2016 북미통계학술대회 참가 결과 보고
(Joint Statistical Meeting 2016)

2016년 8월

통계청

통계개발원 조사연구실

목 차

1. 출장 개요

○ **학술대회명** : 「**북미통계학술대회**[1], JSM 2016」

- 개최기간 : 2016. 7. 31(일) ~ 8. 4(목)
- 개 최 지 : 미국, 시카고

* 북미통계학술대회(Joint Statistical Meeting) : 미국통계학회(ASA), 캐나다통계학회(SSC), 영국통계학회(RSS), 한인통계학회(KISS) 등 170개 이상의 학회, 대학, 통계기관 등이 공동으로 협력하여 개최하는 대규모 학회.

○ **출장기간** : 2016. 7. 31(일) ~ 8. 6(토)/ 5박 7일

○ **출 장 자** : 박민정 사무관, 조사연구실

○ **출장 목적**

- JSM은 52개 이상의 국가에서 6000명 이상의 참가자들이 600개 이상의 세션들에 참가하는 대규모 합동 학술대회로, 통계적 방법론 이론 및 응용뿐만 아니라 통계학의 경계를 허무는 다양한 영역의 주제까지 다룸
- 이에, 국가통계와 조사방법론을 중심으로 최신의 국제적 연구 동향을 파악하고 습득하며 빅데이터 관련 최신 연구동향도 파악하고자 함
- 더불어 조사연구실의 연구 내용을 KISS(한인통계학회)에서 주관하는 세션에서 발표하여 향후 연구 협력 체계 구축을 도모하고자 함
- 기타 해외 전문가들과 네트워크 형성을 통한 국제적 연구협력체계 구축

1) 북미 지역에서 미국과 캐나다가 연합하여 양국의 여러 도시에서 번갈아 열리는 학술대회여서 북미통계학술대회로 번역함

2. 주요 내용 및 시사점 (대주제별)

❒ Big Data Issues (총 3개 세션, 10개 주제 발표)

337 Fusion Learning and Combining Inference from Diverse Complex Data Sources

#1. Model Calibration Utilizing Summary Level Information from External Big Data, Nilanjan Chatterjee, The Johns Hopkins University 외 3인

↪ Model Building using Data Integration (제목변경)

질병관련 자료에서 요약통계 모수를 추정할 때, 외부의 빅데이터 원천들을 활용하여 정확도를 높일 수 있다. 이에 관한 이론과 시뮬레이션 결과를 발표하였다. 두 자료의 통합은 메타 자료 분석 과정을 활용해 이루어진다. 제안된 제약된 추정량(constrained MLE)를 사용하는 일반적인 이론과 추정에 관하여 2016년 JASA에 게재된 논문의 내용 일부를 발표하였다.

#2. Efficient Bayesian Inference on Genetic Association, Hélène Ruffieux, Ecole Polytechnique Fédérale de Lausanne 외 3인

↪ Efficient Inference for genetic association studies with multiple outcomes

유전체 자료의 연관성 분석을 할 때 고차원 자료들을 결합하여 추론을 하는 문제는 많은 어려움을 가지고 있다. 그 중 고차원으로 인한 성긴 회귀 모형(sparse regression model)을 위해 베이지안 접근을 연구하였다. 동시에 예측변수와 반응변수를 감지 및 선택하는 방법을 제안하였다.

#3. Generalized Fiducial Inference for Massive Heterogeneous Data, Jan Hannig, The University of North Carolina at Chapel Hill

↪ Fusion Learning for Inter-Laboratory Comparison

일반적으로 정확도를 추정하거나 나타내기 위한 도구로 신뢰구간, bootstrap, Bayesian posterior, generalized fiducial distribution 등이 있지만, 다양한 원천의 정보를 결합하기 위해 confidence distribution(CD)를 사용할 수 있다. CD의 fiducial combination을 제안하고 대용량 자료에 어떻게 쓸 수 있는지 등을 설명하였다. 베이지안이 철학적으로 설득력을 얻는 것처럼 fiducial 접근이 20년 후에 그러한 의의를 가질 것이라 예상한다고 주장하였다.

#4. Fusion Learning from Complex Data Sets to Efficient Goal-Directed Individualized Inference, Regina Liu, Rutgers University 외 1인

여러 자료 원천들을 함께 이용해 추론하고 다양한 분야의 결정 과정(decision-making

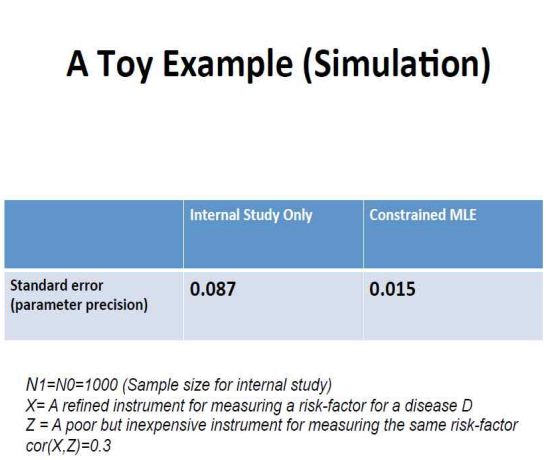
process)에 활용할 때 퓨전학습(fusion learning) 기법을 활용한다. 특히 유용한 정보를 추출하고 결합(merging)하기 위해 CD(confidence distribution) 접근을 활용한다. 예를 들어 이러한 접근 방식으로 두 종류의 항공기 착륙거리 및 높이 비교를 빅데이터를 활용하여 할 수 있음을 보였다.

(337) 정리: 다양한 출처의 대용량 자료를 결합하기 위한 방법론

빅데이터는 다양한 출처를 가지고 있고 대용량이므로 이들을 결합하여 의미 있는 정보를 만들어내기 위해 효율적인 방법론에 대한 연구가 필요하다. 이 세션에서는 새로운 통계 모형의 제안 및 활용, 베이지안 접근, 퓨전학습이론과 그 중 fiducial 접근에 대하여 발표되었다. 대상 자료는 질병 자료, 유전체 자료, 항공기 운항 자료 등의 대용량 자료이다.

국가통계 방법론에서 베이지안 접근이나 퓨전학습이론이 아직은 수용하기 쉽지 않으나, 일반적인 빅데이터 분석에 이러한 접근이 폭넓게 이루어지고 있음을 기억할 필요가 있다. 기존 조사자료에 외부 빅데이터를 결합해 요약통계량의 정확도를 높이는 연구(#1)는 관련 과제를 수행하게 될 경우 참고할만하다고 판단된다.

