북미통계학술대회 참가 결과보고

Joint Statistical Meetings 2015

2015. 9.

통계개발원

목 차

Ⅰ 출장 개요

❑ 참가 회의

○ 회의명

- 북미통계학술대회(JSM 2015*)

  * Joint Statistical Meetings : 미국 통계학회(ASA), 캐나다 통계학회(SSC), 영국 통계학회(RSS), 한국통계학회(KISS) 등 10여개 통계학회가 공동으로 참가하는 대규모의 통계연합학술모임으로, 매년 북미지역에서 개최

  * JSM 사이트 : www.amstat.org/meetings/jsm/2015

○ 회의 기간 및 장소

- 기　간 : 2015. 8. 8(토)~8. 13(목) (*참가기간: 11~13)
- 개최지 : 미국 시애틀

○ 참가 규모

- 702개 세션이 구성, 6일간 여러 개의 세션이 병행적으로 발표
- 미국, 캐나다, 유럽, 아시아 등 각 통계학회 및 기관과 연구자를 중심으로 통계 전반에 관련된 전문가 7,000여명 참석
  - 미국 내 한국인 통계학자로 이루어진 KISS*에서도 33개 세션에 참가

  * Korean International Statistical Society

❑ 출장자

- (조사연구실) 김경미 주무관

❑ 출장 수행 내역

○ 통계개발원 수행 연구논문 발표를 통한 연구 성과 확산(붙임1 참조)

- **제목** : The Analysis of Pilot Survey Data for the 2020 Rolling Census in Statistics Korea(#444 Government Statistics Section, 8.11.)

  · (연구목적) 순환센서스 도입 검토를 위해 실시된 시험조사 결과 자료를 이용하여 조사항목별 연간 공표 가능한 수준을 검토

  · (연구내용) 추정에 이용할 개인 및 가구 가중치를 개발하고 인구, 가구, 주택 부문 조사항목별로 누적자료에 대한 추정 및 오차분석을 통해 공표 가능 수준을 검토했으며, 시지역과 군지역의 특성이 다름을 발견

○ 연구동향 파악 및 네트워크 구축

- JSM2015는 'Statistics: Making Better Decisions' 주제로 Big Data 분석 및 모델링, 공간(spatial)·인구 종단자료분석, 각종 조사자료와 행정자료의 연계방법, 인구·가구추계 및 동향, 베이지안 모델링, 비밀보호 등 다양한 주제와 통계를 활용한 올바른 의사결정론을 다각도로 살펴보는 학술장임

- 대규모 합동 학술대회임 만큼 다양한 주제들로 연구이론과 실무기법이 소개될 예정으로 국가통계와 조사방법론을 중심으로 최신의 국제적 연구 동향을 파악하고 습득

- 또한 선진통계국 전문가들과 인적 네트워크 구성을 통한 국제적 연구협력체계 구축

Ⅱ 주요 내용

❑ JSM에서는 의학, 환경, 사회, 경제 전반에 걸쳐 통계와 관련한 다양한 논문들이 발표되고, 북미를 중심으로 다수의 국제통계학회와 학계와 민간이 함께 참가해 대규모로 개최됨

주요 분야

분야별

- Biometric Section
- Biopharmaceutical Section
- Business and Economic Statistics Section
- **Government Statistics Section**
- Health Policy Statistics Section
- Quality and Productivity Section
- Social Statistics Section
- **Survey Research Methods Section**

주제별

- Section on Bayesian Statistical Science
- Section on Medical Devices and Diagnostics
- Section on Nonparametric Statistics
- Section on Physical and Engineering Sciences
- Section on Risk Analysis
- Section on Statistical Computing
- Section on Statistical Consulting
- Section on Statistical Education
- Section on Statistical Graphics
- Section on Statistical Learning and Data Mining
- Section on Statistical and the Environment
- Section on Statistical in Defense and National Security
- Section on Statistical in Epidemiology
- Section on Statistical in Genomics and Genetics
- Section on Statistical in Imaging
- Section on Statistical in Marketing
- Section on Statistical in Sports
- Section on Teaching of Statistics in the Health Sciences

