I n d i a n a U n i v e r s i t y
University Information Technology Services

어떻게 자료를 입력하고 처리할 것인가: 자료분석 기초*

How Do I Input and Manipulate Data for Quantitative Analyses?

Hun Myoung Park, Ph.D.
kucc625@indiana.edu

Last modified on November 2009

University Information Technology Services
Center for Statistical and Mathematical Computing
Indiana University
410 North Park Avenue Bloomington, IN 47408
(812) 855-4724 (317) 278-4740
http://www.indiana.edu/~statmath

어떻게 자료를 입력하고 처리할 것인가: 자료분석 기초

이 문서는 초보자에게 어떻게 개념을 정의하고, 변수를 측정하고, 측정된 자료를 어떻게 자료분석 소프트웨어에 입력하고, 입력된 변수를 어떻게 정제하여 가공하고, 분석이 끝난 자료를 어떻게 처리하는가를 소개한다. 따라서 개념, 측정수준, 변수, 자료구조, 파일형식, 자료보관 등을 설명하지만, 실제 어떻게 자료를 분석하고 해석하는 것은 포함하지 않는다.

1. 무엇을 연구할 것인가?

연구하려는 대상對象과 그것을 보는 관점觀點을 먼저 생각해 보자. 연구대상은 설명하고자 하는 모집단 population 을 의미하는데, 구체화하여 분명하게 정의해야 한다. 수준을 따져 태양계인지, 대륙인지, 나라인지, 시도인지, 사람인지를 정한다. 공간을 고려하여 한국인인지 영국인인지 결정하고, 시간을 고려하여 백제시대 사람인지 고려시대 사람인지 특정해야 한다. 또한 연구목적에 따라 다르겠지만, 두리뭉실 한국인이라고 말하기보다는 보통 "2005 년 현재 한국인 20 대", "2005 년 현재 한국인 20 대 미혼남녀" 등이 더 낫다. 그렇다고 "2005 년 현재 한국인 20 대 미혼남녀"인 아랫마을 사는 갑돌이와 갑순이로 한다면 정말 구체화된 연구대상이긴 하나, 의미있는 대상이 되기 어렵다. 보통 사람들이 궁금해하고 관심을 보일만한 연구대상이 아니기 때문이다. 어쨌든 연구가치가 있는 대상을 수준, 공간, 시간 등에 대하여 정확하게 정의해 놓아야 한다.

연구대상을 분석하는 기본 단위를 분석단위 unit of analysis 라고 한다. 연구대상이 "2005 년 현재 한국인 20 대 미혼남녀"라면 분석단위分析單位는 20 대 미혼남녀 개개인이 될 것이다. 개개인뿐만 아니라 동네, 면, 군, 도 등을 분석수준으로 생각해볼 수도 있다.분석단위는 자료판 data set 을 구성하는 기본 단위인 관측단위 unit of observation 와 대개 일치하는데, 반드시 그러하지는 않다. 시도 단위로 측정된 20 대 미혼남녀에 관한 자료가 있다면 관측단위는 개개인이 아니라 시도이다. 단위가 서로 다른 경우에 추론 문제 ecological inference 가 생길 수 있다.

모집단 전체를 조사하는 일은 대개는 힘들고, 돈도 많이 들고, 시간도 많이 필요하다. 때로는 불가능하기도 하다. 모집단 일부만을 조사하여 모집단 전체가 어떠한지를 알 수 있다면 좋을 것이다. 그 모집단 일부는 물론 모집단 전체 특성을 반영할 수 있도록 대표성이 있어야 한다. 그런 모집단 일부를 표본 sample 이라고 한다. 1987 년 4.13 호헌조치 때 전두환이 대통령 직선제에 대한 국민의견을 알아보기 위해 이순자, 전경환, 노태우, 장세동에게 물어봤다면 큰 문제가 있다. 1987 년 현재 한국인이 모집단인데, "전두환 똘마니"들은 모집단의 특성을 대표하는 표본이라고 볼 수 없기 때문이다. 대운하나 4 대강 사업에 대한 국민의 의견을 알아보는 공청회에 대운하에 반대하는 사람을 참석하지 못하게 하는 것도 대표성이 없는 엉터리 표본을 가지고 자위自慰하는 짓이다. 모집단이 어떠한지와 전혀 관계없는 "어쨌거나 내 맘대로"일 뿐이다. 모집단과 표본은 대개 같은 분석단위를 가진다. 표본을 분석하여 모집단 특성을 추론해내는 학문을 통계학이라고 한다.

분석대상을 어떻게 볼 것인가? 어떤 입장에서 분석할 것인가? 연구 대상을 보는 나름의 관점 perspective 이 있어야 한다. 자신의 안경을 통하여 사물을 보는 것처럼, 대상을 보는 나름의 입장이 있어야 한다. 이론틀 framework 이나 이론 theory 이 그러한 것이다. 대상의 모든 면면을 한꺼번에 연구할 수 있으면 좋겠지만 현실적으로 가능하지 않다. 가치판단이 개입되어 있기는 하나 민주주의 틀에서 4 대강 사업을 바라볼 수도 있고, "삽질주의" 관점에서 볼 수도 있다. 설악산을 보고 싶다면 오색에 가도 되지만, 천불동계곡이나 백담사나 미시령에 가볼 수도 있다. 바라보는 곳에 따라 설악산의 다른 모습을 확인할 수 있을 것이다. 설악산 전체를 한눈에 보고 싶다면, 그것은 "하나주의자"들이나 빨간색 혹은 "콩사탕"만 봐도 거품물고 발작하는 "꼴통 어린이"들의 희망사항이요 부질없는 욕심일 뿐이다.

그러면 분석대상의 무엇을 볼 것인가? 관심있는 대상의 성질性質이나 현상現象을 생각해 보자. 가지고 있는 관점이나 이론理論에서 강조되는 성질과 현상이 있을 것이다. 자신이 특별히 관심을 가지고 있는 성질이나 현상이어도 괜찮다. 민주주의 틀에서 보면 4 대강 사업이 얼마나 정당한 절차를 거쳤는지를 따져볼 것이며, 일반 시민에게 어떤 도움을 주는지 해악을 끼치는지를 관심있게 볼 것이다. 설악산이면 한계령에 올라 불타오르는 단풍을 눈여겨 볼 수도 있고 암벽에 살고 있는 소나무의 식생을 살펴볼 수도 있다. "한국인 20 대 미혼남녀"라면, 전통혼례에 대한 그들의 선호라든가 "차떼기 추억"에 대한 분노를 알고 싶어할 수 있다. 돌에 대해서 말한다면, 크다 작다, 무겁다 가볍다, 검다 붉다, 거칠다 부드럽다 등을 생각할 수 있다. 그런 성질이나 현상에 절차 정당성, 사회 비용과 편익, 단풍, 식생, 선호도, 분노, 크기, 무거움, 색깔 같은 이름을 붙여보자. 이렇게 대상의 성질이나 현상에 붙인 이름을 개념 concept 이라 부르자. 이런 과정을 개념화 conceptualization 라고 한다.

개념槪念 자체는 만질 수도 볼 수도 없다. 추상화된 언어로 존재하기 때문이다. 그래서 사람들이 경험할 수 있도록 개념을 정의하여 측정하는 과정이 필요하다. 개념을 측정하여 담는 그릇을 변수變數 variable 라고 부른다. 자, 어떻게 개념을 변수화하여 측정할 것인가.

2. 어떻게 개념을 측정할 것인가?

연구관심이 되는 개념을 측정하는 방법은 여러가지가 있다. 우선 기존 문헌에서 얻을 수도 있다. 책이나 논문이나 보고서에 나온 자료를 참고할 수 있다. 신문을 들춰보면 설악산 단풍이 시작되는 시기를 연도별로 찾아낼 수 있다. 이미 처리된 자료를 참고할 수도 있고 처리되지 않은 원자료를 얻을 수도 있다. 심한 경우에는 여기 저기 흩어져 존재하는 자료를 모아 짜깁기를 해야 한다. 차떼기 추억에 관한 여러 설문조사를 이리저리 다듬어서 자료로 활용할 수 있다. 운이 좋으면 다른 사람이 만들어 놓은 자료를 그대로 얻을 수도 있다. 미시간대학에 있는 ICPSR (Inter-University Consortium for Political and Social Research)에서는 실제 연구에 사용된 자료파일을 아스키 형식 ASCII text 로 제공하고 있다. General Social Survey (GSS), American National Election Studies (ANES), Current Population Survey (CPS), Pew Internet and American Life Project 에서 수집한 자료는 인터넷에 공개되어 있다.

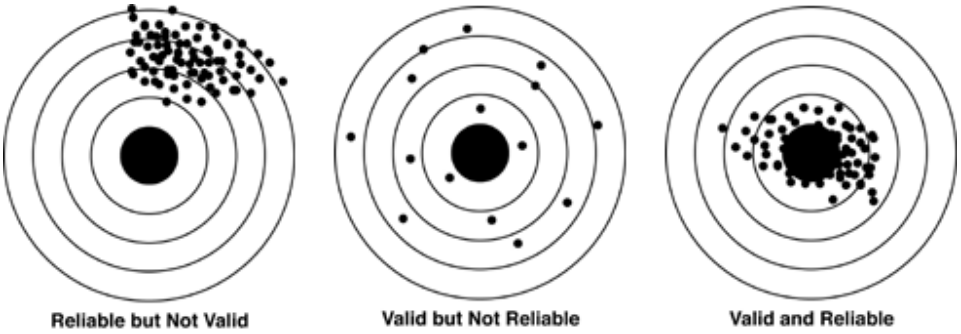
하지만 자신이 자료를 얻어야 하는 경우가 많다. 남이 만들어 놓은 자료를 이용하는 것보다 시간도 많이 걸리고 다리품도 많이 팔아야 하고 비용도 많이 들 것이다. 잘 만들어진 자료가 얼마나 소중한지를 몸소 체험할 수 있을 것이다. 우선 연구자가 직접 관측치 observation 를 재는 방법이 있다.[1] 또한 관측치의 반응을 설문조사 questionnaire survey 를 통해 알아볼 수도 있으며, 얼굴을 맞대고 물어서 interview 개념을 측정할 수도 있다. 현장에 나가서 관측치가 하는 행동을 관찰할 수도 있으며, 연구자 자신이 현상에 참여하여 participant observation 자신의 경험을 적어낼 수도 있다. 흔하지는 않지만 잘 통제된 실험실 상황 experimentation 에서 측정할 수도 있다.

[1] 여기서 관측치 observation 는 관측값 values of variables 에 해당하는 觀測値가 아니라 관측해서 나온 물건이나 개체라는 뜻으로 쓴다. 우리말에 "치"는 값 (수치), 몫 (한달 치) 등의 뜻으로도 사용하지만, 놈 (어린 치)이나 어느 곳에서 나는 물건 (고추는 청양치가…)이라는 의미로도 쓰인다.

측정을 하기 위해서는 개념을 구체화하여 정의할 필요가 있다. 개념을 측정할 수 있도록 조작화 operationalization 혹은 변수화해야 하는 것이다. 그 개념이 무엇인지 특정하고, 어떻게 재는 것인지 operational definition 를 정해야 한다. 개념을 재는 잣대 scale 와 측정 단위 unit 를 구체화해야 한다.

돌의 길이라고 한다면, 가장 긴 쪽으로 할 지 중간쯤 되는 위치에서 재는 것인지를 명확하게 해야 한다. 또한 cm 로 잴지, inch 로 잴 것인지를 정해야 한다. 나이를 정의한다고 했을때, 한국식 나이인가 서구식 나이인가, 실제 출생일 기준인가 주민등록증에 기록된 출생일 기준인가를 정해야 한다. 또한 연 단위로 할 것인가 월단위로 할 것인가, 아니면 일단위, 시간단위, 혹은 분단위로 할 것인가를 밝혀야 한다. 나아가 실제 측정된 나이로 분석할지 청년층—장년층—노년층으로 구분할지도 생각해야 한다. 그래서 개념의 성질이나 현상을 측정한 변수는 어떤 특정한 방법으로 측정되어 어느 특정 단위로 표시되어야 한다. 예컨대, 개인소득이면 2005 년 1 월 1 일 현재 세무서에 신고된 해당연도의 개인소득을 백만원단위로 표시한 것으로 정의할 수 있다. 그냥 개인소득이 100 이라고 하면 언제적 소득인지, 월소득인지 연소득인지, 100 원인지 100 달러인지 100 억원인지 알 수 없기 때문이다.

그림 1. 타당성과 신뢰성



Source: http://www.ccnmtl.columbia.edu/projects/qmss/meas_valrel.html

얼마나 잘 측정했는가는 흔히 타당성 validity 과 신뢰성 reliability 으로 평가한다. 타당성은 원래 측정하고자 하는 것을 얼마나 정확하게 측정했는가를 따진다. 신뢰성은 같은 측정도구를 사용했을때 얼마나 일관되게 개념을 측정할 수 있는가를 묻는다. 좋은 측정도구라면 누가 언제 어디서 측정하더라도 비슷한 stable and consistent 결과를 낼 것이다. 그림 1 을 보라. 첫번째 그림은 신뢰성은 있으나 타당성은 떨어진다. 과녁 중심을 맞추지 못하고 있다. 두번째 그림은 신뢰성도 떨어진다. 연습이 부족한지 원하는 곳을 겨냥해도 마음대로 맞추기 어렵다. 세번째는 신뢰성과 타당성 모두 높은 경우다. 대체로 과녁 중심을 향하고 있고 크게 벗어나지 않는다. 측정 타당성과 신뢰성을 따지기 위해서라도 개념을 적절하게 조작화하고 측정수준과 측정단위를 구체화할 필요가 있다.

3. 측정수준 level of measurement

개념을 측정하는 수준 혹은 잣대 scale 를 생각해 보자. 측정 수준에는 이름잣대, 순서잣대, 등간잣대, 비율잣대가 있다. 측정수준이 높을수록 더 많은 정보를 포함한다.

3.1 이름, 순서, 등간, 비율 잣대

이름잣대 nominal scale 는 관측값으로 이름을 사용한다. 사과, 배, 감, 밤, 대추와 같은 구분을 생각해 보라. 관측값 사이에는 순서도 없으며 의미있는 차이도 생각할 수 없다.[2] 사과를 1 로, 배를 2 로, 감을 3 등으로

[2] 물론 사과를 배보다 더 좋아한다거나 감이 대추보다 크다는 것이 연구질문에서 중요하다면, 그것은 과일을 구분하는 것이 아니라 과일에 대한 선호나 크기를 순서잣대로 측정하는 것이다.

