Journal of the Korean
Data & Information Science Society
1999, Vol. 10, No. 2, pp. 501 ~ 510

PHP3를 이용한 웹상에서의 통계분석

황진수 [1] · **엄대호** [2]

요 약

컴퓨터의 발달과 더불어 멀티미디어 산업은 급속히 발전하고 있고, 인터넷 또한 폭발적으로 확산되면서 우리의 컴퓨터 환경을 바꾸어 놓고 있다. 통계학 분야에서도 마찬가지로 인터넷을 이용한 기초통계 교육의 필요성이 대두되고 있다. 본 논문에서는 스크립트 언어인 PHP3을 이용하여 웹상에서 동적인 그래프를 통한 기초 자료 분석 및 간단한 검정을 구현하였다. 또한 데이터베이스의 자료와 연동하여 웹상에서의 설문조사 및 결과를 제시하였다. PHP3는 서버에서 수행이 되며 Apache 웹서버에서 모듈형태로 연계되어 있어 기존의 CGI에 비하여 빠른 처리속도를 얻을 수 있는 스크립트언어이며 인터넷상에서 많은 활용을 기대할 수 있다.

주제어: PHP3, 스크립트언어, 통계분포 모듈

1. 서론

최근 컴퓨터의 발달과 더불어 급속히 발달하고 있는 멀티미디어 산업, 그리고 폭발적으로 확산되어지고 있는 인터넷은 우리의 컴퓨터 환경을 바꾸어 놓고 있다. 이러한 발달은 교재와 칠판만으로 진행되어 오던 전통적인 통계학 강의의 형태를 컴퓨터를 활용한 시뮬레이션과 인터넷상에서 진행되는 가상교육까지 포함하는 효과적인 통계교육으로 발전시키는 하나의 해결책을 마련해 주었다. 기존의 통계분석 패키지인 SAS나 SPSS, S-plus, MINITAB 등과 같은 대부분의 통계관련 소프트웨어들은 이용자의 요구에 따라 다양한 결과를 제시해 주고 있으나, 일차적인 통계분석이나 교육과정에 손쉽게 적용하기에는 대용량이며 구입하는데 비용도 많이 들어 통계학을 공부하는 일반학생들과 비전문가들에게 적당한 소프트웨어라고 할 수 없다. 근래에 들어와 SAS나 S-plus와 같은 통계패키지들이 웹상에서 사용자가 손쉽게 쓸 수 있도록 도와주는 프로그램을 개발하여 시판하고 있으나, 이들은 기본적으로 사용자가 해당 소프트웨어를 가지고 있고 사용할 수 있다는 전제하에서 웹

[1](402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과 부교수
[2](402-751) 인천광역시 남구 용현동 253, 인하대학교 통계학과 석사

상에서 인터페이스만 가능케 하는 CGI 프로그램이며 인터넷상에서의 분석으로는 규모가 크고 비용이 많이 드는 단점을 가지고 있다. 이러한 이유에서 경제적이고 통계교육에 보다 적합한 소프트웨어 개발은 필요하다고 여겨진다.

오래 전부터 통계관련 소프트웨어 개발에 대한 필요성은 강조되어져 왔으며, 개발 또한 활발히 진행되어져 왔다. 상업용인 경우에는 특정 분야에 대한 소프트웨어 형태로 개발되고 있으며, 근래에 와서는 인터넷의 보급으로 웹상에서 CGI를 이용한 개발이 활발히 이루어지고 있다. 국내에서도 많은 인터넷 통계프로그램이 자바를 이용하여 개발되고 있으나 자바애플릿을 이용하는 경우 클라이언트에서 수행되므로 이용자의 환경에 따라 영향을 받을 수 있고, 네트웍의 전송속도도 큰 문제가 될 수 있다. 경우에 따라서는 상용소프트웨어는 아니지만 Xlisp-stat과 같은 통계소프트웨어를 클라이언트가 가지고 있어야만 이용할 수 있는 형태로 개발되어지기도 한다.