참가 및 후원

국제 통계학회

- American Statistical Association(ASA)
- Statistical Society of Canada(SSC)
- Royal Statistical Society(RSS)
- International Chinese Statistical Association
- International Indian Statistical Association
- Korean International Statistical Society(KISS)
- International Society for Bayesian Analysis(ISBA)
- International Statistical Institute
- RTI International 등

학계

- JASA 등 다수 저널
- Columbia University, Harvard University, Iowa State University, North Carolina State University, Ohio State University, Oregon State University, Penn State University, Rice University, Rutgers University, Southern Methodist University, Texas A&M University, University of North Carolina at Chapel Hill, Tsinghua University, University of California-Berkeley, University of California-LA, University of Illinois at Urbana-Champaign, University of Michigan, Yale University Alumni Reception, University of Washington, University of Waterloo, University of Wisconsin-Madison, University of Minnesota 등

민간

- Amazon ▪ Google ▪ SAS ▪ Stata ▪ Westat 등

❑ 국가통계 및 조사통계와 관련된 발표를 중심으로 참여

○ 총 702개의 세션이 개최,

- 이 중 국가통계와 관련한 세션이 100여개,
- 조사통계와 관련한 세션이 50여개,
- 국가통계이면서 조사통계와 관련한 세션이 30여개로 구성

○ 조사방법론 전반에 관한 다양한 주제의 논문들이 발표

- 조사설계, 표본설계, 추정 등 전통적인 조사방법 영역과 더불어 각 분야에 대한 베이지안 방법론의 적용과 빅데이터, 행정자료 활용, 다양한 자료원에 관한 연구 및 자료연계와 무응답 대체 방법, 자료의 비밀보호, 데이터의 품질 관련 연구 등

주요 주제(국가통계 및 조사통계)

- Survey Design
- Adaptive Design, Weighting and Design effect, Sample Allocation
- Bayesian Approaches
- Small Area Estimation
- Time Series
- Administrative Data
- Big Data
- Complex Longitudinal Data
- Multiple Sources of Data
- Missing Data, Nonresponse, Imputation
- Online Survey
- Data Linking
- Statistical Quality
- Confidentiality, Disclosure, Data Privacy

○ 빅데이터는 여전히 높은 관심을 가지고 있는 분야였으며, 10여개의 세션에서 40여개의 논문이 발표되었음

- 이와 더불어 빅데이터, 행정자료 등의 다양한 자료원에 대한 세션도 열려 이에 대한 주제들도 논의되었음

○ 무응답대체(imputation)방법에 대해서도 다양한 적용사례들이 발표되었는데 기존의 방법보다 모형에 기반한 방법들의 적용과 이에 대한 장점들이 많이 소개되고 있었음

○ 자료원이 다양해짐에 따라 자료연계(record linkage)분야에 대해서도 활발히 연구되고 있음

- 다양한 자료를 서로 연계한 사례들이 발표되고, 조사자료와 행정자료를 연계한 경우, 확률적 연계 방법을 사용한 경우 등 다양한 방식으로 연구되고 있음

○ 미국 센서스국에서는 2020 CENSUS를 대비하여 적응적 조사 설계, 행정자료의 활용, 무응답 대체 등 많은 연구가 진행되고 있으며, 이에 대해 단독으로 하나의 세션이 구성되어 발표되었음

Ⅲ 세부 내용

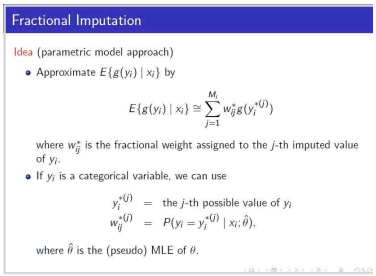
1 무응답 대체 방법

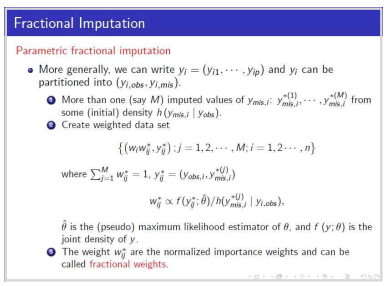
(1) Fractional Imputation Method for Missing Data Analysis in Survey Sampling (아이오와 주립대학)

