2016년 총 조사 오차 국제워크숍 참가 결과 보고

2016. 11.

조사관리국
표 본 과

목 차

Ⅰ 출장 개요

◇ **참가 회의 : 2016년 총 조사 오차 국제워크숍**
(2016 International Total Survey Error Workshop; ITSEW)
◇ **개최 지역 : 호주(시드니), 호주 통계청 주관**
◇ **여행 기간 : 2016년 10월 8일 ~ 2016년 10월 13일(4박 6일)**
◇ **참가자 : 표본과 한혜은 주무관**
* ITSEW는 2005년 워싱턴에서 개최한 이례 2008년부터 매년 개최하고 있는 국제 워크숍으로 조사자료를 이용하여 통계를 작성하는 과정에서 발생하는 모든 오차를 주제로 연구이론과 실무기법 소개(WESTAT, RTI, NISS 후원)

□ 참가 목적

○ 조사자료를 이용하여 통계를 작성하는 과정에서 발생하는 모든 오차를 주제로 각국 전문가들의 논문 발표 및 토론을 진행하는 국제워크숍에 참가하여 최근의 동향을 파악하고 국내 통계작성에 활용

□ 수행 내용

○ 워크숍 기간 동안 주제별 발표 세션에 참가하여 **총 조사오차 관련 최신의 연구 방법과 각국의 현황 파악**

- Paul Biemer의 「Can TSE Save Survey Science in a Big Data World?」를 비롯하여, 총 조사오차의 구조, 조사표 설계와 측정오차, 조사과정자료를 활용한 오차 축소 방법, 혼합조사에서 총 조사오차를 최소화하는 방법 등의 주제로 참가국의 연구결과 공유 및 향후 연구방향 논의

□ 업무 활용 방향 및 향후 계획

○ 무응답을 고려한 표본설계 및 추정방법 개선에 활용
○ 조사과정자료를 활용한 총 조사오차 최소화 방안 연구
○ 행정자료 활용, 데이터 연계, 조사자료 통합 등 자료를 통합하여 통계를 작성하는 방안 연구
○ 표본연구회를 통해 청내 업무 관련자에게 지식 공유

Ⅱ 회의 개요

1. 회의 기간 : 2016년 10월 9일 ~ 10월 12일

2. 주관 기관 : 호주 통계청(ABS)

○ 총 조사오차 국제워크숍(International Total Survey Error Workshop; ITSEW)은 2005년 워싱턴에서 개최한 이례 2008년부터 매년 개최하고 있는 국제워크숍으로 조사자료를 이용하여 통계를 작성하는 과정에서 발생하는 모든 오차를 주제로 연구이론과 실무기법 소개(WESTAT, RTI, NISS 후원)

3. 회의 내용

□ 개최 목적

○ 조사자료를 이용하여 통계를 작성하는 과정에서 발생하는 모든 오차를 주제로 참가국의 통계청 및 통계작성기관의 실무자들이 **연구 결과를 발표**하고 향후 연구 방향에 대한 **자문을 구하고 경험 공유**

□ 주요 연구

○ 기조 연설 : Can TSE Save Survey Science in a Big Data World?
(Paul Biemer, RTI International and University of North Carolina)

- 조사연구를 지속·발전시킬 방안을 찾기 위해 빅데이터를 활용한 통계작성 과정과 조사통계 작성과정을 총 조사오차(TSE Framework) 관점에서 비교

○ 일반 세션

- 행정자료를 활용한 통계 작성 과정의 품질관리 체계(뉴질랜드, SNZ)
- 조사과정자료와 지리정보를 활용하여 조사원의 업무 경로 결정(미국, Westat)
- 적응 표본설계에서의 품질 관리(호주, ABS)
- 전국예방접종률조사(NIS)의 총 조사오차의 분포 연구(미국, NORC)
- 표본설계, 조사표 개선 등이 오차에 미치는 영향 시계열 분석(호주, ABS)

