Journal of the KSQC
Vol. 13, No. 1, 1985

마이크로 컴퓨터에 의한 統計資料分析에 관한 硏究

A study on statistical data analyis by microcomputers

朴 聖 炫*

ABSTRACT

First of all, the necessity of statistical packages, and the strengths and weaknesses of microcomputers for statistical data analysis are examined in this paper. Secondly, some statistical packages available for microcomputers in the international market are introduced, and the contents of two statistical packages developed by the author are presented.

1. 통계팻키지 (Statistical Package) 의 필요성

産業, 技術, 學問 등이 점점 發展됨에 따라서 각 분야에서 처리하여야 할 資料 (data) 의 양이 방대하여지고, 또한 신속성을 아울러 요구하게 되었다. 그러므로 적은 량의 데이타에 대한 초보적인 통계처리를 除外하고는 사람의 손으로 처리한다는 것이 거의 不可能하며, 컴퓨터의 사용이 불가피하게 되었다. 데이타를 電算 處理하여 分析하기 위해서는 그 目的에 맞는 프로그램 을 작성하여야 한다. 그러나 매 작업마다 프로그램을 새로이 作成한다는 것은 非能率的인 일이며, 또한 프로그램을 짤 수 있는 人力도 제한되어 있으므로 어리운 일이 될 것이다.

빈번히 사용되는 統計資料處理方法은 그 種類가 제한되어 있으며, 그 처리형식이 일정하기 때문에 컴퓨터 프로그래밍이 可能하다. 이런 必要性에 의하여 開發된 것이 소위 말하는 통계팻키지 (statistical package) 이다. 이것은 각 통계처리에 적합하도록 미리 프로그램들이 작성되어 있으며, 여기에 資料를 入力시키고, 요구하는 資料處理方法을 명시하는 내용을 入力시키면, 통계 결과가 出力되게 되어 있다.

2. 마이크로 컴퓨터의 장점

몇 년 전부터 대량으로 생산·판매되고 있는 마이크로 컴퓨터는, 몇 가지 장점으로 인하여 향후 統計分析에 상당히 큰 비중을 가질 것임에 틀림

* 서울대학교 자연과학대학 계산통계학과 교수
(본 연구는 1984년 - 1985년도 문교부 학술연구 조성비에 의하여 수행되었음.)

없다. 이들 장점들을 들어보면 다음과 같은 것이다.

1) 값싸게 개인이 구입할 수 있을 정도의 가격으로까지 저렴화될 것이다.

2) 플로피 디스크(floppy disk) 등의 보조기억장치의 발달로 인하여 길고 복잡한 통계계산 프로그램을 저장할 수 있으며 언제든지 간편하게 사용 가능하다.

3) 대화식(interactive)으로 입력·출력 및 작업처리가 가능하므로 단계적으로 통계분석이 요망되는 경우에는 매우 편리하다.

4) 그래픽(graphics) 기능이 다양한 마이크로 컴퓨터가 출현하면서 통계그래프나 圖表등을 시각적으로 이해하기 편리하도록 출력시킬 수 있다.

5) 마이크로 컴퓨터회사 또는 소프트웨어 하우스 등에서 통계프로그램에 대한 연구 및 발표가 증대되고 있으며, 사용자는 이들 중에서 선택하여 값싸게 구입할 수 있을 것이다.

6) 마이크로 컴퓨터는 학생들의 통계교육에 매우 편리하다.

3. 마이크로 컴퓨터의 문제점

마이크로 컴퓨터도 統計分析을 할 때 문제가 전혀 없는 것은 아니다. 그러나 이런 문제점들은 향후 해결될 가능성이 높다고 믿어진다. 마이크로 컴퓨터가 갖는 몇 가지 문제점을 열거하면 다음과 같다.

