통 계 - 2005 - 04 - 13
무응답처리 교육과정

무응답처리 실무론

서비스업동향과

통 계 – 2005 – 04 – 13
무응답처리 교육과정

무응답처리 실무론

서비스업동향과

머 리 말

이 보고서는 본인이 뉴질랜드 통계청(Statistics New Zealand)에 근무하면서 습득한 무응답 처리기법 가운데서 기초적이면서 범용적으로 쓰이는 imputation 방법을 통계청 인구조사과의 2005센서스 무응답처리기법 개발에 파견근무 하여 개발업무에 참여한 직원들을 지도하면서 정리한 내용들을 재편집하여 작성한 자료입니다.

제1장은 무응답 처리개요로서 무응답처리를 위해 기본적으로 알아야 할 기본 용어와 처리 준비사항 및 무응답처리방법 등을 개략적으로 설명하였고,

제2장은 가구조사에 주로 쓰이는 방법들을 이용측면에서 프로그램 이용법과 함께 소개하였고,

제3장은 사업체조사에서 주로 쓰이는 대표적인 방법을 간단히 소개하면서 imputation 방법선택 및 imputation class 결정에 고려할 사항을 그간의 경험을 중심으로 작성하였습니다.

또한 부록에는 본인이 Statistics New Zealand에서 개발한 SAS program의 source도 함께 수록하였으니 관심 있는 독자들이 계속 기능을 추가하였으면 하는 바람입니다.

아울러 앞으로 한국 통계청에서 본격적으로 imputation 처리기법을 개발할 때 이 보고서가 조금이나마 기여를 했으면 하는 기대도 가져 봅니다.

마지막으로 이 보고서가 나올 수 있도록 개발에 참여하여 많은 충고를 아끼지 않으신 김형석사무관, 양경진씨, 정남수씨, 송일규씨, 강영민씨에게 이 자리를 빌어 감사 드립니다.

2005. 4

송 순 관
통계청 인구조사과 · Statistics New Zealand

목 차

제 1 장 : 무응답 처리 개요

제 2 장 : 주요 imputation방법 소개

부 록

제 1 장

무응답 처리 개요

1. 무응답 종류

무응답이란 조사표에 기입할 조사사항을 기입치 않은 것을 말하며 여기에서 응답자가 조사항목의 응답거부, 조사자가 응답자를 조사기간에 전혀 만나지 못하여 조사표 전체를 공백으로 한 것을 포함하는데, 인구 센서스와 같은 면접조사,인터넷조사 및 응답자 직접 기입하는 조사방법이 혼합형태를 이루고 있는 조사에서는 다양한 무응답 형태가 발생될 것으로 예상되어진다.

이에 관계된 무응답 종류를 크게 두 가지로 나누어보면, 조사단위 무응답(Full nonresponse/Unit nonresponse) 및 조사항목 무응답(Partial non response/Item nonresponse)으로 나눌 수 있는데 이들에 대하여 개념들을 간략히 정의해 보도록 하고 이들에 대한 처리 방법들을 생각해보고자 한다.

1.1 조사단위 무응답(Unit nonresponse/Full nonresponse)

가구조사에서 나타나는 조사단위 무응답 종류는 크게 가구전체의 무응답(Household nonresponse)과 가구원 일부의 무응답(Household member nonresponse)으로 크게 나눌 수 있으며, 사업체 조사에서 나타나는 조사단위 무응답은 조사단위 전체가 무 응답하는 형태를 나타난다.

1.2 항목 무응답(Item nonresponse/Partial nonresponse)

항목 무응답은 모든 통계조사에서 여러 가지 이유에서 발생한다. 항목 무응답은 응답자가 특정 항목에 대하여 응답을 거부하거나 혹은 응답자의 항목에 대한 이해 부족으로 인하여 그 항목에 대한 응답기피 등으로 인하여 발생할 수 있다. 때로는 응답된 내용을 사용하기는 너무 부실하여 그 응답된 내용을 수정하는 것 보다는 오히려 무응답으로 처리 하는 것이 바람직한 경우도 발생할 수 있다.

무응답 항목은 에디팅 단계에서 오류리스트에 포함 시키지 말아야 한다. 만약 항목 무응답을 에디팅 단계에서 오류리스트에 포함시키면 에디팅 요원으로 하여금 에디팅 부실행위라는 평가를 회피하기 위하여 임의로 무응답 내

용을 임의로 채워넣는 사례가 발생될 우려가 있어 그 항목에 대한 편기(bias) 결과를 초래할 위험이 있기 때문이다. 때문에 무응답 항목은 입력 및 에디팅 단계에서 무응답이 그대로 보존되도록 철저한 에디팅 전략이 필요로 한다.

이러한 항목 무응답을 발취할 수 있는 전략이 수립되어서 imputation단계에서 이용할 수 있도록 사전준비가 자료처리 단계에서 고려되어야 한다. 항목 무응답은 입력단계에서 결측 치(missing)로 남겨두면 입력요원이 그 항목에 대하여 입력을 하지 않았는지 아니면 항목이 무응답 이었는지를 구분할 방법이 없어서 반듯이 입력단계에서 무응답 코드를 부여하여 무응답 항목에 대하여서는 명확한 무응답항목과 입력요원의 입력누락을 반듯이 구별할 수 있어야 한다.

1.3 종단면조사 무응답(Wave nonresponse)

한 조사단위를 반복하여 조사하는 종단면조사와 같은 경상조사의 경우에는 응답자의 응답된 시계열을 유지하면서 조사된 조사자료가 한 특정한 조사시점에 조사단위 전체가 자료를 제공하지 않았다 하더라도 이를 단위 무응답으로 처리하지 않고 한 항목의 시계열이라는 자료의 각도에서 한 시계열 부분이 무응답 된 것으로 보아 항목 무응답으로 처리되는 경우도 있다.

1.4 이용할 수 없는 응답 항목(Unusable response)

조사된 자료를 처리하다 보면 간혹 우리는 응답된 내용이 너무도 지리멸렬하게 응답되어서 잘못된 항목들을 에디팅을 하기에는 너무도 복잡하여서 응답된 항목들의 내용을 오히려 무응답으로 간주하여 imputation기법으로 처리 하는 것이 용이할 경우가 있다.

2. 무응답발생 형태

조사대상자 가운데 무응답그룹이 어떤 특징을 가지고 무응답으로

나타났는지를 무응답처리 전략수립에 반영되어야 한다. 그 무응답그룹의 형태와 특성을 이해하고 나서 그에 따라 적절한 imputation방법과 imputation class결정에 이 특성을 반영하여야만 정도 높은 자료를 생산할 수 있다. 그 대표적인 형태가 다음 두 가지가 있다.

2.1 무작위 무응답(Uniform nonresponse)

이는 일명 MCAR(Missing Complete At Random) 혹은 MCR(Missing Complete Random)로 표기되는데 이는 무응답이 완전히 어떠한 조사상황 혹은 특정한 질문항목에도 관계없이 무응답이 무작위로 발생하는 경우를 무작위 무응답이라 한다.

현실적으로 거의 이러한 형태의 무응답이 발생하는지를 분석하기에는 용이하지는 않다. 무응답그룹의 크기가 상대적으로 크지 않으므로 전체 결과에 큰 영향이 미치지 않고, 어떤 특수한 분석이 무응답 그룹에 대하여 이루어 지지 않는 다면 무응답그룹을 이 무응답 모델로 가정하여 적절한 imputation방법을 결정 하는 것도 문제가 없을 것으로 생각되어진다.

2.2 종속된 무응답(Non uniform nonresponse)

이는 일명 MAR(Missing At Random) 혹은 MR(Missing Random)로 표기되는데 이는 무응답이 어떤 특수한 항목과 관련이 있어서 이에 종속적으로 나타나는 무응답 사례를 종속된 무응답이라 한다. 선행 항목이 다음에 물어볼 항목에 응답영향을 주는 경우가 한가지 그 대표적인 예가 될 것이다.

예를 들면 개인의 수입항목을 질문하고 나서 다음에 지출항목을 질문하였다면 이는 수입이라는 질문항목에 응답기피 현상때문에 다음의 지출항목을 응답할 의사가 충분히 있는 응답자 임에도 불구하고 그 지출항목의 응답을 기피할 경우가 있을 것이다.

이러한 형태의 무응답은 imputation전략수립에 이러한 관련변수가 보조변수 혹은 imputation class변수로 반영되도록 하여야 한다.

3. 무응답처리 방법들

조사가 완료된 후에 조사자료에서 나타난 무응답을 어떠한 방법으로 처리하는 방법들이 있는지를 생각해보도록 하자.

3.1 오직 응답된 자료만 이용(Use only responses)

이 방법은 아주 간단한 방법으로서 어떠한 무응답에 대한 자료를 생산할 대책도 수립하지 않고 자료처리를 수행하는 것이다. 거의 모든 통계분석도구에서는 이를 결측 치(missing) 처리가 가능하다. 조사자료 처리는 아주 신속하게 이루어 지지만 이는 무응답의 크기 정도에 따라서 추정에 대한 편기(bias)가 발생될 위험이 있다.

무응답이 전체 조사에 대하여 무시할 정도로 아주 미미한 경우는 그런대로 생각해볼 방법이지만 완전한 조사설계에 기초를 둔 추정은 불가능하니 무응답처리 기법에 투입될 자원이 없는 경우가 아니면 추천할 방법이 아니다.

3.2 가중치 조정 방법(Re-weighting)

이 방법 또한 어떤 무응답에 대하여 자료를 생성하지 않는 간단한 방법 가운데 하나이다. 추정결과의 편기를 최소화 하가 위하여 응답자수만을 고려하여 가중치를 재 조정하는 방법이다(re-weighting). 대다수 통계분석도구에서 이를 수용할 수 있도록 설계되어 있다.

무응답자가 표본의 크기에 비하여 아주 큰 영향이 없다면 아주 간단한 방법으로 단위무응답(unit nonresponse)을 처리하는데 자주 쓰이는 방법이다. 하지만 항목무응답(item nonresponse)을 처리하기에는 항목별로 가중치가 재조정되어야 하므로 많은 가중치관리와 분석의 복잡성이 수반되는 방법이므로 항목무응답에는 적용되는 방법이 아니다.

3.3 무응답자 대체 법(Nonresponse substitution)

이는 표본조사의 단위무응답에서 주로 사용하는 방법으로서 당초 설계된 표본의 크기를 유지하기 위하여 단위 무응답이 발생하면 이를 다른 대상자로 대체하여 조사하는 방법(unit substitution)으로 완전한 표본의 크기를 유지시키는 방법이다.

이 방법은 완전한 데이터를 구성하는 장점은 있지만 항목 무응답은 별도로 해결할 방법을 생각해야 한다. 만약 다른 대상자를 조사당시 대체한다면 조사원으로 하여금 발생될 대상자선정에 편기가 발생할 수 있고, 조사 후에 대체가 이루어 진다면 별도의 추가 조사기간이 필요하여 자원을 배분해야 하는 문제가 발생된다. 대체 대상을 선정하는 과정에서도 추출확률계산(inclusion probabilities)이 복잡하여 최종 가중치 계산과정이 복잡해 질 수도 있다.

3.4 무응답자 질의(Follow-up)

주로 항목 무응답을 해결하기위한 수단으로 사용하는 방법이다. 조사도중 조사요원이 누락된 항목을 발견하였을 때 혹은 조사기간이 종료된 후에 주로 에디팅 요원이 누락된 항목을 보완하기 위하여 응답자와 연락을 취하여 무응답부분을 해결하는 방법이다.

자료수집 비용적인 측면에서 가능하다면 이 방법이 질적인 자료를 확보하는데 가장 좋은 방법이지만 비용과 질적자료의 득실을 고려한다면 바람직한 방법은 아닌 듯 하다. 그러나 imputation방법 연구 측면에서 무응답자들의 특성을 연구하는 방법으로는 일부 무응답자에게 질의를 시도해볼 필요가 있다고 생각되어진다.

일부 서방국가에서는 사회구조가 복잡 다양한 형태로 변화하고 있고 개인정보의 누출을 꺼리는 조사환경이 점점 팽배해 가면서 국민들 응답부담 경감이라는 국가통계조사 기본정책을 수호하기 위하여 응답자를 떠난 조사표에 대해서는 가능한 재 접촉은 금기로 조사지침이 되어있다. 이 방법을 선택 하는 데는 사회적환경과 통계조사의 관련 법규사항 및 자원의 가능성을 사전에 검토되어야 한다.

3.5 Imputation 방법적용

이는 무응답에 대한 자료를 합리적인 방법을 동원하여 그럴듯한 값(plausible value)을 창조하여 무응답자 혹은 무응답항목에 창조된 값을 할당하는 방법이다. 값을 창조하는 방법은 동일응답자의 응답된 항목을 이용하는 방법(deductive/deterministic method) 혹은 응답자모두의 응답된 내용을 이용하여 확률적인 방법(stochastic method)으로 값을 창조하는 방법이 있다.

이는 모든 조사대상자에 대하여 모든 항목에 값이 존재하여 완전한 자료를 구성하게 되므로 분석자 모두가 동일한 결과를 항시 산출하는 큰 장점이 있다. 하지만 무응답 크기의 정도에 따라서 창조되어 할당된 값이 추정에 편기를 가져올 위험이 있을 수 있다. 또한 창조된 무응답 값이 항목과 항목간의 논리적인 관계를 해칠 수도 있으니 imputation후에 자료의 에디팅을 해야 하는 새로운 자원 할당을 고려해야 한다.

3.6 무응답 사전방지(Nonresponse prevention)

이는 무응답이 발생하기 이전에 조사의 기획단계에서부터 무응답 방지전략을 수립하는 것이다. 조사의 기획가운데 한 분야인 조사표설계 단계에서 고려하여야 할 몇 가지 중요한 사항만 살펴보면 다음과 같다.

- 간단명료한 질문형식
- 적당한 질문의 길이
- 가능한 전문 약어 용어의 회피
- 자료수집방법을 고려한 질문작성
- 간단명료한 지시사항
- 폐쇄형 혹은 구간형식의 질문 최대화

조사항목 선정 시 현실적으로 수집 가능한 조사항목인가를 신중히 검토하고 선정된 조사항목에 대해서는 응답자가 이해할 수 있는 정도로 명확한 정의를 하였는가를 아울러 점검하여야 한다. 그러하지 않으면 질문항목에 이해의 난이도에 따라서 응답자는 응답을 회피하는 경향이 있기 때문이다.

본격적인 조사표설계 이전에 먼저 행하여야 할 중요한 몇 가지를 고려해 보면 다음과 같다.

- 모든 관계자들의 의견청취
- 과거의 동일 혹은 유사조사의 경험평가
- Focus group research
- 시험조사의 실시 등

이외에도 조사설계, 표본설계 및 자료수집 단계에서 행하여야 할 사항들은 여기서는 생략하기로 한다.

4. Imputation을 위한 준비사항

어떠한 방법을 적용한 훌륭한 조사기획이 있었더라도 무응답이 발생 하는 것은 현실적으로 피할 수 없다. 무응답처리 방법결정 및 자료처리 단계에서 준비할 사항들이 있는데 여기서는 간단하게 자료처리 단계에서 준비하여야 할 사항들을 소개하고자 한다.

- 조사단위의 조사 및 입력시스템에서 고려하여야 할 사항은 조사단위의 조사참여를 구분할 수 있는 참여코드를 설계하여야 한다. 이를 바탕으로 하여 단위무응답을 처리할 수 있는 근거가 되는 것이다. 예를 들면 응답(1), 부재(2), 불응(3), 대리응답(4), 응답불능(5) 등으로 가구에 대한 조사 참여코드를 생각해 볼 수 있다.

- 입력자료 코드부여시 반듯이 무응답 코드를 부여하여 입력 후에 무응답 여부를 구분할 수 있는 장치를 마련한다. 예를 들면 조사표에서 성별(sex) 변수를 수집 하였다면 이 성별에 대하여 남자(1), 여자(2) 및 무응답(9)을 코드 설계하여 입력시스템에서 이를 처리 하도록 설계한다.

- 조사표의 자료입력 완료 후에 원시자료 보관시스템은 무응답의 코드를 찾아 내어 무응답 변수(Non Response flag: NR변수)를 각 대상변수에 부여한다. 예를 들면 성별(sex)은 NR변수 sex_nr 혹은 nr_sex 등을 생성하여 무응답이 확인된 조사단위의 sex변수는 sex_nr변수에 무응답여부의 표시를 부여한다(예: sex_nr='y'). Imputation처리 시스템은 이 변수를 이용하여 언제든지 관심변수에

대한 무 응답 대상여부를 구분한다.

- Imputation처리 시스템을 수행 한 다음에 어떤 방법으로 imputation을 수행하였는지 구분하기위한 imputation방법 flag가 필요하다. 예를 들면 성별 변수에 대한 imputation방법 flag변수는 sex_imp로 생성할 수 있다. 이는 imputation방법 여부 및 방법에 대한 점검을 후에 분석 하기위한 장치이다.

- 자료보관에 대한 여유공간 및 자료처리에 큰 문제가 없다면 imputation class정보와 무응답자에게 값을 부여한 값 기증자(donor)를 구분할 수 있는 ID도 함께 보관 해야 한다. 그러나 이는 권고 사항이고 앞에서 언급한 무응답 코드, NR변수 및 imputation flag는 반듯이 자료처리시스템에 고려하기를 권고 한다.

