응용통계연구 제 7권 2호
1994년 9월 pp. 279-288

한국형 통계패키지 개발 연구[1)]

이정진[2)], 강근석[3)]

요약

현재 국내에서 많이 사용되고 있는 통계패키지는 SAS, SPSS 등 모두 외국산 패키지이다. 영어로 된 이 외국산 패키지들을 통계전문가가 아닌 일반인들이 배우기는 매우 힘이 든다. 그리고 대부분의 일반인들은 이 외국산 통계패키지를 비싸게 임대 또는 구입해서 그림을 그리거나 간단한 테이블을 만드는 극히 제한된 작업만 하고 있다. 본 연구에서는 이러한 일반인들을 위한 한국형 통계패키지의 개발과정을 소개하여 새 패키지 개발에 관심있는 사람들에게 도움을 주고자 한다.

1. 서론

금세기 들어서 급속도로 발전한 컴퓨터과학은 우리들이 사는 세계를 점점 더 복잡하고 다양해지는 정보화 사회를 이루어 내고 있다. 이러한 정보화 사회에서는 어떻게 다량의 정보를 유효 적절하게 만들고 또 사용하느냐에 따라 각 개인, 단체, 기업, 나아가 국가의 성패가 달려있다. 그래서 다량의 정보를 신속 정확하게 처리하여 분석하는 통계패키지의 사용은 최근 급성장하게 되어 요즈음은 국가 공공기관이나 대학, 연구소, 대기업, 그리고 웬만한 중소기업에서도 통계패키지를 사용하게 되었다. 현재 국내에서 많이 사용되고 있는 통계패키지들은 세계적으로 유명한 SAS, SPSS, BMDP, MINITAB 등과 같은 제품들이다. 이 제품들은 수 십년간에 걸친 연구개발과 운용경험이 축적되어 나온 우수한 제품들임에 틀림이 없다. 하지만 영어를 쓰지 않는 일반 한국인이 외국산 패키지를 사용하려면 간단한 표를 하나 작성하려고 하여도 매우 힘든 일이다. 또 기능이 매우 다양하고 전문적인 기법까지 다루고 있는 이들 패키지를 구입하여도 실제적으로 대부분의 일반 사용자들은 그림이나 테이블을 만드는 간단한 목적의 자료처리에만 이용하고 있다. 이러한 통계패키지의 구입과 임대에 소비되는 외화가 현재 적어도 년간 몇 십억원이 되고, 이 금액은 앞으로도 급증되리라 예상된다. 이에 대한 책임은 외국산 패키지를 대체할만한 국산 통계패키지를 아직까지 개발하지 못한 한국의 통계인들에 있다고 생각한다.

본 연구의 목적은 당장에 SAS나 SPSS와 같은 훌륭한 패키지를 만드는 것은 불가능하지만 우선 일반인을 대상으로 하는 개인용 컴퓨터를 위한 한국형 통계패키지 개발을 하려는 것이다. 구체적으로, 일반인들이 자료처리 분석시에 가장 많이 사용하는 통계분석기법을 포함하는 패키지 개발을 여기서 연구하였다. 본 연구에서는 이러한 일반인들을 위한 한국형 통계패키지의 개발과정을 소개하여 새 패키지 개발에 관심있는 사람들에게 도움을 주고자 한다.

2절에서는 통계패키지 개발방향 및 전략을 소개하고, 3절은 개발과정의 중요부분을 하나 하

1) 이 논문은 1992년도 교육부지원 한국학술진흥재단의 자유공모과제 학술연구 조성비에 의해 연구되었음.
2) (156-743) 서울특별시 동작구 상도동 1-1, 숭실대학교 통계학과.
3) (156-743) 서울특별시 동작구 상도동 1-1, 숭실대학교 통계학과.

나씩 설명하였고, 4절에서는 패키지의 특징있는 결과출력 예를 살펴본 후, 5절에서 결론및 향후 과제에 관한 제안을 하였다.