#1 발표내용 인용[2)]

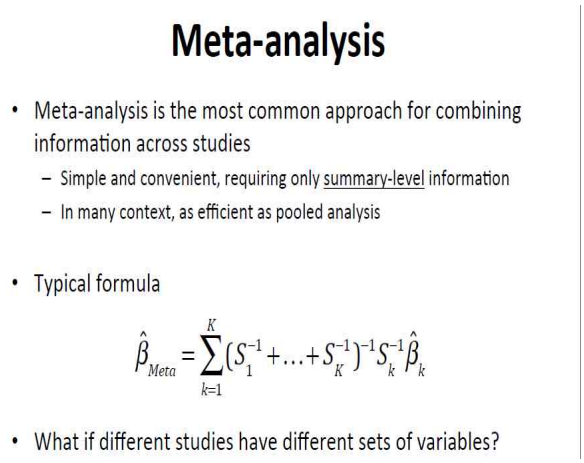


A Toy Example (Simulation)

	Internal Study Only	Constrained MLE
Standard error (parameter precision)	**0.087**	**0.015**

N1=N0=1000 (Sample size for internal study)
X= A refined instrument for measuring a risk-factor for a disease D
Z = A poor but inexpensive instrument for measuring the same risk-factor
cor(X,Z)=0.3

조사자료 이외에 외부 빅데이터를 결합해 요약통계량의 정확도를 높이는 시뮬레이션 결과



Meta-analysis

- Meta-analysis is the most common approach for combining information across studies
  - Simple and convenient, requiring only summary-level information
  - In many context, as efficient as pooled analysis
- Typical formula

$$\hat{\beta}_{Meta} = \sum_{k=1}^{K} (S_1^{-1} + \ldots + S_K^{-1})^{-1} S_k^{-1} \hat{\beta}_k$$

- What if different studies have different sets of variables?

메타 분석 설명 중 일부 내용

---

2) JSM에서 발표 내용은 외부로 유출시키지 않으며, 발표자가 허용하는 경우만 발표 자료를 시스템에 업로드하는 speakers룸에서 접근이 가능하다. 사진 촬영도 금지되어 있으나, 참가자들은 관심 있는 슬라이드들을 부분적으로 사진 촬영을 하며, 많은 경우 발표 자료는 발표자에게 학회이후 이메일로 요청하여 받을 수 있다.

468 Uncertainty Estimation for Massive Data Sets

#1. How to Ask Questions of Huge Data with Few Samples, David Dunson, Duke University

↪ Flexible Bayes Inference from Huge data: Modularization

매우 고차원인 자료를 다루기 위해 차원을 축소하는 문제를 베이지안 관점에서 다루고 새로운 알고리즘을 제안하였다. 비모수적 베이즈 변수 선택을 하며 핵심 아이디어는 모듈화(modularization)이며 시뮬레이션 결과를 보여주었다.

#2. Divide-and-Conquer and Statistical Inference, Michael I. Jordan, University of California at Berkeley

병렬 연산에서 하둡(hadoop)이나 스파크(spark)와 같은 플랫폼은 큰 규모의 자료를 처리한다. 이런 플랫폼을 통계적 개념과 유기적으로 이해하고 활용할 필요가 있다. The bag of little bootstrap(BLB) 아이디어는 병렬 연산 계산속도와 정확도를 획기적으로 개선한다. 유사한 관점에서 빅데이터 처리에서 최적화(optimization) 해를 찾는 방안도 설명하였다.

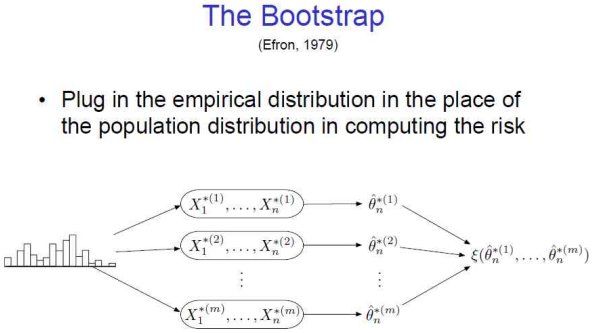
#3. A Variational Bayesian Analysis of Stochastic Gradient Methods, Matt D. Hoffman, Princeton

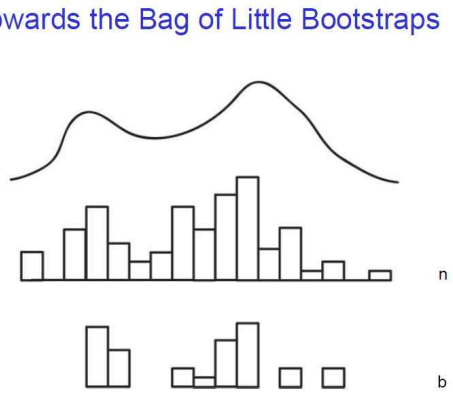
SGM(Stochastic gradient method/descent)는 통계학에서 최적화 해를 찾는데 널리 사용되고 있는 기법인데, 다양게 파생된 베이지안 알고리즘들을 설명하였다.

(468) 정리: 대용량 자료에서 차원 축소와 계산 성능 향상

이 세션에서는 빅데이터, 고차원 대용량 자료를 다루는 최신 방법들이 주로 베이지안 관점에서 연구되고 있는 흐름을 파악할 수 있었다. 향후 통계기관에서 고차원 빅데이터를 분석하게 될 경우, 차원 축소의 문제를 해결할 때 베이지안 방법론을 수용할 필요가 있다. 빅데이터 공급시 플랫폼에서 BLB알고리즘을 적용하는 것은 효율적이어 보이며 실제 적용 사례가 있는지 확인하여 참고할만하다고 판단된다.

#2 발표내용 인용





전통적인 부츠트랩 아이디어 | BLB의 직관적인 이해를 돕는 그림

634 Analysis, Storage, and Privacy for Big Data

#1. Distributed Data Analysis at Scale, Tom Peterka, Argonne National Lab

분자역학 등 과학 분야 대용량 자료를 다룰 때, 자료를 블락으로 나누어 분석하는 기법들을 소개하고 논의하였다. 분포 추정, 정확도 등을 다루었다.

#2. Storage Issues and Assessment Arising from Large Scale Simulations, Emily Casleton, Los Alamos National Laboratory 외 2인

대용량 자료를 다루기 위해서는 자료 저장의 문제 역시 많은 연구가 필요하다. 공간 자료를 저장할 때 분할(partitioning)을 어떻게 할지 연구하고, 몇몇 거리 함수들에 대한 제안된 분할 기법의 효과를 해양 자료 시뮬레이션을 통해 보였다.

