series, that they had more sinister vocations.

3. Constancy in the Reduction of Portliness

In the Aesop fable the wolf certainly demonstrates her constancy of in the reduction of portliness or "constancy of purpose" as Deming called it.

There is a cautionary note, though, about the wolf to emphasize here that comes from the traditional retellings. As was mentioned, most of these different versions have the wolf defeated and, indeed, in many cases dying a horrible death. Certainly that can happen to those who follow the path of quality as unarmed, as the wolf was. Her good problem-solving appetite, good statistical tools (or forensic claws), and good intentions were simply not enough. She had to have sustained upper management support (in this case from the story teller.)

What to do, then, if you are a survey quality wolf? Well if I were to choose only one source of advice on survey quality it would obviously have to be Biemer and Lyberg (2003). But you could also try the advice of Juran (1988). He emphasizes that you need to fool an organization's immune system if you are to achieve quality. I have found that this can work with survey Data family problems. While it cannot be confirmed, Juran may have been dealing mainly, unlike Deming, with problem sheep and not problem pigs. I am not sure, but could Juran be the source of the expression "A wolf in sheep's clothing"?

Since I expected to be asked about the name of the wolf, I traveled widely, including to the Orient (e.g., Ishikawa 1990), to find more quality ideas After all I had given you the names of the three pigs and owed you, dear reader, the rest of the story. Anyway, after long meditation and years of study, I am fairly certain that the wolf has the noble, albeit unusual name of *Kaisen* or "Continuous Improvement." I should have expected this but, I must admit, was still surprised at how obvious the name was, once known.

4. Next Meals

Usually my papers end with something called "Next Steps." Maybe "Next Meals" would be a better goal here -- certainly for those of you aspiring to be quality wolves someday. Anyway writing or even reading this paper can give one an appetite for quality problems, so I could end with the one word chant Data, Data, Data – immortalized in *The Graduate* (Hoffman 1978).

Some authors in closing similar papers (Barnum 1881) have used the phrase "Bring on the Clowns." I do not recommend clowns, however. They are not nearly as good eating as pigs or, if you must, sheep. The chant "Bring on the Pigs" would have worked, except for its late 1960s associations (Daley 1968). So **Data** is my watchword. "Bring on the Data!"

The editors, after rejecting the current paper several times as being too plausible, finally accepted it but reminded me that Aesop's fables always end with a moral. And I had to give one here! Unaccountably, in the versions of the three little pigs that I found, no satisfactory moral was ready at hand. What I came up with, as a quality moral, was the need in survey production problem solving to employ what in other settings (Stewart 2005) has been called "lean cuisine." Since I am out of space and time, a better resolution

of this issue is left to you dear reader. Please share whatever you come up with – be it, an "Eat-In," a "Take-Out" or some other form of "Take-Away."

5. References

Angst, B. and Newman, A. E.(1973). What Me Chance To Worry?. San Francisco: Long Gone Press. For the generally accepted view, however, see Brackstone G.(1999). "Managing Data Quality in a Statistical Agency." Survey Methodology, 25, 139–150.

Barnum. P.T. (1881). Zum Dreisatz in Noetherschen Ringen. Journal fur Irrelevanztheorie (17), Schweinfurt. This reference cannot be completely authenticated but the historical record is clear that Barnum was in Germany in the 1880's. Whether he had the elephant Jumbo with him is unknown, so only one author has been given.

Biemer, P. and Lyberg, L. (2003). Introduction to Survey Quality. New York: Wiley. For a fully operational example of top survey quality, see the work of Arthur Kennickell and his colleagues in the Survey of Consumer Finances, found at http://www.frb.gov/

Cooper, M. and Lyberg, L. (2005), The Use of Paradata in Survey Research. International Statistical Institute Meetings April 2005, Sydney. Paradata is data about the processes with which plain old data are obtained. Response rates would be an example. Number of contacts on a case before obtaining a response would be another. Paradata are generally considered to be only a subset of metadata – metadata being the broader concept..

Daley, R. (1968). The Recent Chicago Democratic National Convention. Understandably only privately published – a memory thankfully now fading fast.

Deming, W. E. (1986). Out of the Crisis. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study. It must be acknowledged that, except for this one cite, the other references to Deming in this article are not substantiated anywhere in the extant quality literature but were told to the author by a generally unreliable source.

Dippo, C. and Sundgren, B. (2000). The Role of Metadata in Statistics. See also Colledge, M., and E. Boyko. 2000. Collection and Classification of Statistical Metadata: The Real World of Implementation. Both of these papers were presented at the Proceedings Second International Conference on Establishment Surveys in Buffalo.

Hoffman, R, (1978). From an unpublished essay entitled The Future of Plastic and its Data Derivatives. By the way, this reference arguably is hearsay and, except for the desire for completeness, would not have been included.

Ishikawa, K. (1990). Introduction to Quality Control. Tokyo: 3A-Corporation.

Juran, J. M. (1988). Juran on Planning for Quality. New York: Free Press. While an attempt was made to substantiate the Juran references to sheep in this volume, time did not allow my legion of unnamed graduate students to do their usual superb job.

Kyle, S. (2005) Mathematical Fables Revisited. New Age: Wild West. As detailed in this reference, the story comes originally from a lost book of Euclid's Elements. The

Euclid version is only in a fable form, later attributed to Aesop. Confusing matters the story of the three pigs is not found in the definitive 1937 Aesop edition in the Harvard Classics. Instead interested readers should look at Aesop's collaboration with Gregory Maguire, See also Zhu, Lang, and the Jiuzhang Suanshu (as translated by Cao Xueqin) on HardtoFindPress.com.

Robin, L. (in press). PoomJil HyangSahng HahnDahGoYo Hahn Kook Tong Gye Hock Bo**.** Unfortunately, in searching for this important Korean reference I have so far come up empty-handed but I have cited it anyway, in the unlikely event it should ever actually appear.

Rubin, D. (1976). Inference and Missing Data. Biometrika**.** See also Kwanisai, M. (in Press). Zvekuita pasina humbowo. The second (Zimbabwe) reference, because of its expected importance (although not read), has been included also. The literature on missing data is, of course, enormous with many other citations to be found in French, Spanish – indeed in most languages. However, in keeping with the topic, it seemed better (except for the two cites already given) to treat further references as missing at random, even though many are nonignorable.

Scheuren, F. (2005). Seven Rules of Thumb for Nonsampling Error in Surveys. National Institute of Statistical Science (NISS) Total Survey Error Conference, Washington, March 2005. For still more advocacy in the use of survey paradata, see also Scheuren, F. (forthcoming). The Role of Paradata at all Stages of Survey-Going from Concept to Completion, a paper to be presented at the Fall 2005 Statistics Canada Methodology Conference.

Stewart, M. (2005) Recipes from My Jail Cell. Now, while this reference has not been read, I did hear it sung to the old tune "If I was an angel, over these prison walls I would fly."

Womack, J., Jones, D., and Ross, D. (1991). The Machine that Changed the World: The Story of Lean Production. HarperPerrennial. Sadly, again, despite this being a wonderful book (and highly recommended), there is no mention of the Three-Pigs Theorem, not even anything about lean bacon, whether American or better yet Canadian.

Received April 2005
Revised April 2005
Revised April 2005
Revised April 2005
Revised April 2005
Revised April 2005
Revised April 2005

The Hidden Menace – Measurement Errors in the Absence of Measurement

Colm O'Muircheartaigh[1]

This article demonstrates the magnitude, and interprets the significance, of measurement errors in the crucial case where the number of elements successfully measured is zero. While major effort has been devoted to peripheral issues such as sampling and nonresponse, this core problem at the intersection of these areas has been ignored. The article also outlines the historical reasons underlying this reprehensible failure.

Key words: Inconsistency; Stuft; Gap.

1. Introduction

There has been a regrettable failure on the part of the survey profession to pay sufficiently serious attention to a major threat to the validity and usefulness of survey research in general. This neglect can be contrasted with the sometimes excessive attention (and consequent respect) paid to issues of nonresponse (Fisher 1921, Groves 1956); sampling (Dalenius 1955; Kish 1954, 1965); and process quality (Student 1899, Lyberg and Biemer 1990, 1991, 1992)[2].

Even when researchers have considered measurement error, the emphasis has been misguided. With only one exception (Lyberg 1962), which includes everything, all of the published papers and bibliographies concentrate on cases where the number of elements measured is greater than zero, i.e., $n_m > 0$; the critical case of $n_m = 0$ is ignored. It is this major gap in the literature that this article addresses.

One of the reasons that the area has been neglected is the mistaken belief that nonresponse is a problem. Fisher (1921) was an early victim of this belief, though it could be argued that it was his struggle with nonresponse that led to the development of experimental design as a field (Fisher 1921, 1924). We now realize (Murphy and O'Muircheartaigh 2004) that the real problem is not nonresponse but nonresponse error. This powerful reconceptualization is best illustrated by Groves (2006), the title of whose paper (74 pages, including 137 diagrams and three formulae) says it all.[3]

[1] NORC and Harris School of Public Policy, University of Chicago, U.S.A.

[2] Indeed, what serious scientist knows what "process quality" means?

[3] "Response rates are irrelevant"

2. The Problem

Consider the standard total error model, based on work by Hansen and his colleagues at the U.S. Bureau of the Census. From this model we can derive a number of indicators of data quality. The index of inconsistency, I, is probably the most robust measure. The index summarizes quality as the proportion of the measurement that is attributable to measurement error; this is related to the signal to noise ratio in engineering. Below two different approaches to the explication of the problem are presented: (i) a calculus approach; and, (ii) for those not comfortable with calculus, an applied approach.

2.1. A calculus approach

For reasons of space, the editors have decided not to present the mathematical formulas in the body of the paper. They can be found in the special supplementary volume published at the end of each year.[1] A non-mathematical interpretation of the mathematical approach is presented in the next section.

