기술혁신학회지 제5권 제2호
2002년 7월 pp. 228-244

과학기술연구개발활동조사의 개선방안*

- 기업부문을 중심으로 -

Policies for Improving the Survey of Research and Development in Science and Technology: The Case of Industrial Sector

유 승 훈** · 문 혜 선***

〈目 次〉

<Abstract>

The survey of research and development (R&D) in science and technology (S&T) covers the current status of R &D activities in S&T in Korea, and provides a basis for decision making regarding S&T policy. Continuous improvement of the survey is widely needed to present reliable national basic statistics. Therefore, the purpose of the study is two-fold: to introduce sampling survey method in industrial sector and to make statistical technique to deal with non-response data from industrial sector. To these ends, first, case studies of the United States and Japan are illustrated. A new sampling design for the R&D survey is proposed and implementing stratified random sampling scheme is suggested. Moreover, statistical analysis of the non-response data is dealt with. Based on several screening criteria, we develop a new imputation method suitable for the R&D survey and also provide more detailed implementation plan. Various solutions to a problem arising from non-response item are also presented. Finally, some implications of the results are discussed.

Key words: R&D activities, Sampling framework, Stratified random sampling, Non-response

** 호서대학교 경상학부, shyoo@office.hoseo.ac.kr
*** 한국과학기술기획평가원, hsmoon@kistep.re.kr

Ⅰ. 서 론

과학기술연구개발활동조사는 우리나라 과학기술 분야의 연구개발현황에 대한 기초통계조사로서 1963년 경제기획원의 최초 조사이래 매년 실시되어 왔으며, 1999년부터는 한국과학기술기획평가원(KISTEP)에서 조사 실무를 담당하고 있다. 이 조사는 통계법 상의 지정통계(제 10501호)일 뿐 아니라, 조사결과는 우리나라의 연구개발관련 공식 통계로서 OECD에 보고되고 있으며, 과학기술정책 결정 및 계획과정에 기초자료로 제공되고 있어 매우 중요한 위치를 차지하고 있다.

이 조사의 대상은 <표 1>에 제시되어 있는 바와 같이, 크게 자연과학분야의 공공연구기관, 대학, 의료기관 및 대학부속병원, 기업체의 4개 부문으로 구분된다. 이 중에서 공공연구기관 및 대학부문은 우리나라 전 기관에 대한 전수조사가 실시되고 있으며, 회수율도 95% 이상으로서 미회수 기관에 대한 통계처리만 보완된다면 조사방법상 별다른 문제는 없다. 의료기관은 현재 종합병원급 의료기관인 병상수 80개 이상의 병원[1]을 대상으로 조사를 하며 회수율은 95%이다. 따라서 병상수 80개 미만 병원의 연구개발실적이 있을 경우 연구개발통계가 과소계상될 수 있으나, 의료기관의 연구개발활동이 전반적으로 미미하고[2] 실제 여건상 일정규모 이상의 병원에서 연구개발 수행 가능성이 높다는 것을 가정한다면, 조사방법으로 인해 누락되는 비중은 크지 않을 것으로 보인다. 통계의 신뢰성 제고를 위해 우선적으로 보완되어야 할 부분은 기업부문으로서, 현재 전년도 연구개발활동조사 응답기업 또는 부설연구소 및 연구전담부서 보유기업을 대상으로만 조사를 실시하고 있으며, 회수율은 63%이다.

물론 연구개발실적보유기업 및 매출액 1,000대 기업에 대해서는 전수조사 및 100% 회수가 이루어지고 있어 주요 연구개발활동은 집계되고 있지만, 전년도 연구개발활동 조사 응답기업 또는 부설연구소 및 연구전담부서 보유기업에 대해서만 조사를 수행하는

〈표 1〉 2000년 연구개발활동조사 현황

구 분	조 사 대 상	대상기관 수	회수기관 수 (회수율)
공공연구기관	- 국공립시험연구기관, 정부출연 연구기관, 비영리법인연구기관	225	217 (96%)
의료기관	- 병상수 80 이상의 병원	464	440 (95%)
대 학	- 2년제 이상 대학	362	342 (95%)
기 업 체	- 전년도 연구개발수행기업 - 연구소 및 전담부서보유기업	7,350	4,620*(63%)

주 : 미회수 기업(2,730개 기관)들 중 1,604개 기업은 2000년도 중에 설립된 기업으로서 2001년 활동조사(조사대상기간: 2000.1.1~200.12.31)의 분석과정에서는 제외하였으며, 나머지 미회수 기업(1,126개 기업)이 기업부문 총 연구비에서 차지하는 비중은 약 3.5%[3]이다.

1) 현재 종합병원 기준은 100병상 이상이나, 기존 조사기준과의 일관성유지를 위해 80 병상 이상을 대상으로 한다.
2) 2000년 조사에서 연구개발수행 의료기관은 50개이며, 우리나라 총 연구개발비의 0.17%에 해당된다.
3) 산업기술진흥협회의 "사업실적 및 계획보고서" 자료 기준.

것은 표본선택(sample selection)의 문제를 야기할 수 있다. 즉, 연구개발활동이 과소하게 평가될 수 있는 여지를 안고 있는 것이다. 따라서 기업체의 경우, 모집단을 정확하게 파악한 후에 표본조사 기법을 적용하는 것이 바람직해 보인다.

아울러 기업체의 경우 조사표 회수율은 63%에 이르고 있는데, 조사에서의 무응답은 조사의 정확도를 결정하는 중요한 요인임에도 불구하고 이에 대한 통계적 처리가 제대로 되지 못한다면 조사결과를 집계하는 데 있어 오류를 초래하게 된다. 따라서 우리나라의 연구개발현황을 정확히 집계하고, 조사의 신뢰성을 제고하기 위해서는 현재 조사기준에 포함되지 않는 기업을 대상으로 한 표본조사 방법의 도입과 조사표 미회수 기업에 대한 무응답 처리 방안 도입이 필요하다.

본 연구의 내용은 크게 3가지로 구성된다. 먼저 Ⅱ장에서 기술선진국인 미국과 일본의 연구개발통계조사 현황을 조사하여 이를 우리나라 과학기술연구개발활동조사개선방안 마련의 배경자료로 활용한다. 둘째, Ⅲ장에서 연구개발활동조사의 표본조사 도입방안을 검토한다. 연구개발활동조사의 특성을 고려한 표본조사 방안을 연구하는 것이다. 이를 위해 기업부문의 모집단에 대해 층화표본추출 도입 방안을 검토한다. 층화의 기준 설정과 함께 각 층별로 표본의 수를 결정하는 방안을 제시한다. 마지막으로, Ⅳ장에서 무응답 기업에 대한 통계처리를 연구하여 도입방안을 제시함으로써 조사결과의 신뢰성을 제고하고자 한다.

Ⅱ. 선진국의 사례

1. 미국의 기업부문 R&D 활동조사

1) 조사의 개요 및 표본추출틀

기업 R&D 활동조사는 미국 내 산업 R&D 활동에 대한 주요 정보원으로 미국 국립과학재단의 과학자원통계국(Division of Science Resources Statistics)이 담당하고 있으며, 부처간 협약에 의해 조사는 미 상무부 산하 조사국(Census Bureau)이 수행하고 있다. 조사는 R&D를 수행하는 것으로 알려진 제조업 또는 비제조업의 대표표본 및 R&D를 수행하는 것으로 추측되는 기업들로부터 추출된 대표표본을 대상으로 이루어진다. 자료는 기업단위로 취합되며, 조사내용은 R&D의 유형(기초, 응용, 개발), 기업규모, 과학자와 공학자의 고용수준, 장비, R&D 지출, 연방 R&D 기금 지원액, 지리적 위치, 북미 산업분류 시스템, R&D에 종사하는 과학자와 공학자, 매출, 판매, R&D 재원(기업 또는 연방) 등이다.