2. 통계패키지 개발방향 및 전략

가. 개발방향

통계패키지 개발의 장기적인 목적은 세계시장에서 SAS, SPSS 등과 경쟁할만한 패키지를 만드는 것이다. 하지만 이 제품들은 수 십년간에 걸친 연구개발과 운용경험이 축적되어 나온 우수한 제품들이기 때문에, 당장에 모든 분야에서 이들보다 우수한 제품을 만드는 것은 거의 불가능한 일이다. 그래서 한국형 통계패키지의 개발방향은 1차적으로 우리나라의 통계전문가가 아닌 일반인을 대상으로하는 패키지를 개발하여 보기로 하였다. '어떠한 패키지를 만들어야 일반인들이 SAS나 SPSS 등보다 우수하다고 하겠는가?' 라는 구체적인 방향은 다음과 같이 설정하였다.

-- 일반인이 많이 필요로 하는 통계분석기법은 SAS/BASIC나 SPSS/BASE에서 다루어지는 분석기법이면 충분하다. 즉, 그림을 이용한 자료정리, 도수분포표, 다원분할표, 통계량계산 등 기초자료분석과, 추정 및 가설검정, 회귀분석, 분산분석기법을 패키지에서 다루기로 한다.

-- 일반인이 사용하기에 SAS, SPSS 등 보다 편리하고, 분석방법이 다양하고, 결과출력이 이해하기 쉬우면서 모양도 좋은 패키지를 만들기로 한다.

-- 패키지는 다량의 자료처리가 가능하여야 하며 시스템의 신속성과 정확성이 유지되어야 한다.

이와 같은 개발방향대로 1단계 통계패키지가 개발되면 축적된 기술을 이용하여 차후에 고급 통계분석 등을 추가하여 세계시장에서 경쟁할만한 패키지를 만들어 보려고 한다.

나. 개발전략

일반인이 사용하기에 외국산 패키지보다 우수한 통계패키지를 개발하기위한 방향을 구체화하기위한 전략에는 다음과 같은 것이 있다. 전략의 촛점은 '통계분석내용에는 크게 차이가 있을 수 없으나, 사용자의 편리성을 높이는 패키지를 만들자' 라는 것이다.

-- 한글의 입출력이 자유로운 패키지가 되어야 한다. 외국산 패키지가 우리나라 사용자들에게 가장 불편한 점은 한글로서 입출력이 불가능하다는 점일 것이다.

-- 패키지의 운용시스템은 풀다운메뉴(pull-down menu) 형식이, 자료의 입출력방법은 스프레드쉬트(spreadsheet) 방식을 이용하는 소프트웨어를 개발하는 것이 사용자에게 편리하다. 이러한 시스템 디자인 방식이 사용자에 제일 편리한 것으로 연구 보고되고 있다.

-- 대형컴퓨터용보다 PC용, PC중에서도 우선은 매킨토시용보다는 우리나라에서 제일 많이 사용되는 IBM/PC 호환기종 소프트웨어를 우선 개발한다. 개발되는 통계패키지는 시장에 나와 있는 어떠한 IBM/PC의 호환기종에서도 사용이 가능하게끔 하여야 한다.

-- IBM/PC의 운영시스템중 우선은 DOS시스템에서 작동되는 패키지를 개발한다. 여러 시스템이 장단점을 지니고 있는데, 한글 WINDOWS를 이용하면 한글폰트를 제작할 필요가 없고, 여러 가지 그래픽 사용자 환경(Graphic User Interface)을 제공받을 수 있어 소프트웨어 개발이 쉬울 수 있다. 하지만 현재 우리나라의 대부분의 컴퓨터 사용자는 DOS시스템을 가지고 있고 이에 익숙하여 있기 때문에 WINDOWS용 개발은 많은 일반인에게 도움을 주지 못할 수 있다. 따라서 DOS시스템에서 그래픽환경을 구사하는 패키지를 개발하면 이들 외국산 패키지보다 우수할 수 있다.
-- 그래픽 카드는 IBM/PC의 호환기종에 가장 많이 보급되어 있는 Hercules와 VGA카드를 지원하면 PC 사용자의 90%이상이 이 시스템을 사용할 수 있다.
-- 통계분석의 결과출력은 가능하면 그래프를 많이 이용하여 사용자가 쉽게 결과 분석을 할 수 있도록 한다. SAS나 SPSS 등은 분석결과에 대한 값의 출력에 치중하고 있고, 결과분석에 도움이 되는 그림출력은 미진하다. 이는 WINDOWS버전도 마찬가지이다. 따라서 같은 결과라도 출력을 잘 설계하면 외국산 패키지보다 우수할 수 있다.