2.2. An applied approach

Consider the index of inconsistency. This is the ratio of the simple response variance (SRV) to the simple total variance (STV); the STV is the sum of the SRV and the simple sampling variance (SSV). Both the numerator and the denominator are functions of sample size. As sample size decreases the simple response variance (SRV) converges to its expected value. However, as sample size decreases the simple sampling variance (SSV) decreases, as the variability of the observations approaches zero. The sampling variance reaches zero in the limit when the sample size reaches zero. [For non-technical readers, when you have no sampling, you cannot have sampling variance.]

Consequently the index of inconsistency, I, the ratio of the SRV to the STV, approaches 1 as the sample size decreases, and when the size of the observed sample (n_m) reaches zero, the ratio reaches this maximum. Consequently, and paradoxically, measurement error is not only still relevant when you have no measurement, it is the only issue. Of course, one approach worth considering (based on *stuft*) is that used in the context of sampling by O'Muircheartaigh and Magilavy (1979).

3. Conclusion

This article points out a little understood consequence of the research on nonresponse. Far from providing us with reassurance that all is well when we maximize nonresponse (or minimize response rates) as suggested by the Dutch-Swedish collaborative project, this article demonstrates that the true outcome is that the importance of measurement error is increased. Consequently when the limiting value of 100% nonresponse is achieved, we simply create an even greater threat to the integrity of our research. The threat is exacerbated by the complete absence of data to illustrate it.

The time has come to set up a working group on measurement error to replace the nonresponse group, whose work is done. Such a working group could, in just a few

[1] *Deleted sections of published papers*. JOS Supplement, 2006. Available on request.

decades, provide a useful agenda for further research on this topic. It is not for nothing that we call it the Hidden Menace.

4. References

Dalenius, T. (1955). Sampling in Sweden.

Fisher, R.A. (1921). Nonresponse in Tea-tasting Surveys. Survey Methodology, Preprints, 12-31.

Fisher, R.A. (1924). Experiments: A New Approach to Tea-tasting. Journal of the Royal Statistical Society, 87, 1-27.

Groves, R. (1956). Misunderstanding "No": The Identification of Refusals in Kindergarten Research. Sociological Theory, 23, 562-589.

Groves, R. (2006). Response Rates Are Irrelevant. Journal of Official Statistics, 22, 1-137.

Kish, L. (1954). Differentiation in Metropolitan Areas. American Sociological Review, 19, 388-398.

Kish, L. (1965). Survey Sampling. New York: John Wiley and Sons.

Lyberg, L. (1962) Nonsampling in Sweden: A Bibliography [work carried out by T. Dalenius].

Lyberg L. and Biemer, P. (1990). Some Thoughts on Process Quality. Journal of Official Statistics, 6, 1-14.

Lyberg L. and Biemer, P. (1991). More Thoughts on Process Quality. Journal of Official Statistics, 7, 1-23.

Lyberg L. and Biemer, P. (1992). Previously Unpublished Thoughts on Process Quality. Journal of Official Statistics, 8, 1-35.

Murphy, W. and O'Muircheartaigh, C. (2004). Maximizing Unwarranted Assumptions in Nonresponse Research. Journal of the Royal Statistical Society, Series C, 48, 241-273.

O'Muircheartaigh, C. and Magilavy, L. (1979). Get *Stuft*: A General Answer to Unwelcome Complexity. Disciplinary Readings in Survey Research, 1, 1-99.

Student (1899). Process Quality in St James' Gate Brewery in Dublin: Some Data. Industrial Quality Control, 2, 17-57.

Student (1912). The Irrelevance of Data to Statistical Theory. Journal of the Royal Statistical Society, 75, 1-27.

Received April 2005

Population Design and Survey Quality

Thomas Körner[1]

Until very recently, survey research has been dominated by sampling based methods. The complementary approach based on population design focuses on the quality of the target population rather than that of the sample. The resulting survey procedures make life much easier for survey researchers and help reducing many sources of error, including nonresponse and coverage errors. This article introduces random as well as nonrandom methods for population design and discusses their implications on data quality.

Key words: Quality of target populations; probabilistic design method; self fulfilling samples; journal anniversaries.

1. Introduction

In the last decades, survey research focussed almost exclusively on sampling techniques. The measurement of the quality of survey results was often reduced to the quality of the sample, namely the error due to selecting a sample instead of the entire population. The perspective was slightly changed by scholars like Lyberg and Biemer (2003) who pointed out the various types of nonsampling errors (like coverage errors or nonresponse errors). This change certainly widened the perspective, but did not change the sample-centred approach in principle.

The research carried out in this tradition tended to focus too narrowly on the quality of sample, be it related to the sampling variance or the bias. With the quality of the target population another feature largely affecting the quality of the survey as well as the quality of life of statisticians was ignored.

It is evident that the characteristics of the target population are complementary to those of the sample. Consequently, describing the total survey error as a function of the sampling error and the nonsampling error neglects the role of population quality. The 2001 *International Conference on the Quality of Target Populations* (Q2001) succeeded in gathering population design experts from Europe and the world. Q2001 helped establishing a new dimension of survey methods, taking the perspective of population design rather than that of survey design.

This article tries to summarize some of the approaches developed in the last years, culminating in the latest developments of the probabilistic design method (PDM).

[1] Federal Statistical Office Germany, 65180 Wiesbaden, Germany Email: thomas.koerner@destatis.de
The opinions expressed in this article do necessarily not reflect those of the author's institution or of the author himself.

2. Population Design Methods

In analogy with sampling methods one can distinguish random and nonrandom population design. The main difference is that nonrandom methods define the population *ex ante* whereas the random population design approach uses *ex post* target populations.
The most important nonrandom population design methods are quota based approaches. In quota population design, the composition of the target population is specified on the basis information available regarding the characteristics of respondents and nonrespondents. Here, the results of nonresponse studies are a valuable input (see Lyberg et al. 1997). If necessary, a screening prior to the survey can provide additional information. The information available enables researchers to tailor the population according to the readiness for responding to survey questions. This largely facilitates fieldwork efforts and minimizes the costs for interviewer training and follow-up processes. Data collection is carried out until a predefined number of statistical units (satisfying the predefined selection criteria) have been interviewed.

However, satisfying the quotas is still prone to some uncertainty, as some units might refuse participation despite the careful selection of the quota criteria. For this reason, random population design has been developed. According to this method, also referred to as the probabilistic design method (PDM) or self fulfilling sample method, survey units are selected at random with *ex post* population design. After the completion of the fieldwork and the data processing the inclusion rules of the population are defined. Thus, the population design relies on the empirical evidence of which units did *de facto* respond. Constructing the sample *ex post* leads to a very high quality of the survey results, which can be easily shown by the use of a number of quality indicators, like the unit nonresponse rate or the coverage error.

3. Conclusions

Population design methods can make life easier for statisticians. Similar to quota sampling methods, quota population design is easier to administer. From the point of view of the users it has the advantage of a higher predictability of survey results (given that the selection criteria have been chosen carefully), a quality component too often ignored. In PDM surveys the field costs are considerably higher. However, this drawback is more than compensated by the fact that coverage errors and nonresponse errors in PDM surveys no longer play a role of any importance. Further advantages include a reduced preparation time for the questionnaire development, enabling us to make full use of the ugly design method (UDM) for mail, telephone, and internet surveys. A wider use of population design methods can finally help statisticians to find the time for an appropriate preparation of anniversary parties of important scientific journals.

4. References

Lyberg, L. (ed.) (2001). The International Conference on the Quality of Target Populations (Q2001), Stockholm, Sweden, 14-15 May 2001, Proceedings.

Lyberg, L. et al. (eds.) (1997). Survey Measurement and Population Quality. New York: Wiley

Lyberg, L. and Biemer, P. (2003): Introduction to Survey Quality, New York: Wiley.

Received April 2005

The Average Quality Pyramid in the European Statistical System – New Developments on Quality Dimensions in the European Union

Werner Grünewald and Håkan Lindén[1]

Quality in statistics is at the centre of statistical work in the European Union. Its definition was to a large extent influenced by a fairly unknown scientist, researcher and official. The article shows the influence on the definition, the outcome of his work and proposals for (his) further improvement.

Key words: Quality management; pyramid; chaos.

1. Introduction

The definition of quality in statistics currently used in the European Union consists of six dimensions: relevance (R), accuracy (A), punctuality and timeliness (PT), accessibility and clarity (AC), comparability (COM) and coherence (COH) [Eurostat 2003][2]. Though there is a general agreement on the different aspects of quality, no consensus could be reached so far on their hierarchical relations. Brackstone [Brackstone 2001] gives one hierarchal view, but we will show with the help of the path breaking work of a fairly unknown Swedish, 60 years old scientist, researcher and official, that the hierarchy of the quality dimensions is much more complicated. Ranking the different quality aspects with respect to the different roles of a human being will lead to different quality dimensions hierarchies and shows the need for an extension of the current approach.

2. The View of a Survey Specialist

For a survey specialist, it sounds logical to single out (A) as the more or less only important quality dimension. All other dimensions play a minor role. This view is based on the fact that the survey specialist is, indeed, a specialist in defining all possible survey errors (for details see Biemer and Lyberg 2003) without really being in a position of assessing the total survey error.

---

[1] The authors are members of a well-known but confidential organisation. The opinions expressed are those of an unknown referee and do not necessarily reflect those of the authors.

[2] The former idea of including a seventh dimension of 'completeness' was given up as this definition is already so complete.

3. The View of a Political Adviser to the Director General of a Statistical Office

The role of the political adviser to the Director General of a statistical office in terms of quality management sees the dimensions (PT) and (COH) in front, followed by (A) and (R), and (COM) and (AC) being least relevant. Coherence is of particular importance for the continuity of quality management in a statistical office as this role might consume a few Directors General.

4. The View of a Professor in Statistics

Looking at the view of a professor in statistics, (R) and (AC) are on top, (A) and (COH) being next and (COM) and (PT) at the end of the ranking. The importance of (R) and (AC) is based on the fact that this role often provides really popular and full-house courses, though their success is also a function of the age of the course leader, in particular with respect to the percentage of female students attending.