4. 참가자 현황 : 13개 국가, 약 50명 참가

○ 참가국 : 호주, 미국, 캐나다, 뉴질랜드 등/ 한국

○ 기관 : 통계청, 통계작성기관, 대학, 연구소

- 호주(Australian Bureau of Statistics), 미국(Westat, RTI International, University of Michigan, NORC at the Univercity of Chicago), 캐나다(Statistics Canada), 뉴질랜드(Statistics New Zealand), 독일(IAB), 핀란드(Statistics Finland), 싱가포르(Statistics Singapore), 피지(Fiji Bureau of Statistics) 등

Ⅲ 향후 계획

1. 활용 방향

○ 행정자료 활용, 데이터 연계, 조사자료 통합 등 확장된 통계작성방안 연구

- 자료 통합 및 활용, 통합된 자료의 품질평가 방안 연구

○ 조사과정자료(paradata)를 활용한 품질 제고 방안 마련

- 조사과정자료를 활용한 현장조사 업무량 최적화 방안 연구
- 적응표본설계 등 무응답을 고려한 표본설계 방법 연구

○ 국가통계의 신뢰도 제고를 위한 자료 분석

- 조사표 개선, 표본개편 등과 관련한 통계의 시계열 안정성 검토

2. 향후 계획

○ 매년 워크숍에 참가하여 국제적 연구협력체계를 구축하고 통계품질 제고를 위한 표본설계 및 추정방안 개발에 활용('17년, 독일 개최 예정)

부록 주요 발표내용(요약)

☐ Can TSE Save Survey Science in a Big Data World?

(Paul Biemer, RTI International and University of North Carolina)

○ **개요**

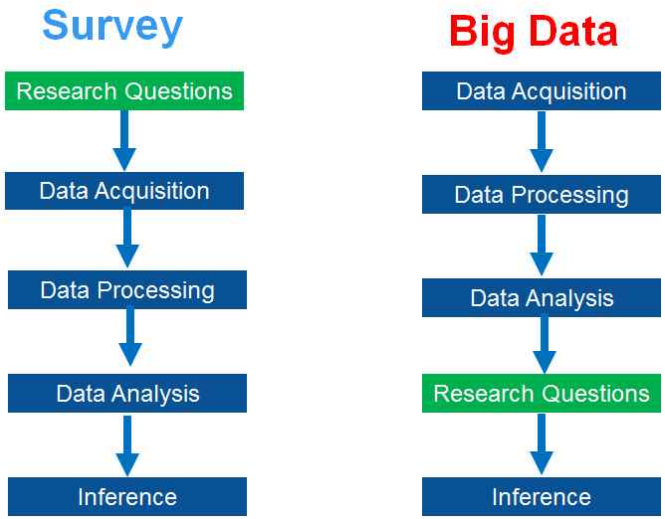
- 조사연구를 지속·발전시킬 방안을 찾기 위해 빅데이터를 활용한 통계작성 과정과 조사통계 작성과정을 총 조사오차(TSE Framework) 관점에서 비교

○ **빅데이터의 유형**

- 소셜미디어에서 생산되는 정보
- 웹 스크리닝을 통해 수집되는 자료(예, trivago)
- 교통카메라, CCTV 등 비디오 자료
- 소매업의 상품구매 이력, 신용카드 사용 자료
- 인터넷 포털 사이트의 검색 자료
- 정부에서 생산하는 자료

