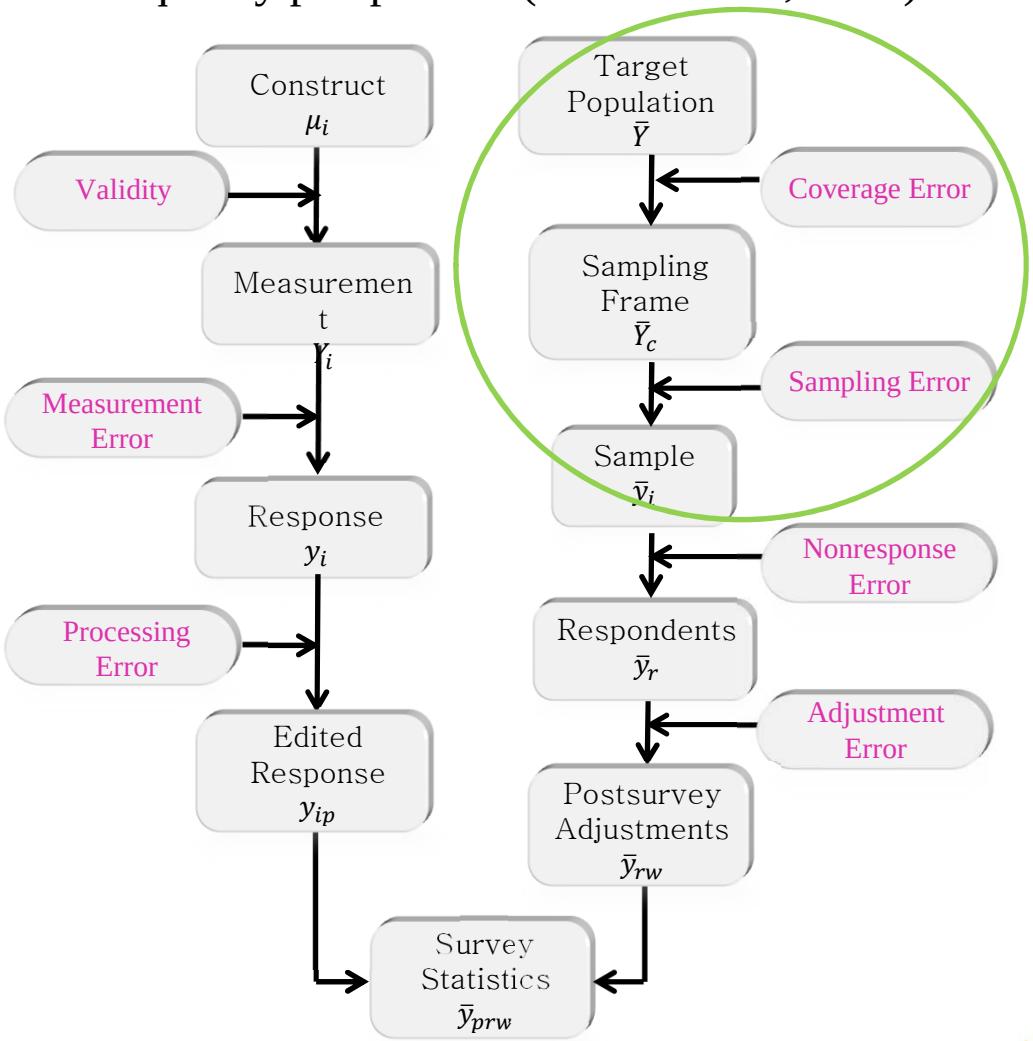
4. Area Probability Sampling using IT

Sources of Errors in Surveys

❖ Survey lifecycle from a quality perspective (Groves *et al.*, 2009)



dongguk UNIVERSITY SHPRC

Survey Overview

❖ Survey on Environmental Pollution and Health

- ➢ Sponsor
  National Institute of Environmental Research
- ➢ Collector
  Survey and Health Policy Research Center, Dongguk Univ.
- ➢ Purpose
  To understand recognition of environmental health problem and real condition of environmental disease by using a scientific sample survey
- ➢ Target Population
  199,328 households near Industrial Complexes in Incheon
- ➢ Sample Size
  606 households
- ➢ Sample Design
  Four-stage area sampling, within-household random selection
  IT-based approach
- ➢ Mode of Administration
  CAPI (computer-assisted personal interviewing)