5. The View of An International Statistician

For an internationally active statistician, all quality dimensions are of high importance – except (AC). This role does not require access to data and their understanding. It is enough to highlight the concepts and leave it to the later user not to understand how to use the concepts proposed. By-products of this role are the collection of latest news (called rumours) and know-how back to the home institution and the share of best (or worst) practices…

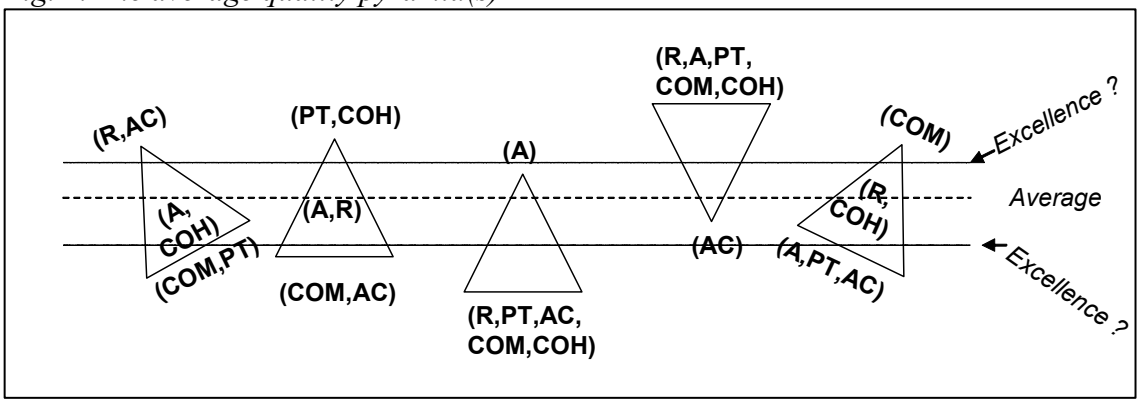
6. The View of the Chairman of a European Expert Group

For the role of the chairman of a European expert group, (COM) is by far the most important quality dimension, followed by (R) and (COH) at the second level and (A), (PT) and (AC) at the bottom. The outstanding importance of (COM) is caused by the European efforts to improve comparability of European data and aggregates though these efforts often result in fairly large numbers of recommendations for further work.

7. Conclusions and Recommendations

Combining the rankings of the six quality dimensions in the different roles leads to a complete chaos. All dimensions are either on top, in the middle or at the bottom (see figure below).

Fig. 1. The average quality pyramid(s)



No clear ranking is possible. The ranking depends on the role. It is therefore recommended to

a) Either restrict the use of the quality dimensions to just one role per person,

or

b) Extent the current set of quality dimensions by a seventh one capturing the personality of the user.

8. References

Biemer, P. and Lyberg, L.E. (2003). Introduction to Survey Quality", Wiley, New Jersey.

Brackstone, G. (2001). Managing Data Quality: The Accuracy Dimension. Paper presented at the International Conference on Quality in Official Statistics, Stockholm, Sweden, May 2001.

Eurostat (2003). Definition of quality in statistics. Document n° Eurostat/A4/ Quality/03/General/Definition.

Received April 2005

Survey Inequality and Official Statistics: A Casual Approach to Privacy Preservation

Stephen E. Fienberg[1] *and Miron L. Straf*[2]

The survey quality movement has radically changed the nature of official statistics. In this article we reverse the approach using tools from the recent literature on casual modeling, such as directed optimal graphs (DOGs), to study survey inequality. We develop a new method of imputation that allows us to replace traditional sample surveys by anonymized fake census records.

Key words: Multiple obfuscation; DOGs and other statistical animals; informative nonresponse; latent tendencies; survey quality, survey shmality.

1. Introduction

The quality movement has radically changed the nature of official statistics (see Lyberg and Sundgren, 2005). In this article we reverse their approach and study survey inequality using tools from the recent literature on casual modeling, e.g., directed optimal graphs (DOGS), as illustrated in Figure 1.

2. Casual Model

Following Freedman (2004), we applied Burridge's (2003) extension to Rubin's model for casual inference and imputation and developed a multiple obfuscation method for survey nonresponse. We begin with a hierarchical linear model at multiple levels (See Goodman, *Analysis of the Paths Less Traveled By*), and adopt a Dirichlet process prior with base measure that is based on the empirical inequality among survey responses. This is designed to produce a proper posterior from which we can sit and contemplate survey quality and well as generate anonymized fake records.

[1] Department of Statistics, Center for Automated Learning and Discovery, and Cylab, Carnegie Mellon University, Pittsburgh, PA 15213-3890, U.S.A. fienberg@stat.cmu.edu

[2] Division of Behavioral and Social Sciences and Education, The National Academies, Washington, DC 20001, U.S.A. mstraf@nas.edu

Acknowledgments: The preparation of this article was supported in part by a grant from the Humane Society of Pittsburgh. We thank David Freedman whose figure inspired our example and simulation model and two referees who clearly missed the point of the article.

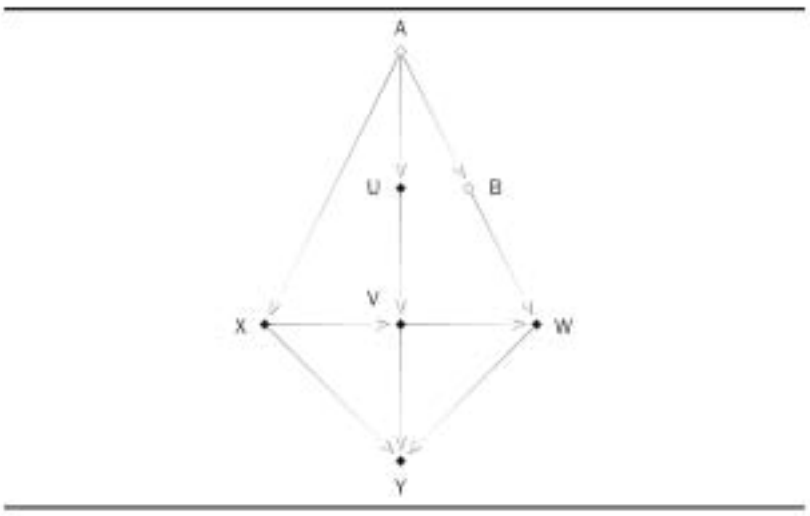


Fig. 1: DOG Model with Indigenous Variables U, V, W, and X, RepresentingSurvey Inequality and Latent Variables A and B Which Measure the UnderlyingUnobservable Survey Quality. Source: Freedman (2004).

Tourangeau, in his consummate exposition (*Don't Ask Me*), has identified three sources of why responses to surveys differ, i.e., are *unequal*: First, respondents don't know what's going on. Second, interviewers don't know what's going on. Third, users don't care what's going on. We represent these sources as indigenous variables (U, X, V, and W, respectively) in the DOG model of Figure 1. Our model captures how these sources of inequality add up through the acyclic cognitive pathways. Although this model is not the best one for our purposes (see Freedman 2004) who explains why the assumptions are untenable), there is a bias among survey researchers towards its use. We employ latent class analysis (represented by latent variables A and B in Figure 1) to estimate this bias and correct for it. Finally we impugn nonresponse with Rubin's multiple obfuscation methods (1987) by drawing on our posterior, if the nonresponse is ignorable. If it is not, we forget about it.

Let y_{ijklmn} be the response of respondent i to survey j in sample k to question l to interviewer m in replicate n. Then, according to the inclusion-exclusion rule:

$$\sum (y_{\text{ijklmn}})^2 + \lambda \geq 0,$$

where λ is a penalty for bad survey questions. The implication of this relationship is that survey errors are not going to get any better than they already are. Or, as Lars Lyberg has phrased it: "Get over it." If the response $y$ijklmn cannot be released without violating the confidentiality of respondent i, we suggest using the nearest neighbor rule. Ask the nearest neighbor i' to guess what response would be received from i. If i' guesses correctly, use that response. Otherwise, go to the next nearest

neighbor.

We applied our model and methodology to the Longitudinal Swedish Survey of Male Massage Therapists, and generated 5 replicates of fake responses for female massage therapists. These replicates satisfy the privacy-preserving data-mining criteria outlined in Dalenius (1986).

3. Conclusion

Our methodology illustrates the extent to which survey quality in Sweden has gone to the DOGs.

3. References

Burridge, Jim (2003). Information Preserving Statistical Obfuscation. Journal of Official Statistics, 13, 321–327.

Dalenius, T. (1986). Finding a Needle in a Haystack or Identifying Anonymous Census Records. Journal of Official Statistics, 2, 329–336.

Freedman, D.A. (2004). Casual Modeling or Why We Should Have Adjusted the 1990 U.S. Decennial Census. True Confessions, 48, 36-50.

Lyberg, L. and Sundgren, B. (2005). Production of Official Statistics: Concepts and Methods. Springer-Verlag, New York, forthcoming.

Rubin, D.B. (1987). Multiple Impugnation: Reasoning From My Posterior. Wiley, New York.

Received April 2005

Perfect Quality in the Swiss Deep Sea Fishing Survey

David A. Marker and David R. Morganstein[1]

Many statistical agencies have been applying the concepts of Total Quality Management to their surveys in an attempt to improve the accuracy of the results. This article reports on the amazing quality of one of the most recent surveys conducted by the Swiss Federal Statistical Office.

Key words: TQM; LARS.

1. Introduction

Inspired by the America's Cup victory of the Swiss yacht Alinghi, the Swiss Federal Statistical Office (BFS) decided in 2003 to measure the success of Swiss sailors around the world. Thus began the Swiss Deep Sea Fishing Survey, commonly referred to as the Latest Aquaculture Results Survey (LARS), which attempts to measure the Swissfish yield from all the oceans of the world.

Responding to the efforts of the European Statistical System to improve quality in all activities of national statistical offices (Eurostat 2001), BFS applied the techniques of Total Quality Management (TQM) to the LARS. This article reports on the results of using control charts to measure LARS' quality.