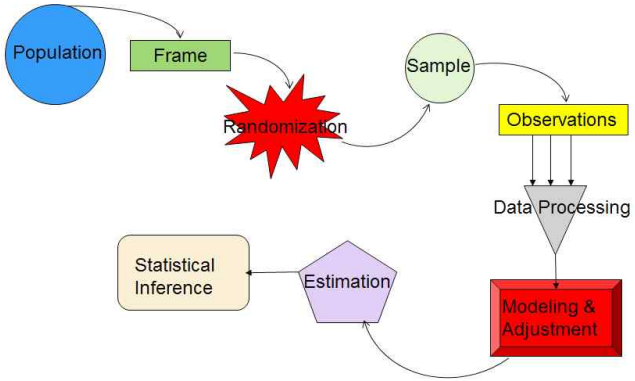
○ **조사와 빅데이터를 활용한 통계의 추론과정 비교**

- 조사통계는 연구주제(조사목적)에 맞게 자료수집 계획을 수립하고 통계를 작성하는 반면, 빅데이터를 활용한 통계는 자료분석을 통해 문제 도출
- 빅데이터는 자료 수집 전에 연구주제, 분석 계획을 수립하기 어려움



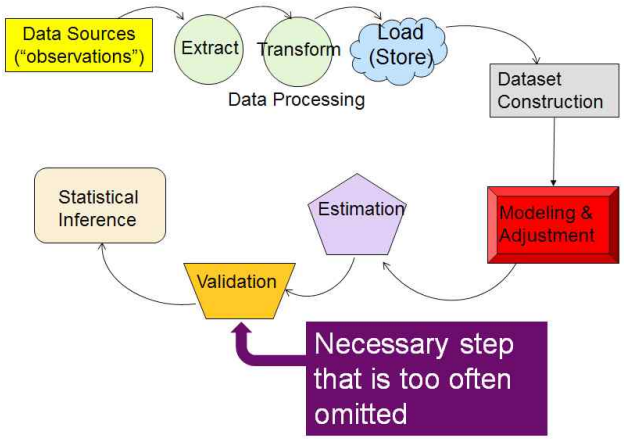
○ **조사통계의 추론 과정**

- 모집단을 대표하는 표본추출틀을 구축하고 확률추출방법으로 표본 선정
- 자료 수집 및 처리, 모형화·보정, 추정량 산출의 과정을 거쳐 추론
- 모형화·보정은 조사통계의 가중값 작성 과정에 해당



○ **빅데이터를 활용한 통계의 추론 과정**

- 각각의 원천으로부터 수집된 자료를 분석 목적에 맞게 추출, 변환, 저장하는 과정을 통해 데이터 구축
- 모형화·보정, 추정량 산출, 검증의 과정을 거쳐 추론
- 많은 경우, 검증 과정을 생략하지만, 빅데이터는 통계분석을 목적으로 수집된 자료가 아니기 때문에 반드시 필요