입력한다고 해도 그 숫자는 편의상 이름을 대신할 뿐이다. 성별, 찬반과 같이 관측값이 두 개 밖에 없는 특별한 경우는 "두쪽변수" binary variable 라고 부른다.[3]

순서잣대 ordinal scale 는 관측값 순서가 의미있는 경우이다. 1 학년, 2 학년, 3 학년 등으로 분류하거나, 행정수도이전이 관습헌법에 위배된다는 판결에 대해 적극 반대한다—반대한다—찬성한다—적극 찬성한다로 물을 수 있다.[4] 축구경기 후 개별 선수들에 대한 평가를 Excellent (10)—Very good (8)—Good (6)—Fair (4)—Poor (2) 식으로 물을 수도 있다. 우열 정도에 따라 백합반—장미반—들꽃반으로 나눴다고 해도 마찬가지이다. 이런 경우에는 순서를 말할 수 있지만 관측값 사이에 의미있는 차이가 있다고 말할 수 없다. 즉, "적극 반대"와 "반대" 간 차이가 "반대"와 "찬성" 간 차이보다 작거나 크다는 식으로 말할 수 없다. 이름잣대와 순서잣대는 흔히 분류변수 categorical variable 혹은 마디변수 discrete variable (마디 단위로 끊는다는 의미에서)라고 부른다.

등간잣대 interval scale 는 계량분석에서 가장 자주 사용되는 측정수준이다. 관측값 사이에 순서도 있을 뿐만 아니라 관측값 간 차이가 의미있게 해석될 수 있다. 어디서든 눈금 간격이 같은 것으로 (등간격) 보기 때문이다. 키, 몸무게, 소득수준, 국민총생산 등을 생각해보라. 가구소득이 1 억원이면 3 억원보다 적으며, 그 차이는 11 억원과 13 억원 차이와 같다. 순서잣대와는 달리 적극 반대 (-2)와 반대 (-1) 간 차이가 찬반아님 (0)과 찬성 (1) 차이와 같다고 보는 것이다. 비율잣대 ratio scale 는 등간잣대 중에서 의미있는 영 meaningful zero point (absolute zero)이 존재하는 경우이다. 나이가 영이라면 아직 생명이 생기지 않았거나 (한국식) 태어나지 않았다 (서구식)는 의미이다. 연소득이 영이라면 소득이 전혀 없다는 뜻이다. 키와 몸무게도 현실에서 관찰하기는 불가능하다 해도 의미있는 0 이 있다.

비율잣대가 등간잣대와 다른 점은 의미있는 영이 있다는 것과 비율을 계산할 수 있다는 점이다. 섭씨 Celsius 는 표준기압 standard atmospheric pressure 에서 물이 어는 상태를 0 도로, 물이 끓는 상태를 100 도로 해서 정의한 온도잣대이다. 섭씨攝氏 영도는 자의로 정한 기준점일 뿐이다. 반면에 절대온도 Kelvin scale of absolute temperature 는 의미있는 절대영도絶對零度 (물질이 열에너지를 갖지 못하는 상태)을 가지고 있다. 따라서 섭씨나 화씨 Fahrenheit 는 등간잣대이나 비율잣대는 아니다. 섭씨 30 도가 섭씨 15 도보다 두 배 뜨겁다고 말할 수 없다 (그 차이가 100 도와 85 도 차이와 같다고 말할 수는 있지만). 반면에 비율잣대에서는 30K 가 15K 보다 두배 뜨겁다 (열에너지가 많다)고 말할 수 있다. 물론 그 차이는 100K 와 85K 차이와 같다.

등간잣대와 비율잣대로 측정된 변수를 연속변수 continuous variable 이라고 부른다. 마디변수와는 달리 어떤 두 값 사이에 무수한 다른 값이 존재할 수 있다는 의미이다. 개념상 등간잣대와 비율잣대를 엄밀하게 구분할 수는 있지만 실제 자료분석에서는 그 실익이 많지 않다. 따라서 이 글에서는 그냥 등간잣대로 부른다.

하나 둘 세는 자료는 마디 변수로 0 과 자연수로만 표시된다. 이를 "세는자료" (event) count data 라고 부른다. 월별 미군사망자수, 연도별 왜놈들 노략질수, 중정과 안기부에서 조작한 분기별 공안사건수, 일별 딴나라 차떼기수나 성폭력 건수 등을 생각해 보라. 세는자료는 흔히 등간잣대로 취급하곤 한다. 하지만, 평균과 분산이 같지 않은 overdispersion 과 같은 독특한 행태에 주의를 기울여야 한다. 예컨대, 저명한 학술잡지에 논문을 싣는 수를 생각해 보라. 대부분이 0 일 것이니 논문 하나를 싣는 것이 얼마나 큰 고비가 되는가… 일단 논문을 한번 싣게 되면 100 편을 내는 것이나 101 편을 내는 것이나 별 차이가 없는 것이다. 단순한 숫자 이면에 담긴 이런 엄청난 차이, 특히 0 과 1 사이 를 세심하게 고려해야 한다.

3.2 연구질문에 따른 측정수준

[3] "둘 중 하나"를 선택하는 상황인데, 마늘 하나에 붙은 쪽을 세는 느낌으로 두 쪽, 세 쪽, 네 쪽 등으로 부르고자 한다.
[4] 만일 자유롭게 답하게 해놓고, "경국대전經國大典에 의거 왕을 능멸한 대역죄를 물어 삼족을 멸해야 한다," "관습법에 따라 모든 문화와 제도 (상투, 가마, 호롱불, 노비, 과거제도 등)를 15 세기로 되돌려야 한다," "관습헌법이 시작된 조선왕조 이전 역사는 중국역사로 넘겨줘야 한다," "노망난 그 영감탱이들을 관습법대로 똥지게에 얹어 깊은 산 속에 버려야 한다," "멍석말이라도 해서 이참에 꼴통들 버르장머리를 고쳐놔야 한다," "앞으로 남자는 재판관이 될 수 없도록 관습법을 바꿔야 한다," "관습헌법교를 국교國敎로 삼아 제정일치祭政一致 사회로 돌아가야 한다" 등으로 단순히 범주화한다면 순서잣대가 아니라 이름잣대라 할 수 있다.

측정수준을 크게 네 가지로 구분하였지만 언제나 그 구분이 명확한 것은 아니다. 엄밀한 의미에서 연속인 continuous 변수를 상정하기는 어렵다. 자연상태에서 존재하더라도 인간이 그 현상을 객관 잣대로 측정하는 순간 엄밀한 연속성은 사라지게 된다. 어떤 대상에 대한 느낌은 연속이라 하더라도 그것이 0 부터 100 까지 정의된 잣대로 기록한다고 생각해 보라. 시간을 분, 초, 백분의 1 초로 나눈다 해도 마찬가지다. 길이나 무게나 시간을 아무리 쪼갠다 한들 인간이 측정한 것은 엄격히 말하면 분류변수나 세는 자료 (백만분의 1 초, 백만분의 2 초….) 라 할 수 있다. 연속이라는 말 자체가 계속 나누고 쪼개는 일이 부질없다는 것을 의미한다. 따라서 측정은 현상 그대로가 아닌 인간이 만든 (등간)잣대에 의해 "단순화된 인공물"이 될 수 밖에 없다.

또한 현실에서 등간잣대와 순서잣대를 명백히 구분하기 어려운 경우가 많다. 순서잣대로 보고 1 학년, 2 학년, 3 학년으로 나누는 것은 "관습법"에 사로잡힌 고정관념일 수 있다. 학교를 다닌 실제 달 수로 따지거나 출석한 날 수로 따지거나 수업을 들은 시간으로 따지면 더 정확한 구분을 할 수 있다. 설문조사에서 흔히 적극반대—반대—찬성—적극찬성을 묻고 각각 1, 2, 3, 4 로 입력하여 순서잣대로 사용하지만, 만일 입력한 숫자에 (어거지를 써서) 실제 의미있는 차이를 부여한다면 등간잣대로 사용하겠다는 뜻이다. 아예 -10 부터 10 까지 20 개 단위로, 나아가 0 부터 100 까지 느낌을 정확하게 표시하라면 순서잣대라기보다는 등간잣대라 할 수 있다. 나이를 10 대—20 대—30 대 등으로 표시하면 순서잣대나 이름잣대가 되겠지만, 연, 월, 일 등으로 표시하면 세는 자료 (등간잣대)가 된다. 이러한 경우 등간잣대와 순서잣대는 명확하게 갈리지 않는다. 찬성 87.4003, 나이 24 년 6 개월 28 일 6 시간 등과 같이 정확성이 있는 자료를 쉽게 측정할 수만 있다면 등간잣대로 분석할 수 있다. 다만 연구대상을 분석하는데 얼마나 적절한가는 항상 생각해야 한다.

어떤 개념 혹은 요인 factor 이나 "숨은 변수" latent variable 를 직접 측정하기 어려운 경우, Likert-scale 로 질문을 만들어 "드러난 변수" manifest variable 측정하곤 한다. 예컨대, 경상도의 민주주의퇴화民主主義退化를 측정하기 위해 "술먹고 욕하고 사람을 패도 경상도 출신이면 찍는다," "술먹고 여자가슴을 주물러도 우리 대구 사람이면 상관없다," "광주 빨갱이들을 무찌르고 나라를 구한 전두환 장군을 백담사에 유폐시킨 것은 한국 민주주의의 치욕이다," "박정희 각하의 적통인 박근혜 영애令愛와 그 자손이 왕위를 잇는 것은 진리요 천륜이다," "아무리 똑똑하고 유능해도 전라도와 친한 사람은 경상도 출신이라도 절대 찍지 않는다," "김대중이 거짓말쟁이고 빨갱이인 것은 성경에도 나와있다," "대구 지하철 사고는 전라도 정권의 음모 陰謀다" 등을 물을 수 있다. 그리고 개개 질문의 평균을 구하거나 요인분석 factor analysis 을 통해 요인점수 factor score 를 구하여 숨은 변수의 값으로 사용한다. 개별 질문은 순서잣대이지만 개별 질문을 종합한 점수는 등간잣대처럼 되는 것이다. 이러한 요인분석은 측정 수준 간 울타리가 높고 단단한 것이 아님을 보여준다. 이러한 방법을 사용하는 것이 적절한지 아닌지는 또 다른 문제이다. 요컨대, 측정수준은 언제나 명확한 것도 아니며, 미리 정해진 것도 아니며, 절대 불변하는 것도 아니라는 점에 유의해야 한다 (Long 1997: 2). 먼저 관련된 이론理論과 연구질문硏究質問이 어떠한 것인가를 따져야 한다.

The levels of measurement cannot be isolated from the theoretical and substantive context in which the variable is to be used (Charter 1971: 12).

연구질문 research question 이 학년 순서와 무관하다면 1 학년—2 학년—3 학년은 그저 이름잣대일 뿐이다. 백만원 단위로 측정된 개인소득을 상—중—하로 구분하였다면 더이상 등간잣대가 아니다. 순서잣대가 되거나, 순서마저 고려하지 않는다면 이름잣대가 될 뿐이다. 성별은 흔히 이름잣대로 두쪽변수에 사용되지만, 소위 "경국대전 풍風"으로 남녀간 순서를 엄밀히 따지거나 실제 의미있는 차이를 가정한다면 순서잣대나 심지어는 등간잣대도 될 수 있다. 나이도 보통은 등간잣대로 간주하지만, 때에 따라서는 20 대—30 대—40 대처럼 순서잣대나 이름잣대로 사용할 수도 있다. 문제는 얼마나 연구질문에 합당한 것이냐에 달려있다. 개념, 변수, 잣대, 측정단위는 연구질문과 밀접한 관계가 있다.

3.3 측정수준과 자료분석 방법

측정 수준은 자료를 분석하는 방법를 정한다. 보통 자료분석은 등간잣대나 비율잣대를 대상으로 한다. 이름잣대나 순서잣대에 비해 의미있는 정보를 많이 가지고 있기 때문이다. 회귀분석나 모평균비교(t-test/ANOVA) 등은 등간잣대나 비율잣대로 측정된 변수 (종속변수)를 설명하는 방법이다.

표 1. 측정 수준과 자료분석 방법

	순서	차이	비율	유형	주요 자료분석 방법
이름잣대 nominal	X	X	X	마디	Chi-squared test, Lambda, Kappa
순서잣대 ordinal	O	X	X	마디	Chi-squared test, Wilcoxon Rank Sum
등간잣대 interval	O	O	X	연속	Regression, T-Test, ANOVA
비율잣대 ratio	O	O	O	연속	Regression, T-Test, ANOVA
세는자료 count data	O	O	O	마디	Poisson regression

반면에 이름잣대나 순서잣대에는 평균이나 분산과 같은 일반 통계량을 사용할 없다. 1, 2, 3 등으로 입력한 이상 그런 통계량 statistics 을 계산하는 일은 쉬우나 의미있는 해석을 할 수는 없기 때문이다. 그런 식으로 입력한 사과, 배, 감, 밤, 대추의 평균이나 분산을 가지고 어쩌자는 것인가? 그것이 parametric 인지 (모집단의 평균, 분산과 같은 특성이 알려진 경우) nonparametric 인지 따져서 무엇을 하자는 것인가? 이러한 마디변수들은 Pearson Chi-squared test 와 같이 관측값이 나타난 횟수 frequency 나 순서가 분포된 모양을 따져볼 수 있을 뿐이다. 세는 자료는 변수값 간 순서와 차이가 의미를 갖지만 마디변수로 측정된다.

4. 어느 잣대로 어느 정도까지 측정할 것인가?

어느 잣대로 어느 정도까지 측정하는 것이 연구질문에 합당하고 좀더 정확하고 신뢰할 수 있는 자료를 비용을 덜 들이고 얻을 수 있는가? 측정수준을 선택하기 위해서는 연구질문 뿐만 아니라 정확성, 신뢰성, 효율성 등을 동시에 고려해야 한다. 물론 이들 간 엇박자 trade-off 가 있기 때문에 가장 좋은 타협점이 어디인가를 찾아야 한다. 따라서 중요한 것은 잣대 자체가 아니라, 그러한 측정수준이 특정한 현실상황에 얼마나 의미있는가, 얼마나 효율성있게 측정할 수는 있는가, 얼마나 정확하게 현상을 측정할 수 있는가 하는 것이다.

4.1 정확성, 신뢰성, 효율성

나이를 측정한다고 생각해 보자. 우선 2005 년 1 월 1 일 현재 서양식 나이라고 정의해 보자. 가장 정확한 나이라면 어머니 몸에서 분리된 정확한 순간에서부터 계산한 시분초 (예컨대 99 만 시간 55 분 11 초) 시간일 것이다. 최초로 세포가 만들어진 순간부터 따진다 해도 마찬가지다. 과연 이런 자료가 연구질문에 답하기 위해 필요한가? 과연 이런 자료를 수집하는 것이 가능한가? 가능하다 해도 얼마나 정확하고 신뢰할 만한가? 또 얼마나 많은 시간과 돈을 투자해야 하는가?