2. Control Charting Total Fish Weight

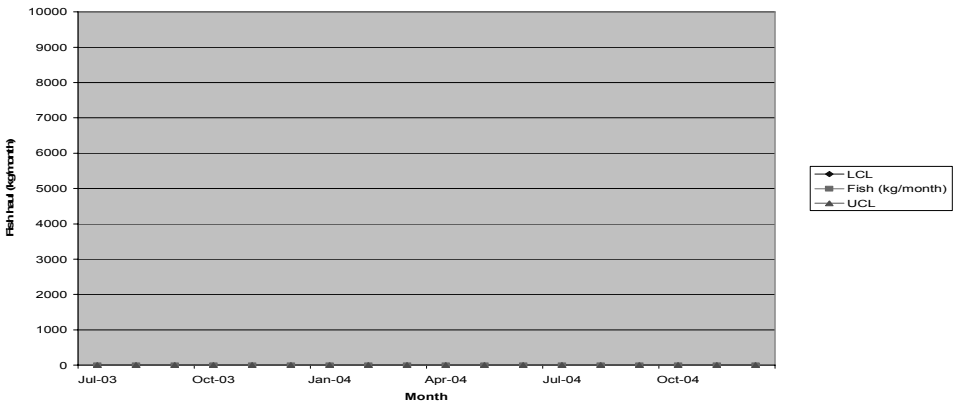
Morganstein and Marker (1997) suggest a variety of methods that should be used to improve the quality of surveys. One key method to control quality and measure improvements is the control chart. The control chart measures the same characteristic over time to see how repeatable the data are. This long-term average demonstrates the expected value in the future, but the key addition from the control chart is that control limits demonstrate the range of values that can be expected, if they system remains stable. If none of the plotted values exceed the control limits or exhibit nonrandom patterns the system is said to be stable. To improve the quality of a system, one should try to reduce the variability, thus shrinking the width of the control limits.

The control limits are typically set at plus and minus three standard errors around the observed mean. Relying on the central value theorem, if the system remains stable then only 1 out of 400 values would be expected to fall outside these limits. Clearly the smaller the standard error, the tighter the limits, the more consistent the resulting system. This is how BFS used TQM to improve quality.

[1] Westat Inc., U.S.A.

Figure 1 shows a control chart of the deep sea fishing haul (kg/month) of the Swiss fleet as measured by LARS. The survey began in July 2003, so this represents the first 18 months of the survey. Amazingly, the figure shows that not only are the fishing totals in statistical control, the upper and lower control limits are equal! There is no variation left in this system, a statistical system with perfect quality!

Figure 1. LARS control chart



3. Recommendations

Further investigation of this unusual result identified the probable reason for this result was that the Swiss don't have a deep-sea fishing fleet, so there are never any boats to do any fishing. Other national statistical offices are encouraged to follow the lead of the Swiss and identify systems like LARS with perfect quality so they too can impress their government finance ministers.

4. References

Eurostat (2001). Report of the Leadership Expert Group, International Conference on Official Statistics, Stockholm, Sweden.

Morganstein, D.R. and Marker, D.A. (1997). Continuous Improvement in Statistical Agencies, in Survey Methodology and Process Quality, Lyberg et al. (eds). Wiley and Sons.

Received April 2005

Control of the Coding Operation in Statistical Investigations – Some Contributions

Daniel Thorburn[1]

An important part of the work at a statistical agency is coding of the respondent's answers. Three aspects on coding are discussed. We derive a coding scheme that retains most of the information in the raw answers while speeding up the statistical estimation procedure. The control of the disclosure risk is discussed in connection with RSA-codes and finally we suggest a method to code sensitive data under the constraint of gender equality.

Key words: Disclosure control; gender coding; process control; public key.

1. Introduction and Notations

Coding is an important issue in statistical surveys and there exists a vast literature on the subject. One seminal work that should be mentioned in particular is the path breaking, not to say breath taking, dissertation by Lyberg (1981). Some other important works in this field are the information theory presented by Kullback (1959) and the DaVinci Code presented by Brown (2003). Singh (1999) is an excellent review. There exist many well known codes e.g. Code Napoleon (law, 1804), EAN (products), Enigma (Germany), ICD (mortality) and IIC (trade). Sweden has a long-standing history in the field of coding. Apart from Lyberg we can mention Beurling, who forced the German codes during the Second World War (Beckman, 1997) and the former minister for foreign affairs, Sandler (1965). There are many aspects of the problem of coding control but we will only discuss three aspects shortly, the information contents of coded data, data security and gender aspects.

2. Information Contents

A good coding system will retain all the information in the original data, while protecting the anonymity of all elements in the sample. A good measure of the information is decrease in entropy (c.f Kullback-Leibler information, Kullback, 1959)

$$-\sum_{i=1}^{k} p_i \ln(p_i), \qquad (2.1)$$

where k is the number of classes and p_i is the proportion of elements in the *i*-th class. It is easily seen, for instance using Lagrange multipliers, that for a fixed number of classes

[1] Department of Statistics, University of Stockholm, Sweden.

the information is maximised when all the probabilities are equal. It follows that the coding system should be chosen so that all the groups have the same frequency. This coding scheme has also the advantage that all tables can be filled out in advance with equal numbers in each cell, which speeds up the data processing and makes it possible to publish the statistics almost before the data collection is finished.

3. Disclosure Control and Data Security

Coding is also important from the point of data security, since a good code minimises the risk for disclosure. A new concept in this field is "one way coding" with public and closed keys. Everyone with access to the public key can code the message but only those who have access to the closed key can decipher it. It would be too long to go into the theory here. It suffices to mention that it uses RSA-codes, which are based on factorisation of huge prime numbers (Gardner 1997). This concept can be used e.g. in personal interviews on sensitive data such as the respondents' sex life. The interviewer gives the public key to the respondent who uses it and codes his answers. Since the interviewer does not have access to the closed key he has no possibility to find out the sex of the respondent. Only the computer at the statistical agency can decode the message before computing the statistics. This means that the respondent can feel fully protected.

However, one might argue that it is not satisfactory that persons at a statistical agency who have no interest in the subject should have access to the closed key. It is only a few statistical users, e.g. governmental agencies that must be able to decode the statistics. Thus we suggest that the agency produces statistics from coded data. In other words the published reports will be coded. Only the ultimate users of the statistical figures have access to the closed key and can decipher the statistics before making their decisions for the benefit of the public. This is in line with a long tradition. During the 18th century e.g., the Swedish population size was considered a state secret and was only revealed to those users that really needed the figure (c.f. Elvius 1744).

4. Gender Aspects

Statisticians, who always use the figure 1 to code females and 0 for males should rightly be criticised from the point of gender equality as well as those that always use 1 for men and 0 for women. A better way is to let 1 denote a female in every second survey and a male person in the others. This practice has, however, some disadvantages. It is sometimes difficult to decide which survey is performed first, e.g. if the data are collected during the same period. Another disadvantage is that some surveys may be bigger than others and thus the proportion of females and males denoted by 1 may still not be equal. A third disadvantage appears when data from two surveys are merged into one. Instead we suggest the convention that female persons born during the period January to June should be coded as a 0 and males by a 1 and the other way around for those born during the second half year. In this case the equality will be complete for a couple like Lars and Lilli who were born in December and June, respectively. Both will be counted as zeroes.

5. Discussion

We have tried to show that there are many important aspects on coding. Without a careful control of the coding operation anything may go wrong. Let us give a final example, which would not be detected without a good control system. One may think that Lars' work should be classified as NACE 22.130: "Editing and publication of journals", but a more suitable code in Sweden is (SCB 1992) is NACE 15:320: "Production of alcoholic and nonalcoholic beverages like whisky, beer, lemonade, and juice". (In Swedish: Produktion av alkoholhaltiga och alkoholfria drycker som whisky, öl, saft och JOS).

6. References

Beckman, B., (1996). Svenska Kryptobedrifter, (in Swedish), Bonniers, Stockholm.

Brown, D. (2003). The Da Vinci code, Bantam Press, London.

Code Civil de Francais, (ou Code Napoleon), (1804), Paris.

Elvius, P (1744). Förtekning uppå barnens årliga antal som äro födda uti U**** stad under de sist förflutne 50 åren, Kungliga Vetenskapsakademiens handlingar. (List on the yearly number of children who are born in U**** township during the last recent 50 years, Annals of the Royal Academy of Science, Stockholm.

Gardner, M., (1997). A new kind of cipher that would take a million years to break, Scientific American.

Kullback, S. (1959). Information theory and statistics. Wiley.

Lyberg, L., (1981). Control of the coding operation in statistical investigations – Some contributions, PhD-dissertation at Stockholm University, Urval 13, SCB, Stockholm.

Sandler, R. (1965). Chiffer, 2nd edition, Prisma, Stockholm, (1st edit 1943)

SCB (1992). Swedish Standard Industrial Classification 1992, MIS-1992, Statistics Sweden, Stockholm.

Singh, S. (1999). The Code Book, (Swedish translation), Norstedts, Stockholm

Received April 2005

Pure Imputation for Statistical Surveys

Peter Lynn[1]

In this article, we describe a technique referred to as "pure imputation." The ideas underpinning pure imputation are not new but the conceptual framework has not previously been developed. We set out the principles of the technique and then apply them in a realistic simulation study. We demonstrate that pure imputation has a number of important advantages compared to other commonly used imputation procedures. Notably, it is not influenced by measurement error or other random vagaries in observed data, and it allows the researcher to constrain distributions in the completed (imputed) data set to conform with auxiliary information or external targets.

Key words: Data generation; fabrication; missing data; WILD-GUESS.

1. Introduction

Data often contain missing values. This is true of many kinds of data, including those collected by means of surveys and those collected by administrative processes. When the data are to be used for statistical estimation, the missing values represent a challenge. The analyst must decide how they should be treated. A common solution is to "impute", or insert, a value for each missing datum. If this is done for every missing datum pertinent to a particular estimation, then the estimation can proceed using the completed data (original data, in cases where data was present, and imputed data, in cases where data was originally missing) using standard complete-data techniques.