본 연구에서는 인터넷상에서 여러 형태의 통계적 추론을 할 수 있는 프로그램을 별도의 통계소프트웨어의 도움없이 웹상에서 브라우져만을 가지고 서버 상에서 실행되도록 구현하였다. 동적인 그래프를 통한 탐색적 자료분석, 신뢰구간의 의미를 확인할 수 있는 동적인 그림, t검정과 χ^2검정에 대한 간단한 내용을 우선 구현하여 보았다. 그리고 위 프로그램에 대한 설문조사를 실시하고 MySQL를 이용하여 데이터베이스를 구축한 후, 구축된 자료를 동적인 그래프를 통해 확인할 수 있는 간단한 관계형 데이터베이스 시스템(RDBMS)을 구축해 보았다. 또한 일반통계학 교재에 부록으로 제시되는 다양한 분포의 확률분포표를 웹상에서 제공하였고, 다양한 분포에서 난수를 발생할 수 있도록 하였다. 기본 개발언어인 PHP3은 웹과 여러 데이터베이스 등과 연동이 자유롭고 편리하게 되어 있어 현재 개발된 프로그램에 데이터베이스를 연동시키는 작업이 수월하다. 동적인 그래프는 graphics library 함수를 이용하였고, Apache 웹서버에 PHP3가 모듈(mod_php3)형태로 되어 있어 기존의 CGI 프로그램에 비해 아주 빠른 반응속도를 얻을 수 있다. 이 프로그램은 이미 개발되어진 라이브러리를 손쉽게 적용할 수 있으므로 기존에 개발된 C 언어 통계프로그램이나 새로운 프로그램을 추가하여 다양한 사용자를 위한 웹상의 통계프로그램으로 발전할 수 있으리라 여겨진다. 본 프로그램은 인터넷을 이용할 수 있는 환경이라면 웹브라우저 종류에 관계없이 언제, 어디서라도 사용할 수 있으며 모든 내용을 한글화하여 통계전공자가 아니더라도 결과물을 쉽게 이해할 수 있도록 구성하였다.

2. 프로그램 개발도구, 구조 및 구성내용

2.1 PHP 및 개발시스템의 전체구조

PHP는 본래 Personal Home Page tools의 약자였으나 지금은 PHP is a Hypertext Preprocessor로서 정의하고 있다. PHP는 현재 3.0버전이나 성능과 속도가 개선된 4.0버전(PHP 4.0, 일명 Zend)이 현재 베타버전으로 나와있다. PHP는 웹상에서 여러 가지 CGI 프로그램을 개발하는데 사용되고 있다. 또한 다양한 데이터베이스와 연동이 가능하여 웹상에서 데이터베

이스의 활용에 특히 편리하다. 직접 연동이 가능한 데이터베이스로는 MySQL, PostgreSQL, mSQL, Solid, Sybase, Oracle, Informix 등의 RDBMS(Relational DataBase Management System)가 있고, 그 이외의 데이터베이스는 ODBC(Open DataBase Connectivity)를 이용하여 연동이 가능하다. PHP는 Perl과 마찬가지로 무상으로 인터넷(http://www.php.net)상에서 얻을 수 있고 UNIX용뿐만 아니라 NT용도 무상으로 제공되므로 platform에 상관없이 다양한 웹상의 응용프로그램 개발에 적합한 언어라고 할 수 있다.

자바애플릿이 클라이언트 측에서 실행되는 반면 PHP는 서버 측에서 모든 계산이 이루어져 Apache 웹서버를 통해 사용자에게 전달된다. 텍스트 형태로 전달되므로 자바애플릿에 비해 전송속도는 빠르지만, 반복되는 작업이 많은 대규모 작업에서는 매번 전송이 이루어지므로 사용자측에서 실행이 되는 자바애플릿에 비해 작업속도가 느리다. 그러나 이러한 문제도 PHP4.0에서는 해결이 되어서 큰 규모의 프로그래밍에서도 사용할 수 있는 스크립트언어로 발전되고 있다.

통계계산 부분에 있어서 간단한 사칙연산은 PHP에 내장되어 있는 수학연산자를 사용하였으며, 여러 가지 통계분포의 누적확률값 계산과 난수발생 함수는 Brown, Lovato와 Russell(1994)이 개발한 DCDFLIB(Library of C routines for Cumulative Distribution Function, Inverse and Other Parameters)와 RANDLIB(Library of C routines for Random Number Generation)를 PHP에 링크 시켜 계산하였다. 또한 통계관련 그래프는 Thomas Boutell이 개발한 graphics library 함수
(http://www.boutell.com/gd)를 링크 시켜 여러 가지 형태의 통계관련 그래프를 그리는데 활용하였다.