- 앞서 권고한 무응답관련 flag변수들을 일반 이용자에게도 제공할 것인가는 충분한 검토가 사전에 있어야 할 것이다. 무응답으로 인하여 생산된 자료 때문에 전체 조사자료의 질을 의심할 여지가 있으므로 이들을 분간할 수 있는 무응답관련 flag변수들은 최종 자료형성 파일에 제공되지 않는다. 이 무응답관련 flag변수들이 있는 조사자료는 향후에 내부적으로 비표본 오차 및 유사종류조사의 발전을 위한 연구자료로 이용이 될 것이다.

5. Imputation 방법 분류

Imputation방법 분류는 어디에서나 쓰이는 정형화된 분류용어는 없지만 참고문헌 및 학술지 등에서 출현하는 용어를 이용하여 방법적인 면에서 상대적인 개념으로 이용되는 용어를 비교하여 대표적인 3가지 카테고리를 이야기 할 수 있다.

- 연역적인 방법 대 확률적인 방법(deductive/deterministic/logical method versus stochastic method): 전자는 주로 동일 응답자 내에서 응답된 항목을 중심으로 무 응답한 항목의 값을 계산 혹은 추론하는 방법이고 후자는 무 응답한 항목을 응답한 응답자들의 응답항목을 중심으로 통계적인 방법을 적용하여 무 응답한 항목의 값을 이끌어 내는 방법이다.

- **핫덱 방법 대 콜덱 방법**(Hot Deck method versus Cold Deck method): 이 방법은 imputation방법을 적용할 때 주로 어떠한 자료를 이용하느냐에 기준을 두고 이야기 하는 방법이다. 전자는 현재의 조사자료를 이용하여 무응답 항목의 값을 유도하는 방법이고 후자는 과거의 조사자료 및 과거의 유사 자료를 이용하여 현재의 무응답 항목의 값을 유도하는 방법이다.

- **단일 값 대체 방법 대 복수 값 대체 방법**(Single imputed value method versus multiple imputed value): 전자는 현실적으로 주로 통계조사에서 적용하는 방법으로 무 응답된 항목의 값을 오직 하나만 산출 해내어 마치 응답한 항목의 값처럼 분석에 직접 쓰이도록 하는 방법이고 후자는 무 응답한 한 항목에 여러 개의 값을 산출하여 상황에 맞게 그 산출된 값들을 사용하는 방법으로 주로 imputation연구에 쓰이는 방법이다.

구체적인 방법들 용어를 사용하여 분류하자면 다음과 같은 대표적인 용어들이 있다. 여기서 방법들에 관한 구체적인 설명은 생략하고 이 가운에 몇 가지 중요한 방법들은 다음 장에서 좀더 구체적으로 설명을 하고자 한다.또한, 방법을 한글로 번역 하지 못하고 직접 원어를 사용하고자 하니 정형화된 한글용어는 독자에게 남기기로 한다.

- Logical/deductive imputation
- Mean imputation*
- Ratio imputation*
- Regression imputation
- Model imputation
- Probability imputation*
- Previous value/Historical imputation*
- Unit-trend imputation
- Group-trend imputation*
- Cold deck imputation
- Hot deck imputation*
- Nearest neighbour imputation
- Nearest neighbour trend imputation
- Imputation with residuals imputation

- *Moving average method**

노트: * 다음 장에서 자세히 방법론을 설명할 부분.

6. Imputation과 editing관계

에디팅과 imputation을 정확히 구분하여 정의 하기는 지금까지 많은 논란이 있어 왔다고 생각되어진다. "잘못 응답된 항목에 대한 자료를 옳다고 추정하는 값으로 수정하는 것이 에디팅이고 무 응답한 항목에 자료를 창조하는 것이 imputation이다". 라는 아주 간단한 정의가 공감이 가는 해석인 듯 하다.

그러나 imputation에서 응답된 내용을 기초로 하여 무 응답한 내용 혹은 잘못 응답 되었다고 추정되는 값을 무시하고 옳다고 추정되어지는 값을 유도하였다면 이를 에디팅으로 보아야 하는가 아니면 imputation으로 보아야 하는가는 논란이 있을 수 있다. 왜냐하면 응답된 항목의 값은 무시하고 새롭게 자료를 창조하기 때문이다. 이러한 작업절차를 imputation의 범주로 분류한다면 이를 연역적 방법(deductive/deterministic method)으로 분류하기도 한다.

용어에 어떤 정의를 부여 하든가에 종국적인 목적은 양질의 응답자료를 생산하는 일련의 작업 절차임에는 분명한 사실이다. 때문에 필자는 여기서 용어의 혼란을 피하기 위하여 다음과 같이 에디팅과 imputation을 정의하여 사용하고자 한다.

- 에디팅(Editing): 응답된 자료를 정비하는 절차로 응답된 내용을 기초로 하여 잘못 되었다고 추정되는 값(일부항목 무응답 포함)을 수정 혹은 값을 창조(유도)하는 작업을 의미 한다.

- Imputation: 에디팅 단계에서 해결하지 못한 무 응답한 항목에 대하여 확률적인 방법(stochastic method)을 적용하여 값을 창조하는 작업을 의미한다.

하지만 이 두 방법을 철저히 분리하여 단계별 작업처리를 할 수는 없고 편리에 따라 imputation과 editing을 무응답 처리 작업상 혼용하여 사용하기도 한다.

제 2 장

주요 Imputation방법 소개

1. 방법 소개 배경

이 장에서 소개 하고자 하는 방법은 학계등에서 발표하는 여러 imputation관련 연구보고서에 주로 이용되는 복잡한 수식 및 공식 등은 피하고 방법이 어떻게 작동하느냐의 실무적인 측면에서 방법의 개념설명과 어떻게 실제 업무에 적용하는가 하는 이용방법 측면에서만 소개하고자 한다.

Imputation에서 사용되는 용어들은 아직 정착이 이루어지지 않은 이유로 적용면에서는 거의 유사한 방법들이 여러 참고문헌 및 연구보고서에서는 간혹 다른 이름들로 소개되고 있다.

본 보고서에서는 현재 뉴질랜드 통계청 및 기타 통계기관에서 주로 가구조사와 사업체조사에서 주로 사용하는 대표적인 방법으로서 이름이 부여된 방법들만 소개 하고자 한다.

2. 가구조사에 주로 쓰이는 방법 소개

다음에 제안된 세 종류의 Imputation방법들은 뉴질랜드 통계청(Statistics New Zealand)에서 가구조사 무응답처리로 주로 사용되고 있는 방법들인데 가구조사관련 무응답처리기법이 필요한 독자는 이 방법들을 검토하여 변수의 특성에 따라 적절하게 이 방법을 적용함에는 별 문제가 없을 것으로 생각되어 다음 세가지 방법들을 소개 하고자 한다. 여기서는 방법들에 대한 이론적 배경은 독자의 연구 몫으로 남겨놓고 업무에 이용방법 측면에서만 소개하고자 한다.

현재 뉴질랜드 통계청 경상가구통계조사의 표준화된 지침서에 무응답 처리기법으로 이 방법들을 사용하도록 권장하고 있다. 이 권장된 방법을 적용하기 위하여 자체 개발된 정형화된 SAS macro program (IMPMethodWizardVersion2.sas developed by Soon Song)과 그의 이용방법을 간단히 소개하고 부록에는 program source도 함께 제공하니 관심 있는 통계조사의 imputation 개발작업에 참고를 하여 주었으면 한다.

관심 있는 분들은 이 program source을 면밀히 분석하여 더욱 능률적인 program으로 기능을 다할 수 있도록 구조와 기능을 보강하여 한국 통계청

imputation기법 보급에 기여를 하였으면 하는 기대를 하면서 다음 방법들을 소개 하고자 합니다.

2.1 Probability method

이 방법은 관심 있는 항목에 대하여 카테고리별 분포도를 이용한 방법으로 무응답이 카테고리별로 무작위로 발생(MCR/MCAR) 하였다는 가정 아래에서 카테고리별 사례수가 점유하는 분포의 정도에 따라 무응답자를 확률적으로 배분하는 방법으로서 연속변수(예: 소득변수) 보다는 불연속변수(예: 성별변수)에 주로 사용하는 방법으로 알려져 있다.

2.1.1 방법 설명

이 방법을 이해하기 위하여는 분포도의 생성이 먼저 필요로 한다. 예를 들면 다음 테이블과 같이 성별 '남자(1)'의 관심변수(X:5개 카테고리)가 다음과 같이 각 카테고리별로 분포를 이루고 있다고 하자. 성별이 '1' 가운데 1명의 무응답이 발생 하였다면 전체 성별 '1' 가운데 카테고리 'A' 중에서 무응답이 발생할 확률은 10%(0.10) 카테고리 'B' 중에서 발생할 확률은 15%(0.15) 등으로 우리는 무응답자의 발생 정도를 각 카테고리의 분포의 정도로 기대 할 수 있을 것이다.

만약 성별 '1' 중 무응답자를 균등확률분포 0<f(x)<1에서 무작위로 값 하나를 선택하여 이 확률 값을 카테고리의 누적확률 범위에 속하는 카테고리값을 무응답자의 변수 X의 카테고리의 값으로 할당하는 imputation방법을 probability method라 한다.

남자(1)

X	빈도수	누적빈도수	구성비	누적구성비	누적확률 범위
'A'	10	10	0.10	0.10	0.00<=R<0.10
'B'	15	25	0.15	0.25	0.10<=R<0.25
'C'	20	45	0.20	0.45	0.25<=R<0.45
'D'	30	75	0.30	0.75	0.45<=R<0.75
'E'	25	100	0.25	1.00	0.75<=R<1.00

예를 들면 성별 '1'인 한 무응답자가 무작위로 추출한 확률변수 값이 0.042라면 이 무응답자의 X 변수의 응답한 카테고리값을 'A'로 할당할 수 있을 것이다.

2.1.2 Imputation class

카테고리의 분포도를 어떻게 산출하느냐 하는 문제에 봉착하게 되는데 무응답발생 형태가 완전 무작위(MCAR/MAR)로 가정한다면 전체조사 자료에서 생성된 분포도를 사용할 수 있지만 현실적으로 이는 불가능한 경우이므로 이와 유사한 환경을 구축하여 분포도를 산출한다. 이를 imputation class라는 용어로 정의하여 조사 단위들이 동질적인 특성이 있는 그룹으로 분해하여 분포도를 구하도록 하는 것이다.

Class변수는 주로 관심변수 X를 잘 설명 해줄 수 있는 보조변수를 class변수로 선택하여 class변수의 조합으로 형성된 카테고리별로 각각 분포도를 산출하여야 하는데 이 class변수들은 주로 관심변수 X를 동질성(homogeneity) 있게 분류 할 수 있는 분류변수 이어야 바람직하다.

여기서 사용되는 관심변수의 분포도는 조사당시의 응답자들을 이용하여 주로 구하여 지는데 전체 조사대상자에 비하여 무응답자가 무시하기에는 너무 많은 비율을 차지하는 경우에는 과거조사결과 혹은 유사한 다른 조사의 결과를 이용 하여 분포도를 구하여 이용할 수도 있다.

예를 들면, 인구센서스는 과거의 경험으로 보아 항목별 무응답비율이 크지 않을 것으로 기대되므로 금번 조사 응답자로부터 산출하더라도 문제가 없을 것으로 사려 된다. 만약에 어떤 관심 항목에 대하여 무응답자의 비율이 무시하기는 곤란한 정도(예:50%이상)가 발생된다면 과거 센서스의 분포도를 이용함이 타당할 것으로 생각되어진다.

2.1.3 '%IMPProb' Program 이용방법

이는 Probability method을 원활히 수행하기 위하여 작성한 SAS macro program이다. Imputation class별 분포도를 구하는 방법은 처리 데이터

중에서 응답자만 발췌하여 분포도를 구하여 무응답처리 하도록 고안된 program으로 다른 조사에서 생산된 분포도는 이용할 수는 없는 한계가 있으니 이용에 주의를 기울이기 바란다.

각 macro의 format과 parameter의 사용방법은 다음과 같다.

Format: %IMPProb(indata=,key=,impvar=,impclass=),

Essential parameters for IMPProb

indata=dataset (관심변수의 무응답 flag인 response='N'가 포함된 데이터임);

key=unique key(조사단위를 확인할 수 있는 ID);

impvar=an imputing variable(imputation하고자 하는 관심변수);

impclass=imputation classes(imputation class 변수들) ;

2.2 Hot Deck method

이 방법은 연속변수 및 불연속변수에 범용적으로 쓰이는 가구관련 통계조사의 무응답 처리기법으로 주로 사용하는 imputation방법인데 이 방법을 이용하는 통계기관마다 사용하고자 하는 목적에 따라 약간 다른 의미로 사용되고 있는데, 여기에 소개된 방법은 가장 전통적은 Hot Deck개념을 적용한 방법으로 사용하고자 한다.

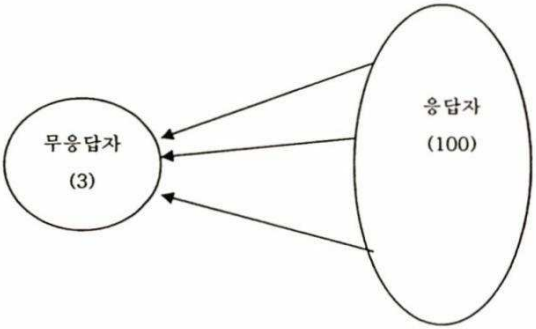
2.2.1 방법 설명

관심변수(imputing variable)X의 응답자들을 한곳에 모아놓고(creation a donor pool) 그 응답자들 가운데서 무응답자수 만큼 무작위로 추출하여 무응답자(receiver) 한명한명에게 무작위로 응답자(donor)를 한명한명 할당하여 응답자의 해당변수의 값을 일대 일로 무응답자에게 할당하는 방법이다.

응답자를 추출하는 방법에 따라서 Random Hot Deck 혹은 Sequential Hot Deck등으로 구분하는데 여기에 소개된 방법은 Random Hot Deck의 개념으로 Hot Deck 방법을 설명 하고자 한다.

2.2.2 응답자 무작위 추출방법

응답자와 무응답자를 관심변수(imputing variable)를 잘 설명할 수 있는 변수(imputation class variable)의 카테고리별로 각각 그룹화하여 응답자 그룹에 각 응답자에게 균등확률분포(uniform distribution)에서 생성된 무작위 확률 값($0<f(x)<1$)을 각각 할당하고, 무응답자 그룹 또한 각 무응답자에게 균등확률분포에서 생성된 무작위 확률 값을 각각 할당하여 이들을 각각 할당 받은 확률 값을 순서로 나열한 다음 무응답자수에 해당하는 개수를 응답자의 상위부터 배열된 순서로 하나하나 할당하는 방법이다.



예를 들면, '학력상태' 변수를 imputation하고자 한다. 이 조사에서 '성별', '연령'이 '학력상태'를 설명할 수 있는 변수라 가정하고, 남자(1)이며 연령이 15-19세 그룹의 응답자가 100명 무응답자가 3명이라면 응답자 100명에 대하여 균등분포에서 생성된 확률 값을 할당 하여 sorting을 하고, 3명도 균등분포에서 생성된 확률 값을 할당 하고 sorting을 하여 다음 테이블과 같이 응답자와 무응답자의 그룹이 확률 값에 따라 정렬되었다고 하자.

응답자 그룹					무응답자 그룹				
ID	Imputation class		학력 상태	확률 값	ID	Imputation class		학력 상태	확률 값
	성별	연령	(x)			성별	연령	(x)	
A11	1	15-19	1	0.002	A51	1	15-19	.	0.541
A31	1	15-19	3	0.012	A19	1	15-19	.	0.620
A22	1	15-19	2	0.035	A33	1	15-19	.	0.655

A54	1	15-19	5	0.120					
A44	1	15-19	2	0.223					
...	...	...	...	...					

응답자를 처음부터 3명을 무응답자에게 차례로 각각 한명씩 할당하는 방법이다.

즉, 응답자 'A11'의 학력상태 '1'은 무응답자 'A51'에게 할당하고, 응답자 'A31'의 학력상태 '3'은 무응답자 'A19'에게 할당하고 마지막으로 응답자 'A22'의 학력상태 '2'은 무응답자 'A33'에게 할당한다.

주의: Imputation class변수들은 응답자와 무응답자가 모두 이용 가능한 변수들(응답 값 혹은 imputed한 값)에 의하여 구분되어질 수 있음.

2.2.3 Imputation class

Imputation class변수는 관심변수 X에 대하여 강한 설명력이 있는 변수이어야 하며 또한 카테고리변수가 주로 이용되고 있다. 연속변수를 사용하고자 하면 이를 카테고리변수로 변형하여 사용하는 것이 바람직하다.

여기서 정의한 Hot Deck방법은 응답자를 without replacement로 사용하기 때문에 한 imputation class에 속하는 응답자수는 항시 무응답자수 보다 커야만 다음 macro program은 작동되도록 설계되어 있으므로 imputation class변수 결정시 사전에 응답자수 및 무응답자수의 크기가 충분한 검토가 선행되어야 한다.