사람들에게 정확한 시분초 단위로 나이를 묻는다면 대부분이 "잘 모른다"로 답할 것이다. 즉, 정확하게 답하기 불가능하거나 "확실히 부정확한" 답변을 적어낼 것이다. 사실 나이는 몇 시간, 몇 일, 혹은 몇 달 차이를 따져봤자 대부분 별 소득이 없다. 물론 소아 연구에서는 몇 개월, 몇 일이 더 적절할 수 있다. 그래서 실제 연구에서는 흔히 나이를 연 수로 따지거나 그냥 생년월일을 묻는다 (나중에 계산하면 된다). 정확한 개월수를 묻는 것보다 많은 사람들이 답할 것이다. 만으로 나이를 셈하는 것보다는 그냥 생일을 답하는 것을 더 쉽게 생각할 것이기 때문이다. 경우에 따라서는 10 년 단위로 끊어서 (10 대—20 대—30 대…) 물을 수도 있다. 개인소득도 정확한 금액이 아니라 천만원 미만, 1—2 천만원 등으로 범위를 나눠 물어보는 것이 더 효과가 좋을 수 있다. 세금 문제도 있고 해서 정확한 액수를 밝히기를 꺼리는 경향이 있기 때문이다. 순서잣대로 보고 1, 2, 3 등을 할당해서 사용할지, 아니면 중간값 (천만원, 천오백만원, …)을 취하여 등간잣대로 처리할지는 연구자가 판단할 일이다.

서울역에서 광화문까지 거리라면 몇 리 혹은 몇 Km 면 족하지 몇 센치미터 혹은 몇백만분의 1 cm 까지 따져서 무엇에 쓰겠는가. 그 정도 정밀함을 따지기 위해서는 어디를 기준으로 재야 하는가, 누가 어떻게 잴 것인가를 두고 생산성없는 논쟁만 벌여야 할 것이다. 그냥 대충 걸어서 10 분 거리라고 해도 될 일을 가지고… 또한 찬반贊反을 가리는 문제를 두고 0 부터 100 까지 오호惡好를 정확한 숫치로 밝히라고 하면

과연 얼마나 많은 사람들이 심각하게 고민을 하여 보고할 것인가? 아마도 답하기를 꺼리거나, 아무렇게나 적어내거나, “그까이꺼 대--충” 거의 비슷한 숫자 (25, 50, 90 등)를 써낼 것이다 . 어쩌면 찬성 아니면 반대로 답하라는 것보다 정확성이 떨어질는지도 모른다. 정확성을 높이기 위해서는 대체로 비용과 시간이 많이 든다. 하지만 정확성을 일정 수준 이상으로 높이는 일이 현실에서 불가능한 경우가 있다는 것도 알아야 한다.

4.2 사생활과 사회윤리 문제

사생활이나 사회윤리와 관련된 질문은 여러가지 의미에서 주의가 필요하다. 정확한 개인소득을 요구한다면 실제 소득보다 적게 쓰거나 답하기를 거부할 것이다. 몇 원이 아니라 백만원 단위로 써달라 해도 마찬가지이다. 길을 가로막고 나이찬 여자에게 숫처녀인지 아닌지를 묻는다고 해보자. 자료를 얻기는 커녕 뺨맞기 십상이다. 너도 나도 벗지못해 안달이 난 것같은 요즘 세상에 가슴—허리—엉덩이 크기를 묻는 것은 좀 나아 보인다. 하지만 직접 자를 들고 재려 한다거나 느끼한 눈빛으로 아래 위를 훑어보는 날에는 당장 정강이를 채이기 십상이다. 설령 설문지를 돌려 어거지로 자료를 얻었다 한들 얼마나 쓸모가 있을까? (가운데 숫자만 축소하고 양쪽 수치를 부풀릴 것으로 생각한다면 지나친 추측일까?)

뜬금없이 “학교”에 몇번 다녀왔느냐 (세는 자료)를 대놓고 묻는다고 생각해보자. 많은 사람들이 불쾌감을 느낄 것이고, 뒤가 구린 사람이라면 아예 회피할 것이고, 좀 성깔이 있는 사람에게 걸리면 멱살잡이라도 당할 것이다. 전과경력이 있는지 없는지 (두쪽변수)만을 말해달라면 좀 나아보이지만, 좋은 자료를 얻기 힘들다는 면에서 별반 차이가 없다. 그렇다면 공무원들에게 뇌물을 받은 적이 있는지, 얼마나 짭짤하게 받았는지를 묻는 것은 어떻겠는가? 아무리 정중하게 물어도, 억원 단위로 대충 답해달라고 부탁을 해도 마찬가지다.

어느 잣대가 제일 좋은 것인가? 이것은 좋은 질문이 아니다. 나이면 등간잣대고 학년이면 순서잣대라고 못박을 필요가 없다. 상황에 따라서 나이를 순서잣대나 이름잣대로 사용할 수도, 학년을 이름잣대로 사용할 수도 있다. 또한 연구 엄밀성에 사로잡혀 무조건 정확성을 고집하는 것도 능사가 아니다. 시간과 돈이 많다고 해도 등간잣대로 정확하게 측정하는 것이 항상 가능한 것도, 바람직한 것도 아니다. 이 문제는 통계 전문가나 지도교수가 답할 일이 아니다. 어느 잣대로 보는 것이, 또 어느 정도 수준으로 측정하는 것이, 어떤 방법으로 측정하는 것이 연구질문에 답하는데 적절한가를 스스로에게 물어야 한다. 이 문제에 대하여 자신이 세상에서 제일 잘 알고, 또 알아야 하는 사람이기 때문이다. 그렇기 때문에 연구자는 항상 부질없는 “하나”에 대한 집착을 버리고 열린 마음으로 균형감각을 유지해야 한다.

5. 변수와 Random 변수

변수 variable 는 개념을 측정가능하게 정의한 것이다. 구체화된 측정 방법과 측정 단위가 명시되어 있어야 한다. 수학이나 계량 방법론으로 말한다면 변수는 관측치에 따라 여러 값을 가질 수 있는특성을 갖는다. 관측값이 일정하지 않고 관측치에 따라 변한다. 수강생의 키를 측정했다면 사람마다 다른 측정치를 기대할 수 있다. 이와 대비되는 것은 상수 constant 인데, 파이 pi 3.141592 와 같이 항상 같은 값을 갖는다. 어느 특정 시점에서 모집단 population 의 평균은 변수가 될 수 없지만 표본 sample 의 평균은 변수가 된다. 모집단의 평균은 하나인데 (그래서 변하지 않는데), 표본의 평균은 표본에 따라 다른 값을 가질 수 있기 때문이다. 같은 모집단을 측정한다 해도, 오늘 측정한 표본이 다르고, 내일 측정한 것이 다르다. 내가 측정한 표본의 평균과 다른 사람이 측정한 표본의 평균이 다르기 마련이다. 그래서 상수는 확정성 deterministic 을 가진 반면 변수는 확률성 stochastic 을 가졌다.

5.1 Random 변수

변수는 그 값이 변하기 마련이데, 그 변화하는 양태는 여러가지가 있다. 그 중에 방법론에서 의미있는 양태는 *제멋대로 randomly* 변하는 것이다. 여기서 제멋대로라는 뜻은 *아무렇게나 arbitrarily* 변한다는 것이 아니라 한 관측치가 다른 관측치의 영향을 받지 않고 나름의 논리 data generation process (DGP)대로

행동한다는 것을 의미한다.[5] 어떤 관측값이 다른 관측값에 영향을 전혀 주지 않는 독립성 independence 을 가지고 있음을 말한다. 따라서 미리 어떤 값이 나올지 예측할 수 없다. 물론 나름대로 논리를 가진 "제멋" (정규분포든 F 분포든 간에)이 있기 때문에 그 전체 특성 (예컨대, 평균과 분산)을 알 수는 있다. 이러한 변수를 random 변수라고 부른다. 계량 자료 분석은 이런 random 변수를 상정하고 각종 이론을 전개한다. Random 변수가 아니라면 널리 사용되는 자료분석 이론을 적용할 수가 없다. 따라서 앞으로 변수라 함은 특별히 명시하지 않으면 random 변수를 말한다.

Random 변수는 우선 종속변수 dependent variable 와 독립변수 independent variable 로 구분한다. 전자는 관심을 가진 현상을 측정했다는 의미에서 반응변수 response variable 라 부르기도 하고, 방정식의 왼쪽에 나타난다 하여 왼쪽변수 left-hand side (LHS) variable 라고도 부른다. 회귀방정식 regression equation Y=a +bX1+cX2 를 생각해 보라. 독립변수 X1 과 X2 는 종속변수 Y 를 설명한다는 의미로 설명변수 explanatory variable 혹은 오른쪽변수 right-hand side (RHS) variable 로 부른다. 이 외에 종속변수에는 영향을 미치기는 하나 (그래서 모델에 포함되어야 하나) 주된 관심사가 아닌 독립변수를 특별히 통제변수 control variable 혹은 covariate 라고 부른다.[6]

그러면 어느 변수가 중요한가? 우리가 설명하고자 하는 것이 종속변수인 이상 종속변수가 가장 중요하다. 따라서 종속변수의 행태를 분석하고 이에 근거하여 모델 model 을 만드는 일이 가장 중요하다. 그래서 어떤 모델에서 변수라고 하면 흔히 종속변수를 말한다. 독립변수는 종속변수를 설명하는 도구일 뿐이다. 독립변수를 중요하다고 보는 입장은 모델을 낚시질하거나 사냥하는 data fishing 경우이다. 논문을 출판하는 데 급급하여 아무 생각없이 소프트웨어를 닥달하여 통계 유의성 statistical significance 이 있는 모델을 찾아 헤매는 일로 이론 가치가 전혀 없다. 누군가가 같은 방법으로 자료를 수집한다면 쉽게 다른 결론을 내릴 수 있고, 그 모델이 예측한 것은 현실과 같이 움직일 가능성이 별로 없기 때문이다.

5.2 분류형 종속변수와 제한된 종속변수

어떠한 측정수준으로 종속변수를 측정했는가, 어떤 과정으로 측정했는가, 측정된 결과가 어떠한지는 자료분석에 꼭 필요한 정보다. 이러한 정보를 토대로 어떻게 자료를 분석할지, 어떤 모델을 사용할지를 정하기 때문이다. 분석과정 (특히 해석)에서 어떤 점에 주의를 기울여야 할지를 알게 된다.

보통은 등간잣대로 종속변수를 측정한다. T-test, ANOVA, 여러가지 선형회귀모형을 자유롭게 사용할 수 있다. 이름잣대나 순서잣대로 측정한 경우에는 선형회귀분석보다는 logit 이나 probit 모델을 사용한다. "두쪽변수"이면 binary logit/probit regression, 순서변수면 ordinal (ordered) regression, 이름변수면 multinomial regression, conditional logit, nested logit 모델을 상황에 따라 사용한다. 종속변수가 세는 변수이면 Poisson regression 이나 Negative binomial regression 모델을 사용한다. 이러한 것들을 흔히 분류형 종속변수 모델 categorical dependent variables model 이라고 부른다.

하지만 종속변수를 원하는 대로 측정하기 불가능한 경우가 종종 있다. 측정과정에서 오차나 실수가 없다고 해도 여러가지 문제가 생길 수 있다. 먼저 측정값을 얻을 수 없는 경우 missing value 를 생각해 보자. 설문조사에서 민감한 질문에 대하여 응답자가 대답을 거부하는 경우가 많다. 0 이라고 답한 것과 답하지 않은 것은 분명한 차이가 있다. 무응답이 지나치게 많으면 제대로 자료를 분석하기 어렵다. 이런 경우 "채워넣기" imputation 를 하여 분석을 진행할 수 있다.

어떤 종속변수는 측정과정에서 일정한 값으로 잘리거나 아예 측정에서 배제될 수도 있다. 흔히 제한된 종속변수 limited dependent variables 라고 부르곤 한다. 먼저 "잘린" censored 것은 종속변수가 어떤 범위에 걸쳐있는 실제 관측값이 특정값으로만 관찰되는 경우이다. 어떤 값 밑으로, 위로, 혹은 아래 위 모두 잘릴 수 있다. 측정기계가 미세한 공해수준을 제대로 측정하지 못하고 일정한 값 (최소값)으로만 보고한다고 생각해

[5] 예컨대, 자료발생기 DGP 의 논리가 정규확률분포 normal probability distribution 라면 random 변수는 평균값에 가까운 값을 가질 확률이 크고 평균에서 멀어진 값 (음수이거나 양수이거나)을 가질 확률이 적어진다. Poisson distribution 이라면 음수는 발생하지 않으며 평균값보다 큰 값이 발생할 확률이 정규확률분포보다 천천히 줄어든다 (꼬리가 길다).
[6] ANCOVA (analysis of covariance)에서는 마디변수 외에 통제 목적으로 연속변수를 독립변수로 사용한다.

보라. 공해정도가 .001 이나 .01 이나 모두 1 로 측정되는 경우이다. 국민들의 충격과 무장봉기를 두려워하여 900 억원 이상되는 차떼기는 모두 900 억으로 발표했다고 생각해 보라. 이를테면 개인소득이 10 억 이상인 사람은 백억이든 천억이든 모두 10 억으로 기록하고, 천만원 미만인 경우 (쑥스러우니까) 모두 천만원으로 적었다고 해보라.

어떤 관측치는 종속변수와 독립변수 모두 측정되지 못할 수도 있다. 이런 경우는 관측에서 배제되었다고 truncated 말한다. 표본이 모집단이 아닌 모집단의 일부에서 뽑힌 경우 이러한 문제가 생긴다. 반면 잘린 censored 경우 독립변수는 관찰할 수 있다. 예컨대, 개인소득이 일정 수준 threshold point 미만인 저소득층에 대해 연구한다면 일정수준 이상 되는 개인에 대한 정보는 수집되지 않는다. 차떼기 수준 (예컨대, 9 백억원) 이상이 되는 사례만 관찰하여 불법탈법 돈선거 (모집단)를 연구한다고 생각해 보라. 물론 표본은 랜덤샘플링 random sampling 으로 얻는다. 하지만 배제된 자료는 모집단 (국민전체 혹은 불법탈법 선거)을 대표하는 표본이 될 수 없다. 그렇다면 현재 관측된 표본 (저소득층 자료)을 이용해서 어떻게 전체 모집단 (국민전체)의 특성을 추정할 것인가?

표본을 의도를 가지고 선택한 경우 sample selection or incidental truncation 도 있다. 예컨대, 저소득층의 개인소득을 연구한다고 했을 때 국민학교졸업 미만을 대상으로 표본을 뽑았다고 생각해 보라. 앞서 얘기한 전두환의 표본도 마찬가지다.[7] 표본이 랜덤하게 뽑히지 않은 것이다. 학력이 높을수록 개인소득이 높다면 어떻게 선택된 자료를 가지고 모집단의 개인소득을 추정할 것인가? 응답자나 참여자 스스로 통제집단 control group 에 들어갈 지 실험집단 treatment group 에 들어갈 지를 정하는 "자기선택" self-selection 도 마찬가지이다. 예컨대, 실직자를 위한 직업훈련교육에 참여할 지 말지를 스스로 정한다고 생각해 보라. 만일 참여자가 참여하지 않은 사람과 구별되는 특성 (예컨대, 강한 의지)이 있고 이것이 설명하고자 하는 변수 (구직)에 영향을 미친다고 생각해 보라.

