통계 소프트웨어의 특징 비교

박진우

수원대학교 통계정보학과

Characteristics in Softwares for Statistical Analysis

Park, Jinwoo

Dept. of Applied Statistics, University of Suwon

1. 서 론

"미래에 유능한 시민이 되기 위해서는 읽고 쓰는 능력만큼이나 통계적 사고방식이 필요하게 될 것입니다." (H. G. Wells)

위 말은 지식정보화 시대에 통계의 중요성을 강조하는 말이다. 통계학을 뜻하는 'statistics'라는 단어가 영국의 브리태니커 사전에 최초로 등재된 것이 1799년이라는 사실에서 알 수 있듯이 통계학은 불과 200 년 정도의 역사를 지닌 학문이라고 할 수 있다. 이처럼 그다지 오랜 역사를 지녔다고 보기 어려운 학문이 어떻게 그렇게 빠른 속도로 여러 분야에 확산되어 사용될 수 있었을까? 또한 과연 통계를 아는 것이 교양인의 필수조건이 되는 그런 시대가 올 것인가?

통계학은 불확실성이 포함된 현상을 기술하는 학문이다. 따라서 과거 결정론적(deterministic)인 사고방식이 지배하던 시대에는 통계학이란 학문이 존재할 수가 없었다. 그러나 소위 과학혁명 이후 자연과학이나 사회과학에서 나타나는 불확실성(uncertainty)에 주목하게 되면서 현대적인 의미의 통계학이 태동하게 되었다. 가우스(Gauss)는 천체 관측 데이터를 보면서, 갈톤(Galton)은 유전학 관련 데이터를 보면서, 퀘틀레(Quettlet)는 장병들의 신체 관련 데이터를 보면서 각각의 데이터에 포함된 불확실성을 오차의 확률분포라는 개념으로 설명하였는데 이들의 이러한 사고방식은 근대 통계학의 기초를 다지게 되었다. 오늘날 각 학문 분야에서 계량적인 데이터에 근거한 과학적 방법론이 널리 자리 잡게 되면서 통계학은 약방의 감초처럼 각 분야에 급속하게 스며들었다.

여러 학문 분야에서 통계적 방법론이 중요한 도구로 활용되기 시작하면서 각 분야의 전공자들에게는 통계를 작성하고 활용할 줄 아는 능력이 필요하게 되었다. 그러나 이런 필요에 의해 통계학 입문 과목을 한 두 학기 수강한 비전공자들이 공통적으로 통계학이 복잡하고 어렵다는 소감을 말하곤 한다. 통계학의 기본개념을 이해하기 보다는 복잡한 수식 계산에 치여서 통계학에 대한 자신감을 잃어버리는 경우가 허다한 실정이다. 따라서 실제 상황에서 구체적인 데이터가 주어졌을 때 이 데이터를 이용하여 필요한 통계를 적절히 계산하고 그 결과를 적절하게 활용하는 면에서 부담을 느끼게 된다. 이럴 경우 통계에 대한 바른 이해 없이 다른 유사한 연구에서 사용한 통계 - 가령, t-검정, ANOVA, p-값 등 -를 기계적으로 계산하여 그대로 사용하기 쉽다.

그런데 통계 응용의 분야에 획기적인 전환을 불러오는 사건이 발생하게 되는데 그것은 바로 PC의 보급과 함께 PC용 통계계산 소프트웨어들이 개발된 것이다. 통계학 비전공자 뿐 아니라 전공자들에게 있어서도 통계 계산은 매우 까다롭고 어려운 문제였는데 이를 해결하기 위해 이미 1960 년대부터 SPSS와 같은 전문적인 통계 소프트웨어들이 개발되었다. 하지만 당시의 소프트웨어들은 모두 대형 컴퓨터 기종을 위한 것이기 때문에 일부 전문가들을 제외

하고는 여전히 사용하기 어려운 형편이었다. 그러나 1980년대 PC가 개발되어 보급되면서 여러 연구자들이 PC를 위한 통계 소프트웨어들을 개발하였는데 SPSS, SAS, Minitab, R 등 다양한 종류가 세계적으로 널리 사용되어 오고 있다. 여러 종류의 PC용 소프트웨어가 서로 경쟁하게 되면서 꾸준히 업그레이드되어 오늘날에는 통계 전문가가 아닌 일반인들도 사용할 수 있는 도구가 된 것이다.

이 글의 목적은 우리나라에서 현재 가장 널리 쓰이고 있는 통계 소프트웨어들인 SPSS, SAS, R을 간단히 소개하고 그 특성을 비교, 설명하는 것이다. 대표적인 소프트웨어들의 특성을 안다면 비전문가의 입장에서 통계 소프트웨어를 이용하려고 할 때 매우 유용하리라고 생각한다. 개개의 통계 소프트웨어들은 각각의 특징이 있고 장, 단점이 있으므로 사용자의 특성에 따라 각자에게 적합한 소프트웨어가 있을 것이기 때문이다. 이 글에서 각각의 통계 소프트웨어에 대해 체계적이고 전문적인 비교를 하는 것은 아니다. 다만 비전문가 내지는 초보자의 입장에서 느낄 수 있는 차원의 비교를 하는데 국한한다.

2. 통계 소프트웨어

계량적 연구에서 연구자는 연구 대상으로부터 연구하고자 하는 특성을 나타낼 수 있는 데이터를 관측하게 된다. 이렇게 해서 얻어진 데이터에 담겨 있는 정보들을 효과적으로 요약하고 정리할 뿐 아니라 해석하기 위해서 다양한 통계적 방법들이 이용된다. 그런데 통계 이론이나 데이터 분석에 익숙하지 않은 연구자에게는, 얻어진 데이터를 효과적으로 정리하는 일이나 데이터로부터 연구목적에 맞는 통계를 계산하여 해석하는 일이 결코 쉬운 일이 아니다.

통계 소프트웨어는 간단히 말하자면 통계학에서 개발된 대부분의 통계공식들을 컴퓨터 언어를 이용하여 프로그램화한 것으로서 일종의 자동 계산기라고 할 수 있다. 자동차를 사용하기 위해 모든 운전자들이 자동차 구조나 원리들을 다 알아야 하는 것이 아니듯이 통계를 계산하기 위해 모든 사람들이 통계 이론을 배워야만 하는 것은 아니다. 일반인들이 단지 간단한 조작법만을 배워 쉽게 다양한 고급 통계들을 계산할 수 있도록 해주는 것이 바로 통계 소프트웨어이다.

통계 소프트웨어의 종류에는 여러 가지가 있지만 모든 소프트웨어들은 대체로 공통적인 기능을 가지고 있다. 통계 소프트웨어의 중요한 기능으로는 첫째, 데이터 처리(data processing) 기능을 들 수 있다. 즉, 데이터의 입력, 수정, 가공, 저장, 회수 등을 담당하는 기능이다. 둘째, 포괄적 통계 데이터분석 기능이다. 입력된 데이터에 대하여 다양한 그래프를 그리는 것, 거의 대부분의 공인된 통계적 분석기법을 처리하는 것 등을 담당하는 것이다. 최근 들어서 통계 소프트웨어들은 단순히 통계 데이터분석을 위한 도구로 그치는 것이 아니라 종합적인 해법(solution)을 제공하는 패키지로서까지 발전하였다. 하지만 이 글에서는 통계 데이터분석에 국한하여 논의를 진행하기로 한다.