3. 통계패키지 개발과정

통계패키지의 개발은 통계학을 연구하는 사람, 특히 통계계산(statistical computing)분야에 관심있는 사람에게는 이론적으로 큰 문제점이 없다. 다만 이러한 이론을 개발방향대로 '사용자들에게 쉽고 유용하게'끔 설계하여 구현하는데 있어서 이러 저러한 방법으로 시도해보아 수정하는 'trial-error'에 수 년간 많은 사람들의 끊임없는 노력이 요구되는 것이 하나의 큰 문제점이다. 본 연구도 90년 8월부터 시작되어 지금까지 계속되어 왔는데 한글을 입출력하는 통계패키지의 개발은 국내에서는 처음이어서 여러 가지 시행착오를 겪으면서 진행되어 왔다. 이 절에서는 그간의 개발과정의 큰 줄거리를 소개하고 어려웠던 문제들이 무엇인지 알아본다.

가. 통계계산 및 그림모듈의 개발

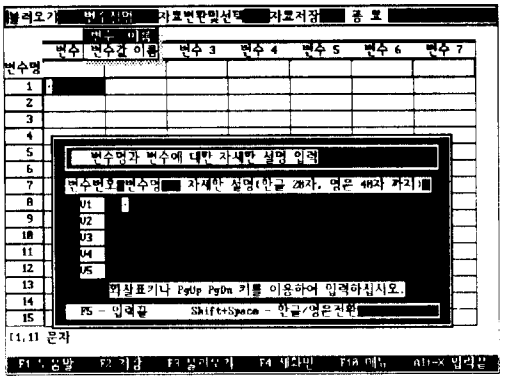
제일 처음 착수한 프로그램은 패키지의 통계분석 프로그램에 사용하기 위해 필요한 각종 분포함수에 관련된 것이다. 구체적으로 이항분포, 포아송분포, 초기하분포, 정규분포, t-분포, 카이제곱분포, F-분포, 베타분포, 감마분포 등의 분포함수식의 계산, 누적확률의 계산, 백분위수의 계산 모듈이 필요하다. 이러한 통계계산 모듈은 statistical computing에 지식이 있는 사람들에게는 큰 문제점이 아니고 이미 문헌에 공개되어 있는 알고리즘을 코드화 하거나 (참고문헌 [10]), 시중에 공개되어 있는 프로그램(참고문헌 [9])을 비교 검토하여 개발한 후 정확도를 조사하여 이용하였다. 이밖에 통계분석 결과의 출력을 위해서 필요한 히스토그램, 상자그림, 산점도 등을 화면의 어디에 위치하든지, 크기가 어떠하든지 마음대로 조절할 수 있는 일반화된 모듈로 자체 개발하였다.

나. 한글입출력 모듈의 개발

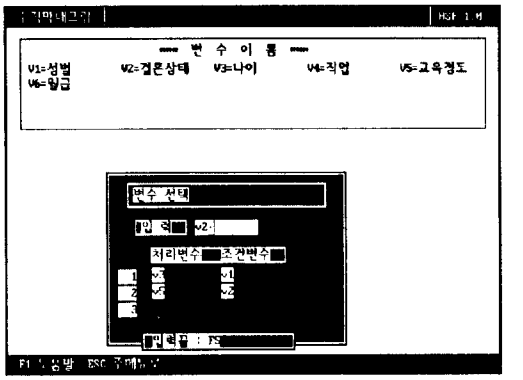
개발되는 통계패키지는 시장에 나와 있는 어떠한 IBM/PC의 호환기종에서도 사용이 가능하게끔 하여야 한다. 제일 중요한 문제는 PC 기종에 관계되지 않고 한글의 입출력을 할 수 있어야 하는데, 이를 위해서는 한글카드를 이용하지 않고 한글폰트를 개발하여 프로그램으로 한글을 스크린에 보이게 하고 프린터에 출력되도록 하여야 한다. 한글의 입출력은 초기단계에서 가장 어려웠던 문제이었다. 다행히 한글 오토마타에 관해서는 많은 연구가 되어 있고(예: 참고문헌 [11]) 프로그램 소스도 공개되어 있는 것이 많아 이를 토대로 자체적으로 개발하였다. 한글폰트는 정부에서 일반인이 사용할 수 있도록 공표한 것도 있고 또 시중에 공용으로 유통되는 폰트들이 있는데 여기에 통계학에서 필요로 하는 50개 정도의 특수문자를 더 개발하여 사용하였다.