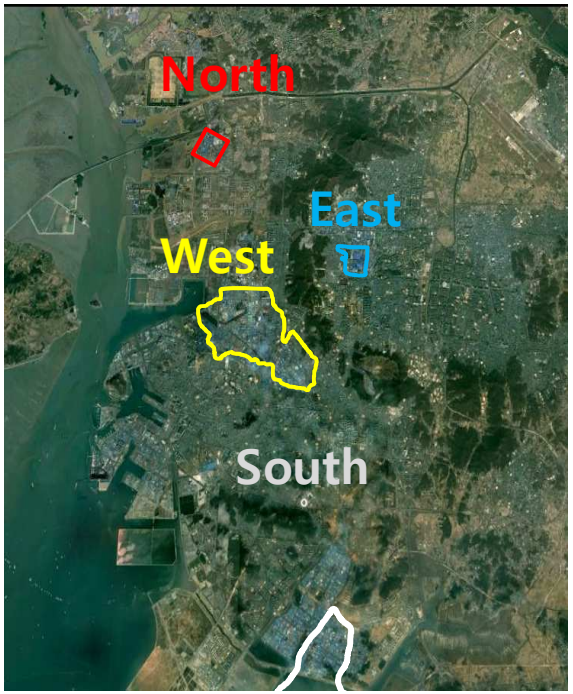
❖ Advantages of IT (Information Technology)

- Available from Internet information services including
  - Statistical Navigator (Statistics Korea)
  - New Address Information System (Ministry of Security and Public Administration)
  - Electronic map services in web-portal sites
- IT provides useful information such as
  - Address of every structure with their location
  - Type of structure
  - Number of dwellings in the ED
  - Surroundings

Sample Design

❖ Locations of Industrial Complexes in Incheon

Satellite map of Incheon

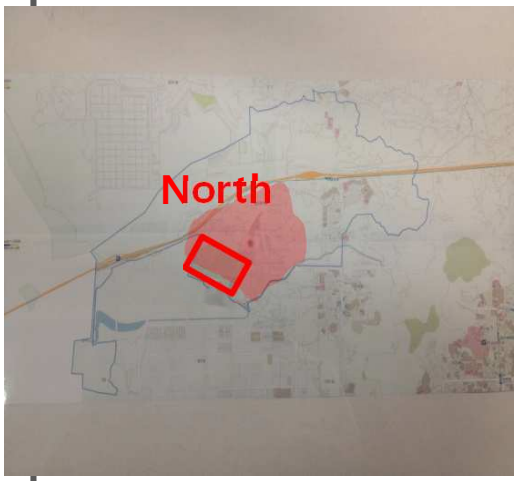


Enlarged map of Incheon

Sample Design

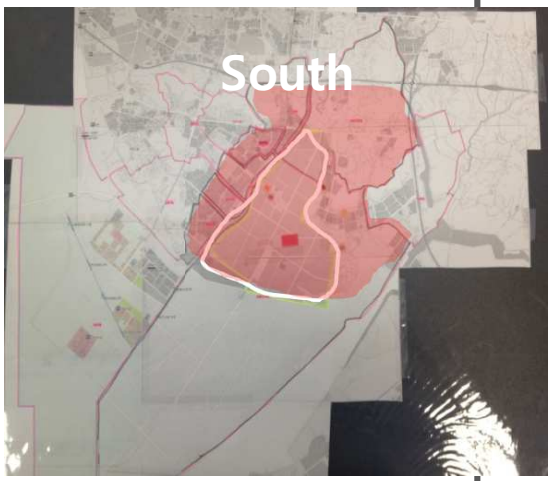
❖ Defining Target Areas

➢ Statistics Korea provides the map with geographical boundaries of “Dongs” and enumeration districts (EDs)



North industrial complex
(서부산단)

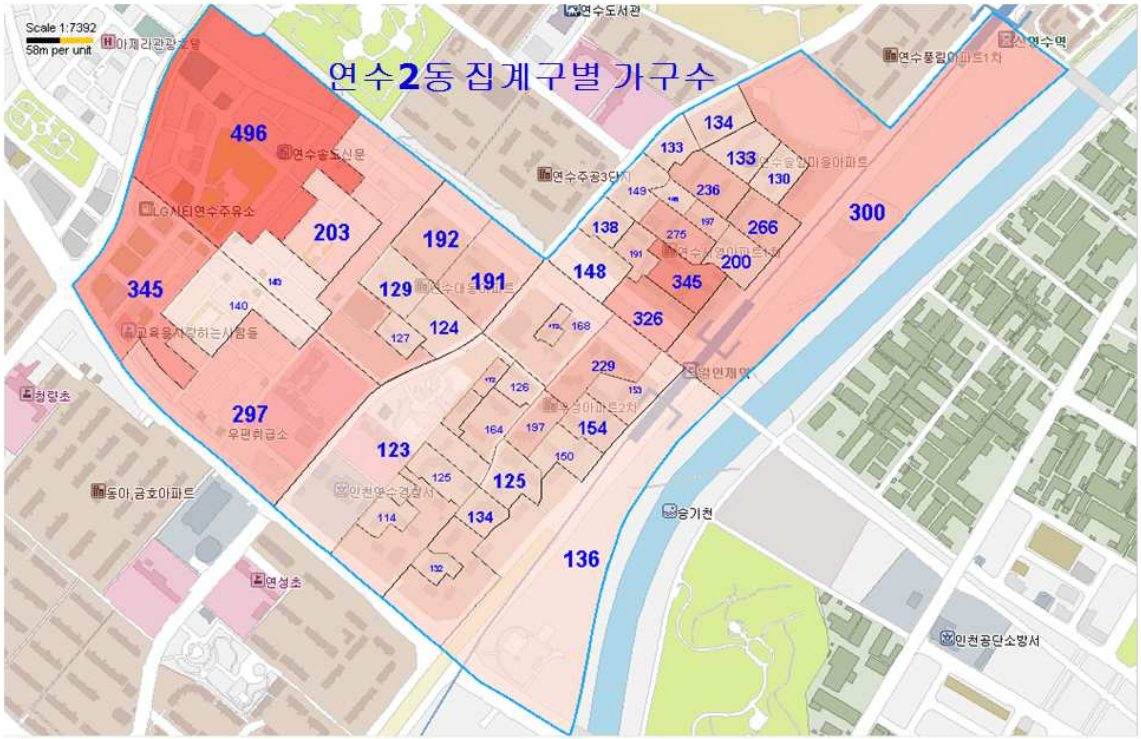
East and West Industrial Complexes
(부평수출산단,
주안(일반,기계,수출)산단)