2.2.4 '%HDeck' Program 이용방법

이는 Hot Deck method을 원활히 수행하기 위하여 작성한 SAS macro program이다. 이 macro에 사용될 imputation대상 관심변수(imputing variable)는 연속변수(예: 수입 혹은 나이) 이어야 하는데 그 이유는 응답자들 중에서 이상치(outlier)를 사전에 제거하고자 하는 과정에서 평균 및 표준편차를 계산하는 과정이 program이 내장되어 있기 때문이다.

만약, 카테고리변수를 관심변수로 사용하고자 하면 그 변수를 연속변수로 변환하여 처리하여야 하고 또한 카테고리의 이상치 제거는 사실상 무의미하므로 이를 편법으로 작동치 않도록 하기 위하여 매개변수(parameter) alpha 값을 상당히 크게 (예:alpha=999)주면 이상치는 제거함이 없이 모든 응답자를 imputation class(donor pool)에 남아 있게 하여 작업을 진행할 수 있다.

이상치 제거는 2시그마를 자동으로 부여하였고 만약에 2보다 적은 값의 alpha를 부여하면 많은 이상치 응답자가 imputation class에서 제거되고, 큰 값을 부여하면 상대적으로 적은 숫자의 이상치 응답자가 imputation class에서 제거된다.

각 macro의 format과 parameter의 이용방법은 다음과 같다.

Format: %Hdeck(indata=,key=,impvar=,impclass=,alpha=);

Essential parameters for Hdeck:

indata=dataset (관심변수의 무응답 flag인 response='N'가 포함된 데이터임);

key=unique key(조사단위를 확인할 수 있는 ID);

impvar=an imputing variable(imputation하고자 하는 관심변수);

impclass=imputation classes(imputation class 변수들) ;

alpha=이상치 제거수준(default=2)

2.3 Hierarchical Hot Deck method

필자의 견해로는 이 방법은 Hot Deck method와 Nearest Neighbor method의 개념을 통합하여 응용된 방법으로 Hot Deck method에서 한 카테고리(a donor pool/an imputation class/최종cell)에 무응답자수에 비하여 응답자수가 적을 때 응답자 선택을 비복원(without replacement)으로 수행한다면 카테고리와 카테고리의 통합(collapsing)과정이 필요하게 되는데 주로 카테고리의 통합은 한 class변수를 사전에 적절히 통합하여 하나의 새로운 class변수로 생성한 다음 이들의 순서를 적절히 배열하여 자동적으로 무응답처리를 할 수 있도록 고안된 일종의 Hot Deck imputation 방법이다.

미국 Census Bureau에서는 이 방법을 주로 Hot Deck이란 용어로 사용하고 있다(Donald B.Rubin(1987), Multiple Imputation for Non Response in Surveys, Wiley Series, p157). 어떤 연구 보고서에는 이 방법을 Flexible Matching Imputation method라고도 한다.

Hot Deck method은 만약 한 donor선택을 비복원(without replacement) 방법을 사용한다면 imputation class에 응답자수가 무응답자수보다 커야 하는 제한이 있는 관계로 class변수의 구조와 특성을 사전에 충분히 검토해야 하는데 이 방법은 한 class변수에 속하는 응답자수에 제한이 없으므로 상관관계가 있다고 생각되는 무제한 class변수의 개수를 사용할 수 있는 편리성은 있다. 때문에 가장 많은 변수들의 속성이 유사한 응답자를 선택할 수 있다는 개념으로 Nearest Neighbour Method의 방법을 도입하였다고 할 수 있다. 하지만 class변수의 중요정도에 따라 imputation class형성에 변수의 순서를 부여해야 하는 어려움이 있다.

2.3.1 방법 설명

Imputation작동은 모든 class변수들을 조합하여 형성된 최단 cell로부터 시작하여 무응답자는 응답자의 그룹으로부터 비복원(without replacement) 무작위로 선택된 응답자로부터 관심변수(imputing variable)의 값을 할당 받는다.

만약에 해당 최단cell에 무응답자수가 응답자수보다 많은 경우는 일부 무응답자는 응답자로부터 값을 할당 받지 못하게 될 것이다. 이 문제를 해결하기 위하여 최단에 배치된 한 class변수를 제거하고 나머지 class변수들로 형성된 다른 최단의 cell로부터 무응답자는 응답자의 값을 할당 받는 처리절차를 다시 시작한다. 이때 무응답자에게 한번 기여한 응답자는 제외하고 기여를 기다리는 응답자들만 구성된 새로운 imputation class의 그룹으로 구성한다.

이러한 절차를 계속 반복하게 되면 imputation class의 그룹화에 관계없이 전체 응답자수가 전체 무응답자수 보다 많은 한 어느 단계에서는 모든 무응답자는 응답자로부터 관심변수의 값을 할당 받게 될 것이다. 그러므로 전체 무응답자의 수가 전체 응답자의 수보다 큰 경우에는 이 방법을

적용하는 데는 한계가 있다.

이 방법의 구체적인 예를 들면 imputation class변수가 sex(2개의 카테고리) * education(5개의 카테고리) * marriage(4개의 카테고리)로 구성되었다면 imputation class로 형성된 최단cell수는 최대한 40개, 이 40개의 각각 cell에서 무응답자가 응답자로부터 값을 할당 받는 작업을 시작한다. 만약 일부 무응답자가 값을 할당 받지 못하면 다음 단계에서는 최단 class변수인 marriage인 class변수를 제거하여 형성된 새로운 sex(2)*education(5)의 imputation class 10개의 cell에서 무응답자는 각각 값을 할당 받는 작업을 하게 된다. 이러한 class변수 제거 반복 작업을 모든 무응답자가 값을 할당 받을 때 까지 수행한다.

Level의 정의: Imputation class변수들의 결합으로 생성된 층인데 이들에 대하여 각 층별로 부여된 단계별 imputation class의 이름을 level이라 정의한다. 예를 들면 앞의 예에서 3개의 변수로 구성된 imputation class는 sex*education*marriage로서 level1은 sex변수만의 층, level2는 sex*education변수의 결합된 층, level3는 sex*education*marriage의 변수의 결합된 층들을 level로 지칭한다.

어떻게 Hierarchical Hot Deck이 작용하느냐를 정의된 level의 개념을 도입하여 좀더 구체적인 예를 들어보면 다음과 같다 (도표참조).

- level3에서 sex(1)*education(대졸)*marriage(미혼)인 한 cell에서 2개의 무응답단위는 2개의 응답단위로부터 무작위로 무 응답된 항목(들)의 값을 할당 받는다. 2개의 무응답 단위는 여전히 응답자의 값을 할당 받기를 기다리고 있다.

- Level3에서 sex(1)*education(대졸)*marriage(기혼)인 한 cell에서 2개의 무응답 단위는 8개의 응답 단위로부터 무작위로 2개의 단위를 선택하여 무 응답된 항목(들)의 값을 할당 받는다. Level3에서 이 cell에 대한 무 응답할당 처리 절차가 종료된다.

- Level2에서 sex(1)*education(대졸)인 새로운 한 cell에서는 할당을 한 4개의 응답단위를 제외한 6개의 응답단위가 무응답 단위를 기다리는 그룹으로 구성되어 있고 2개의 무응답단위는

응답단위로부터 값을 기증을 기다리고 있다. 여기서 2개의 무응답 단위는 6개의 응답단위로부터 무작위로 2개의 단위를 선택하여 무응답된 항목(들)의 값을 할당 받는다. Level2에서 2개의 잔여 무응답단위의 값 할당 처리절차가 종료된다. 이로서 모든 무응답 처리절차가 마무리된다.

Level1	Level2	Level3	응답자 수	무응답자 수
Sex	Education	Marriage		
1	대졸	미혼	2	4
1	대졸	기혼	8	2

2.3.2 Imputation class

Imputation class변수는 관심변수(imputing variable)X와 상관관계가 높은 카테고리 변수들을 사용해야 하고 상관이 가장 높은 변수는 상위 class(예:level1)로 시작하고 상관이 가장 약한 변수는 하위 class(예:level5)로 지정하여 최단 class변수 제거가 필요한 경우 가장 먼저 상관이 가장 약한 변수를 제거대상이 되게 함으로서 가장 강한 상관관계의 변수는 최후까지 imputation class cell형성에 기여 하도록 한다. 이 class변수의 level순서의 결정은 통계적인 방법을 사용하기는 용이하지 않으므로 관심변수와 class변수에 상당한 지식이 있는 subject matter 전문가와 상의하여 순서를 결정하기를 권하고자 한다.

Imputation class변수는 동일한 변수에서도 통합된 카테고리의 정도에 따라 순서를 달리 부여하여 같은 속성 class변수 내에서도 imputation class cell의 통합(collapsing)절차를 자동으로 처리하도록 할 수도 있다. 예를 들면 imputation class인 sex * age1(각 세 연령그룹)에서 연령변수를 즉시 제거한다면 연령에 대한 모든 기여정보를 상실하게 되므로 연령 class변수를 제거하기 중간단계에 새로운 age5(5세 연령그룹) 카테고리 변수를 형성하여 level3의 imputation class를 sex*age5*age1로 형성한다면 age1의 class변수를 제거하여도 연령의 class변수는 imputation class형성에 기여를 할 수 있다. 이러한 동일 class변수의 변환으로 imputation class cell 통합(collapsing)을 자동으로 이루어 지도록 imputation class를 설계 할 수도 있다.

2.3.3 이상치 제거 class(Outlier class)

Hot Deck method에서는 응답자를 할당 받는 최단 imputation class cell에서 응답자 할당작업이 수행되기 이전에 이상치 제거작업을 수행하도록 설계되었다. 이 방법에서도 최단 imputation class cell에서 이상치를 제거한다면 너무 상세한 최단의 imputation class cell이어서 이상치를 통계적으로 제거하기에는 응답자수가 충분하지 않을 가능성이 많다. 이러한 관계로 이상치 제거는 응답자 전체중에서 관심 있는 class변수들로 새롭게 imputation class 대신에 outlier class를 형성하여 그 형성된 최단의 cell로부터 이상치를 제거토록 한다.

이상치 제거를 목적으로 하는 outlier class 형성은 주로 분석하고자 하는 관심 있는 cross table에 사용하는 주요 보조변수를 사용함이 바람직 하다고 생각된다. 이는 imputation으로 생성된 자료가 후에 cross table에 예상치 않았던 이상한 값으로 출현될 가능성을 사전에 예방하고자 하는 의도에서 비롯된 것이다.

2.3.4 '%HHDeck' program 이용방법

Hierarchical Hot Deck method을 원활히 수행하기 위하여 작성한 SAS macro program이다. 이 macro에 사용될 imputation대상 관심변수(imputing variable)는 연속변수(continuous variable, 예: 수입 혹은 나이) 이어야 운영할 수 있도록 program이 설계되었다. 그러므로 카테고리변수와 같은 불연속변수를 imputation 관심변수(imputing variable)로 사용하고자 하면 그 변수를 연속변수로 변환하여 처리 할 수 있다.

카테고리변수에 대하여 이상치 제거는 작업진행상 무의미하므로 이상치 제거 절차의 작동을 방지하기 위하여 매개변수(parameter)의 alpha 값을 상당히 크게 (예:alpha=999)주면 이상치 제거 절차 없이 모든 응답자를 donor pool에 남아 있게 할 수 있다.

각 format과 macro parameter의 이용방법은 다음과 같다.

Format: %HHDeck(indata=,key=,impvar=,noclass=,varlist=,OutlierClass=,alpha=);

Essential parameters for HHdeck:

indata=dataset (관심변수의 무응답 flag인 response='N'가 포함된 데이터임);

key=unique key(조사단위를 확인할 수 있는 ID);

impvar=an imputing variable(imputation하고자 하는 관심변수);

noclass=number(class 변수들의 갯 수)

varlist=class variables (class변수들: 변수들은 동일한 data형태임, 모두 수치 혹은 문자)

outlierclass=class variables(관심 있는 이상치 제거 그룹)

alpha=이상치 제거수준(default=2)

2.4 이용방법 예제

```
*다음 %inc 부분은 반듯이 이용자 환경에 맞게 프로그램을 복사하여야 한다.
%inc 'c:₩2005censusimputation₩programs₩IMPMethodWizardVersion2.sas';
*********************************************;
** need to set response flag before processing ;
*********************************************;
*---IMPProb-----------------;
** setting response or nonresponse;
data master; set master;
     PersonID=GI_sido||GI_sigun||GI_dong||GI_josag||GEO_no||GA_no||GA_wonno;
     response='Y';
     if NR_IN_sex='N' then response='N';
run;
%IMPProb(indata=master,
         key=PersonID,
         impvar=IN_sex,
         impclass=GI_sido GI_sigun);

*************************************************************;
```

```
** An imputing variable for Hdeck and HHdeck should be numeric;
** All imputation class variables should be same data format like numeric or alphanumeric;
*************************************************************;

*---Hdeck-------------------;
data master; set master;
      response='Y';
      if NR_IN_sex='N' then response='N';
      var=1*IN_sex;
run;
%Hdeck(indata=master,
        key=PersonID,
        impvar=var,  /* var은 alpha numeric인 IN_sex로부터 변환된 numeric변수 */
        impclass=GI_sido GI_sigun GI_dong,
        alpha=999);

*---HHdeck-------------------;
 data master; set master;
       response='Y';
       if NR_IN_age1='N' then response='N';
 %HHdeck(indata=master,
          key=PersonID,
          impvar=IN_age1,  /* IN_age1은 numeric 변수 */
          noclass=6,
          varlist=GI_sido GI_sigun GI_dong GI_josag IN_mar IN_edu),
          outlierclass=IN_sex,
          alpha=2);
 run;
```

2.5 결과표 해석 및 program source

결과표를 이용하는 방법은 부록[1]에 정리하여 놓았고 program source는 부록[2]에 수록하였으니 프로그램에 관심 있는 독자는 한번 검토하여 더욱

진보된 기능을 첨가한 프로그램으로 발전하였으면 하는 바람입니다.

3. 사업체조사에 주로 쓰이는 방법 소개

이 장에서는 사업체관련 항목 무응답처리방법 개발에 도움이 되었으면 하는 기대에서 현재 뉴질랜드통계청에서 경상사업체 조사에 주로 쓰이는 방법을 간단히 예제 중심으로 소개하고자 한다.

여기에 사용한 방법들은 가구조사에서 사용한 방법과 같이 Statistics New Zealand에는 정형화된 SAS program은 개발된 것은 없고 사업체조사 자료 입력시스템에서 사용한 Lotus Notes 언어로 imputation 방법을 구현하여 사용하고 있으므로 앞으로 소개할 imputation방법에 대한 SAS program은 독자의 몫으로 남기고자 한다.

3.1 Historical method

이 방법은 주로 표본으로 선정된 동일사업체를 반복하여 조사하는 월,분기, 년간 등 경상조사 과정에서 발생하는 무응답 사업체에 적용하는 방법으로 동일사업체의 응답된 과거자료를 이용하여 현재의 조사시점에 발생된 무응답을 imputation 하는 수단으로 사용하는 방법이다.

이는 두 가지 방법으로 구분되는데 하나는 과거시점의 응답된 자료를 아무 변형 없이 그대로 현재 조사시점에 무 응답된 항목의 값으로 사용하는 (Historical method without FMF) 것이고 다른 한 방법은 과거와 현재 조사시점사이에 항목의 시계열변화율을 적용하여 무 응답한 항목의 값을 추정하는 방법을 사용하는(Historical method with FMF) 것이다.

설명의 편의를 도모하기 위하여 다음과 같이 기호를 정의하고자 한다.

t = 현 조사시점, $t-p$ = 전 조사시점

$x_{i,t}$ = i사업체의 t시점에 응답한값

$w_{i,t}$ = i사업체의 t시점에 사용한 승수

$x_{i,t-p}$ = i사업체의 $t-p$시점에 응답한값

$w_{i,t-p}$ = i사업체의 $t-p$시점에 사용한 승수

c = *imputation class(cell)*

$$FMF_c(Forward\ Movement\ Factor) = \frac{\sum_i x_{c,i,t} * w_{i,t}}{\sum_i x_{c,i,t-p} * w_{i,t-p}}$$

여기서 p = 1,2,3,4,5.....

3.1.1 Historical method without FMF

이는 현 조사시점의 무응답 항목을 과거조사의 응답한 항목의 값을 그대로 사용하는 방법이다. 즉, 무응답한 현 조사시점의 i사업체의 값 : $\hat{x}_{i,t} = x_{i,t-p}$

3.1.2 Historical method with FMF

이 방법은 historical method without FMF 보다는 경상 사업체 표본조사에서 범용적으로 사용되는 방법으로 알려져 있다. 무 응답한 현재 조사시점의 값을 imputation하는 절차로서 먼저 FMF를 구하고 이 FMF를 응답한 전조사 시점의 항목의 값에다 곱하여 값을 구하여 imputation을 할 수 있다. 즉,

무응답한 현 조사시점의 i사업체의 값 : $\hat{x}_{c,i,t} = x_{c,i,t-p} * \mathrm{FMF}_c$

여기서 FMF계산에 투입된 항목의 값은 두 조사기간에 모두 응답한 표본 사업체들만 사용하는 것이 관례인데 경우에 따라서는 전 조사시점에 imputation한 값을 사용할 수도 있다. 따라서 현 조사시점에 새로 표본에 반영된 신규사업체는 과거자료가 없으므로 계산에 제외하는 것이 당연하다고 생각된다.