다. 자료 입출력모듈의 개발

입력자료를 만들고 이를 출력하는 부분은 스프레드쉬트(SPREADSHEET) 형식으로 사용자가 편리하게끔 설계하기로 전략은 세웠지만, 이 부분의 모듈은 참고문헌이나 예제 프로그램이 부족하고 대부분의 회사에서 스프레드쉬트 프로그램의 DOS버전에 대한 소스는 비밀로 하기 때문에 자체개발에 많은 어려움을 겪었다. 패키지 개발에 착수한 이래 이 모듈은 끊임없이 이러 저러한 방법으로 시도한 후 다시 수정하는 'trial-error'의 대표적인 부분이었다. 초기단계에서는 터보 C에서 제공하는 기본 스프레드쉬트 예제를 변형하여 자체개발하였지만 개발이 진전됨에 따라 독창적으로 사용자가 편리하게끔 변형시켰다. 이 모듈은 패키지에 사용되는 시스템 화일과 서로 맞물려 있어 신속 정확한 자료처리와도 연관되어 있다. 본 패키지는 고심끝에 나무구조(tree structure)형태의 자료관리 시스템파일을 만들어 스프레드쉬트와 연계되도록 설계하였다. 여기서의 또 다른 문제점은 DOS환경에서 지원받을 수 있는 메모리의 기본용량은 640K 밖에 되지 않기 때문에 '어떻게 적은 메모리용량을 효율적으로 관리하여 많은 자료에 대한 입출력을 신속히 할 수 있는가?'이다. 여러 가지 통계패키지의 우열에 대한 기준중의 하나가 바로 이 기능이다. 본 패키지는 확장메모리(extended memory specification)를 이용해서 자료의 입출력 관리를 하도록 하였는데, 자료의 최대수는 3만개, 변수의 최대수는 100개까지 가능하다. 이러한 자료입출력 모듈이 본 통계패키지의 최대 노우하우라고 할 수 있는데, 프로그램의 양도 방대하고 아직은 제품도 시장에 나와 있지 않은 단계라 더 이상 공개하여 자세히 설명하기가 곤란하다. <그림 1>과 <그림 2>는 한글 입출력 모듈과 자료입출력 모듈을 혼합하여 패키지에서 실제 자료를 입력하거나 자료분석 변수를 입력하는 부분의 예이다.



<그림 1> 스프레드쉬트 형태의 자료입력



<그림 2> 자료분석 변수의 입력

라. 시스템 운용에 필요한 모듈 개발

시스템 운용에 필요한 모듈은 크게 한글및 영문폰트 제어용, 그래픽화면 제어용, 프린터출력 제어용, 각종 메뉴디자인 제어용 모듈로 구분할 수 있는데 메뉴의 형태는 사용자에게 편리한 풀-다운(pulldown) 메뉴방식으로 하기로 한 것은 이미 언급하였다. 이러한 시스템 운용에 관한 모듈의 개발기법은 많이 연구되어 있고 문헌도 다양하여 여러 가지 기법을 비교 검토하여 자체 개발 하였다. 다만 프린터출력 프로그램은 워드프로세서를 자체 개발하는 것과 같은 대규모 작업이어서, 편법으로 텍스트(text)와 그래픽의 혼합형태의 화일을 만들어 기존의 "한글" 등 워드프로세서를 이용하여 편집과 인쇄가 가능하도록 하였다.