- **(연구배경)** 무응답 혹은 여러 가지 이유로 누락된 데이터를 포함한 자료를 적절한 방법으로 무응답 대체 하여 완전한 데이터 파일을 제공하면 추정값의 편의를 감소시키고 모든 사용자가 일관된 분석 값을 가질 수 있음

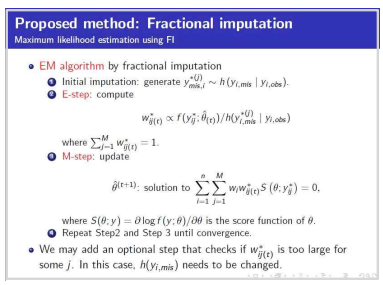
- **(연구내용)** 모수적 모형의 개념을 바탕으로 fractional weight을 이용하여 무응답 대체하는 방법을 소개하고 시뮬레이션 과정을 공유

Fractional Imputation





MLE 추정



- **(연구결과)** 전체 표본 추정값에 대한 가중치를 나누고, 서로 다른 목적의 자료 사용자들 사이의 일관성을 유지함에 있어서, 설문조사 방식의 자료에서 fractional Imputation (FI) 방법이 유용하다고 제안

(2) SRMI Multiple Imputation in the CPS ASEC
(미국 센서스국, U.S. Census Bureau)

- (**연구배경**) 일반적으로 센서스국은 소득과 관련한 무응답대체방법으로 핫덱(Hot deck)방법을 이용해 왔으나 본 논문은 모델 기반의 무응답대체 방법론을 제시하고자 함

- (**연구내용**) CPS(Current Population Survey)-ASEC(Annual Social and Economic Supplement) 자료에 순차적 회귀 다중 무응답대체(Sequential Regression Multiple Imputation, SRMI)방법을 적용하여 무응답을 대체하고, 소득 및 빈곤과 불평등과 관련한 항목에 대해 각각의 공식통계와 평균, 중앙값, 분산 등을 비교

- (**연구결과**) SRMI방법은 기존의 핫덱 방법보다 공변량을 추가하거나 불확실성을 고려할 수 있다는데서 모델 기반의 방법론에서 오는 유연성이 있는 점이 장점이라고 제언

(3) Exact Balanced Random Imputation(프랑스 통계청, INSEE)

- (**연구배경**) 무응답이 많이 발생할 경우 유효 샘플수가 감소하는 효과가 나타나며, 효율적인 무응답대체 방법의 연구가 필요

- (**연구내용**) 변수의 분포를 유지하면서도 무응답대체 분산은 높이지 않는 무응답대체방법으로 Balanced Imputation 방법(도너 혹은 잔차가 무응답 대체 분산을 제거하는 형식으로 랜덤하게 선택되어 하나하나 채워지도록 하는 Cube method)을 제안

- (**연구결과**) 평균제곱일관성(mean square consistency)을 계산하고 시뮬레이션한 결과로 본 방법론을 뒷받침하고 있음

2 자료연계

(1) SED/STAR Metric Record Linkage (American Institutes for Research)

- **(연구목적)** 조사자료와 행정자료의 연계를 통해 국립과학재단(NSF, National Science Foudation)의 박사과정 자금 지원의 효과 등을 파악할 수 있는 자료의 생성

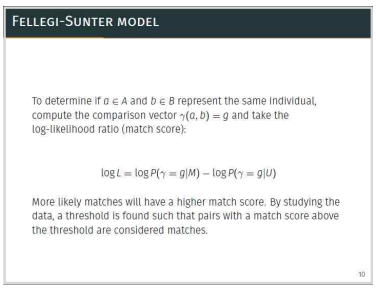
- **(연구내용)** SM(STAR Metrics)와 SED(Survey of Earned Doctorates) 자료를 연계하여 자료를 생성