Imputation class(cell)는 사업체들의 특성상 매우 동일하다고 판단되는 사업체들의 그룹으로 형성하도록 한다. 예를 들면 동일한 산업분류 혹은 동일한 종사자규모의 사업체로 구성된 표본 층으로 imputation class를 구성하여 과거와 현재 조사시점 사이에 관심항목에 관하여 시계열증감 추세의 변화가 잘 반영이 되도록 class를 정의할 필요가 있다.

경우에 따라서는 두 조사시점을 모두 응답한 사업체라 하더라도 FMF계산 과정에서 제외 시키는 사업체가 있을 수 있다. 두 조사시점에 너무도 응답값의 차이가 많이 나는 대상 사업체는 FMF를 상당히 크게(혹은 작게) 하는데 큰 기여를 하기 때문이다. 예를 들면 계절 사업체를 그 한가지 예로 한 시점에는 미미한 판매액이고 어느 시점에는 계절성으로 인하여 상당한 판매액을 응답하는 사업체라면 이런 사업체는 그 산업분류(imputation class)에 일반적인 증감추세에 대하여 큰 변화를 기여하기 때문이다.

어떤 수준의 FMF의 값이 이상적일까? 조사종류별로 차이가 있겠지만 뉴질랜드 통계청의 연간 사업체 조사(Annual Enterprise Survey)를 예를 들자면 여러 해의 과거자료 연구를 거쳐서 FMF의 값의 한계를 0.667<=FMF<=1.5로 사용하고 있다. 이는 FMF 계산결과가 1.5이상이 되면 이를 1.5로 처리하고 0.667이하이면 이를 0.667로 고정하여 사용하겠다고 내부적으로 정의한 규칙이다. 따라서 FMF 한계의 결정은 조사별로 연구를 하여 사용하는 것이 바람직하다고 생각되어진다.

FMF 계산 및 무응답 값 추정 예제

어떤 imputation class(cell) 'c'는 3개의 사업체로 구성되었다고 정의하고 그 사업체중 'CC' 사업체가 현재 조사시점만 무 응답 하였다고 가정하고 전 조사시점의 자료가 이용 가능하다면,

사업체	전 조사시점			현 조사시점		
	판매액($x_{t-1,i}$)	승수($w_{t-1,i}$)	추정치(x*w)	판매액($x_{t,i}$)	승수($w_{t,i}$)	추정(x*w)
AA	100	5	500	200	5	1000
BB	100	5	500	200	5	1000
CC	200	5	1000	.	5	.

BB (NR) ' ' 300 5 1500

FMF는 $FMF = \left(\frac{1000+1000}{500+500}\right) = 2$ 와 같이 계산 될 수 있을 것이다. 따라서 사업체 'CC'의 현 조사시점의 무 응답한 판매액의 값은 400=200*2로 imputation할 값을 추정치할 수 있을 것이다.

3.2 Regression method

이 방법은 때로는 ratio method 라고도 한다. Imputation하고자 하는 관심변수와 강한 상관관계가 있는 보조변수(auxiliary variable)가 있는 경우에 이 방법을 이용할 수 있다. Imputation에 관심 있는 변수는 무응답으로 조사되고 보조변수가 동일조사내의 응답된 항목, 이용 가능한 행정자료 혹은 표본을 설계한 모집단등에서 이용 가능하다면 이들의 상관관계를 이용하여 무응답한 변수에 대하여 무 응답한 값을 추정하여 imputation할 수 있을 것이다.

뉴질랜드 통계청에서는 국세청의 세무자료를 공유하기 때문에 이를 보조정보로 활용하여 사업체조사의 무응답 항목을 ratio method를 이용하여 무응답항목의 추정 값을 도출하는데 이를 연간 사업체 조사 혹은 경상 사업체조사에 주로 이용된다.

한국 통계청에서는 기초통계조사에서 조사되는 종사자수 혹은 판매액이 이용될 가능성이 있는 보조정보 이므로 연구할 필요가 있을 것으로 생각 되어진다.

다음의 기호는 설명의 편의를 위하여 다음과 같이 정의 하고자 한다.

$x_i = i$사업체의 추정하고자 하는 변수에 응답한값
$y_i = i$사업체의 보조변수의 값(동일조사혹은다른source)
$w_i = i$사업체에 승수
$c = imputation\ class(cell)$

$$\hat{r}_c(\text{Regression Factor}) = \frac{\sum_i x_{c,i} * w_i}{\sum y_{c,i} * w_i}$$

무 응답한 항목의 값을 imputation하는 방법은 다음과 같다.

무응답한 i사업체의 값: $\hat{x}_{c,i} = y_{c,i} * \hat{r}_c$

여기서 보조변수의 값이 '0'인 혹은 무 응답한 사업체는 계산절차에서 제외하는 것이 내부규칙으로 되어있다.

계산절차 예제

어떤 imputation class(cell) 'c'는 4개의 표본사업체로 구성되었다고 정의하고 그 표본 사업체중 'DD' 사업체가 현 조사 시점에 판매액을 무 응답 하였고 보조변수만 응답하여 이용가능 하다고 가정하면,

사업체	판매액(x_i)	승수(w_i)	추정치($x*w$)	보조변수(y_i)
AA	100	5	500	50
BB	100	5	500	48
CC	200	5	1000	105
DD	.	5	.	98

먼저 Regression factor, $\hat{r}_c = \left(\frac{100*5+100*5+200*5}{50*5+48*5+105*5}\right) = \frac{2000}{1015} = 1.97$ 로 계산될 것이다. 따라서 사업체 'DD'의 현 조사시점의 무 응답한 판매액의 값은 193=1.97*98 로 imputation할 값을 추정할 수 있을 것이다.

3.3 Mean method

이 방법은 유사한 표본 사업체들을 한 그룹으로 구성하여(imputation class) 무 응답한 표본 사업체들에 대해서는 형성된 그룹에서 응답자들로부터 계산된 단순평균값 혹은 가중평균값을 각각 무응답자에게 할당하는 방법이다.

신규사업체가 금번 조사시점에 처음 표본으로 추출되어 조사에 투입 하였지만 신규 사업체로부터 무응답이 발생하여 과거 조사시점 자료가 존재하지 않고 다른 source의 보조정보가 이용가능하지 않는 경우에 주로 사용하는 방법으로 알려져 있다.

Mean method는 사업체 조사의 무응답처리 방법가운데 마지막으로 선택되는 처리기법으로 알려져 있다. 그 이유가운데 하나가 표본오차를 과소 추정케 하는 단점이 있기 때문이다.

$x_i = i$사업체의 추정하고자 하는 변수에 응답한값

$w_i = i$사업체에 승수

$c = imputation\ class(cell)$

$$\hat{\bar{x}}_c(\text{평균값}) = \frac{\sum_i x_{c,i} * w_i}{\sum_i w_i}$$

무 응답한 항목의 값을 imputation하는 방법은 다음과 같다.

무응답한 i사업체의 값 : $\hat{x}_{c,i} = \hat{\bar{x}}_c$

계산절차 예제

보편적으로 쓰이는 가중평균법을 예시로 들고자 한다.

어떤 imputation class(cell) 'c' 에 4개의 표본사업체로 구성되었다고 정의하고 그 사업체중 사업체 'CC' 가 현 조사시점에 처음 표본 사업체로 선택되었는데 무 응답 하였다고 가정하면,

	전 조사기간		현 조사기간		
사업체	판매액(x_i)	승수(w_i)	판매액(x_i)	승수(w_i)	추정치(x_i w_i)
AA	100	5	200	5	1000
BB	100	5	200	5	1000
CC	.	.	.	5	
DD	.	.	200	5	1000

Imputation class 'c' 에 대한 가중평균 $\hat{x}_c = \left(\frac{1000+1000+1000}{5+5+5}\right) = 200$ 이므로 무응답 사업체 'CC'의 현 조사시점의 판매액은 200으로 imputation할 값을 구할 수 있을 것이다.

주의: Mean imputation은 가중치를 조정하는 방법(weighting up)과 동일한 결과를 가져온다. 위의 예제에서 응답한 AA, BB, DD사업체를 새로 조정된 가중치 $\hat{w}_c = \left(\frac{5+5+5}{3}\right) = 6.6667$ 을 이용하여 추정한 총계 추정치는 동일한 결과를 가져오는데 이 가중치 조정방법을 사용하는 데는 주의가 필요로 한다.

한 변수를 취급하는 조사의 추정에는 문제가 없는데 만약 여러 개의 변수를 조사하는 사업체조사에서 여러 항목 무응답을 동시에 처리해야 하는 상황에는 모든 항목에 대하여 조정된 가중치를 각각 계산하여야 하는 절차상 복잡함이 따르기 때문에 이 방법을 사용하지않고 각 변수에 대한 mean imputation방법을 사용하는 것이다.

3.4 Moving average method

이 방법은 현재 뉴질랜드 통계청의 경상사업체조사에서 사용하지 않는 방법이지만 필자가 연구한 바에 의하면 historical method와 거의 유사한 정확성을 가져오는 연구결과를 가져왔기 때문에 여기서 간단히 소개하고자 한다.

이는 단순한 시계열 이동평균방법으로 조사된 과거시계열자료를 이용하여 3개월 혹은 3분기 응답자료를 가지고 단순이동평균을 구하여 현재의 조사시점에 무 응답한 값을 추정하여 이용하는 방법이다.

3개월이동평균 : $\hat{x}_t = \left(\frac{x_{t-3} + x_{t-2} + x_{t-1}}{3}\right)$

4. Imputation방법 결정에 고려사항

어떤 imputation방법을 선택해야 하는 전통적인 선택의 기준은 없고 조사특성과 변수성격에 따라 적절히 선택하여야 하는데 어떤 방법을 선택하든 조사집단의 분포와 분산을 유지하는 것을 대원칙으로 하여야 한다. 무응답 집단의 성격도 아울러 고려하여 이들이 조사집단에 어느 정도의 영향을 미치는지도 사전에 검토하여 방법결정에 반영토록 하여야 한다.

4.1 어떤 imputation방법을 사용할 것 인가를 결정하기는 용이하지는 않다. 조사의 특성과 관심변수의 성격에 따라 여러 방법들을 적절히 혼합하여 사용하는 것이 바람직하다.

4.2 사업체 조사에서도 가구부문에서 소개한 Hot Deck 및 Hierarchical Hot Deck방법을 사용할 수 없을까? 하는 의문이 제기되는데 의문의 여지도 없이 사업체조사에서도 이들 방법을 사용할 수 있다. 단지 조사종류의 특성에 따라 사용할 여부를 판단하여야 할 문제라 생각되어진다.

4.3 경상사업체조사(월간, 분기 등)에 주로 사용하는 방법으로는 대표적으로 ① historical method를 추천 할 수 있고, 추정하고자 하는 변수에 대하여 이용 가능한 보조변수가 상관관계가 강하다면 ② regression(ratio) method도 권하고자 한다. 관심변수의 특성에 따라 두 가지 방법을 혼용하여 사용할 것을 권한다. 그리고 이 두 가지 방법이 적용 불가능할 때 ③ mean method를 최후로 선택 하였으면 한다.

4.4 Imputation방법 개발단계에서는 반듯이 subject matter전문가들과 프로그램개발 요원들과도 반듯이 방법설정단계에서 토의하여야 한다. 이는 방법 구현에 예상치 못한 문제를 사전에 대처할 수 있기 때문이다. 특히 시스템개발 복잡성을 수반하는 훌륭한 imputation방법을 선택하는데 있어서는 비용과 질적인 자료생산에 득실을 고려하여 최종방법을 결정해야 한다.

4.5 무응답 단위에 대한 것도 방법결정에 아울러 고려 해야 한다. Unit non response 와 item non response에 대한 imputation방법이 다르게 선택되어야 할 경우도 발생하기 때문이다.

4.6 Imputation 방법결정에 가장 고려해야 할 사항 가운데 하나는 방법이 간단하여 누구나 이해할 수 있고 실행하기위한 시스템개발이 용이 해야 한다. 복잡한 방법을 도입하여 imputation system tool을 성역화 시키고 일반 이용자들에게는 black box취급을 받으면 imputation방법 발전에는 기여를 할 수 없기 때문이다.

4.7 무응답형태도 방법 결정에 고려해야 한다. 만약 무응답이 대체로 완

전 무작위(MCAR: Missing Complete At Random)로 발생하였다면 어떤 방법을 적용하여도 문제가 없을 것으로 생각되어진다. 하지만 무응답이 어떤 변수에 종속되어 무작위(MAR: Missing At Random)로 발생하였다면 이는 관련변수를 연관시키는 방법이 결정되어야 할 것이다.

5. Imputation class 결정 고려사항

Imputation class는 앞에서 언급한 바와 같이 여러 다른 용어로 사용하고 있다. 예를 들면 imputation cell, imputation pool, donor pool등으로 상황에 따라 미미한 의미차이가 있다. 무응답 관련 책자와 학술지를 읽을 때 그 의미를 정확히 상황에 맞게 이해하여야 할 것으로 생각한다. 그러면 imputation class 변수는 어떻게 결정하고 어떠한 방법이 imputation class 변수들의 순서를 결정하는데 주로 쓰이는지를 간단히 소개 하고자 한다.

5.1 Imputation class변수 발굴은 반듯이 subject matter 전문가와 협의하여 관심변수(imputing variable)와 관련이 있는 모든 변수를 동일조사 내의 관련조사항목 혹은 다른 source의 유사변수 중에서 발굴하도록 한다.

5.2 Imputation class변수들을 조합하여 형성되는 한 카테고리에 응답자수 및 무응답자수를 고려하여야 한다. 만약 Hot Deck방법이 사용되고 응답자선택(donor selection)이 비복원(without replacement)으로 이루어 진다면 반듯이 한 카테고리에 응답자수가 무응답자수 보다는 커야 한다. 그렇지 않으면 카테고리와 카테고리의 통합(collapsing)전략을 사전에 수립하여야 한다.

5.3 Hierarchical Hot Deck 방법을 사용할 경우는 중요 imputation class 변수를 상위 level로 시작하고 덜 중요하다고 생각되는 imputation class 변수를 마지막 level로 순서를 배열하여야 한다. 중요 imputation class 변수를 최후까지 class형성에 기여토록 하기 위함이다.

5.4 Imputation class변수를 결정하는데 사용되는 어떤 정형화된 통계적인 분석방법은 소개되어 있지 않다. 관심변수(imputing variable)와 반듯

이 상관관계가 있는 보조변수를 선택하는 것은 필수 사항이다. 그러면 모든 대상 대상변수를 class변수로 사용할 것인가 하는 문제에 봉착하는데 이를 선택하는 주된 기준은 어떤 변수를 주로 통계표생산에 카테고리변수로 이용하여 통계표를 작성하는가도 class변수 선택의 한가지 기준이다.

5.5 필자가 뉴질랜드 통계청에서 결정하던 한가지 방법은 tree analysis의 일종인 SAS macro procedure인 CHAID analysis를 통하여 중요변수 순서정도를 대략 결정한 다음 그 변수들을 나름대로 조합하여 여러 형태의 class변수를 구성한 다음 간단한 simulation을 한 뒤에 가장 적합하다고 생각되는 변수의 조합을 imputation class변수로 결정한 경험이 있으니 관심 있는 독자는 한번 시도해볼 가치가 있다고 생각된다.

6. 종단면조사(Longitudinal Survey)에 쓰이는 방법 소개

종단면 조사에서 쓰이는 방법에 대해서는 구체적인 설명은 하지 않고 방법들 용어만 소개하고자 한다. 방법면에서는 앞에서 소개한 방법과 동일 및 거의 유사하고 단지 종단면조사에서 적용하는 용어로만 정의 된 것이 주를 이루고 있으니, 자세한 내용은 독자에게 남기고자 한다. 다음에 소개된 방법들의 용어는 뉴질랜드 통계청 종단면가구조사 SoFIE(Survey of Family Income and Employment) 개발당시 검토 되었던 용어들 인데 실제로 적용한 방법은 앞 절에서 소개된 가구조사에 주로 쓰여진 세가지 방법(Probability, Hot Deck, Hierarchical Hot Deck)이다.

- Mean imputation method
- Modal response imputation method
- Hot deck imputation method
- Flexible matching imputation method
- Carry-over imputation method
- Carry-over with random R imputation method
- Carry-over with population R imputation method
- Regression imputation method
- Predictive mean imputation method
- Little & Su imputation method

- Use of retrospective and proxy data

노트: 위의 방법들 용어는 Statistics New Zealand에 근무하던 필자의 동료인 Marry-Anne Heys씨가 종단면조사(SoFIE)의 imputation기법개발에 관한 연구보고서에서 발췌한 용어들이니 관심 있는 독자는 부록[3]을 참조 하시기 바랍니다.