There are many ways to choose the value to impute. Usually, the method involves identifying values that are both likely and plausible, typically by comparison with other units represented in the data set, for which values are present. A common method is the so-called hot-deck method[2]. In this procedure, for each case with a missing value of a target variable y, a case with a present value serves as a "donor" and donates its value of y to the first case, the "recipient"[3]. The donor is constrained to match the recipient on a small set of data items known as the "matching variables." For example, if y is economic

[1] University of E-sex, Colchester, UK.

[2] It is believed that the hot-deck method is so called because of its origins on the deck of a Caribbean Cruise liner in the 1930s, where two American graduate statisticians – Garry S. Bell and Reg R. Ballsy - were working as waiters. Each morning, they would tour the sun deck taking orders for lunch, but became frustrated at the time it often took to locate passengers who were inconveniently not occupying their usual lounger. They developed a way of imputing the orders of missing passengers, based on known characteristics such as their nationality, gender, and previous days' orders (The story was recounted by a Scandinavian passenger, Lerr Slybag, according to Beletristický 1962).

[3] This process is unlike other donor processes, such as heart donation, in that the donor case retains its own value. The mechanism is perhaps more like that of "cloning" (e.g., O'Diss 1998).

activity status of an individual, the matching variables may be age, gender and education, on the grounds that these three variables collectively explain much of the variation in activity status (or, more commonly, on the grounds that "this is the way we've always done it"). A later extension in the field of imputation methods was "multiple imputation", which involves imputing not just one value for each missing datum, but several possible values. The motivation for this has been described as a well-intended desire to make data management and analysis sufficiently complex that the risk of data being used by non-statisticians is minimised and thereby to improve the job security of statisticians.

2. Limitations of Existing Imputation Methods

- Researchers and data users have been known to express frustration at the limitations of existing imputation methods. These limitations include the following:
- It is difficult to constrain the method to produce desired distributions, particularly where these are very different from those in the observed data (cf., calibration weighting);
- It is difficult to incorporate constraints relating to complex combinations of variables;
- It is difficult to prevent measurement error in the observed values from influencing the imputed values;
- It is just difficult.

Pure imputation has been proposed as a method to overcome some of these limitations.

3. Pure Imputation: Conceptual Development

One of the earliest attempts to generate data where it was missing was Verrückt's (1959) Wanton Imputation of Likely Distributions (WILD) method. However, critics observed that this method had a rather weak control mechanism and did not always produce data that were of practical use. The Generation of Useable Empirical Statistical Summaries (GUESS) process (Fictício and Keksitty 1963) partly addressed this criticism by constraining the generated distributions to suit a preconceived policy initiative. The two methods have been productively combined in what is now known as the WILD-GUESS method.

However, standard application of WILD-GUESS involves the production of a simple statistic such as a percentage, a difference in percentages, or the economic impact of a new government policy. The method has not been used to produce micro-data. Pure imputation is a natural extension, applying the philosophy of WILD-GUESS to the imputation of micro data. In most commonly used forms of imputation, the statistician specifies a model for data generation. The observed data are then used to generate imputations following the model. In other words, the data-generating machine (DGM) consists of a model and observed data as inputs. Imputations are consequently sensitive to the model specification (the correctness of which can almost never be tested) and to errors in the observed data (which can almost never be detected). Pure imputation avoids

both the need to rely on a model and the need to rely on observed data by utilising the statistician as the DGM. Furthermore, the method is completely flexible as the DGM is allowed to choose values at will. This flexibility permits the incorporation of almost any kind of desired constraint. Furthermore, the process can be iterative. If the first pass imputations produce results that are not quite as desired, changes can be made. This can be done as many times as necessary. The result is an imputation method that is free from the excessive influence of observed values or statistical models and can be relied upon to produce results to suit any pre-conception or prejudice.

4. Case Study

We have tested Pure Imputation in a simple but realistic setting. Four statisticians were chosen to act as DGMs. Although the statisticians were simple, we felt this to be realistic. They were each set an identical imputation task, to create a data set of 100 cases and 10 variables, under some simple constraints. For brevity, we refer here to just 3 of the variables and two of the constraints. This provides sufficient illustration of the results. (We reported these results previously, but here adopt the RE-SPEW technique (Slygrerb 1980; see also 1981, 1982, 1983, ….) The first constraint was that the data set should contain 50 economists and 50 statisticians (variable `OCCUPATION`). The second was that a simple test of differences in mean equivalised pay (ratio of variables `PAY` and `HOURS`) should show that statisticians earn significantly less per hour than economists. The DGMs were provided with no data and no models that might have influenced their imputations.

Some minor operational difficulties were encountered. One of the DGMs took a very large number of iterations before he managed to obtain the required distribution of the variable `OCCUPATION`. This was ascribed to poor numeracy skills and lack of familiarity with a computer keyboard, problems which could easily be overcome in an ideal world where NSIs are able to recruit people with appropriate abilities. Another DGM failed to submit results to timetable, despite two reminder mailings following Dillman's Laboured Design Method.

The results obtained from the 3 responding DGMs were most encouraging, however. All managed to demonstrate the assumed earnings differential unequivocally. In addition to meeting the statistical requirements, we believe that the method is also highly cost-effective. The cost of producing the results was considerably less than that which would have been incurred if we had had to go to the trouble of carrying out a troublesome and inconvenient survey (though it would have been more had we paid statisticians a decent salary).

5. Conclusions

Pure imputation is well suited to obtaining the results you want. It can be applied in almost any setting. Though it is perhaps best avoided when others have real data. Only moderate skills are required to implement the procedures – even most statisticians can do it.

6. References

Beletristický, Pavel (1962). Contributions of Czech-Caribbean interactions to early development of statistical methods. Annals of the Silesian Society for Statistical History, 21, 4, 581-612.

Fictício, I.M. and Keksitty, Metoo (1963). Generation of Useable Empirical Statistical Summaries. Portuguese Journal of Practical Methods for Government Advisors, 7, 1, 32-38.

O'Diss, Ivor Predge (1998). Baa-king Mad: meet Dolly the clone sheep. The Planet, 1 April 1998, 1.

Slygrerb, Al (1980). REpeating Selected Parts of Earlier Work (RE-SPEW), pp. 105-114 in Academic Career Progression for Dummies.

Verrückt, Ikbin (1959). Go WILD in the Country. Munich: Bow Wow Wow verlag.

Received April 2005
Revised April 2005

Allowing Nonresponse May Give You a Better Estimate

Jan Wretman[1]

This is an example to demonstrate that, contrary to what we teach our students, a small amount of nonresponse may sometimes give you a better estimate, especially when the nonresponding person deviates a lot from the rest of the population.

Key words: Nonresponse; statisticians; estimate.

1. Introduction

Say that a statistician wants to carry out a sample survey in order to estimate the mean, of a population, consisting of $N = 100$ persons. He has no auxiliary information about the population, so he decides to select a simple random sample of size $n = 10$, and to use the sample mean as an estimator of the population mean. He also decides to use all possible resources, regardless of cost, to get response from all persons in the sample. No nonresponse will be allowed.

Let $U = \{1, 2, \ldots, k, \ldots, N\}$ be the population, and let y_k be the value of the study variable; $k = 1, 2, \ldots, N$. The statistician wants to estimate the population mean $\bar{y}_U = (1/N)\sum_{k\in U} y_k$. In our example we have $N = 100$, and we assume that the population values of the study variable are as follows:

Person	y-value
$k = 1, 2, \ldots, 33$	$y_k = 1$
$k = 34, 35, \ldots, 66$	$y_k = 2$
$k = 67, 68, \ldots, 99$	$y_k = 3$
$k = 100$	$y_k = 52$

Thus, the mean and variance of the study variable in the population are

$$\bar{y}_U = \frac{1}{100}\sum_{k\in U} y_k = \frac{250}{100} = 2.5$$

$$s_U^2 = \frac{1}{99}\sum_{k\in U}(y_k - \bar{y}_U)^2 = \frac{2541}{99} = \frac{77}{3}$$

[1] Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden.
Email: jan.wretman@stat.su.se

We now assume that the persons labeled $k = 1, \ldots, 99$ are decent people who would love to become survey respondents. The person labeled $k = 100$, however, is a certain Mr. *A*, who would be an obstinate nonrespondent if he was selected. He also has an extremely high value on the study variable. Let us consider two alternative strategies to deal with Mr. *A*.

Strategy I: No nonresponse allowed. This is the strategy that our statistician has chosen. If Mr. *A* is selected, he will be persuaded to respond, with an enormous amount of work, and at a high cost. The population mean will be estimated by the sample mean

$$\bar{y}_s = \frac{1}{n}\sum_{k \in s} y_k \qquad (1)$$

where $s \subset U$ is the sample of ten persons selected from *U* by simple random sampling. Using standard results for simple random sampling, it is seen that, under Strategy I,

$$E_I(\bar{y}_s) = \bar{y}_U = 2.5$$

$$V_I(\bar{y}_s) = MSE_I(\bar{y}_s) = (1 - \frac{n}{N})\frac{s_U^2}{n} = (1 - \frac{10}{100})\frac{77/3}{10} = 2.31$$

Strategy II: Nonresponse allowed. If Mr. *A* is selected, we will be content with getting no answer from him. Let s^* be the response part of the sample *s*, that is, if Mr. *A* is a member of *s*, then he will not respond, and thus $s^* = s \setminus \{100\}$. And if Mr. *A* is not a member of *s*, then all members of *s* will respond and then $s^* = s$. Let the population mean now be estimated by the mean of the *responding* persons in the sample,

$$\bar{y}_{s^*} = (1/n_{s^*})\sum_{k \in s^*} y_k \ , \qquad (2)$$

where $n_{s^*} = 9$ if Mr. *A* is a member of *s*, and $n_{s^*} = 10$ otherwise.