(1) 하드웨어(hardware)에 따라 달라지는 언어

현재 대부분의 마이크로 컴퓨터가 BASIC언어를 주종의 언어로 사용하고 있으나, 이 BASIC 언어도 하드웨어 생산업체가 다르면 모두 동일하지 않다. 예를 들면 Apple Ⅱ에 짜여진 BASIC 프로그램이 TRS-80에 그대로 사용되지 않고 약간의 수정작업을 거쳐야 한다.

따라서 어느 한 기종에 맞도록 개발된 팻키지는 다른 기종에 그대로 사용할 수 없으므로 사용상 문제점이 많다. 그러나 이러한 문제점도 생산업체간 서로 호환성 있는 소프트웨어의 개발에 깊은 관심을 가지고 있으므로, 앞으로는 차츰 개선될 것이 틀림없다.

(2) 기억용량(memory size)상의 제한성

통계분석에서는 종종 상당히 큰 RAM이 요구되는 경우가 있다. 예를 들면 중회귀분석에서 독립변수의 수가 10개 이상이면 행렬의 역행렬을 구하는 과정에서 상당히 큰 기억용량이 필요하게 된다.

이러한 문제점은 값싼 마이크로 컴퓨터에서는 언제나 존재하는 문제지만, 과학기술의 발달과 더불어 차츰 해소되리라 믿는다.

(3) 계산속도(computational speed)상의 문제

아직도 마이크로 컴퓨터의 계산속도는 중형이상의 컴퓨터에 비하여 상당히 느린 것은 사실이나 과거 수년간 마이크로 컴퓨터의 계산속도는 급진적으로 개선되고 있으며, 대부분의 통계계산은 소비자에게 만족스런 계산속도를 제공하고 있다고 생각된다.

(4) 그래픽(graphics) 문제

아직도 그래픽 기능을 못가진 마이크로 컴퓨터가 많이 있고, 가지고 있더라도 정교하지 못한 경우가 많으나, 이 분야는 마이크로 컴퓨터의 기능에서 가장 빠르게 보강되어 가고 있다고 생각된다. 최근에는 색깔을 나타내는 그래픽(color graphics)도 출현하고, light pen으로 입력이 가능해지고, hard-copy plotter가 나타나는 등, 이 방면에서 눈부신 성장을 거듭하고 있다. 필자가 믿기에는 그래픽 기능은 향후 마이크로 컴퓨터가 갖는 가장 큰 장점이 될 것이다.

4. 마이크로 컴퓨터용 통계팻키지의 소개

다음에 이미 개발되어 시판되고 있는 마이크로 컴퓨터용 통계팻키지 중에서 필자가 생각하기에 좋다고 생각되는 9가지만 예를 들어 보기로 한다. 여기서 CP/M(control program for microcomputers)이라고 쓴 것은 CP/M을 사용하여 돌아가는 컴퓨터에서는 사용 가능하다는 뜻이고, PET는 Commodore 社에서 만든 컴퓨터이고, TRS-80은 Radio Shack 社에서, Apple은 Apple 社에서 제작된 컴퓨터를 뜻한다.