#3. Differentially Private Data Synthesis Partitioning for Big Data, Claire McKay Bowen, University of Notre Dame 외 1인

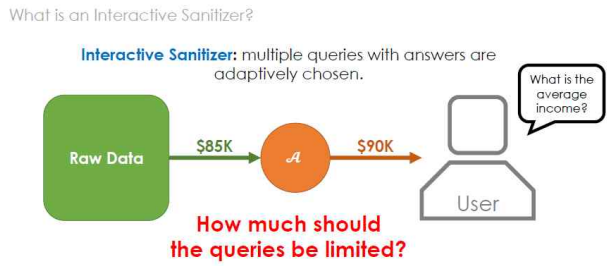
빅데이터를 다루기 위해 풀어야할 도전 과제는 비밀보호 문제이다. 과거 기법의 한계를 벗어나기 위해 차등적 정보보호(differential privacy) 개념이 제안되어 있는데, 쿼리와 관련해 이를 어떻게 풀어갈지 연구하였다. DIPS(differentially private data synthesis)를 이용한 자료 배포를 제안하고, 빅데이터에서 노출의 문제를 예를 들어 설명하였다.

(634) 정리: 대용량 자료의 저장과 비밀보호된 자료 공급

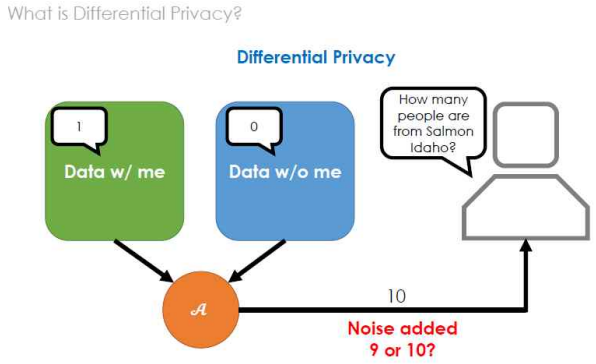
이 세션에서는 과학 분야 대용량 자료를 다루기 위한 기법, 자료 저장을 위한 기법, 그리고 비밀보호 이슈를 다루었다. 과학 분야 이슈는 국가통계방법론과 큰 관련을 찾기 어렵고, 대

용량 공간 자료 저장을 위한 최신 기법으로 분할에 관한 이슈들이 다루어지고 있음을 참고할만하다. 마지막으로 빅데이터에서 노출의 문제와 DIPS를 설명한 연구는 현재 수행 중인 과제의 참고 자료로 바로 활용할 필요가 있다.

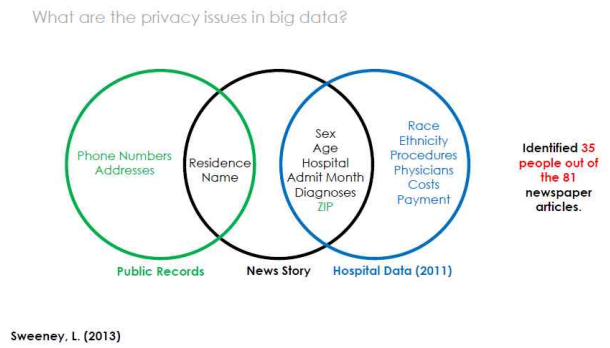
#3 발표내용 인용



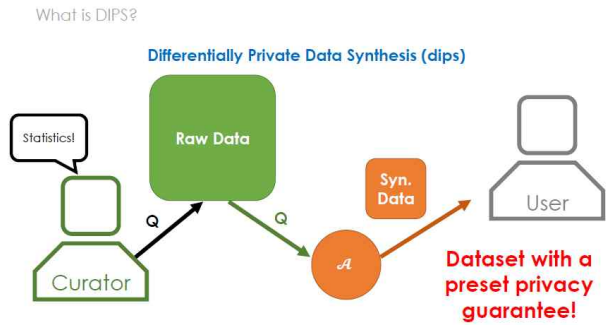
노출 제어를 위한 쿼리 제한 기준 설정의 문제



Differential privacy의 직관적인 이해를 돕는 그림



빅데이터에서 노출이 일어날 수 있는 사례



DIPS의 직관적인 이해를 돕는 그림

❒ Disclosure Control Issues (총 3개 세션, 9개 주제 발표)

152 Cyber Security in Support of National Defense and Global Security

#1. Estimation of True Quantiles from Quantitative Data Obfuscated with Additive Noise, Bimal Roy, Centre for Cryptology and Security 외 1인

일종의 매스킹 기법인 Data Obfuscation을 응용한 연구 결과 발표. 이를 이용하여 잡음을 추가하는 수학적 이론을 설명하고, 균일분포, 정규분포, 라플라스 분포를 따르는 자료에 잡음을 추가한 후 분위수들에 대해 추정값을 비교하였다.

#2. Enabling Privacy Preserving Machine Learning at Scale, Farinaz Koushanfar, UCSD

보안 문제는 data science 접근보다는 컴퓨터 과학적 접근과 관련되고 Bayesian networks, Petri nets, 암호학, Categroy theory 접근 등의 모형이 있다. 센서를 이용하여 보안 문제를 해결하려 할 때 프라이버시 문제가 발생한다. 이를 해결하기 위한 컴퓨터 과학의 이론을 약간 소개하였다.

(152) 정리: 자료 보안

자료에 잡음을 추가하여 개인 정보 노출을 방지할 수도 있으며, 이를 위해 Data Obfuscation 기법을 활용할 수 있다. 연구 결과에 따르면 실무에 적용하기에는 분위수 추정값 기준으로 정보손실이 크다고 판단된다. 한편 물리적 보안 시스템에서 센서를 이용할 때 프라이버시 침해 문제가 발생할 수 있는데, 과거에는 "좋고 빠르고 저렴한" 방안을 찾는데 주력하였으나, 최근에는 프라이버시 문제로 "안전하고 쓸모있고 저렴한" 방안을 찾는 기조가 되었다.

#1 발표내용 인용

Table: Showing True and Estimated Quantiles, $\epsilon = 200$

Alpha	"TRUE"	"Estimated"
"0.1"	580.8	536.873
"0.2"	612.8	588.339
"0.3"	645.2	629.159
"0.4"	675.6	666.
"0.5"	700	700.941
"0.6"	727	740.717
"0.7"	750	774.214
"0.8"	786	815.547
"0.9"	826.6	866.275

노출제어(Data Obfuscation) 처리 결과 분위수 추정값에 대한 시뮬레이션 결과, 정보손실 발생이 크다.

#1. An Integrated Approach to Providing Access to Confidential Social Science Data, Jerome Reiter, Duke University

노출제어 최신 이론인 synthetic data 및 differential privacy에 대한 내용을 예제와 함께 정리하여 발표하였다. 참가자들이 전체 흐름을 배우는 유익한 시간이었으며, 두 방안의 향후 연구 방향 설정을 위한 참고 자료로 활용할 예정이다. 일원화된 시스템을 통한 자료 배포의 시너지에 대해 강조하였다.

