캐나다 국제 심포지엄 참가 결과 보고

Statistics Canada's 23rd International Symposium
November 1 to 3, 2006, Ottawa

2006. 11.

통계개발원
경제통계실

<차 례>

Ⅰ. 출장개요

1. 출 장 자 : 통계개발원 경제통계실 정선경, 이은정

2. 출장기간 : 2006. 10. 30. ~ 11. 5. (7일간)

3. 출 장 지 : 캐나다 오타와

4. 출장목적 :

○ 제23차 조사방법론 국제 심포지엄 참가

- 우리 청에서 관심이 높은 분야에 대한 선진기법(행정자료 이용, 이상값 처리, 매스킹, 소지역추정 등)에 대한 선진연구방법 습득

Ⅱ. 제23차 조사방법론 국제 심포지엄

1. 회의개요

ㅇ 명 칭 : Statistics Canada's 23rd International Symposium on Methodological Issues(제23차 조사방법론 국제 심포지엄)

- 부 제 : Methodological Issues in Measuring Population Health (인구보건통계에 관한 조사방법론)

ㅇ 주 관 : Statistics Canada

ㅇ 기 간 : 2006. 11. 1. ~ 11. 3. (3일간)

ㅇ 장 소 : Chateau Cartier Resort, Gatineau, Quebec, Canada

ㅇ 참가자 : 12개국 1국제기구(WHO) 약 400명 참가

- 주로 캐나다 통계청 및 논문 발표자 중심

2. 회의주제 및 일정

□ 2006. 11. 1.(수)

1) 인구보건통계에서의 자료 연계(Record linkage)

○ 자료 매칭의 목적
○ 자료 연계가 이용되는 분야
○ 자료 연계의 과정
○ 보건통계 자료를 이용한 자료연계 실습

2) 보건통계의 종단자료에 관한 분석 방법

○ 종단조사 자료를 위한 모델 개요
  - growth curve, GEE, stochastic process models
○ 관찰 연구 및 causality에 관한 토론
○ 모델 적용 원리 및 실제 조사자료에 적용 방법
○ 소프트웨어 검토

□ 2006. 11. 2.(목)

1) 보건통계의 소지역 추정

○ 캐나다 보건통계조사에서의 소지역 추정에 관한 선형·비선형 혼합 모델
○ 소지역 추정에서의 영향력 있는 관찰값 및 이상값 처리
○ 소지역에서의 제한된 인구 변위에 대한 Robust Bayesian 이론
○ 소지역 평균에서 사용되는 MSE의 Robust 추정

2) 보건통계의 종단자료 분석

○ Life History 자료의 추정 에러
○ 인구보건의 종단자료 분석에 관한 방법
○ 보건통계의 종단자료 분석에 준한 복합 표본 설계 개요
○ 단순 및 복합 표본의 표본오차 및 사망자수 추정

3) 상이한 출처로부터의 자료연계(Record linkage)

○ 인구센서스 자료 이용에 관한 질적인 연구
○ 센서스 데이터의 지역자료를 표본조사 및 행정자료로 연계
○ 정밀하고 가능성 있는 자료연계

4) 가중법, 추정 및 이상값 발견에 관한 논의

○ Weight Trimming에 관한 변수 선정 모델
○ 미국의 직업 질병 조사에서의 이상값 발견 및 처리 방법

□ 2006. 11. 3.(금)

1) 보건통계자료 분석방법

○ 복합 표본조사 분석에 관한 Bootstrap 방법
○ 표본조사의 회귀계수 추정
○ 보건 행정자료를 이용한 관찰 연구

2) 표본론 : 이론과 적용

○ 캐나다 보건추정조사(Canadian Health Measures Survey)의 표본설계
○ 국가 보건 및 영양실태조사(National Health and Nutrition Examination Survey)의 표본설계

3) 보건통계 조사 기획 및 수행

○ 국제 질병 연구의 개념 및 분석
○ 세계 보건 조사로부터의 경험 : 보건 자료 추정에서의 국제적인 비교

3. 차기대회 일정

○ 2007년 심포지엄은 미국 ASA와 Third International Conference on Establishment Surveys(ICES-Ⅲ, 제3차 사업체통계조사 국제회의)을 공

동 개최하는 것으로 대체

○ 일　　시 : 2007. 6. 18.~21.