○ **오차 및 주요 특성별 비교**

- **구성 오차(Specification Error)** : 조사통계는 연구 목적에 맞게 주요 개념을 정의하기 때문에 오차가 작고, 빅데이터는 연구자가 주요 개념을 정의하고 자료를 정제하는 과정 필요
- **포함 오차(Coverage Error)** : 표본추출틀의 포함률은 일반적인 조사에서 높은 수준이며 미포함 오차는 추정을 통해 보정 가능, 단 희소한 집단에 대한 포함률이 낮기 때문에 이런 경우 빅데이터 활용이 도움
- **표본 오차(Sampling Error)** : 조사통계는 추출확률을 알 수 있고 오차의 한계 추정 가능, 빅데이터는 자료의 수가 매우 크기 때문에($n \approx N$) 전수조사 수준의 정도를 확보할 수 있으나 확률 구조를 알기 어렵고 모집단의 관심 단위와 연계 불가
- **측정 오차(Measurement Error)** : 조사통계는 조사표 설계 측정오차를 최소화할 수 있으며, 컴퓨터 기반 조사의 경우 실시간 내검 가능, 빅데이터는 자료의 크기가 방대하기 때문에 측정오차가 전체 오차에 미치는 영향은 크지 않다고 볼 수 있지만, 자료의 원천에 따라 체계적인 오차가 있을 수 있음
- **무응답 오차(Nonresponse Error)** : 조사통계는 응답성향과 관련된 무응답 메카니즘에 대한 모형화가 가능하고 적합도가 높지만 무응답률이 지속적으로 증가하는 추세, 빅데이터는 무응답률, 무응답 메카니즘을 파악할 수 없어 무응답 오차 측정 불가
- **모형/추정 오차(Modeling/Estimation Error)** : 조사통계는 보조정보를 이용하여 무응답 조정, 사후보정이 가능하지만, 보조정보가 결측이거나 품질이 낮은 경우 영향을 받음, 빅데이터의 경우에도 외부 자료를 이용하여 결측자료을 보완하거나 포함 오차를 보정할 수 있지만 분석단위를 외부 자료와 연계할 수 없다는 한계
- **비용(Costs)** : 우편, 웹, 전자조사와 같은 자계식 조사의 경우 면접조사 보다 비용이 적지만 기본적으로 조사통계는 자료수집에 많은 비용 소요되지만, 빅데이터는 자료 수집 비용이 적지만 자료처리 및 데이터 구축, 자료 분석 등에 추가적인 비용 소요 가능
- **시의성(Timeliness)** : 조사통계는 조사기획, 수행, 자료처리 및 분석에 소요되는 시간을 예측할 수 있지만 많은 시간(수개월 또는 수년)이 소요되는 반면 빅데이터는 실시간으로 자료를 수집할 수 있음, 단 자료수집체계가 규칙·규격화(routinize)되지 않은 경우 오히려 상당히 많은 시간이 소요될 수 있음
- **접근성·명확성(Data Accessibility/Clarity)** : 조사통계는 약간의 컴퓨터 활용능력이 필요하며, 자료에 대한 설명이 문서화되어 있음, 공개된 자료는 익명으로 활용할 수도 있지만 특정 자료의 경우 정보보호의 차원에서 접근이 제한되는 경우가 있음, 빅데이터는 실시간으로 자료를 활용할 수 있으나 분석을 위해 높은 수준의 컴퓨터 활용능력이 요구되며 자료의 구조와 자료 수집과정을 이해하기 어렵고, 자료의 소유자만 접근할 수 있음
- **추론(Inferential Power)** : 조사통계 자료는 다양한 변수에 대해 종단, 계층적 구조를 갖고 있으나 복합설계의 경우 분석이 어렵고 결측값에 대한 추정 필요, 빅데이터는 시·공간적 범위가 넓고 희소한 집단에 대한 추정이 가능하지만 종단분석, 연관성 분석을 위한 데이터 구축이 어려움

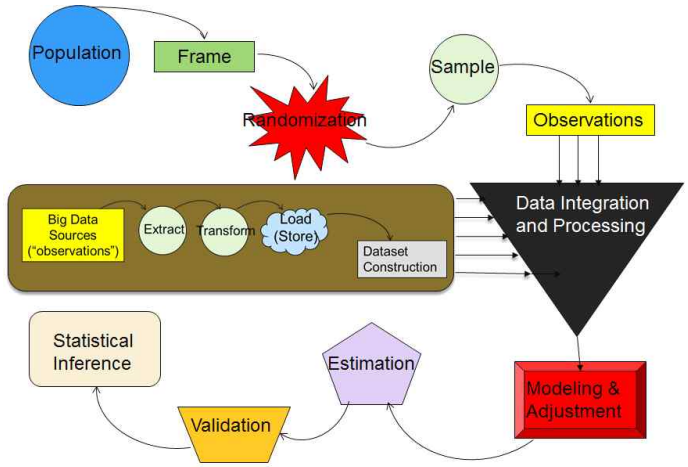
○ **오차 및 주요 특성별 점수**

- 10가지 주요 차원에 대해 각각의 통계작성방법을 비교한 결과 전체적으로 조사통계가 우수하지만 표본오차, 비용, 시의성 측면에서 빅데이터 활용 통계가 우위

Criterion	Survey Science	Big Data Science
Specification Error	9	6
Coverage Error	7	3
Sampling Error	8	8
Measurement Error	7	5
Nonresponse Error	4	3
Model/Estimation Error	7	6
Costs	4	8
Timeliness	4	8
Accessibility/Clarity	8	4
Inferential Power	7	5
Total	**65**	**56**

