Journal of Korea Multimedia Society Vol. 8, No. 7, July 2005(pp. 1000-1011)

가이드 맵과 인터랙티브 시각화를 이용한 의료 통계분석 시스템

이돈수[†], 최수미[††]

요 약

본 논문에서는 통계에 대한 지식이 부족한 임상 의학자들이 보다 쉽고 정확하게 데이터를 분석할 수 있도록 표본 데이터의 분포에 따라 적절한 분석 방법을 제시해주고, 분석 과정을 아이콘들의 트리로 구성한 가이드 맵을 제공하는 의료 통계분석 시스템을 개발하였다. 개발 시스템은 일반적으로 활용되는 통계 방법, 반복측정 자료에 활용되는 통계 방법, 생존분석 등 의료 분야에서 자주 사용되는 분석법들을 포함하고 있다. 또한 3차원 글리프를 이용하여 결과를 인터랙티브하게 보여주고, 불확실성을 시각화함으로써 분석된 결과를 더욱 쉽게 이해할 수 있도록 하였다.

A System for Medical Statistical Analysis Using Guide Maps and Interactive Visualization

Don-Soo Lee[†], Soo-Mi Choi[††]

ABSTRACT

This paper presents a system for medical statistical analysis that helps medical professionals analyze clinical data more easily and accurately. It is able to recommend proper methods according to the distribution of sample data, and provides guide maps composed of icons for the understanding of the process of analysis. Besides general statistical analysis, it includes commonly-used statistical methods for medical fields, such as survival analysis and methods for repetitive measurements. The results of analysis are interactively displayed by 3D glyph-based visualization with uncertainty.

Key words: Medical Statistics(의료 통계), Visual Guide(비쥬얼 가이드), Data Visualization(데이터 시각화), Uncertainty Visualization(불확실성 시각화)

1. 서 론

임상에서 환자를 대하면서 많은 시간과 노력을 기울여 얻어진 관찰 결과 혹은 실험을 통해서 얻어진 의료 데이터를 정확하게 분석하는 것은 매우 중요하다. 대부분의 임상 의학자들은 SAS™나 SPSS™와 같은 전문적인 통계 소프트웨어들을 사용하거나 통계 전문가에게 의뢰하여 분석하는 경우가 많다. 그러나 비전문가가 SAS™나 SPSS™와 같은 통계 소프트웨어를 배우는 데는 많은 시간이 걸리며 전문적인 통계 지식의 부족으로 실험이 잘못 분석되거나 결과 해석에 어려움을 겪는 경우가 많다[1,2]. 그러므로 통계에 비전문가인 임상 의학자들이 데이터를 쉽고 정확하게 분석할 수 있도록 자주 사용하는 의료 통계분석 기능을 제공하고, 처리 과정 및 결과를 쉽게 이해할 수 있도록 해주는 인터페이스 및 시각화에 대한 연구가 필요하다.

※ 교신저자(Corresponding Author) : 최수미, 주소 : 서울시 광진구 군자동 98(143-747), 전화 : 3408-3754, FAX : 3408-3662, E-mail : smchoi@sejong.ac.kr
접수일 : 2004년 1월 26일, 완료일 : 2005년 1월 14일
[†] 준회원, 세종대학교 컴퓨터공학부
(E-mail : chunshanghwa@hotmail.com)
[††] 정회원, 세종대학교 컴퓨터공학부 조교수
※ 본 연구는 보건복지부 보건의료기술 연구개발사업의 지원에 의하여 이루어진 것입니다.(02-PJI-PG3-51312-0003).

이와 관련된 연구들을 살펴보면, 먼저 의료 통계분석 시스템인 MedCalc™[3]을 들 수 있다. MedCalc™는 의료 통계분석에 특화된 시스템으로 의학 분야 실험에서 자주 사용되는 분석법을 중점적으로 만들어진 것이 특징이다. 메뉴와 셀 작업표의 배치가 범용 통계 분석 시스템 보다 의료 분야에 편리하도록 잘 정리되어 있으며, 다양한 통계 그래프들을 생성하는 기능도 제공하므로 사용하기에 편리하다. 그러나 통계 지식이 부족한 사용자가 잘못된 데이터를 입력할 경우 문제점을 미리 지적해주는 등의 올바른 통계 분석 방법을 제안하는 가이드 기능을 제공하지는 않고 있다.