South industrial complex
(남동국가산단)

Sample Design

❖ Identifying the number of households and local residents in target area



Statistics Navigator (Statistics Korea)
– South Industrial Complexes

Sample Design

❖ Four-stage Area Sample Design

	First stage	Second stage	Third stage (listing)	Fourth stage
Sampling Unit	Dong	ED	Chunk	Segment
Sampling Method	Sampling for selecting 2 Dongs from each stratum	Sampling for selecting several EDs from each selected City	Simple random sampling of 2 chunks from each selected ED	Systematic sampling of a segment from each selected chunk
Selection Equation	$\frac{2Mos_{h\alpha}}{\sum_{\alpha} Mos_{h\alpha}}$	$\frac{d_{h\alpha} Mos_{h\alpha\beta}}{\sum_{\beta} Mos_{h\alpha\beta}}$	$\frac{2Mos_{h\alpha\beta\gamma}}{\sum_{\gamma} Mos_{h\alpha\beta\gamma}}$	$\frac{Mos_{h\alpha\beta\gamma\delta}}{\sum_{\delta} Mos_{h\alpha\beta\gamma\delta}}$

Chunk: a set of 24 HU's
Segment: a set of 4 HU's

$$f_h = \frac{2Mos_{h\alpha}}{\sum_{\alpha} Mos_{h\alpha}} \times \frac{d_{h\alpha} Mos_{h\alpha\beta}}{\sum_{\beta} Mos_{h\alpha\beta}} \times \frac{2Mos_{h\alpha\beta\gamma}}{\sum_{\gamma} Mos_{h\alpha\beta\gamma}} \times \frac{Mos_{h\alpha\beta\gamma\delta}}{\sum_{\delta} Mos_{h\alpha\beta\gamma\delta}} = f$$

Sample Design

❖ Listing Operation

- Important to reduce non-coverage error
- Follows the steps:
  - Visiting the selected cluster
  - Verifying geographic locations of the structures
  - Drawing a sketch map of the structures
  - Describing characteristics of structures
- These tasks require a considerable amount of time, effort and cost
- The accuracy of listing depends on field staffs

Sample Design

❖ Listing Operation (cont.)

Map for the selected ED

magnified map for the above map

Listing Sheet

동_리 계구:

NO. : _____

작성자: _____

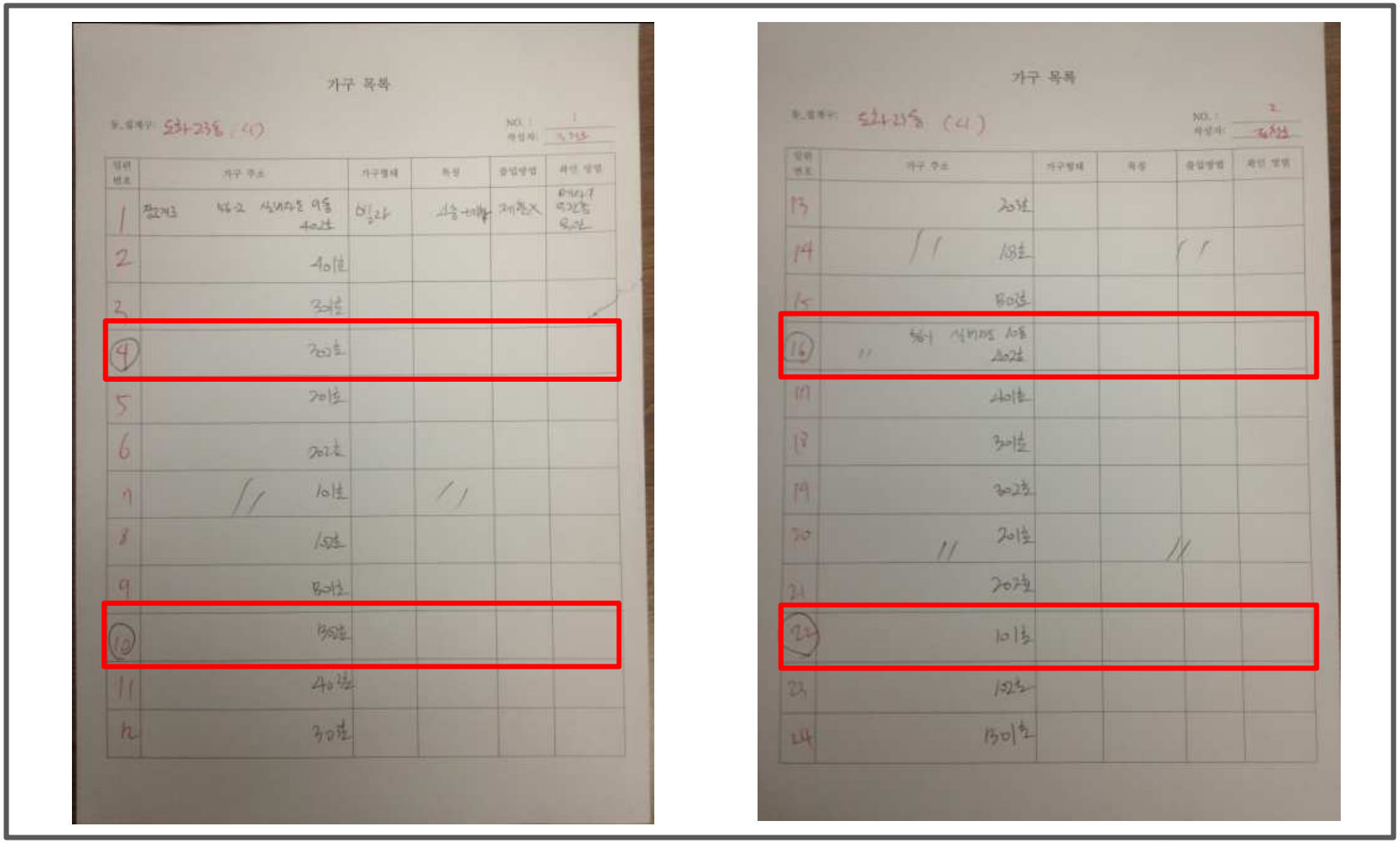
일련 번호	가구 주소	가구형태	특징	응답 방법	확인 방법

❖ Selecting a segment

➢ Dividing chunks into compact segments of size 4 and making segments heterogeneous

	Ordered number of households			
Segment 1	1	7	13	19
Segment 2	2	8	14	20
Segment 3	3	9	15	21
Segment 4	4	10	16	22
Segment 5	5	11	17	23
Segment 6	6	12	18	24

❖ Selecting a segment (cont.)