우리나라에서 가장 널리 사용되는 대표적인 통계 소프트웨어로는 SPSS, SAS, Minitab, R 등이 있다. 가장 먼저 세계적으로 널리 보급된 것은 SPSS이다. SPSS는 애초에 사회과학 연구를 위해 개발된 통계 소프트웨어인데 오늘날에는 사회과학에 국한하지 않고 모든 분야에서 사용할 수 있도록 발전되었다. 이에 비해 SAS는 1966년 모든 분야의 통계 데이터분석을 목적으로 개발되었다. 한편 R은 최근에 개발된 소프트웨어로서 다른 소프트웨어들과는 달리 무료이며, 소스가 공개된 독특한 것으로서 급속하게 그 영향력을 높여가고 있는 소프트웨어이다.

통계 소프트웨어들은 PC가 발명되기 이전인 1960년대 말 대형 기종을 위한 프로그램으로 개발되었다. 그러나 1980년대 초에 PC가 나타나자 PC용 소프트웨어로 개발되어 1980년대 중반 이후 전 세계로 퍼지기 시작하였다. PC 하드웨어의 급속한 개발에 발맞추어 통계 소프트웨어들도 개정을 거듭하였는데 SPSS의 경우 현재 제13판, SAS의 경우 제9판의 버전이 출시된 상태이며, 최근에 개발된 R도 제2판이 나와 있다.

각각의 통계 소프트웨어들을 비교하기 위해서는 기능, 속도, 용량, 안정성, 편의성, 가격 등 여러 가지 점들을 검토할 수 있다. 하지만 이 글은 통계 소프트웨어를 전문적, 학술적 수준에서 비교하는 것이 목적이 아니다. 또한 오늘날 대부분의 통계 소프트웨어들은 일반적인 사용자들이 요구하는 수준 이상의 전문적인 고급 통계 분야의 이론까지 모두 다루고 있으므로 어느 소프트웨어를 쓰든 기능이나 안정성 등의 측면에서는 문제가 없다고 해도 과언이 아니다. 따라서 일반적인 사용자 입장에서는 어느 것이 더 사용하기 쉽고 편리하냐, 어느 것이 더 저렴하냐 하는 점이 더중요한 관심사가 될 수 있으므로 이런 점에 초점을 맞추어 설명하기로 한다.

3. 통계 소프트웨어의 소개

여기서는 우리나라에서 일반인들이 가장 널리 사용하는

표 1. 7세 어린이의 신체 측정 데이터

ID	성별	키	몸무게	ID	성별	키	몸무게
1	1	1237	23.6	21	2	1280	28.2
2	1	1238	25.1	22	2	1220	21
3	1	1255	35.2	23	2	1245	21.1
4	1	1226	21.5	24	2	1198	26.8
5	1	1324	33.8	25	2	1200	19.4
6	1	1231	25.1	26	2	1221	23.9
7	1	1237	27.9	27	2	1292	24.7
8	1	1132	21.5	28	2	1162	21.7
9	1	1208	24.4	29	2	1135	18
10	1	1257	27.3	30	2	1220	30.5
11	1	1278	26.1	31	2	1123	20.3
12	1	1143	20.5	2	2	1212	21.9
13	1	1228	23.1	33	2	1178	22.5
14	1	1183	20.8	34	2	1279	24.2
15	1	1202	25	35	2	1250	29.2
16	1	1184	20.6	36	2	1164	21.3
17	1	1220	25	37	2	1200	26.2
18	1	1120	18.2	38	2	1081	19.4
19	1	1185	21.7	39	2	1257	25.9
20	1	1176	25.3	40	2	1222	24.5

대표적인 통계 소프트웨어들을 간단히 소개하고, 각 소프트웨어의 특징을 쉽게 파악할 수 있도록 동일한 예제에 대해 각각의 소프트웨어로 실제 데이터분석 하는 방법을 소개하고자 한다. 각종 소프트웨어를 배우는 데 있어서 '백문(百聞)이 불여일타(不如一打)' 이다. 즉, 직접 자판을 두드리며 프로그램을 직접 해보는 것이 가장 빨리 배울 수 있는 지름길이라는 뜻이다. 따라서 구체적인 예제를 가지고 각 소프트웨어의 특징을 이해할 수 있도록 하겠다.

1) 예 제

다음의 [표 1]에는 우리나라 7세 남자와 여자 어린이 각각 20명씩의 키와 몸무게 데이터가 기록되어 있다. 이 데이터를 가지고 남녀 각각 키, 몸무게, 그리고 이 두 변수의 함수로 된 BMI지수의 분포를 파악하고, 중요한 기초통계량들을 계산하려 한다. 마지막으로는 남녀 간의 차이를 검토하려 한다.

분포를 파악하기 위해서는 히스토그램이나 줄기와 잎 그림, 상자그림 등을 그려 한 눈으로 파악할 수 있게 하는 것이 바람직하다.

중요한 기초통계량으로는 각 변수별로 평균, 표준편차, 최대값, 최소값 등을 생각할 수 있다. 더 나아가서 키와 몸무게, 그리고 BMI지수(body mass index)라는 변수들 상호간의 상관계수 등을 구해보는 것도 좋다.

마지막으로 남자와 여자의 차이를 검토하기 위해서는 각 변수별 평균의 차이를 검정해야 한다.

2) SPSS

(1) 개요

SPSS는 Statistical Package for the Social Sciences를 줄인 말로서 애초에 사회과학 분야의 데이터분석을 위해 개발된 통계 소프트웨어인데 오늘날에는 사회과학 분야에 국한하지 않고 다른 여러 분야의 통계 분석에도 널리 활용되고 있다. 1960년대 말에 정치학자인 Norman Nie에 의해 최초로 개발될 때는 대형 컴퓨터용으로 만들어졌는데 PC가 널리 보급된 이후 PC용 소프트웨어도 개발되어 현재 제13판이 출시되어 있다. 현재 우리나라 대부분의 사용자들은 제10판, 제11판, 제12판, 제13판을 사용하고 있다.

SPSS에서 나타나는 창(windows)에는 데이터 편집창, SPSS 뷰어창, 도표 편집창, 피벗표 편집창 등이 있다. 데이터 편집창은 분석하고자 하는 데이터를 입력, 가공, 편집하는 작업을 실행하는 창이며, 나머지 세 개의 창은 분석과 연관된 창이다. SPSS 뷰어창은 데이터를 이용해서 실행한 분석 결과가 나오는 창이며, 도표 편집창과 피벗표 편집창은 실행 결과 얻어진 각종 표나 그림을 편집할 수 있도록 하는 창이다. 따라서 SPSS 소프트웨어는 크게 데이터입력 부분과 데이터분석 부분으로 구분되어진다고 할 수 있다.

[그림 1]은 SPSS 편집창에 위 예제에 소개한 데이터를 입력한 그림이다. 데이터 입력창은 엑셀과 같은 스프레드 시트 양식으로 되어 있으며 [그림 1]의 아래에 표시한 탭 부분에 나온 바와 같이 'Data View'와 'Variable View' 두 종류의 시트로 구성되어 있다. 'Data View' 시트는 데이터를 입력하는 시트이며, 'Variable View'는 입력된 데이터의 각 변수들의 속성 등을 지정하는 시트이다. SPSS 편집창에 원 데이터를 직접 입력하는 경우 이 데이터 세트를 저장하면 sav라는 확장자가 붙는다. 한글 파일의 경우 hwp, 엑셀 파일의 경우 xls라는 확장자가 붙는 것과 마찬가지로 SPSS 데이터 파일의 경우 sav라는 확장자가 붙게 되는 것이다. 한편 엑셀에서 입력하여 저장하였거나 기타 txt 파일 형태로 저장된 외부 파일을 직접 불러 바로 SPSS 데이터 편집창에 불러 올릴 수 있으므로 매우 편리하다.