○ **빅데이터를 활용한 조사통계 작성 체계(Unified Data Paradigm)**

- 자료처리 과정에서 빅데이터를 통해 구축한 데이터와 통합하여 분석하고,
- 빅데이터를 활용한 통계작성과정과 마찬가지로 추정값에 대한 검증 필요
- 자료 연계 시 발생하는 문제(예, 불일치 자료의 처리방법 등)에 대한 방안 마련 필요



○ **통합된 통계작성 체계 평가**

- 조사통계와 빅데이터를 통합한 통계작성 체계는 대부분의 차원에서 각 자료의 긍정적 요소가 반영되지만 자료를 통합하는 과정에 높은 수준의 컴퓨터 기술과 방대한 분석이 요구되기 때문에 접근성 차원의 점수는 낮아짐

Criterion	Survey Science	Big Data Science	Unified Data Paradigm
Specification Error	9	6	**9**
Coverage Error	7	3	**7**
Sampling Error	8	8	**9**
Measurement Error	7	5	**6**
Nonresponse Error	4	3	**8**
Model/Estimation Error	7	6	**8**
Costs	4	8	**7**
Timeliness	4	8	**7**
Accessibility/Clarity	8	4	**5**
Inferential Power	7	5	**8**
Total	**65**	**56**	**74**

○ **통합된 통계작성 체계의 효용**

- 새로운 통계작성 체계는 다음의 상황에 따라 활용 가능성과 효용성이 결정됨
- 연구자가 빅데이터 활용에 따른 복잡성 증가와 고난도의 업무를 대처할 수 있는가
- 빅데이터에 접근하고 활용하는 통계과정에서 발생하는 장애 요소가 해소되는 않은 기존의 환경 하에서 새로운 체계가 실현 가능한가
- 기밀유지, 정보보호 등의 문제가 빅데이터 활용에 제약을 줄 것인가
- 빅데이터의 품질이 고품질의 조사데이터와 통합함으로써 개선될 것인가

□ A quality framework for statistical design and outputs using administrative data

(Felipa Zabala, Statistics New Zealand)

○ 개요

- 행정자료를 활용한 통계 작성 시 필요한 품질관리 체계 소개
- 통계작성을 위해 활용되는 행정자료는 조사자료와 마찬가지로 오차의 관점에서 품질을 평가해야 하지만, 행정자료가 생성된 본래의 목적에 맞게 평가되어야 함
- 뉴질랜드 통계청에서는 품질 보고서 작성 지침을 제정하고 행정자료와 행정자료를 통합하여 생성한 자료의 품질을 측정하는 과정에서 고려해야 하는 요소 제시
  「Guide to reporting on administrative data quality」

○ 행정자료 품질관리 지침 제정 목적

- 통계작성에 활용되는 다양한 원천의 자료에 대한 이해도를 높일 수 있는 체계 제시
- 행정자료, 조사자료, 행정자료와 조사자료를 통합한 자료를 통해 생산된 통계의 품질과 관련하여 각각의 자료의 강점과 제약 설명

○ 품질관리 지침의 활용

- 모든 자료 수집 과정에서 측정오차는 발생되며 한 가지 원인으로 발생하지 않지만 지침서에서는 오차의 관점(error frame)에서 데이터셋과 산출물의 장점과 약점을 종합적으로 설명
- 품질을 판단하기 보다는 자료 본래의 생성 목적에 맞게 강점과 약점 제시
- 품질 측정 결과는 작성된 통계의 일관성을 점검하거나 향후 통계작성에 필요한 자료 수집 계획을 수립할 때 활용 가능