다음 그림은 Apache 웹서버에 장착된 PHP3 모듈과 데이터베이스 서버와의 연동 관계를 도식화한 그림이다. 여기에서 데이터베이스 서버와 웹서버는 물리적으로 한 시스템에 공존할 수 있고 별도의 서버로서 존재할 수도 있다.

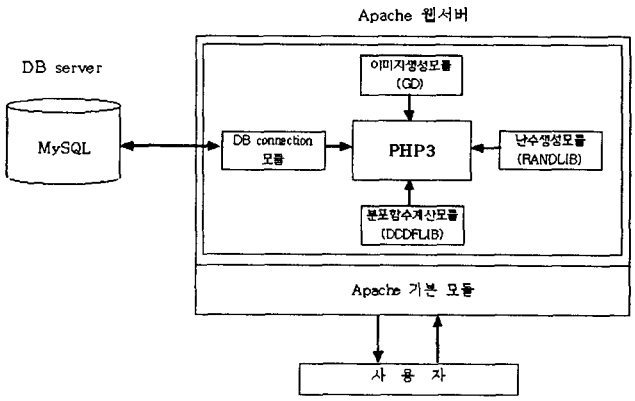

그림 1 : 시스템 구조도

2.2 구성내용

본 논문에서는 인터넷상에서 별도의 소프트웨어 없이 기초통계 교육 프로그램을 PHP3를 이용하여 구현하였다. 현재는 표와 그림을 통한 분석 (EDA), 신뢰구간 시뮬레이션, 두 모집단 비교분석, 범주형 자료분석과 온라인 설문조사 및 결과분석등을 포함하고 있다.

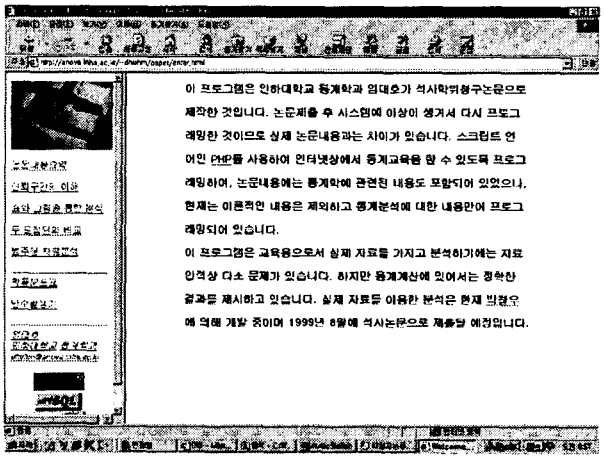

그림 2 : 시작 화면

다음은 위에서 제시된 각각의 프로그램에 대한 자세한 설명이다.

2.2.1 표와 그림을 통한 분석

목록에서 표와 그림을 통한 분석을 선택하면 그림 3와 같은 화면이 제시되고 우측 프레임은 다시 상하 프레임으로 나누어져서 윗부분에 소목록에 제시되어 있다. 소목록은 자료의 형태에 따라 범주형과 연속형으로 나누어져 있으며 범주형의 경우는 돗수분포표와 원형그래프, 막대그래프, 연속형의 경우는 기초통계량과 점도표, 돗수분포표, 히스토그램, 줄기-잎그림, 상자그림 등이 있다. 히스토그램을 예로 살펴보면 그림 3의 하단 화면과 같이 자료를 입력할 수 있는 상자가 제시되면 기본자료가 나타나 있다. 물론 이용자가 제시된 자료를 삭제하고 새로운 자료를 입력하여 히스토그램을 그릴 수도 있다.