○ 장　　소 : Hyatt Regency Montreal-Montreal, Quebec, Canada

※ 2006년 12월 18일부터 참가 등록 가능

Ⅲ. 주제발표 및 회의내용

(http://www.statcan.ca/english/conferences/symposium2006/pdf/finalprogram.pdf)

1. 인구보건통계에서의 자료 연계

Record Linkage in Studies of Population Health - An Overview

By Karla Fox, Department of National Defence, Partricia Whitridge, Elections Canada

□ 매칭(Matching)의 종류

○ 통계적 매칭(Statistical Matching)

○ 정밀한 매칭(Exact Matching)

□ 통계적 매칭(Statistical Matching)

○ 특징

- 같은(same) 단위라기 보다는 비슷한(similar) 단위의 자료 연계가 기대됨
- 그렇다고 같은 단위의 연계를 배제하지는 않음
- 동일한 자료 보다는 비슷한 성격에 준하여 만듬
- 결합분포를 생성함

○ 방법

- Monte Carlo 방법
- Gibbs 샘플링
- 무응답처리

○ 에러 발생

- 에러는 결합분포가 정확한 조건으로 지정되지 않았을때 발생함
- 예를 들어, 가구 타입에 의한 가구소득 분포와 독신 분포는 독신남만 보고자 할때는 분포가 잘못 결합된 경우임

□ 정밀한 매칭(Exact Matching)

○ 특징

- 정밀함(Exact) : 완벽하게 매치됨
- 계층적(hierarchical)으로 정확
  - · 서로 다른 매칭 층을 주고 다중 패스를 사용
  - · 가장 규칙이 엄격한 것부터 시작

예) Pass 1 : First name, Middle name, Last name, Street Number and Name, PC, Sex
Pass 2 : First name, Last name, Street Number and Name, PC, Sex
Pass 3 : First initial, Last name, Street Number, PC, Sex
Pass 4 : Last name, Street Number, FSA, Sex

○ 사용 소프트웨어

- Netrics 소프트웨어
  - · 이분그래프로 이론적 접근
  - · 자료가 속해 있는 데이터베이스에서 동일한 것을 찾아냄

○ 에러 발생

- 에러는 개인이 그 정보에 매치가 안된 경우 발생(부정확한 매치 또는 불일치)
- 예를 들어, 58세 John Smith를 사망자 record에 연계할 때 살아있는 John Smith를 사망자 레코드에 연계함(부정확한 매치)
  또한, 58세 J. Smith를 사망자 레코드에 연계할 때 58세 남자 이름 J. Smith를 파일에서 찾을 수 없음(missed)

□ 자료연계에 관한 이론과 실제

○ 데이터의 질 : 모든 데이터는 지저분하다!!

- 데이터의 출처가 어디인지?
- 데이터는 어떻게 수집되었는지?
- 파일은 이미 자료처리가 되었는지?
  그렇다면, 누구에 의해? 무슨 이유로?
- 식별자(데이터 항목에 이름을 부여하여 일시적으로 규정하거나, 그 데이터의 어떤 특성을 표시하기 위해서 사용하는 기호)는 있는지?
- 식별자들 가운데 불일치는 없는지?

○ 캐나다 통계청의 자료연계

- 모든 연구는 규정된 review 과정을 거쳐야 함
- 자료연계의 목적은 통계적이고 분석적이어야 함
- 통계법을 준수하여야 함
- 명백한 비용절감과 응답자 부담 경감, 실행가능하여야 함
- 대중의 관심이 있어야 함

□ 자료연계(Record linkage) 과정(Step)

○ 식별자의 결정

- 자료연계 수행의 효율성은 아래와 같은 기준을 만족시키기 위하여 얼마나 잘 항목들을 추출하는지에 달려있음
  · 불변, 보편, 합리, 경제
  · 단순, 이용가능, 잘 알려진, 정확성, 독특함

○ 데이터 Cleaning

- 아래와 같이 의문을 가짐
  · 식별자가 재 사용되나?
  · 개개인의 성이 변하고 움직이면 어떻게 되나?
  · 가구 또는 가계를 위한 식별자인가?
  · 가구구조가 변하면 어떻게 되나?