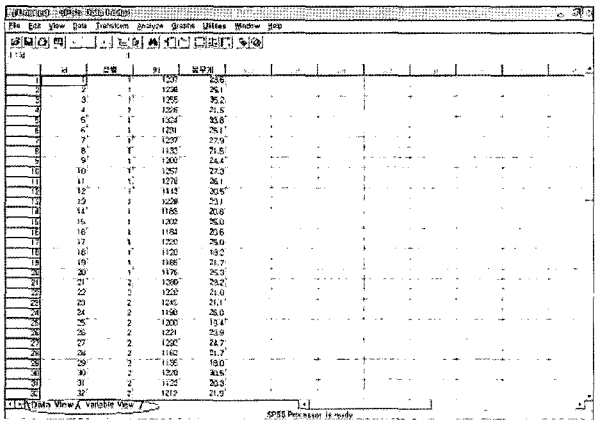


그림 1. SPSS 편집창 화면

분석하려는 데이터를 SPSS의 asv 파일로 저장하고 나면 다음으로 이 데이터에 대해 여러 가지 분석 작업을 할 수 있다. [그림 2]는 SPSS의 메뉴들을 나타낸 그림인데 동그라미로 표시한 'Data', '변환(T)', 'Analyze', 'Graphs' 메뉴들이 자료분석 과정에서 널리 사용되는 메뉴들이다. 'Data', 'Transform'은 데이터의 가공, 편집 등을 위해 활용되는 반면, 'Analyze', 'Graphs'는 통계적 분석 및 각종 그래프를 그리기 위해 사용된다. 자세한 사용방법은 아래의 예제 프로그래밍에서 보다 구체적으로 소개할 것이다.

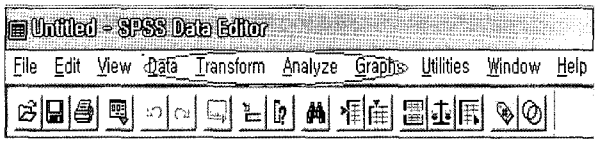


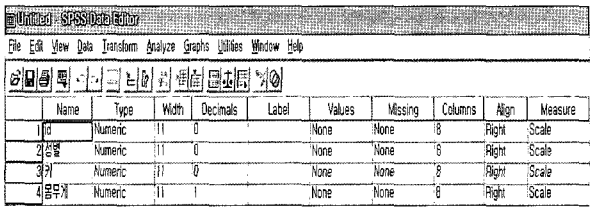
그림 2. SPSS 메뉴 창

(2) 예제 프로그래밍

[표 1]에 나온 7세 어린이 데이터를 SPSS 소프트웨어를 사용하여 구체적으로 분석해보기로 하자. 먼저 분석에 합당한 형태로 데이터 파일을 가공하는 작업을 한 후 가공된 데이터 파일에 대해 원하는 각종 데이터 분석을 실시해 보자.

① 데이터 편집

앞에서 소개한 [그림 1]은 [표 1]의 데이터를 SPSS 편집창에 입력한 그림인데 입력변수는 ID, 성별(남=1, 여=2), 키, 몸무게의 네 개 변수이다. 먼저 이 변수들의 속성을 정의해보자. 데이터 편집창의 'Variable View' 시트를 연 후 각 변수의 이름과 유형, 자리수, 소수점이하자리 등을 [그림 3]과 같이 정의하면 된다.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	11	0		None	None	8	Right	Scale
2	성별	Numeric	11	0		None	None	8	Right	Scale
3	키	Numeric	11	0		None	None	8	Right	Scale
4	몸무게	Numeric	11	1		None	None	8	Right	Scale

그림 3. 변수의 정의

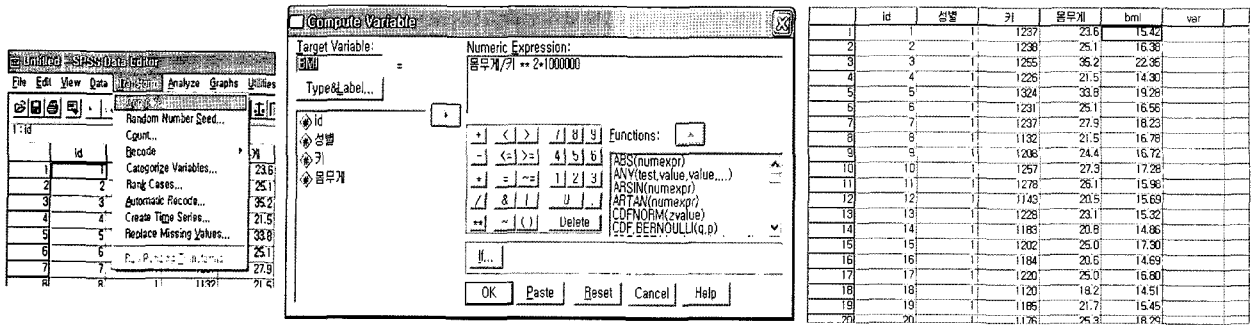
다음으로는 이 변수들을 이용하여 BMI지수라는 새로운 변수를 생성해야 한다. 참고로 BMI지수는 다음과 같이 정의되는 지수이다.

$$BMI = 몸무게(kg) / \{키(m) \times 키(m)\} \quad <식\ 1>$$

여기서 주의할 점은, [표 1]의 데이터에서는 키의 단위가 밀리미터(mm)인데 반해 [식 1]에서는 키가 미터(m) 단위라는 사실이다. BMI 지수라는 변수를 계산하기 위해서는 'Data View' 시트로 가서 'Transform' 메뉴를 선택하고 그 중 '변수계산(C)'을 클릭한다 (그림 4의 (a)). 그러면 [그림 4]의 (b)와 같은 대화상자가 나타나는데 '대상변수'라는 칸에 BMI라는 변수 이름을 기입하고 '숫자표현식(E)' 칸에는 [식 1]의 함수를 기입한 후 '확인'을 클릭하면 된다. 데이터에서 키가 밀리미터 단위로 입력되어 있으므로 이것을 미터 단위로 고쳐주기 위해 숫자표현식에 1,000,000 이라는 수를 추가로 곱해주었다. 그러고 나면 [그림 4]의 (c)와 같이 데이터 편집창에 BMI라는 새로운 변수가 생성되어 나타난다.