○ 지침서의 주요 내용

- 오차 구조(error framework) 하에서 행정자료가 자료의 본래의 생성 목적과 확장된 활용 목적에 얼마나 부합하는지 점검하고자 할 때 확인해야 하는 사항 제시
- 서로 다른 원천의 자료를 통합하여 활용할 때 발생할 수 있는 문제점 설명
- 주요 품질 요소와 관련된 수량화한 지표들과 측정 방법 소개
- 지침에서 포괄하지 못한 사항들, 특히 특정 통계에 필요한 기술적으로 복합한 품질관리 척도의 개발과 관련된 아이디어 제시
- 오차의 구조로 설명하지 못하는 품질 관련 문제들은 데이터셋 관점에서 강조하고 향후 업무에 필요한 사항들 탐색
- 지침서, 품질 평가 지표, 관련 양식은 뉴질랜드 통계청 홈페이지에 게시

○ 지침서에 포함된 사항

- 오차의 구조(error framework) : 다양한 자료와 산출물에 적용하기 위해 활용
- 적용 사례 : 오차의 구성요소별로 적용한 품질 측정표 견본
- 오차 체계를 적용한 품질관리 실행 계획과 향후 업무 방향
- 통계설명자료 견본 : 데이터셋의 품질 평가에 유용한 주요 정보를 엑셀 시트로 제공
- 품질 척도의 상세한 목록 : 오차 유형별 품질 지표와 척도의 목록을 두 가지 유형으로 제공하며 자료의 특성에 맞게 선택하여 활용
- 지침에 사용한 주요 용어에 대한 설명과 목록 제공

○ 품질 관리를 위해 점검해야 하는 사항들

- 오차의 구조를 적용한 품질관리방법은 거의 모든 자료와 산출물에 적용 가능하지만, 이를 이해하고 적용하는 과정이 어렵고 복잡함
- 자료의 품질을 평가하기 이전에 다음의 사항들 점검 필요

1) 오차의 구조를 적용하고자 하는 목적은 무엇인가
   수집된 행정자료의 품질 평가, 산출물의 품질 평가를 위한 보다 나은 척도 개발, 행정자료를 활용할 때 품질 관점에서 설계에 영향을 주는 요소 이해, 행정자료와 산출물의 품질과의 절충 요소에 대한 이해 등 명확한 목적 정립 필요
2) 작성하고자 하는 통계와 가장 연관성이 높은 자료는 무엇이며, 가장 중요한 목적은 무엇인가
3) 검토 중인 데이터셋이 이전에도 활용된 적이 있는가
   이전의 작업이 현재의 업무 시간을 단축하는데 도움이 되는가, 통계설명자료 작성 또는 데이터의 정제가 다른 업무에게 이미 이뤄졌는지 내부 시스템을 통해 확인
4) 각각의 데이터셋에 있는 변수는 무엇인가, 추후 활용할 가장 중요한 변수는 무엇인가
5) 검토 중인 자료들이 대표하는 모집단은 무엇인가, 개인, 가구, 사업체 등 기본 단위 확인
6) 최종 데이터셋 생성을 위해 원자료의 변수를 어떤 방식으로 변환(transform)하거나 결합(combine)할 것인가
7) 원자료의 기본단위를 최종 데이터셋의 통계단위로 어떻게 전환(convert)할 것인가
8) 행정자료의 활용과 관련하여 원자료에 대해서 알고 있거나 예측하고 있는 품질관련 문제는 무엇인가

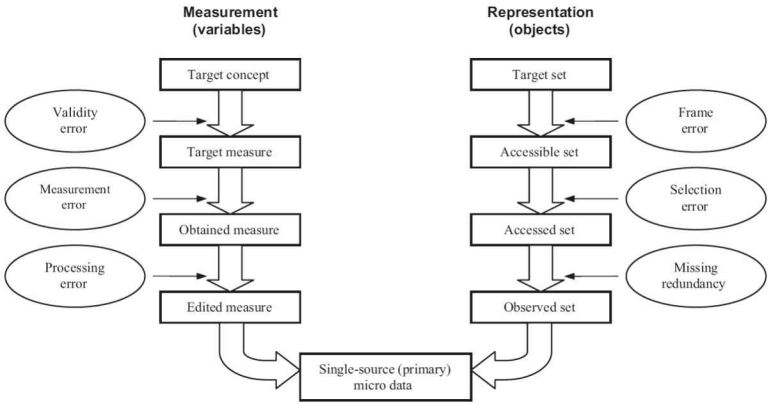
○ 통계설명자료 양식(metadata information template)