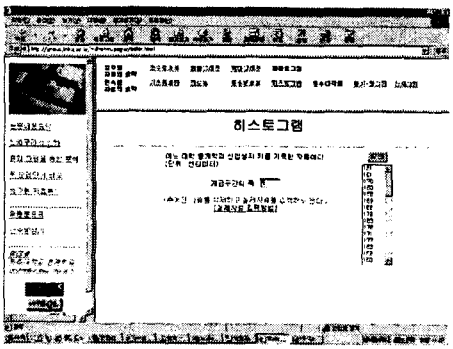

그림 3 : 히스토그램 자료입력

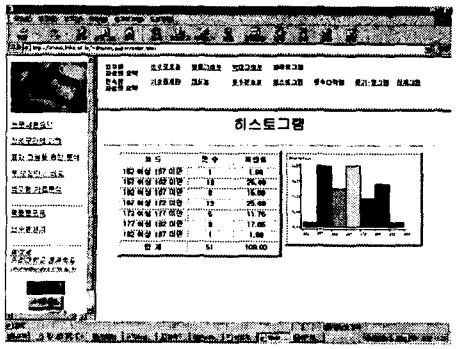

그림 4 : 히스토그램 그리기

2.2.2 신뢰구간의 이해

목록에서 신뢰구간의 이해를 선택하면 그림 5의 화면이 제시되고 실습버튼을 누르면 그림 6과 같은 화면이 제시된다. 그림 5에서는 신뢰구간에 대한 간단한 내용이 서술되어 있고, 그림 6에서는 시뮬레이션을 통해서 이를 실습할 수 있다. 같은 화면 내에서 계속해서 반복 실습할 수 있으며 난수 발생함수를 사용함으로써 매번 다른 그림을 전송 받을 수 있다.

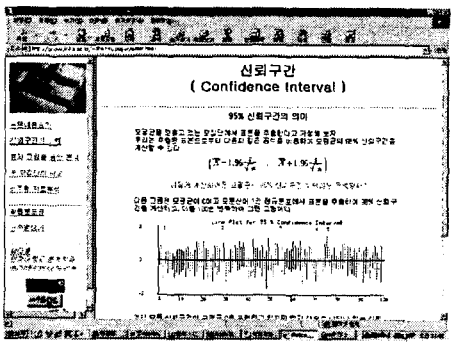

그림 5 : 신뢰구간의 이해 내용

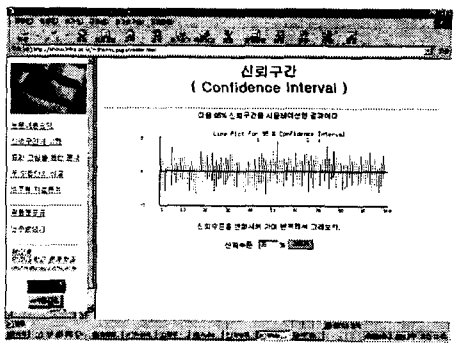

그림 6 : 시뮬레이션

2.2.3 두 모집단의 비교

목록에서 두 모집단의 비교를 선택하면 그림 7과 같은 화면이 제시되고 두 집단의 자료를 입력할 수 있는 상자가 나타난다. 여기서도 히스토그램에서와 마찬가지로 기본자료가 제시되어 있고 새로운 자료 또한 입력이 가능하다. 그림 8은 결과화면을 보여주고 있다. 결과는 표본의 크기가 작은 경우에 필요한 가정과 평균과 표준편차 등의 기초통계량, 등분산과 이분산 경우의 t검정 결과, 등분산가정에 대한 F검정결과 등이 제시되어 있다. 모든 검정의 결과는 검정통계량과 함께 유의확률값(p-value)을 제공하여 사용자가 결정을 할 수 있도록 하였다.