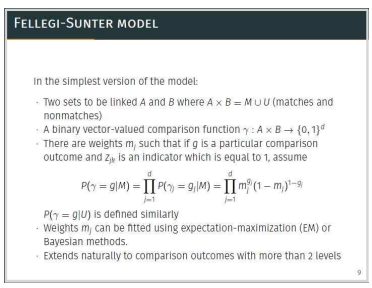
  ① SM : 미국의 부처 간 프로그램 자료로 정부의 R&D 지원과 관련된 데이터(어느 대학, 누구에게 지원금이 나갔는가 하는 자료)

  ② SED : 미국에서 박사학위를 취득한 사람을 대상으로 한 연간 조사 (인구통계학적 특성, 박사 후 계획 및 직업 등의 조사항목 포함)

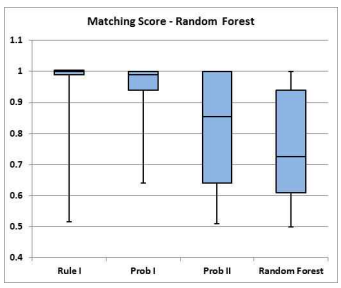
- **(연계방법)** Fellegi - Sunter 모델을 통한 연계 방법으로 매치스코어를 로그 우도비를 이용하여 확률적 연계

Fellegi – Sumter 모델





매치스코어



- **(연구결과)** 두 자료를 연계한 자료의 생성을 통해 박사과정의 정부 지원금과 박사 후 직업의 관계 및 각 대학별 특성 분석 등을 가능케 함

(2) NCHS Record Linkage Program (미국 질병통제예방센터, CDC)

- **(연구목적)** 미국의 건강, 보건, 질병 및 사망 등과 관련한 조사자료 및 행정자료를 통합 연계한 시스템의 제공

- **(연구내용)** 국민건강 면접조사(NHIS), 국민건강 영양조사(NHNES), 요양원 조사(NHHS) 등의 조사자료와 사망, 퇴직, 메디케어 서비스 등의 행정자료를 스코어링 시스템을 이용하여 연계

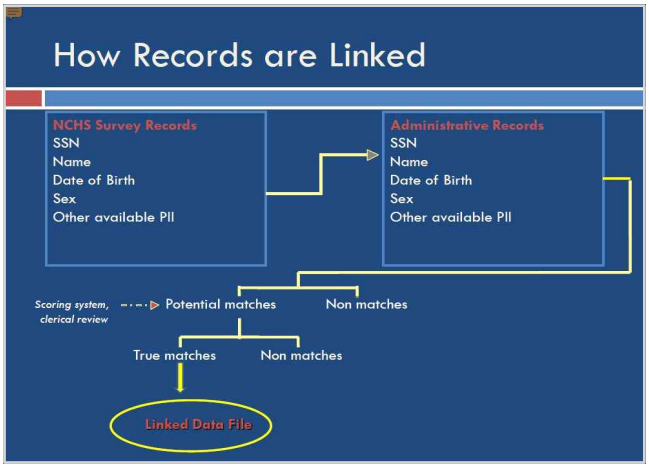
① 국민건강 면접조사(National Health Interview Survey, NHIS) : 건강상태 및 장애, 질병 및 장애의 정도, 보험, 의료 서비스 이용, 예방 접종, 건강을 위한 행동 등의 항목

② 국민건강 영양조사(National Health and Nutrition Examination Survey, NHNES) : 질병, 위험 요인, 영양 모니터링, 인체 측정 등

③ 요양원 조사(Nursing Home Survey, NHHS) :

· 시설 - 크기, 소유권, 진행 프로그램, 요금 등

· 개인 - 인구통계학적 특성, 건강상태, 약물투여 현황, 지출액 등



④ 행정자료 : 국립사망지수의 사망항목, 사회보장국(SSA)의 퇴직 및 장애관련 항목, 메디케어센터(CMS)의 메디케어 관련 등록 및 청구 자료 등