관심모집단은 미국에서 R&D를 수행하는 모든 기업들로 구성된다. 조사국은 유급직원을 보유한 3백만 개 이상의 사업체들에 대한 정보를 담고 있는 SSEL (The Standard Statistical Establishment List)을 작성하는데, 표본추출틀(sampling framework)에는 비농업산업으로 분류된 모든 영리기업들이 포함된다. 1992년 이전의 표본추출틀은 산업별로 그 기준이 다르게 적용되었는데, 예를 들어 1987년의 표본추출틀은 특정 규모와 산업의 범주에 포함되는 154,000개의 기업을 대상으로 하고 있다. 그러나 1992년부터 새로운 산업들이 표본추출틀에 포함되었고, 규모에 대한 기준이

상당히 낮아지면서 전 산업에 고르게 적용된 결과, 5인 이상의 종업원을 보유한 약 2백만개의 기업들이 표본조사(1998년도 표본을 포함하여)의 대상이 되고 있다.

2) 표본의 설계

먼저 각 기업을 제조업과 비제조업의 2개 부문으로 구분한 다음, 각 부문 내에서 3자리수(3-digit)의 표준산업분류(SIC) 코드 수준으로 표본추출층(sampling strata)을 만들어 분류한다. 1994년 이전에는 항상 조사대상에 포함되는 확실성 기업(certainty company)의 기준으로 종업원 1,000명 이상의 기업을 선정하였으나, 실제조사결과 R&D 지출 수준의 변동폭이 매우 크며 많은 기업들이 R&D 지출을 하지 않는 것으로 보고되어 이러한 규모기준(size criteria)은 폐지되었다. 1994년부터는 전년도 조사에 포함되었던 기업에 대해서는 R&D 지출의 크기를, 전년도 조사에 포함되지 않은 기업에 대해서는 R&D 지출 추정액의 크기를 기준으로 하여 확실성 기업 여부를 결정하고 있다. 1996년부터 총 표본 내 확실성 기업의 수를 제한하기 위해, 확실성 기업의 기준을 총 R&D 지출 1백만 달러에서 5백만 달러로 상향 조정하였다.

1994년 이후부터는, 표본추출틀을 대기업군과 소기업군으로 분할하고 있는데 총 종업원수가 이에 대한 기준이 되고 있다. 예를 들어, 제조업의 경우는 종업원수 50명 이상의 기업을 대기업군으로, 비제조업의 경우에는 종업원수 15명 이상의 기업을 대기업군으로 구분하고 있다. 1998년도 조사의 경우, 대기업군에는 약 55만개, 소기업군에는 약 130만개의 기업이 포함되어 있었다.

1996년도 이후로, 조사의 최종 표본추출틀에서 한 가지 더 조정된 것은 1992년~1994년도 조사에서 R&D 지출이 없다고 응답한 2자리수 SIC 산업들(zero industries)을 식별하는 것이며, 비록 이들 산업이 조사대상에는 포함되었지만 매우 적은 수의 기업들만 해당 산업에서 임의추출하고 있다.

1995년도 조사에서는, 대기업군과 소기업군 모두 각각 40개의 층을 구성하여 각 층에서 표본을 추출하였으나, 매 연도마다 소기업군의 영향으로 인해 산업별 집계치의 변동성이 커지는 문제가 있어 이를 극복하기 위해 1996년도부터는 소기업군의 경우 제조업과 비제조업의 2개 층으로 줄여 집계함으로써 자세한 산업별 자료는 집계되지 않으며 제조업 및 비제조업으로만 집계되고 있다.

3) 자료의 수집 및 무응답의 처리

조사표는 매년 3월에 각 기업대표에게 발송하여 5월 15일까지 작성할 것으로 요구한다. 5회에 걸쳐 우편으로 조사표의 반송을 요청하며, 전년도 조사에서 밝힌 R&D 지출의 순위상 300대 기업에 대해서는 응답을 하지 않는 경우 전화독촉도 병행한다. 조사표는 두 가지 종류를 사용한다. 잘 알려진 대규모 R&D 수행기업에게는 RD-1 양식의 조사표를 보내는데, 매출액, 총 고용인원, 과학자와 공학자의 고용수준, 재원별 R&D 지출(연방정부, 자체, 기타), 연구특성(기초연구, 응용연구, 개발연구), 외국에서의 R&D 지출 등의 정보에 대해 기입하길 요구한다. RD-1 양식을 받는 기업들은 전년도 조사에 참여하였으므로 전년도에 조사된 전산입력 자료가 참고로 제공되는데, 전년도 통계치에 문제가 있다면 수정할 것도 요구받는다. 반면에 소규모 R&D 수행기업과 처음으로 표본에 포함된 기업에게는 RD-1A 양식의 조사표를 받게 되는데, RD-1 양식과 비교할 때 기업에 대한 연방정부의 R&D 지원액, 지출항목별 R&D 비용, 주별 국내

R&D 지출액, 에너지관련 R&D 지출액, 국가별 해외 R&D 지출액의 5개 문항이 제외된다.

무응답은 조사에 참여하지 않아 하나의 관찰단위 전체가 무응답인 단위 무응답(unit nonresponse)과 조사에는 참가했지만 특정한 문항에 대해서 대답을 하지 않는 항목 무응답(item nonresponse)으로 구분된다. 단위 무응답과 관련하여, 1998년에 조사된 기업들 중에서 13.2%의 기업들이 응답하지 않았는데, 대체로 무응답 기업에 대한 다른 정보로부터 R&D 자료를 추정할 수가 있기 때문에 단위 무응답의 문제는 무시할만한 것으로 판단된다. 항목 무응답의 경우, 기업들에게 실제 자료가 가용하지 않은 경우에는 정보를 추정하도록 요구하지만 1998년도 조사의 경우 항목 무응답율은 1.6%에서 68.3%에 걸쳐 발생하였으므로, 추가조사를 통해 자료를 수집하거나 아니면 결측치가 발생한 항목의 값을 무응답 기업이 속한 산업의 해당 항목의 평균변화율을 이전 조사에서 보고되거나 대체된 값에 적용하여 그 값을 계산하는 대체(imputation) 기법을 사용한다.

2. 일본의 기업부문 R&D 활동조사

조사는 우편을 통해 이루어진다. 설문지는 표본조사대상 기업에게 직접 발송되어 관련 내용을 기재한 후 다시 우편으로 반송받는다. 1999년도 조사의 경우, 기업은 자본금 규모와 일본의 표준산업분류에 따라 30개의 층(strata)으로 세분화된다. 각 층에서 특정 개수의 기업을 조사대상으로 선정한다. 자본금 규모는 크게 10억엔 이상과 10억엔 미만으로 구분된다. 표준산업분류의 대분류는 농업, 임업, 어업, 광업, 건설업, 제조업, 전기 · 가스 · 열공급 · 수도업, 운수 · 통신업, 방송 · 컴퓨터 프로그램 · 기타 소프트웨어 서비스업이다. 표집비율(sampling ratio)은 1/1에서 1/400에 이른다.[4)]

자본금 10억엔 이상의 기업에 대해서는 전수조사를 하며, 비록 자본금이 10억엔 이하라 하더라도 전년도에 R&D 활동을 수행한 기업에 대해서도 전수조사를 한다. 아울러 공기업 및 신설기업 중 의약품 제조업, 소프트웨어업에 해당하는 기업도 전수조사를 한다. 반면에 전년도에 R&D 활동을 수행하지 않았거나 전년도 조사에 포함되지 않은 기업 중에서 자본금 10억엔 이하인 기업들에 대해서는 표본조사를 한다. 아울러 자본금 10억엔 이하의 신설기업도 표본조사의 대상이다.