다른 관련 시스템으로는 셀 작업표 상의 수치 데이터간의 상관관계 또는 수식을 그래프로 보여주는 Natto View™[4]를 들 수 있다. Natto View™는 스프레드 시트(spread sheet) 상의 셀 데이터 간에 숨겨져 있는 상관 관계를 삼차원으로 보여주는 기능을 제공한다. 포커스된 셀이 위로 올려짐에 따라 단계 구조로 셀 데이터 간의 수식과 상관 관계가 사용자에게 보여지게 된다. 이러한 방식은 간단한 수식 관계를 이해하는 데는 적합 하지만 복잡한 삼차원적 셀 관계의 경우 시각적으로 구분하기에 어려운 단점이 있다. 또한 셀 간의 관계 위주의 시각화로 분석 과정을 포함하고 있지는 않다.

ViSta™[1]는 통계 지식이 있는 사용자 뿐만 아니라 비전문가를 대상으로 설계된 시스템이다. Vista™의 특징은 구조화된 그래픽 인터페이스로 WorkMap과 GuideMap을 제공하고 있다는 점이다. WorkMap은 데이터 분석 과정을 정리하여 보여주고, GuideMap은 통계 지식이 부족한 초보자들에게 올바른 분석을 할 수 있도록 가이드 해 주는 역할을 한다. 이렇게 ViSta™는 분석 과정을 쉽게 이해할 수 있도록 중간 단계의 인터페이스 개념을 제시하고 있다. 그러나 범용 통계분석을 지원하고자 개발된 시스템으로 상용 통계 시스템과 마찬가지로 초보자가 배우는 데에는 복잡하다. dBSTAT™[16]는 통계지식이 부족한 일반인을 위해 설계된 통계 소프트웨어로서 통계 마법사 기능을 통해 간편한 분석을 할 수 있도록 하였다. 그러나 분석법의 적절성 평가를 해주는 기능은 없다.

본 논문에서는 의료 통계에 자주 사용되는 분석법들을 쉽게 사용할 수 있도록 분석 과정을 아이콘들의 트리로 구성한 가이드 맵을 제공하고, 표본 데이터의 갯수 및 분포에 따라 적절한 분석 방법을 가이드 해주는 의료 통계분석 시스템을 개발하였다. 또한 의료 분야에서 자주 사용되는 2차원 그래프 이외에 3차원 글리프 (glyphs)[5,6]를 이용하여 데이터 결과를 인터랙티브하게 볼 수 있도록 하고 통계 방법에 내재한 불확실성을 가시화하는 방법을 개발하였다.

이어지는 논문의 구성은 다음과 같다. 2장에서는 개발된 의료 통계분석 시스템의 특징을 설명하고, 전체 구성 및 주요 모듈에 대하여 기술한다. 3장에서는 구현된 결과를 제시하고 마지막으로 4장에서는 결론을 맺는다.

2. 의료 통계분석 시스템 설계

2.1 시스템의 특징 및 구성

모집단의 분포 형태와 데이터의 특성에 따라 통계 분석 방법을 신중히 선택해야 정확한 결과를 예측할 수 있음에도 불구하고, 통계 분석에 대한 지식이 부족한 사용자들은 종종 자주 사용하거나 쉬운 분석 방법을 적용하여 결과를 도출하곤 한다.

이러한 문제점을 보완하기 위하여 개발 시스템은 다음과 같은 특징을 갖는다. 첫째, 아이콘 트리를 사용한 가이드 맵은 데이터 입력에서 결과 해석까지의 분석 과정을 보여준다. 이때 분석 진행 과정에 대한 텍스트 해설 기능을 포함하며, 분석이 끝나면 텍스트 보고서와 그래프로 분석 결과를 해석해 준다. 이는 통계 지식이 부족한 사용자들이 잘못된 분석 방법을 사용하는 것을 방지하도록 하며, 진행 과정을 시각적으로 보여주어 이해를 증진 시키는 기능을 한다.