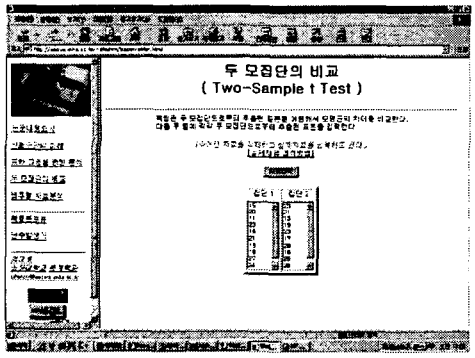

그림 7 : 두 모집단의 비교 자료입력

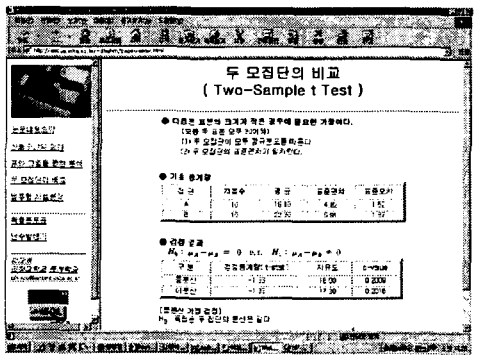

그림 8 : 두 모집단의 비교 결과

2.2.4 범주형 자료분석

목록에서 범주형 자료분석을 선택하면 그림 9과 같은 화면이 제시된다. 두 모집단의 비교와 마찬가지로 자료를 입력할 수 있는 상자가 있고 더불어 체크박스가 오른쪽에 제시되어 있다. 체크박스는 관측돗수와 퍼센트, 열 퍼센트, 행 퍼센트 등의 교차표를 만들 때 필요한 통계량과 기대값과 편차, 셀카이제곱값, 카이제곱검정 등의 χ^2검정에 필요한 통계량으로 구성되어 있어, 원하는 통계량만 선택하여 계산할 수 있게 하였다. 체크박스를 선택하게 되면 해당되는 통계량은 그림 10 화면에 나타나게 된다. χ^2검정에서와 마찬가지로 검정통계량과 함께 유의확률값(p-value)이 제시된다.

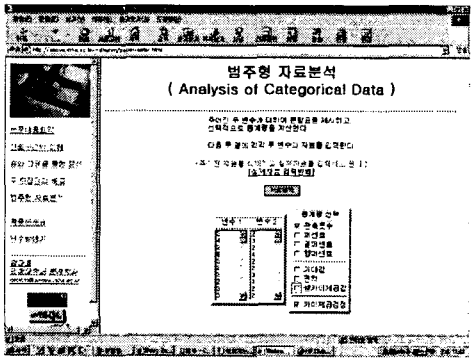

그림 9 : 범주형 자료분석 자료입력

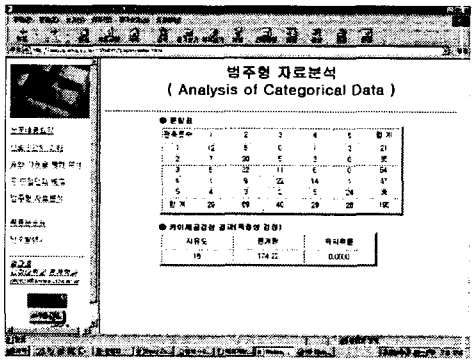

그림 10: 범주형 자료분석 결과

2.2.5 설문조사

일반적으로 개발되어지는 설문지 형태를 그림 11과 같이 HTML형식으로 구성한다. 응답자는 설문을 모두 마치면 설문입력을 위한 버튼을 누르게 되고 다음 화면에서는 응답한 설문을 확인하게 된다. 확인 후 자료입력 버튼을 누르면 MySQL로 구축된 데이터베이스에 입력이 된다. 누적된 설문자료를 원그림과 막대그림을 통해 결과를 확인할 수 있다. 그림 12은 어떤 문항에 대해 막대그림을 그림 결과이다.

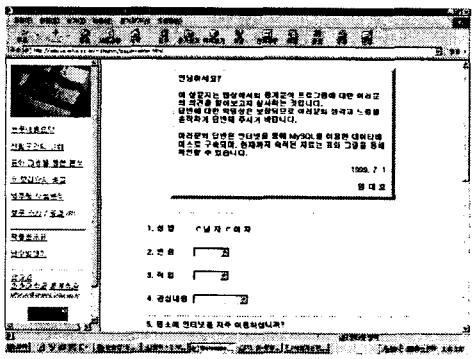

그림 11 : 설문지

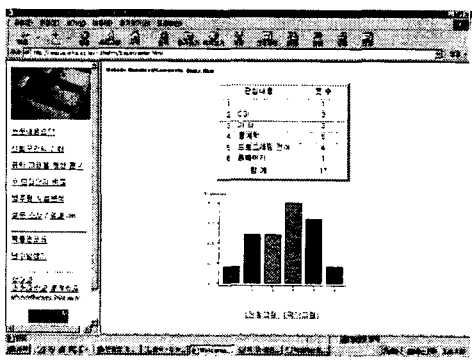

그림 12 : 설문결과

데이터베이스 구축과 구축된 데이터베이스의 활용은 모두 PHP에 내장되어 있는 MySQL를 위한 함수를 사용함으로써 손쉽게 프로그래밍할 수 있다. 여기에서 이용된 RDBMS인 MySQL을 비상업용인 경우에 무상으로 사용할 수 있는 프로그램으로서 안정성과 속도 면에서 가장 좋은 평가를 받고 있으며 PHP와 연동하여서 가장 많이 쓰이는 데이터베이스 프로그램이다.