최종적인 조사대상 기업의 개수는 1999년 조사의 경우 약 12,400개에 달한다. 이 중에서 약 80%로부터 설문지를 회수했으며, 2,500개 무응답 기업 중에서 부표본(sub-sample) 800개를 다시 선정하여 모두 설문지를 회수하였다. 자료를 집계시, 조사결과를 각 층의 추출율의 역수로 곱하여 추계하였다. 무응답 기업 중에서 다시 선정된 기업인 부표본에 대해서는, 부표본비율의 역수를 결과에다 곱한다. 연구기업과 대상에 대해서는 모든 기업에 대해 조사된 결과를 단순하게 합하면 된다.

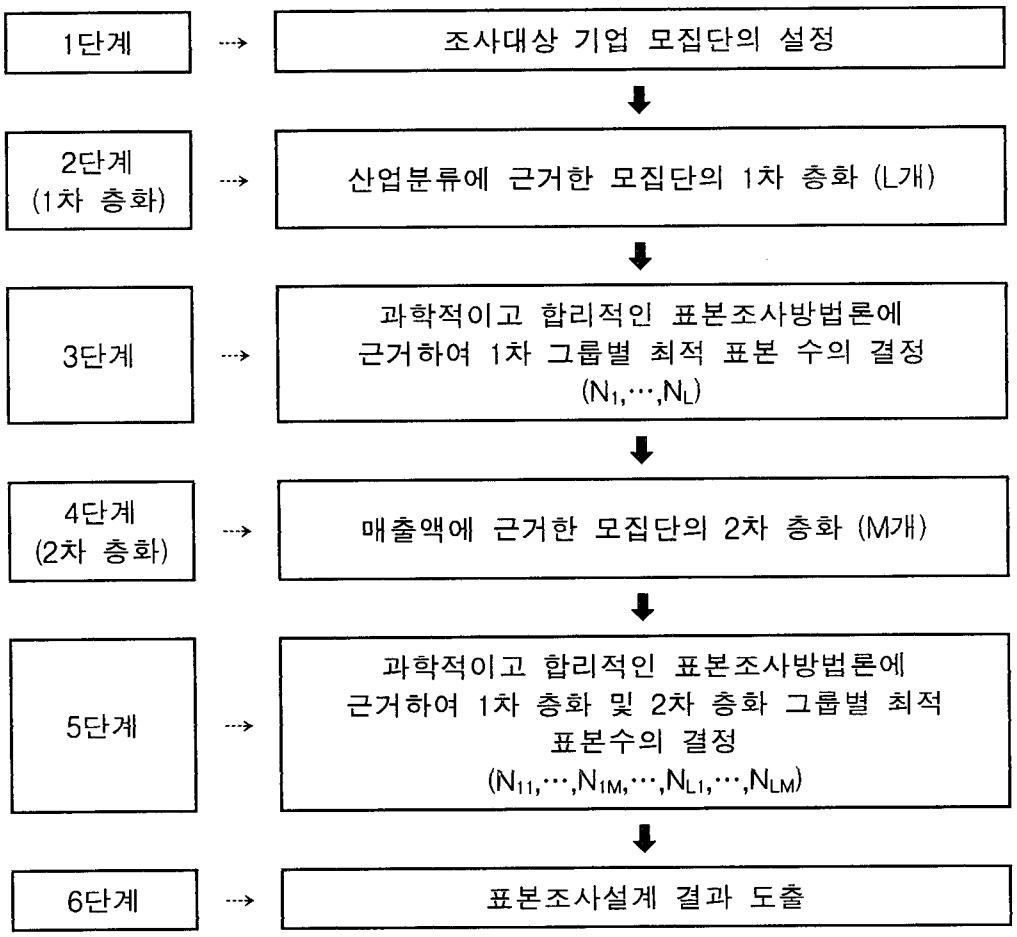
Ⅲ. 표본조사의 도입방안

1. 표본조사 도입의 개요

서론에서 언급했듯이 기업부문 과학기술연구개발

4) 보다 자세한 내용은 지면의 제약상 생략하지만 필요시 첫 번째 저자에게 요청가능하다.

활동조사의 경우 표본조사의 도입이 요구되고 있으며, 기술선진국인 미국 및 일본의 사례를 보더라도 표본조사를 사용하고 있다. 본 연구에서는 표본조사의 방법으로는 층화표본추출(stratified random sampling)의 적용을 제안한다. 본 연구에서 제안하는 층화표본추출법의 운용절차를 도식적으로 정리하면 [그림 1]과 같다. 먼저, 1단계에서는 조사대상 기업 모집단을 정의한다. 이를 위해서는 기업정보를 담고있는 다양한 원천의 자료들을 검토해야 한다. 2단계는 산업분류를 기준으로 1차 층화를 수행하고, 다음 단계에서 1차로 층화된 L개의 그룹, 즉 각 층에 대해 적절한 개수의 표본을 설정한다. 이를 위해서는 과학적이고 합리적인 표본조사방법론에 근거하여 최적 표본수가 도출되어야 할 것이다. 4단계에서는 1차로 층화된 그룹에 대해 2차 층화를 한다. 이 때 층화의 기준으로는 종업원 수, 자본금, 매출액 등을 고려할 수 있는데 본 연구에서는 매출액을 기준으로 할 것을 제안한다. 5단계에서는 표본조사방법론에 근거하여 1차 및 2차 층화된 그룹별 최적 표본수를 결정한다. 마지막 6단계는 1단계에서 5단계까지의 과정을 통해 도출된 최적 표본조사 설계결과를 도출하여 운용가능한 형태로 제시한다.



〔그림 1〕 표본조사기법의 도입

2. 표본조사의 설계

1) 층화의 방법[5)]

본 연구에서는 앞서 언급하였듯이, 실증연구에서 그리고 중소기업청(2001), 중소기업연구원(2001), 한국은행(2001) 등 실제조사에서 가장 널리 사용되는 층화표본추출 방법을 이용하고자 한다. 이 방법을 사용하는 목적은 유사한 성질을 갖는 원소들을 그룹으로 묶어서 분석하는 데 있다. 기본적 원칙은 층안에 있는 원소들은 가능한 한 같은 성질을 가져야 하며, 서로 다른 층에 있는 원소들은 가능한 한 서로 달라야 한다는 것이다(김종호, 1998; 남궁평, 2001; 신민웅 · 이상은, 2001).