마지막으로 수명자료 duration data or survival time data 는 어떤 상태가 지속되는 시간을 관찰한 것이다. 예컨대, MP3 연주기가 얼마나 오래 고장나지 않고 동작하는가, 환자가 암에 걸린 뒤 (혹은 어떤 수술을 받은 후) 얼마나 살 수 있는가? 특정 시점에서 관찰하면 이미 수명을 다한 것도 있고, 아직 동작을 멈추지 않는 것도 있을 것이다. 전자는 정확한 수명을 얻을 수 있지만, 후자는 자료가 잘려서 censored 정확한 수명을 관측할 수 없다 (지금까지 얼마동안 동작했나를 알 수 있을 뿐이다).

이와같은 제한된 종속변수를 일반 방법 (예컨대, 일반선형회귀)으로 분석한다면 그 결과는 신뢰할 수 없으며 큰 의미를 주지 못한다. 따라서 잘린 자료를 위한 Tobit 모델, 선택된 자료를 위한 Heckman 모델과 Propensity score matching method, 수명자료를 위한 duration model 등이 주목받고 있다.

5.3 한변수법과 여러변수법

모델에 사용된 종속변수 수에 따라 한변수법 univariate 과 여러변수법 multivariate 으로 나눈다. 독립변수가 몇개가 되었는지는 별로 고려 대상이 아니다. 그냥 독립변수 묶음이 있다고 생각할 뿐이다. 흔히 독립변수가 하나인 회귀모형을 단순회귀 simple regression, 여러개인 것을 다중회귀 multiple regression 라고 구분한다. 이것은 설명을 위한 교육목적 상 구분이지 실제 이론으로 치면 전혀 의미가 없다. 혹자는 다중회귀와 여러변수회귀 multivariate regression 를 같은 의미로 사용하기도 하는데, 이는 개념을 혼동한 것일 뿐 전혀 근거가 없다. 여러변수회귀는 종속변수가 두 개 이상인 흔치 않은 회귀모형으로 독립변수 수와는 아무런 관계가 없다. Bivariate probit model 과 같이 종속변수가 두 개인 경우에는 특별히 두변수법 bivariate method 이라고 부른다.

5.4 변수형 variable type

[7] 모집단과 표본은 일관성이 있어야 한다. 전두환의 표본이 정당화되려면 모집단을 일반 국민이 아닌 "전두환과 그 똘마니"로 바꾸면 될 것이다. 독재자들은 걸핏하면 나라니 국민이니를 들먹이면서 (모집단) 힘으로 다른 사람들을 누르고 자기 생각을 (자의로 선택한 엉터리 표본) 밀어붙이곤 한다. 무식하거나 정직하지 않은 (그래서 음흉하고 사악한) 것이다. 진실을 알고 싶지도 않고, 또 진실이 알려지는 것을 두려워하기 때문에 어거지를 쓰는 것이다. 물론 일반 연구에서도 모집단을 대표하지 못하는 표본으로 부당하게 모집단에 관한 결론을 내리는 우를 종종 범하기도 한다.

변수형을 고려하는 일은 자료를 처리하는데 매우 중요하다. 아무리 CPU 성능이 탁월하고 메모리가 싸진다 해도 효율성과 효과성을 버릴 수 없기 때문이다. 자료분석 프로그램은 대부분 정수형, 실수형, 문자형, 날짜형, 논리형 변수를 처리할 수 있다. 보통 숫자형 numeric 인 정수형 integer 과 실수형 float or real 이 가장 많이 사용되고 있다.[8] 특별한 경우가 아니면 숫자형 변수를 사용하고, 그 중에서도 정수형이나 바이트형 byte 을 사용하는 것이 좋다. 숫자형 (정수형)이 자료를 처리하기 쉽고 메모리도 적게 차지하기 때문이다. 이름과 같이 꼭 문자로 써야 할 경우에는 물론 문자형 character or string 변수를 사용한다. 하지만 자료분석을 위해서 반드시 변수가 숫자형이어야 하는 소프트웨어도 있음을 기억해 두라. 숫자형과 문자형 변수는 길이를 최소화하여 메모리와 연산력 computing resources 을 아낄 필요가 있다. 예컨대, 두쪽변수나 4-point Likert 잣대라면 2 바이트 정수형이나 바이트형(1 바이트)이면 충분하다.

날짜형 date 변수는 연산하고 출력하는데 많은 부담을 주기 때문에 반드시 필요한 경우가 아니라면 문자형으로 처리하고 날짜형식으로 입력한다. 예컨대, yyyy-mm-dd 나 mm/dd/yyyy 와 같이 입력하는 것이 좋다. 논리형 boolean ("불리언")변수 역시 꼭 필요한 경우가 아니라면 정수형을 취하고, 1 (TRUE) 과 0 (FALSE) 로 입력하는 것이 좋다. 필요한 경우 변수 형식을 변환하는 함수를 이용하여 바로 날짜형이나 논리형으로 바꿀 수 있기 때문이다.

표 2. 자료처리 소프트웨어 비교

	SAS/BASE 9.1	Stata 11.0 SE	LIMDEP 9.0	R 2.9	SPSS 17.0
OS	UNIX/Windows	UNIX/Mac/Win	Windows	UNIX/Mac/Win	Mac/Win
1st Interface	Batch	Interactive,batch	Point-and-Click	Interactive,batch	Point-and-Click
Batch File	.sas	.do	.lim		.sps
Data Editor	Point-Click	.edit	Point-Click	> edit()	Data view
Casesenstive	No	Yes	No	Yes	No
Size	1.5GB*	245MB	25MB	75MB	650MB
# Variables	2 billions[9]	32,767	900		32,767
# Observations	Memory	Memory	Limited		2 billions
Var Name	32 bytes	32 bytes	8 bytes		32 bytes
String	32,767 bytes	244 bytes	N/A		255bytes
Label	256 bytes	80 bytes	N/A		255 bytes

* When all modules such as SAS/STAT, SAS/IML, SAS/ETS, SAS/SQL, and SAS/OR are installed.

6. 변수이름을 어떻게 할 것인가?

이름을 어떻게 정할 것인가는 컴퓨터에서 중요하다. 파일, 변수, 배열 array, 라이브러리 library, 매크로 macro, 함수 function, 변수뜻말 variable label 이름을 적절히 정해야 하기 때문이다.[10] 특히 변수이름은 자료분석과 직접 관련이 되기 때문에 시간을 가지고 잘 생각해야 한다. 급한 마음에 아무렇게나 변수 이름을 정하다 보면 자료분석이 고통스러지기 십상이다. 또한 한글로 변수이름을 쓰는 것은 아직까지 여러가지 문제가 있기 때문에 영문으로 사용할 것을 권한다. 다음과 같은 중요한 규칙을 꼭 숙지하라.

가. 문자(A-Z, a-z)와 숫자 (0-9) 그리고 underscore _만으로 이름을 만든다.[11]
나. 첫글자는 반드시 문자로 한다.[12]

[8] 소프트웨어에 따라서 긴정수형 long integer 과 심각한 정확성을 요하는 경우에 double-precision 을 사용할 수도 있다. SAS 에서 일반 정수형은 2 바이트 byte 이고, double 형은 기종에 따라 차이가 있지만 최고 8 바이트나 된다.
[9] SAS 엔진 8.xx 까지는 32,767 (2^{15})개까지만 지원한다.
[10] 소프트웨어마다 용어도 다르고 허용하는 길이도 다르다 (표 2). 예컨대 SAS 에서 변수, 배열, 매크로, 매크로 변수, formats, informats 는 32 자까지 허용하지만 함수나 CALL routine 이름은 16 자까지만 사용한다. 반면 Librefs, Filerefs, Engines, password 이름은 8 자까지만 허용한다. Stata 에서 변수뜻말은 80 자, 변수값뜻말 value label 은 32 자까지 허용한다.
[11] 특수문자, 예컨대, -, 공란 space, ~, !, @, #, $, %, ^, &, *, (,), {, }, [,], <, >, ?, / 등은 사용하지 않는다.
[12] Underscore 는 _N_과 같은 system 변수 (예약어) 에 사용되곤 하기 때문에 첫글자로 사용하지 않는 것이 좋다.

다. 가능하면 8 자 이내로 하되 짧을수록 좋다.[13]
라. 명령어 등에 사용되는 예약어 reserved word 를 사용하지 않는다.
마. 공란 대신 underscore 를 사용하라.
바. 변수내용을 연상시킬 수 있는 이름이 좋다.
사. 일관성과 체계성을 갖는 것이 좋다.[14]
아. 가능하면 소문자를 사용하고 대소문자를 구별하여 사용한다.[15]
자. 두쪽변수인 경우 변수값 하나를 변수이름으로 사용하길 권한다.

가장 흔한 실수는 변수이름에 공백을 사용하거나 (US citizen) 숫자로 시작하는 경우 (2002_sale)이다. 특히 Excel 에서 자료를 입력할 때 열 column 이름을 그런 식으로 적어놓으면 자료처리 소프트웨어에서 그 파일을 읽는 일이 힘들어진다.

또한 긴 설문문항 자체 (How would you feel if …)를 변수이름으로 하려는 용맹스런 사람도 가끔씩 있다. 심지어 중복된 변수이름을 사용하는 사람도 있다. 같은 반에 삼식이가 여러 명 있을 수는 있어도 한 자료판에서 변수이름 삼식이는 오직 하나만 허용된다. 이런 용례를 허용하는 소프트웨어가 있다 해도 일반인의 상식에 비추어 봐도 어처구니 없는 짓이다. 파일이나 디렉토리 이름에 공백이 들어가면 특히 UNIX 에서 눈물나는 경우가 생긴다는 점에 유의하라.[16]

표 3. 좋은 변수이름과 나쁜 변수이름

좋은 예	나쁜 예	설명
gnp2002	gnp-2002; gnp#2002	특수문자를 쓰지말라
real_int	real interest rate	공백대신 _로 연결하라
score1; gnp2003	1st_score; 2003gnp	숫자로 시작하지 말라
reg_out; glm1	REG; glm; ttest	예약어를 쓰지 말라
invest; interest	xxx; yyy; zmdje;	의미있는 이름을 쓰라
male; black	gender; race	관측값 하나를 사용하라
score1; score2; score3	math, physics, math_1, math02	일관성과 체계성을 유지하라
Citizen	Are_you_a_US_citizen?	가능한 짧게 하라
income; intUS03	INCOME; Int_Us2003;	가능한 소문자로 써라

설문조사를 한다면 다음과 같이 일련번호 대신에 설문지에 아예 변수이름을 적어두는 것도 좋은 방법이 될 수 있다. 또한 자료입력을 쉽게 하기 위해 설문문항 배열을 조정하거나 문항번호 (무응답 포함)를 붙여넣을 수도 있다.

[male] *How do you identify your gender?* (Male/Female)

7. 자료구조: 관측치와 변수

자료를 입력하기 전에 소프트웨어에서 보는 자료구조를 생각해 보자.

[13] 표 2 에서 보듯 자료처리 소프트웨어는 보통 32 자까지 허용하지만 변수이름을 가능한 짧게 만들고 변수뜻말이나 값뜻말을 써서 설명을 붙이는 것이 정석이다. 하지만 너무 뜻말에 의존하면 출력물이 복잡해지기 때문에 뜻말도 간단하게 최소한으로 사용하는 것이 좋다.
[14] 그래야 배열이나 Wild card 를 사용하기 쉽다. 예컨대, score1-score10, score??, vote* 등이다.
[15] Stata, R, GAUSS 와 같이 대소문자를 구별하는 case-sensitive 소프트웨어가 있으며, 대소문자를 구별하면 batch file 을 읽는데도 도움이 되기 때문이다.
[16] 이런 의미에서 Microsoft 는 사람들에게 기본을 무시하고 모든 것을 "쉽게 쉽게" 하려는 못된 습성을 길러주고 있다. Login 을 왜 해야 하면서 투덜대거나, Stata 사용자라면서 파일을 읽어오는 명령어를 모른다는 황당한 사람들을 양산하고 있다. 기본을 알고 point-and-click 을 사용하는 것은 可하나 그 반대는 不可하다.

7.1 관측치와 변수

OpenOffice 나 Excel 이나 Quattro Pro 에서 작업판 Worksheet 를 생각해 보라. 작업판에서 행 row 은 관측치 observation, case, or subject (피실험자) 를 의미한다. 관측치는 자료를 모아서 소프트웨어에서 분석하는 기본 관측 단위 unit of observation 이다. 분석단위를 어떻게 할 것인가와 자료판의 관측단위를 어떻게 정할 것인가는 매우 중요한 문제이다.

데이타베이스 database 에서 관측치 observation 는 자료개체 record 혹은 entity 라고 부른다.[17] 반면 열 column 은 관측치의 성질을 담고 있는 변수 variable 를 의미한다. 그래서 데이타베이스에서 변수를 attribute 나 field 라고 부른다. 예컨대, 이름, 학번, 성별, 키, 몸무게 등을 말한다. 그림 2 에서 왼쪽은 관측치와 변수가 어떻게 자료판에 배열되어 있는지를 보여준다. 오른쪽은 실제 자료처리 소프트웨어 Stata 에서 관측치와 변수가 어떻게 구성되는지를 보여준다. 오른쪽 그림에서 점 period 으로 입력한 관측값 (변수값)은 응답없음 (무응답) missing 을 의미한다.

그림 2. 자료판을 구성하는 관측치와 변수

$$Table\begin{cases} & \mathrm{var}_1 & \mathrm{var}_2 & \cdots & \mathrm{var}_k \\ obs_1\{ & \cdot & \cdot & \cdots & \cdot \\ obs_2\{ & \cdot & \cdot & \cdots & \cdot \\ \cdots\cdots\{ & \cdots & \cdots & \cdots & \cdots \\ obs_n\{ & \cdot & \cdot & \cdot & \cdot \end{cases}$$

```
+-----------------------------------+
|   id   age0   age   male   interest |
|-----------------------------------|
| 1025     29     1      0       1.00 |
| 1026     40     3      1       3.50 |
| 1027     27     1      0          . |
| 1028     34     2      .       5.00 |
| 1029     35     2      1       4.00 |
… …        …     …      …          …    …
| 1226     50     4      1       3.25 |
```

데이타베이스에서는 이러한 작업판 하나를 자료판 table 으로 부르고, 자료판이 여러 겹 있는 것을 자료집 database 라고 부른다. 예컨대, 학생 개개인의 개별정보 (학번, 이름, 키 등)는 자료항목 attribute 이고, 개인의 개별 정보를 모은 것은 자료개체 record 이고, 학년 전체 학생의 정보는 자료판 table 이고, 여러 학년의 자료판이 모인 것이 자료집이다. 자료항목이 0 차원 (점) 세계라면, 자료개체는 1 차원 (선)이고, 자료판은 2 차원 (평면)이고, 자료집은 3 차원 공간인 셈이다.