마. 통계분석 프로그램 개발

위에서 설명한 모듈의 개발이 완료된 후 이 모듈을 이용하여 실제 통계분석 프로그램을 개발하였다. 이 단계에서는 '개발된 여러가지 모듈을 이용하여 어떻게 통계분석기법의 결과를 화면에 효과적으로 나타내느냐?' 하는 디자인의 문제이다. 역시 여러 가지 패키지의 장단점을 비교하여 이러 저러한 방법으로 시도해보아 수정하는 'trial-error'에 많은 시간을 소비하였다. 이밖에 통계분석 프로그램 개발에서 발생하였던 문제점은 사용자가 많은 종류의 통계분석을 요구(예: 수 백개의 분할표나 조절변수가 있는 상자그림 등)하였을 때 '어떻게 메모리를 효율적으로 사용하여 한 번에 많은 분석을 신속하게 하느냐' 하는 점이었다. 이 문제는 1차원 배열(one-dimension array)의 이용과, 소팅(sorting)이 자유로운 시스템화일의 설계로 해결하였다.

4. 결과출력 디자인의 예

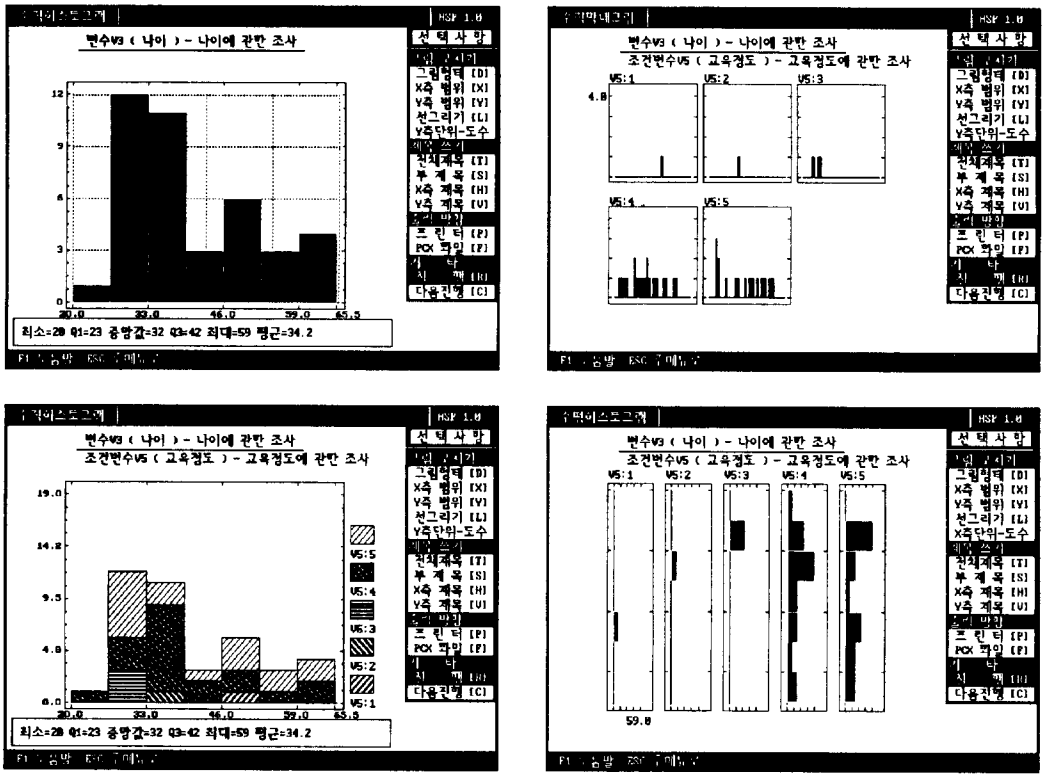
패키지에 포함된 통계분석기법의 종류와 이중에서 SAS나 SPSS와 비교하여 특징있는 결과출력 디자인을 일부 소개하면 다음과 같다.