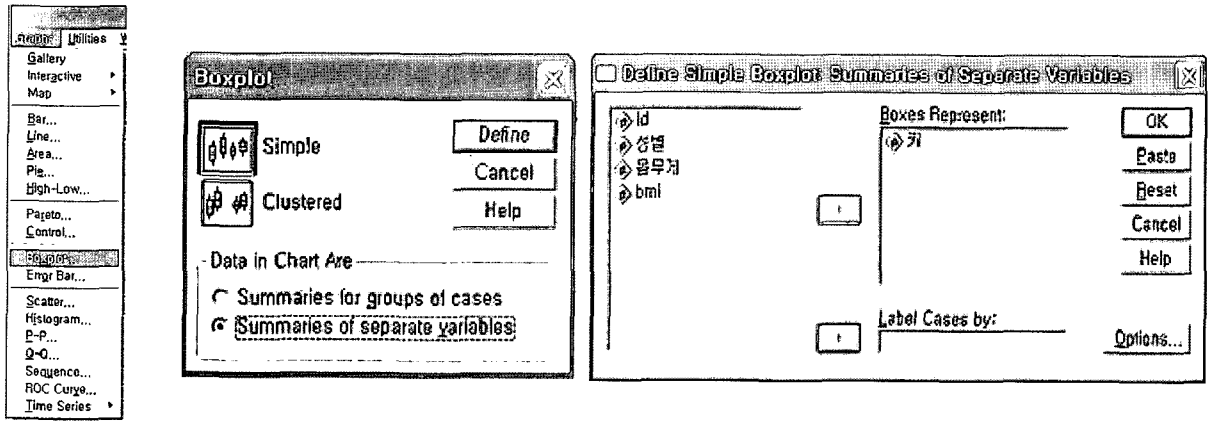
② 그림 그리기

이제 [그림 4] (c)의 데이터를 가지고 통계적 분석을 실시해보자. 먼저 키의 상자그림을 그려보자. 'Graphs' 메뉴



(a) 변수변환 메뉴 (b) 새로운 변수생성을 위한 대화상자 (c) 편집창에 새로운 BMI 변수가 생성된 화면

그림 4. 변수관련 메뉴 및 대화상자



(a) Graph 메뉴 (b) 상자그림 첫 대화상자 (c) 상자그림 두 번째 대화상자

그림 5. 상자그림을 위한 메뉴와 대화상자

를 클릭하면 [그림 5] (a)에서와 같이 SPSS에서 제공하는 다양한 그래프 종류들이 열거되는데 그 중 'Boxplot'을 선택하면 먼저 [그림 5] (b)의 대화상자가 나타난다. 성별로 별도로 그림을 그리지 않고 모든 데이터를 합쳐서 하나의 그림을 그리는 경우 'Summaries of separate variables'를 선택한 후 'Define'을 클릭하면 다시 [그림 5]의 (c) 대화상자가 나타난다. 'Boxes Represent'라는 칸에 키라는 변수를 선택해주면 된다. [그림 6]은 이렇게 키에 대한 상자

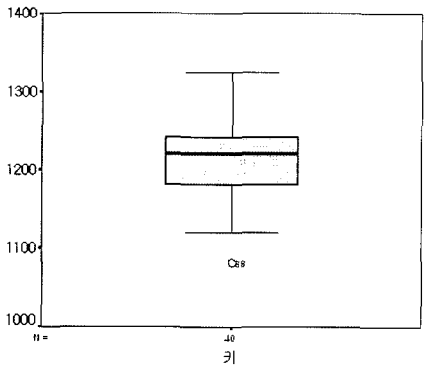


그림 6. 키에 대한 상자그림

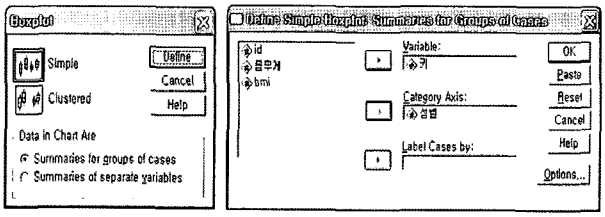


그림 7. 그룹별 상자그림을 위한 대화상자

그림 화면이다.

위에서는 남녀 구분 없이 하나의 상자그림으로 나타내었는데 이번에는 성별로 각각 상자그림을 그려보자. 이 경우에는 [그림 7]의 왼쪽과 같이 Boxplot 메뉴를 선택 후 처음으로 나타나는 대화상자에서 'Summaries for groups of cases'를 선택해야 한다. 그러면 다시 오른쪽과 같은 대화상자가 나타나는데 'Variable'은 키, 'Category Axis'는 성별을 선택하면 된다. [그림 8]은 성별 상자그림을 그린 결과이다. 위와 비슷한 과정을 반복하면 몸무게와 BMI지수에 대한 상자그림도 그릴 수 있다.

③ 기초통계량 계산

이번에는 SPSS를 이용하여 각 변수들의 평균, 표준편

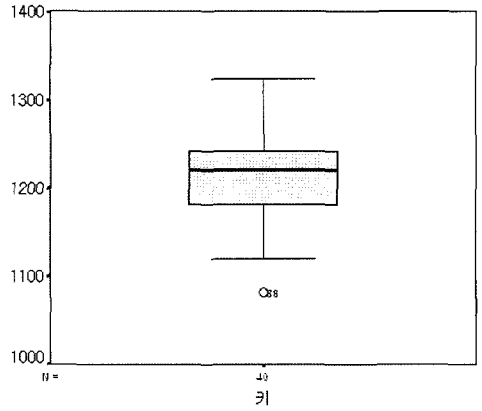
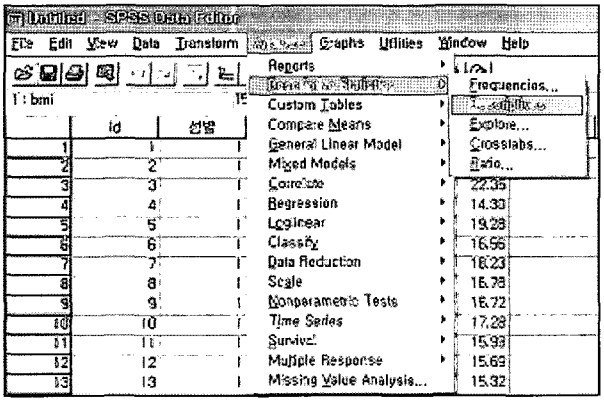
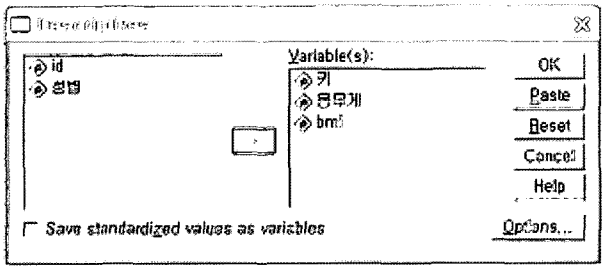


그림 8. 성별 키의 상자그림



(a)



(b)

그림 9. Analyze 메뉴상자와 Descriptives 대화상자

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
키	40	1081	1324	1210,08	51,802
몸무게	40	18,0	35,2	24,060	3,8472
BMI	40	13,47	22,35	16,3668	1,86949
Valid N (listwise)	40				

그림 10. 키, 몸무게, BMI 변수들의 기초통계량 계산 결과

차, 최대값, 최소값 등과 같은 기초통계량을 계산해보자. [그림 9]의 (a)에서 보듯이 'Analyze' 메뉴에서 'Descriptive Statistics'를 택하고 이어 'Descriptives'를 선택하면 [그림 9]의 (b)와 같은 대화상자가 나타난다. 이 대화상자