- 통계설명자료의 양식은 자료의 품질을 체계적으로 이해하는데 도움
- 서로 다른 원천의 자료를 비교하고 관련된 정보를 표준화된 양식으로 기록하기에 편리함
- 일반적인 사항 : 작성 기관, (본래의)수집 목적, 포함된 변수, 자료의 기간
- 모집단 : 목표 모집단, 관리 모집단(admin population), 기본 단위(reporting units) 등 포함률(coverage)과 관련된 정보
- 변수 : 변수 설명, 측정하고자 한 개념 기록

- 수집 방법 : 시간·지연 정보, 자료 수집 방법
- 그 밖에 원 자료를 이해하는데 도움이 되는 정보 기록

○ 1단계 오차 구조(phase 1 of the error framework)

- 행정자료를 포함한 통계작성에 필요한 자료들의 품질을 평가하고, 발생 가능한 오차를 파악하고자 할 때 적용할 수 있는 틀 구성, 단일 자료(single dataset) 적용 목적
- Li-Chun Zhang(2012)의 조사통계의 오차 체계를 바탕으로 구성
- 1단계는 자료가 고유의 목적에 맞게 생성되었는지 단계별로 점검하기 위한 틀로써 조사자료와 행정자료에 공통적으로 적용 가능
- 분석에 활용되는 자료가 다양할 경우 각각의 자료에 대하여 개별적으로 적용
- 품질 평가를 위해 아래 그림의 오차와 연관된 요소들(Box)을 모두 확인



<측정 측면 : 변수>

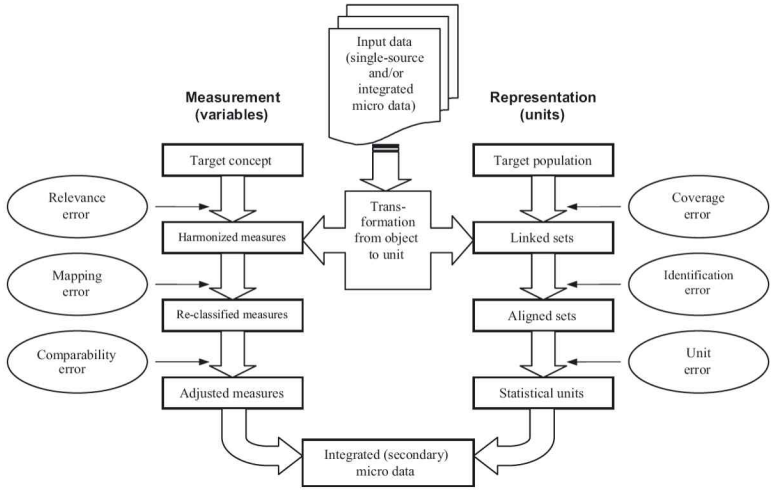
- 타당성과 관련된 오차(validity error) : 행정자료의 항목(조사항목)이 최초 목적한 정보를 수집(개념을 측정)할 수 있도록 구성되지 못한 경우 발생
- 측정 오차(measurement error) : 조사 또는 행정 양식 등을 통해 수집된 값과 실제값의 차이로 인해 발생
- 처리 오차(processing error) : 조사자료 또는 행정자료의 값의 오류를 점검하고 수정하는 과정에서 발생