What about the bias, variance, and mean square error of the estimator (2) under Strategy II? Omitting details I claim that the conditional sampling design, given that $n_{s^*} = 9$, is equal to simple random sampling of 9 persons from the reduced population $U^* = \{1, 2, \ldots, 99\}$. Also, the conditional sampling design, given that $n_{s^*} = 10$, is equal to simple random sampling of 10 persons from U^*. We then find that

$$E_{II}(\bar{y}_{s^*}) = P(n_{s^*} = 9)\, E_{II}(\bar{y}_{s^*} \mid n_{s^*} = 9) + P(n_{s^*} = 10)\, E_{II}(\bar{y}_{s^*} \mid n_{s^*} = 10)$$

$$= \frac{1}{10}\bar{y}_{U^*} + \frac{9}{10}\bar{y}_{U^*} = \bar{y}_{U^*} = 2$$

where $\bar{y}_{U^*} = (1/99)\sum_{k \in U^*} y_k$.

So the estimator (2) is not unbiased under Strategy II. It has the bias

$$B_{II}(\bar{y}_{s*}) = E_{II}(\bar{y}_{s*}) - \bar{y}_U = 2 - 2.5 = -0.5 .$$

Next, we look at the variance. With a similar reasoning as above we find that

$$V_{II}(\bar{y}_{s*}) = P(n_{s*} = 9)\, V_{II}(\bar{y}_{s*} \mid n_{s*} = 9) + P(n_{s*} = 10)\, V_{II}(\bar{y}_{s*} \mid n_{s*} = 10)$$

$$= \frac{1}{10}(1 - \frac{9}{99})\frac{S_{U*}^2}{9} + \frac{9}{10}(1 - \frac{10}{99})\frac{S_{U*}^2}{10}$$

$$= \frac{1}{10}(1 - \frac{9}{99})\frac{33/49}{9} + \frac{9}{10}(1 - \frac{10}{99})\frac{33/49}{10} \approx 0.06$$

where $S_{U*}^2 = (1/(N-2))\sum_{k \in U*}(y_k - \bar{y}_{U*})^2 = 33/49$.

The accuracy of the estimator (2), as measured by its mean squared error, is

$$MSE_{II}(\bar{y}_{s*}) = V_{II}(\bar{y}_{s*}) + [B_{II}(\bar{y}_{s*})]^2 = 0.31$$

2. Summary

To sum up, the somewhat easy-going Strategy II gives an estimator with some bias. But at the same time it gives an estimator with considerably better overall accuracy than Strategy I, as measured by the mean square error. Strategy II also has two more advantages over Strategy I. First, it is less expensive. Second, the possible values of the estimator will all be within the range 1.00 – 3.00, while, under Strategy I, there is a probability of 0.1 that the estimator will take values in the range 6.1 – 7.9 (which will happen as soon as Mr. *A* is selected). So, instead of chasing Mr. *A*, the statistician in this example should be glad to be spared the answer from him.

Received April 2005

The Problem of Mythomaniacs in Statistical Surveys

H. C. Andersen[1], Eva Elvers[2], Ulf Jorner[3], Karl F. H. Münch-Hausen[4], and N.N. Vantroen[5]

The problem of surveying mythomaniacs is reviewed, and a new estimator of the proportion of mythomaniacs is proposed. This estimator avoids some earlier defects of confounding and hiding true values.

Key words: Propensity for Lying; true value; confounding.

1. Importance of Problem

Several authors have addressed the problem of inaccurate answers to survey questions and the underlying causes, cf., Biemer and Lyberg (2003). However, one root cause for erroneous answers has been virtually neglected i.e., the existence of mythomaniacs or compulsive liars.

In fact, there is even no generally accepted figure of the percentage of the population that suffers from this condition. The estimate of 9,543% given by the International Association of Mythomaniac Statisticians, IAMS, is widely thought to be an exaggeration (IAMS 1999).

2. The Problem Restated

Mythomania has several definitions, but we have chosen the current WHO definition (WHO 1992-94):

An excessive or abnormal propensity for lying and exaggerating

Thus, an inherent problem in estimating the proportion of mythomaniacs in a population is to handle the variation between responses, especially with respect to overstatements.

[1] University of Odense, Denmark.
[2] Statistics Sweden, Sweden.
[3] Statistics Sweden, Sweden.
[4] Univ. of Hannover, Germany.
[5] University of Lieburgh, England.

A recent breakthrough is the refocusing of the problem due to Hauser (2003). Using the concept of Propensity for Lying (PL) he defines the Mean Unbalancing of Mythomania (with obvious notation) as

$$MUM = \int f(y)PL(y)dy \qquad (2.1)$$

He gives the Hauser estimator for the SRS case (again with obvious notation) as

$$HE = 1/k \sum w(x)\ \ p(x)-p^*(x) \quad \text{over a set of questions} \qquad (2.2)$$

where corroborative information can be obtained through secondary sources. Hauser also gives estimators for more complicated designs as well as standard errors.

While obviously a major contribution, HE has been criticized as measuring not only the effects of mythomania, but of other sources of errors as well, c.f., Benign (2004). To overcome this confounding, we propose the new Synthesised Hauser Estimator[1]

$$SHE = 1/k \sum w(x)\ \ p(x)-p^*(x)/\check{s}(x) \qquad (2.3)$$

where š(x) is a separate, synthetic set of quasi-estimates aimed at the differentiation of pure mythomaniacal effects.

To give a simplistic example of the nature of š, consider the question
"Do you habitually lie to survey questions?"
Nonmythomaniacs will answer "No", while e.g., a mythomaniac with PL= 0.5 would answer "No" only half of the times. However, a mythomaniac with PL=1 would also answer "No" all the time, just like the truthful respondent.

In order to retrieve information in the face of this type of ambiguities, the approach of synthesising was applied, using rather natural assumptions of e.g. monotonous concave functions.

Thus … [2]

[1] Although we of course hope that it will be known as the Andersen-Elvers-Jorner-Münch-Hausen-Vantroen estimator, AEJMV.

[2] The mathematics of the next two pages was considered well above the level of our readers and is thus omitted for clarity (Ed.).

3. An Explorative Survey

An explorative survey is suggested to illustrate the superiority of the SHE estimator.

Adequate questions like "Did you respond to our previous survey" are needed. This has a de-confounding effect.

As a minimum a panel design with three parts should be used:

- a panel that had responded,
- a panel that had not responded, and
- a panel that had not been surveyed.

Inaccurate answers will still occur, especially among mythomaniacs.

However, with š(x) and (2.3), the mountain is reduced to a molehill and some facts come into view, as seen in (2.28). Similarly, repeated trumped-up replies will fall flat through (2.49). To be honest we are convinced that the use of š(x) will bring us considerably closer to the truth and shed new light on this important area.

4. References

Benign, U.N. (2004). Can Hauser be trusted? In Proceedings from the 5th APOOR Conference, Delphi.

Biemer, P.P. and Lyberg, L.E. (2003). Introduction to Survey Quality. New York; Wiley

Hauser, C. (2003). Compulsive Lying – A New Approach. In Zeitschrift für Umfragen, Nürnberg.

IAMS (1999). Chairman's address, 4th Meeting of the International Association of Mythomaniac Statisticians, Lieden University Press.

WHO (1992-94). International Classification of Diseases, Rev 10, Ch V.

Received April 2005

The 'Crossbow' Procedure Revisited: A Consumer Research International Methodological Experiment[1]

Lars L. Mintcoin and Norma L. Clitsin[2]

In this article we summarise briefly our recent work exploring the continuing relevance of the 'Crossbow' procedure for estimating precision in an experiment or survey. This work is founded in our belief that new developments in survey methodology should seek evolution - building on the past - rather than revolution - reinventing a wheel that has always, within predictable limits, worked.

Key words: Precision; bolts from the blue; bolts from other sources; the role of women; epitaphs to survey statisticians.

1. Introduction

This article arises from two worries. The first is the continuing presentation in journal articles of increasingly complex procedures designed to simplify the admittedly complex and potentially expensive task of variance estimation in survey research.

We find that many of these papers ignore what has gone before and, in turn, are ignored in papers that follow. It may be that the pressure to publish 'new' ideas is to blame for this lack of real progress. Or the algebra – perhaps too much for most of you? Or, most likely, the fact that whole papers are lost in the kind of tiresome fog found in the second sentence of the paragraph above. (You may have missed that sentence; if so, go back and try it again.)

There are, of course, exceptions where presentation standards cannot be blamed (see, for example Rust and Brick in this issue). Then, perhaps, we should question whether any serious progress is being made.

This is our second driver – our concern that 'new' methods may not be advances. We are especially moved by our rediscovery of the 'Crossbow' procedure for estimating precision in 'one-shot' studies – studies where there is little opportunity for replication. Use of this procedure has in fact been recorded since at least the early nineteenth century but has not been noted widely in the research community outside Switzerland. (There are, however, reports of slightly later use of the procedure elsewhere in the then Habsburg

[1] The Consumer Research International Methodological Experiment program is involuntarily funded by the community.

[2] The authors (both Gemini) acknowledge the absence of their regular collaborator (and fellow anagram) Dr. Sydney Skew. They thank Martin Collins for his thoughtful comments on an earlier version.

Austro-Hungarian Empire. Commentators differ as to whether or not this follow-up could be deemed a success.)

2. The Origins of the Procedure

The origins of the 'Crossbow' procedure are not entirely clear. And its first publication is debatable. Most reliably perhaps, it is attributed to the Hungaro-Swiss research team of Kish and Tell (19xx). Certainly it is the work of these authors that we have used to guide our own experiments.

3. Our Early Experiments

In our first series of experiments, Mintcoin was the experimenter and Clitsin the willing object. We found that our version of the 'Kish and Tell' procedure – needing minimal adjustment for today's Apple hardware – was reliable. Experiments with an Apricot found in the dungeon *(Eds note: please replace"dungeon" with "archive")* were less conclusive.

In the main Apple-based test we found that the 'Kish and Tell' procedure could be expensive, especially in terms of respondent incentives (where $100k seemed to be the norm). But it was revealing. The risk of bias remained unknown, but repeated (and repeated) replication lent a strong component of perceived reliability. The procedure certainly seemed to add to the credibility of results, especially in tabloid presentations.