〈표 2〉 통계청 표분산업분류의 재분류

층화코드	산업분류명칭	층화단위
2	광업	10, 11, 12
3	음식료품	15
4	담배	16
5	섬유	17
6	의복	18
7	가죽, 신발	19
8	목재및나무	20
9	펄프종이제품	21
10	인쇄출판	22
11	코크스, 석유정제품및핵연료	23
12	화합물및화학제품(의약품제외)	24
13	의약품	242
14	고무및플라스틱	25
15	비금속광물	26
16	제1차금속산업(제1차철강산업제외)	27
17	제1차 철강산업	271
18	조립금속제품	28
19	기타기계장비	29
20	컴퓨터및사무용기기	30
21	전기기계	31
22	전자(반도체, 전자부품제외)	32
23	반도체및전자부품	321
24	의료정밀측정광학기기	33
25	자동차및트레일러	34
26	기타운송장비(선박, 항공기제외)	35
27	선박	351
28	항공기, 우주선 및 부품	353
29	가구제조업	361
30	기타제품제조업	369
31	재생용가공원료생산업	37

5) 표본설계를 위해서는 먼저 모집단이 정의되어야 하는데, 아직까지 연구개발활동조사에서는 모집단 리스트 구축에 필요한 시간, 비용 제약 등 현실적인 어려움으로 인해 모집단이 확보되지 못한 상황이다. 따라서 본 연구에서는 향후 연구개발활동조사의 모집단이 정의되었을 때 직접 적용될 수 있도록 표본조사방안을 설계하는데 중점을 두고자 하며, 추가적으로 모집단 확보를 위한 방안은 V장에서 제시하고자 한다.

우선 표본의 설계를 위해 1차 층화 기준을 결정해야 한다. 본 연구에서는 한국은행, 중소기업청 등의 사례를 참고하여 산업별 분류를 1차 층화의 기준으로 삼고자 한다. 현재 과학기술연구활동조사에서 활용할 수 있도록 통계청에서 채택하고 있는 광공업부문의 표준산업분류를 적절하게 통합 및 분리하면서 재분류하면 <표 2>와 같이 정리될 수 있다.[6] 1차 층화된 군의 개수는 총 30개로서, 너무 세분화된 분류는 표본설계에 있어서 바람직하지 않으므로 비교적 유사한 산업은 통합하는 것을 원칙으로 하되, 과학기술연구활동조사에서 충분히 활용할 수 있도록 산업을 적절하게 세분화한다.

2차 층화는 매출액, 총자산액, 자본금의 세 변수와 R&D 지출액 사이의 상관계수를 계산하여 상관계수의 값이 가장 큰 변수를 기준으로 하는 것이 바람직하다. 2001년에 수집된 기업체 연구개발활동조사 자료를 이용하여 상관계수를 계산한 결과는 <표 3>에 요약되어 있다. 매출액, 총자산액, 종업원수, 자본금의 4개 변수와 R&D 지출액과의 상관계수는 각각 0.7470, 0.0562, 0.6807, 0.2544로 계산되었다. 매출액의 경우가 0.7470으로 가장 크며, 종업원수의 경우가 0.6807로 두 번째로 큰 값을 가진다. 반면에 총자산액은 0.0562로 대단히 작은 상관계수 값을 가진다. 이로써, 2차 층화에 적절한 변수로는 매출액을 상정할 수 있다. 매출액 규모를 10억원 미만, 10억원~100억원 미만, 100억원~1,000억원 미만, 1,000억원 이상의 4개 군으로 분류한다. 이러한 분류는 KISTEP의 과학기술연구활동조사보고서 상에서 매출액 규모별 현황을 집계하고 있는 것과도 일관성을 유지한다.

3) 표본추출틀의 정립

2000년도 과학기술연구활동조사의 경우 매출액 1,000대 기업은 100% 회수를 하고 있다. 이것은 일본에 있어서도 마찬가지이다. 일본은 자본금이 일정 규모 이상이 되는 기업들에 대해서는 표본추출율을 100%로 하여 전수조사를 하고 있다. 따라서 매출액이 일정 규모(1,000억원) 이상이 되는 기업들에 대해서는 전수조사를 하는 것이 바람직한 것으로 판단된다.

통상 층화표본추출 방식에서는 1차 층화기준에 따른 기준별 표본업체수를 결정한 후에, Neyman 배분방식을 이용하여 이를 다시 2차 층별로 배분하며 본 연구에서도 이러한 방식에 따른다(이기재 · 전종우, 1997; 윤기주 · 박상언, 1998; 박진우, 1999; 황진수, 1999; 류제복 · 이계오 · 김영원, 2001). 우선 산업별

〈표 3〉 각 주요 변수들간의 상관계수

구 분	매 출 액	총자산액	종업원수	자 본 금	R&D 지출
매 출 액	1.0000				
총자산액	0.1035	1.0000			
종업원수	0.8131	0.0875	1.0000		
자 본 금	0.4252	0.0700	0.4098	1.0000	
R&D 지출	0.7470	0.0562	0.6807	0.2544	1.0000

6) 본 연구에서는 실제조사에서 우선적으로 통계청의 광공업센서스 기업을 모집단으로 하여 표본조사를 시범적용할 수 있도록 광업 및 제조업을 대상으로 산업분류를 층화하였다.

표본업체수를 다음 식에 근거하여 결정한다.

$$n_i = \frac{(\sum_j N_{ij} \cdot S_{ij})^2}{N_i^2 D^2 + \sum_j N_{ij} \cdot S_{ij}^2} \quad (1)$$

단, $n_i = i$ 산업의 표본 사업체수

$N_i = i$ 산업의 부모집단 사업체수

$N_{ij} = i$ 산업 j층(매출액) 부모집단 사업체수

$S_{ij} = i$ 산업 j층(매출액) 모표준편차

$D = d/z$ 으로 정의되는 목표정도로 d는 허용오차 및 z는 신뢰계수

여기서, 신뢰도를 95%로 하면 z는 1.96이 된다. 이는 추정량이 95%의 확률을 가지고 추정오차의 한계내에 놓이도록 한다는 의미이다. 허용오차(d)는 각 산업의 부모집단 사업체들의 평균값의 5%로 한다.

각 산업 내에서의 매출액 규모별 표본업체수의 배분은 다음 식에 근거하여 결정할 수 있다.

$$n_{ij} = \frac{N_{ij} \cdot S_{ij}}{\sum N_{ij} \cdot S_{ij}} \cdot n_i \quad (2)$$

단, $n_i = i$ 산업의 표본 사업체수

$n_{ij} = i$ 업종 j층 표본 사업체수

$N_{ij} = i$ 업종 j층 부모집단 사업체수

$S_{ij} = i$업종 j층 표준편차

3. 모집단 추계방법 및 표본의 관리

1) 층별 모집단 추계

설계된 표본조사에 따라 회수된 조사표에 근거하여 i 번째 산업의 j 번째 매출액 규모 모집단에 대한 추정치는 다음 식과 같이 계산된다.

$$X_{ij} = \sum_{k=1}^{n_{ij}} x_{ijk} \cdot \frac{N_{ij}}{n_{ij}} \quad (3)$$

여기서,

$X_{ij} = i$산업 j번째 매출액 규모에 대한 모집단 추정치

$x_{ijk} = i$ 산업 j 번째 매출액 규모 k번째 표본조사업체에서 수집된 관측값

$n_{ij} = i$ 산업 j 번째 매출액 규모군의 표본 사업체수

$N_{ij} = i$ 산업 j 번째 매출액 규모군의 모집단 총 사업체수

이로써 $X_{ij}(i = 1, 2, \cdots, 31, j = 1, 2, \cdots, 4)$의 모집단 추정치를 얻게 된다. 다음으로 i번째 산업군에 대한 모집단 추정치를 다음 식으로 구한다.

$$X_i = \sum_{j=1}^{4} X_{ij} = \sum_{j=1}^{4} \sum_{k=1}^{n_{ij}} x_{ijk} \cdot \frac{N_{ij}}{n_{ij}} \quad (4)$$

여기서, $X_i = i$ 번째 산업군에 대한 모집단 추정치

$X_{ij} = i$ 산업 j 번째 매출액 규모에 대한 모집단 추정치

이제 $i = 1, 2, \cdots, 31$ 에 대해 X_i의 값을 얻게 된다.