팻키지 이름	공 급 처	팻키지의 주내용	언 어	기 종	가 격
AIDA	Appli-Tech Software Services Broad Oak Accrington BB5 2DJ ENGLAND	그래프, 시계열, 회귀분석, 자료정리, 결측치 취급, 각종의 도표작성	Pascal	Apple	£ 450
A.STAT	Rosen Grandon Assoc. 296 Peter Green Road Tolland Connecticutt 06084 U.S.A.	각종의 자료변환, 기술통계, 도표작성, 회귀분석	Basic	CP/M Apple PET	$ 125
DATA-X	Patrick Royston 85 Canfield Gardens London NW6 ENGLAND	데이타 editing, 비모수통계, 분산분석, 회귀분석, 기술통계	Basic	PET	£ 350
Dynacomp	Dynacomp Inc. 1427 Monroe Ave. Rochester New York, 14618 U.S.A.	회귀분석 분산분석	Basic	Apple TRS-80 PET	$ 238
Microstat	Ecosoft PO Box 68602 Indianapolis Indiana 46268 U.S.A.	데이타관리, 기술통계, 가설검정, 분산분석, 비모수통계, 확률분포	Basic	CP/M	$ 250
SAM	International Software PO Box 160 Welwyn Gdn city Herts AL8 6TQ ENGLAND	데이타관리, 각종의 도표, 회귀분석, 인자분석, 분산분석, 가설검정	Basic	Apple PET CP/M	£ 335
STATPAK	Northwest Analytical Inc. PO Box 14430 Portland Oregon 97214 U.S.A.	파일관리, 기술통계, 회귀분석, 비모수, 분산분석, 가설검정, 확률, 난수, 도표작성	Basic	CP/M	$ 500
MASS	Westat Associates 60 Bruce Street Nedlands, West Australia 6009	데이타관리, 기술통계, 각종의 도표, 가설검정, 회귀분석, 판별분석, 비모수통계	Pascal	CP/M	미지임
Statistics Pac	Creative Discount Software 256 S Robertson Suite 2156 Beverly Hills California 90211 U.S.A.	데이타관리, 곡선적합, 확률계산, 일반통계	Basic	TRS-80 Apple	$ 100

5. SQC 팻키지

앞에 소개된 9가지 통계팻키지는 세상에 알려진 수십 가지 중에서 극히 일부에 지나지 않으며, 마이크로 컴퓨터의 급진적인 보급과 더불어 이런 프로그램들의 활용성은 더욱 증대될 것으로 생각된다. 우리나라에는 이런 프로그램들이 아직 소개되어 있지 않으며, 앞으로 마이크로 컴퓨터가 더 많이 보급됨에 따라서 통계팻키지의 필요성이 반드시 대두되리라 믿는다. 향후 이러한 수요에 일부만이라도 충족시켜 주기 위하여 필자가 과거 1년간에 걸쳐서 서울대학교 계산통계학과 학생들과 같이 금성사의 Mighty 마이크로 컴퓨터를 이용하여 개발한 통계팻키지를 소개하기로 한다.

이 팻키지는 금성사의 마이크로 컴퓨터 Mighty (구체적으로는 모델 GMC-2010에서 개발되었음)을 사용하여 개발된 것으로, 보조기억장치로서 5.25 inch의 미니 플로피 디스크(mini floppy disk)의 용량 306KB을 사용하여 작성되었다. OS(operating system)로는 세계적으로 널리 알려진 CP/M을 사용하였으며 기존의 모든 Mighty 기종과 호환성 있게 작성된 것이다. 프로그래밍 언어로는 BASIC 언어가 사용되었는데 CP/M 상에서는 이를 MBASIC으로 명명하여 사용하고 있다. 이 팻키지는 대부분의 일반통계를 전산화하였으나, 이중에서 특히 統計的 品質管理(statistical quality control) 方法을 전산화한 것이 많으므로 「SQC」라고 명명하였다. 이 팻키지는 금성사 OA 사업부에서 저렴한 값으로 구입할 수 있다.

5.1. 팻키지의 내용

이 팻키지는 기업이나 연구소 등에서 품질관리를 수행할 때 흔히 사용되는 모든 통계적 방법을 전산화한 것으로 모두 5개의 부문(category)으로 나누어져 있고 어느 한 부문을 임의로 택하여 사용할 수 있다. 각 부문은 또한 다음과 같이 여러 개의 프로그램으로 구성되어 있으며 사용자가 임의로 하나의 프로그램을 택하여 사용할 수 있다. 이 팻키지는 이 책에서 소개되는 통계 프로그램 중에서 발췌하여 작성된 것으로 대부분의 중요 프로그램들이 포함되어 있다.

(1) SQC의 기본적인 도구 (basic tools for SQC)

데이타의 그래프에 의한 기본적인 분석방법으로 다음의 5개 프로그램 중 하나를 택하여 사용할 수 있다.