Survey Instrument and Data Collection Method

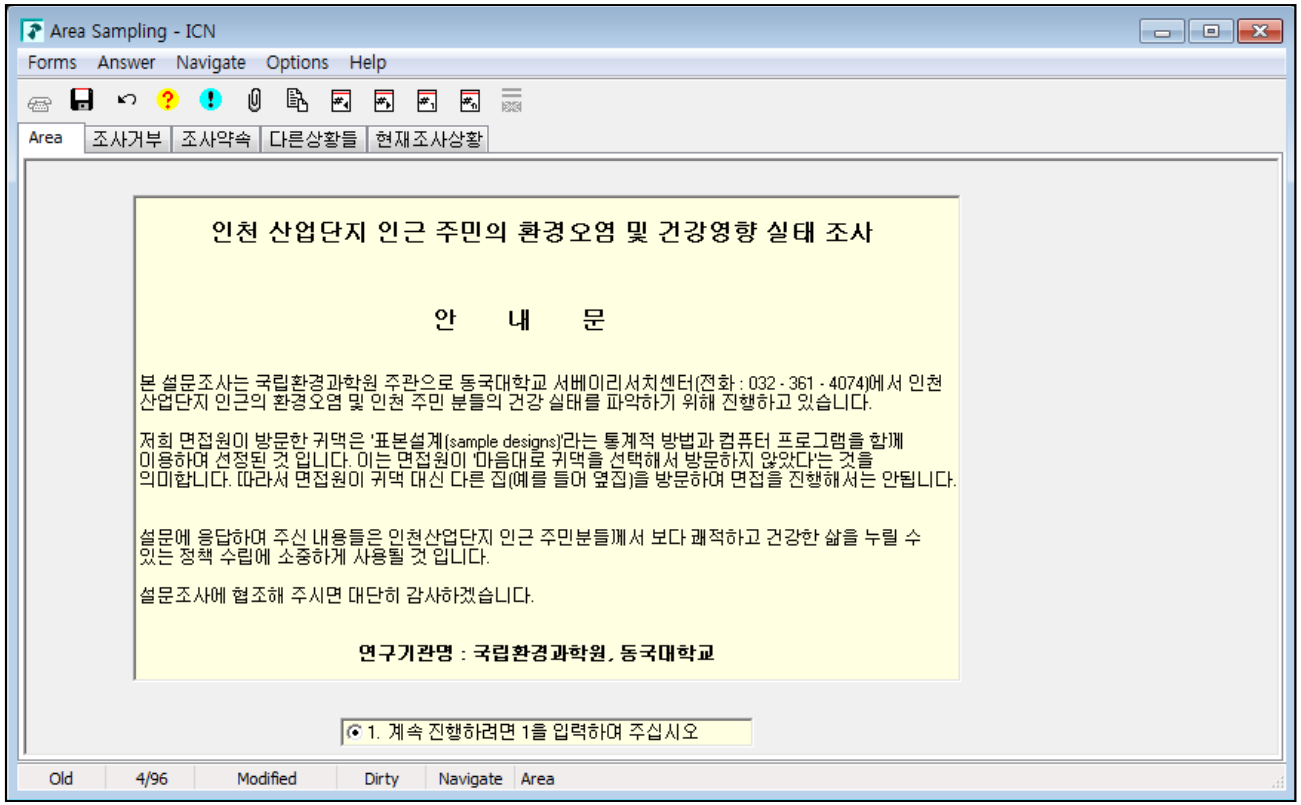
❖ Questionnaire Design

- 7 revisions after construction:
  - Following up on prior surveys of other industrial complexes
  - Construction of initial instrument
  - Revision by expert group on survey questionnaire design (1^{st})
  - Revision by expert group on occupational environment (2^{nd})
  - Revision by expert group on survey questionnaire design (3^{rd})
  - Revision by expert group on occupational environment (4^{th})
  - Revision by expert group on survey questionnaire design (5^{th})
  - Revision for easy manipulation of CAPI (6^{th})
  - Final revision after pretest (7^{th})
- Focuses on building and correcting questionnaire
  - Clarity
  - Comprehensiveness
  - Acceptability

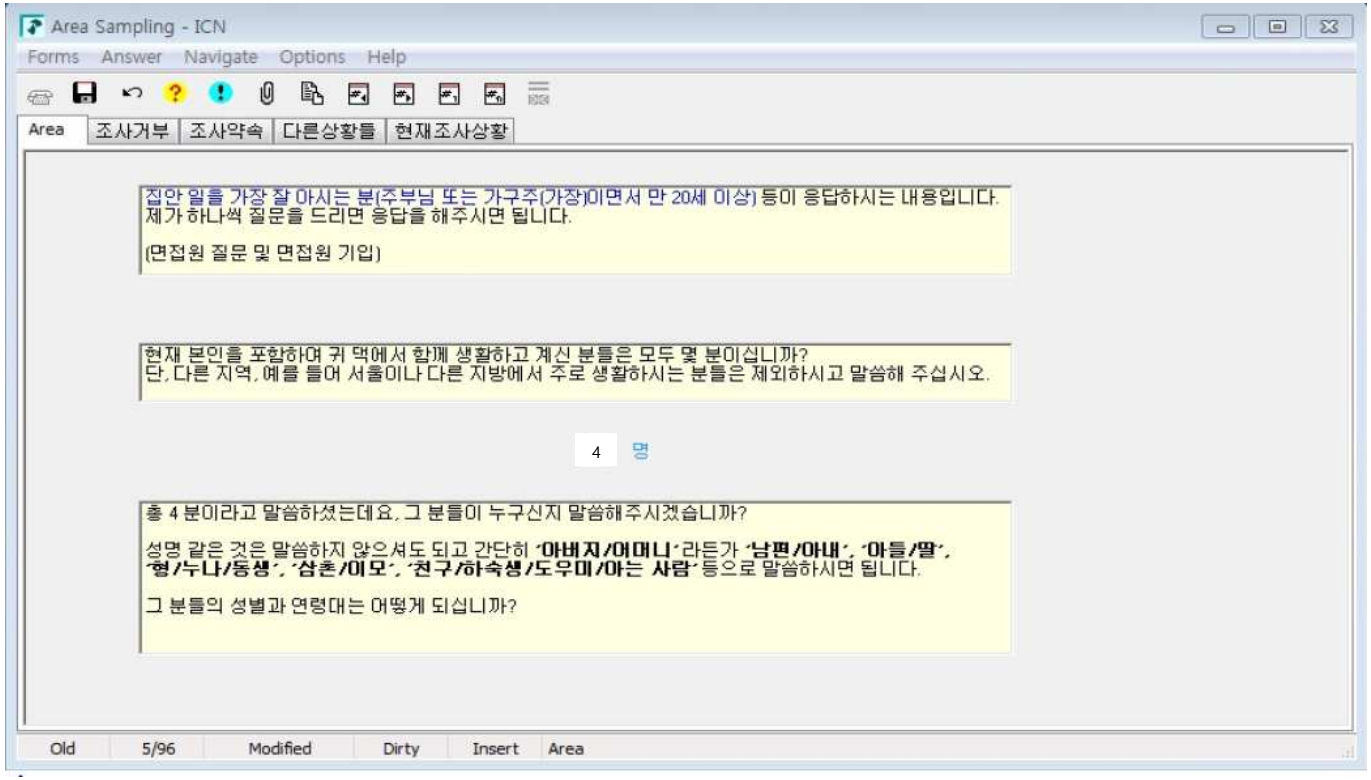
❖ CAPI (Computer-Assisted Personal Interviewing)