Bit → byte → word → field (attribute) → record (entity, observation, or case) → table → database[18]

스프레드시트의 작업판 사이에 정보를 주고 받을 수 있는 것처럼, 자료집에서 자료판 table 을 연결해서 정보를 참조하는 일이 흔하다. 자료판을 연결 joining 하기 위해서는 각 관측치의 고유값을 갖는 식별변수 identification field 가 필요하다. 그렇지 않으면 어느 관측치와 어느 관측치를 연결해야 하는지를 모르기 때문이다. 고유변수 unique variable 는 중복되지만 않는다면 일련번호일 수도 있고, 학번과 같은 고유번호일 수도 있다. 물론 여러 변수 (예컨대, 학년과 반번호)를 사용하여 고유한 관측치를 정할 수도 있다. 그림 2 의 오른쪽에서 id 라는 변수를 보라. 식별변수는 자료를 추적하거나 다른 자료판과 연결하기 위해서라도 반드시 필요하다.

7.2 자료구조와 자료입력

이제 자료구조를 생각해 보자. 연구자가 원자료를 입력하는 자료구조와 소프트웨어에서 요구하는 자료판 data set 의 자료구조가 있다. 후자는 실제 소프트웨어에서 자료를 분석하기 위해 필요한 자료구조이다. 하지만 전자가 후자와 반드시 일치해야 할 필요가 없다. 얼마나 효율성 있게, 실수를 줄이면서 입력할 수 있는가와 얼마나 쉽게 입력된 자료를 해당 소프트웨어에서 원하는 자료구조로 읽어들일 수 있느냐가

[17] 관측치는 독립성을 가져야 하며 특별한 경우가 아니면 같은 비중으로 처리되어야 한다. 한 관측치가 다른 관측치의 영향을 받는다거나, 근거없이 한 관측치가 다른 관측치보다 더 많은 비중을 가진다면 자료분석에서 중대한 문제가 된다.
[18] Bit 는 1 과 0 을 저장하는 기본단위이며, byte 는 보통 8bit 가 모여 한 문자를 표시하는 저장단위이며, word 는 8bit, 16bit, 32bit, 64bit 머신 등과 같이 컴퓨터가 한꺼번에 처리할 수 있는 정보단위이다. Word 는 컴퓨터 기종에 따라 다르다.

관건이다. 입력해야 할 자료의 양, 형태, 소프트웨어의 기능, 연구자의 소프트웨어 사용 능력 등을 고려해서 결정해야 한다.

Randomized block design 실험을 했다고 해보자. 블락 block 이 셋이고 처리 treatment 가 넷이라고 해보자. 어떻게 원자료를 입력할 것인가? 자료처리 소프트웨어는 대개 그림 3 의 오른쪽과 같은 자료구조를 원한다. 변수 block (1-3)와 treat (1-4)는 일정한 형태가 있음을 주목하라. 또한 이러한 자료구조가 개체와 시간으로 배열된 시공간자료 panel data 의 자료구조와 같음을 상기하라.[19]

가장 단순한 방법은 오른쪽처럼 입력하는 것이다. 관측치가 12 개 (= 3 X 4) 정도일 때는 해볼만한 일이지만 시공간자료時空間資料처럼 개체 (예컨대, 회사나 나라)가 한 100 개, 관측시점 (예컨대, 분기나 연도)이 한 50 개쯤 되면 삽들고 땅파는 노동이 된다. 그림 3 에서 왼쪽 첫번째 예를 보라. 블락번호만 있고 처리번호는 생략했다. 두번째 예는 한 자료줄 data line 에 관측치를 두 개를 입력했다. 세번째는 아예 블락번호까지 생략했다. ENTER 치는 것도 귀찮아서 자료줄 하나에다 관측값만 나열할 수도 있다. 공란도 귀찮아 아예 숫자로만 한줄에 57689875882634…와 같이 입력할 수도 있다. 이러한 고정형 fixed format 은 자료항목의 길이가 일정한 경우에만 쓸 수 있으며, 변수 이름과 위치 등을 정의해둔 자료입력사전 data dictionary 을 별도로 작성해야 한다.

그림 3. 원자료 입력시 자료구조와 소프트웨어의 자료판 자료구조

```
1 57 68 98 75
2 88 26 34 85
3 98 68 77 96

1 57 68 98 75 2 88 26 34 85
3 98 68 77 96

57 68 98 75
88 26 34 85
98 68 77 96

57 68 98 75 88 26 34 85 98 68 77 96

576898758826348598687796
```

```
     +------------------------+
     | block   treat   result |
  1. |     1       1       57 |
  2. |     1       2       68 |
  3. |     1       3       98 |
  4. |     1       4       75 |
  5. |     2       1       88 |
  6. |     2       2       26 |
  7. |     2       3       34 |
  8. |     2       4       85 |
  9. |     3       1       98 |
 10. |     3       2       68 |
 11. |     3       3       77 |
 12. |     3       4       96 |
     +------------------------+
```

어느 방법이 제일 좋은가? 손으로 입력하는 것으로만 친다면 마지막 방법이 제일 효율이 좋다. 하지만 자료입력을 검사하는 것까지 고려한다면 세번째가 제일 낫다. 실험한 결과를 적은 표대로 입력하기 때문에 검사하기 쉬운 까닭이다. 다음으로는 사용할 소프트웨어에서 지원하는가 하는 문제다. 다행히 SAS 나 Stata 모두 왼쪽 다섯 가지 자료형태를 오른쪽 자료구조로 읽어올 수 있다. 얼마나 쉽게 읽을 수 있는가는 소프트웨어와 사용자의 능력에 달린 문제다. 만일 어느 소프트웨어를 사용할 지 확실하지 않거나 나중을 위해 자료를 저장하는 것이라면 오른쪽 자료구조를 사용하는 것이 좋다.

8. 어떻게 자료를 입력할 것인가?

가장 먼저 관측치를 구별할 수 있도록 고유한 식별번호 unique identification number 를 입력해야 한다. 어느 관측치가 실제 어느 자료에서 나왔는지 추적할 수 있어야 하기 때문이다. 예컨대, 자료가 엉뚱하게 입력되었을 경우 관측치를 찾아서 확인할 필요가 있다. 당연한 것 같으면서도 많은 사람들이 너무 가볍게 생각하고 있는 점이다. 변수이름은 id 도 좋고 serial 도 상관이 없다. 학번이나 주민등록번호가 있다면 그것을 응용해도 (예컨대, 입학연도나 출생연도를 취하여 번호를 만든다면) 무방하다. 아니면 관측치 (설문지)에 일련번호를 적어 놓고 그 번호를 변수에 입력하면 된다.

[19] 시공간자료 時空間資料는 공간자료 cross-sectional data 와 시간자료 time-series or longitudinal data 가 합쳐진 형태이다. 같은 사람들을 대상으로 같은 설문조사를 매년 실시하여 자료를 쌓아놓았다면 시공간자료라 할 수 있다. 반면 개별 연도 자료는 공간자료 空間資料라 할 수 있다. 50 년간 한국의 GNP 를 모아두었다면 시간자료 時間資料라 할 수 있고, OECD 회원국의 50 년간 GNP 자료라면 시공간자료라 할 수 있다.

자료는 쉽게 숫자로 기록할 수 있는 것도 있고 기호로 표시하거나 글로 설명해야 하는 것도 있다. 전자는 숫자자료 quantitative 이고 후자는 묘사자료 qualitative 이다. 백만원 단위로 측정한 개인소득이나 100 점 만점으로 표시된 학업성적은 쉽게 숫자화할 수 있다. 어느 나라의 제도특성은 독재/민주주의/제국주의 등으로 설명할 수는 있지만, 50% 민주주의, 90% 민주주의와 같이 숫자로 표시하기는 어렵다.

8.1 숫자자료 입력

숫자로 표시된 자료를 얻었다면 그대로를 변수에 입력하면 된다. 물론 정수형이나 실수형 변수를 취한다. 예컨대, 성적이 90 점이면 "90"으로 입력한다. 개인소득이 2 천만원이고 측정단위가 백만원이면, "20"을 입력하면 된다. 1,000 단위를 나타내는 컴마는 생략하는 것이 좋다. SAS 와 같은 소프트웨어는 컴마가 들어있는 숫자를 쉽게 읽을 수 있지만 다른 소프트웨어에서는 불편할 수 있기 때문이다.

숫자가 매우 크거나 매우 작다면 변수의 측정단위를 적절히 조정하여 입력하는 것이 좋다. 심각한 정확성이 요구되지 않는다면 숫자를 단순화하는 것이 좋다. 예컨대, 987,654,321,000 보다는 987,654.321 이나 9,877 억이 훨씬 낫다. 0.000000000789 보다는 0.789 로 측정단위를 조정하여 입력하는 것이 좋다. 숫자가 너무 크면 용량초과 overflow 문제가, 너무 작으면 0 으로 처리되는 문제가 생길수 있기 때문이다. 계산이 불가능하거나 전혀 엉뚱한 결과를 초래할 수 있다.

8.2 묘사자료 입력

숫자로 표시할 수 없는 자료는 어떻게 처리해야 하는가? 제도특성을 묘사하거나 대통령탄핵에 대한 의견을 적도록 했다고 생각해 보라. 가장 단순하고 무식한 방법은 문자형 변수를 취하여 그대로를 입력하는 것이다. 하지만 묘사한 내용이 관측치에 따라 길이가 천차만별이고 어떤 것은 소프트웨어가 허용하는 길이를 넘을 수도 있다. 또 관측치 수가 많으면 자료파일은 황당하게 커질 수 있다. SAS 에서 32,000 바이트 문자형으로 (영어로 3 만 2 천 자까지 쓸 수 있는) 변수 하나를 정의하고 자유답변형 질문 open-ended question 1 만명분을 입력한다고 상상해보라! 32K 바이트면 실수형 변수 8 천 개 (=32,000/4)를 저장할 수 있는 양이다. 1 만명 분이면 그런 문자형 변수 하나를 저장하는데만 320MB (=32K X 10,000)가 필요하다는 소리다. 따라서 묘사한 내용 그대로를 적는 것은 꼭 필요한 경우가 아니라면 미련한 짓이기 십상이다.

그러면 어떻게 하는 것이 좋은가? 묘사 내용을 몇 개 단어로 요약하거나 몇 개 범주로 유형화해보자. 대통령탄핵에 대한 의견이라면, 의회쿠데타다, 수구 꼴통들의 이유없는 저주이자 난동이다, 사과하면 탄핵안하고 안하면 탄핵한다니 말도 안된다, 그렇게까지 해야 하나, 장난삼아 한번 해본 것이다, 재수없게 생긴 고딩의 버르장머리를 고쳐놔야 한다, 상고전성시대를 종식하고 SKY 독점시대를 열었다, 의회민주주의의 승리다, 구국을 위한 대딩들의 위대한 결단이다… 등으로 나눌 수 있을 것이다. 주의해야 할 것은 어떤 묘사이든 간에 반드시 범주 하나에 속할 수 있도록 해야 한다.[20] 범주화를 했다면 묘사 그대로 대신에 범주를 관측값으로 입력하면 된다. 하지만 이것도 변수가 길어지게 되어 좋은 방법이 아니다. 흔히 우리는 범주에 번호를 붙여 그 번호를 대신 입력하곤 한다. 예컨대, "의회쿠데타다"는 1, "말도 안된다"는 2 와 같은 식으로 붙인다.[21] 필요하면 나중에 변수값에 뜻말 value label 을 붙여주면 범주이름으로 입력한 것과 비슷한 효과를 낼 수 있다.

8.3 설문지 자료 입력

설문지에 답변을 입력하는 것도 마찬가지이다. 이미 범주화된 항목에 번호를 붙이고 그 번호를 입력하면 된다. 예컨대, "적극 반대한다"는 1, "반대한다"는 2, "찬성한다"는 3 과 같은 식으로 번호를 붙인다. 성별,

[20] 어떤 묘사가 어느 범주에도 속하지 않는다거나 not thoroughly exhaustive, 한개 이상의 범주에 속한다면 not mutually exclusive 범주화는 잘못된 것이다.

[21] 왜 하필 1, 2, 3 인가? 물론 아무렇게나 (예컨대, 77108, 6.25, -125) 붙여도 상관은 없지만 번호체계를 단순화하고 최소화하는 것이 좋다. 아무렇게나 붙이다보면 나중에 자신도 어느 것이 어느 것인지를 몰라 눈물빼야 한다. 큰 정수, 실수, 음수 등은 피하는 것이 좋다. 답변이 많아서 묶을 필요가 있다면, 1.1, 1.2…, 2.1, 2.2., 3.1, 3.2… 등으로 할 수도 있다.

찬반과 같이 둘 중 하나를 입력해야 하는 경우에는 1 혹은 0 으로 표시하는 것이 좋다.[22] 예컨대, 6 절에서 언급한 설문문항은 변수이름를 male 로 하고, 남자는 1 여자는 0 로 입력한다. 다음과 같은 설문문항을 생각해 보자.

[vote2] ***How many times have you voted so far?***

1. None
2. Once or twice
3. Three to five times
4. More than five times
99. I do not know

변수이름은 "vote2" (투표에 관한 두번째 질문이라는 뜻으로)로 할 것이고, 답한 항목에 붙여진 번호를 입력할 것이다. 변수뜻말은 질문이 길지 않기 때문에 "How may times have you voted?"로 주고, 변수값말은 항목 그대로를 넣어줄 것이다. 그런데, 왜 잘 모른다는 "99"를 썼을까?

8.4 여러 개를 선택하는 질문은 어떻게 입력하나?

자료 입력을 할 때 초보자들을 곤혹스럽게 하는 유형은 하나가 아닌 여러 개를 선택하도록 하는 질문 multiple response question 이다. 예컨대, "해당되는 항목을 순서대로 4 개까지 고르시오."라고 묻고는 1 번부터 한 20 번까지 항목을 늘어놨다고 생각해 보자. 한 술 더 떠서 해당되는 모든 항목을 고르라는 "잔인한" 경우도 가끔씩 있다. 자, 어떻게 입력할 것인가? 변수를 20 개 (a1-a20)를 만들어 놓고 그 항목을 선택했으면 1, 아니면 0 을 넣을 것인가? 아니면 변수는 하나로 하고 관측치를 4 배로 (관측치 하나당 네 개씩) 늘릴 것인가?

두 방법 모두 "아니올시다"이다. 첫번째 방법은 메모리 낭비일 뿐만 아니라 분석하기 위해 변수를 가공해야 한다. 두번째 방법은 자료중복으로 메모리 낭비가 더 심할 뿐만 아니라 그 변수를 제외한 다른 변수를 분석하는 것이 곤란하다. 게다가 네 개 모두를 답하지 않은 관측치가 있을 경우 자료를 종합하는데 큰 문제가 생긴다. 관측치가 독립되지 않고 같은 비중으로 처리되지 않기 때문이다. 정답은 변수를 네 개 (예컨대, choice1-choice4)를 만든 뒤, 첫번째 선택한 항목 (의 번호)을 choice1 에 입력하고, 두번째는 choice2 에, 세번째는 choice3 에, 네번째는 choice4 에 넣는 것이다. 아래 예에서 두번째 사람은 첫번째부터 세번째까지 각각 1, 4, 3 항목을 선택했고 네번째 항목을 선택하지 않았다.

```
     +---------------------------------------+
     | choice1   choice2   choice3   choice4 |
     |---------------------------------------|
  1. |       9         1         5         7 |
  2. |       1         4         3         . |
  3. |       6         7         4         1 |
  4. |       8         7         .         9 |
  5. |       7         3         2         5 |
     +---------------------------------------+
```

이 방법에서 선택한 네 가지 항목 전체 횟수를 계산하기 위해서는 자료를 잘라서 붙여야 stacking up 하지만, 이것이 자연스럽고 훨씬 효율성이 높다.[23] 반복하여 측정하는 경우 repeated measure 도 마찬가지로 반복한 수만큼 변수를 만들고 해당 관측값을 입력하면 된다.