1) 히스토그램
2) 파레토그림
3) 산점도
4) 각종의 그래프(막대그래프, 꺾은선그래프, 띠그래프)
5) 각종의 관리도($\bar{x}$ - R관리도, x - R_s관리도, p관리도 등)

(2) 통계적 가설검정과 신뢰구간 추정 (statistical hypothesis testing and confidence interval estimation)

다음과 같이 7개의 프로그램으로 되어, 이중 하나를 택하여 사용할 수 있다.

1) 모평균의 검정과 신뢰구간 추정
2) 모비율의 검정과 신뢰구간 추정
3) 모분산의 검정과 신뢰구간 추정
4) 두 모집단의 모평균 차의 검정과 신뢰구간 추정
5) 두 모집단의 모비율 차의 검정과 신뢰구간 추정
6) 두 모집단의 모분산비의 검정과 신뢰구간 추정
7) 분할표에 의한 독립성 검정

(3) 상관분석 및 회귀분석 (correlation analysis and regression analysis)

다음과 같이 4개의 프로그램으로 구성되어, 이중 하나를 택하여 사용할 수 있다.

1) 상관분석
2) 단순 및 곡선 회귀분석
3) 중회귀분석
4) 변수 선택에 의한 회귀분석

(4) 최적조건을 찾는 통계분석 (statistical analysis for optimal conditions)

다음중 하나를 택하여 사용할 수 있다.

1) 반응표면분석

2) 혼합물 실험분석

(5) 실험계획법 (experimental designs)

실험계획법에서 가장 많이 사용되는 다음의 3가지 실험계획법이 하나의 프로그램으로 작성되어 있다. 인자의 수의 입력에 의해 1,2,3원 배치로 나누어진다.

1원 배치법 : 반복수가 같거나 다른 경우

2원 배치법 : 반복이 없거나 있는 경우

3원 배치법 : 반복이 없거나 있는 경우

5.2 팻키지의 사용방법

이 「SQC」 팻키지의 사용방법은 다음 순서에 의하여 실시하면 된다.

1) 「SQC」 diskette 을 disk drive 에 넣고, monitor, disk drive 및 프린터의 전원을 킨다.

2) RETURN 키를 세 번 누른다.

3) A>란 표시가 나오면 MBASIC SQC 를 부르고 RETURN 키를 누른다.

4) 「SQC」 팻키지에 대한 간단한 설명이 나타나고, 다시 RETURN 키를 누르면 앞에서 설명된 5개부분의 설명이 화면에 나타난다.

이 때 원하는 부분의 번호(1, 2, 3, 4, 5번 중의 하나)를 누른다.

5) 그러면 선택된 부문에 속해 있는 프로그램의 이름들이 화면에 나타나고, 이때 원하는 프로그램의 번호를 누른다.

6) 선택된 프로그램의 이름이 나타나고, 화면에서 대화식으로 요구하는 필요한 데이타를 입력시키기 시작한다.

7) 입력 중에 (Y/N) 을 물어보는 것은 yes/no 의 물음으로, Y 또는 N의 입력은 대문자로 하여 주어야 한다. 따라서 CAPS LOCK 버튼을 눌러주고 시작하는 것이 편리하다.

8) 출력시키기 위해서는 출력파일 (output file) 이름을 지정하여 저장한 후에 SYSTEM 을 type 하여 CP/M 상으로 돌아오고 A> 표시가 나타나면 다음 중 택일하여 사용한다.

a) 화면에 출력시키고 싶으면 TYPE '출력파일 이름'

b) 프린터로 출력을 뽑아보고 싶으면 PIP LST : = '출력파일 이름', 예를 들면, 만약 출력파일 이름은 B:PQ로 하였다면 다음과 같이 입력하여 프린터에 출력을 뽑아볼 수 있다.