가. 그림을 이용하는 기초자료분석

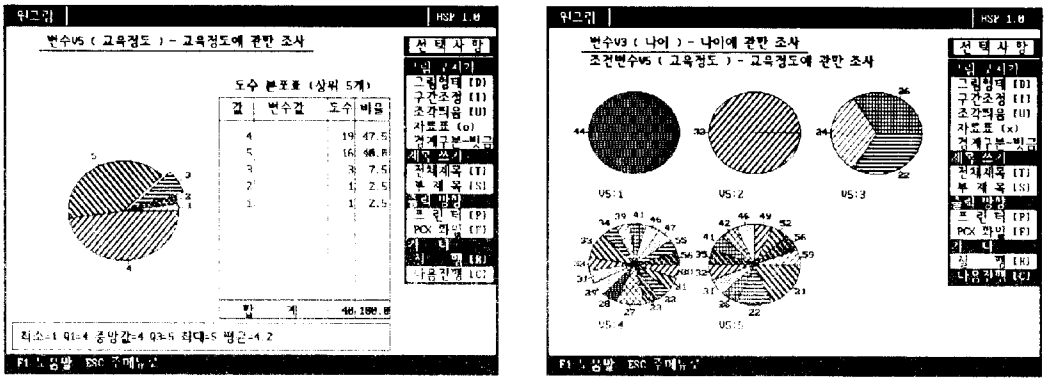
막대그래프(bar graph), 히스토그램(histogram), 선그래프(line graph), 레이더챠트(radar chart), 원형그림표(pie chart), 줄기잎그림(stem-leaf plot), 나무상자그림(box-whisker plot), 산점도(scatter diagram), 산점도 행렬(scatter matrix), 정규분포에 대한 Q-Q그림(Quatile-Quantile plot) 등이 개발되었다. 그래픽 사용자환경(Graphic User Interface: GUI)을 이용하였기 때문에 본 패키지의 그림들은 DOS버전의 SAS나 SPSS 보다는 훨씬 정교하고 다양하다. <그림 3>-<그림 5>는 히스토그램, 원형그림표, 산점도와 산점도행렬의 여러 출력형태 예이다.

나. 도표및 기초통계량 계산

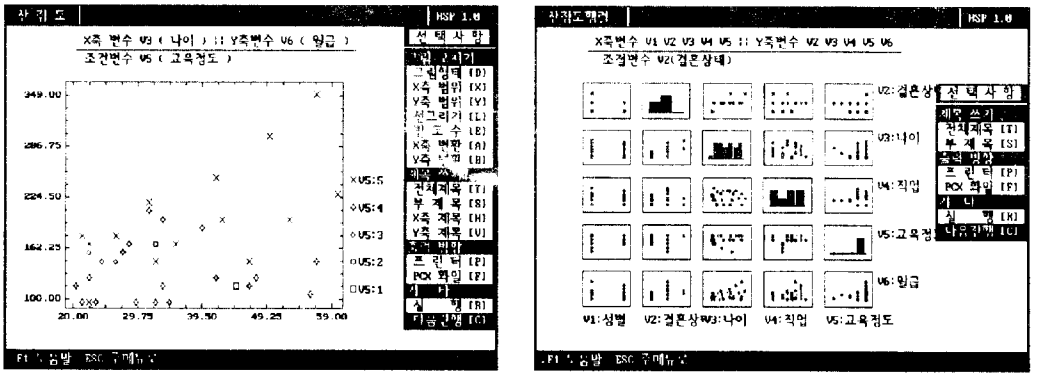
도수분포표(frequency), 다원도수 분할표(cross table), 평균 및 분산표(mean and variance table), 기초통계량(descriptive statistics) 등의 분석기법이 패키지에 포함되었다. 이러한 도표는 항상 그룹변수(group variable) 또는 조절변수(control variable)의 설정이 가능하다. 출력의 디자인은 SAS나 SPSS등과 유사한데 기초통계량 출력시 히스토그램과 상자그림을 같이 출력하여 사용자가 분석을 용이하게 하도록 설계하였다.