2) 전수조사 결과와의 결합

앞에서 매출액 1,000대 기업 등에 대해서는 전수조사를 수행할 것을 제안하였다. 따라서 앞서 제시된

절차에 따라 표본 응답값들을 산업별로 집계하여 부모집단에 대한 값을 추계하고, 이 값을 기업체에 대한 전수조사 결과와 합산하여 모집단에 대한 값을 산정한 후 이에 기초하여 각종 분석지표를 산출하게 된다.

산업별 모집단 추계치 = 표본의 의한 부모집단추계치 + 전수조사치

$$Z_i = X_i + B_i = \sum_{j=1}^{4} X_{ij} + B_i = \sum_{j=1}^{4}\sum_{k=1}^{n_{ij}} x_{ijk} \cdot \frac{N_{ij}}{n_{ij}} + B_i \quad (5)$$

여기서, $Z_i = i$ 번째 산업에 대한 추계치

$X_i = i$ 번째 산업군에 대한 모집단 추정치

$B_i = i$ 산업에 대한 전수조사치

$X_{ij} = i$ 산업 j 번째 매출액 규모에 대한 모집단 추정치

$x_{ijk} = i$ 산업 j 번째 매출액 규모 k 번째 표본조사업체에서 수집된 관측값

$n_{ij} = i$ 산업 j 번째 매출액 규모군의 표본사업체수

$N_{ij} = i$ 산업 j 번째 매출액 규모군의 모집단 총 사업체수

3) 표본의 관리

과학기술연구활동조사는 1회용이 아니라 매년 시행된다는 점에서 조사대상이 시간의 경과에 따라 변동하므로 이에 따른 표본관리가 필요하다. 특히, 시간의 흐름에 따라 기업들 중에는 휴·폐업, 신설 또는 이전하는 등 다양한 변동이 있을 수 있다. 또한 급속한 산업 발달로 인해 규모가 급하게 변하는 경우도 있다. 따라서 이러한 모집단의 변화를 반영하면서 대표성을 갖는 표본을 유지하기 위해서는 모집단의 변동을 지속적으로 파악하여 변동 내용이 표본에 반영될 수 있도록 관리 체제를 갖추는 것이 요구된다. 즉, 매년 조사되는 새로운 조사결과로부터 신규 사업체나 휴·폐업된 사업체를 파악하여 모집단을 지속적으로 보완하고, 모집단 사업체의 증감이 발생한 경우에는 산업분류별, 매출액 규모별 표본사업체 추출률을 변경하여 표본이 보완되도록 해야 한다. 부득이하게 표본사업체에 대한 휴·폐업, 전출 등에 의해서 조사가 불가능한 경우에 대비하여 20%의 예비표본을 추가하여 표본 사업체 명부를 작성하는 것이 좋은 방안이다.

Ⅳ. 무응답자료의 처리

1. 무응답 자료의 처리 원칙

무응답 자료처리를 위해서는 여러 가지 무응답 자료 처리 기법 중 연구에 사용할 특정한 기법을 결정하는 것이 필요한데, 이 결정에 있어서 가장 중요한 것은 어떤 근거에 의해서 처리기법의 체계를 형성할 것인가를 명확하게 밝히는 것이고, 이 근거는 방법론 결정에 있어서 경제적·사회적·통계적 원리라고도 할 수 있다. 방법론 결정 이전에 이것이 명확하게 제시되지 않는다면 무응답 자료 처리에 있어서 일관성을 상실하기 쉽다(김유진·이승욱, 1994; 김영원·조선경, 1996; 염준군·손창균, 1998).

따라서 무응답 자료 처리기법의 마련에 있어서 몇 가지 중요한 원칙들을 세우고자 한다. 아래에서 자세하게 설명될 여러 원칙을 받아들이는 데 있어서 한 가지 주의해야 할 사실은 이 원칙들이 서로 어느 정

도 상충(trade-off) 관계에 있을 수 있다는 것이다. 즉, 어느 하나의 원칙을 강조하다 보면 다른 기준을 소홀히 할 수 있으며, 여러 개의 원칙을 동시에 만족시킬 수 없다는 것이다. 따라서 본 연구에서는 모든 원칙을 동일한 중요도로 받아들이기보다는 상황에 따라 여러 원칙들을 적절하게 우선순위를 정해 받아들이고자 한다. 이와 관련하여 크게 4가지 기본 원칙을 설정하였다.

첫째, 과학적인 기준에 근거한 무응답 자료 처리기법의 확립이다. 표본조사론의 견지에서 볼 때, 문헌에서 소개된 기법에 근거하여 무응답 자료처리 기법을 결정해야 한다. 따라서 임의로 추론되거나 문헌에서 잘 발견되지 않는 기법의 선택은 지양한다. 둘째, 단순하고 명시적이며 기법의 운용이 용이한 무응답 자료 처리기법의 정립이다. 아무리 이론적으로 뛰어난 기법이라 하더라도 실무자가 운용하기에 지나치게 복잡하거나 애매 모호하면 안 된다. 실무자가 자료분석 또는 집계시 비교적 짧은 시간에 쉽게 운용할 수 있는 무응답 자료 처리기법을 선택해야 한다. 세번째 원칙은 정책당국자가 이해하고 수용할 수 있는 공평하고 합리적인 무응답 자료 처리기법의 도출이다. 과학기술부 해당 부서의 정책당국자들이 충분히 이해하고 수용할 수 있는 취지를 가진 무응답 자료 처리기법을 도출해야 한다. 마지막 원칙은 주요 선진국에서 운용되고 있는 실태를 반영한 무응답 자료 처리기법의 선택이다. 기술선진국에서 운용되고 있는 무응답 자료 처리기법에 대해 살펴보고 이를 참고한다면 우리에게도 유용할 수 있다.

2. 단위 무응답 자료의 처리

단위 무응답의 문제를 다루는 전략을 수립하는 데 있어서 단기적인 전략과 장기적인 전략으로 구분하여 살펴볼 필요가 있다. 단기적으로는 무응답 기업에 대해 다른 정보를 수집하여 중요한 변수들을 추정하는 방법을 사용하는 것이 바람직해 보인다. 이렇게 되면 단위 무응답의 문제는 결국 항목 무응답의 문제로 귀결되게 된다. 장기적으로는 조사대상 수가 적은 부문은 회수율을 100%에 가깝게 유지하는 전략이 요구되며 조사대상 수가 많은 부문은 표본조사를 도입하는 전략이 요구된다.[7] 즉, 조사대상 수가 적은 시험연구기관, 의료기관, 대학의 경우에는 회수율을 100%로 올릴 필요가 있다. 아울러 기업체의 경우에는 표본조사 방법을 도입하고 표본조사시에 무응답 기업에 대해서는 추가표본을 추출하여 보완하는 것이 요구된다.