#2. The Challenge of Reproducible Science and Privacy Protection for Statistical

Agencies, John M. Abowd, U.S. Census Bureau/Cornell University

노출제어 대안으로 거론되고 있는 synthetic data와 differential privacy는 서로 경쟁적인 방법으로 인식하여 왔으나, 이 발표를 통해 둘의 의미가 서로 다름을 파악하였다. 노출제어 대안 마련은 어려운 문제로 올바른 방안 선택에 대한 고민은 지속되어야 하며, 발표를 통해 소개된 세부 사항들은 향후 과제에서 심층 검토할 예정이다.

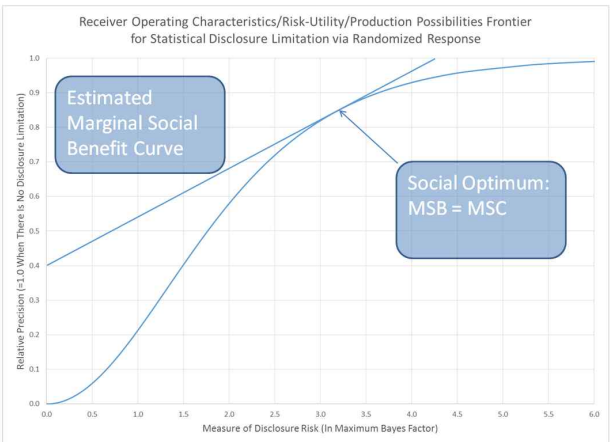
#3. Spatio Temporal Change of Support with Application to American Community Survey Multi Year Period Estimates, Scott H. Holan, University of Missouri 외 2인

ACS 자료의 multi-year 추정량을 얻고자할 때 구역 변화(change of support) 문제를 해결해야 한다. 이를 위해 data model, process model, random effects parameterization에 관한 이론적인 내용을 발표하고, 가구소득 중앙값 3년 추정값을 지도에 표현한 결과를 보여주었다.

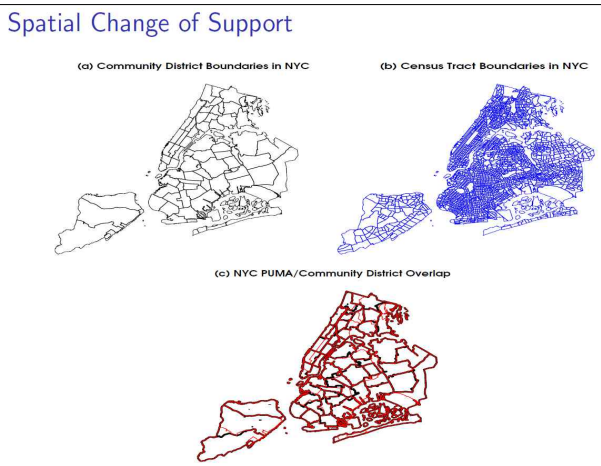
(218) 정리: 국가통계 노출제어 연구의 최근 흐름

노출제어 분야의 손꼽히는 두 대가의 발표를 통해 최신 이론의 흐름을 느낄 수 있었고, 향후 과제 수행에 세부 내용을 입수하여 활용할 예정이다. 세 번째 발표는 노출제어 분야는 아니나, 패널 자료의 추정량에 관한 이론적 내용과 실제 활용 사례로 향후 참고할만하다.

#2 및 #3 발표내용 인용



노출위험과 정보손실 상충성(trade-off)에 대한 고민



추정을 위해 해결해야하는 구역 불일치 현상

527	Innovations in Disclosure Avoidance at the U.S. Census Bureau

#1. Controlling Identification Disclosure Risk in Microdata Release Through Unbiased

Post Randomization, Cheng Zhang, The George Washington University 외 1인

매스킹 기법 중 하나로 널리 쓰이고 있는 PRAM을 마이크로데이터에 적용하여 혼인상태의 빈도수를 기준으로 자료 유용성을 측정하였다. 노출제어를 학위과정으로 전공한 사람은 세계적으로도 드문데, 박사과정 학생이 관련 기법을 공부한 결과를 발표한 것은 이 분야의 연구 활동 범위가 커지고 있다는 고무적인 신호로 볼 수도 있다.

#2. Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau, Amy Lauger, U.S. Census Bureau 외 2인

미센서스국에서 ACS마이크로자료를 가지고 synthetic data를 만들어보는 연구 결과를 발표하였다. CART 대신에 Dirichlet 모형을 사용하고 두 결과를 비교하였다. 치우친 분포를 가지는 임금 자료 다루기의 어려움도 설명하였다. 이 분야의 권위자인 Reiter 교수의 도움을 받아 관련 연구를 시작하고 간단한 결과를 발표하였다.

#3. Estimating Regression Parameters from a Sensitive Variable with Noise Multiplication, Marlow Lemons, U.S. Census 외 2인

외부 공격자는 일부 정보를 아는 특정 개인을 찾으려는 유형과 무조건 식별(re-identify)하려는 유형으로 나뉠 수 있다. AHS(American House Survey)에 대해 식별된 자료를 생성하려는 공격자에 대해 지역변수 등이 활용될 수 있어 위험하며, 여러 다른 자료를 이용해 연계하는 공격자에 대한 후속 연구를 계속해 나갈 것이다.

#4. Challenges Facing the Disclosure Review Board at Census, William Wisniewski, U.S. Census Bureau 외 1인

노출제어를 위한 위원회(Disclosure Review Board)는 품질 좋은 자료를 관련 법규를 지켜가며 제공하기 위해 노력하고 있다. 위원회의 역할은 다양하므로 위원들은 관련 법뿐만 아니라 방법론에도 능숙해야하고, 위원회는 구체적인 체크리스트를 가진다. 체크리스트는 공공이용 마이크로데이터 파일, county 단위 빈도표, 경제 관련 총계표 등에서 반올림 규칙 등등의 세부 사항들로 구성된다.

(527) 정리: 미센서스국의 최근 노출제어 현황

미센세서국에서는 ACS 마이크로자료를 이용해 최신 기법인 synthetic data 생성에 관한 내부 연구를 시작하였고, 지역 변수가 중요한 AHS에서 연계로 인한 노출위험성을 줄이기 위한 연구를 계속 진행해 나갈 것이다. 한편, 위원회인 DRB를 통해 통합된 가이드라인을 가지고 노출제어 처리를 하기 위한 노력을 지속하고 있다. 학계에서 노출제어 관련 연구를 학위과정에서 진행하는 사례가 생겨나고 있다. 자료별로 노출제어 방식이 달라지므로 한국 통

계청도 다양한 방향에서 노출제어 연구를 진행해 나갈 수 있다면 좋겠다.