부 록

1. Imputation 방법별 결과진단을 위한 Outputs

1.1 Probability method output

1. Producing random number boundaries: 분포도에 관련된 output
2. Finding imputed value: 개인 ID별 imputed 값
3. Imputation summary: Imputation한 결과에 대한 요약표임.

IMPProb outputs

1. Producing random number boundaries

GI_sido	GI_sigun	(A) IN_sex	(B) Responding units	(C) Distribution	(D) Random number boundary
34	010	1	14993	0.49978	0<Random=<0.4998
34	010	2	15006	0.50022	0.4998<Random<1
			==========		
			29999		

2. Finding imputed value

Obs	GI_sido	GI_sigun	(A) ID	(B) Random number	(C) Random number boundary	(D) imputed value
1	34	010	9912	0.84414	0.4998<Random<1	2

3. Imputation summary: imputing variable=IN_sex impclass=GI_sido GI_sigun

Obs	GI_sido	GI_sigun	(A) Total	(B) Response	(C) Nonresponse	(D) Imputed value	(E) NotFound imputed
1	34	010	30000	29999	1	1	0
			=====	========	======	=======	========
			30000	29999	1	1	0

1.2 Hot Deck method output

1. Summary of response and non-response: 이상치를 제외한 응답자수, 무응답자수 및 응답률을 보여주는 결과표. 응답자수의 충분여부를 진단할 수 있음. 응답자수가 무응답자수 보다 많은지 적은지를 진단하는 단순 참고 지표임.
2. Does system find all donors? : Imputation한 결과와, 성공여부를 판단할 수 있는 지표임.

HDeck outputs

1.Summary of response and non-response

Gl_sido	Gl_sigun	Gl_dong	(A) Total sample	(B) Outliers	(C) Responses (excldуding outliers)	(D) NonResponses	(E) Resp_Rate (%)	(F) (Resp_Rate>40% & Rs>=10 units)
34	010	11	24269	12096	12172	1	100	Yes
34	010	12	5731	2834	2897	.	100	Yes
			30000	14930	15069	1		

2.Does system find all donors?

Gl_sido	Gl_sigun	Gl_dong	(A) Donors	(B) NonResponse	(C) No of donors	(D) Find donors?
34	010	11	12172	1	1	Yes, found all donors
34	010	12	2897	.	.	Yes, found all donors
			15069	1	1	

1.3 Hierarchical Hot Deck method output

1. Summary of response and non-response: Hot Deck의 1번 리스트와 동일한데 여기서는 이상치 제거 class를 별도로 지정한 관계로 응답자 무응답자수의 크기가 실제 imputation진행과정에 문제성이 있는지를 진단할 수 있는 결과표임.
2. Donors information: 각 계층별로 무응답자가 취한 donor의 수를 요약한 표임. 가장 이상적인 donor의 발견은 높은 level(예: level 6)에서 많은 donor를 발견한 것 임.

HHDeck outputs

1.Summary of response and non-response

IN_sex	(A) Total sample	(B) Outliers	(C) Responses (exclduding outliers)	(D) NonResponses	(E) Resp_Rate (%)	(F) (Resp_Rate>40% & Rs>=10 units)
1	14993	5306	8958	729	92.5	Yes
2	15007	5545	8689	773	91.8	Yes
	30000	10851	17647	1502		

2.Donor information

(A) Find from	(B) Class list	(C) Donors
		1502
Level_01	GI_sido	.
Level_02	GI_sido GI_sigun	.
Level_03	GI_sido GI_sigun GI_dong	.
Level_04	GI_sido GI_sigun GI_dong GI_josag	113
Level_05	GI_sido GI_sigun GI_dong GI_josag IN_mar	336
Level_06	GI_sido GI_sigun GI_dong GI_josag IN_mar IN_edu	1053

2. Program Source: ImpMethodWizardVerson2.sas

```
**********************;
* The ImpMethodWizard.sas developed by Soon Song and revised it with Version2
  based on original program. Recommend to use the revised program;

**********************;

*%HHDeck(indata=,key=,impvar=,noclass=,varlist=,OutlierClass=,alpha=);
*-------------------------------------------------;
*Essential parameters for HHdeck:
*     Indata=original dataset for imputing process;
*     key= unique key for sample unit identification;
*     impvar=an imputing variable;
*     noclass=number of class for class variables;
*              note: variables should be same data type like integer or alphanumeric;
*     varlist=class variables :
*     outlierclass=class variables you want to delete;
*     alpha=deleting point for outliers;
*-------------------------------------------------;


*-------------------------------------------------;
*%IMPProb(indata=,key=,impvar=,impclass=);
*-------------------------------------------------;
*Essential parameters for IMPProb:
*     indata=dataset for imputing process;
*     key=unique key for sample unit identification;
*     impvar=an imputing variable;
*     impclass=probability calculaition class;
*-------------------------------------------------;

*-------------------------------------------------;
*%Hdeck(indata=,key=,impvar=,impclass=,alpha=);
*-------------------------------------------------;
*Essential parameters for Hdeck:
*     indata=dataset for imputing process;
*     key=unique key for sample unit identification;
*     impvar=an imputing variable;
*     impclass=class variable list;
*     alpha=deleting point for outliers;
*-------------------------------------------------;
```

```
*********************************;
*          COMMON ROUTINES      ;
*********************************;

*-------------------------------------------------------------------------
--;
* Outlier processing from responding group:
*-------------------------------------------------------------------------
--;
%MACRO Outlier(class=,alpha=);
   *---------------------------------------;
   ** finding outliers:
   *---------------------------------------;
   proc summary nway missing data=_resp;
        class &Class;
        var var;
        output out=_total mean=mean std=std;

   proc sort data=_resp;by &Class;
   run;
   data _resp _outlier;merge _resp _total; by &Class;
        if std>0 then d=abs(var-mean)/std;
        big=2.0;
        If &alpha>0 then big=&alpha;
        if d>big then
            do;output _outlier;
               delete;
            end;
        output _resp;
        drop d big mean _freq_ std _type_;
   run;

   *-----------------------------------------------;
   *     summary of response nonresponse & outlier;
   *-----------------------------------------------;
   proc summary data=_resp nway missing;
        class &class;
        output out=_respsum(rename=(_freq_=Resp));
   run;
   proc summary data=_nonresp nway missing;
        class &class;
        output out=_nonrespsum(rename=(_freq_=NonResp));
   run;
   proc summary data=_outlier nway missing;
```

```
        class &class;
        output out=_outliersum(rename=(_freq_=outliers));
    run;
    data _sum; merge _respsum _nonrespsum _outliersum;by &class;
        prop=round(100*resp/sum(resp,nonresp),.1);
        donor='No ';
        if resp>=10 and prop>40   then donor='Yes';
        total=sum(nonresp,resp,outliers);
    run;
    proc print data=_sum noobs split='*';
        title1 '1.Summary of response and non-response';
        var &class total outliers Resp NonResp prop donor;
        sum total outliers resp nonresp;
        label total='(A)*Total*sample' outliers='(B)*Outliers'
              resp='(C)*Responses*(excldуding*outliers)' nonresp='(D)*NonResponses'
              prop='(E)*Resp_Rate*(%)'
              donor='(F)*(Resp_Rate>40% *& Rs>=10 units)';
    run;
%MEND Outlier;

*-------------------------------------------------;
 Data to separate between respons and non-response;
 -------------------------------------------------;
 %MACRO Resp_NonResp(indata=);
    *-------------------------------;
    *-- Two working datasets
          1. resp=response group
          2. nonresp=non-response group;
    *-------------------------------;
    data _nonresp _resp; set &indata;
         response=upcase(response);
         if response='Y' then output _resp;
                         else output _nonresp;

    run;
%MEND Resp_NonResp;

*-------------------------------------------------------------------------
--;
*Delete working datasets to avoid any side effects for the next process ;
*-------------------------------------------------------------------------
--;
%MACRO Datasets(delete=);
    proc datasets lib=work nolist;
         delete &delete;
    run;
```

```
%MEND Datasets;
*********************************;
*------END COMMON SECTION      ;
*********************************;




*********************************;
*          SECTION A            ;
*********************************;
*-------------------------------------------------------------------------
-;
* Macro moudle for hierarchical hot deck method                          ;
*        Alias name=record matching or flexible matching method          ;
*-------------------------------------------------------------------------
-;

*-------------------------------;
*Matching process sub routiones;
*-------------------------------;
%MACRO MATCH(no=,last=);
   %MACRO impclass(number);
        %do x=0 %to &number %by 1;
            class&x
        %end;
   %MEND impclass;
   data _nonresp; set _nonresp;
        random=ranuni(&no);
   data _resp; set _resp;
        random=ranuni(&no);
   run;
   proc sort data=_nonresp;by %impclass(&no) random;
   proc sort data=_resp;by %impclass(&no) random;
   run;
   *-----------------------------------;
   *Finding most matching records      ;
   *-----------------------------------;
   data _match; merge _nonresp(in=a) _resp(in=b); by %impclass(&no);
        if a & b;
        keep %impclass(&no);
   run;
   proc sort data=_match nodupkey;by %impclass(&no);
   *---------------;
   * to find donors;
```

```
*---------------;
data _nonhit; merge _match(in=a) _nonresp(in=b);
            by %impclass(&no);
      if first.&last then key=0;
      key+1;
      if  a & b;
      keep _ID key  %impclass(&no);

data _resphit; merge _match(in=a) _resp(in=b); by %impclass(&no)  ;
      if first.&last then key=0;
      key+1;
      donor=_ID;
      impvalue=var;
      if a & b;
      keep %impclass(&no) key donor impvalue;
run;
Data _impvalue; *--- clear just errors for processing;
data _impvalue; merge _nonhit(in=a) _resphit(in=b); by %impclass(&no)    key;
      if &no<10 then FindFrom='Level_0'||left(&no);
                else FindFrom='Level_'||left(&no);
      if a & b;
      keep _ID donor impvalue findfrom;
run;
*-----------------------------------------------------;
* delete the nonrespondent who had an imputed value;
*-----------------------------------------------------;
proc sort data=_impvalue(keep=_ID) out=_t;by _ID;
proc sort data=_nonresp;by _ID;
data _nonresp;merge _nonresp(in=a) _t(in=b);by _ID;
      if a & b then delete;
run;

*-------------------------------------------;
* delete the respondent who donated a value  ;
*-------------------------------------------;
proc sort data=_impvalue(keep=donor) out=_t(rename=(donor=_ID));by donor;
proc sort data=_resp;by _ID;
data _resp;merge _resp(in=a) _t(in=b);by _ID;
      if a & b then delete;
run;
%MEND MATCH;
*---------------;


%MACRO Set_Variables(indata=,impvar=,varlist=, noclass=);
```

```
data _master; set &indata;
      array _a(*) class1-class&noclass;
      array _v(*) &varlist;
      do i=1 to &noclass;
         _a(i)=_v(i);
      end;
      class0='0';
      var=&impvar;
      original=var;
      keep _ID class0 class1-class&noclass var response;
run;
*---- matching between class alias variables and actual variables;
data _a; set _master(obs=1);
         array _a(*) class1-class&noclass;
         array _v(*) &varlist;
         do i=1 to &noclass;
            _a(i)=_v(i);
         end;
         keep class1-class&noclass &varlist;
proc transpose data=_a out=_b;

data _class(keep=class no) _var(keep=var no1 rename=(no1=no)); set _b;
      if _n_<=&noclass then
         do;no+1;
            i=1*compress(translate(_name_,' ','class'));
            if i<10 then class='Level_0'||left(i);
                    else class='Level_'||left(i);
            output _class;
         end;else
         do; no1+1;
             var=_name_;
             output _var;
         end;
data _var; set _var;
      length classlist $150.;
      retain classlist ' ';
      classlist=trim(classlist)||' '||trim(var);
      keep no classlist;
data _class; merge _class end=last _var;by no;  *-- class dataset to use in
Final_Summary step;
      Level=class;
      keep level classlist;
run;
%MEND Set_Variables;
```

```
*-------------------------------------------------;
* Main routine for iteration mathing process;
*-------------------------------------------------;
%MACRO Finding_Match(NoClass=);
   *----------------;
   %MACRO CheckNonR;
        data _null_;
           call symput('NoNonR',0);
        data _null_;set _nonresp end=last;
           if last then call symput('NoNonR',_n_);
        run;
   %MEND CheckNonR;
   data _outdata; *-- clean output data;
   *----------------;
   %do I=&NoClass %to 0 %by -1;
       %CheckNonR;
       %If &NoNonR>0 %then
           %do; %MATCH(no=&I,last=class&i);
                data _outdata;set _outdata _impvalue;
           %end;
    %end;

    *-- summary of matching by level;
    proc summary data=_outdata(firstobs=2) missing;
         class findfrom;
         output out=_find(rename=(_freq_=Donors));
    data _find; merge _find _class(rename=(level=findfrom));by findfrom;
    run;
    proc print data=_find noobs split='*';
         title '2.Donors information';
         var FindFrom classlist Donors;
         label findfrom='(A)*Find from'
               classlist='(B)*Class list'
               donors='(C)*Donors';
    run;
%MEND Finding_Match;

*-------------------------------------------------;
%MACRO HHDeck(indata=,key=,impvar=,noclass=,varlist=,OutlierClass=,alpha=);
    data _master; set &indata;
         _ID=&key;
         var=&impvar;
    run;

    %resp_nonresp(indata=_master);
    %outlier(Class=&outlierclass,alpha=&alpha);
```

```
data _master; set _resp _nonresp;

%Set_Variables(indata=_master,impvar=&impvar,varlist=&varlist,noclass=&noclass);
%Resp_NonResp(indata=_master)
data _savenonresp; set _nonresp;
*-------------------------------------------------;
%Finding_Match(Noclass=&noclass);
*-------------------------------------------------;

*-- finished finding donors--------------;

* The following steps are matching original dataset;

proc sort data=_outdata;by _ID; *-- outdata is donors;
proc sort data=_savenonresp;by _ID;
data _savenonresp; merge _savenonresp(in=a) _outdata(in=b); by _ID;
      Donor_For_&impvar=donor;
      &key=_ID;
      if a;
      keep &key impvalue FindFrom donor_for_&impvar;
run;
*proc print data=_savenonresp;
*      title1 " NO of classes: &noclass ";
*      title2 " class variables: &varlist";
run;

proc sort data=&indata; by &key;
data &indata; merge &indata(in=a) _savenonresp(in=b); by &key;
      if a & b then &impvar=impvalue;
      drop impvalue;
run;
%datasets(delete=_a _b _t _nonresp _resp _class _var _master _nonhit _resphit
                  _savenonresp _match _impvalue _find _total _outdata _outlier
                  _outliersum _respsum _nonrespsum _sum);
run;
title ' ';
run;
%MEND HHDeck;
*-------------------------------------------------;
*******************************;
*          SECTION B          :
*******************************;

*-----------------------------------------------------------------;
* Macro moudle for Probability method                            :
*-----------------------------------------------------------------;
```

```
%MACRO IMPProb(indata=,key=,impvar=,impclass=);

   %if "&impclass" ^="" %then %do;
   *-- setting the last class variable;
   data _null_;
        retain one_var 'Y';
        length a $200 b $30;
        a="&impclass";
        a=trim(a);b='';
        len=length(a);
        do i=len to 1 by -1;
           if substr(a,i,1)>=' ' then b=compress(b||substr(a,i,1));
           if substr(a,i,1)=' ' then
              do;one_var='N';
                 goto next;
              end;
        end;
        if one_var='Y' then goto next;
        return;
     next:b=reverse(b);
     call symput('lastclass',b);
     run;
   %end; %else %do;
   data _null_;
         impclass='AllSampleUnits';
         call symput('impclass',impclass);
         call symput('lastclass',impclass);
   run;
   data &indata;set &indata;
         AllSampleUnits=1; ** defining an artificial variable;
   run;
   %end;
   run;


   data _temp; set &indata;
         var=&impvar;              ** imputing variable;
         _ID=&key;                 ** idenfication key;
         response=upcase(response);
   run;

   *-- assign random number;
   proc sort data=_temp;by &impclass &lastclass;

   data _temp; set _temp; by &impclass &lastclass ;
         retain seed 0;
```

```
      if first.&lastclass then seed=seed+1;
      random=ranuni(Seed);        *random seed is key point for randomising data;
      drop seed;
run;

**-- response & nonresponse;
%resp_nonresp(indata=_temp);

**-- distribution;
proc summary data=_resp nway missing;
      class &impclass var ;
      output out=_prop(rename=(_freq_=count));
run;
proc summary data=_prop nway missing;
      class &impclass;
      var count;
      output out=_tot sum=total;
run;
data _prop;merge _prop _tot;by &impclass;
      p=count/total;
run;
data _dist; set _prop; by &impclass &lastclass var;
      retain cump 0 ;
      if first.&lastclass then cump=0;
      low=cump;
      cump=sum(cump,p);
      high=cump;
      bound=compress(round(low,.00005)||'<Random=<'||round(high,.00005));
run;
**-- for information;
proc print data=_dist noobs split='*';*by &impclass;
      title '1.Producing random number boundaries';
      var &impclass var count p bound;
      sum count;
      label var="(A)*&impvar" bound='(D)*Random*number*boundary'
            p='(C)*Distribution' count='(B)*Responding*units';
run;

*----- take imputed values;
data _non; set _nonresp;
      low=random;
      keep _ID response &impclass low;

proc sort data=_non;by &impclass;
data _impdata; set _dist(keep=&impclass var low high bound) _non;
```

```
proc sort data=_impdata;by &impclass low _ID;
data _impvalue; set _impdata;
      retain tvar tlow thigh tbound;
      if response ^in('N') then tvar=var;
      if high>. then
         do; thigh=high;
             tlow=low;
             tbound=bound;
         end;
       if response='N' & (low<thigh) then
         do;var=tvar;
            random=low;
            bound=tbound;
            output;
         end;
        keep &impclass _ID var random bound;
   run;
  proc sort data=_impvalue;by _ID;

  proc print data=_impvalue split='*';
       title '2.Finding imputed value';
        var &impclass _ID random   bound var;
       label   _ID='(A)*ID' random='(B)*Random*number'
               var='(D)*imputed*value' bound='(C)*Random*number*boundary';
  run;

  *------ produce final data and flag;
  proc sort data=&indata;by &key;

  proc sort data=_impvalue(rename=(_ID=&key));by &key;
  data &indata _temp; merge &indata(in=a)
                        _impvalue(in=b keep=&key var rename=(var=impvar)); by &key;
        if b then
           do;&impvar=impvar;
              Imp_&impvar='Y';
           end;
        output;
        drop impvar;
  run;

  *****-- statistics for process;
  data _temp; set _temp;
        if response='N' then nonresp=1;
                        else resp=1;
        if imp_&impvar='Y' then imputed=1;
  run;
```

```
  proc summary data=_temp nway missing;
       class &impclass;
       var resp nonresp imputed;
       output out=_temp sum=;
  run;
  data _temp; set _temp;
       dif=sum(nonresp,-imputed);
  run;
  proc print data=_temp split='*';
       title "3.Imputation summary: impvariable=&impvar    impclass=&impclass";
       var &impclass _freq_ resp nonresp imputed dif;
       where nonresp>.;
       sum _freq_ resp nonresp imputed dif;
       label _freq_='(A)*Total' resp='(B)*Response' nonresp='(C)*Nonresponse'
             imputed='(D)*Imputed*value' dif='(E)*NotFound*imputed';
  run;

  %if "&impclass"="AllSampleUnits" %then %do;
  data &indata; set &indata;
       drop allsampleunits;
  run;
  %end;

  %datasets(delete=_resp _nonresp _temp _dist _prop _non _impdata _impvalue _tot);
  title ' ';
  run;
%MEND IMPProb;



*******************************;
*          SECTION C          ;
*******************************;
*-------------------------------------------------------------;
* Macro moudle for HOTDECK method                             ;
*-------------------------------------------------------------;
%MACRO HDeck(indata=,key=,impvar=,impclass=,alpha=);
*-------------------------------;
%if "&impclass" ^="" %then %do;
   *-- setting the last class variable;
   data _null_;
        retain one_var 'Y';
        length a $200 b $30;
        a="&impclass";
        a=trim(a);b='';
        len=length(a);
```

```
      do i=len to 1 by -1;
         if substr(a,i,1)>=' ' then b=compress(b||substr(a,i,1));
         if substr(a,i,1)=' ' then
            do;one_var='N';
               goto next;
            end;
      end;
      if one_var='Y' then goto next;
      return;
    next:b=reverse(b);
    call symput('lastclass',b);
    run;
  %end; %else %do;
  data _null_;
       impclass='AllSampleUnits';
       call symput('impclass',impclass);
       call symput('lastclass',impclass);
  run;
  data &indata;set &indata;
       AllSampleUnits=1; ** defining an artificial variable;
  run;
  %end;
  run;
*-------------------------------------;
   *Setting _ID and imputing vairable;
   data _temp; set &indata;
        _ID=&key;
        var=&impvar;
        response=upcase(response);
   run;
   proc sort data=_temp;by &impclass;
   run;
   *-----------------------------------------;
   * Separate response and non-response;
   *-----------------------------------------;
   data _temp; set _temp; by &impclass &lastclass;
        retain seed 0;
        if first.&lastclass then seed=seed+1;
        random=ranuni(seed);
        drop seed;
   run;

   %resp_nonresp(indata=_temp);
   %outlier(class=&impclass,alpha=&alpha);

   **-- sum data from outlier routine;
```

```
data _pool:merge _resp(in=a) _sum(in=b):by &impclass;
      if a & b & nonresp>0   then output;
      keep &impclass var _ID resp nonresp random;
run;
proc sort data=_pool:by &impclass random;

*-----------------------------------------;
* Finding donor and assigning key randomly;
*-----------------------------------------;

data _donor; set _pool; by &impclass;
   retain key 0 sampsize obsleft;
   if first.&lastclass then
      do; key=0;
          sampsize=Nonresp;
          obsleft=resp;
      end;
   if ranuni(1000) <sampsize/obsleft then
      do; key=key+1;
          donor=_ID;
          output;
          sampsize=sampsize-1;
      end;
   obsleft=obsleft-1;
   keep &impclass var donor key;
run;

*-----------------------------------------;
* assign key to non respondents;
*-----------------------------------------;
proc sort data=_nonresp:by &impclass random;
data _non; set _nonresp:by &impclass;
      retain key;
      if first.&lastclass then key=0;
      key=key+1;
      keep _ID &impclass key;
run;

*-----------------------------------------;
* meging donor and non respondent by key;
*-----------------------------------------;
data _non; merge _non(in=a) _donor(in=b):by &impclass key;
      if a & b;
run;

*-----------------------------------------;
```

```
  * creation final data;
  *----------------------------------------;
  proc sort data=_nonresp(drop=var);by _ID;
  proc sort data=_non; by _ID;
  run;
  data _nonresp; merge _nonresp(in=a) _non(in=b); by _ID;
       &impvar=var;
       if b then donor__ID_&impvar=donor;
       drop donor key;
  run;

  proc summary data=_non nway missing;
       class &impclass;
       output out=_impsum(rename=(_freq_=non));
  run;

  data _temp; merge _nonrespsum _impsum _sum;by &impclass;
       if nonresp=non then FindDonors='Yes, found all donors     ';
                      else FindDonors='No,  not found all donors';
  run;
  proc print data=_temp noobs split='*';
       title '2.Does system find all donors ?';
       var &impclass resp nonresp non FindDonors;
       sum resp nonresp non;
       label resp='(A)*Donors' nonresp='(B)*NonResponse' non='(C)*No of donors'
             finddonors='(D)*Find*donors?';
  run;
  data &indata; set _nonresp _resp _outlier;
       drop _ID var random;
  run;
  %datasets(delete=_temp _resp _nonresp _respsum _nonrespsum _impsum _sum _pool
                   _donor _non _total _outlier _outliersum);
  title ' ';
  run;
%MEND HDeck;
```