8.5 관측값이 없는 경우

관측값이 없거나 모른다고 답하거나 판명하기 어려운 경우에는 응답없음 missing 으로 처리한다. 잘 모른다거나 답하기 싫다는 것도 missing 으로 간주하기도 한다. 숫자형 변수에서 응답없음은 보통 점

[22] 1 과 2, 다른 숫자, 혹은 문자 (남자, 여자)도 상관은 없으나 자료분석 소프트웨어나 분석기법에 따라서는 꼭 1 과 0 으로 해야만 하는 경우가 있다.

[23] 소프트웨어마다 이런 유형의 질문을 처리하기 위한 방법이 마련되어 있다. SAS 나 Stata 에서 그런 작업을 쉽게 할 수 있는 statement 와 command 가 있으며, SPSS 는 자료를 건들지 않고 새로운 내부 변수를 만들어서 편리하게 횟수를 계산할 수 있게끔 해준다.

period 으로 입력하거나 특정한 값 (예컨대, 99 나 9999) 을 넣는다. 소프트웨어에서 점 대신에 특별한 문자나 숫자를 missing 으로 인식하게 할 수 있다. 문자형 변수에서는 자료입력을 그냥 건너뛰면 된다.

설문조사에서 흔히 5-point 혹은 7-point Likert 잣대를 많이 사용한다. 그런데, 대칭구조로 된 질문 (예컨대, 매우 반대—반대—중립—찬성—매우 찬성) 에서 가운데 항목으로 "그저 그렇다"와 "관심없다" 같은 것을 넣으면 문제가 될 수 있다. 그런 답변과 "모른다" 혹은 응답없음을 구별하기가 현실적으로 힘들기 때문이다. 따라서 대칭구조로 된 질문은 4-point 이나 6-point 잣대를 사용할 것을 권한다.

9. 자료파일 형식

자료는 여러가지 형식으로 파일에 담을 수 있다. 가장 흔한 형식은 아스키 형식 ASCII text 인데, 스프레드시트 spreadsheet 와 데이타베이스 database 도 많이 사용하고 있다.

9.1. 구분자 아스키 형식

아스키 ASCII (American Standard Code for Information Interchange)는 미국정보교환표준코드인데, 컴퓨터에서 사용되는 알파벳, 숫자, 제어문자, 그림문자 256 개 (=2^8)를 정의해 놓은 것이다. 예컨대 0 은 48 번, A 는 65 번, a 는 97 번, +는 43 번, Enter 는 13 번, Space 32 번 식이다.[24] 아스키 형식으로 된 파일은 자료항목과 자료 항목을 구분하는 구분자 delimiter 만을 가지고 있다. OpenOffice Writer (.odt)나 하안글 파일 (.hwp)은 글자 뿐만 아니라 글자의 크기와 글자꼴 font 과 색깔 정보까지 저장하고 있다. 하지만 일단 아스키 형식으로 저장되면 크기나 색깔 등은 모두 사라진다. 아스키 파일에서는 오직 글자만 남기 때문이다. 따라서 아스키 파일은 가장 단순하고 기본적인 형식이면서 저장공간을 가장 적게 차지한다. 또한 거의 모든 소프트웨어가 읽을 수 있기 때문에 호환성 문제가 생기지 않는다.

아스키 형식은 구분자 delimiter 를 무엇으로 사용하느냐에 따라 여러 종류로 나뉠 수 있다. 흔히 자료를 공백 space 로 구분하지만, 열을 맞출 필요가 있으면 탭 tab 을 사용하기도 한다. 이들을 각각 공백으로 구분한 space-delimited 아스키 형식, 탭으로 구분한 tab-delimited 아스키 형식으로 부른다. 공백으로 구분한 아스키 형식은 특별히 자유형 free format 으로 부른다. 자료가 적고 단순한 형식을 가지고 있을 때 유용하다.

자료가 공백을 포함하면 컴마 comma 를 구분자로 이용할 수 있는데 comma-delimited, 이는 스프레드시트에서 많이 사용하는 CSV (Comma Separated Values) 형식이다. 자유형과 함께 가장 널리 사용되는 아스키 파일형식이다. 물론 다른 특수문자, 예컨대 #, $, &, @를 구분자로 사용할 수도 있으며, 때에 따라서는 두 개 문자 이상을 동원할 수도 있다. 표 4 는 가장 많이 사용하고 있는 세 가지 아스키 형식을 보여주고 있다. 탭은 Carriage Return 과 같이 제어문자로 사용되기 때문에 그 자체는 문서편집기에서 보이지 않는다

표 4. 구분자에 따른 아스키 형식

공백 space 으로 구분된 아스키	탭 tab 으로 구분된 아스키			컴마로 구분된 아스키(CSV)
Park 87 40	Park	87	40	"Park",87,40
Kim 85 100	Kim	85	100	"Kim",85,100
Hwang 89 25	Hwang	89	25	"Hwang",89,25

아스키 파일에 자료를 저장하기 위해서는 흔히 vi, eMacs, pico, nano, Notepad 와 같은 일반파일편집기를 사용하거나, oXygen 이나 WinEdit 같이 HTML 이나 프로그램소스를 편집하기 위한 특별파일편집기를 사용할 수도 있다. OpenOffice Writer, WordPerfect, MS Word 같은 전문문서편집기를 사용할 때는 입력이 끝난 다음에 아스키 파일로 변환해야 한다. CSV 형식이나 특수문자를 구분자로 사용할 경우에는 편집기로

[24] ASCII 는 대형컴퓨터에서 사용하던 IBM 의 EBCDIC (Extended Binary-Coded Decimal Interchange Code)와 구분된다. 예컨대, Space 는 아스키에서는 32 번이고 EBCDIC 에서는 64 번에 할당되어 있다.

직접 입력하기보다는 스프레드시트나 데이타베이스를 사용하여 자료를 입력하고 파일변환을 하는 것이 훨씬 낫다.

9.2 고정형 아스키 파일

자료가 일정한 길이와 형식으로 정렬되어 있다면 열 위치로 자료항목을 구분하는 것이 낫다. 고정형 아스키 형식 fixed format 은 구분자 없이 자료의 위치로 자료항목을 구분한다. 예컨대, 변수가 한자리 정수나 문자, 같은 형식으로 된 실수 real number (예컨대, xx.xx)로 되어 있다면 딱 알맞은 방법이다. 이 방법은 많은 자료를 가장 효율성이 있게 저장할 수 있다. 다만 변수가 너무 많으면 파일편집기에 따라 불편할 수도 있으며, 자료를 검사하는데 어려움이 있을 수 있다. 고정형은 열의 위치로 자료항목을 구분하기 때문에 자료파일 외에 변수이름, 변수의 위치, 변수형을 정의한 입력사전 data dictionary 이 반드시 필요하다. 대신에 문자형 변수에 ‘, ’, “, ” 등 어떤 문자가 들어가도 상관이 없다.[25] 다음은 Stata 의 infix 명령어에서 사용하는 입력사전파일 (data.dct)이 어떻게 생겼는지를 보여준다.

```
infix dictionary using d:\data\data.txt {
     long id       1 - 7
     str name      11-40
     int male      41-41
     str birth     51-60
     float income  80-89
}
```

첫번째 단어는 변수형, 두번째는 변수이름, 세번째는 읽어올 관측값의 위치를 말한다. 따라서 Stata 는 data.txt 를 불러온 뒤, 1 번부터 7 번 열을 읽어서 긴정수형 long integer 변수 `id` 에 저장하고, 11 번부터 40 번까지를 읽어서 문자형 string 변수 `name` 에 저장한다. 개인소득은 80-89 에서 읽어서 실수형 변수 `income` 에 저장한다. 날짜형식 (mm/dd/yyyy)인 생일자료를 문자형 변수 `birth` 로 읽었다는 점에 주목하라.

고정형 아스키 파일은 자료를 입력할 때 뿐만 아니라 출력해서 보관할 때에도 많이 사용된다. 호환성 면에서 가장 유리하며 많은 자료를 효율성 있게 보관할 수 있기 때문이다. ISPSR 에 가서 자료를 찾아보라. 고정형 아스키 파일과 입력사전 뿐만 아니라 SAS 와 SPSS 에서 바로 읽을 수 있도록 batch 파일을 제공하고 있다. Batch 파일은 변수이름을 정의하고, 어느 변수를 몇번째 열부터 몇번째 열까지 어떠한 형식으로 읽어야 하는지를 명시하고, 변수뜻말과 변수값말을 정의하는 것을 포함하고 있다.

9.3 스프레드시트

VisiCalc 에서 출발한 스프레드시트 spreadsheet 는 Lotus 1-2-3, Quattro Pro, Excel, OpenOffice Spreadsheet 를 거치면서 많은 각광을 받고 있다. 스프레드시트에서는 손쉽게 자료를 입력할 수 있을 뿐만 아니라 간단한 자료분석까지 할 수 있다. 많은 자료분석 소프트웨어는 스프레드시트 파일을 자료변환없이 바로 읽어들일 수 있다. 일단 자료를 입력했다면 쉽게 여러가지 아스키 파일형식으로 바꿀 수도 있다. 특히 커마로 구분된 CSV 형식과 깊이 연관되어 있기 때문에 일반 상황에 권하는 자료입력 방법이다. 다만 파일이 아스키 형식에 비해 몹시 커지고, 자료분석 소프트웨어에서 지원하지 않으면 먼저 아스키 파일로 변환해야 하는 불편함도 있다.

자료분석 소프트웨어에서 제공하는 자체 편집기도 스프레드시트를 닮은 인터페이스를 가지고 있다.[26] 자체편집기는 자료를 입력한 이후에 자료변환없이 바로 분석을 할 수 있는 것이 장점이다. 하지만 자료의 양이 많거나 복잡한 경우에는 그리 좋은 입력방법은 아니다. Stata 빼고는 자료를 수정한 것을 기록하기 어렵다는 문제도 있다. 최근에는 SPSS Data Entry 처럼 설문양식을 작성하거나 전문으로 자료입력을

[25] 예컨대, CSV 형식에서는 문자형 자료는 보통 큰인용부호(“…”)로 감싸게 되는데, 그 안에 컴마나 인용부호가 들어가면 처리하는데 애를 먹게 된다(e.g., “…, then I read Schumacher’s “Small is Beautiful.””).

[26] 그래서 스프레드시트 자료를 그냥 복사하여 자체 편집기에 붙여넣기를 하는 사람들도 있다. Microsoft 식 “쉽게 쉽게”에 익숙한 사람들에겐 상식이 되겠지만, 자료가 중요하고 양이 많다면 결코 권하는 방법이 아니다.

도와주는 자료입력프로그램이 각광을 받고 있다. 입력자가 범할 수 있는 실수를 줄이고 검사할 수 있는 기능을 가지고 있기 때문이다. 하지만 별도로 구입하여 붙여야 add-on 하기 때문에 부담이 있다.

표 5. 자료분석 소프트웨어가 읽을 수 있는 자료파일 형식

	SAS/BASE 9.1	Stata 11 SE	LIMDEP 9.0	R 2.9	SPSS 17.0
ASCII (Space)	O	O	O	O	O
ASCII (Tab)	O	O	O	O	O
ASCII (Comma)	O	O	O	O	O
ASCII (fixed)	O	O	O	O	O
ASCII (Other)	O	O		O	O
Lotus 1-2-3	O		O		O
Quattro Pro					
EXCEL	O		O		O
dBase III+/IV	O			O	O
Foxbase/FoxPro	O				
Access	O				

9.4 데이타베이스

대용량 자료를 입력해야 하는 경우에는 데이타베이스가 유리하다. IBM DB2, Oracle, Sybase, MySql, PostgreSQL, Microsoft SQL Server 등과 같은 서버용 데이타베이스엔진 뿐만 아니라 dBase III+, dBase IV, FoxPro, Access, Paradox 등과 같은 데이타베이스 클라이언트를 사용할 수 있다. 다양한 입력양식을 지원하고, 자료중복을 최소화하고, 자료 무결성 integrity 을 점검함으로써 대용량 자료를 좀더 쉽고 정확하게 입력할 수 있다. 또한 표준화된 SQL (structured query language) 같은 질의어로 원하는 정보를 다양한 방법으로 뽑아내어 처리할 수 있는 장점이 있다.

데이타베이스는 자료의 양이 방대하거나 복잡한 경우에 주로 사용한다. 여러가지 자료판 (예컨대, 학생, 교수, 과목)을 만들어서 관계형 데이타베이스 relational database 로 서로 연결해서 정보를 뽑아내야 할 때 유용하다. 입력할 때 일어날 수 있는 실수를 막을 수 있는 기능을 사용할 수 있으며 (예컨대, 성별에 1 이나 0 외에 입력할 수 없도록 한다), 다양한 방법으로 자료를 검사할 수 있다. 다만, 데이타베이스에 관한 지식이 어느 정도 필요하며 파일이 매우 커진다는 약점도 있다. 또한 자료분석 소프트웨어에서 지원하지 않으면 자료를 입력한 후에 아스키 파일로 변환해야 하는 불편함도 있다. 최근에는 온라인 설문결과를 바로 각종 파일로 저장하거나 데이타베이스 서버에 저장하는 방법을 사용하기도 한다. 표 5 는 주요 자료분석 소프트웨어가 지원하는 파일형식을 정리하고 있다.

9.5 어떤 파일 형식으로 입력할 것인가

그러면 어떤 파일형식으로 입력하는 것이 가장 좋은가? 자료의 양이 많은지 적은지, 구조가 복잡한지, 얼마나 자장공간을 차지하는지, 자료분석 소프트웨어에서 얼마나 지원하는지, 파일을 변환해야 하는지, 자료입력과정에서 일어나는 실수나 자료의 일관성 등을 검사하기에 편한지 등을 고려해서 결정해야 한다.

자료 양이 많지 않다면 자유형 아스키 형식이나 스프레스시트를, 많다면 고정형 아스키 형식이나 데이타베이스를 사용하는 것이 좋다. 문자형이 많고 복잡하다면 데이타베이스가 유리할 것이다. 스프레드시트와 데이타베이스는 아스키 파일에 비하여 덩치가 커진다는 것을 유념해야 한다. 많은 자료를 효율성있게 저장하려면 고정형 아스키 형식이 좋다. 성능이 좋은 문서편집기가 필요할 것이다. 다른 소프트웨어와 호환성을 고려하면 아스키 형식을, 자료 수준을 엄밀하게 관리할 필요가 있다면 데이타베이스에 자료를 입력하는 것이 좋다. 물론 데이타베이스를 구축하기 위해서는 관련 지식과 기술을 가지고 있어야 한다.