❒ Visualization and Graphics Issues (총 2개 세션, 10개 주제 발표)

54	Recent Advances in Information Visualization

#1. Using Maximum Topology Matching to Explore Differences in Species Distribution Models, Jorge Poco, University of Washington

종(species)의 분포를 모형화하고 모형들 사이에서 유사성을 토폴로지 이론을 활용하여 측정한다. 분포의 깊이, 높이, 넓이 등을 수치화하고 고차원 자료에 적합한 장점을 가지고 있으므로 토폴로지를 이용한다.

#2. Automatic Selection of Partitioning Variables for Small Multiple Displays, Anushka Anand, Tableau Research

관심 있는 구조의 특징을 잡아내기 위한 시각화 방법을 제안. 예를 들어 입학 허가 비율에 대한 졸업 비율의 분포를 사립학교와 공립학교로 나누어 표현하거나, 졸업 비율이 높은 집단, 중간 집단, 낮은 집단으로 나누어 시각화하면 좀 더 많은 정보가 간결하게 전달된다. 또한 시각적 패턴이나 알고리즘 평가 모두 달라진다. 이렇게 군집화할 수 있는 시각적 패턴 특징들을 cognostics라 부르며, 이를 이용하여 작은 규모의 복수 시각 자료(multiple displays)를 제공할 수 있다.

#3. Visualizing Statistical Mix Effects and Simpson's Paradox, Zan Armstrong, Freelance Data Visualization

구글의 데이터 시각화 연구팀과 공동 연구한 결과를 발표. 임금 중앙값과 낮은 임금 근로자에 대한 자료를 수집할 때 정보손실이 많고, 자료를 나누어 조각(segment)별로 분석할 때 평균이 어떤 조각과도 반대로 움직이는 현상인 심슨의 역설이 나타난다. 적절한 정보가 잘 드러나지 않는다면 각종 차트는 쓸모가 없으며, 분석 결과 시간에 따른 변화를 나타내는 시각화가 중요하다. 이를 위해 comet chart를 제안한다. comet chart를 활용하여 노동력 크기별로 실업율을 표현하거나, 고용인구별로 임금 중앙값을 교육수준별 등으로 나타내어 그룹별 군집별 변화를 한눈에 볼 수 있다. 6개의 주요 측도(metrics)들을 한 차트에 나타내는 장점이 있다. (research.google.com/pubs/pub42901.html)

#4. ImMens: RealTime Visual Querying of Big Data, Zhicheng Liu, Adobe Research

빅데이터를 어떻게 시각화하고 자료와 대화(interact)할 것인가는 오래 연구할 문제이다. 게다가 변수까지 많으면 더욱 복잡한 문제가 된다. 한 변수에 대해 천 개 자료만 산점도를

그려도 보기에 적절하지 않으므로, 빅데이터 시각화는 표본추출, 모형 결과, binned aggregation 등의 조치가 필요하다. binned aggregation을 위한 ImMens를 소개하였다. (github.com/uwdata/imMens)

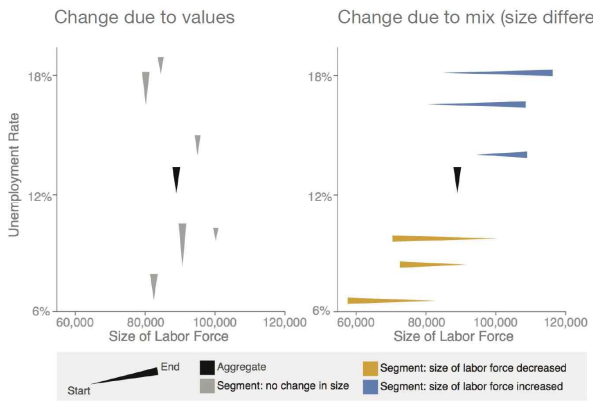
#5. An Algebraic Process for Visualization Design, Gordon Kindlmann, The University of Chicago

자료의 변화가 시각화 결과 잘 드러나도록 algebraic process를 제안, 수학적 접근을 설명

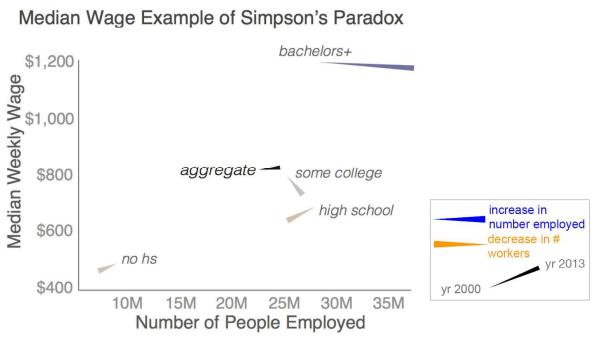
(54) 정리: 자료 시각화 최신 기법

빅데이터 시대에 자료 공급을 위해 효율적인 시각화를 위한 연구 결과들이 발표되었다. topology 혹은 algebraic process 등의 수학적 접근도 있으나, cognostics라는 시각적 패턴에 근거해 자료를 나누어 시각화하거나, comet chart를 활용하거나 binned aggregation 아이디어에 근거한 시각화를 할 수 있다. 국가통계와 관련하여 자료 제공시 좀 더 세부적 단위로 나누어 차트를(small muptiple displays) 제공하거나 comet chart를 활용하는 방안은 향후 관련 연구에서 참고할 가치가 있다고 판단된다.

#3 발표내용 인용



comet chart의 기본 표현, 다양한 정보 포함 가능



그룹별로, 전체로 통계 변화를 표현

695 Methods and Applications of Statistical Graphics

#1. Visualization of Latino Political Participation in Nebraska, Farhana Luna, University of Nebraska Omaha 외 2인

선거에서 라틴 인구 증가의 영향을 예측하기 위해 다양한 출처의 복잡한 자료를 생성하고, 이들을 시각화하고, 패턴을 탐색하기 위해 R패키지 Shiny를 사용한 결과를 발표하였다.

#2. Use of Phase Plots to Explore Financial Data, Mahbubul Majumder, University of Nebraska Omaha 외 1인

위상의 시각화는 금융 상태 예측에 도움을 주기 때문에 금융 자료 시각화를 위해 위상(phase)을 시각화하는 연구 결과를 발표하였다.

#3. Improved Simulation for Exponential Random Graph Models for Social Network Analysis, Junchi Guo, The George Washington University 외 1인

ERGMs(Exponential Random Graph Models)는 소셜 네트워크 분석에 효율적이어서 널리 사용되고 있으며, 2015년 논문에서 보인 MCMC 샘플링을 사용할 때 층화 샘플링을 적용하여 Metropolis Hasting을 더욱 효율적이게 한다는 연구 결과에 더하여, 좀 더 세부적인 발전 사항들을 다루었다.