3. 결 론

본 논문에서는 스크립트 언어인 PHP3를 이용하여 웹상에서 동적인 그래프를 통한 기초적 자료분석및 간단한 검정과 데이터베이스를 활용한 웹상의 설문조사를 구현하였다. 신뢰구간의 의미를 눈으로 확인할 수 있도록 난수발생함수(RANDLIB)를 이용하여 동적인 신뢰구간 그래프를 구현하여 보았고 통계분석에 필요한 유의확률(p-value)은 DCDFLIB를 PHP3에 접목시켜 구하였다. 동적인 그래프는 graphics library 함수를 이용하였으며, 설문조사의 경우에는 MySQL 데이터베이스를 이용해 RDBMS를 구축하였다.

RANDLIB는 정규분포뿐만 아니라 통계학에서 쓰이는 거의 모든 분포함수를 포함하고 있으므로 신뢰구간뿐만 아니라 이항분포의 포아송분포 수렴을 포함하는 중심극한정리등 이론적으로는 설명하기 어려운 내용을 시뮬레이션을 통해 구현할 수 있으나 본 논문에서는 t검정과 χ^2검정만 구현 해 보았다. 그러나 분산분석이나 회귀분석 등 많은 통계분석분야에 응용할 수 있으며 새롭게 확장하는 것도 어렵지 않으리라 기대된다. 데이터베이스를 연동한 프로그램은 설문조사의 방법으로 간단하게 소개하였지만, 사용자가 실제자료를 입력하여 데이터베이스로 구축한 후 시간을 가지고 자료분석을 하는 간단한 통계패키지 수준의 홈페이지를 구축할 수 있다. 이것을 응용한다면 사용하기도 어렵고 몸집만 큰 기존의 통계패키지를 대신해서 사용자 수준과 내용에 맞는 다양한 통계처리 서비스를 인터넷상에서 제공할 수 있을 것으로 기대되어진다.

참 고 문 헌

1. 구자홍 외 6 (1997). 통계학 -미니탭을 이용한 분석-, 자유아카데미, 서울.

2. 김영훈 (1996). 알기 쉬운 CGI 활용, 정보문화사, 서울.

3. 백영균, 설양환 (1997). 인터넷과 교육, 양서원, 서울.

4. 안기수, 허문열 (1997). 인터넷을 이용한 통계 교육과 컨설팅의 현황, 한국통계학회 논문집, 제4권 2호, 473-489.

5. 안기수, 허문열 (1998). 멀티미디어와 통계 소프트웨어를 활용한 회귀분석 학습 시스템, 응용통계연구, 제11권 2호, 389-401.

6. 이정진, 강근석, 이윤오 (1992). 통계학 교육용 한글 소프트웨어 개발 연구, 응용통계연구, 제5권 1호, 81-91.

7. 한경수, 안정용 (1996). 저작도구를 이용한 통계교육용 멀티미디어 소프트웨어 개발 연구 -주사위 게임과 카드 게임-, 응용통계연구, 제9권 2호, 73-82.

8. Brown, B. W. and Lovato, J. (1994). *RANLIB.C. Library of C Routines for Random Number Generation.* Department of Biomathematics, M.D. Anderson Cancer Center, University of Texas, Houston.

9. Brown, B. W., Lovato, J. and Russell, K. (1994). *DCDFLIB.C. Library of C Routines for Cumulative Distribution Functions, Inverse, and Other Parameters.* Department of Biomathematics, M.D. Anderson Cancer Center, University of Texas, Houston.

Statistical Analysis on the Web Using PHP3

Jinsoo Hwang[3] · Daiho Uhm[4]

Abstract

We have seen a rapid development of multimedia intustry as computer evolves and the internet has changed our way of life dramatically in these days. There are several attempts to teach elementary statistics on the web but most of them are based on commercial products. The need for statistical data analysis and decision making based on those analysis is growing. In this article we try to show one way of reaching that goal by using a server side scripting language PHP3 toghether with extra graphical module and statistical distribution module on the web. We showed some elementary exploratory graphical data analysis and statistical inferences. There are plenty of room of improvements to make it a full blown statistical analysis tool on the web in the near future.

All the programs and databases used in our article are public programs. The main engine PHP3 is included as an apache web server module so it is very light and fast. It will be much better when the PHP4(ZEND) will be officially out in terms of processing speed.

Key Words and Phrases: PHP3, GD, statistical distribution module

[3]Associate Professor, Department of Statistics, Inha University, 402-751, Inchon, Korea
[4]Department of Statistics, Inha University, 402-751, Inchon, Korea