둘째, 프로그램이 시작될 때 위저드 형태의 다이얼로그를 통해 사용자가 선택한 데이터 분포 및 데이터 형에 따라 적합한 분석 방법을 제시해 준다. 이와 함께 분석 도중에 데이터 무결성, 데이터 형, 적합도 검사를 통해 올바르지 못한 결과가 예상될 경우 경고 메시지를 출력하고 올바른 분석을 할 수 있도록 해준다.

셋째, 통계 분석 시 나타나는 잔차와 같은 불확실한 데이터를 3차원 글리프를 이용하여 시각화함으로써 의사결정에 도움을 주고, 처리 결과 역시 3차원 인터랙티브 시각화 방법을 통하여 결과 해석을 직관적으로 할 수 있도록 해준다.

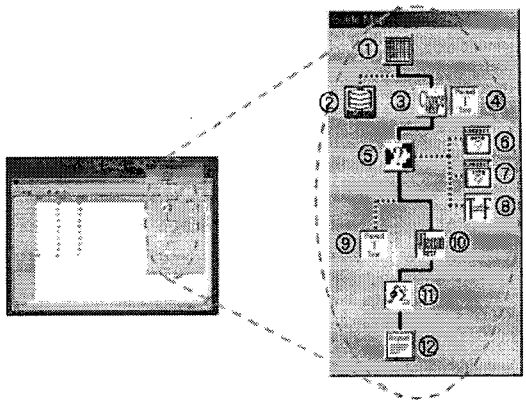
넷째, 본 시스템을 통하여 통계 처리한 내용에 대하여 SAS™ 코드 생성을 원하는 경우 자동으로 코

드를 생성해 준다. 이 기능을 통해 SAS™를 모르는 사용자도 SAS™ 소프트웨어에서 제공하는 기능을 쉽게 이용할 수 있다.

전체 시스템의 구성도는 그림 1과 같다. 통계분석 시스템은 의료 데이터의 분석을 위해서 필요한 일반적인 통계 방법, 반복측정 자료에 활용되는 통계 방법, Kaplan-Meier 생존곡선과 같이 의료 분야에서 자주 사용되는 통계 방법들을 포함한다. 데이터 입력 형식은 실제로 병원에서 널리 사용되고 있는 데이터베이스인 MS Access™ (.MDB) 형식을 따르고, Excel™로 작성된 데이터도 복사 기능을 제공함으로써 쉽게 불러들여 사용할 수 있도록 하였다. 비쥬얼 가이드 시스템은 데이터의 적합도 평가 모듈의 검사를 거쳐 부적절성이 발견되면 경고 메시지를 출력하여 실수를 방지하고, 전체 처리 과정을 아이콘들의 트리로 표현한 가이드 맵을 생성하는 기능을 한다. 인터랙티브 시각화 시스템은 분석법별 오차 계산 모듈을 통한 불확실성 시각화[7]와 3차원 글리프 기반의 인터랙티브 시각화[8,9]를 제공함으로써 분석법의 적절성을 평가하는 의사결정을 돕는다. 통계 처리 과정이 끝나면 결과해석 모듈에서 텍스트, 2D/3D 그래프, SAS™ 코드를 제공한다.

2.2 비쥬얼 가이드

비쥬얼 가이드 시스템은 진행되는 검정법에 데이터가 적합한지를 검사하는 부분과 분석 단계를 시각화하는 부분으로 나누어진다. 그림 2는 셀 작업표[10]에 입력된 데이터와 가이드 맵 윈도우를 보여준다. 오른편의 가이드 맵은 내부적으로 수행한 적합도 검사 결과, 현재 분석이 진행되고 있는 경로, 시스템에서 제안하는 경로를 보여줌으로써 분석 과정을 쉽게 이해할 수 있도록 해준다. 가이드 맵에서 점선은 지나온 경로를 나타내고, 실선은 현재 진행되는 경로를 나타낸다. 가이드 맵은 아이콘을 사용하여 텍스트만을 사용할 때 보다 단계별 처리 기능을 훨씬 더 쉽게 기억할 수 있게 해준다. 또한 셀 작업표가 항상 보이도록 반투명 윈도우를 사용하고 필요하지 않은 경우에는 화면에 보이지 않도록 할 수 있다.



