0-1. 통계패키지 비교

데이터 분석에 컴퓨터를 이용한 통계적 분석이 갈수록 중요해지고 있다. 많은 수의 개체를 다루는 연구나 역학적인 연구에서 데이터 분석은 필수적이다. 그러한 분석을 수작업이나 계산기로 할 수 있다 하더라도 얼마나 많은 시간이 걸릴 것인가?
이러한 작업을 편리하게 도와주는 여러 가지 통계 소프트웨어들이 개발되어 사용되고 있으며, 통계 분석을 위해 개발된 컴퓨터 프로그램을 통계 패키지라고 한다.

통계 패키지를 크게 상용과 비상용으로 분류할 수 있는데, 우리들이 사용해 보았거나 이름이 유명한 것들은 대부분 상용 패키지이다.
SAS, SPSS, Minitab, Stata, LISREL, Medcalc, BMDP, MATLAB, Genstat 등 같은 것들이 이에 해당한다.

비상용 통계 패키지로 유명한 것은 여기서 다루는 R과 Epi-Info가 있다.
R은 프로그램의 source가 모두 공개된 open source 프로그램인 반면, Epi-Info는 source가 공개되지는 않았지만 누구나 사용할 수 있는 공용 프로그램이다.

이중 SAS는 행동과학 분야에 자주 이용되어 왔으며, SPSS는 사회과학과 마케팅에서 흔히 사용되며, BMDP(Bio-Medical Data Package)는 의학에서 흔히 사용된다. 그러나 최근에는 특화된 분야에서만 사용되는 프로그램은 줄어드는 추세여서 이러한 구분을 하는 것도 사실 무의미하다. 오히려 특수한 분석절차만이 가능한 프로그램이 만들어 지고 있으며, 또한 이러한 프로그램도 다른 프로그램에 통합되어 가는 추세에 있다.

1. SAS

SAS란 Statistical Analysis System (통계 분석 시스템) 또는 Strategic Application Software(전략 응용 소프트웨어)의 약자로서 미국 North Carolina에 있는 SAS 연구소에 의해 1970년대 초에 개발된 범용 통계분석 패키지

여타의 자료분석 응용 프로그램과의 차이는 SAS가 다른 자료분석 응용 프로그램보다 사용자가 지정해 주지 않아도 컴퓨터가 미리 알아서 처리하는(default) 방법이 적어 사용자가 적절한 분석방법을 일일히 지정해 주어야 하는 경우가 많아 다소 어렵고 까다로운 점은 있으나 다른 프로그램보다 훨씬 정밀한 결과를 산출해 낸다는 장점이 있다.
SAS를 간단히 소프트웨어라고 부르지 않고 시스템이라고 부르는 이유는 여러 개의 단위 소프트웨어들로 구성되어 있기 때문이다.

이제는 SAS는 다른 설명을 붙이지 안고 고유명사로 그냥 SAS라고 부릅니다.

2. SPSS

SPSS(Statistical Package for the Social Science)는 원래 사회과학분야의 데이터 분석을 위한 컴퓨터 프로그램들의 모음집으로 출발하여 1969년에 처음으로 SPSS가 발표되고, 1975년에 SPSS Manual이 발간되면서 세상에 알려지기 시작하였다.

SPSS의 가장 큰 특징은 사용의 편리함에 있다고 할 수 있다. SPSS가 내걸고 있는 이슈가 'Real stats, Real easy'인 만큼 사용의 편이성이 이 프로그램의 가장 큰 장점이라고 할 수 있다. 특히 원도우용 프로그램의 경우 어지간한 통계치는 마우스를 클릭하는 것만으로도 충분히 얻을 수 있다. 그래서 과거 통계를 어려워했던 연구자들도 쉬워진 SPSS를 사용하여 직접 자료분석을 하고 있는 실정입니다.(의학, 보건학, 의류학, 영양학, 공학, 농학 등 거의 모든 영역에서 사용 중입니다.)

오늘날에는 사회과학과 자연과학 등 거의 모든 학문분야에 있어서의 통계분석뿐만 아니라 일반회사나 증권회사, 은행 등의 금융기관에 있어서 각종자료의 정리 및 보고서 작성 등을 할 수 있는 기능을 갖게 되었다.