<그림 3> 히스토그램의 여러 가지 변형형태



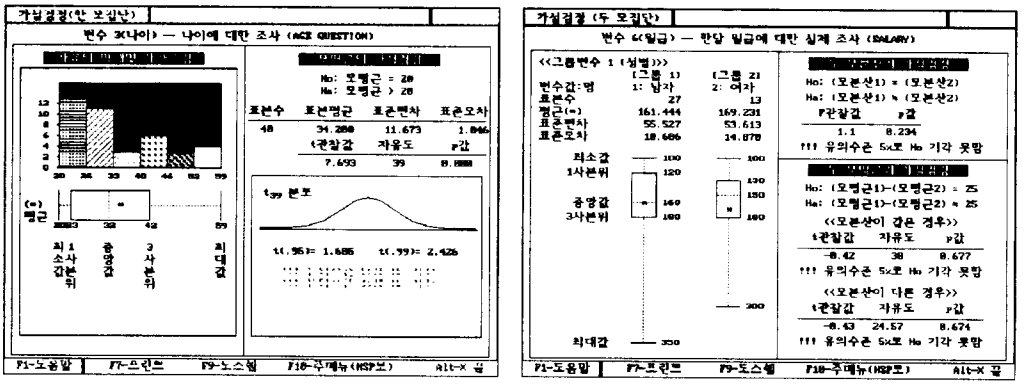
<그림 4> 원형그림표의 변형형태



<그림 5> 산점도와 산점도 행렬

다. 추정및 검정, 분산분석, 회귀분석

추정, 한 모집단의 모평균의 검정, 두 모집단의 모평균 및 모분산에 대한 검정, 대응비교 검정, 일원분산분석, 이원분산분석, 다중회귀분석등의 통계기법이 패키지에 포함되어 있다. 이러한 통계분석의 결과출력도 통계량의 계산은 다른 패키지와 비슷하지만 여러 가지 그림을 같이 출력하여 사용자가 분석을 용이하게 하도록 설계하였다. <그림 6>은 한 모집단과 두 모집단의 평균에 대한 가설검정의 출력형태인데, 히스토그램, 상자그림, t-분포그림에서의 관찰통계량의 위치, 상자그림을 이용한 두 모집단의 비교 등 여러 가지 그림을 보여 주어 사용자의 분석을 도와주도록 설계하였다.



<그림 6> 한 모집단과 두 모집단의 평균에 대한 가설검정의 예

5. 결론및 향후과제

시스템 개발에 여러 패키지를 비교 분석하여 나름대로 사용자에 편리한 시스템을 개발하려고 노력은 했으나 부족한 점이 많이 있을 것 같다. 많은 분들의 조언을 부탁드린다. 그리고 본 연구에서 개발된 통계패키지는 각종 고급 통계기법은 포함되어 있지 않아 전문적인 통계분석을 요하는 학자나, 연구원에게는 불충분할 것이다. 하지만 본 통계패키지만으로도 기존의 SAS 나 SPSS 를 수입하여 사용하고 있는 사용자의 상당수를 만족시킬 수 있다고 생각된다. 또 이와같은 한글 통계패키지는 대학이나 기타 교육기관에서 통계학을 교육시키는데는 충분히 잘 활용될 수 있다고 생각된다.

당장에 해야될 일은 향후 PC 시스템의 발전에 맞게 WINDOWS용 소프트웨어로의 전환이 될 것이다. 이 시스템을 이용하면 다작업기능, 그래픽 사용자 환경(GUI: Graphical User Interface) 등 다양한 형태의 바람직한 통계소프트웨어를 개발할 수 있을 것이다. 앞으로의 과제는 학자나 연구원과 같은 전문인들을 위한 통계패키지 개발이 될 것이다. 일반용 통계패키지를 개발한 경험이 있기 때문에 이제 어떻게 하면 이 사업을 추진할 수 있을 것인지 방향은 잡을 수 있다. 하지만 이 사업은 한 두사람의 힘만으로는 될 수 없고 국내의 모든 통계인들이 합심하여야 가능한 일이다. 조만간에 중지를 모아 전문가용 통계패키지 개발에 착수하였으면 한다.