<대표성 측면 : 대상>

- 구성 오차(frame error) : 행정자료(accessible set)가 분석하고자 하는 관심 집단(target set)을 모두 포괄하지 못하는 경우 발생
- 선택 오차(selection error) : 행정자료에 포함되어야 하지만, 실제로 활용 가능한 자료(accessed set)에 포함되지 않은 자료가 있는 경우 발생
- 누락·중복 오차(missing/redundancy error)

○ 2단계 오차 구조(phase 1 of the error framework)

- 이미 생성되어 있는 자료들을 활용하여 새로운 통계를 작성하고자 할 때, 발생할 수 있는 오차 설명
- 서로 다른 모집단의 자료를 통합하거나 여러 데이터 셋을 융합할 때 적용
- 2단계의 측정과 관련된 오차는 서로 다른 자료에 포함된 변수들을 융합하고 조정하는 과정에 초점을 맞추고 있으며, 대표성과 관련된 오차는 원 자료로부터 통계분석을 위한 데이터셋과 통계단위를 생성하는 과정에 초점
- 이 단계는 단일 자료에도 적용 가능하지만, 원 자료의 품질 문제와 통계를 생산하는 과정에서 의도하지 않게 발생하는 오차를 구분해야 함



<측정 측면 : 변수>

- 타당성과 관련된 오차(relevance error) : 1단계의 validity error와 유사하지만, 자료를 통합하기 위해 각각의 자료에 포함된 유사 항목들의 개념을 조정(harmonized measures)하는 과정에서 발생하는 오차
- 구조화 오차(mapping error) : 변수의 값을 조정된 개념에 맞게 변환(re-classified measures)하는 과정에서 발생하는 오차
- 일관성 오차(comparability error) : 각각의 자료에서는 오류가 없지만 통합 데이터셋에서는 내적 일관성이 결여되는 문제 발생, 추가적인 변수값 보정(adjustment) 필요

<대표성 측면 : 단위>

- 포함 오차(coverage error) : 조사통계에서의 포함오차와 유사한 개념으로 자료 연계를 통해 생성된 데이터셋이 관심 모집단을 얼마나 대표하는가와 관련된 오차, 데이터셋의 단위가 통계작성 단위와 일치하지 않을 수 있음

- 개별화 오차(identification error) : 자료의 단위와 관련된 오차로써, 통합자료(aligned sets) 생성 시 자료들의 단위가 일치하지 않아 상위 단위로 통합하는 등 단위를 조정하는 과정에서 발생하는 오차
- 단위 오차(unit error) : 통합자료의 단위가 아닌 통계작성을 위한 단위를 생성하는 과정에서 발생하는 오차

<참고문헌>

Statistics New Zealand (2016). "Guide to reporting admin data quality" from http://www.stats.govt.nz/methods/data-integration/guide-to-reporting-on-admin-data-quality.aspx

Zhang, L-C (2012). "Topics of statistical theory for register-based statistics and data integration." *Statistica Neerlandica* 66: 41-63, DOI: 10.1111/j.1467-9574.2011.00508.x.

□ ITSEW 프로그램

날짜	세션	주제
Day1 (10/10)	S1	Keynote Presenter: Dr. Paul Biemer
	S2	Applying TSE or Quality Frameworks to Research, to administrative or transactional data
	S3	Multimodal or web questionnaire design and respondent engagement to minimise TSE
	S4	Questionnaire design and measurement error
Day2 (10/11)	S5	Using paradata to reduce non-sampling error
	S6	Minimising the effect of selection or response bias on statistical inference
	S7	TSE applications
	S8	Measuring TSE in the production and compilation of statistics
	S9	Assessing the quality of administrative data, including in comparison with survey data
Day3 (10/12)	S10	Measuring measurement error and impacts of measurement change
	Reflections	Learnings from ITSEW 2016 – Planning for ITSEW 2017
	Farewell	Concluding remarks / Farewell : Dr. Siu-Ming Tam

※ 발표자료는 2016년 총 조사오차 워크숍 홈페이지에 게시 예정
https://consol.eventsair.com/QuickEventWebsitePortal/itsew2016/itsew16