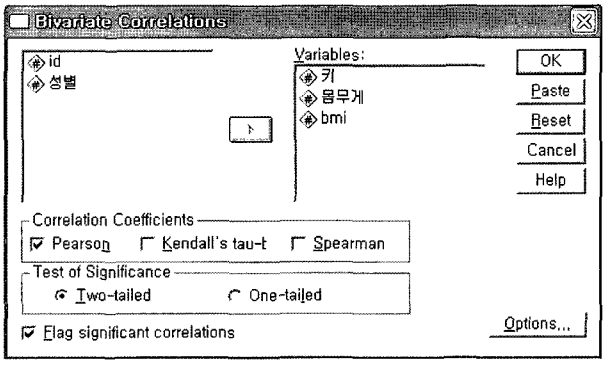


그림 11. 상관계수 계산을 위한 대화상자

Correlations

		키	몸무게	BMI
키	Pearson Correlation	1	,693**	,225
	Sig. (2-tailed)	.	,000	,163
	N	40	40	40
몸무게	Pearson Correlation	,693**	1	,856**
	Sig. (2-tailed)	,000	.	,000
	N	40	40	40
BMI	Pearson Correlation	,225	,856**	1
	Sig. (2-tailed)	,163	,000	.
	N	40	40	40

**. Correlation is significant at the 0.01 level (2-tailed).

그림 12. 세 변수들 상호 간의 상관계수 행렬

의 'Variables' 칸에 통계량을 구하고자 하는 변수들을 모두 선택하여 나타낸 후 실행을 시키면 [그림 10]과 같은 결과가 나타난다.

④ 상관계수 계산

키, 몸무게, BMI 지수 등 세 변수들 사이의 상관계수의 계산은 'Analyze' 메뉴 중 'Correlate'를 선택하고 이어서 'Bivariate'을 선택하여 할 수 있다. [그림 11]과 같은 대화상자가 나타나면 'Variables'에 키, 몸무게, bmi 세 변수를 선택한 후 를 클릭하면 [그림 12]의 결과가 나온다.

⑤ 성별 평균 차이 검정

마지막으로 각각의 변수들의 평균이 성별로 서로 차이가 나는지를 검정해보자. 이 경우는 독립표본에 의한 t-검정의 문제가 된다. 따라서 [그림 13]의 (a)와 같이 'Analyze' 메뉴 중 'Compare Means'를 선택한 후 'Independent Sample T Test'를 클릭한다. 그러면 (b)의 대화상자가 나오는데 'Test Variables'에는 검정하고자 하는 변수들인 키, 몸무게, bmi를 기입한다. 한편 'Grouping Variable'에는 성별을 선택해주고 다시 그 아래에 있는 단추를 누른 후 Group 1과 2에 각각 그 변수값인 1과 2를 선택해 주면 된다.

다음의 [그림 14]는 세 변수에 대한 성별 평균 차이에 대한 t-검정 결과이다. 이 분석결과에 의하면 세 변수 모두

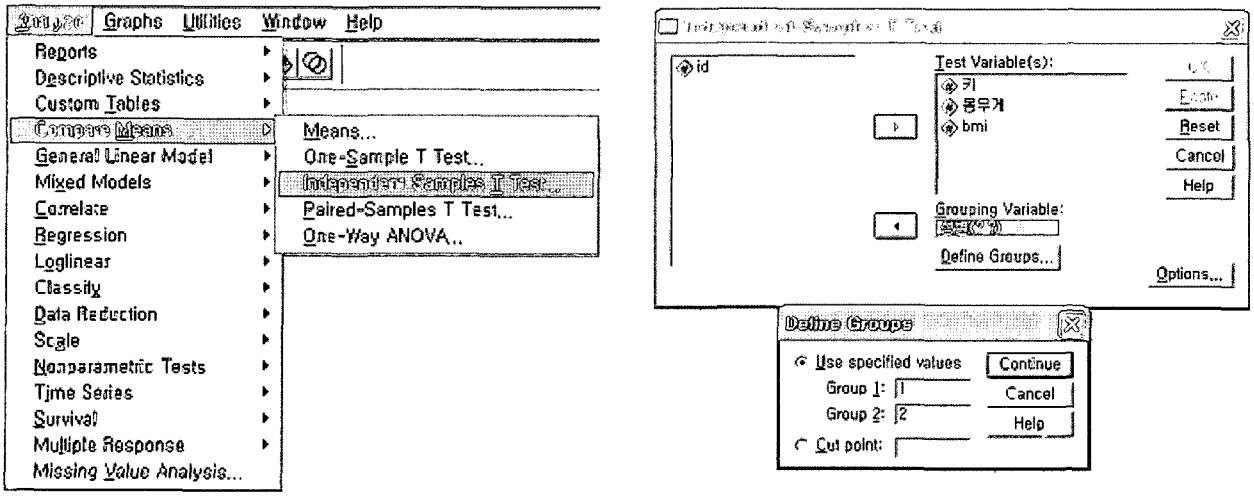


그림 13. 독립표본 t-검정 메뉴 및 대화상자

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
키	Equal variances assumed	.125	.726	.377	38	.708	6.25	16.564	-27.293	39.793
	Equal variances not assumed			.377	37.627	.708	6.25	16.564	-27.284	39.784
몸무게	Equal variances assumed	.025	.876	.360	38	.385	1.050	1.2207	-1.4211	3.5211
	Equal variances not assumed			.360	36.620	.385	1.050	1.2207	-1.4242	3.5242
BMI	Equal variances assumed	.008	.854	.821	38	.417	.4876	.59367	-.71421	1.69842
	Equal variances not assumed			.821	37.874	.417	.4876	.59367	-.71424	1.69845

그림 14. 세 변수에 대한 성별 평균에 대한 차이 검정 결과

에 대해 성별로 평균값은 유의한 차이를 나타내보이지 않는다는 결론을 내릴 수 있다.

(3) 참고사항

SPSS는 오늘날 세계에서 가장 널리 사용되는 통계 소프트웨어 중의 하나이다. 특히 사회과학 분야에서는 독보적인 소프트웨어라고 할 수 있는데 이렇게 널리 사용되는 이유는, 첫째 사용하기 쉽고 간편하다는 점을 들 수 있다. 특히 비전문가들도 어렵지 않게 배워 쓸 수 있다는 장점이 있다. 둘째로는 널리 쓰이는 통계 소프트웨어 중 가장 오랜 역사를 지니고 있어 많은 사용자들을 가지고 있다는 점이다. 통계 소프트웨어들 중 메뉴를 클릭하여 사용하는 방식으로는 거의 표준이라고 해도 과언이 아닐 것이다. Minitab이라는 소프트웨어도 널리 사용되는 통계 소프트웨어인데 외형적으로 보기에 SPSS와 큰 차이가 없다. 국내 최초로 개발된 한글 통계 소프트웨어인 S-Link도 SPSS와 유사한 체계를 지닌 소프트웨어이다. 즉, SPSS 사용법만 잘 알면 다른 유사한 통계 소프트웨어도 쉽게 활용할 수 있는 것이다.

그러나 SPSS와 같이 메뉴를 클릭하는 방식의 소프트웨어가 가지는 약점으로 지적되는 것은 원 자료에 대한 통계적 출력 결과를 제어하기가 어렵다는 점이다. 가령, 원자료의 평균(χ)과 표준편차(σ)를 계산한 후 각각의 데이터가 $\chi \pm 3\sigma$범위 내에 있는지를 SPSS를 이용하여 파악하려 할 때 매우 불편함을 알 수 있다. 또한 SPSS는 사회과학에서 널리 사용되는 통계적 방법들에 대해서는 상대적으로 강점을 지니는데 비해 의학, 보건학이나 품질관리 분야의 통계적 방법들에서는 약점을 보였는데 지속적인 업그레이드를 통해 이런 문제들을 많이 해결해가고 있다.