#4. Confident Class Micromaps for Visual Analytic Inference, Daniel Carr, George Mason University 외 1인

마이크로맵은 지역을 분류하기 위해 신뢰구간을 사용하며, 하나의 변수에 대하여 3층 맵(three-class map)이 생성된다. 이차, 삼차 신뢰 층 맵 디자인을 위해 기법 확장을 연구한 세부 결과를 보여주었다.

#5. Integrating LargeScale MultiOmics Data to Achieve Causal Inference in Observational Studies, Azam Yazdani, The University of Texas Health Science Center at Houston 외 2인

122개 metabolites 중에서 인과 네트워크를 추론하기 위해, 백만개가 넘는 유전학 자료로 구성된 데이터를 분석한 결과를 발표하였다.

(695) 정리: 그래프와 같은 통계적 시각화 툴의 활용

다양한 분야의 자료에 대하여 시각화 기법을 활용한 사례들을 발표하였다. JSM은 invited

session, topic-contributed session, contributed session들로 이루어지는데, 695번 세션과 같은 contributed session은 작은 연구들이 발표되는 경우가 많고, 동질성이 적은 발표들이 한 세션에 묶이는 경우가 많아 종합적인 관점 형성이나 자료 수집에 적합하지 않다. 특별한 이유가 없다면 향후 JSM참가자들은 invited 혹은 topic-contributed session 위주로 참석하는 것이 바람직하다고 판단된다.

❒ Survey Methodology Issues (총 3개 세션, 11개 주제 발표)

25	Employer list linking methods, implementation, and usage of probabilistic matches for enhancing workforce statistics

#1. Robustness of employer list linking to methodological variation 외 8인

미국 센서스국에서 개별 조사로 얻어진 고용주 정보를 고용주 행정 자료와 확률 연계하는 연구 결과를 발표하였다. 구체적으로는 ACS(American Community Survey)의 직업 정보와 LEHD(Longitudinal Emplyer Household Dynamics)에서 얻어진 행정 자료를 매칭시켰다. 매칭할 때, 기업명을 표준화하는 모형을 제안하고 비교한 것이 주요 연구 결과이며, 이 때 표준화된 자료의 패턴 파일, Jaro-Winkler string comparator, 혹은 python review tool layout 등을 이용하였다. 한편, 연구에서 10% ACS 표본 자료를 사용하였는데, ACS의 8%는 직접적인 id가 존재하지 않는다. 때문에 확률적 레코드 연계를 진행하고, 이를 위해서 Fellegi-Sunter 모형과 로지스틱 모형을 구현하였으며, 매칭 모형을 훈련하기 위해 참값 집합을 개발(정리)하기도 하였다. 여러 시도들의 결과를 분석한 결과를 보여주었다.

#2. Two Perspectives on Commuting and Workplace: A Microdata Comparison of Home to Work Flows Across Linked Survey and Administrative Files, Andrew Green, Cornell University/U.S. Census Bureau 외 2인

앞 발표와 같은 ACS-LEHD 표본을 매칭하여 만들때 블락 샘플(blocked sample)에 로짓 모형을 적합하고 이를 통한 매칭 확률을 이용한다. 인자 재구성을 통해 여러 매칭 결과를 비교한다. 매칭 자료를 이용해 LEHD의 LODES 및 ACS의 직장과 거리(Journey to work) 통계량 사이의 차이를 살펴보고, 매칭 자료가 unit-to-worker imputation을 향상시키는데 이용될 수 있음을 확인하였다.

#3. Developing Job Linkages for the Health and Retirement Study — Kristin McCue, U.S. Census Bureau 외 6인

매칭할 때 이용하는 block에 따른 매칭 비율을 검토한 결과를 발표하였다. 블락 전략은 정확 매칭을 배제할 수 있으므로 언제 적용할지를 검토할 필요가 있고, 소득 기록을 사용하여 매칭을 하면 좀 더 품질 좋은 블락을 만들 수 있다.

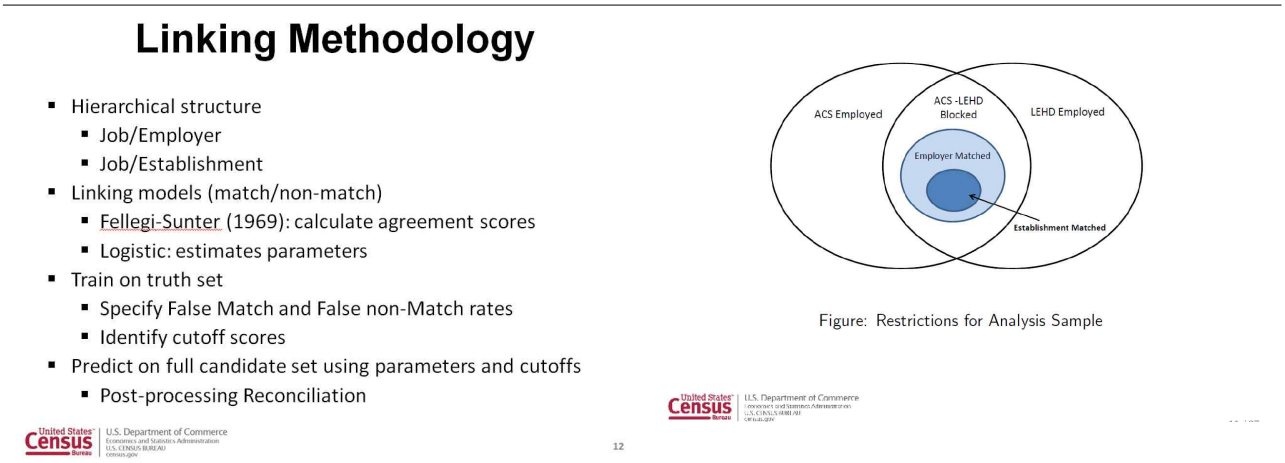
#4. Comparing Survey and Administrative Earnings: An Application of Employer List Linking, Lori Reeder, U.S. Census Bureau 외 1인

SIPP(Survey of income and program participation)을 행정 자료와 결합하여 직업 수준 자료(job-level dataset)을 만들고자 한다. 매칭 사례 연구로 세부 과정들과 비교 결과를 발표하였다.