10. 외부자료 읽기와 자료변환

자신이 직접 자료를 입력할 수도 있지만 이미 입력된 자료파일을 소프트웨어로 읽어들이는 경우도 생각해볼 수 있다. 또한 자료변환 프로그램으로 자료파일을 한 형식에서 다른 형식으로 변환할 수도 있다.

10.1 ODBC 를 통하여 읽어오기

SAS, Stata, R 같은 자료처리 소프트웨어는 ODBC (Open Database Connectivity)와 네트웍 프로토콜을 통해서 자료를 읽어올 수 있다. 개방형데이타베이스연결 ODBC 은 다양한 자료형식을 불러올 수 있는 function call 표준이다. 일단 Quattro Pro, Excel, dBase, FoxPro, Access 의 ODBC 드라이버 driver 를 사용해 자료원이름 DSN (Data Source Name)을 정의해 놓으면, SAS 와 Stata 와 같은 소프트웨어에서 바로 읽고 수정할 수 있다. ODBC 가 자료분석 스프트웨어와 자료파일 사이에서 정보교환을 매개하는 것이다. 다음 `odbc` 명령어는 `cancer2000` 으로 정의된 DSN 에서 `cancer` 라는 자료판 table 을 읽어온다.

```
. odbc load, table("cancer") dsn("cancer2000") clear
```

10.2 Network 에서 읽어오기

SAS, Stata, R 등은 http (hypertext transfer protocol)와 ftp (file transfer protocol)와 같은 네트워크 프로토콜을 통하여 손쉽게 자료를 읽어올 수도 있다. 이러한 기능은 네트워크 환경에서 작업을 하는 요즘같은 시대에 잘 어울린다. 특히 서버 server 에 저장된 대단히 큰 (예컨대, 100GB), 그래서 복사해오기 곤란한 자료파일을 읽어올 때 유용하다. 예를 들면, 다음 `use` 명령어는 http 프로토콜을 통하여 웹사이트에 저장되어 있는 Stata 자료파일을 바로 읽어들일 수 있다.

```
. use http://mypage.iu.edu/~kucc625/documents/cancer.dta, clear
```

10.3 XML 파일에서 읽어오기

SAS, Stata, R 은 XML 형식을 읽어올 수 있다.[27] 꾸미기 언어로 된 문서는 형식으로는 아스키 파일이지만 내용으로 보면 자유형이나 고정형 아스키 형식과 다르다. 글자 내용 외에 Markup 을 이용하여 문서정보와 글자를 꾸미는 정보를 덧붙이기 때문이다. 흥미롭게도 XML 에서는 자료항목이 Markup 으로 정의되기 때문에 마치 구분자 아스키 형식처럼 자료를 읽어오기 쉽다.

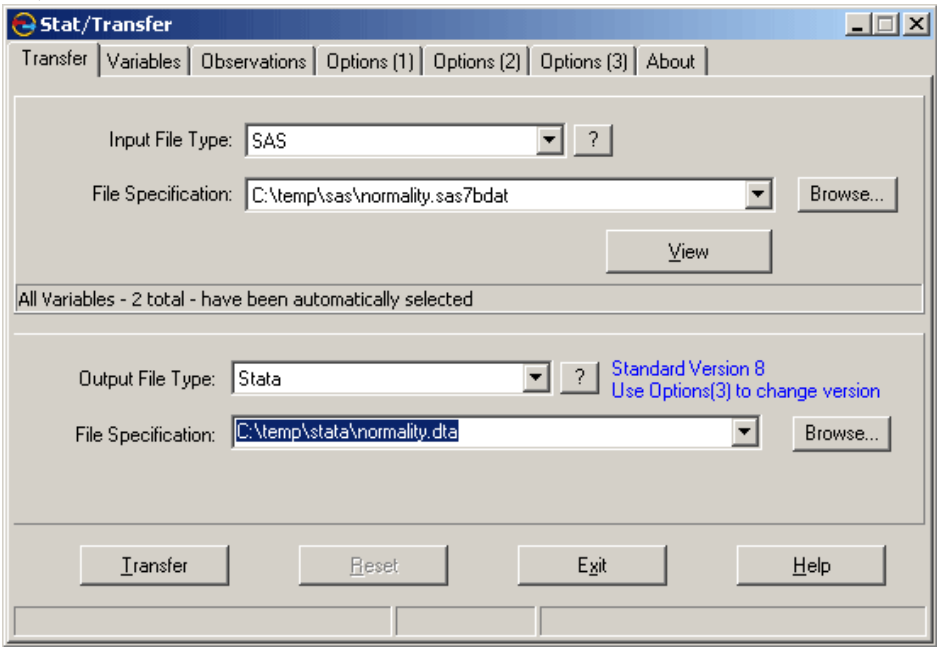
최근 꾸미기 언어로 된 문서가 많아지고 웹문서에서 자료를 뽑아내야 할 경우가 많이 생겼다. 예컨대, 서버에서 직접 설문조사를 하는 경우 응답 (혹은 분석내용)을 XML 문서로 출력하여 웹에서 바로 확인할 수 있도록 하고 있다. XML 로 문서를 작성하여 웹에 올리거나 웹에서 XML 로 작성된 자료를 구하는 경우가 많이 있기 때문에 중요성이 커지고 있다.

10.4 다른 파일 형식과 자료변환

SAS, Stata, SPSS 등 소프트웨어는 모두 고유한 자신의 파일형식을 가지고 있다. 이들은 아스키 형식을 읽을 수 있을 뿐만 아니라, Import 기능을 사용해서 다른 소프트웨어 파일형식을 읽거나, 스프레드시트와 데이타베이스 파일을 읽어올 수 있다. 특히 SAS 는 거의 대부분의 파일형식을 지원하며, 다양한 방법으로 복잡한 아스키 파일까지 읽을 수 있는 큰 장점을 가지고 있다 (표 5). R 은 read.xport, read.dta, read.spss 등의 함수를 써서 SAS transport 형식, Stata .dta 파일, SPSS .sav 파일을 읽을 수 있다. SPSS 는 Excel 이나 dBase 뿐만 아니라 SAS 와 Stata 자료파일을 바로 읽을 수 있다.

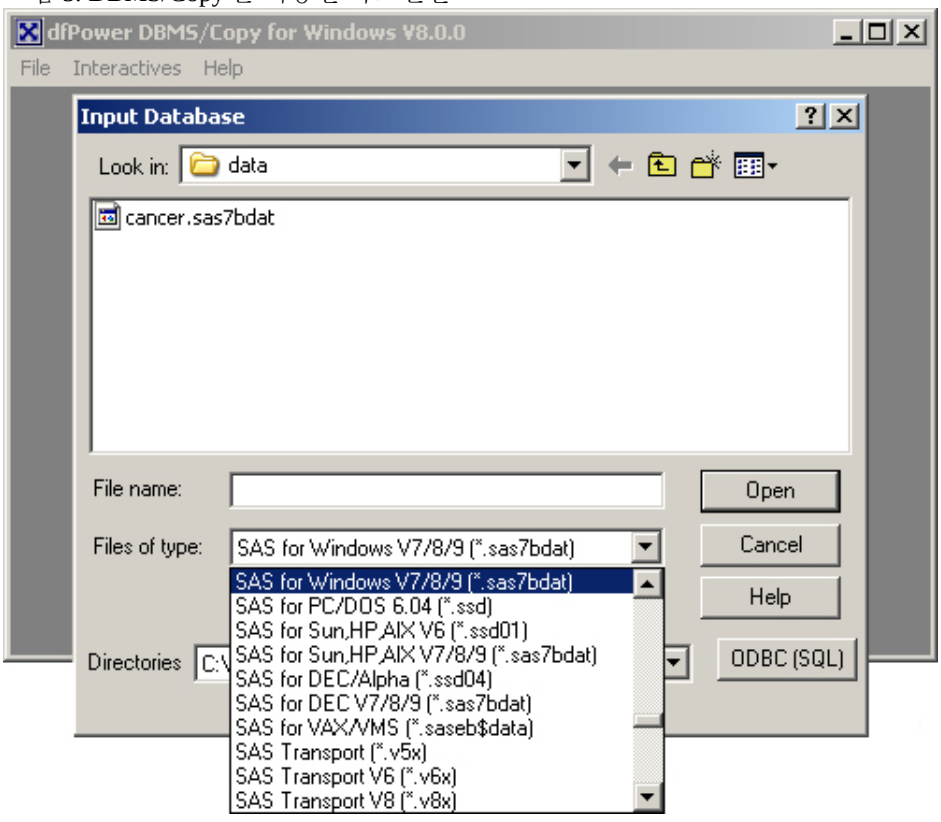
[27] "꾸미기 언어" Markup Language 는 문서이름, 저자 등 문서에 관한 정보와 글자꼴, 글자 크기 등 글자에 관한 정보를 저장할 수 있게 한다. 웹 World Wide Web 에서 사용하는 HTML (Hypertext Markup Language), XHTML (Extensive HTML), XML (Extensible Markup Language) 등 뿐만 아니라 문서편집을 위한 LaTex 나 PostScript 도 꾸미기 언어에 포함된다. *http://en.wikipedia.org/wiki/Markup_language*

그림 4. Stat/Transfer 를 이용한 자료변환



자료분석 소프트웨어에서 지원하지 않는 파일형식을 읽거나, 자료분석 소프트웨어간 혹은 컴퓨터 기종 간 파일 변환이 필요하거나, 옛날 자료파일을 읽어야 한다면 자료변환 유틸리티를 사용하는 것이 좋다.

그림 5. DBMS/Copy 를 이용한 자료변환



http://www.masil.org

가장 많이 사용하고 있는 유틸리티는 Stat/Transfer (http://www.stattransfer.com/)와 DBMS/Copy (http://www.dataflux.com/)인데, 각종 파일형식을 서로 변환해 준다(그림 4, 5).예컨대, Paradox (.db)나 Quattro Pro V7 (.wb3) 자료파일을 SAS 9.1 (.sas7bdat) 형식으로 바꾼다든가, SAS 6.12 자료파일 (.sd2)을 Stata 8.0 (.dta)로 바꿀 수 있다. 또한 Stata 8.0 파일을 Stata 6.0 파일로 바꾸거나, IBM AIX 용 SAS 6.12 자료형식 (.ssd01) 을 DEC 용 SAS 자료형식 (.ssd04)으로 변환할 수 있다. 자료변환 유틸리티는 변수뜻말과 변수값말을 변환해 주기도 한다.

11. 어떻게 입력한 자료를 정제하고 가공할 것인가?

자료를 입력했다고 해서 분석할 준비가 된 것은 아니다. 쌀과 채소와 고기를 샀다고 해서 바로 밥을 먹을 수 있는 것이 아니다. 못먹는 것을 골라내고, 깨끗하게 다듬고 씻고, 알맞은 크기로 썰고 다지고, 삶고 찌고 구워야 비로소 입에 들어갈 수 있는 것이다. 물론 잘 차려진 밥상이 있다면 숫가락만 들면 된다. ICPSR 에서처럼 이미 처리가 잘 된 자료라면 해당 자료처리 소프트웨어에서 읽기만 하면 된다. 하지만 대부분은 자료분석을 하기 전에 여러가지 방법으로 입력한 자료를 정제精製하고 적절하게 가공할 필요가 있다. 그렇지 않으면 분석을 했다 해도 실효성과 신뢰성이 떨어지기 때문이다.

특히 유념해야 할 것은 입력한 자료를 정제하고 자료를 변환하는 전 과정을 기록해 두는 일이다. 변수를 어떻게 입력 coding 을 했으며, 어떻게 스프트웨어에서 읽었으며, 어떻게 자료를 검사하여 고쳤는지, 어떻게 변수를 조작하였는지를 소상하게 적어두어야 한다. 소위 로그 log 파일을 작성하여 보관하지 않는다면 그 자료가 정당하게 만들어졌는지를 보여줄 길이 없다. 또한 한달도 지나지 않아 자신이 어떤 과정을 거쳐 자료를 만들었는지를 까맣게 잊게 될 것이고, 나중에 필요하다 해도 두번 다시 그 자료를 똑같이 만들어낼 수 없을 것이다. 간단히 말해서 log 파일 없이 자료를 분석했다면 그 결과를 신뢰할 수 없으며, 분석자체가 과학이라고 말할 수 없다. 이런 의미에서 SAS 와 Stata 는 자료처리 과정을 기록해 둘 수 있는 매우 훌륭한 도구라 할 수 있다.

11. 1 자료 선별

자료처리 전과정에 걸쳐 자료의 순도를 떨어뜨릴 수 있는 위험이 도사리고 있다. 처음부터 측정을 잘못할 수도 있고, 측정한 것을 잘못 기록할 수도 있다. 설문조사에서 응답자가 불성실하게 답변하는 경우가 적잖이 있다. 아무렇게나 답하거나, 한가지 답으로 일관하거나, 또는 답을 건너띄는 경우를 생각해 보라. 우연이든 실수에서든 설계된 조건에 맞지 않은 상황에서 설문작성이 되었을 수도 있다. 따라서 연구질문에 비추어 쓸만한 관측치를 먼저 선별해야 한다.

11. 2 자료 입력과 정제

선별된 자료를 입력하는 과정을 생각해 보자. 사람들이 하는 일이기 때문에 실수가 없을 수 없다. 자신의 자료를 직접 입력하는 것이 아니라 남이 입력해주는 경우에는 아무래도 실수가 더 많을 것이다. 백만단위에서 숫자하나가 잘못 입력되었다고 생각해 보라! 따라서 의도하지 않은 실수를 줄일 수 있는 여러가지 방법을 동원할 필요가 있다. 예컨대, SPSS Data Entry 나 데이타베이스의 picture 기능처럼 허용하는 변수값 범위를 정하거나 일정한 숫자나 문자만 받아들이게끔 할 수도 있다. 두 사람에게 자료를 나누어 입력하고 서로 점검하게 할 수도 있다. 자세한 설명은 Long(2009)의 6 장을 보라.

자료가 입력되었으면 자료를 정제해야 data cleaning 한다. 황당한 관측값이나 논리적으로 맞지 않는 자료를 찾아내어 사실관계를 확인해야 한다. 숫자로 입력한 경우 합이나 평균이나 분산을 계산해서 모난 값 (특이값) outlier 이 있는지 확인해 본다. 큰 단위 숫자나 소수점 위치나 부호 (양수이어야 하는데 음수가 있는지) 등을 확인해 본다. 이름변수나 순서변수인 경우 횟수표 frequency table 를 만들어서 원하지 않는 값이 들어 있는지를 확인한다. 특히 횟수가 너무 많거나 너무 적은 경우를 눈여겨 보라. 또한 논리 일관성이 없는 관측치가 있는지도 검사해 볼 필요가 있다. 예컨대, 최종학력이 고졸이라고 답을 하고 대학전공에도

답을 했다고 생각해 보라. 모난값이나 의심스러운 관측값이 발견되면 자료원본이나 설문서를 대조하여 자료를 수정한다.