SYSTEM

PIP LST : = B : PQ

5.3 파일의 관리

Mighty 컴퓨터는 기본적으로 2대의 disk drive 와 같이 작동하게 되어 있는데 각 drive 의 이름을 A, B(0,1으로 표시되어 있음)라 한다. 「SQC」 팻키지를 사용할 때 2대의 drive 를 모두 사용할 수 있다. 예를 들면, drive A 에서는 프로그램을 읽기만 하고, drive B 에서는 데이타와 출력파일을 저장하고 필요에 따라 사용할 수 있다. 이 때는 사용하고자 하는 drive 를 지정해야 한다.

이 「SQC」 팻키지는 디스크용량 306 K 를 차지하고 있고 임의로 사용 가능한 것은 12KB 밖에 남아 있지 않으므로 많은 양의 데이타나 출력은 drive B 의 diskette 에 저장하는 것이 바람직하다.

프로그램의 수행결과는 반드시 출력파일에 저장되어야 하는데, 이 때 출력파일명을 지정해 주어야 한다. 컴퓨터가 출력파일명을 물어보면 다음과 같이 할 수 있다. (「SQC」 팻키지 diskette 가 drive A에 들어 있고, 파일명은 임의로 정할 수 있으나 SHP.OUT이라고 가정하자).

출력파일 이름은

SHP.OUT (이 경우에는 drive A에 SHP.OUT이란 이름으로 저장됨)

A: SHP.OUT (위와 동일)

B: SHP.OUT (이 경우에는 drive B에 SHP.OUT이란 이름으로 저장됨)

만일 「SQC」 팻키지를 drive B에서 읽어들

인 것이면 A와 B를 바꾸어 주어야 하며 데이타 파일명도 같은 방식으로 지정할 수 있다.

데이타 파일명의 뒤에 DAT를 붙이고, 출력파일에는 OUT를 붙이면 구분, 삭제, 복사 등의 파일관리가 용이하다.

6. SDA 팻키지

이 팻키지는 삼보 컴퓨터의 8 bit 용 마이크로 컴퓨터인 Tri Gem20에서 개발된 것으로 데이타의 통계적 분석에 많이 사용되는 통계적 방법을 대부분 망라하고 있으며, 모니터를 사용하여 사용자를 대화식(interactive)으로 유도하면서 유용한 정보를 추출할 수 있도록 작성되어 있다. 이 팻키지는 24개의 통계프로그램이 2장의 diskette에 12개씩 나누어 저장되어 있으며, 각 diskette에는 12개의 프로그램이 4개의 category로 나누어져 있다. 단 diskette 1에는 용량 140 KB 중에서 106 KB가 사용되고 34 KB가 data 및 output 저장용의 free space로 남아있으며, diskette 2에는 132 KB가 사용되고 8 KB가 free space 로 남아 있다.

이 팻키지의 구입은 삼보컴퓨터(주)의 소프트웨어 개발과에서 구입할 수 있다. 여기서 SDA의 의미는 통계자료분석(Statistical Data Analysis)을 의미하는 약자이다.

6.1 팻키지의 내용

두 장의 diskette (SDA-1 , SDA-2)에 담겨 있는 내용은 다음과 같다. 이 프로그램은 영어로 되어 있으므로 영어로 소개하기로 한다.

SDA-1 (diskette 1)

1. Basic statistical analysis & graphical presentation
   - (1) Basic statistics (program BASIC STATISTICS)
   - (2) Histogram (program HISTOGRAM)
   - (3) Pareto diagram (program PARETO DIAGRAM)
   - (4) Scatter diagram (program SCATTER DIAGRAM)
   - (5) Several graphs
     - circle graph (program CIRCLE GRAPH)
     - strip graph (program STRIP GRAPH)
     - bar and line graph (program BAR & LINE GRAPH)

2. Statistical estimation and hypothesis testing
   - (1) Population mean, Population proportion, Population variance — (program ESTIMATION 1)
   - (2) Difference of two population means, Difference of two population proportions, Difference of two population variances — (program ESTIMATION 2)

3. Goodness-of-fit test
   - (1) By contigency table (program CONTIGENCY)
   - (2) By probability distribution (program GOODNESS)

4. Random variate generation (program SIMULATION)

   (uniform, normal, gamma, chi-square, F, binomial, Poisson)

SDA-2 (diskette 2)

1. Correlation and regression analysis
   - (1) Correlation analysis (program CORRE-

LATION ANALYSIS)

(2) Simple and curvilinear regression analysis (program SIMPLE & CURVILINEAR REG.)