- Advantages
  - Less processing error and measurement error
    - No errors are allowed to branching/skip logic
    - Errors and missing responses are not allowed by direct input into database
    - Easy implementation in listing household members
  - Gains in cost and time efficiency
    - No efforts are needed for computerization of survey responses
    - Reduction in response burden and interviewer workload
  - Convenient survey management
    - Easy supervision on interviewers and interview process
    - Easy notification to interviewers when further notice required
    - A way to interviewers look more professional to sample interviewees

❖ CAPI (Computer-Assisted Personal Interviewing)



❖ CAPI (Computer-Assisted Personal Interviewing) (Cont.)



❖ Monitoring and Managing Paradata

- Paradata
  - Data related to field survey processes captured including call or visit records, interviewers observations, time stamps, travel and expense information, and many others
- Use of paradata
  - Post-survey use
    - Optimal call or visit scheduling for subsequent surveys or survey waves
    - Evaluating and reducing nonresponse by comparing early respondent to late respondent
    - Assessment of net quality gains by examining of the relationship between nonresponse and measurement error
  - Process control
  - Monitoring and managing of survey (especially, ongoing survey)

❖ Monitoring and Managing Paradata (cont.)

- Example: Interim outcome codes of the results of each Household visit
  - Complete / Partial Complete
  - Appointment by Respondent/Household member
  - Refusal by Respondent/Household member
  - Unavailable Respondent/Household member
  - Absence of Respondent
  - Empty or Vacant House
  - No allowance for entry
  - Identified as out-of-scope in House or Household
  - No Existence of the house

- Helpful for determining whether follow-up is needed or not
- Repeated follow-up up to 10 times for samples to maximize contact rate
- Used in weight adjustment and calculation of response rate or cooperation rate

Strategies for Enhancing Survey Quality

❖ Recruitment and Training of Interviewers

- Recruitment
  - Interviewers were recruited through online and offline
  - Most of interviewers were living near target industrial complexes or could easily approach to those places
- Interviewer training
  - Interviewers were trained for 9 days, totally 72 hours (8 hours per day)
  - The roll and importance of interviewers as well as interviewing skills and briefs of the survey were understood for them to have responsibility and confidence for the successful implementation

Strategies for Enhancing Survey Quality

❖ Recruitment and Training of Interviewers (cont.)

- Training contents
  - Outlines of the survey
  - Sample design and the way to listing operation
  - Pre-visit and identification of each responsible place
  - Administrative cooperation and guidelines for how to conduct survey
  - Understandings of the questionnaire and each survey question
  - How to run CAPI system and to record interim outcomes of the survey
  - Some possible embarrassing situations and how to deal with unexpected things
  - Repeated mock interviewing until the interviewers were thought to be well trained

❖ Requesting Administrative Cooperation

➢ New administrative cooperation for reducing nonresponse

Step 1	• Visiting City Office and “Gu” (District) Offices
Step 2	• Cooperation with “Dong” (Neighborhood) Offices
Step 3	• Cooperation with the Head of a“Tong”(primary division of a“Dong”)
Step 4	• Arranging schedules to visit sample households
Step 5	• Sending pre-notification letter
Step 6	• Using strategies for contacting households

Postsurvey Adjustment

❖ Weighting Adjustments

➢ Definition

- ➢ Compensation of sample cases that are underrepresented in the final data set by using some information about the target or frame population, or response rate information on the sample

➢ Adjustments applied to the initial weight (6 steps)

Step 1	• Initial weight
Step 2	• Removal of Out-of-scope Households
Step 3	• Household Non-response Adjustment
Step 4	• Creation of Person-Level Weight
Step 5	• Person Non-response Adjustment
Step 6	• Post-stratification by gender, age and location