3. 항목 무응답 자료의 처리

항목 무응답의 처리를 위해서는 간접적으로 누락치를 추정하는 대체기법을 운용하는 것이 가장 적절해 보인다(배성우 · 오형재, 1999). 직접적으로 추정하는 방법은 지나치게 복잡하여 실무에서 활용하기 어렵기 때문이다. 대체기법의 개발 및 적용을 위해서는 항목의 무응답이 있는 기업이 이전년도의 조사에 포함되었는지 여부를 판단하는 것이 요구된다. 즉, 이전년도에 조사된 기업의 경우와 조사에 한 번도 포함되지 않았던 기업의 경우로 구분하여 살펴본다.

7) 일차적으로 무응답기업에 대해서는 중요항목에 대한 전화조사를 실시하거나 전술한 표본조사방법을 적용한 2차 표본조사 등을 활용함으로써 회수율을 최대한 높이는 것이 가장 중요하다.

1) 이전년도의 조사에 포함된 기업의 경우

만약 당해연도의 조사에서는 중요 항목에 대해 응답을 하지 않았지만 이전년도의 조사에서는 응답한 자료가 있다면 비율 대체기법의 운용이 비교적 편리하고 유용하다. 이 기법은 R&D 지출과 같은 주요 총량 변수를 추정하기 위해 인플레이터(inflator) 또는 디플레이터(deflator)를 사용한다. 아울러 이 총량 변수의 값은 알고 있는데 세부 항목의 값을 모른다면 세부 항목으로의 배분을 위해 기준 분포로서 이전년도에 조사된 자료를 이용한다. 이 기법은 보다 구체적으로 다음과 같은 수학적 형태를 취한다.

$$\hat{y}_{ik_t} = \hat{B}_{k_t} \cdot y_{ik_{t-1}} \tag{6}$$

여기서, $\hat{y}_{ik_t} = t$ 연도의 기업 i 에 있어서 주요 변수 y_k의 추정값

$y_{ik_{t-1}} = t-1$ 연도의 기업 i 에 있어서 주요 변수 y_k 의 실제값

$\hat{B}_{k_t}$ = 주요 변수 y_k 에 대한 인플레이터 또는 디플레이터

인플레이터 또는 디플레이터는 무응답으로 인한 대체가 전혀 이루어지지 않은, 즉 해당 항목을 완전하게 응답한 기업들의 자료만을 이용해서 계산된다. 만약 r 을 유사한 군에 속하는 기업의 개수라고 한다면 $\hat{B}_{k_t}$는 다음과 같이 계산될 수 있다.

$$\hat{B}_{k_t} = \frac{\sum_{j=1}^{r} y_{jk_t}}{\sum_{j=1}^{r} y_{jk_{t-1}}} \tag{7}$$

여기서, $\hat{B}_{k_t}$ = 주요 변수 y_k 에 대한 인플레이터 또는 디플레이터

$y_{ik_t} = t$ 연도의 기업 i 에 있어서 주요 변수 y_k 의 실제값

$y_{ik_{t-1}} = t-1$ 연도의 기업 i 에 있어서 주요 변수 y_k 의 실제값

R&D 지출과 같은 관심대상 변수에 대한 인플레이터 또는 디플레이터는 산업별로 분리되어 계산된다. 또는 산업별 · 매출액 규모별로 보다 세분화하여 계산될 수 있다. Ⅲ장에 제시된 표본설계의 기준들을 활용할 수 있다. 즉, 주요산업별로 31개 군, 매출액 규모별로 4개 군이 총 124 그룹에 대해 서로 다른 인플레이터 또는 디플레이터를 계산할 수 있다. 다시 말해서 무응답 기업에 대한 현재 연도의 추정치를 얻기 위해 전년도의 주요 변수들에 대해 대체계수(imputation factors)를 적용한다. 완전 무응답 기업에 대해서는, 모든 주요 변수들이 추정되어야 할 것이다. 반면에, 부분적인 무응답 기업에 대해서는 누락된 주요 변수들만 추정하면 된다. 한 그룹 안의 응답된 자료의 개수 50개 이하로 지나치게 작아서 유의미한 대체계수를 얻는 것이 어렵다면, 합리적인 대체계수를 계산할 수 있도록 인접한 그룹과 통합하는 것이 요구된다. 예를 들어, 인접한 산업군 간에 또는 인접한 매출액 규모군에 대해 자료를 통합할 수 있다.

따라서 본 연구에서는 이와 같은 비율대체 기법을 주요변수들을 대체하는 데 사용할 것을 제안한다. 그런데 R&D 지출과 같은 주요 변수는 용도에 있어서 기초분야 대 응용분야, 자금원에 있어서 자체조달 대 외부조달 등 세분화하여 구분할 필요가 있게 된다.

즉, 주요 변수들은 수직적인 구조를 취하고 있어 상당히 많은 수의 하위 항목으로 구성되어 있다. 다른 예를 들어, 자체사용연구개발비의 사회경제적 목적별 구성은 농림수산, 산업개발, 에너지, 원자력, 교통, 통신, 도시 · 지역개발, 환경보전, 보건, 공공개발 · 서비스, 지구 및 대기, 지식 증진, 우주개발, 국방 등으로 세분화되어 있다. 다음의 관계식을 이용하여 주요 변수들로부터 하위항목들에 대한 정보를 이끌어내는 것이 필요하다.

$$\hat{y}_{ij_t} = \hat{y}_{ik_t}\left(\frac{y_{ij_{t-1}}}{y_{ik_{t-1}}}\right) \quad (8)$$

여기서, $\hat{y}_{ij_t} = t$ 연도의 기업 i 에 대한 세부변수 y_j 의 추정값

$\hat{y}_{ik_t} = t$ 연도의 기업 i 에 대한 주요변수 y_k 의 추정값

$y_{ij_{t-1}} = t-1$ 연도의 기업 i 에 대한 세부변수 y_j 의 실제값

$\hat{y}_{ik_t} = t-1$ 연도의 기업 i 에 대한 주요변수 y_k 의 실제값

그런데 특정 기업에 있어서는, 세부변수에 대한 전년도 실측치가 없을 수 있다. 이러한 기업에 대해서는 전년도 이전의 자료를 활용하거나 이마저 여의치 않다면 비슷한 위치에 있는 다른 기업의 자료를 활용하는 것이 필요하다.

2) 조사에 한 번도 포함되지 않았던 기업의 경우

앞서 언급하였듯이, 본 연구에서는 특정 기업에 있어서 주요 변수에 대한 값이 없다면 이의 대체를 위해 이전 연도의 주요 변수를 부풀리거나 혹은 축소시키는 비율 대체기법을 제안했다. 하지만, 이전 연도의 자료가 없는 기업에 대해서는 이러한 대체 기법을 적용하는 것이 불가능하다. 이러한 경우에는 <표 4>에 제시된 바와 같이 크게 네 가지 방안을 고려할 수 있다(Little, 1982; Rubin, 1987).