3. Potential imputation method for a longitudinal survey

Written by Mary_Anne Heys, Statistics New Zealand

1. Objective

This document attempts to identify and describe longitudinal imputation methodology that is currently being used and/or researched by other statistical agencies. It is intended to promote discussion and may be amended/enhanced as issues are investigated. The information in this document is a result of a literature search on the topic, all references are listed at the end of the document.

2. Introduction

For cross-sectional surveys, unit non-response is generally compensated for by weighting, and item non-response is often dealt with by imputation methods such as hot-deck, mean, etc. The time dimension associated with longitudinal surveys makes it more difficult to develop imputation strategies. The problem is that units can respond to some but not all waves of data collection. From a longitudinal perspective, wave non-response can be thought of as a set of item non-responses in the longitudinal record (suggesting imputation in suitable). However, from a cross-sectional perspective it can be thought of as unit non-response (suggesting re-weighting is suitable).

Imputation uses deterministic and probabilistic methods to supply missing data for individual items for interviewed persons, and all data for non-interviewed persons in interviewed households. In general, imputed data for the second type of non-response would be used for cross-sectional data only. No single set of procedures is optimum for all kinds of analyses, but most users of the data would prefer to have some (identifiable and documented) adjustments for non-response (Jabine et al, 1990).

In SIPP, non-response is broken down into three types, household/family (unit) non-response, person (within unit) non-response, and item non-response (Jabine et al, 1990). LSID outputs will include household, family, and individual data so we can categorise non-response in a similar way. In addition, household/family non-response and person non-response can be considered item non-response in the longitudinal record of the person.

Looking at non-response from another angle, there are three types of wave non-response: attrition, re-entry, and late entry. Attrition occurs when an individual leaves the survey at any wave, other than the first, and never responds again in any subsequent waves. Re-entry is when an individual is a non-respondent for a period of one of more waves, but responds to one or more waves preceding the period of non-response and to one or more waves following the non-response. Late entry occurs when an individual is a non-respondent to the first one or more waves, then responds on the following wave and all subsequent waves until the end of the panel (note that it had been decided that individuals who are non-respondents for the first wave of LSID will be not be considered OSMs in

future waves: see (Wave 1 person non-response), this decision may be re-considered). Certain imputation methods will be more appropriate for different types of non-response. The appendix contains brief summaries of the overseas longitudinal surveys mentioned in this document.

3. Issues not covered in this paper

Although weighting and imputation are closely linked areas, this document does not cover weighting techniques. For ideas on longitudinal survey weighting see Steve's LSID weighting document: . Another way of dealing with non-response which can help improve quality of imputed data is using admin data. Admin data has the potential to provide another possible source of skeletal data for missing waves (Kalton et al, 1993), and also to deal with item non-response. Tax data from the IRD is currently being investigated as a possible source of admin data. If it does become available it could assist in the imputation of income data in LSID. This issue is being investigated by other members of the LSID team.

4. Imputation methods

In addition to the following methods, proxy data about non-respondents may be provided by other members of their family, and non-respondents may be asked retrospective data in their next responding wave. These types of data could be used to improve imputations. See section 5 for more details.

4.1 Mean imputation method

Under this method, the mean of the values provided by respondents is assigned to the non-respondents. If the mean is calculated using all respondents it is called overall mean imputation. This method will underestimate variance and may result in invalid confidence intervals.

Under class mean imputation, individuals are grouped into classes and the mean calculated using all respondents in the class is assigned to all non-respondents in the same class. This method is deterministic. The distortion of the distribution of the values is smaller using this method instead of overall mean imputation. Class mean imputation is actually an analysis of variance model (Eurostat (author not specified), 1995).

4.2 Modal Response Imputation (for categorical variables)

This is a possible procedure for imputing categorical variables. This method assigns a modal response category among respondents who gave the same response on a previous wave. A relationship between the response in wave 1 and the response in wave 2 for respondents in both waves can be identified (i.e. the proportion in category x in wave 1 who are in category y in wave 2, etc.).

The imputed value can be that which the greatest proportion of respondents in the same category in wave 1 are in wave 2 (deterministic). This understates change for a fairly

stable variable. The alternative is to assign wave 2 response with known probability based on distribution of responses of those who gave the same response on wave 1 (stochastic). This procedure can also be modified by grouping these people based on other auxiliary information (Kalton, 1986).

4.3 Hot deck method

This method can uses cross-sectional data to divide the individuals into classes. The variables used to define the classes (imputation matrices) vary depending on the variable being imputed, and can include age, race, sex, income, occupation, and education. For each missing value, the value reported for a person with similar characteristics is substituted (Jabine, 1990, Mack and Waite, (date unknown)). This method results in more realistic variability in the imputations which is an improvement on the mean imputation method.

The donor can be chosen at random (random hot-deck imputation) or sequentially (sequential hot-deck imputation) from the respondents in the same class. Under the latter method, all records within a class are ordered sequentially according to some variable. Then the donor for a non-respondent is the value of the previous respondent out of that class in the sequential list (Eurostat (author not specified), 1995). This method is a stochastic version of the class mean imputation method, i.e. the residual added to the class mean is the donor's deviation from the class mean (Eurostat 1995). Classes can be formed using cross sectional or longitudinal data. That is, respondents and non-respondents are grouped into classes based on their response to; a) other variable(s) in the current wave, or b) the response given in the previous wave to the variable being imputed.

Random hot-deck imputation is currently used in SIPP, and ECHP to impute selected items for interviewed people using cross-sectional data. ECHP uses the same donor for variables that are linked, but not necessarily for all variables. This helps to preserve known relationships between variables (Eurostat 1996). Eurostat states that this method has a smaller effect on changes in distributions and on the under-estimation of variance than deterministic methods (Eurostat, 1996). Random hot deck is superior to sequential hot deck if the dataset is not too large: the performance is comparable but sequential method often uses the same donor, causing possibly some loss of precision (Eurostat 1995).

Hot-deck can be also used to impute the change, i.e. the difference between the donor's current and previous (or subsequent) values. The change can then be added directly to a non-missing value from the previous (or subsequent) wave, or another wave's non-missing value can be proportionally altered (Heeringa et al, 1986). The larger the classes, the less accurate the imputation, but obviously the classes can't be too small (Kalton, 1986). For SIPP, each class must contain 30 or more respondents.

The cold deck method is a variation of hot-deck imputation where imputations are based on sources other than the current wave, i.e. longitudinal data (SSW (Särndal et al, 1992)).

Users of SIPP data have commented that the cross-sectional imputations exaggerate

income changes and their magnitude (Jabine et al 1990), this will not necessarily be the case for the longitudinal hot-deck method but there are implementation issues to consider when using longitudinal data for co-habitants, e.g. record linking. Record link refers to the linking of records (datasets) for the same unit (household or individual) across waves. Verma (1996) details a record linking method for ECHP. Data for all OSMs will be linked across waves regardless.

4.4 Flexible matching method

The flexible matching method uses multivariate forward stepwise linear regression to find a set of 'matching variables'. The matching variables are ranked in order of importance and each non-respondent is then matched to a respondent using as many of these variables as possible. If a match is not made then the least important variable is dropped and the match is tried again, and so on until a match is achieved (Tremblay, 1994). The missing values are then replaced by the matched respondents values. Note that this method is equivalent to collapsing imputation cells in a case where there are no potential donors in an imputation cell.

Advantages are that this method may be good for people with unusually variable records. Disadvantages are that this method needs to be monitored, there may not be suitable matches for all non-respondents, and it is operationally more complex (Williams et al 1996).

This method is essentially a modified sequential hot-deck method that matches incomplete cases to complete cases on a hierarchical bases, i.e. the donor chosen is the nearest neighbour (Tremblay, 1994). It is perhaps more suitable for item non-response than unit non-response because a reasonable number of matching variables must be available. Data from the previous wave could be used to find a suitable donor for unit non-response.

4.5 Carry-over method (a.k.a. Longitudinal direct substitution)

The carry-over method is the situation where a wave non-respondent's responses on a missing wave were assigned the values of that non-respondent's responses to the same items on the most recent earlier wave for which data was available. The is effective for variables which are stable over time and retains the relationships between responses that occur on the wave used for imputation; provided that these relationships do not change over time, this is an attractive feature (Kalton et al, 1985). The carry-over with random R (4.4.1) and the carry over with population R (4.4.2) methods are variations.

4.5.1 Carry-over with random R method

This method is used to edit cross-sectional imputed data in SIPP and involves imputing data from previous and subsequent waves. It is only applied to imputed interviews that are bounded on each side by a self or proxy interview (Mack and Waite). It is applied as follows:

A value r is randomly assigned to each non-respondent's household for each missing wave (r =0, 1, 2, 3, or 4). The first r reference months within the missing wave receive their imputed amounts from the last reference month of the preceding wave and the remaining 4-r reference months receive their imputed amounts from the first reference months of the subsequent wave (because SIPP has a four month wave length) (Tremblay, 1994).

Figure 1. Diagrammatic representation of random carry-over method where r = 1

wave 1	wave 2		
(response)	(non-response, this wave is imputed)		
month 4	month 1	month 2	month 3

If this method were to be applied to LSID, it would involve selecting a [0-1] random number (from a symmetrical distribution) and using the value from the previous wave to impute for the current wave if the random number is less than 0.5, or using the value from the following wave otherwise.

Advantages of this method are that it is simple to implement and the data is conducive to multiple analytic purposes. In studies was shown to be good at reducing total error. Disadvantages are that it forces stability in responses for non-respondents which can lead to underestimates of between wave change (Williams et al, 1996). In general, carry-over imputation procedures appear to fail to track net changes in means or proportions when these vary over time, i.e. as expected, these methods appears to underestimate change. These methods should be used carefully on variables where extreme outliers can occur (Jabine et al, 1990).

4.5.2 Carry-over with population R method

Like the random carry-over method, the population carry-over method involves imputing data from previous and subsequent waves. However, the interviewed wave chosen to impute for the non-response is determined by a probability mass function defined by the probabilities associated with patterns found in the interviewed population. That is, the distribution of across wave changes (differences greater than zero) in the amounts between waves is derived using actual values from the population. Based on the distribution, the previous or subsequent non-missing wave data is copied into the missing wave (Tremblay, 1994).

The advantages of this method are that imputed data reflects patterns of within wave changes and total error is reduced. The disadvantage is that this method is more complex. Studies where the first month of wave 2 is used to supply data to a missing wave 1, and similarly for the last wave showed that the quality of the resulting imputed data is lower than for the middle waves (Williams et al, 1996). In general, carry-over imputation procedures appear to fail to track net changes in means or proportions when these vary over time, i.e. as expected, these methods appears to underestimate change. These methods should be used carefully on variables where extreme outliers can occur (Jabine et al, 1990).