(3) Multiple regression analysis (program MULTIPLE REGRESSION)

2. Analysis for optimal conditions

(1) Response surface analysis (programs RESPONSE SURFACE ANALYSIS, RSP)

(2) Mixture experiments analysis (programs MIXTURE ANALYSIS MIXTURE PLOTTING)

3. Experimental designs

(1) One-way classification (program ONE-WAY)

(2) Two-way classification (program TWO-WAY)

(3) Three-way classification (program THREE-WAY)

4. Control charts

(1) XBAR-R control chart, X-RS control chart — (program CCFV)

(2) P control chart, PN control chart, C control chart — (program CCFA)

6.2 팻키지의 특징

(1) 모니터에 의해 대화식으로 입력과 출력이 이루어지므로 입출력이 용이하고, 출력은 CRT로 할 수도 있고, 프린터로도 할 수 있도록 option을 주고 있다.

(2) 각각의 diskette에서 프로그램이 MENU 식으로 나열되어 있으므로 원하는 프로그램을 언제든지 쉽게 선택하여 사용할 수 있으므로 operating이 매우 용이하다.

(3) 각 프로그램의 중간과정에서 다음과 같은 DATA MODE를 수시로 사용할 수 있도록 되어 있다.

1. INPUT : 데이타 입력을 원할 경우
2. REVIEW : 데이타를 확인하고 싶은 경우
3. MODIFICATION : 데이타를 수정하고 싶은 경우
4. SAVE : 데이타를 diskette에 저장하고 싶은 경우
5. LOAD : 데이타를 diskette에서 꺼내고 싶은 경우
6. QUIT THE DATA MODE: DATA MODE를 떠나서 다음 단계의 작업을 수행하고 싶은 경우

따라서 이 팻키지는 데이타 관리에 있어서 자유자재로 입출력, 확인 수정등이 가능하다.

6.3 팻키지의 사용방법

이 팻키지의 구성은 2장의 diskette로 되어 있고, 첫째가 SDA-1이고 두번째가 SDA-2 이다.

SDA-1은 SDA Package I의 hello program이다. Diskette을 booting 시키면 먼저 이 program이 실행되는데, 이것은 SDA Package의 개략적인 설명을 포함하고 있다. 또한 이 program을 통해 SDA Package I의 여러 program 등을 선택적으로 RUN 시킬 수도 있다. program의 선택은 다음과 같이 한다.

SDA-1이 실행되면, 먼저 title과 전체 Package의 소개가 나오며, 어떤 Key든지 누르면 main menu로 넘어간다. main menu는 2 page로 되어 있으며 화면에는 한번에 하나의 page만 보인다. 다른 page를 볼려면 space-bar를 누르면 되고, 4개의 category 중 하나를 선택했으면 그 번호를 누르면 된다. 일단 선택이 되면 그 category의 sub menu가 나오고 다시 선택하여 입력하면 원하는 program이 실행된다.

4개의 category와 각 category의 sub menu

는 앞의 1,2절의 내용에서 소개된 것과 동일하다. 각각의 category 에서 sub menu 중에 (9) RETURN TO MAIN MENU 를 선택하면 4개의 category 중 하나를 다시 선택할 수 있다.

SDA-2는 SDA Package Ⅱ의 hello program 이며 사용방법은 SDA-1과 동일하다. 여기에서도 main menu에 4개의 category가 있고, 각 category 에 sub menu 가 있다.