① 셀작업표 아이콘, ② 데이터베이스 아이콘, ③ 통계분석 선택 아이콘, ④ Paired T-test 아이콘, ⑤ 적합도 검사 아이콘, ⑥ 데이터 무결성 검사 아이콘, ⑦ 데이터 형 검사 아이콘, ⑧ 통계 가정 성립 검사 아이콘, ⑨ Paired T-test 아이콘, ⑩ 비모수적 통계방법 아이콘, ⑪ 계산중 아이콘, ⑫ 결과해석 아이콘

그림 2. 비쥬얼 가이드 맵

가이드 맵에 나타나는 아이콘들은 컨트롤이 가능

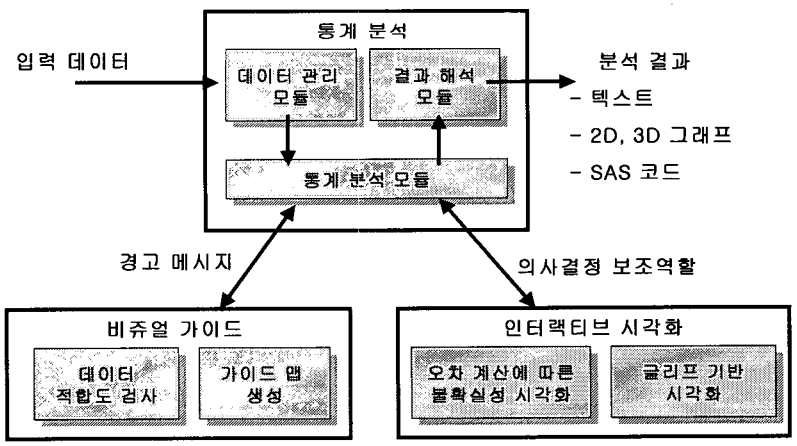


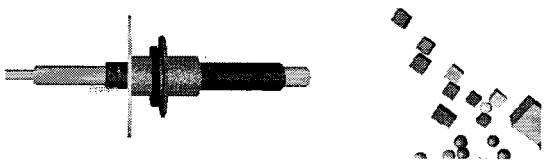
그림 1. 의료 통계분석 시스템의 구성도

한 아이콘들과 현재 상태 표시를 위한 아이콘들로 분류된다. 컨트롤 아이콘은 사용자가 아이콘을 클릭함으로써 실행이 가능한 아이콘으로 셀작업표 아이콘, 데이터베이스 아이콘, 통계분석 선택 아이콘, 결과해석 아이콘 등이 있다. 상태 아이콘은 작업의 진행 상황을 표시해 주는 아이콘으로 테이블 선택 아이콘, 필드 선택 아이콘, 선택된 분석 방법 표시 아이콘, 분석 중 아이콘 등이 있다.

검사부분은 데이터 무결성 검사, 데이터 형 검사, 분석 방법의 적합도 검사로 분류된다. 데이터 무결성 검사는 셀 작업표 중간에 공백값 같은 손실 값이 없는지를 검사한다. 즉, 데이터가 없을 경우 "." 을 입력하고 통계 결과에 반영되지 않도록 한다. 데이터 형 검사는 데이터가 현재 선택된 분석법에 적합한 형인지를 검사한다. 마지막으로 분석 방법의 적합도 검사는 데이터의 분포 형태와 표본 수를 검사하여 진행중인 분석법의 가정에 적합한지를 평가 한다. 위의 세 가지 검사 중 어느 한 가지가 부적합할 경우 경고 메시지를 출력하여 사용자의 실수를 방지하도록 하였다.

2.3 인터랙티브 시각화

글리프란 위치, 형태, 색, 투명도와 같은 속성을 갖는 그래픽 객체를 의미한다. 이러한 속성들은 데이터의 여러 차원을 동시에 표현함으로써 다변량 데이터의 시각화에 많이 사용되어진다[6,9,11]. 그림 3(a)는 평균, 분산, 구간의 통계적 의미의 정보를 3차원 글리프로 보여주는 예이다. 가장 큰 링은 데이터의 평균을 의미하고, 중간 링은 그룹의 평균, 가장 작은 링은 그룹의 중앙값을 의미한다. 각 튜브들은 1-3 quartile, 1-9 decile, min-max intervals을 나타낸다. 이러한 글리프 객체들은 시점 변환이나 수치 변환과 같은 인터랙티브한 조작이 가능하여 데이터 분석이 용이하다. 다변량 데이터의 가장 간단한 표현 방법으로 그림 3(b)와 같은 플레닛(planets)을 들 수 있다.