- 어떤 파일도 100% 정확하지 않음
  · 논리적으로 타당한 값, 범위, 코드 등을 체크
  · 특수 문자(‘%*)^@&$#_-)가 있는지 체크

○ 접근방법에 대한 결정

- 정밀한(Exact) 매칭으로 할 것인지 통계적(Statistical) 매칭으로 할 것인지 결정
  · 각각의 파일에 같은 개개인이 있는지?
  · 인구가 비슷한지?
  · 파일을 링크할 식별자가 충분한지?

○ 파일 연계

○ 에러의 발견

- 두 종류의 에러(잘못된 쌍, 불일치된 쌍)
- 잘못된 링크(False Links)와 불일치 링크(Missed Links)는 각각 Manually reviewing a sample (Barlett, 1993)와 모델링(Armstrong and Mayda, 1993)에 의해 추정함

○ 분석

- 연계 비율의 효과를 문서화 함

□ 사용 소프트웨어

○ LINKS : 레코드 linkage 패키지
○ GRLS : Generalized Record Linkage System
○ Link King : 레코드 연계 및 통합 소프트웨어
○ Netrics : 지능적인 레코트 매칭 소프트웨어
○ Identity Search Server
○ SureMatch

2. 캐나다 보건추정조사의 표본설계

Sample Design of the Canadian Health Measures Survey(CHMS)
By Suzelle Giroux, Statistics Canada

□ CHMS 조사의 목적

○ 신체측정, 구강검진, 심관혈관상태, 근골격상태, 혈액검사, 소변검사 등 신체의 측정에 있어서 캐나다인의 건강과 관련된 정보를 획득하기 위함

○ 영양, 흡연습관, 음주여부, 신체상태, 과거병력, 인구·사회학적인 변수 등과 같은 캐나다인의 건강과 관련된 중요 정보 획득을 위함

○ 건강과 관련된 문제점을 좀 더 잘 평가하기 위하여

□ 목표 모집단

○ 개인 집에서 살고 있는 6세부터 79세까지의 개인

○ 아래와 같은 경우는 제외됨
- 인디언 특별구역과 왕실 소유지에 살고 있는 사람
- 캐나다 군대의 정규멤버
- 시설 거주자
- 오지

□ CHMS의 특징

○ 자료수집 방법 단계

① CAPI를 이용하여 응답자의 거주지를 방문하여 조사
- 신체 측정 결과를 확인할 수 있음

② 응답자가 CHMS 크리닉을 방문
- 신체 측정, 혈액, 소변 검사를 위하여 방문

- 일주일 동안 가속도계(accelerometer)를 달고 있음
- 여행경비를 부담
- 결과는 응답자에게 보고서 형식으로 제공됨
- 크리닉 장소(site)는 응답자가 여행 가능한 거리에 설치
- 크리닉 수용인원은 6주동안 350명의 응답자를 검사
- 크리닉은 한 곳(site)에서 다른 곳으로 이동

○ 자료수집은 2년이상 걸림

(※ CHMS mobile clinic)

○ 전수추정을 위해 성별 5개의 age 그룹이 필요함
(6~11, 12~19, 20~39, 40~59, 60~79)

○ 표본수는 5,000명의 응답자가 동일하게 10개 age-sex 그룹으로 분포
- 한 위치마다 350명의 응답자를 조사하므로 15곳(site)이 방문됨

□ 표본틀 선택

○ 수집장소(site)는 거리와 인구밀도에 따라 결정

- 한 장소마다 최소 10,000명
- 도시 : 센터부터 장소(site)까지 최대 50킬로
  시골 : 센터부터 장소(site)까지 최대 100킬로

○ 도(province) 경계와 Census Metropolitan Areas 경계내에 있어야 함

○ 장소(site) 설치할 때 노동력조사(LFS)의 지역 프레임을 이용함

□ 장소(site) 추출

지역	6-79세 인구수 (2001 Census)	Proportional Allocation	모집단 site수	방문할 site수
Atlantic	2,061,425	382	36	1
Quebec	6,560,375	1,217	50	4
Ontario	10,248,545	1,901	61	6
Prairies	4,538,970	842	77	2
B.-C.	3,540,000	657	33	2
합계	**26,949,315**	**5,000**	**257**	**15**

□ person 추출

○ 각 사이트 내 거주지 추출
○ 인터뷰 중 가구원 목록을 작성한 후 지원자 중 랜덤하게 person 추출

□ 향후 과제

○ person에 대한 표본설계 마무리
○ 2006년 가을 기술적인 테스트후 2007년 1월 총연습
○ 2007년 2월말 조사 시작

3. 캐나다 지역보건조사의 사이클 통합

Combining Cycles of the Canadian Community Health Survey(CCHS)