❖ Contact Distribution of Samples Completed

➢ Number of visits for completed or uncompleted households

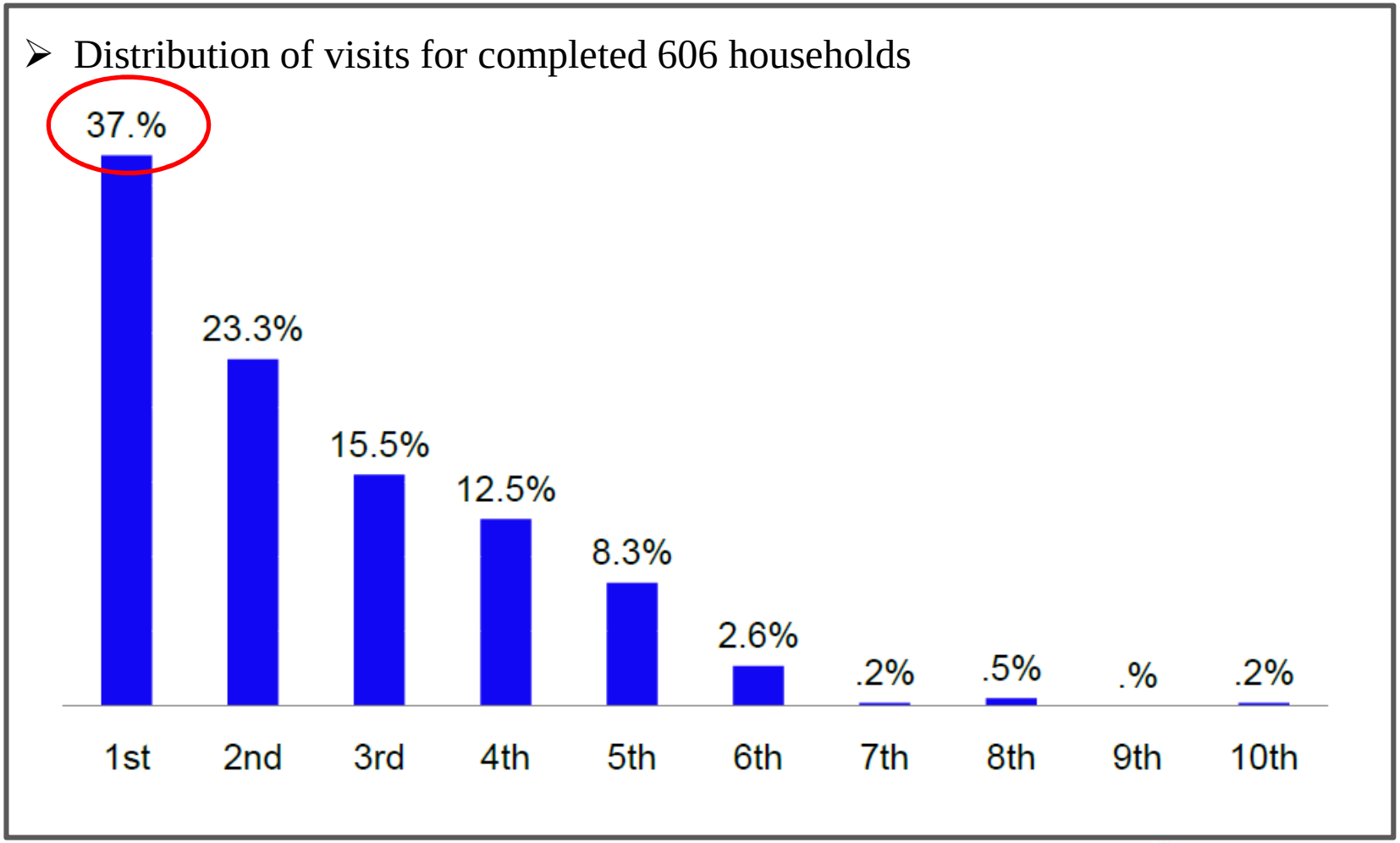
	# of HHs	# of visits
Completed households	**606**	**1,478**
Uncompleted households	**1,082**	**3,911**
Total	**1,688**	**5,389**

➢ Average number of visits per response to complete

Completed households	Total number of visits for households	Average number of visits per response
606	**5,389**	**8.9**

❖ Contact Distribution of Samples Completed (cont.)

➢ Distribution of visits for completed 606 households



Results

❖ AAPOR/WAPOR Response Rates

	Rates
RR1	36.1
COOP1	97.6
REF3	2.2

❖ Comparison between Population and Sample Distributions

	Sample		Population	
	Frequency	Percent	Frequency	Percent
Male	376	49.7	308,195	50.1
Female	407	50.3	307,380	49.9
Total	783	100.0	615,575	100.0

	Sample		Population	
	Frequency	Percent	Frequency	Percent
4-12	122	13.3	102,260	18.0
20-64	540	77.7	417,478	73.5
65 or higher	121	9.0	48,043	8.5
Total	783	100.0	567,781	100.0

❖ Differences of responses according to the number of cumulative visits

	Visits						(%)
	1	2	3	4	5	6	> 6*
Asthma	0.92	0.98	1.08	1.21	1.12	1.09	1.08
Allergic Rhinitis	5.51	6.11	5.78	5.74	5.56	5.46	5.42
Allergic Conjunctivitis	4.09	3.44	3.18	2.82	2.61	2.54	2.52
Cardiovascular Disease	1.41	1.07	1.22	1.41	1.36	1.43	1.42
Atopic Dermatitis	1.57	1.31	1.47	1.51	1.56	1.59	1.58
Thyroid disease	0.90	0.69	0.72	0.61	0.89	0.95	0.95

5. Advanced Sampling Methods

References

Brewer, K. R. W. (1975). A simple procedure for πpswor. *Australian Journal of Statistics*, 17:166–172.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer, New York.

Chen, S. X., Dempster, A. P., and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81:457–469.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, Rennes.

Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85:89–101.

Fan, C. T., Muller, M. E., and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computer. *Journal of the American Statistical Association*, 57:387–402.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14:333–362.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Madow, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20:333–354.