(a) Tukey Glyph (b) Planet Glyph

그림 3. 3차원 글리프

세 개의 선택된 변수들은 공간상의 x, y, z 위치로 표현되어지고, 추가적인 정보는 유닛을 표현하는 글리프에 의해서 나타낼 수 있다. 다양한 뷰는 데이터 간의 관계를 쉽게 파악할 수 있도록 해준다.

통계 분석법에 데이터를 적용시킬 때 산출되는 잔차(residual)는 분석법의 적절성을 평가하기 위해 필요한 정규성, 등분산성 평가를 위해 유용하게 쓰여질 수 있다[13,14]. 단순 선형회귀 분석법을 예로 들면 잔차는 $e_i = y_i - y'_i$ 로서 오차(error)라고도 불리운다. 또 다른 방법은 종속변수의 총 변동 중 회귀모형에 의해 설명되는 변동 비율을 살펴보는 것이다. 회귀직선에 의해 종속변수의 총 변동 정도를 나타내는 측도는 (식 1)과 같고, 이를 결정계수(coefficient of determination)라 한다.

$$R^2 = \frac{SSR}{SST} \tag{1}$$

결정계수의 범위는 $0 < R^2 < 1$ 로서 만약 모든 측정값들이 회귀직선상에 위치하여 회귀직선에 의한 변동 SSR이 총 변동 SST와 동일하게 되면 $R^2 = 1$ 이 된다. 그러나 뚜렷한 회귀직선 관계가 없어 추정된 회귀직선의 기울기가 0에 가깝게 되면 $y' = \bar{y}$ 로써 SSR이 0에 가깝고 따라서 R^2도 0에 가깝다. R^2의 값이 크면 클수록 유용한 회귀직선이며, R^2의 값이 작을 때에는 비록 회귀직선이 매우 유의하다는 검정 결과가 나왔다 하더라도 추정된 회귀직선은 종속변수를 제대로 설명하지 못하므로 유용하다고 할 수 없다.

일반적으로 어떤 자료에 대해 회귀모형이 적절한지 알 수 없으면서 회귀모형을 우선 자료에 적합시키는 경향이 있다. 한편, 회귀모형의 적절성을 잔차(residual)로 부터 검토할 수가 있는데, 구체적으로 단순회귀모형에서 요구되는 직선관계, 정규성, 독립성, 등분산성의 가정을 잔차(residual)로 부터 검토한다[14]. 그림 4는 잔차(residual) $e_i = y_i - y'_i$를 회귀직선상의 추정값 y'_i에 대해 그려본 전형적인 네 가지 형태를 보여준다. 그림 4 (a)는 잔차(residual)들이 0을 중심으로 수평대(horizontal band)의 형태를 지니며 오차항의 등분산 가정이 성립됨을 보여준다. 그러나 그림 4(b)는 y'_i가 증가함에 따라 e_i의 값들이 넓은 폭으로 흩어지므로 오차항의 등분산 가정이 성립되지 않는다. 그림 4(c)는 y'_i 값이 증가하면서 e_i의 값들

이 음의 값에서 양의 값으로 변하다가 다시 음의 값을 가진다. 이와 같은 그림은 회귀모형이 직선식 보다는 2차 곡선식을 가정해야 함을 말해준다. 마지막으로 그림 4(d)는 잔차(residual)를 시간축에 대해 그린 경우이며, 이는 오차항의 독립성이 성립되지 않음을 나타낸다. 그림 4(d)에서 보면 시간에 따른 효과가 직선식으로 변해가므로 시간이란 독립변수가 회귀모형에 추가되어야 함을 말해준다.

본 연구에서는 잔차(residual)를 통계적 범주의 불확실한 데이터[7]로 보고 사용자가 현재의 잔차(residual) 모델과 일반적으로 잘 알려진 오류 모델과 비교 검토해 봄으로써 분석법의 신뢰도를 평가할 수 있는 기능을 추가하였다. 이 때 3차원 글리프를 이용하여 사용자가 잔차(residual) 데이터를 쉽게 이해할 수 있도록 하였다. 모든 분석법에 이 방법을 적용하기에는 무리가 있으나 불확실성을 내포하고 있는 분석법의 경우에 유용하게 사용될 수 있다.