SPSS와 관련하여 참고할 만한 문헌으로는 서울대학교 통계학과(2005), 이학식 외(2005), 허문열 외(2003), 정영해 외(2003) 등을 들 수 있다.

3) SAS

(1) 개 요

SAS란 Statistical Analysis System의 약어로서 컴퓨터를 통계적 분석을 위해 미국 North Carolina 주의 SAS Institute에서 개발된 소프트웨어이다. 1976년 회사가 설립되어 상업적으로 판매되기 시작했는데 현재 제9판까지 개발되어 시판되고 있으며 현재 우리나라에서는 제6판, 제8판, 제9판이 주로 사용되고 있다.

SAS의 주요기능으로는 정보의 저장, 회수, 데이터의 수정과 프로그래밍, 포괄적 통계자료의 분석, 파일의 관리 및 조작 등을 들 수 있다. SAS가 SPSS와 가장 크게 다른 점은 메뉴를 클릭하여 분석을 시행하는 방식이 아니고 직접 프로그램 문장을 작성하여 프로그램을 시행하는 방식을 취한다는 점이다.

SAS의 시작화면이 [그림 15]에 나와 있는데 프로그램 창, 로그 창, 출력 창의 세 가지로 구성되어 있다. 프로그램 창은 분석을 위한 프로그램을 작성하는 창이다. 프로그램 창에 프로그램을 다 작성한 후 프로그램을 실행시키면 각 명령의 수행상태를 설명하는 문장들이 로그 창에 저절로 나타난다. 만일 프로그램 상 오류가 발생하면 그 오류 메시지가 로그 창에 나타나게 된다. 마지막으로 출력 창은 프로그램의 수행 결과가 기록되는 창으로서 출력 창에 나타나는 출력 결과를 이용하여 분석을 할 수 있다.

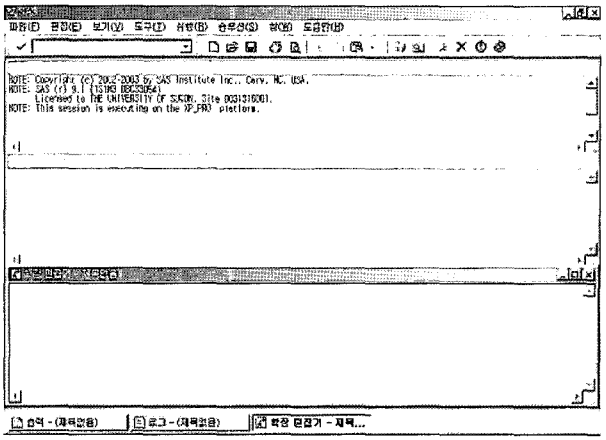


그림 15. SAS의 시작화면

SAS 프로그램은 크게 Data 단계와 Proc 단계로 구분된다. Data 단계는 데이터를 입력하고 편집, 가공하는 단계로서 SPSS에서 스프레드 시트형의 데이터 편집창에서 행하는 작업들을 수행하는 단계이다. Proc 단계는 Data 단계에서 만들어진 데이터 세트에 대해 각종 통계 분석을 수행하는 단계이다.

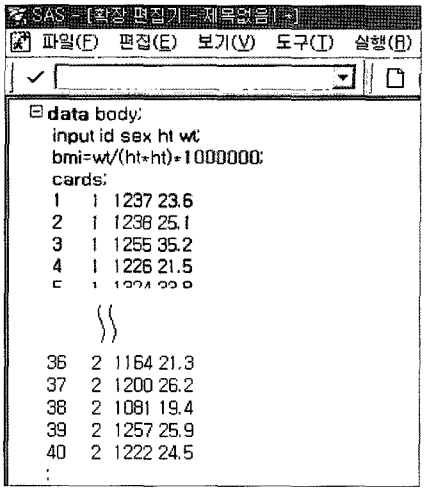


그림 16. 데이터 입력

(2) 예제 프로그래밍

① 데이터 입력과 편집

[표 1]의 데이터를 SAS 데이터 세트로 만드는 프로그램이 [그림 16]에 나와 있다. 이 프로그램에서 input 문은 입력할 변수들을 나타내는 명령인데 id, sex, ht, wt 라는 이름의 변수명으로 다음에 나올 숫자를 기억하라는 뜻이다. 한편 input 명령에 있는 변수 ht와 wt를 이용하여 새로운 변수 bmi를 생성하라는 명령이 그 아래 문장에 나와 있다. 'cards;'라는 문장 이후로는 데이터를 입력하여야 한다. 데이터를 입력할 때 변수들 사이에는 빈 칸을 두어야한다. 40개의 데이터를 다 입력한 후 그 다음 행에 세미콜론(;)을 찍어 데이터 입력이 끝났음을 표시한다. 이 명령을 시행시키면 SAS 내부에서는 body라는 이름의 데이터 세트를 생성시키게 되는데 이 데이터 세트에는 id, sex, ht, wt, bmi 라는 이름의 다섯 개 변수가 저장된다.

② 그림 그리기

[그림 16]에서 소개한 body라는 SAS 데이터 세트를 가지고 키, 몸무게, BMI지수 각각의 상자그림을 그려보자. 프로그램 창에서 이미 작성한 [그림 16]의 명령 아래에 다음과 같은 명령을 추가하여 실행하면 상자그림이 각 변수별로 그려진다.

```
proc boxplot data=body;
  plot wt*sex;
  plot ht*sex;
  plot bmi*sex;
run;
```

③ 기초통계량 계산

변수들의 기초통계량을 계산하는 명령은 proc means 라는 명령이다. body라는 데이터 세트를 가지고 wt, ht, bmi 세 변수의 기초통계량을 계산하는 프로그램은 다음과 같다.

```
proc means data=body;
 var wt ht bmi;
run;
```

한편 성별로 각각 기초통계량을 계산하려면 proc sort 라는 명령을 사용하여 데이터를 정렬한 후 sex라는 변수에 대해 각각 proc means를 적용시키면 되는데 그 프로그램은 다음과 같다. 이후 모든 SAS 프로그램의 실행 결과는 SPSS의 출력결과와 유사하므로 지면 관계상 생략하기로 한다.

```
proc sort data=body;
 by sex;
run;
proc means data=body; by sex;
 var wt ht bmi;
run;
```

④ 상관계수 계산

여러 변수들 상호간의 상관계수를 계산하려면 proc corr 라는 명령을 사용하는데 세 변수의 상관계수를 구하는 프로그램은 다음과 같다.

```
proc corr data=body;
 var wt ht bmi;
run;
```

⑤ 성별 평균 차이 검정

마지막으로 성별로 키, 몸무게, BMI 지수의 평균이 유의한 차이가 나는지를 검정하려면 proc ttest 명령을 사용한다. 여기서는 비교하려는 그룹을 지정해주기 위해 class 라는 명령을 사용하며 작성된 프로그램은 다음과 같다.

```
proc ttest data=body;
 class sex;
 var ht;
run;
```

실제 SAS 프로그램을 할 때에는 위의 모든 단계들을 별개로 실행시키는 것이 아니라 모든 과정을 한 프로그램으로 만들어서 한번에 동시에 수행하는 것이 일반적이다.