- **(연구결과)** 연계된 자료를 이용한 다양한 자료 분석 제안

* 자료의 비밀보호를 위해 CDC내 RDC(Research Data Center)에서만 분석 가능

· 인종/민족이나 사회경제적 지위에 의한 사망률 패턴 분석, 건강결과와 위험 요인의 상관관계 분석, 장애와 사망률과 관련한 만성 질환 혹은 비만의 효과, 자기보고 대 관리 기록의 검증 등

3 미국 2020 CENSUS (미국 센서스국, U.S. Census Bureau)

- 2010년 인구센서스 실시 후, 무응답의 후속작업(Nonresponse FollowUp operation, NRFU)에 대한 비용이 약 16억 달러가 발생함에 따라 미국 센서스국은 이러한 NRFU의 작업 부하를 감소시키고 보다 효율적으로 인구센서스를 실시할 수 있는 전략을 마련하고자 노력

- JSM2015에서 2020 Census에 대한 주제만으로 하나의 세션을 구성하여 현재까지의 성과 및 향후 계획을 공유

- 시험조사를 통해 반응적 조사설계, 행정자료의 활용, 무응답대체방법 등 다양한 검토를 진행중

2013년	2014년	2015년
2013 Census Test ▪ Philadelphia, PA ▪ October – December 2013 ▪ 2,077 housing units in sample ▪ 4 treatments – 2 involved administrative records ▪ Results and findings – Treatment 1 (fixed) – 7.8% removal of vacants and 31.3% removal of occupied units prior to NRFU – Treatment 2 (adaptive design) - 8% removal of vacants and 31.4% removal of occupied units prior to NRFU – Consider using additional sources to designate vacant housing units – Consider relaxing some rules used to match persons with occupied housing units	**2014 Census Test** ▪ Montgomery County, MD and Northwest Washington, DC ▪ June – September 2014 ▪ July 1, 2014 Census Day ▪ 151,759 housing units in sample – 65.7% self-response rate ▪ 46,247 housing units in NRFU ▪ 4 treatments in NRFU – 2 involved administrative records ▪ Results – "Option 2" or "hybrid" treatment – 32.6% of NRFU workload removed – "Option 3" or "full" treatment – 62.3% of NRFU workload removed	**2015 Census Test** ▪ Maricopa County, AZ ▪ April 1, 2015 Census Day ▪ 155,000 housing units in sample ▪ 60,000 estimated housing units in NRFU ▪ 3 treatments in NRFU – 2 involve administrative records ▪ An evaluation follow up of approximately 5,000 cases will occur where the NRFU response does not match administrative records

(1) Adaptive Design Research for the 2020 Census

- 반응적 조사설계를 통해 조사의 효율적 운영방안 마련
- CATI 방법의 효율적 활용 검토 및 테스트

	Fixed	Adaptive
Admin records identify vacants and enumerate occupied units before fieldwork	**Treatment 1** Administrative records • Remove cases from workload Telephone • If number, CAPI interviewers call • All numbers called twice Priority • *None* Number of visits • Three personal visits before proxy	**Treatment 3** Administrative records • Remove cases from workload Telephone • If number, CATI before field • CATI call procedures Priority • Propensity models determine priority Number of visits • Three personal visits before proxy
Admin records do not identify vacants and enumerate occupied units before fieldwork	**Treatment 2** Administrative records • *Not used* Telephone • If number, CAPI interviewers call • All numbers called twice Priority • *None* Number of visits • Three personal visits before proxy	**Treatment 4** Administrative records • Determine level of effort (number of contacts) Telephone • If number, CATI before field • CATI call procedures Priority • Propensity models determine priority Number of visits • If administrative record, one personal visit before proxy • Three personal visits before proxy

(2) Administrative Record Reseach to Resuce Contacts in the 2020 Census

- 조사 시 사전에 행정자료를 활용할 수 있는 방안 검토
- 3가지 디자인을 마련하여 시험조사를 진행

Design 1	Design 2	Design 3
		

- 이를 어떠한 행정자료가 필요한지에 대해 검토하고 또한 응답자가 행정자료 활용에 대해 어떠한 반응을 보이는지도 연구

Ⅳ 시사점

❍ (컨퍼런스 전반)