11. 3 자료변환 資料變換

자료분석에서 입력된 변수 그대로를 사용하는 경우는 드물다. 많은 경우 분석하는데 적절한 변수를 만들어낼 필요가 있다. 먼저, 변수이름이 적절하게 붙여졌는지를 확인하자. 필요하면 이름을 바꾸고 수정내용을 입력사전에 반영한다. 그리고 변수뜻말을 붙이고 변수값말을 달아보자. 최대한 짧고 간단하게 만드는 것이 상식이다. 입력사전에 기록해 두는 것도 잊지 말자. 적절한 기술통계를 사용하여 제대로 변수뜻말과 변수값말이 달렸는지를 꼭 확인하라.

먼저 반대순서로 reverse ordered 질문한 항목이 있으면 원래순서로 바꿔준다.[28] 다음 `recode` 명령어는 반대순서로 된 변수 `vote9` 을 역산逆算하는 예를 보여주고 있다.

```
. recode vote9 1=4 2=3 3=2 4=1 *=., gen(vote9rev)
```

생년월일로 입력이 되었다면 나이를 계산해야 할 것이다. 또 나이를 그대로 쓰지 않고 10 대—20 대—30 대 식으로 구분하려면 또다시 가공을 해야 한다. 개인소득이라면 인플레이션을 고려하여 실질소득을 계산하거나 선형관계로 바꾸기 위해 로그함수를 취해야 할 상황도 있다. 가처분소득을 사용해야 한다면 개인소득에서 세금낸 것을 빼서 새로운 변수에 저장해야 한다. 너무 크거나 너무 작은 값도 바꿔주고, 측정단위(예컨대, 환율, 미터법)를 의미있게 통일해줘야 한다. 어떤 공식을 적용해서 분석에 필요한 변수를 만들 수도 있다. 드러난 변수 manifest variable 로 숨은 변수 latent variable 를 측정했다면, 요인분석을 통하여 숨은 변수를 추정해야 한다.

자료 변환을 할 때에는 원래 변수를 직접 조작하지 말고 변환결과를 반드시 새 변수에 저장해야 한다. 나중에 원래 변수를 사용하거나 다른 방법으로 변환해야 할지도 모르기 때문이다. 또한 무슨 목적으로 어떠한 변수를 어떠한 가정에서 변환하였는지를 입력사전에 기록해 두어야 한다. 필요하면 나중에 자료변환이 적절했는지 점검할 수도 있으며, 변환이 복잡한 경우에 참고해야 하기 때문이다. 아예 변환에 사용된 script 를 별도로 보관하거나 입력사전에 script 자체를 복사해 넣는 것도 좋은 방법이다. 이는 다른 사람에게 자료처리과정을 투명하게 보인다는 점에서도 의미있는 일이다.

11.4 변수순서를 조정하고 입력사전을 마무리하자

필요한 변수변환이 다 되었으면 이제 변수순서를 생각해 보자. 의미있는 순서로 정렬되었는지, 비슷한 변수끼리 묶였는지 점검해 보라. 분석에 필요한 주요 변수가 찾기 어려운 위치에 있는지도 따져보라. 경우에 따라서는 (변수가 수천 개가 있다면) 변수찾다가 날새는 경우도 있음을 기억하라. 자료파일이 너무 크고 복잡하다면 원본파일은 그냥 두고 필요한 변수만 떼어내어 새로운 자료파일을 만들 수도 있다. 변수순서를 조정했다면, 지금까지 자료를 정제하고 가공한 과정을 입력사전이 제대로 반영하였는지를 확인한다.

12. 어떻게 처리된 자료파일을 보관할 것인가?

입력된 자료를 깨끗하게 씻고 요리하여 성공적으로 분석을 마쳤다고 생각해보자. 많은 사람들이 돈을 들여 모은 자료를 그냥 구석에 방치해 두는 것이 현실이다. 다른 사람들과 같이 사용해서는 안된다는 강박관념이 있는 것인지, 보여줄 수 없을 만큼 자료분석에 자신이 없는 것인지 알 수는 없다. 하지만 정부기관, 대학, 비영리 재단을 포함해서 쓸만한 자료를 모아놓고 제공하는 곳이 많지 않다는 것은 여러가지를 시사해준다. ICPSR 과 같은 자료도서관이 아니더라도 최소한 학위논문에 사용된 자료를 모아서 공개하는 것이 여러모로 필요하다.

[28] 응답자가 성실하게 답하는지 점검하기 위해 일부러 항목의 순서를 바꾸는 경우가 종종 있다. “적극 반대—반대…” 순으로 묻고 뒤에서 같은 질문을 반대로 표현하여 내용상 “적극 찬성—찬성…” 순으로 물을 수 있다.

12.1 꼭 보관해야 할 것

자신이 모아서 처리한 자료를 남에게 보여준다고 생각해 보자.[29] 무엇이 필요할 것인가? 먼저 자료 자체가 필요할 것이다. 고정형 아스키 파일 fixed format ASCII text 로 저장하는 것이 일반적이다.

자료자체만으로는 무엇인지 모르기 때문에 입력사전 data dictionary 이나 입력설명서 codebook 가 필요할 것이다. 제 9 절에서 설명된 Stata 예와 같이 입력사전에다 자료를 읽기 위하여 변수형, 변수이름, 변수값의 위치, 읽는 형식과 같은 정보만을 포함할 수도 있다. 하지만 입력설명서에 이러한 정보 외에 자료를 어떠한 목적으로 누가 수집했는지, 모집단은 어떠한 것이며, 어떻게 표본을 정했는지, 어떤 과정을 거쳐 자료를 수집했는지를 상세하게 기록해 둘 필요가 있다. 자료를 어떻게 정제하고 변수를 어떻게 가공하였는지도 밝혀두어야 한다. 자료파일의 특이한 점이나 연구자가 주의해야 할 사항을 꼼꼼하게 적어두는 것이 좋다. 또한 변수별 평균, 분산, 횟수와 같은 기술통계량까지를 입력설명서에 포함시키면 매우 유용할 것이다.

가능하다면 많이 사용되고 있는 자료처리 소프트웨어에서 바로 읽을 수 있도록 batch 파일을 제공해야 한다. 즉, 그 batch 파일만 실행하면 원래 연구를 수행했던 사람과 똑같은 조건에서 분석을 할 수 있어야 한다. 따라서 script 에는 변수정의 뿐만 아니라 변수뜻말과 변수값말을 포함하는 것이 좋다. 물론 자료정제와 변수변환과정까지 담아야 한다. 간단한 자료라면 batch 파일 자체가 입력사전을 포함할 수도 있다. 하지만 자료 덩치가 크고 (변수도 많고 변수뜻말도 많고) 변수변환이 복잡하다면 별도 파일로 보관하는 것이 바람직하다. 이런 경우 전체 로그파일을 덧붙이는 것도 좋은 방법이다. 최소한 SAS 와 Stata 용 배치파일을 제공하는 것이 좋다.[30] 따라서 다음과 같은 정보를 제공해야 제 3 자가 자료를 제대로 사용할 수 있다.

- 자료파일 data set file
- 입력사전 data dictionary 이나 입력설명서 codebook
- Batch 파일: 자료파일을 특정 자료처리 소프트웨어로 읽어들이는 script

12. 2 고정형 아스키 형식으로 저장하자.

자료파일은 흔히 고정형 아스키 형식으로 저장한다. 메모리를 가장 적게 차지하고 호환성이 좋기 때문이다. 물론 공백이나 탭이나 컴마로 구분한 아스키 형식으로 출력해 놓을 수도 있다. 또한 스프레드시트나 데이타베이스로 자료를 제공할 수도 있는데, 아스키 파일에 비해 파일의 덩치가 매우 커질 뿐만 아니라 판 version 에 따라 호환성 문제도 생긴다. 물론 자료 생성과 축적이 대형 데이타베이스에서 이루어지는 것이라면 선택할 여지는 없다. 꼭 스프레드시트나 데이타베이스 파일로 저장해야 한다면 가장 많이 사용하는 소프트웨어 형식 (예컨대, Quattro Pro, Excel, dBase III+, FoxPro, ACCESS) 으로 저장할 것을 권한다.

12. 3 복사본을 만들어 안전한 곳에 저장하자.

개인용컴퓨터든 휴대용 컴퓨터든 망가지거나 잃어버릴 수 있다. 인쇄된 매체든 디지탈 매체든 손실되거나 분실될 위험은 항상 존재한다. HDD, FDD, ZIP 뿐만 아니라 CD, DVD 등도 망가질 수 있음을 알아야 한다. 따라서 언제나 복사본을 만들어 안전한 곳에 보관하는 것을 습관화해야 한다. 현재 Flash memory 가 값싸게 보급되고 있어서 많은 연구자들이 Compact Flash (CF)나 그와 비슷한 매체를 사용하고 있는데, 그런 반도체 매체가 자기매체인 HDD, FDD, ZIP 등에 비해 전기충격에 약하다는 것을 간과하고 있다. 실제로 중요한 자료를 복사본 없이 CF 에 저장하고 작업을 하다가 CF 가 망가지는 바람에 낭패를 본 사람들이 있다. 학위논문을 쓰거나 연구과제를 수행해야 한다면 휴대용 컴퓨터나 CF 에서만 작업하는 일은 절대 해서는

[29] 사생활보호나 기타 법에 관련된 문제를 고려하여 적정한 자료공개 수준을 결정해야 한다. 예컨대, 전체 자료 중 일부 변수 혹은 관측치만 제공하거나 관측값 일부 혹은 전부를 조작하여 제공할 수 있다 (자료제한 restricted data). 또한 자료에 접근할 수 있는 권한을 일정한 사람들에게만 제공할 수도 있다 (접근제한 restricted access).

[30] 진지하게 자료분석을 하고자 하는 연구자라면 SAS 나 Stata 를 사용할 것을 적극 권한다. 자료 입력, 처리, 분석 전 과정에서 다른 자료분석 소프트웨어 (특히 SPSS)와는 차별화된 기능을 제공하고 있다.

안된다. 또한 유출되어서는 안될 자료인 경우 특별한 관심을 기울여 관리할 필요가 있다. 이에 관한 것은 Long(2009) 8 장을 참조하라.

13. 결론

이렇게 해서 많은 사람들이 자료를 공유하게 되면 연구자가 더 책임감을 갖게 될 것이고 연구가 좀더 객관적일 수 있다. 수행한 연구과정을 따라가면 어떻게 연구결과를 얻게 되었는지를 반복 replication 할 수 있기 때문이다. 이처럼 연구과정을 투명하게 드러낸다면 어떤 연구가 어떠한 면 (연구 설계, 측정, 처리, 분석 등)에서 강하고 약한지를 알아낼 수 있을 것이다. 그런 비판과정을 통해 서로 잘못을 깨달을 수 있으며, 좋은 점을 배울 수 있는 것이다.

이렇게 되면 “쉽게 쉽게”에 길들여진 자들의 무책임한 자료분석 때문에 세상이 어지럽게 되는 일이 줄어들 것이다. 이론이나 모델과 무관하게 통계적 유의성만을 쫓아 자료를 낚시질하는 data fishing (순수한 data mining 과 거리가 먼) 족속들이 설 자리를 잃게 될 것이다. 학문발전이나 사회기여가 전혀 없는 단지 “그들만의 자료놀이”가 사라지게 될 것이다. 우리도 조만간 ICPSR 과 같은 훌륭한 자료은행을 운영하게 되기를 바란다.

용어목록

개념 concept | 개념조작화 operationalization
식별변수 identification variable
관측값, 변수값 value of observation | 모난값, 특이값 outlier
관측치 observation, case, subject
독립변수, 설명변수, 오른쪽 변수 independent/explanatory/right-hand side (RHS) variable
데이타베이스, 자료집 database
    자료개체, 관측치 record, entity
    자료항목, 변수 field, attribute
    자료판 table, data set
모집단 population | 표본 sample
변수 variable | 상수 constant
    연속변수 continuous variable
    마디변수 discrete variable
    두쪽변수 binary variable
변수뜻말 variable label | 변수값뜻말 value labels of variables
변수법
    한변수법 univariate method
    두변수법 bivariate method
    여러변수법 multivariate method
변수형 variable type
    실수형 float, real, double-precision
    정수형 integer, long integer, byte
    문자형 string, character
    논리형 boolean
분석단위 unit of analysis | 관측단위 unit of observation
시공간자료 時空間資料 panel data |
    공간자료 空間資料 cross-sectional data
    시간자료 時間資料 time-series/longitudinal data
실험집단, 처리집단 treatment group | 통제집단, 비교집단 control group
아스키 파일 ASCII file
    자유형 free format
    고정형 fixed format
연구질문 research question | 이론 theory | 이론틀 framework
요인분석 factor analysis
    숨은 변수 latent variable
    드러난 변수 manifest variable
응답없음, 무응답 missing | 채워넣기 imputation
입력사전 data dictionary | 입력설명서 codebook
자료발생기 data generation process (DGP)
종속변수, 반응변수, 왼쪽 변수 dependent/response/left-hand side (LHS) variable
    분류형 종속변수 categorical dependent variable
    제한된 종속변수 limited dependent variable
    잘린, 배제된, 선택된 자료 censored, truncated, selected data
    수명자료 duration/survival time data
회귀 regression
    일반선형회귀 ordinary least squares (OLS)
    여러변수회귀 multivariate regression

참고문헌

Burlew, Michele M. 2002. *Reading External Data Files Using SAS: Examples Handbook.* Cary, NC: SAS Institute.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Charter, L. F. 1971. "Inadvertent Sociological Theory." *Social Forces*, 50:12-25.

Greene, William H. 2007. *LIMDEP Version 9.0 Reference Guide*. Plainview, NY: Econometric Software. http://www.limdep.com/

Greene, William H. 2003. *Econometric Analysis, 5th ed.* Upper Saddle River, NJ: Prentice Hall.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables: Advanced Quantitative Techniques in the Social Sciences.* Sage Publications.

Long, J. Scott. 2009. *The Workflow of Data Analysis Using Stata*. College Station, TX: Stata Press.

Maddala, G. S. 1983. *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

R-Project. 2009. http://www.r-project.org/

SAS Institute. 2005. *SAS 9.1 Language Reference: Concepts*, Version 9. Cary, NC: SAS Institute. http://www.sas.com/

Stata Press. 2007. *Stata User's Guide, Release 10*. College Station, TX: Stata Press. http://www.stata.com/

Stata Press. 2007. *Data Management Reference Manual*.College Station, TX: Stata Press. http://www.stata.com/

문서수정

1994. 1 초고 (0.5 판) 고려대학교 대학원 행정학과 (SAS Seminar)
2005. 8 수정 (1.0 판) Political Science and School of Public and Environmental Affairs, Indiana University
2006. 6 수정 (2.0 판) Political Science and School of Public and Environmental Affairs, Indiana University
2009. 2 수정 (2.1 판) University Information Technology Services, Indiana University
2009. 11 수정 (2.2 판) University Information Technology Services, Indiana University