(3) 참고사항

SAS를 이용한 예제 프로그래밍을 통해 확실히 느꼈을 것으로 생각하는데, SAS는 SPSS와는 달리 사용자가 직접 명령문을 직접 작성하여 실행시켜야 한다. 따라서 SAS를 이용하려면 기본적인 SAS 명령문이나 문법들을 어느 정도 알고 있어야한다. SPSS를 위시한 모든 통계 소프트웨어들이 초기에는 모두 SAS와 같이 사용자가 명령문을 직접 작성하는 방식으로 만들어졌다. 그러다가 여러 차례 개정이 이루어지는 과정에서 대부분의 소프트웨어들은 사용자 편리를 위한 메뉴 방식으로 변해갔는데 SAS는 여전히 초기의 방식을 고수하고 있다.

메뉴 방식과 직접 프로그램을 작성하는 방식은 상대적으로 장, 단점을 지니고 있다. 초보자나 비전문가들의 입장에서는 메뉴 방식의 SPSS나 Minitab이 쉽고 편리한 반면, 통계 소프트웨어를 전문적으로 사용하는 전문가 입장에서는 오히려 프로그램을 직접 작성하는 SAS 방식이 더 편리할 수 있기 때문이다.

SAS와 관련하여 추천할 만한 참고문헌으로는 조인호(2004), 성내경(2004), 김기영 외(2003), 송문섭 외(2002) 등이 있다. 한편 인터넷의 다음 까페 같은 곳을 방문하면 SAS 사용자 그룹이 활발하게 활동하고 있으므로 여러 가지 실제적인 도움을 얻을 수 있다.

4) R

(1) 개 요

R이라는 통계 소프트웨어는 SPSS나 SAS 등의 다른 범용 소프트웨어에 비해 매우 색다른 소프트웨어이다. 다른 소프트웨어들이 모두 1960년대 또는 70년대에 개발되어 상업적으로 널리 판매되어온 소프트웨어인데 비해 R은 1990년대 들어서 개발되었으며 무료로 웹을 통해 누구든지 내려 받을 수 있는 소프트웨어라는 점이다. 미국의 벨 연구소 연구원들에 의해 통계적 분석을 위한 새로운 언어인 S라는 언어가 개발된 바 있다. S라는 언어를 사용하여 만든 통계 소프트웨어로 S-plus와 R이 있는데, S-plus는 다른 통계 소프트웨어와 마찬가지로 상업적으로 판매되고 있지만 R은 무료이며 그 소스가 공개되는 독특한 프로그램이다. R의 웹 주소는 www.cran.r-project.org 인데 거기에 접속하여 누구든 쉽게 프로그램을 다운 받을 수 있다.

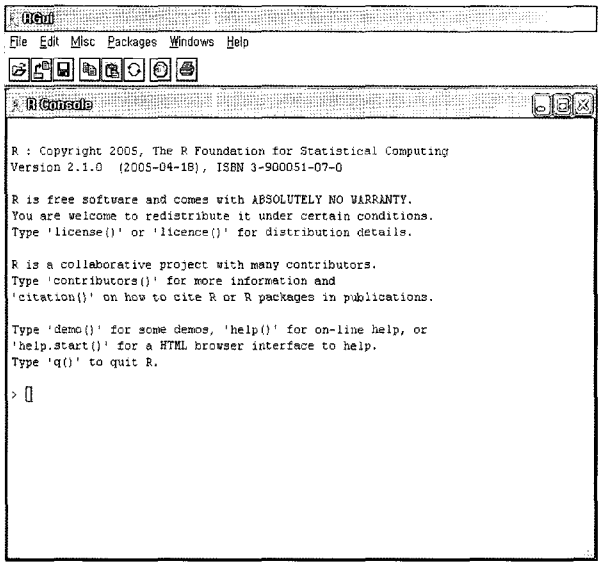


그림 17. R의 시작화면

R은 SAS와 마찬가지로 사용자가 프로그램을 직접 입력하여 실행하는 방식의 소프트웨어인데 SAS와 크게 다른 점은 대화형 프로그램으로서 하나의 명령을 입력하면 바로 그에 따른 결과가 화면에 나타난다. [그림 17]은 R의 시작화면을 나타낸다. 그림을 그리는 명령을 내릴 경우 별도의 그림 창이 뜨게 되지만, 결과가 텍스트인 경우 바로 R의 프로그램 화면에 실행결과가 나타나게 된다.

R은 무료 소프트웨어이므로 입출력에 대한 사용자 편리성은 다른 소프트웨어에 비해 떨어지지만 본질적인 계산 능력이나 프로그램 수행 능력은 결코 뒤지지 않으며 오히려 전문적인 연구자의 측면에서 볼 때 더욱 우수한 면이 많다. PC 운영체제인 윈도우즈의 상업성에 대항하여 리눅스라는 무료 공개 프로그램이 개발된 것과 마찬가지로 R은 기존 통계 소프트웨어들의 상업성에 대항하여 나타난 무료 공개 프로그램이라고 할 수 있다.

R에서는 모든 데이터를 하나의 객체(object)로 인식한다. 사용자가 객체의 이름을 붙여 저장한 후 그 이름의 객체를 가지고 각양의 통계함수들을 사용하여 통계적 분석을 내린다는 면에서 SAS와 유사성을 지닌다고 할 수 있다. R의 웹 사이트에 접속하면 R 사용법에 대한 각종 문서들도 무료로 다운 받을 수 있으므로 이를 이용하여 간단한 사용법을 배울 수 있을 것이다.

(2) 예제 프로그래밍

① 데이터 입력과 편집

R에서 데이터를 입력하는 방법은 여러 가지가 있는데 여기서는 가장 단순한 형태로 입력하는 방식을 소개하기로 하겠다. [표 1]의 데이터는 네 개의 변수를 지니므로 먼저 각각의 변수들을 하나의 객체로 하여 다음과 같이 입력한다. 데이터 입력을 위한 R의 함수는 'c( )'이다. BMI 지수를 나타내는 변수도 바로 만들 수 있다. 제일 마지막 행은 다섯 개 개별 변수를 body라는 하나의 행렬로 묶는 것을 나타낸다.

```
sex=c(1,1,1,1, ……, 2, 2)
ht=c(1237,1238, ……, 1257,1222)
wt=c(23.6, 25.1, ……, 25.9, 24.5)
bmi=wt/ht^2 * 1000000
body=data.frame(id,sex,ht,wt,bmi)
```

② 그림 그리기

키에 대한 상자그림을 R을 이용하여 그리는 명령문은 다음과 같다. 아래의 명령문은 각각의 변수에 대하여 성별 상자그림을 그리라는 명령이다. 같은 요령으로 몸무게나 BMI지수에 대한 상자그림도 그릴 수 있다.

```
boxplot(ht)
boxplot(ht~sex)
```

③ 기초통계량 계산

변수들의 평균은 mean( ), 표준편차는 sd( ), 다섯수치 요약값은 summary( ) 함수를 사용하여 구할 수 있다. 각각의 변수에 대해 구할 수도 있지만 이 경우에는 여러 변수를 합친 객체인 body 만을 이용하여 mean(body)라는 명령을 내리면 body라는 객체에 포함된 sex, ht, wt, bmi 등 모든 변수들의 평균을 한꺼번에 구해준다. [그림 18]은