(25) 정리: 미센서스국의 자료 연계 연구 현황

미국 센서스국에서 자료 연계 관련 연구 결과를 모아서 발표하였다. 같은 자료에 대하여 세 번의 발표가 있었고, 다른 자료의 연계 사례 발표 한 건이 있었다. 블락 샘플에서 매칭 모형을 구축하여 확률 연계하는 것에 관하여 복수의 발표가 이루어진다는 것은 그만큼 연구 자원이 많다는 것으로 생각된다. 통계청에서도 자료 연계 실무 적용이 가능할 수 있는 자료들을 이용해 최신 기법들을 적용, 비교, 분석하는 연구를 진행하면서 향후 실무 활용을 위한 밑거름을 다지는 것이 중요하다고 여겨진다.

미센서스국의 발표내용[3] 인용



주로 검토한 연계 모형 목록 | 블락 샘플

115	Tackling the Challenges of Missing Data in Surveys: Applying Methods and Assessing Uncertainty

#1. Estimating the Variance Due to Hot Deck Imputation for Product Value Estimates in the 2017 Economic Census, Katherine Jenny Thompson, U.S. Census Bureau 외 2인

3) 미센서스국에서는 JSM 발표 자료를 홈페이지에 게시한다.

먼저 경총(Economic Census) 자료를 설명하고, 모델 기반 대체와 핫덱 대체 과정을 비교적 상세히 보여준 후, 각각에 대한 분선 추정량을 설명하였다. 베이지안 부츠트랩도 검토하였다. 랜덤 핫덱과 NN(Nearest neighbor)핫덱, 모델 기반 대체와 핫덱의 결합 등에 대하여 MSE와 bias를 검토하고 각 방안을 비교하였다.

#2. Variance Estimation for Product Value Estimates in the 2017 Economic Census Under the Assumption of Complete Response, Matthew Thompson, U.S. Census Bureau 외 1인

무응답 대체를 다루었던 앞의 발표에 이어, 표본 자료와 전수 자료가 섞여 구성되어 있는 경총의 표본과 사후 층화를 다루었다. 분산 추정 방법으로 CHIP(Chipperfield-Preston), Mirror Match, BWO(Without replacement bootstrap) 및 FPBB(Finite population Bayesian Bootstrap)을 검토하였다. FPBB방법이 bias 등의 측면에서 좀 더 효율적인 것으로 판단된다. 향후 후속 연구 예정이다.

#3. Using Auxiliary Marginal Information to Deal with Nonignorable Missing Data, Mauricio Sadinle, Duke University/National Institute of Statistical Sciences 외 1인

결측치는 대부분 MAR(missing at random) 처리되고 있으나, ICIN(itemwise conditionally independent nonrespons)를 검토하였다. MAR, ICIN, PMM(pattern mixture model)에 관한 이론과 비교 시뮬레이션 결과를 보여주었다. ICIN이 MAR가전에 대한 대안이 될 수 있으며 결측 패턴과 함께 사용될 수도 있다.

#4. Semiparametric Fractional Imputation Using Empirical Likelihood in Survey Sampling, Sixia Chen, University of Oklahoma 외 1인

대체 기법은 NN, 핫덱, NP(nonparametric) 및 다중대체(multiple impuatation)이 있는데 최근 제안된 fractional imputation 기법을 확장한 내용을 발표하였다. 몇몇 모형에 대하여 시뮬레이션 결과를 비교하였다.

(115) 정리: 무응답 및 결측값 처리에 관한 최근 흐름

무응답이나 결측값을 처리하는 실무 사례로 미센서스국의 경총에 관한 두 발표를 통해 실무에서도 베이지안 부츠트랩을 이용하여 분산을 추정하는 실험이 이루어지고 있음을 확인할 수 있었다. 다음으로 흔히 사용되는 MAR을 대체할 수 있는 효율적인 측면이 많은 ICIN 기법은 향후 검토해볼 가치가 있다고 여겨진다. 가장 최신 이론으로 fractional impuation이 학술적 측면에서 연구 중임을 확인하였다.

398	Recent Developments in Survey Sampling: Session in Honor of Graham Kalton's 80th Birthday

#1. Practical Issues Related to Model Based Small Area Estimation, J.N.K. Rao, Carleton University

Graham은 조사 방법론 통계학 분야에 공로가 크며, 세부적으로는 분산 추정, sampling rare population, fractional imputation 등에서 업적이 많다. 모형 기반 소지역 추정, Beta 표본추출 모형을 제안하였고, 최근 빅데이터에서 GPS 자료를 이용한 소지역 추정을 연구하고 있다. 국가통계기관의 중요한 임무는 효율적이고 유용한 기술 통계량 생산에 있다고 보는 견해를 가지고 있다. 이러한 관점에서 관련 연구를 평생 지속해왔다.

#2. Recent Developments in Survey Design for Rare and Hard to Survey Populations, Steven G. Heeringa, University of Michigan

Graham의 공헌들 중, rare population survey 혹은 HTS(hard to survey population)에 관한 내용을 발표하였다. HTS 표본 설계의 전체적인 그림을 설명하고, 확률 표본 추출에 관한 내용 중 GIS와 storm data를 이용한 층화, multi-phase 설계에서 에러 요인 등을 설명하였다. 그 외에 간접 표본 추출 방식으로 location sampling, network sampling을 소개하였다.

#3. Recent Developments in Fractional Imputation, Wayne Fuller, Iowa State University

Graham의 공헌들 중, fractional imputation에 관한 내용을 수리적 내용과 함께 소개하였다. 핫덱 대체부터 시작하여 fractionally weighted 분산, 잭나이프 분산 추정, 모수적 fractional imputation 등을 차례로 설명하고 간단한 예제를 보여주었다.

(398) 정리: Graham Kalton의 조사방법론 분야에서의 공헌

이 세션은 Graham Kalton의 업적을 기리기 위해 조사방법론 분야의 대가들이 모여 그의 공헌과 의미를 분야별로 설명한 세션이다. 특히 rare population 영역의 방법론을 접할 수 있었고, 최신 fractional imputation에 관한 내용을 배울 수 있었다.

❒ Presentation

617	Topics in Statistical Methods and Applications

이 세션은 한인통계학회(KISS)에서 주최한 세션으로 한국인들이 발표를 하여 네트워크를 도모할 수 있다. 그러나 각 주제들이 서로 연관성이 없어 참석자가 적어 통계청 연구 결과

홍보의 효과가 미흡한 아쉬움이 있었다. 향후 JSM에서 발표는 topic-contributed session에서 할 수 있다면 바람직할 것이다.