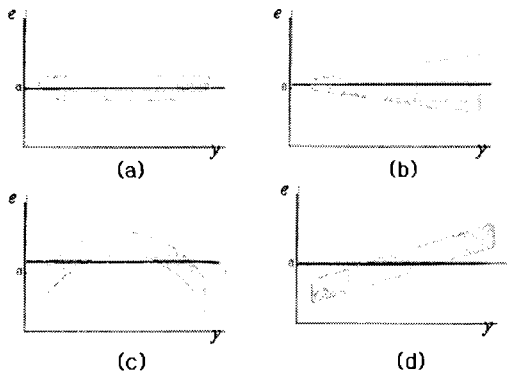


그림 4. 전형적인 잔차(residual) 형태

2.4 통계 분석 및 결과 해석

임상 데이터에 자주 사용되는 통계 방법들은 크게 일반적으로 활용되는 통계 방법과 반복 측정 자료에 활용되는 통계 방법으로 분류 할 수 있다[13]. 일반적으로 활용되는 통계 방법에는 쌍을 이룬 자료의 T-검정, 쌍을 이루지 않은 자료의 T-검정, 상관분석, 단순선형 회귀분석, 다중선형 회귀분석, 일원 배치법, 이원 배치법, 랜덤화 블록법, 카이제곱 검정, 생존분석 등이 있다. 반복 측정 자료에 활용되는 통계 방법에는 일변량 분석, 다변량 분석, 한 모집단 한 반복요인, 여러 모집단 한 반복요인, 여러 모집단 두 반복요인 등이 있다. 또한 통계적 분석법은 크게 모수적(Parametric) 방법과 비모수적 방법으로 나누어질 수 있는데, 일반적으로 많이 쓰이는 방법은 모수적 방법의 범주에 속한다. 비모수적 방법의 큰 장점은 그 이름이 암시하듯이 모집단의 분포형태에 대해 가정하지 않으므로, 데이터가 정규분포 가정이 성립되지 않을 경우에 안전하게 사용될 수 있다. 가장 일반적으로 쓰이는 비모수적 통계검정법으로 Wilcoxon 순위합 검정법과 부호순위 검정법, Kruskal-Wallis 검정법을 들 수 있다.

본 연구에서는 이러한 통계 분석 모듈을 구현하였고, 단순 선형 회귀 분석, 상관분석 등의 데이터에 대한 통계처리 결과를 해설 또는 다양한 그래프를 통하여 보여 준다. 또한 Kaplan-Meier 생존곡선과 같이 의료용에 특화된 내용을 그래프 형태로 제공함으로써 사용자로 하여금 쉽게 분석할 수 있도록 하였다. 이외에도 본 시스템에서는 SAS™ 사용법을 모르더라도 결과를 인용할 수 있도록 SAS™ 프로그램 코드를 결과 해설과 함께 제공하였다.

3. 구현 및 결과

3.1 비쥬얼 가이드 결과

개발 시스템은 Windows 2000 운영체제 하에 비쥬얼 C++ 언어를 이용하여 구현하였다. 구현 결과의 예로 다음은 정신 안정제의 효과를 알아보기 위하여 쌍을 이룬 자료의 T 검정을 데이터에 적용한 결과이다.

실험 예>

신경증 환자를 치료하는 새로운 정신안정제의 효과를 알아보기 위하여 환자마다 일주일은 안정제를 투여하고, 다른 일주일 동안은 가짜약(placebo)을 투여하는데 그 순서는 랜덤하게 정하였다. 주말마다 환자에게 질문표를 배부하여 그 응답을 토대로 안정점수를 매겼다(0~30). 점수가 높을수록 불안감이 심한 경우이다. 본 시스템을 이용하여 안정제의 효과가 가짜약에 의한 효과와 차이가 있을 것이라 예상하고 두 종류의 표본 데이터(표 1, 표 2)에 대하여 쌍을 이룬 자료의 T-검정(paired T-test)을 적용해 보고자 한다.