참 고 문 헌

[1] 강근석외 5인 (1993). 『PC통계학』, 자유아카데미, 서울.
[2] 김우철외 7인 (1990). 『현대통계학』, 제3판, 영지문화사, 서울.
[3] 안현순 (1992). 『터보 C로 구현한 과학기술계산 프로그래밍』, 가남사, 서울.
[4] 이준희, 정내권 (1991). 『컴퓨터속의 한글』, (주)정보시대, 서울.
[5] 한국통계학회 (1987). 『통계용어사전』, 자유아카데미, 서울.
[6] Cooke, D., Craven, A.H., and Clarke, G.M. (1989). *Basic Statistical Computing,* Arnold, London.
[7] Kennedy, W.J. and Gentle, J.E. (1980). *Statistical Computing,* Dekker, New York.
[8] Maindonald, J.H.(1984). *Statistical Computation,* John Wiley & Sons, New York.
[9] SAS Institute Inc. (1985). *SAS User's Guide: Statistics,* 5th Edition, Cary.
[10] Schildt H. (1990). *Turbo C The Complete Reference,* Boland/McGraw-Hill, New York.
[11] SPSS Inc. (1988). *SPSS-X User's Guide,* 3rd Edition, Chicago.

Developing a Korean Statistical Package[4)]

Jung Jin Lee[5)], Gunseog Kang[6)]

Abstract

Most of the statistical packages being used in Korea, such as SAS or SPSS, are imported from foreign countries. Since these packages are written in English, it is not easy for Korean to learn the statistical packages. Also, most of the users except statistician use these expensive packages only to draw pictures and to make tables. We introduce a Korean statistical package which can be used easily for general public.

---

4) This paper was supported in part by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1992.

5) Department of Statistics, Soongsil University, Sangdo-Dong 1-1, Dongjak-Ku, Seoul, 156-743, KOREA.

6) Department of Statistics, Soongsil University, Sangdo-Dong 1-1, Dongjak-Ku, Seoul, 156-743, KOREA.

참 고 문 헌

[1] 강근석외 5인 (1993). 『PC통계학』, 자유아카데미, 서울.
[2] 김우철외 7인 (1990). 『현대통계학』, 제3판, 영지문화사, 서울.
[3] 안현순 (1992). 『터보 C로 구현한 과학기술계산 프로그래밍』, 가남사, 서울.
[4] 이준희, 정내권 (1991). 『컴퓨터속의 한글』, (주)정보시대, 서울.
[5] 한국통계학회 (1987). 『통계용어사전』, 자유아카데미, 서울.
[6] Cooke, D., Craven, A.H., and Clarke, G.M. (1989). *Basic Statistical Computing,* Arnold, London.
[7] Kennedy, W.J. and Gentle, J.E. (1980). *Statistical Computing,* Dekker, New York.
[8] Maindonald, J.H.(1984). *Statistical Computation,* John Wiley & Sons, New York.
[9] SAS Institute Inc. (1985). *SAS User's Guide: Statistics,* 5th Edition, Cary.
[10] Schildt H. (1990). *Turbo C The Complete Reference,* Boland/McGraw-Hill, New York.
[11] SPSS Inc. (1988). *SPSS-X User's Guide,* 3rd Edition, Chicago.

Developing a Korean Statistical Package[4)]

Jung Jin Lee[5)], Gunseog Kang[6)]

Abstract

Most of the statistical packages being used in Korea, such as SAS or SPSS, are imported from foreign countries. Since these packages are written in English, it is not easy for Korean to learn the statistical packages. Also, most of the users except statistician use these expensive packages only to draw pictures and to make tables. We introduce a Korean statistical package which can be used easily for general public.

4) This paper was supported in part by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1992.

5) Department of Statistics, Soongsil University, Sangdo-Dong 1-1, Dongjak-Ku, Seoul, 156-743, KOREA.

6) Department of Statistics, Soongsil University, Sangdo-Dong 1-1, Dongjak-Ku, Seoul, 156-743, KOREA.