3. MINITAB

기초통계학을 수강하는 학생을 위해 펜실베니아 대학에서 1972년 개발되었으나 그 후 공학, 사회학, 심리학, 경영학 등 자료의 분석을 통해서 연구하는 모든 분야에서 널리 사용하고 있다.
처음에는 범용컴퓨터에 사용되도록 개발되었지만 1982년에 PC에도 사용할 수 있도록 개발되었다.

MINITAB은 SPSS, SAS 등과 같은 통계 프로그램에 비해 차지하는 용량이 적고 상대적으로 사용이 간편하다는 장점으로 널리 사용되고 있다. 또한 수행결과를 살펴보기에는 편리한 반면에 통계 분석적인 측면에서 SPSS, SAS 보다는 떨어지고 있는 실정이다.
산업공학에서 품질관리, 신뢰성분석, 6-sigma등에서 자주 활용된다.

4. R

R은 뉴질랜드 오클랜드대학의 Ross Ihaka와 Robert Gentleman교수가 주도해서 만들어졌는데, 이들의 앞의 이니셜을 따서 R로 명명되었다. S-PLUS의 기초적 프로그램을 작성했던 Chambers 등 많은 통계학 및 컴퓨터 관련 학자들이 R 개발 핵심 팀으로써 R을 만드는데 참가하고 있다.

1997년 8월 R언어를 구체화하기 위한 국제적인 논의가 진행되었으며, 2000년 2월에 버전 1.0.0이 등장해서 현재에 이르고 있다. 또한 많은 통계학 및 계량분석자들이 R을 보다 잘 이용할 수 있도록 응용 소프트웨어를 개발하고 있으며 이를 R에 덧붙혀서 쓰면 매우 유용하다. R로 만들어진 응용 프로그램의 분야로는 금융, 생명공학, 조사, 지리정보 등 거의 모든 것이다.

R은 다양한 통계 기법과 수치 해석 기법을 지원한다. R은 패키지를 추가하여 기능을 확장할 수 있다. 핵심적인 패키지는 R과 함께 설치되며, CRAN(the Comprehensive R Archive Network)을 통해 패키지를 내려 받을 수 있다.
R은 통계 계산과 소프트웨어 개발을 위한 환경이 필요한 통계학자와 연구자들 뿐만 아니라, 행렬 계산을 위한 도구로서도 사용될 수 있으며 이 부분에서 MATLAB에 견줄 만한 결과를 보여준다.

장점

1. 가격이다. 무료 통계 패키지라는 것은 정말 대단한 매력이다.
2. 성능이다. open source 이기 때문에, 개발되어 있지 않은 분석 툴을 직접
   만들어서 배포하는 사람들이 매우 많다.
   개발자들이 전세계에 많이 퍼져 있기 때문에 개발되는 툴들이 매우 많으며,
   그에 따라 성능이 탁월하다.
3. 그래픽 기능이다.
   R에서 기본적으로 제공되는 그래픽 출력물의 형태도 훌륭하지만,
   고품질의 그래픽 출력이 가능하다.
4. 운영체계를 가리지 않고 다양한 플랫폼에서 사용할 수 있다는 점이다.
   R은 윈도우는 물론 맥에서도 거의 유사한 기능을 사용할 수 있으며,
   리눅스를 포함한 다양한 형태의 유닉스 기반 컴퓨터에서 모두 설치 및 사용이 가능하다.

가장 큰 단점은 사용자 비친화성이다. R 안에 포함되어 있는 패키지의 개발자들이 다들 통계에 전문가들이서인지, 도움말을 살펴보면 이러한 형태의 프로그램 및 통계 분석에 상당한 경험이 있는 사람들만을 대상으로 쓰여져 있다는 느낌이 강하게 든다.

그래서 각 대학에서 R프로그래밍을 개설하게 되었습니다.

5. RStudio

통계 컴퓨팅, 그래픽스를 위한 프로그래밍 언어인 R을 위한 자유-오픈 소스 통합 개발 환경(IDE)이다. RStudio는 프로그래밍 언어 콜드퓨전의 개발자 JJ Allaire에 의해 만들어졌다. Hadley Wickham은 RStudio의 수석 과학자이다.