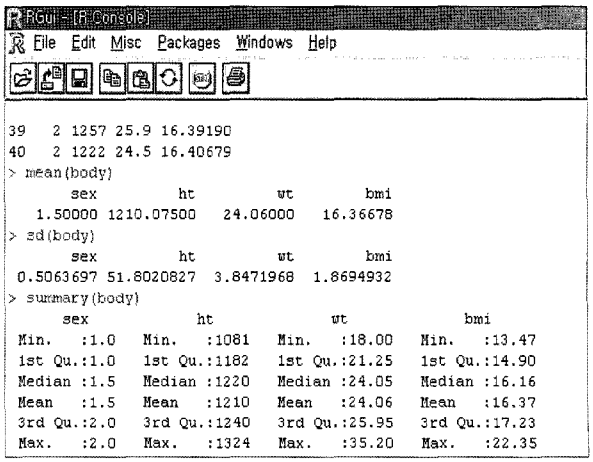


그림 18. R 프로그램으로 구한 기초통계량

```
> t.test(ht~sex)

        Welch Two Sample t-test

data:  ht by sex
t = 0.3773, df = 37.627, p-value = 0.708
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -27.29380  39.79380
sample estimates:
mean in group 1 mean in group 2
        1213.20         1206.95
```

그림 19. R의 t-검정 결과화면

R에서 각 함수와 그 출력결과가 나와 있다.

④ 상관계수 계산

여러 변수들 상호간의 상관계수를 계산하는 함수는 cor( )인데 cor(body)하면 body라는 객체에 있는 모든 변수들 간의 상관계수 행렬을 계산하여 나타내준다.

⑤ 성별 평균 차이 검정

마지막으로 성별로 변수들의 평균이 유의한 차이가 나는지를 검정하려면 t.test( )라는 함수를 사용하면 된다. 다음의 [그림 19]는 성별로 평균키가 같은지에 대한 t-검정 함수 및 그 결과 화면이다. 결과화면에는 t-검정에 대한 p값이 나올 뿐 아니라 두 평균의 차이의 95% 신뢰구간도 나와 있다.

(3) 참고사항

R은 사용자가 직접 프로그램을 짜야 한다는 면에서는 SPSS 보다는 상대적으로 SAS와 비슷하다. 하지만 SAS와는 달리 데이터를 하나의 객체로 처리한다는 점, 아울러 많은 변수를 가진 큰 데이터 세트도 하나의 객체로 처리할 수 있다는 점 등은 R 프로그램이 지닌 장점이다. 그러나 무엇보다도 중요한 장점은 소프트웨어가 무료라는 사실이다. 대부분의 통계 소프트웨어들은 개인이 1년 임대하는데 몇 십만 원에서 몇 백만 원이 소요된다. 그런 까닭에 개인들이 웬만해서는 통계 소프트웨어를 소유하기가 힘든 실정이다. 이런 점을 생각한다면 R처럼 막강한 기능을 수행하는 소프트웨어를 무료로 사용할 수 있다는 것은 아주 큰 장점이라고 아니할 수 없다.

R은 최근에 개발된 소프트웨어이며 상대적으로 국내에 알려지기 시작한 것은 불과 최근 몇 년 전의 일이다. 따라서 아직까지 한글로 된 참고서적이 발간되지는 않았다. 외국서적으로 추천할 참고문헌으로는 등이 있다. R의 웹 싸이트에 가면 좋은 참고 파일들이 많이 올라와 있으므로 이것을 활용하면 매우 유익하다.

4. 맺 음 말

본 글에서는 우리나라 뿐 아니라 세계적으로도 가장 널리 쓰이는 대표적인 통계 소프트웨어들인 SPSS, SAS, 그리고 R을 간단하게 비교해보았다. 각각의 소프트웨어의 특징들을 보다 효과적으로 느낄 수 있도록 하기 위해 구체적인 예제를 가지고 직접 분석하는 방법을 소개하는 방식으로 설명하였다.

SPSS (Minitab, S-link)는 메뉴를 클릭하는 방식으로, SAS와 R은 직접 프로그램을 짜는 방식으로 프로그램을 수행한다는 특징이 있었다. 통계 소프트웨어의 사용에 익숙지 않은 비전문가들의 입장에서 볼 때 상대적으로 SPSS가 사용하기에 편리하다. 또한 통계분석을 아주 빈번하게 수행하지 않고 잊을 만하면 한번씩 사용하는 사용자의 경우 굳이 복잡한 명령들을 다 외울 필요가 없는 SPSS를 사용하는 것이 효과적이다. 반면에 아주 빈번하게 통계분석을 하게 되는 사용자라면 처음 접근은 다소 어려울지라도 SAS나 R을 사용하는 것이 결국에는 더욱 편리하다고 할 수 있다.

다른 측면에서 SPSS나 SAS는 적지 않은 비용을 지불하고 구입해야 하는 소프트웨어인데 반해 R은 무료로 다운 받을 수 있는 프로그램이다. 대규모 통계분석을 항상 하는 사용자라면 몰라도 일년에 몇 번씩 드문드문 통계분석을 하는 사용자인 경우에는 그것을 위해 비싼 소프트웨어를 구입하는 것이 매우 비생산적이라고 할 수 있다. 이런 경우에는 R을 사용하는 것이 바람직하다고 할 것이다.

다양하고 편리한 통계 소프트웨어들이 많이 개발되어 보급된 오늘날에 있어서 통계학이란 더 이상 복잡한 공식놀음이 아니다. 통계학자가 아닌 비전공 통계 이용자의 입장에서 이제는 복잡한 통계 공식을 기억하는 것보다는 통계 소프트웨어를 통해 쏟아져 나오는 각종 수치들의 의미가 무엇인가를 이해하는 것이 더 필요한 시대라고 할 수 있다. 아무쪼록 통계 이용자들이 통계 소프트웨어라는 도구를 잘 활용하여 데이터로부터 수월하게 다양한 통계정보들을 뽑아낼 수 있기를 바란다.

참 고 문 헌

김기영, 강형철, 최병진. (2003). *SAS 입문 및 기초 프로그래밍*. 자유아카데미.

서울대학교 통계학과. (2005). *SPSS 통계분석*. 자유아카데미.

성내경. (2004). *SAS 시스템과 SAS 언어*. 자유아카데미.
송문섭, 조신섭. (2002). *SAS를 이용한 통계자료분석*. 자유아카데미.
이학식, 임지훈. (2005). *SPSS 12.0 매뉴얼*. 법문사.
정영해, 김순흥, 양철호, 조지현. (2003). *SPSS 10.0 통계자료분석*. 광주사회조사연구소.
조인호. (2004). *SAS 강좌와 통계 콘설팅*. 영진.COM.
허문열, 박종선. (2003). *논문작성을 위한 SPSS COOKBOOK*. 교우사.
Crawley, M. J. (2005). *Statistics: An Introduction using R*, *Wiley*.
Verzani, J. (2004). *Using R for Introductory Statistics*. Chapman & Hall.
www.cran.r-project.org

박 진 우

서울대학교 계산통계학과 (학사, 석사, 박사)
한국통계학회 조사통계연구회 부회장
현재 수원대학교 통계정보학과 부교수