4. The Final Experiment

In our final experiment, Clitsin took over the controlling role and Mintcoin acted as the object. The result was disappointing – briefly for Clitsin, but leading to early retirement on health grounds for Mintcoin.

As Clitsin said at the Coroner's Inquest: "It worked just fine with the Apple but was maybe doubtful with the Apricot. I knew we were pushing it to the limits with the Blackberry: the theory always was error-prone on the y-axis. Oh f***[1]".

Received April 2005

[1] *(Authors' note: translation may be needed.)*

Jackknifing the Bootstrap: Tidying Up Some Loose Ends in the Theory and Practice of Variance Estimation

Keith F. Rust and J. Michael Brick[1]

In recent years a number of approaches have been proposed for combining traditional methods of variance estimation for complex surveys. These combined methods endeavor to improve the efficiency of variance estimation. We present a new alternative, the Jackknifed Bootstrap. We investigate its properties using simulations based on data from a survey of retail footwear sales.

Key words: Complex surveys; footwear; fetish; handy tools.

1. Introduction

Traditional methods of variance estimation for complex surveys include Linearization, Balanced Repeated Replication, the Jackknife, and the Bootstrap (Wolter 1985; Rao and Wu 1988). Although these methods all have generally desirable properties in a range of applications, several authors have considered variations of these methods, with the aim of improving the reliability and efficiency of variance estimation in particular applications.

Yung and Rao (1996) introduced the Linearized Jackknife, and showed that it has favourable properties. These were also shown by Canty and Davidson (1999). However, more recently some other proposed alternatives have not proved as successful. Three methods in particular, the Jackboot, the Jockstrap, and Endlessly Repeated Replication (ERR) proved so unsuccessful that they have left no trace in the literature and their authors remain at large. But success has been achieved in the Canadian context through the powerful, though somewhat unsophisticated, Lumberjackknife method (Roots and Python, 2004). A promising though risky method is the Crossbow' Procedure (see Mintcoin and Clitsin, this issue)

In this present article we investigate another alternative variance estimator that combines two existing approaches, in an effort to capitalize on the desirable properties of both. This method we term the Jackknifed Bootstrap. In Section 2 we describe the procedure, its motivation, and some properties. In Section 3 we discuss the results of a simulation study based on data from the Italian Retail Footwear Survey.

[1] Westat, 1650 Research Boulevard, Rockville, MD 20850, U.S.A.

2. The Jackknifed Bootstrap Method

When initially applied to surveys, the Bootstrap Method was often subject to misapplication. As Rao and Wu (1988) pointed out, the data analyst could be easily tripped up as a result of failure to take care of the loose ends of the bootstrap replicates. This results in the likelihood that the analyst will fall short of the goal of reaching valid inferences about the data, often with embarrassing consequences, especially at a well-attended seminar with a high podium.

The proposed method deals with the shortcomings of the bootstrap by an application of the jackknife to the bootstrap replicates. Briefly, the bootstrap replicates are created as usual, but are then trimmed considerably using the jackknife. If the jackknife is applied to the correct extent, then the result is a very tidy and efficient procedure, and one that enables the analyst to proceed sure-footedly through thorny data sets. However, it is important to note that overuse of the jackknife could result in a footloose procedure that comes up short, with the potential to leave the analyst stumbling badly.

Theoretical considerations presented elsewhere (Brick and Rust 2010) indicate that the method is likely to be especially successful when applied to certain kinds of statistical analyses. In particular when applied to Median Polish analyses, and Trimmed Means, the results appear to be much superior to the alternatives.

3. Simulation Study

We considered the relative efficiency of the jackknifed bootstrap, in comparison to the Linearized Jackknife, for an artificial population generated from data from the Italian Retail Footwear Survey (Gucci and Stiletto 2000). The results are summarized in Table 1.

Table 1. Relative efficiency of the Jackknifed bootstrap for estimates from a footwear survey

Estimate	Relative efficiency
Men's Shoes	1.2
Men's Boots	2.0
Athletic Shoes	1.5
Women's Shoes	0.3

The results are very encouraging for laced footwear, but the poor performance for women's footwear leaves us with a note of caution. We conjecture that the asymptotic height of women's heels plays a role in the efficiency of the method.

4. Authors' Note

We wish to counter a referee's suggestion that "Rust" and "Brick", rather than being the authors of a statistics paper, are in fact colour styles from a high fashion footwear

catalogue. We contend that only we could have come up with this research and manuscript.

5. References

Brick, J.M., and Rust, K.F. (2010). Finally, An Actual Justification for the Jackknifed Bootstrap. Journal of Statistical Procrastination. To appear. Maybe.

Canty, A.J., and Davidson, A.C. (1999). Resampling-based Variance Estimation for Labour Force Surveys. The Statistician, 48, 379-391.

Gucci, I.M. and Stiletto, U.R. (2000). The Italian Retail Shoe Survey. Journal of Officially Boring Statistics. Special issue in honor of Imelda Marcos, 111, 23-37.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling Inference with Complex Survey Data. Journal of the American Statistical Association, 83, 231-241.

Roots, O.K., and Python, M. (2004). The Lumberjackknife Method of Variance Estimation. Journal of Purely Canadian Statistics, 1, 1-65.

Wolter, K.M. (1985). Introduction to Variance Estimation. Springer-Verlag: New York.

Yung, W., and Rao, J.N.K. (1996). Jackknife Linearization Variance Estimators under Stratified Multistage Sampling. Survey Methodology, 22, 23-31.

Received April 2005
Revised April 2005

Data Access Recommendations from the Users' Group, Behavioral Econometrica Association[1]

Eleanor Singer[2]

The Users' Group of the Behavioral Econometrica Association, LP proposes a modest set of recommendations for making confidential data more readily accessible to research. Such research is of enormous benefit to society; in order to reap its full benefits, we recommend (1) relaxation of restrictions on the release of confidential data; (2) greater use of fictive data; (3) penalties for recalcitrant statistical agencies and uncooperative respondents.

Keywords: good advice; pearls; swines.

1. Introduction

At its fourth annual meeting, the Users' Group of the Behavioral Econometrica Association, LLP, took up the pressing question of timely access to research data.

Access to detailed data—much of which is collected or funded by government—is necessary for a society to function freely and effectively. Indeed, nations with advanced data collection and information-sharing infrastructures enjoy widespread benefits. Though difficult to quantify, these benefits are obvious. For example, on the basis of detailed data, economists have concluded that people who save more accumulate more wealth, and that people who are healthier live longer.[3]

These benefits are currently threatened by two developments. First, excessive concerns with safeguarding data confidentiality have led to intolerable delays of two months or more in the release of rich, detailed data sets to economists and other users. Second, for unrelated reasons, the public's level of cooperation with surveys, including government surveys, has dropped to unprecedented low levels. The recommendations that follow are designed to address both of these threats to the unfettered exploitation of social research.

2. Recommendations

Recommendation 1: Data produced or funded by government agencies should be made available for research as soon as they have been collected, and preferably earlier.

[1] The helpful comments of Michael Couper on an earlier draft are gratefully acknowledged.
[2] Survey Research Center, University of Michigan, U.S.A.
[3] Smith (2003) erroneously reported that people in poorer health live longer, but this was due to an error in the estimation model used.

Agencies that fail to meet this criterion should have their budget reduced by 1% for each day's delay.

Recommendation 2: The timing and manner of release of research data should be decided by users, especially economists, rather than the statistical or other data collection agencies.

Recommendation 3: Both the public's concerns about confidentiality and the impact of such concerns on survey participation have been exaggerated. Therefore, we recommend that in making decisions about access to research data, the weight given to such concerns should be substantially reduced or eliminated.

Recommendation 4: The significance of high response rates has been greatly exaggerated. Low response rates can be compensated for by judicious use of nonresponse bias studies. Such studies will increase the costs of doing the research, but more of these costs will go to survey research firms rather than to respondents in the form of incentives. We believe this is a net gain.

Recommendation 5: Data masking and multiple imputation techniques produce unacceptable distortions in the models estimated by behavioral econometricians and others. The data produced by such methods are also too difficult to analyze using ordinary methods. Therefore, we recommend that statisticians produce one fictive data set from each survey that can be analyzed using ordinary analytic techniques. The models users want to fit should be used in creating such data sets, thereby increasing the utility of the data.

Such a procedure will also solve the problem of unit and item missing data without the analytic complexity introduced by multiple imputation. Finally, it will solve the problem of declining response rates. Funding agencies such as the National Science Foundation and the National Institutes of Health should support the needed research for creating such data sets.

Recommendation 6: The penalties for willful violations of data confidentiality are currently set too high, especially since no one has documented such a violation. The size of the potential fine prevents full access to confidential data by high school and college students, since neither their parents, nor their schools, are willing to co-sign a bond guaranteeing payment of the fine if a breach occurs. To encourage greater access to research data, such fines should be reduced to zero.

Recommendation 7: Currently, access to some confidential data is limited to research data centers with restricted access. To achieve the research potential and cost-effective operation of these centers, they should (1) broaden the criteria for access; (2) reduce turnaround time for reviewing proposals by eliminating such review; (3) eliminate supervision of researchers at the data center.

Recommendation 8: Statistical and funding agencies should support continuing research to monitor the views of data providers and the general public about research risks and benefits in order to better manipulate them.

Recommendation 9: Basic information about confidentiality and data access given to everyone asked to participate in statistical surveys should include notification about planned or unplanned future uses of the data, possible use by researchers other than those collecting the data, and possible nonstatistical uses of the data.

However, if anyone refuses to participate in the survey as a result of this information, statistical agencies should use available administrative data and appropriate statistical techniques to impute their responses, instead.[1]

Recommendation 10: Eligible sample members who cannot be contacted in a reasonable time or who refuse to participate in a survey for the duration of the field period in spite of repeated callbacks, letters, and monetary incentives should be fined $250,000 or jailed for 5 years, or both. Those who refuse to participate in more than one survey should be sentenced to lifetime participation in an opt-in Internet panel.