0-2. 빅데이터의 정의와 등장 배경

디지털 경제의 확산으로 우리 주변에는 규모를 가늠할 수 없을 정도로 많은 정보와 데이터가 생산되는 '빅데이터(Big Data)' 환경이 도래하고 있다. 빅데이터란 과거 아날로그 환경에서 생성되던 데이터에 비하면 그 규모가 방대하고, 생성 주기도 짧고, 형태도 수치 데이터뿐 아니라 문자와 영상 데이터를 포함하는 대규모 데이터를 말한다.

흔히 우리가 생각하고는 '빅데이터' 라고 하면, 단순히 '방대한 양의 데이터' 라고 생각하게 됩니다. 정확한 '빅데이터'의 뜻은 엄청난 양의 데이터들을 처리(분류)하는 기술' 까지가 정확한 뜻이라고 볼 수 있습니다.

더 정확하게 말하자면 '처리하는 기술'에 초점을 맞추어야 합니다. '처리' 라는 표현에는 데이터들을 수집(Sampling)하고, 수집된 데이터들을 군집화(Clustering) 하여 어떠한 목적에 맞게끔 다시 정제시키는 작업이라고 볼 수 있습니다.

빅데이터

- PC와 인터넷, 모바일 기기 이용이 생활화
- 사용자가 직접 제작하는 UCC를 비롯한 동영상 콘텐츠,
- 휴대전화, SNS(Social Network Service)
- 사진이나 동영상 콘텐츠를 PC를 통해 이용하는 것
- 트위터(twitter)
- 유튜브(YouTube)
- 주요 도로와 공공건물은 물론 심지어 아파트 엘리베이터 안에까지 설치된 CCTV
- 데이터 3법 : 은행, 카드사, 보험사(국민건강보험공단과 민간보험의 자료공유 추진)

NOTE : 개인정보 유출 ???

빅데이터에 평균적으로 사용되는 도구 : R, 파이썬, 하둡

R언어는 정형화된 데이터들을 분석하는 통계 언어입니다. 빅데이터 자료를 기계적으로 처리하기 위해서는 프로그래밍이 필요하며, 데이터 마이닝을 통해 유용한 정보를 추출해내야 하는데, 이 과정에서 데이터를 통계 내고, 해석하고, 분석하고, 의사결정에 이용하기 위해, 사용되는 통계용 언어가 바로 R 언어입니다.

파이썬은 초보자 뿐만 아니라 전문가들까지 사용자층이 매우 다양한 프로그래밍 언어입니다. 오픈소스로 이루어져 있어 접근성이 좋은 것이 특징이며, 문법 자체가 다른 프로그래밍에 비해 어렵지 않다는 장점을 가지고 있습니다.

비정형화되어 있는 데이터들을 유/의미한 데이터로 만드는 작업 후에 다양한 시각적인 그래프로 결과물을 도출해낼 수 있는 것이 바로 이 파이썬 덕분이라고 보시면 됩니다.

하둡은 기존의 관계형 데이터베이스 관리시스템은 고비용이지만, 하둡은 오픈소스로 이루어져 있기 때문에 비용이 거의 들지 않으며, 분산 컴퓨터 방식으로 구축 비용이 저렴하며 데이터 처리 속도가 빠른 장점을 가지고 있는 소프트웨어입니다.

또한 문제가 될 만한 것을 미리 대비하기 위해 매번 하둡 운영한 이후에 결과값을 디스크에 기록하기 때문에, 문제 파악과 해결이 용이하다는 장점이 있습니다.

http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html

클릭해보시면 현재 R이 전 세계에서 사용되는 소프트웨어 중에서 몇 위인지 알 수 있습니다. 1위는 뭘까요???

[과제1] 1. 빅데이터 분석의 개념을 조사해서 정의, 장점, 단점을 서술하시오.

첨부파일 : 학번이름1.hwp **(예 : 2020222260홍길동1.hwp) 또는 word 나 pdf파일**

LMS에 이렇게 올리면 됩니다...

꼭 파일이름을 지켜주세요...

띄어쓰기 안됨....

틀리는 예 : 20202260 홍길동1.hwp, 20202260홍길동과제.hwp

20202260홍길동.hwp, 보고서.hwp, 과제.hwp 안됨)

그래야 제가 저장하면 학번 순으로 출석부 순으로 정리가 됩니다.

혹시나 잘 안되는 학생은 일단 친한 친구한테 일단 물어보고....