첫째, 이전에 조사에 한 번도 참여하지 않거나 완전히 혹은 부분적으로 응답하지 않은 기업에 대해서는 주요 변수들에 대해 0의 값을 부여하는 것이다. 이 방안이 설득력을 얻기 위해서는 이들 기업들은 전체에서 차지하는 비중이 비교적 작아 0의 값을 부여했

〈표 4〉 이전년도 조사에 포함되지 않은 경우의 항목 무응답 처리기법

구 분	내 용
1. 0의 값을 부여	- 가장 단순하고 편리한 처리방식 - 무응답기업이 전체에서 차지하는 비중이 작아야 유의미함
2. 평균대체 방법	- 전체 평균대체방법과 계급별 평균대체방법으로 구분됨 실제의 분포를 덜 왜곡한다는 측면에서 계급별 평균대체방법이 더 선호됨 - 계급별 평균대체방법은 전체 평균대체방법에 비해 작업량이 더 많음
3. 상관관계가 높은 보조변수를 이용한 비율 대체	- 상관관계를 따져 보아 이것이 높은 변수를 이용하여 무응답 자료를 비율 대체 - 상관관계를 계산하여야 하며 층화표본추출시 쉽게 적용 가능
4. 회귀식을 이용한 대체	- 복잡한 회귀분석을 해야 함 - 추정식의 적합도가 높지 않다면 예측오차는 커짐

다 하더라도 조사결과 전체의 전반적인 정확성에 별 다른 영향을 미치지 않아야 한다(김규성, 2000).

둘째, 평균대체방법을 운용할 수 있다. 평균으로 대체하는 방법은 크게 전체 평균대체방법과 계급별 평균대체방법으로 구분된다. 전체 평균대체방법이란 결측치를 응답된 전 기업의 평균으로 대체하는 방법으로 비교적 단순하고 적용이 쉽다. 하지만 이 방법은 원자료의 분산을 지나치게 과소평가하여 실제의 분포를 크게 왜곡시킬 수 있다는 단점을 가지고 있다. 반면에 계급별 평균대체방법은 비교적 유사한 계급, 즉 그룹에 속하는 자료의 평균으로 대체하는 방법으로 전체 평균대체방법에 비해서는 일반적으로 보다 효율적이다. 예를 들어, Ⅲ장에서 제시한 표본설계의 지침대로 124개의 세부 그룹을 조성하여 해당 그룹의 평균으로 무응답 자료를 대체하는 것을 고려할 수 있다.

셋째, 관심대상 변수와 비교적 상관관계가 큰 다른 변수에 대한 정보를 이용하고 응답한 기업에서의 두 변수의 관계를 살펴봄으로써 관심대상 변수의 값을 유추하여 대체할 수 있다. 예를 들어, R&D 지출에 대한 정보는 없지만 매출액에 대한 정보는 구할 수 있는 기업이 있다고 하자. 앞서 분석하였듯이, R&D 지출 변수는 매출액과 상관관계가 비교적 높다. 따라서 R&D 지출과 매출액 모두를 응답한 B 기업의 값을 이용하여 R&D 지출액이 누락된 A 기업에 대해 R&D 지출액을 추정해 낼 수 있다. 이를 수식으로 나타내면 다음과 같다.

$$RD_A = Sales_A \cdot \frac{RD_B}{Sales_B} \tag{9}$$

여기서, RD_A = 기업 A 의 R&D 지출

$Sales_A$ = 기업 A 의 매출액

RD_B = 기업 B 의 R&D 지출

$Sales_B$ = 기업 B 의 매출액

이때, 기업 B 는 한 개 기업을 의미할 뿐만 아니라 다수의 기업도 될 수 있다. 아울러 기업 B 를 표본 전체에 대해 정의할 수 있으며 보다 자세하게 한다면 기업 A 와 유사한 그룹에 속하는 것으로 볼 수도 있다.

넷째, 세 번째 방안을 보다 일반화시킨다면 회귀방정식을 추정하여 누락된 변수를 대체하는 방법도 고려할 수 있다. 예를 들어, 회귀분석을 통해 다음의 식이 추정되었다면 매출액 정보를 이용하여 R&D 지출에 대한 정보를 얻을 수 있다. 현실적인 적용을 꾀할 수 있도록, KISTEP 기업체 조사 자료 4,155개를 이용하여 매출액과 R&D 지출 사이의 관계식을 다음과 같이 몇 가지로 정형화하고 추정하였다. 이 함수형태들은 실증문헌에서 흔히 사용되는 것들로서 각각, 선형, 선형-로그형, 로그-선형, 로그-로그형, 이차형으로 불린다.

$$RD_i = a + b \cdot Sales_i \tag{10}$$

$$RD_i = a + b \cdot ln(Sales_i) \tag{11}$$

$$RD_i = exp(a + b \cdot Sales_i) \tag{12}$$

$$RD_i = exp(a + b \cdot ln(Sales_i)) \tag{13}$$

$$RD_i = a + b \cdot Sales_i + c \cdot Sales_i^2 \tag{14}$$

여기서, RD_i = 기업 i 의 R&D 지출

$Sales_i$ = 기업 i 의 매출액

a, b = 모수 추정치

추정결과는 <표 5>에 정리되어 있다. 추정계수는 유의수준 1%에서 모두 통계적으로 유의하다. 하지만 모형의 적합도를 나타내는 조정된-$R^2(\bar{R}^2)$의 값은 큰 차이를 보인다. 5개 추정식에서 이차형의 추정결과가 0.790으로 가장 높은 적합도를 보이고 있다. 다음으로 선형이 높으며 로그변수를 포함한 형태는 전반적으로 낮은 적합도를 가지고 있다. R&D 지출에 대한 정보가 누락된 자료에 기업에 있어서 매출액에 대한 정보는 구할 수 있다면 이차형 추정결과와 이 매출액 자료를 이용하여 R&D 지출 정보를 추정할 수 있다.

이렇게 수집된 자료만을 이용하여 여러 가지 회귀식을 추정한 후 가장 적합도가 높은 모형을 선정해서 자료를 대체하는 것도 한 가지 유용한 대체기법이다. 다만 회귀분석이라는 복잡한 과정을 거쳐야 하는 것이 단점이라면 단점일 것이다.

아울러 주요 총량변수에 대해서는 응답을 받았는데, 세부 항목 변수에 대해서는 무응답인 경우가 있다. 이 경우에는, 해당 기업이 제공한 다른 주요 변수값을 이용하거나 유사한 기업에서 제공한 자세한 숫자들에 근거하여 필요한 숫자들을 대체한다. 특히, 올해 자세한 자료를 완전히 보고한 기업과 전년도 자료를 이용하여 올해 대체된 자세한 자료를 가지고 있는 기업들의 분포로부터 비례분포계수(proportional distribution factors)를 만든다. 그러한 비례분포계수는 각 유사기업 그룹에 대해 계산될 수 있다. 다음으로 이러한 계수는 주요 변수값으로부터 세부 변수값을 대체하기 위한 기준분포로 사용된다.

V. 결 론

과학기술연구개발활동조사는 국가과학기술정책의 수립과 정부 및 민간 각 부문의 과학기술진흥 및 연구개발계획 수립 등에 필요한 기초자료를 제공할 목적으로 연구개발투자, 연구개발인력 등에 대한 실태를 알아보기 위해 1963년부터 매년 실시되고 있다. 이 조사의 대상은 크게 자연과학분야의 시험연구기관, 대학, 의료기관 및 대학부속병원, 기업체로 구분되는데, 특히 기업부문의 경우 조사대상이 전체 기업을 반영하지 못하고 회수율도 가장 낮다. 따라서 본

〈표 5〉 매출액과 R&D 투자와의 관계식 추정결과

추 정 식	추 정 계 수			
	a	b	c	$\bar{R}^2$
$RD_i = a + b \cdot Sales_i$	-1310.01 (-3.46)	0.0327 (72.43)		0.558
$RD_i = a + b \cdot ln(Sales_i)$	-19751.3 (-9.22)	2551.43 (10.88)		0.027
$RD_i = exp(a + b \cdot Sales_i)$	5.999 (301.9)	0.63E-06 (26.75)		0.147
$RD_i = exp(a + b \cdot ln(Sales_i))$	3.296 (47.97)	6.3154 (41.93)		0.297
$RD_i = a + b \cdot Sales_i + c \cdot Sales_i^2$	1606.03 (6.06)	-0.0049 (-7.62)	0.25E-08 (67.66)	0.790

주 : 괄호 안에 있는 값은 t-통계량이다.