3. References

Rubin, D.B. (2006). Multiple Imputation: Making the Punishment Fit the Crime. Journal of Criminal Statistics, 103, 44-67.

Smith, A. (2003). The Health and Wealth of Nations. London: Routledge.

[1] Rubin, 2006.

Swedish Gentlemen and Norwegian Bullies – Is a Reunification after 100 years Worth Considering?

Trine Dale, Gustav Haraldsen and Øyvin Kleven[1]

This article presents the results from a survey carried out simultaneously in Norway and Sweden in connection with the 100th Anniversary for the dissolution of the union. The survey was on attitudes towards and knowledge about the other people. We found that myths and prejudice are dominant perceptions and that the time has not come for a reunification between the two sister nations.

Key words: Union; dissolution; myth; reunification.

1. Introduction

A hundred years ago the union between Sweden and Norway was dissolved, and lately voices have been raised to promote the idea of a reunification. Most people seem to regard these initiatives as humorous attempts to score points at the cost of the other country.

There have always been close connections between the two countries, but also a love/hate relationship between the two peoples. Sweden has been the big brother and Norway the little brother in many areas. This is still the case even if the power balance has shifted somewhat the last decades. Norwegian economy is more solid than the Swedish, much thanks to the oil. This has resulted in a higher level of living, very little unemployment and higher salaries for most people. The balance has also shifted in sports: Sweden used to achieve better results in important sports like football and alpine skiing, but this is no longer so - a sensitive topic for most Swedes. And, even if Norway is just as advanced as Sweden in many areas, a familiar strategy in Norway is still to say "let's look to what Sweden has done". And then we copy both their successes and their mistakes. Sweden is more oriented towards Europe and European aristocracy and has a tendency to treat its little brother with some arrogance and condescension.

2. Method

In January 2005 Statistics Norway and Statistics Sweden decided to test the perceptions about the other people in a survey conducted simultaneously in the two countries and also whether there was a mood for a reunification in the two peoples.

Attitude questions were posed to the samples in the Omnibus surveys of the two agencies. Most of the questions were based on familiar myths and conceptions. We also

[1] Statistics Norway.

let the respondents listen to statements made in different dialects in the other language to test the level of understanding.

In both countries the data collection method was CAPI, and about 5 000 people responded to each survey.

3. Results: Myths and Prejudice Rule

The results show that many myths still exist and that the ability to understand the other language is somewhat low, especially in Sweden. They also show that the basis for forming a new union between the two countries is weak.

Swedes characterize Norwegians as a rustic, bread-eating people that are easygoing and very (too?) informal. Norwegians are outdoor people that go hiking every Sunday with their "nisseluer" and "lusekofter" on and their "matpakker" in their backpacks. They are also arrogant and nationalistic, hard working and quite competent - but not rule abiding. Very few Swedes think Norwegians are polite.

Norwegians, on the other hand, characterize Swedes as nice, fashionable and good looking. They are extremely polite and very formal, arrogant and rule abiding. Swedes are hard working, competent city people and they are good salesmen. They are stuck up on themselves and lack knowledge and understanding about Norwegian culture and language.

Many stories from the time around the dissolution confirm that many of the myths are historical. (Laache 1941). A story from 1995 illustrates the level of knowledge Swedes have about Norwegian affairs today: A well-known Swedish politician was asked what the name of the Norwegian queen was. His response was: Well, what is the name of King Olaf's wife? At this point, King Olav V had been dead for five years. His wife, who by the way was a Swedish princess, died before Olav became king and was never a queen.

Swedes' basic knowledge about Norway used to be better. In 1905 16,000 delegates to a Pentecost in Sweden could sing the Norwegian National Anthem. The Swedish king could speak Norwegian (Hegge 2004). Today Swedes have trouble understanding basic Norwegian. Very few Swedes could understand the statements in our language tests. Especially the rural dialects proved difficult, but many had trouble understanding “standard Norwegian” as spoken in Oslo as well.

The Norwegians could understand Swedish more easily, but had problems with the dialect from Skaane. Both peoples had trouble placing the dialects geographically. When asked to name the king, queen and prime minister in the other country, the Norwegians could give a correct answer twice as often as the Swedes.

But, Swedes are more positive to a reunification than Norwegians, and the closer to the Norwegian border they live the more positive they are. The main reason given for wanting a reunification was to get hold of some of the oil-money. In the border areas people were very positive. Norwegians were generally against a reunification, especially in rural areas. The main reasons given were that we can manage better on our own and that a union would probably lead to more centralization – like in Sweden.

4. Concluding Remarks

We can conclude that although Norwegians and Swedes like each other quite well and statistics show that we are quite similar in both behaviours and attitudes, both peoples think the neighbouring people is a bit odd. Especially the Swedes are prejudiced, thinking the Norwegians are very rustic, low culture people. For instance, a famous researcher in Sweden, Lars Lyberg, refers to it as "bread-day" when he goes on a one-day visit to Norway – referring to the well established Norwegian custom of bringing lunch packs to work and school. Mr. Lyberg can also serve as an illustration of the Swedish travel pattern. He has travelled all over the world, both in his work and private life. However, he first came to Norway when he was in his 50's after a special invitation.

There is no foundation for a new union in the peoples, so it is probably wiser to stick to well-established patterns of cooperation across the border. However, if we are a little less prejudiced against each other we might learn a thing or two. Swedes could teach Norwegians better manners and Norwegians could teach the Swedes to loosen up a bit and to be less formal. That might prepare the soil for more committing cooperation in the future.

5. References

Hegge, Per Egil and Leif Arne Ulland: Det var i 1905, Andresen and Butenschøn AS 2004.

Laache, Rolv (1941). Nordmenn og svensker efter 1814, H. Aschehoug and Co, Oslo.

Received April 2005

Book and Software Reviews

Books for review are to be sent to the Book Review Editor Joop J. Hoax.

Why Male Surveys Do Not Work; The Total Disaster Method
Carrie Broadsaw
Statistical Package DeSade
Fester Addams

Donna Dilman. *Why Male Surveys Do Not Work; The Total Disaster Method*. New York: Wooley Publishers, 2005. ISBN 0-471-32354-3. 1871 pp. + refs. and index. 1978 USD.

In this book, Donna Dilman explains the difficulties associated with male surveys, and discusses some solutions to these obstacles. The main problem with male surveys is that males tend to be lacking in face-to-face contact, which makes doorstep interactions less successful. In addition, males produce more survey errors at larger costs, a point already made by Groovey (1989). As Donna Dilman explains, these problems can be overcome. The solutions are based on social change theory, which basically teaches males to change into females. This offers a new approach to avoid the face-to-face problems associated with male surveys. The core of this approach is to focus on design and layout factors. The book discusses in detail survey layout and questionnaire make-up based on Max Factor, with some additional attention to Lancome and Guerlain for the international scent. Internet based surveys are treated extensively, which is important because it is well known that males are overrepresented on the Internet.

This book is a valuable contribution to survey methodology, which I recommend to all male survey researchers. Nevertheless, it is not a complete survey handbook, since there are some important issues that are not addressed. One of these is the art of asking questions. Research has shown that in conversation females ask more questions than males, which puts male surveys at a disadvantage. The advice compiled by Cross and Nicks (1981) appears especially relevant in this context. Also, the exclusive attention to male survey methods overlooks the better half of the survey field. Readers who are interested in total survey quality are advised to also obtain the classic treatise by Apec and Yberg (2004).

References

L.J. Apec & L.L. Yberg (2004). Swödish Kvality Ädvices, Őnfortunatöly Pöblished Őnly in Swödish. Stadsholmen: Scandinavian Centralized Bureaus.

John Cross & Steve Nicks (1981). Knowing Me, Knowing You. Asking Questions with Confidence. Ohio: 1910 Fruit Company.

R.M. Groovey (1989*). Surveys Full of Costs and Errors.* New York: Wooley Publishers.

- Reviewed by Carrie Broadsaw

Tom Terrific Software. Statistical Package *DeSade*. (Available from www.TT-Sofware.com). EUR 666, all licenses.

The statistical package DeSade (an acronym of Design-Expected Solutions And Data Exploration) is designed as a set of analysis tools that can be used in different combinations to attack any data set. The most important types of tools are *procedures* and *functions*. Functions are used to transform the data, procedures are used to carry out statistical analyses. In addition to the usual data transformation functions such as calculating squares and roots, DeSade includes some more unusual functions. Analysts will make good use of functions like *slash*, which removes all outliers, *abnorm*, which denormalizes all normal data, and *maim*, a revolutionary multiple imputation method based on little known research by Rittle and Lubin (1990), which starts by removing all cases that are not incomplete. For applied statisticians, a particularly useful function is the function *damn*, which automatically removes all nonsignificant data.

Statistical procedures include standard procedures such as *mean* to calculate averages, and *procustes* to calculate severely trimmed averages. New to this reviewer was the procedure for *sadistic regression* that uses a floating log (*flog*) link to analyze strongly tied data. Again the package includes some tools specially aimed at applied statisticians. The most forceful of these is the procedure *bondage* that uses a brute force number crunching technique to automatically tie up all loose ends in the analysis.

The software is priced competitively, especially since the standard license is unlimited. At an extra EUR 007, the user obtains the additional license to kill. I recommend this package, especially for use in teaching statistical analysis to unwilling students.

References

Rittle, R., and Lubin, D. (1990). Multiply Imputing Everything. Journal of the American Sadistical Association, 198 (11), 1-93.

- Reviewed by Fester Addams

Editorial Collaborators

The editors wish to thank the following referees who have generously given their time, skills, and patience to the Journal of Obnoxious Statistics during the period April 1, 2004- April 1, 2005. An asterisk indicates that the referee served more than once during the period

Boaz, Henriette Pr. *, Meow Institute of Technology, USA
Hox, Heinz Hobbes ***, Catbarra, Australia
Svans, Jip Pelle, University of Meowköping, Sweden

Journal of Obnoxious Statistics

ISBN-10: 9080770825
ISBN-13: 9789080770829
NUR-code: 916