가장 관련 있는 내용은 BLS 조문정 박사의 최신 EDA기법을 사용하여 특정 기법을 활용하기 힘든 자료 분석에서 효과를 볼 수 있다는 내용의 발표였으며, 통계개발원 조사연구실에서는 SGIS에서 인구 특성별로 사용자 선택에 따라 빈도표를 제공할 때 노출제어에 관한 연구 결과를 발표하였다. (제목: Statistical Disclosure Control for Korean SGIS outputs)

【참조1】 발표논문 초록

Title: Statistical Disclosure Control for Korean SGIS outputs

Statistics Korea has disseminated the Census data through Statistical Geographic Information System (SGIS) since 2014. Users can easily access the system on a web-site and obtain frequency tables for each Output Area using the interactive map. However, the Local Area variable, which include Output Areas, has been provided in the public use microdata file for Census. In order to deal with the disclosure by differencing between Local Areas and Output Areas, we need to measure the disclosure risk in a rigorous manner and find an adequate solution to mask SGIS outputs. In this talk, we present the result of our project, which hopefully will be the first step to make a move for statistical disclosure control in the Statistics Korea.

【참조2】 JSM2016 프로그램

○ 총 709개 정규 세션의 개최 일정(괄호는 세션 번호)

	7.31 일 (1~88)	8.1 월 (89~271)	8.2 화 (272~455)	8.3 수 (456~624)	8.4 목 (625~709)
8:30-10:20		(96-140)	(279-323)	(464-508)	(625-667)
10:30-12:20		(141-185)	(324-368)	(509-553)	(668-709)
12:30-1:50	점심시간				
2:00-3:50	(2-44)	(217-260)	(396-440)	(579-623)	
4:00-5:50	(45-87)				

○ 시간대별 주요 세션 정리

날짜	시간대별 주요 세션명
7.31 (일)	● (5) Survey costs and survey designs: trade-offs and advances ● (6) Open Source Statistical Software for Data Science ● **(25) Employer List Linking: Methods, Implementation, and Usage of Probabilistic Matches** ● (27) The First Self-Administered Survey in North Korea: A Glimpse of Self-Esteem of North Koreans Compared with Peers in 53 Other Countries ● (40) New approaches to Small Area/Domain Estimation
	● (46) Modeling, Analysis, and Inference from Survey using Bayesian Methods ● **(54) Recent Advances in Information Visualization** ● (57) The Extraordinary Impact of Janet Norwood on the Federal Statistical System and the Statistical Profession ● (67) The Consumer Expenditure Survey Redesign: Development, Concept Testing, and Evaluation ● (79) Ranking, Post Stratification and Calibration Methods ● (85) Administrative Records & Data Disclosure
8.1 (월)	● (101) Statistical inference with clustered data in survey sampling ● **(115) Tacking the Challenges of Missing Data in Surveys: Applying Methods and Assessing Uncertainty** ● (120) Improving Efficiency and Maintaining High Data Quality: Plans and Early Outcomes for the 2016 Survey of Consumer Finances ● (121) The policy landscape for statistics in the UK and the US ● (128) Statistical Consulting Applications ● (136) Survey modes, including web surveys, phone and multi-mode surveys
	● (141) Some new perspectives in statistical analysis sith incomplete data ● (149) Recent Advances and Challenges of Big Data Inference with Complex Structures ● **(152) Cyber Security in Support of National Defense and Global Security** ● (154) Adaptive Design in Large Scale Sample Survey ● (164) Innovative Uses of Linked Administrative and Survey Data ● (179) Combined data (surveys+administrative data,etc.)
	● **(218) Advances in Statistical Methods for Dissemination and Analysis of Official Statistics** ● (227) Statistical Foundations of Data Privacy ● (236) Multiple Imputation ● (246) Towards Better Communication of Information with Statistical Graphics ● (254) Advances in Small Area/Domain Estimation ● (258) Weighting ● (259) Nonparametric Methods for "Big Data"

날짜	시간대별 주요 세션명
8.2 (화)	● (279) Introductory Overview Lecture: Data Science ● (289) What to do with Messy Data? Four Case Studies ● (298) Census-Operational Design and Methods ● (304) A Roadmap for Promoting Statistical Collaboration ● (319) Statistical Methods for Complex Survey Data
	● (332) Quality of Alternative Sources for Social Economic and Health Data ● **(337) Fusion Learning and Combining Inference from Diverse Complex Data Sources** ● (342) Novel Missing Data Imputation Methods ● (363) Nonprobability/Web Sampling and Data Analysis ● (370) Contributed Poster Presentations: Government Statistics Section
	● **(398) Recent Developments in Survey Sampling-Session in Honor of Graham Kalton's 80th Birthday** ● (398) Estimation and inference for massive data sets ● (405) Recent Advances in High-dimensional Statistics and Computational Methods ● (406) Interactive Visualizations and Web Applications for Analytics ● (408) Bridging BFF (Bayesian/frequentist/fiducial) inferences in the era in the data science ● (415) Data Challenge 2016 Ⅱ ● (436) Adaptive/Innovative Survey Design and Survey Cost ● (437) Inflation, Price Indexes and Labor Statistics ● (440) Missing Data, Imputation, & Calibration
8.3 (수)	● **(468) Uncertainty estimation for Massive Data Sets** ● (470) Design and analysis issues with modern population telephone surveys ● (502) Innovative Statistical Method for Complex Survey Data
	● (520) Machine Learning in Econometrics ● (521) Reflection on Social Surveys' Past and Future ● **(527) Innovations in Disclosure Avoidance at the U.S. Census Bureau** ● (540) Statistical Computing for Machine Learning ● (549) Combined data and data linkage

날짜	시간대별 주요 세션명
8.3 (수)	● (579) Challenges and Opportunities for Analysis of High-Dimensional and Big Data ● (587) Resampling Methods for High-Dimensional Inferences ● (595) Update on Current Population Survey Research ● (600) When the plot is Not the End - Advances in computing and reasoning on data visualization ● (601) Expanding Capacity for the Measurement of Sexual Orientation and Gender Identity in Federal Surveys ● (602) Using Demography and Randomized Response Model to Improve Social Science Research ● (613) Non-response adjustment and nonresponse bias reduction methods ● **(617) Topics in Statistical Methods and Applications** (발표 세션, KISS 주관) ● (619) Sparsity in Record Linkage, Networks, and Privacy: Applications to Official Statistics, Author Disambiguation Data, and the Syrian Conflict ● (620) Sampling & Survey Methods
8.4 (목)	● **(634) Analysis, Storage, and Privacy for Big Data** ● (635) Combining data from multiple sources: Examples from health policy ● (640) Topics in adaptive/responsive survey desgins ● (651) Advances in Sampling and Estimation ● (661) Issues in Estimating and Adjusting for Sampling and Nonsampling Errors ● (666) Sampling & Survey Methods
	● (683) Testing for Data Quality ● (684) Collecting and Analyzing Sensitive Data: Making Lies Naked! ● **(695) Methods and Applications of Statistical Graphics** ● (705) Methods for item missing and unit nonresponse ● (708) Nonresponse & Propensity Scores