4.6 Regression imputation

This method uses auxiliary variables to impute for item non-response. A model is built for the missing variable based on knowledge of the relationship between the wave 1 and wave 2 values of the variable and/or other related variables. The predicted value is imputed using:

$$y_i = f(x_{1i}, x_{2i}, \ldots, x_{pi}) + e_i$$

where f(x) is a function of the p explanatory variables and e_i is an estimated residual (Kalton, 1986).

When the e_i are set to zero, the imputation scheme assigns the predicted means and is deterministic. If they are estimated residuals the scheme is stochastic. Deterministic imputations tend to distort the shape of the distribution of y and attenuate its variance, so stochastic imputation schemes are preferred (Kalton et al, 1985).

The explanatory variables and the error value (optional) may be assigned in a variety of ways (Kalton 1986). For imputing for item non-response the explanatory variables can be responses to other items in the current wave. This method is used for two income tax paid variables in SLID. Eurostat lists the following procedure for choosing the explanatory variables:

First candidates are the variables of interest that are likely to be used in analysis.
Chose variables which maximise the explained variance of the target variable.
Minimise the use of explanatory variables with high percentages of missing values, e.g. discard any variables with > 70% missing values, use those with < 20% missing values, and be pragmatic between these limits (Eurostat 1995).

Alternatively, responses to items in previous waves could be used for unit non-response. Often f(x) is a linear function. One way to choose the parameters is to use least squares estimates based on those who responded in both waves. Sometimes it may be appropriate to force the regression through the origin, resulting in a model of proportional change (Kalton, 1986).

When responses to an item are highly correlated over time, using previous wave data will provide good results (e.g. hourly pay may be correlated between wave 1 and 2) However, knowing this for respondents doesn't guarantee that non-respondents hourly pay will also have high correlation... we would have to assume that the responses were missing at random (Kalton, 1986). A special case of regression imputation is substituting the value from a non-missing wave (e.g. the previous one) on to the missing wave. This may be appropriate for stable variables, in other cases it may understate the change.

4.6.1 Predictive mean matching

This is a variant of stochastic regression imputation (see 4.6) and of hot deck imputation. A regression model is developed using all respondents. This model is then applied to all respondents and non-respondents to calculate predicted values. Each non-respondent is then matched to a respondents with closest predicted values. Alternatively a group of respondents with close predicted values can be assembled and one of the group chosen to be the donor at random. Like hot-deck, this method has the advantage that the imputed values are always feasible because they are actual respondents' values (Eurostat, 1995).

Eurostat (1995) state that predictive mean matching is superior to stochastic regression because it does not rely completely of the correctness of the model. Also, the residual term is always correctly specified under this method and the assumption of homoscedasticity (equal variance) is avoided.

4.7 Little & Su method

The Little & Su method uses a multiplicative model based on row (person) and column (period) effects, i.e. imputation = person effect *period effect * residual. This makes adjustment for how the person differs from the population average, and how the wave differs from the average of the other waves. As such it is a combination of cross-sectional and longitudinal imputation. The residual is donated from a responding individual with similar row effect.

Let aij be the imputed value for variable a, for the i-th non-respondent, in the j-th wave.

$$a_{ij} = (r_i)(c_j)\left(\frac{a_{kj}}{(r_k \cdot c_j)}\right) \quad \text{which is simplified to} \quad a_{ij} = r_i \frac{a_{kj}}{r_k} \quad \text{where} \quad r_i = \frac{1}{m_R} \sum_{h=1}^{m_R} \frac{a_{ih}}{c_h} \quad \text{and} \quad c_j = \frac{m\bar{a}_j}{\sum_{h=1}^{m} \bar{a}_h}$$

where r_i = the row effect for non-respondent i
c_j = the column effect for wave j
r_k = the row effect of the interviewed person k
a_{kj} = the donor amount of the interviewed person k
m_R = the number of interviewed waves
m = the total number of reference waves
$\bar{a}_j$ = the mean amount for wave j for all interviewed people (Williams et al, 1996)

The advantages of this method are that it incorporates information about trend and individual levels, and different missing data patterns don't require separate modelling. It is also easy to implement (according to Little & Su) and analysis showed that it maintains between-wave correlations exhibited by actual data (Williams et al, 1996). Disadvantages are that this method may not be suitable for unit non-response, and possibly the relative size of total error (Kalton, 1986).

Duncan (1992) recommends applying this method for derived variables in which some of the component variables need to be imputed. Tremblay (1994) shows an application of this method to an analysis of longitudinal imputation of SIPP food stamp benefits.

5. Use of retrospective and proxy data

Retrospective data is obtained when a respondent is asked questions about the previous wave (time period) if they were a non-respondent in the previous wave. Collecting retrospective data can reduce the effects of non-response but there are quality issues to be aware of. A SIPP experiment is referred to where analyses showed that fewer transactions were recorded from the missing wave (retrospective) form than from benchmark data (Kalton et al, 1993). Proxy data is obtained when other members of a family are asked to provide data about a non-respondent in the same family. See: (Proxy data required for imputation of missing family/hh income data) for details regarding the collection of proxy data for LSID. Again, the quality of proxy data will not be a good as that obtained from the individuals themselves. It is likely that less detailed information than the full questionnaire (i.e. skeleton data) will be collected in both of these cases.

Both of these types of data have the potential to improve the quality of imputed data. In some cases this data may be of a better quality than any we can impute using other methods so it should be used directly. The non-respondent could also essentially be considered a respondent with item non-response to variables not collected in the skeleton data, but whose available data for the remaining (skeleton) variables is of a poorer quality than the 'true' item non-respondents.

6. Comparing the various methods

The following criteria for assessment of the quality of imputation methods is suggested by Eurostat (Eurostat 1995).

Plausibility of the distribution (mean, variance, outliers). Does the imputation bring the distribution closer to the expected ('true') distribution? Preferably this distribution should be based on knowledge derived from other sources.

Plausibility of relationships (covariances) between variables. These relationships should not be distorted by imputation.

Consistency (edit checks). Imputed variables should be consistent with other variables.

Eurostat also detail the process they followed to chose appropriate imputation methods for round 1 of the ECHP in 1994.

The first decision is deterministic or stochastic methods. Deterministic methods distort the distribution and underestimate the variance more that stochastic methods do. However, they provide maximum likelihood estimates. The kind of analyses that will be performed should be considered. Usually the distributions of variables are of major concern so stochastic imputation was chosen in this case.

The options were then narrowed down to random hot deck within classes and predictive mean matching. Advantages of these methods are that the assigned values are always feasible because they come from respondents, and we get a good estimate of the population distribution. Underestimation of the variance is minimised (because they do not attribute any one value or average to a whole group of non-respondents). Disadvantages are that these methods are not maximum likelihood estimates and no programmes are available within standard statistical software packages.

Predictive mean matching has three theoretical advantages over random hot deck within classes: A) random hot deck categorises continuous auxiliary variables which generally causes some loss of information; B) random hot deck is limited by the fact that there has to be a reasonable number of respondents in each class, this is not a problem for predictive mean matching; C) predictive mean matching allows a larger number of auxiliary variables and more flexibility in the mathematical specification of the model (e.g. numerous transformations can be included). However, high quality modelling is required to realise these advantages.

ECHP used the following methods in round 1:

Random hot deck within classes for imputing discrete variables.

Predictive mean matching for continuous variables where the expected additional amount of work for high quality modelling is justified by the expected quality

improvement

Random hot deck within classes for all other continuous variables. It is considered a 'safe and sound' alternative.

In round 2, historical data (i.e. round 1 data) was used to improve the models for income variables (Eurostat 1996).

The following papers include an evaluation of various method based on SIPP data. The results may be useful for us to consider when choosing LSID method(s).

In Tremblay (1994), an evaluation of the Little & Su method, hot deck method, flexible matching method, carry-over with random R method, and carry-over with population R method against real response data.

Various measure of evaluation are constructed (mean, std dev, correlations, hypothesis tests etc.) for comparing i) the imputed cross-sectional data, and ii) the imputed longitudinal data (cross wave changes only, i.e. the change between two consecutive waves) under the various methods, with the actual data. (All cross wave changes are from real to imputed data or vice versa.)

Results were not particularly conclusive. None of the four methods 'significantly' shines above the rest, the most accurate method differs across the waves and across the various methods. One possible reason for this is that only approx. 8% of all households had data imputed (this proportion was based on observed non-response rates).

Williams et al (1996) also do an evaluation which compares the random carry-over method, population carry-over method, Little & Su method, and flexible matching method against real response data.

The methods of evaluation include between wave correlation coefficients, averages and standard deviations of wave means, the average absolute deviation.

Relative to total error, the carry-over procedures seemed to be favourable. Little & Su and flexible matching performed better in maintaining cross-wave relationships and imputation for relatively large changes between waves. However, the computational burden and relative size of their total error was unfavourable.

As with Tremblay's (1994) research, a uniformly 'best' imputation procedure was not identified.

Kalton et al (1985) examines how well the basic carry-over method (4.4) performs as an imputation scheme using actual values from 1984 SIPP data, some of which are deleted to provide quasi-non-respondents. It is noted that this method of evaluation is problematic because of the probable variation in measurement errors between waves.

The methods of evaluation are i) the mean deviation between actual values and imputed values which is a measure of bias; and ii) the mean square deviation and the root mean square deviation which measure the closeness of the imputed values.

Concludes that the carry-over method should not be applied uncritically with numerical variables because outliers and measurement errors occur and unrealistic longitudinal records may occur.

Imputation may be better that weighting solutions for estimates based on small subclasses, when the loss in effecting sample size matters and when any bias caused

by imputation is less important relative to the sampling error.

Heeringa et al (1986) uses SIPP wage and salary records to compare cross-sectional hot deck imputation with the basic carry-over method.

> The cross-sectional hot-deck method appears to perform only slightly better than chance for categorical wage and salary variables. For continuous variables, this type of imputation does not appear to appreciably alter the distributions of the items, but it does have an impact on both cross-sectional and longitudinal multi-variate distributions. The carry-over method understates change but this may be preferable to the gross overstatement of change from the cross-sectional hot-deck method.

Jabine et al (1990) comment that the direct substitution method has been investigated and, as expected, it appears to underestimate change . However, the cross-sectional hot-deck method tends to over-estimate change. Which is preferable?

7. General points

Routine application of a set of rules for imputation could lead to a dataset that is not improved, possible inferior. ... Good quality results always require considerable amounts of good judgement during the imputation process (Eurostat, 1995).

The statistical need to impute increases with the loss of information caused by the item non-response percentage. When item non-response is low is statistically superfluous to impute. When it is high it can be difficult to avoid multiple use of donor respondents if this type of method is used (Eurostat, 1995).

Note that cross-wave imputation methods depend on a strong inter-wave correlation in the variables of interest (Williams et al, 1996).

Imputation methods are generally limited by the availability of the data item in the previous wave. A technique for cases where the data item is missing in both waves needs to be developed. (adds complexity, reduces quality) (Kalton 1986).

Users of SIPP data have commented that the imputation process does not preserve the known relationships between benefits and the determinants of benefits (Jabine et al 1990). One solution to this is to use the same donor to impute for variables whose relationship is important. This is called record matching and preserves relationships (Eurostat, 1995).

Duncan (1992) stresses that it is important to use methods which preserve the distribution of the variables.

For SIPP, CAPI permits consistency data checks during the interview to reduce the level of post-collection edits and imputation (Huggins et al, 1994).

Eurostat (1995) recommends the SURFOX software package. This software can implement the following techniques: imputation by hand, deductive imputation, random hot deck overall and within classes, predictive mean matching, record matching, and cold deck.

8. Appendix

Brief summaries of overseas longitudinal surveys

SIPP (Survey of Income Program Participation) is run by the US Bureau of the Census. This is a panel survey that collects data from the same units in eight different waves. The units of analysis are individual people for some analyses, and households, families, or other groupings for other analyses. Note that SIPP has a four month wave length, and currently only imputes for missing waves which are bounded on both sides by an interviewed waves, not for two + missing waves, or for the first or last wave of the panel (Williams et al, 1996).

SLID (Survey of Labour and Income Dynamics) is run by Statistics Canada. SLID is a longitudinal panel survey, each panel participates for six years. SLID collects annual data, two collections are carried out per year to collect labour and income data, both have the same reference period (Lavigne et al, 1998). SLID uses the 'nearest neighbour' method to identify donors for some income variables, and linear regression for tax payable.

ECHP (the European Community Household Panel) is run by Eurostat. ECHP is a longitudinal survey conducted in member states of the European Union by Eurostat. It involves annual interviewing of a panel of households and individuals in each country on a wide range of topics related to living conditions. According to Eurostat (1996 - Doc. PAN 74/96), random hot deck within classes and predictive mean matching were used to impute for missing data using cross sectional data. Longitudinal imputation is used only for income from wave 2 onwards. Income is modelled using income from the previous wave as one of the explanatory variables.

9. References

Survey Methods 2 have copies of most of the following documents.

'BCHP User Documentation' Institute for social and economic research, University of Essex

Duncan G (1992) 'ECHP: Quality control measures, weighting and imputation in the EC panel project' Eurostat development group, Doc. PAN 6/92

Eurostat (author not specified) (1996) 'The European Community Household Panel: Survey methodology and implementation', Volume 1, Luxembourg Office for Official Publications of the European Communities

Eurostat (Author not specified) (1996) 'Wave 2 evaluation, longitudinal imputation', ECHP Working group, London 17 - 18 October 1996, Doc. PAN 74/96, Eurostat

Eurostat (Author not specified) (1995) 'Cross-sectional imputation rules, and application to the micro-data files', ECHP Working group, Paris 18-19 September 1995, Eurostat

Eurostat (Author not specified) (1994) 'Agenda Item 6e: Evaluation and Imputation', ECHP Working group "Household Panel", September 1994, Doc. PAN 22, Eurostat

Heeringa S. and Lepowski J. (1986) 'Longitudinal Imputation for the SIPP', Proceedings of A.S.A. section of survey research methods, 1986

Huggins, V. and Fischer, D. (1994) 'The Redesign of the SIPP', U.S. Bureau of the Census, A.S.A. 1994 Proceedings of the section of survey research methods

Jabine, T, King, K and Petroni, R (1990) 'SIPP Quality Profile' U.S. Department of Commerce, Bureau of the Census

Kalton, G. (1986) 'Handling Wave Non-response in Panel Surveys', Journal of Official Statistics, Vol 2, No 3, pp 303 - 314

Kalton, G. and Citro, C. (1993) 'Panel surveys: adding the fourth dimension' Survey Methodology, December 1993, Vol 19, No 2, pp 205 - 215

Kalton, G., Lepowski, J. and Lin, T. (1985) 'Compensating for Wave Non-response in the 1979 ISDP Research Panel" Proceedings of A.S.A. section of survey research methods, 1985, pp 372-377

Lavigne, M. and Michaud, S. (1998) 'General Aspects of the Survey of Labour and Income Dynamics' Income and labour dynamics working paper series, Statistics Canada product number 75F0002M

Mack, S and Waite, P 'Non-response Research Plans for the Survey of Income and Program Participation' No. 206, U.S. Department of Commerce, Bureau of the Census

Pennell S. (1993) 'Cross-Sectional Imputation and Longitudinal Editing Procedures in the Survey of Income Program Participation' The University of Michigan, Survey Research Centre, Institute for Social Research

Särndal, Swensson, and Wretman (1992) 'Model Assisted Survey Sampling', Springer Series in Statistics, Springer-Verlag New York Inc.

'SIPP Quality Profile' 3rd edition 1998, U.S. Department of Commerce, Bureau of the Census

Tremblay, A. (1994) 'Longitudinal Imputation of SIPP Food Stamp Benefits', U.S. Department of Commerce, Bureau of the Census, Proceedings of A.S.A. section of survey research methods

Verma, V. (1996) 'Basic Longitudinal Edits' ECHP doc. PAN 62/96, Eurostat

Williams, T & Bailey, L (1996) "Compensating for Missing Wave Data in the SIPP" A.S.A. 1996 Proceedings of the Section on Survey Research Methods.

12. 표본설계를 처음 시작하는 사람에게 알리고 싶은 글

가. 모집단은 이런 시각으로

표본설계시 가장 기본이 되는 사항은 모집단을 어떻게 정의할 것인가이다. 통계조사는 항상 시점을 중요시하는데 언제를 기준으로 하며 조사 범위는 어디까지 할 것인가? 또 어떠한 내용을 포함할 것인가? 조사단위는 무엇으로 할 것이며, 그 결과 모집단의 층 모수는 얼마인가를 결정하는 것이 중요하다.