연구에서는 기업부문을 중심으로 과학기술연구개발 활동조사의 개선방안을 마련하고자 하였다.

이를 위해 우선 기술선진국인 미국 및 일본의 사례를 살펴보았는데, 체계적인 표본조사 및 무응답 자료처리기법이 완비되어 있음을 확인할 수 있었다. 우리나라도 향후 지속적인 관심과 여건 조성을 통해 이들 국가의 사례를 참고하여 자료의 합리적인 통계처리 방안을 도입해야 할 것이다.

표본조사의 도입을 위해서는 우선 모집단의 목록 확보가 중요하다. 이를 위해서는 전체 기업의 R&D 활동 수행 여부를 알 수 있는 전수조사를 정기적으로 실시하여 모집단 목록을 확보하는 것이 가장 바람직해 보인다. 전수조사가 현실적으로 어려울 경우 통계청, 국세청 등의 협조를 통해 타조사의 리스트를 활용하는 방안도 고려해볼 수 있으나,[8] 이 경우 조사대상단위의 차이, 응답 내용의 차이 등을 충분히 고려하여 모집단 설계에 세심한 주의를 기울여야 한다.[9] 다음으로 정의된 모집단에 기초하여 표본을 추출하는 방법으로서 본 연구에서는 표본추출법으로 층화표본추출법을 제안하였으며, 1차 층화의 기준으로는 산업분류를 2차 층화의 기준으로는 매출액을 제시하였다.

과학기술연구활동조사는 1회성 조사가 아니라 매년 반복 시행되는 조사이기 때문에 시간의 경과에 따라 조사대상이 변동하므로 이에 따른 표본관리가 필요하다. 즉, 폐업, 휴업, 신설, 이전 등의 변동과 급속한 산업 발달로 인한 규모의 급격한 변화 등 모집단의 변화를 지속적으로 파악하여 이러한 변동 내용이 표본에 반영될 수 있도록 관리 체제를 갖추는 것이 요구된다. 따라서 매년 조사되는 새로운 조사결과로부터 신규 사업체나 휴·폐업된 사업체를 파악하여 모집단을 지속적으로 계속 보완하고, 모집단 사업체의 증감이 발생한 경우에는 산업분류별, 매출액 규모별 표본 사업체 추출률을 변경하여 표본이 보완되도록 해야 한다.

조사결과에 적지 않은 영향을 미칠 수 있는 무응답 자료는 적절한 기법을 운용하여 처리되어야 하는 바, 단위 무응답의 경우 단기적으로는 무응답 기업에 대해 다른 정보를 수집하여 중요한 변수를 추정하는 처리방안이, 장기적으로는 전수조사와 표본조사를 병행하여 무응답율을 줄이는 방안이 도입되어야 한다. 항목 무응답을 처리하기 위해서는 이전 연도의 조사에 포함되었던 기업의 경우와 조사에 한 번도 포함되지 않았던 기업의 경우로 구분하여, 전자의 경우에는 본 연구에서 개발된 비율대체기법을 적용하고, 후자의 경우에는 0의 값 부여, 평균대체, 상관관계나 높은 보조변수를 이용한 비율대체, 회귀식을 이용한 대체의 4개 방안의 적용을 고려할 수 있다.

이상의 연구결과는 모집단 정의의 어려움 등으로 인해 완전한 형태를 띄지 못한 측면이 있다. 하지만 과학기술연구개발활동조사가 보다 신뢰성있는 국가지정통계로서의 역할을 다할 수 있도록 한다는 측면에서 파일럿(pilot) 연구로서 몇 가지 유용한 지침을 제공하고 있다고 판단된다. 또한 본 연구에서 제시된 방안을 실제 조사에 도입함으로써 연구개발수행기업과 미수행기업의 산업별 특성 비교, 신규기업의 특성

8) 예를 들어 광공업부문은 통계청의 광공업센서스 기업을, 서비스부문은 도소매서비스업 총조사 기업을, 건설부문은 통계청의 건설업 통계조사기업을 모집단으로 활용할 수 있다.

9) 모집단으로 활용할 수 있는 통계청 조사는 대부분 사업장 단위를 조사단위로 하고 있는 반면, 연구개발활동조사는 기업을 조사단위로 하고 있다.

분석 등 보다 다양한 통계분석이 가능해질 것으로 기대되며 국가 연구개발통계의 신뢰성 제고에 기여할 것으로 생각된다. 추후 보다 완전한 형태의 여러 후속연구가 나오길 기대한다.

참 고 문 헌

김규성(2000), "무응답 대체 방법과 대체 효과", 조사연구, 제1권, 제2호, pp.1-14, 12월.

김영원 · 조선경(1996), "표본조사에서 항목무응답 대체 방법", 한국통계학회논문집, 제3권 3호, pp.145-150.

김유진 · 이승욱(1994), "표본조사에서 무응답처리를 위한 통계기법 고찰", 한국보건통계학회지, 제19권 1호, pp.80-87.

김종호(1998), 『표본조사법』, 자유아카데미.

남궁평(2001),『표본조사 설계와 분석』, 탐진.

류제복 · 이계오 · 김영원(2001), "2001년 국민건강 · 영양조사 표본설계", 응용통계연구, 제14권, 제2호, pp. 289-304.

박진우(1999), "수산물 비계통 생산량 조사를 위한 표본설계 연구", 응용통계연구, 제12권, 제1호, pp. 1-15.

배성우 · 오형재(1999), "표본조사에서 무응답자료처리를 위한 임퓨테이션 방법 등의 비교연구" 서울시립대학교 산업기술연구소 논문집, 제7집, pp.27-29

신민웅 · 이상은(2001),『표본조사를 위한 표본설계』, 교우사.

염준군 · 손창균(1998), "층화표본에서 단위 무응답에 대한 가중치 조정", 품질경영학회지, 제26권, 제3호, pp.82-83.

윤기중 · 박상언(1998), "가축통계 표본조사설계", 응용통계연구, 제11권, 제2호, pp. 233-246.

이기재 · 전종우(1997), "노동통계조사를 위한 표본설계 : 매월노동통계조사, 노동력수요동향조사를 중심으로", 응용통계연구, 제10권, 제2호, pp. 215-226.

이기재 · 최엄문 · 박성현(1998), "전국 지가변동율 조사를 위한 표본설계 연구", The Korean Communications in Statistics, Vol. 5, No. 3, pp. 675-684.

중소기업청(2001),『중소제조업 인력실태조사』.

중소기업연구원(2001),『중소기업 기술통계의 체계화 방안에 관한 연구』, 연구보고서, 중소기업청.

한국은행(2001), 기업경영분석.

황진수(1999), "산업재해의 효율적 분석을 위한 표본설계", 응용통계연구, 제12권, 제2호, pp. 363-374.

Little, R. J. A, (1982), Models for nonresponse in sample surveys, Journal of American Statistical Association, 77, pp.237-250.

Rubin, D. B. (1987), Multiple Imputation for Nonresponse in Surveys, New York: John Wiley & Sons, Inc.