- JSM은 7,000여명이 참가하고 702개의 세션이 개최되는 통계학과 관련한 최대 규모의 학회임
- 학계와 민간, 정부기간 등 다양한 기관에서 적극적으로 참여하고 있으며 특히, 미국 센서스국의 경우 각각의 분야에 참가하는 것 이외에도 2020 Census와 관련한 하나의 세션을 별도로 운영하여 관련된 논문을 발표
- 미국 내 한국인 연구자들이 주축으로 구성된 KISS(Korean International Statistical Society)에서도 33개의 세션에 참석하는 등 활발히 활동하고 있었음

⇨ 통계개발원에서는 순환센서스와 관련한 연구를 꾸준히 진행하고 있으며 이 중 2014년에 진행되었던 연구의 일부를 포스터 발표

⇨ 향후에도 꾸준히 진행된 연구결과를 발표하고, 학회에 참석하는 정부기관 및 KISS와도 네트워크를 구축해 나가는 것이 연구 성과의 확산 및 발전에 도움이 될 것으로 판단

❍ (방법론 분야)

- 국가통계 및 조사통계의 방법론과 활용, 사례에 관한 모든 부문의 주제에 대해 논문이 발표
- 조사설계와 표본설계, 추정과 가중치, 소지역 추정 등의 주제는 꾸준히 연구되고 발표되고 있었으며, 각 분야에 대한 베이지안 방법론의 적용도 활발히 발표되고 있었음

- 빅데이터와 행정자료의 활용, 다양한 자료원에 관현 연구와 더불어 이를 활용한 자료연계와 무응답대체 방법론에 대한 연구도 많음
  - 자료연계는 조사자료와 행정자료를 연계하는 사례도 많이 발표되었으며, 무응답대체방법은 모형에 기반한 방법들의 사례가 많이 발표됨
- 자료의 비밀보호와 데이터의 품질 관련 연구에 관한 세션도 5여개 열렸으며 방법론에 관한 연구뿐만 아니라 적용사례에 관한 논문도 많이 발표됨

【붙임1】 발표자료 (#444 Government Statistics Section, 8.11.)

The Analysis of Pilot Survey Data for the 2020 Rolling Census in Statistics Korea

KyungMi Kim(27kyung@korea.kr), JaeHyuk Choi(leonash@korea.kr)
Statistical Research Institute (http://sri.kostat.go.kr)

Introduction

Motivation

Change of the Korean Census

AS IS (every 5 year)

A short-form Census was completed by everyone, and a more detailed long-form was answered by a 10 percent sample of the population.

TO BE

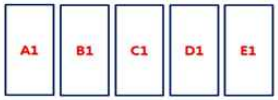
The short-form Census is replaced by register-based Census from the year of 2015. Also, we are preparing the rolling Census for a 20 percent sample of the population for the next Census.

Subject

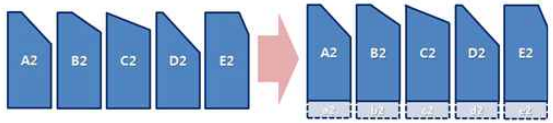
Since 2012 Statistics Korea has conducted pilot survey of the rolling census in two municipalities, **city** and **countryside**

Basic sample design (period: every 5 year)

- Step 1 : Divide the first population into 5 parts.



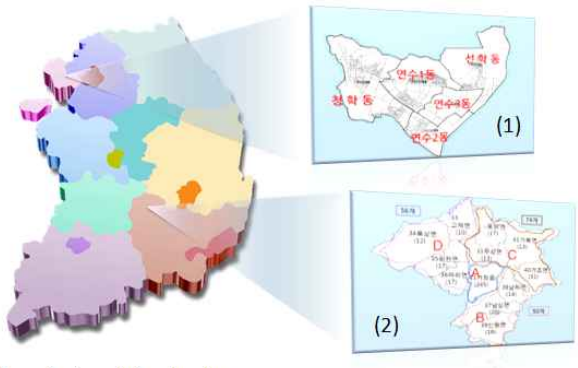
- Step 2 : Modify and complement to the population.