- 서로 다른 조사에서 같은 정보를 수집하고 있을 때 원자료를 통합함으로써 더 유용한 정보를 얻을 수 있음을 CCHS를 예제로 하여 설명
- 캐나다의 CCHS의 경우 지역단위로 130명의 응답자를 대상으로 2년 마다 실시
- 예를 들어, 임신 경험과 같은 드문 특성치에 대한 분석을 할 경우 1 cycle 내의 표본 규모는 충분하지 못하지만 다른 cycle 의 표본을 묶어서 한 표본으로 만들면 충분한 표본 규모를 확보하는 것이 가능함

4. NSDUH 자료에 대한 공개용 마이크로데이터 파일(PUFs) 작성

Creation of Public Use Micro-Data Files for the National Survey on Drug Use and Health (NSDUH)

- NSDUH 자료에 포함되어 있는 민감한 정보에 대한 누출 위험에서 응답자를 보호하기 위해 자료제공 전에는 반드시 누출방지에 대한 규정이 필요하며 정보손실과 누출위험이 동시에 통제되어야 함
- MASSC은 식별변수를 묶는 Micro Agglomeration, 식별변수 값을 대체하는 Substitution, 표본을 줄이는 Subsampling, 대체와 표본감소에 따른 가중치 측정하는 Calibration의 4단계로 구성
- MASSC 처리 된 데이터에 대해서는 추정치 및 표준오차를 측정하고 이를 대조 등 분석을 해야 함

Ⅳ. 참가소감

□ 통계환경의 변화 속에서 국제적으로 새롭게 이슈화 되는 사안들을 알아보고 그것에 대하여 새롭게 인식하는 계기가 되었음

○ 이번 심포지엄 주제가 보건통계이니 만큼 요즘 세계적으로 이슈화되고 있는 보건통계에 대하여 관심을 가지게 되었고, 각 기관에서 생산된 보건통계 및 보건통계생산과 관련된 자료들을 서로 공유할 수 있도록 제도화 해야 할 필요성을 느낌

○ 여러 기관이 개별적으로 보건통계를 생산하고 있으므로 국가통계기관인 통계청에서 보건통계의 품질을 관리할 수 있는 방안에 관심을 기울여야 함을 느낌

○ 국내에서 보건통계를 실제 작성하는 국가통계기관, 병원 및 연구소들이 참여하여 보건통계 조사방법론에 대한 현안을 파악하고 이를 적극 반영하는 계기가 되면 좋을 것으로 생각 됨

□ 급변하는 환경에 부응할 수 있는 새로운 통계기법들을 배워보는 계기가 되었으나 대부분의 주제가 다소 이론적이어서 이해에 어려움 있었음

○ 우리나라의 경우 미국의 NCHS나 캐나다의 CCHS와 같은 국가차원의 보건통계조사가 이루어지지 못 하고 있다고 느낌

□ OECD나 IMF와 같은 국제기구가 아닌 一國의 통계청에서 심포지엄을 제23차례나 개최하고 있다는 사실에 다소 경외감을 갖게 되었으며 우리 청에서도 이런 회의(심포지엄 또는 워크샵)을 정례적으로 개최한다면 우리 청의 위상을 국내외에 드높일 수 있고 소속 직원들의 역량 강화도 독려할 수 있을 것으로 생각 됨

○ 심포지엄 참가도 단순 참여보다는 우리도 통계기법 연구 및 적용 사례를 발표하여 한국의 통계적 역량을 적극적으로 알릴 필요가 있을 것으로 판단 됨

Ⅴ. 기타 사항

□ 캐나다 통계청의 이 심포지엄은 국가통계기관에서 주최하는 연차회의 중의 대표 격으로 소개되고 있음

○ 1984년 이래 캐나다 통계청 주관으로 매년 주제를 바꿔 가면서 개최되고 있으며 주로 미국 등 북미국가들의 참여가 많은 편

○ 참가자는 주로 통계작성기관, 연구소, 대학교수 및 학생들로 구성되어 있으며 40여 편의 논문이 발표 됨

○ 등록비(500$CAN)가 징구되고 심포지엄 자료는 1년 후에 CD 포멧으로 참가자에게 우편으로 우송되며 간행물은 유료로 판매됨