예를 들면 도소매 통계조사는

- 시 점 : 1991년 총사업체 통계조사
- 범 위 : 전국 도소매 사업체가 존재하는 전지역으로 함
- 조사단위 : 도매, 소매업종에 해당하는 개별 사업체(사업체 정의 참고)
- 모 수 : 전국 ○○○개 도소매 사업체

그리고, 모집단에 관한 정보를 어디에서 구할 수 있는 가를 담당실무자들과 협의한 후 그 정보에 포함된 내용중에서 실제로 우리가 모집단을 구성하는데 필요한 사항들을 면밀히 검토한다. 또한 모집단을 구성하는데는 기본 정보로 부터 어떤 항목들을 가져올 것인가가 문제인데 이는 설계하고자 하는 표본틀 (Sampling Frame) 구성을 먼저 염두에 두고 관련항목을 검토하고, 여기에 과거에 행하여진 지침서, 조사표, 전산 file 구성, 전산처리 요령서 등 그 조사와 관련된 모든 사항을 검토하여야만 비로소 표본설계 작업을 시작할 수 있는 지식을 확보할 수 있다.

나. 표본의뢰자의 요구사항 분석

1) 조사를 기획한 실무자가 어떤 목적으로 표본조사를 하고자 하는지를 정확히 파악하여 표본설계 가능성을 타진한다.

- 조사하고자 하는 주된 항목(판매액, 급여액, · · · ·)
- 조사하려는 지역(전국, 시부지역, · · · ·)
- 동원가능한 예산, 특히 조사요원 확보 가능성
- 조사주기(한번, 연속하여 조사, · · · ·)
- 조사방법(면접, 우편, · · · ·)

2) 통계표 작성을 어느 정도 할 것인가를 파악한다
- 전국 단위로 분석
- 지역 단위로 분석
- 개별 조사단위로 분석
- 별다른 분석결과의 계획이 없다면 과거, 혹은 유사 보고서를 참고로 하여 예상 분석결과를 찾아야 한다.

3) 정기적으로 실무자와 회의를 통하여 가시화되어 있지 않은 요구사항을 알아내고 문제점을 파악하고 의문사항이 있으면 수시로 실무자와 협의를 하여 표본설계 이전에 모든 요구사항과 유관사항을 파악해야 한다.
- 유고사업체 대체는 어떻게 할 것인가?
- 행정구역 번호는 언제를 기준으로 할 것인가?
- 산업분류는 신, 구중 어느것으로 할 것인가?

다. 분석하고자 하는 내용 파악

1) 조사기획 실무자는 예산과 조사정도(精度)는 고려하지 않고 때로는 너무 과한 욕심을 부리는 경우가 있다. 조사표본수가 얼마 되지않은 소규모 조사계획에서 분석하고자하는 cross table 이 지나치게 많아 실제로 각 cell에 포함된 숫자가

미미한 경우에는 통계로서의 의미가 전혀 없는데도 불구하고 욕심을 부리는 경우가 있는데, 이런 경우에는 그 부당함을 과감히 지적하여 세분하여 분석하는 것은 무리라는 것을 표본설계자가 일깨워 주어야 한다.

2) 조사 목적에 타당한 분석을 하는지 검토해야 한다. 때로는 조사 목적하고는 별개로, 개인적 흥미로 분석하고자 하는 경우가 있는데, 이렇게 목적에 맞지 않는 항목 삽입으로 조사를 어렵게 할 여지가 있는지를 검토해야 한다.
실제로 응답자는 조사설계자의 취지를 단 몇%도 이해하지 못하므로 되도록 조사에 어려운 항목들은 삭제 시켜야만 용이한 조사가 가능하게 된다.

3) 분석하여 작성하고자 하는 내용을 정확히 알아야만 표본규모를 결정할 수 있고, 또 표본설계 방법론과 추정 방법 등에도 깊은 관계가 있으므로 조사기획 실무자가 분석하려는 내용이 무엇인지 정확히 파악해야 한다.

4) 과거로부터 계속되어온 조사라든가 유사조사 결과가 있으면 이미 발행된 보고서의 분석 table을 보고 업무를 파악하는 것이 가장 바람직하다. 각종 통계 결과표를 보고 표본설계와 추정.규모 등을 생각해야만 한다.

라. 표본규모 결정 요인들

1) 가장 중요한 것은 예산 즉 동원가능한 조사요원 수이다. 아무리 좋은 조사목적이라도 바람직한 표본 규모와 그를 수용할 수 있는 조사요원이 없다면 소용없는 일이다.

2) 무응답, 불응 등 조사가 불가능한 부분을 미리 감안하여 표본규모를 실제보다 여유있게 결정해서 최소한 조사목적에 타당한 분석이 가능하도록 해야 한다. 과거에 실시했던 유사 조사의 유고율을 참조하면 규모를 결정하는데 크게 도움이 될 것이다.

3) 통계결과를 어느 정도로 작성, 공표할 것인가도 표본규모에 중대한 영향을 준다. 전국단위 통계만 작성할 것인가? 아니면 지역단위 통계도 작성할 것인가에 따라 표본규모와 설계방법에 큰 영향을 준다.

4) 분석하고자하는 주요 특성치의 분산정도에 따라 이론적인 표본규모가 결정된다. 특성치의 분산이 크면 당연히 표본이 클 수 밖에 없다.

5) 요구되는 정도(precision)와 신뢰정도(confidence interval)도 표본규모를 결정하는데 중요한 역할을 하는데, 예산과 동원 가능한 조사요원수를 고려하여 실무자와 협의를 하여 결정하여야 한다.

6) 표본규모는 이론상 계산되는 것도 중요하지만 일반적으로 비표본오차가 표본오차보다 계산할 수 없을 정도로 크다는 것을 감안한다면, 이론을 중요시하여 무조건 표본규모를 크게해서 표본오차를 감소시키면 반대로 비표본오차를 감당할 수 없게 된다. 혹자는 조사에 포함된 오차중에서 표본오차를 빙산의 일각이라고 표현하는데 그만큼 조사에 있어서 비표본오차가 엄청나다는 비유일 것이다.

7) 표본설계자들은 최종적으로 표본규모를 어떻게 정할 것인가 하는 문제에 직면하게 되는데 이론상 계산되는 규모와 현실적인 제약(예산및 조사원수)을 비교하면

늘 갈등을 겪게되는데 이럴 때는 이론에 근접한 현실적 숫자를 선택하라고 권하고 싶다.

8) 특히, 사업체 표본조사에서는 이론적으로 바람직한(3σ, 5%) 표본규모는 도저히 표본조사의 잇점(비용, 시간, 인력절감)을 살릴수 없을 정도로 거의 전수조사 규모가 나오게 되는데 이럴 때 설계자는 고민하게된다. 사업체 표본설계에서 이상적인(3σ,5%)이론을 적용시키면 어떤 표본설계 방법론을 이용하더라도 표본규모를 적게할 수는 없다.

9) 사업체 표본설계를 할 경우에는 우리청에서 발간('91.12월)한 〔절사법 표본설계 응용〕을 참고하라고 권하고 싶다. 그 방법이 현실적으로 가장 적용가능한 표본을 구할 수 있는 방법이 아닌가 생각된다.

10) 표본규모 결정의 접근 방법은 두가지가 있다.
한가지 방법은 신뢰구간을 고정시키고 정도를 3, 4, 5 · · · %도로 변화시키면서 현실에 적합한 규모를 산출하는 것이고, 다른 한 방법은 정도를 고정시키고 신뢰구간을 1σ(68%), 2σ, · · · 로 변화시키면서 적당한 표본규모를 결정할 수 있는 Simulation 이 있는데 어느 방법을 선택하느냐 하는 것은 요구자의 취향과 목적에 따라 결정하는 것이 좋다.

11) 그러나, 대부분의 설계 의뢰자가 신뢰구간과 정도의 개념에 다소 인식이 부족하므로 업무 성격에 따라 표본 설계자가 방법을 결정하는 것이 좋다. 사업체 표본

의 경우 일반적으로 68% 신뢰구간에 7~10% 정도를 주어 규모를 계산하면 예상된 수치에 가장 근접한 규모를 계산할 수 있다.

12) 주요 특성치를 무엇으로 정하느냐에 따라 표본규모가 변할 수 있다. 가장 바람직한 것은 조사 목적에 가장 적합한 특성치를 찾아서 그 특성치의 분산을 구하여 표본규모 계산에 사용해야 한다.

13) 특성치는 총조사 자료에서 찾아 이용해야 하는데, 그런 총조사가 없다면 유사한 조사에서 비슷한 특성치의 분산을 사용하거나, 계속되는 조사라면 과거조사의 분산을 구하여 요구 분산을 정하여 표본규모를 계산한다. 실제로 우리청 운수업동계조사 표본조사는 과거 조사의 분산과 현재의 모수로 표본규모를 결정한다.

14) 과거유사조사도 찾을 수 없는 경우는 실제로 가동 가능한 지역별 조사요원수를 고려하여 전체 예상 규모를 할당하여 조사하는 것도 한가지 방법일 수 있으며, 이 결과는 다음 조사의 표본규모 계산의 기초 자료로 이용할 수 있으므로 정보수집 기능도 포함한다고 볼 수 있다.

마. 기본적인 표본이론의 의미를 알자

1) 흔히들 표본이론은 고도의 전문지식이라 생각하고 처음부터 거부반응을 일으키는데, 이는 아주 기본적인 원리를 이해하려고 하지 않는 안이한 생각 때문이다.

2) 표본이론에 접하려면 우선 가장 기본적인 평균과 분산의 의미를 이해하고 여기에 실제로 숫자를 대입하여 현실감을 높여야 한다.

3) 그런 후에 정규분포, t분포 등 분포와 관련된 사항과 표본조사의 근간이 되는 중심 극한 정리를 포함하여 신뢰계수, 정도, 점추정, 구간추정, 추정량과 추정치, 편기(bias), 표본의 크기와 정도와의 관계 등을 이해하고

4) 실제 표본을 추출하는 방법인 단순임의 추출법, 층화 임의추출법, 계통추출법, 집락추출법, 확률비례추출법, 다단추출법 등 추출의 기본개념과 쓰이는 방법 등을 표본이론 및 실무책자를 통하여 기초적인 의미만 숙지하여 두면 나중에 실제로 필요할 때는 책을 찾아 보면 되는 것이다.

5) 이렇게 표본에 대한 내용은 가장 기초가 되는 부분만 이해하고 공식까지 숙지할 필요는 없으며, 실제로 표본을 설계할 때에는 요구한 내용을 면밀히 분석 검토하여 거기에 적합한 공식을 사용할 수 있도록 국내에서 발간된 기본 책자 한 권 정도는 대충 읽어보고 표본이론에 접하는 것이 좋을 듯하다.

6) 표본에 대한 기본 이론이 어느정도 이해가 되었으면, 다음 외국판 서적을 원어로 읽어 내용을 파악하여 그 뜻을 명확히 이해하는 것이 바람직하며, 향후 표본설계를 직접 하게 될 것에 대비하여 외국 표본실무의 이론적용 관계를 간행물 등을 통하여 자료를 모아두는 것도 좋은 방법이라 생각된다.

7) 국내판 표본책자를 소개하면 "표본조사법(표본이론과 실제 : 박제수)"가 표본에 대해 가장 여러부문을 소개했고 "표본조사법(김종호)"은 기본적인 공식들이 노트형식으로 기술되어 있으며, "통계조사론(박흥래)"은 표본이론을 심도있게 다룬 표본이론서 이다. 여기에 "시장조사론(채서일)" "조사방법론(김해동)" 을 추가하면 조사방법과 연관시켜 표본이론을 공부할 수 있으므로 가장 바람직할 것으로 생각된다.

8) 외국서적으로는 가장 기초가 되는 것이 "Sampling lectures(US. Bureau of the Census)" 로서 기본적인 이론을 쉽게 꼭 필요한 것만 수록하였다. 기타 기본 서적으로는 "Sampling techniques(cochran)", "Elimentary Sampling Theory (yamane)", "Survey Sampling(Leslie Kish)" 등이 있으며, 국내 서적을 통해 기본이론에 익숙해져 있는 상태면 외국서적을 보는 데는 영어 실력과는 무관하게 쉽게 읽어나갈 수 있을 것으로 생각된다.

9) 표본에 대하여 좀더 심도있게 실무에 접하고자 하면, 국내에서 발간되는 책자중 표본조사 보고서 첫부분에 수록된 표본설계 개요를 수집하여 읽기 바라며, 또한 UN에서 발간된 각종 보고서에는 외국 기관의 표본설계가 수록되어 있으므로 이를 이용하면 표본실무 숙지가 가능하리라고 본다.

10) 실제 업무를 수행하는데 있어서 표본이론과 실무를 접속하려고 하면 컴퓨터 언어(language)하나 정도는 완전히 숙지하여 두는 것도 표본설계에 도움을 줄 수 있으리라 생각이 든다. 표본설계를 하는데 기본적인 도구로서 컴퓨터를 꼽는데 대량의 기본자료인 모집단을 분석하고 표본추출 방법들을 자유자재로 테스트(test)하려면 기본적인 컴퓨터 언어하나 정도는 완전히 자기것으로 소화해야만 업무를 원활하게 수행할 수 있을 듯 싶다.

11) 기왕에 컴퓨터를 이용할려면, 컴퓨터내에서 통계분석이 직접 가능한 SAS 정도는 숙지하여 두는 것이 좋을 것이라 생각되어 필자가 권하는 통계분석 package이다.

바. 모집단을 분석하자

1) 표본들은 표본설계를 한 후 실제 조사를 할 수 있도록 망부 Listing 까지 고려하여 그 구성항목을 결정해야하므로 사전에 모집단을 철저히 분석해야 한다.

2) 표본들을 작성하면서 오류를 줄이기 위해서는 각 항목에 대한 점검 등 가장 기본적인 사항부터 점검한 후 자기가 직접 작성한 작업절차(program)에 과오가 없는지를 단계적으로 점검해야 한다.

3) 기본적인 사항이란, 예를 들면 시도별 Data 건수라든가 산업분류 등 기본분류에 관계된 사례수를 파악하여 보고서와 상호 비교하여 작업 진행이 제대로 되어가고 있는지를 파악하는 것이며, 기본 보고서가 없는 잠정 자료일 경우 조사된 기초자료의 건수를 파악하여 data건수를 미루어보는 방법을 이용하여 작업 진행과정을 점검해보도록 해야한다.

4) 설계시 주요 특성치로 선정한 조사항목에 대해서는 각 cross 별로 집계하여 설계자가 계산한 방식대로 작업화일이 구성되었는가를 검토하는 것도 매우 중요하다고 생각된다. 이러한 일련의 체크사항이 모두 정비되고 나면, 그 다음 단계로 모집단 자료분석에 임할 수 있게 된다.

5) 필요한 항목들을 설계목적에 따라 구성하였다면, 조사표라든가 조사지침서 등을 검토하여 각 항목에 대한 의미를 명확히 알아야 하며, 또 그 항목들이 전산화일 속에 어떻게 구성되어 있는가를 이리저리 파악해 두어야만 한다. 필자의 경험에 의하면 당연하다고 생각되는 각 자료의 속성중에도 예기치않은 예외가 항시 존재하여 곤란을 겪었던 적이 여러번 있었다.

6) 사실 자료의 속성을 모두 파악했다면, 그 표본설계는 다 끝났다고 보아도 과언이 아닐 정도로 자료의 속성 파악은 표본설계에서 매우 중요한 역할을 한다.

7) 사업체 부문 표본설계 과정중 가장 기본적이며 공통적으로 점검해야 할 사항들

- 행정구역 분류의 사용 시점
- 행정구역(시도별)별 자료 건수
- 시도별 주요특성치 (출하액, 판매액)의 합계
- 산업분류별(대분류) 자료 건수 및 특성치 합계
- 품목별 자료 건수
- 종업원 규모별 자료건수 및 특성치 합계
- 자료중 결손치(missing value)는 있는가? 있다면 처리방법은?
- 층 자료건수 파악 → 자료처리 방법 연구
  - data 의 양에 따라 일괄처리 또는 분할처리 선택

위에 적은 사항들을 점검해가면서, 업무처리 중간중간에 자기가 처리하고 있는 기본 흐름이 정확한지를 항상 실무자들과 협의하거나 기본 보고서를 참고하여 check해 나가면 나중에 낭패를 당하는 일은 없을 것이다.

8) 이제까지 기술한 여러 사항들을 점검하면서 과거 전임자의 설계방법론을 검토하고 그 결과를 현재 실무부서의 담당자들과 서로 협의하면서 표본설계의 문제점 및 표본관리상의 미흡한 점을 개선시킬 수 있는 방안을 연구하면서 모집단을 분석해가면 기존 방법보다 훨씬 개선된 표본설계 방법론을 도출할 수 있다고 생각한다.

사. 기타사항

1) 통계조사용 표본설계 방법론과 여론조사, 시장조사 등의 표본설계 방법론을 상호 비교 연구하는 것도 표본설계공부에 도움이 되리라 생각된다

2) 신문지상에 발표되는 각종 조사결과를 보면서 표본설계는 어떻게 하였으며 또 대표도는 과연 있는지 여부를 생각해 보는 것도 또한 좋은 방법이다. 이러한 방법은 기본적인 모집단 정의를 자주 생각케하고 그 결과치가 대표성이 있을 것인가를 생각하게 되기 때문이다.

3) 통계조사와 여론조사의 조사표 설계에는 아주 상당한 차이가 있으며, 분석 목적도 상이하므로 분석은 어떤 방법으로 하였으며 또 어느 정도(수준)의 분석을 했는가도 에의 주시해서 읽어보면 도움이 되리라 생각된다.