Tillé, Y. (2006). *Sampling Agorithms*. Springer, New York.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B15:235–261.

Survey Methodology (2014)

Optimal solutions in controlled selection problems with two-way stratification

Sun Woong Kim, Steven G. Heeringa and Peter W. Solenberger[1]

[PDF]

Abstract

When considering sample stratification by several variables, we often face the case where the expected number of sample units to be selected in each stratum is very small and the total number of units to be selected is smaller than the total number of strata. These stratified sample designs are specifically represented by the tabular arrays with real numbers, called controlled selection problems, and are beyond the reach of conventional methods of allocation. Many algorithms for solving these problems have been studied over about 60 years beginning with Goodman and Kish (1950). Those developed more recently are especially computer intensive and always find the solutions. However, there still remains the unanswered question: In what sense are the solutions to a controlled selection problem obtained from those algorithms optimal? We introduce the general concept of optimal solutions, and propose a new controlled selection algorithm based on typical distance functions to achieve solutions. This algorithm can be easily performed by a new SAS-based software. This study focuses on two-way stratification designs. The controlled selection solutions from the new algorithm are compared with those from existing algorithms using several examples. The new algorithm successfully obtains robust solutions to two-way controlled selection problems that meet the optimality criteria.

Key Words:

Cell expectation; Probability sampling; Distance function; Optimum array; Linear programming problem; Simplex method.

Table of content

Controlled Selection (Cont.)

8×3 Controlled Selection Problem

Causey, Cox, and Ernst (1985, JASA)

.4	2.0	.0	2.4
1.2	.0	1.0	2.2
.2	.0	.0	.2
1.2	.4	.2	1.8
1.0	.6	.2	1.8
.0	.4	.4	.8
.0	.2	.4	.6
.0	.0	.2	.2
4	3.6	2.4	10

$$E\left(b_{ijk}\middle|i,j\right)=\sum_{B_k\in\mathfrak{B}} b_{ijk}\,p\left(B_k\right)=a_{ij},\quad i=1,\ldots,R,\ \text{and}\ \ j=1,\ldots,C$$

Survey Methodology (2014)

Table 8.4
Comparison of solutions to Problem 2.3

B_k	$p(B_k)$					B_k	$p(B_k)$					B_k	$p(B_k)$				
	N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU		N_2	N_∞	SS	CA	HU
0 2 0 1 0 1 0 0 0 2 0 0 1 1 0 0 1 0 0 0 1 0 0 0	0.2	0.2	0.2			0 2 0 1 0 1 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1			0.11			0 2 0 1 0 1 0 0 0 2 0 0 1 0 0 0 1 0 0 0 1 0 0 1				0.2	
0 2 0 1 0 1 1 0 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 0	0.1	0.2	0.03			0 2 0 1 0 1 1 0 0 1 0 1 1 0 0 0 1 0 0 0 1 0 0 0			0.03			0 2 0 1 0 1 1 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 0 0				0.2	0.2
0 2 0 1 0 1 1 0 0 1 1 0 1 0 0 0 1 0 0 0 1 0 0 0	0.1					0 2 0 1 0 1 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1			0.03			0 2 0 2 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 0 0				0.2	
0 2 0 2 0 1 0 0 0						0 2 0 2 0 1 0 0 0						0 2 0 1 0 1 0 0 0					

Controlled Selection (Cont.)

SOCSLP: Software for Optimal Controlled Selection Linear Programming

Survey Methodology Program
Survey Research Center, Institute for Social Research
University of Michigan

Sun Woong Kim
Steven G. Heeringa
Peter W. Solenberger

SOCSLP is a public-use SAS-based software implementing the suggested algorithm in the paper "Optimizing Solutions Sets in Two-way Controlled Selection Problems" by Kim, Heeringa, and Solenberger (*Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 2002). It provides the following information about and solution to a controlled selection problem:

- The target frequency matrix (the given controlled selection problem): the cell and marginal expectations table.
- The tables of all possible samples (arrays) that meet the constraints on the expectations in the target frequency matrix.
- The distance values for all possible samples (arrays).
- The program code instructing the SAS LP Procedure (PROC LP) to find the solution to the controlled selection problem: the PROC LP setup.
- The output produced by PROC LP.
- The solution to the controlled selection problem.
- The sample (array) selected randomly from the solution.

SOCSLP is currently available for personal computers using the Microsoft Windows and Linux operating systems. It requires SAS version 6.12 or higher.

SOCSLP is freeware. The University of Michigan retains the copyright for SOCSLP and authorizes its use free of charge. See the SOCSLP License Agreement for details. Please report bugs or send comments to Peter Solenberger or Sun Woong Kim.

Download Documentation, Software, and Example

Documentation

Installation Guide
User Guide
Optimizing Solutions Sets in Two-way Controlled Selection Problems

Software

Microsoft Windows	socslp_windows.zip
PC Linux 32-bit	socslp_pclinux32.tgz
PC Linux 64-bit	socslp_pclinux64.tgz

Example

Setup
Description and output

Last updated 6 February 2010

감사합니다