Method and Data

Data Collection

- Area & period :
  - (1) city Incheon Metropolitan city, Yeonsu-gu
    March 2013 ~ December 2014
  - (2) countryside Gyeongsangnam-do, Geochang-gun
    October 2012 ~ September 2014



- Face to face interviewing
- Item(Census Questionnaires) : population(28), household(13), housing(6)

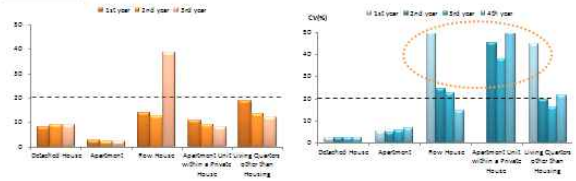
Data Sets

Estimation Period	(1) Yeonsu-gu / city /	(2) Geochang-gun / countryside /
1st year	March 2013, June 2013	October 2012 ~ February 2013
2nd year	September 2013, December 2013	March 2013 ~ July 2013
3rd year	January 2014 ~ June 2014	August 2013 ~ December 2013
4th year		January 2014 ~ May 2014

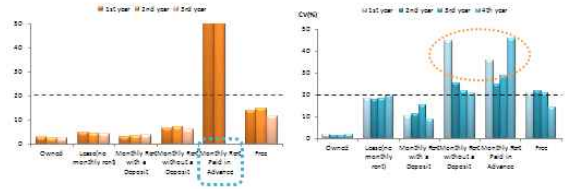
Results

- Cumulative result was estimated by applying the design weights and post-stratifications weights.
- When the RSE(Relative Standard Error) of estimate has a value of less than 20%, it is evaluated as stable.
  - The city has a stable value compared to the countryside.
  - Also, it has regional characteristics in some items.

Housing



Type of Residence



Further Study

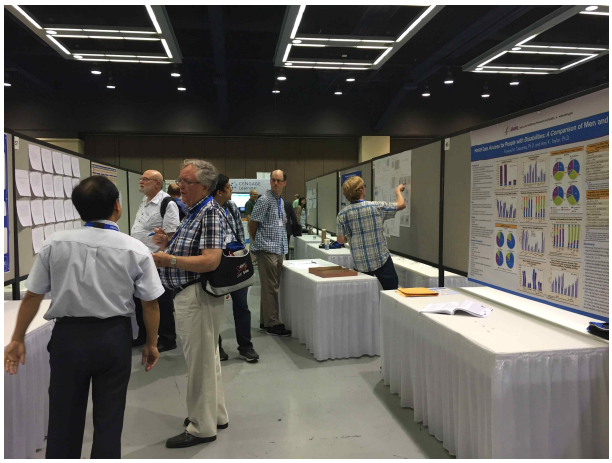
¶ Additional analysis of first pilot survey (2015)
- Socio - demographic analysis and implications.

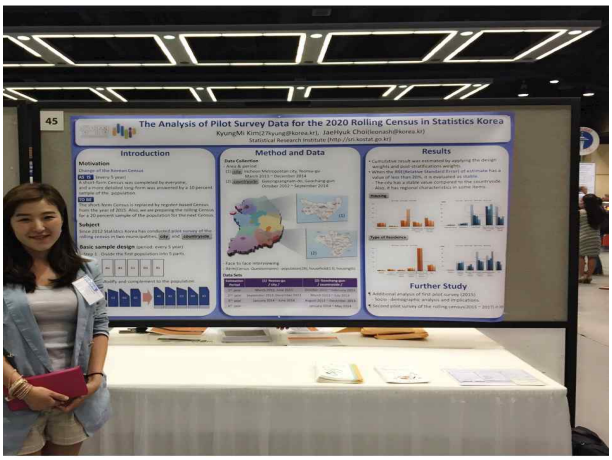
¶ Second pilot survey of the rolling census(2015 ~ 2017) in Jeju

【붙임2】 컨퍼런스 전경

❑ 일반세션

❑ 포스터세션 & 발표





❑ 전시장