먼저 데이터 작성 방법은 필드에 직접 입력하는 방식과 데이터베이스에서 파일을 불러오는 방식 중

표 1. 표본 데이터가 정규분포인 경우

환자	1	2	3	4	5	6	7	8	9	10
안정제(x)	19	11	14	17	23	11	15	19	11	8
가짜약(y)	22	18	17	19	22	12	14	11	19	7

표 2. 표본 데이터가 정규분포가 아닌 경우

환자	1	2	3	4	5	6	7	8	9	10
안정제(x)	19	17	14	17	23	13	15	10	11	8
가짜약(y)	22	18	17	19	22	12	14	11	19	7

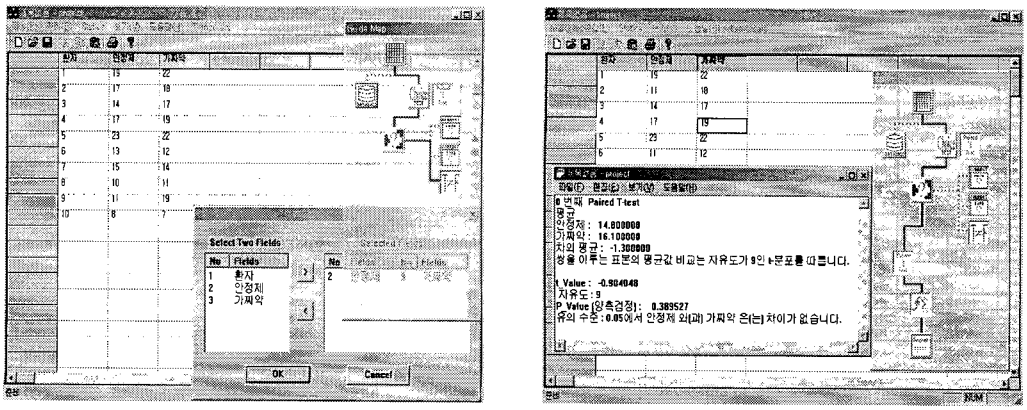
에서 선택할 수 있다. 쌍을 이룬 자료의 T-검정은 실행하기 전에 "모집단이 정규분포 한다"에 대한 검정이 필요한데, 처리하는 표본의 크기가 큰 경우(n>30)에는 정규근사를 이용할 수 있지만 표본의 크기가 크지 않은 경우에는 Shapiro Wilk's Test를 사용하여 모집단이 정규분포를 하는지를 검사하여야 한다. 본 시스템에서는 우선적으로 데이터에 유실이 있는지를 검사한 후 필드의 형을 검사한 다음 문제가 있으면 에러 메시지를 발생한다. 다음으로는 Shapiro Wilk's Test 를 통해 데이터가 정규분포를 하는지 검사한다. 표 1의 데이터와 같이 정규분포를 이루고 있는 경우에는 그림 5와 같이 정상적으로 쌍을 이룬 자료의 T-검정을 수행하게 된다. 즉, 양측 검정으로 구한 p_value의 값이 0.389527이고 "유의수준 0.05에서 안정제와 가짜약은 차이가 없다"라는 결론을 얻을 수 있다.

만일 표 2의 데이터를 쌍을 이룬 T-검정을 적용하려고 하는 경우에는 일단 이 분석법이 가정하고 있는 "모집단이 정규분포를 이루어야 한다"라는 가정에 적합하지 않기 때문에 모집단이 정규분포임을 가정하지 않는 비모수적 방법을 사용하여야 한다. 정규분포가 아닌 표본 데이터에 대하여 T-검정을 적용하려는 경우에는 Shapiro Wilk's 통계량에 대해서 모집단이 정규분포를 하지 않는다는 결과가 나오고 표본의 크기가 30을 넘지 않을 경우 쌍을 이룬 T-검정을 사용하면 결과 값에 오차가 크게 나올 수 있으므로 모집단이 정규분포임을 가정하지 않는 비모수적 방법을 사용하여야 한다. 그러므로 본 시스템에서는 비모수적 방법 중에서 쌍을 이룬 T-검정에 해당하는 Wilcoxon 부호 순위 검정[13,14]을 이용하여 결과를 제공한다.

그림 6(a)는 사용자가 데이터 분석을 위해 선택한 방법에 대해 표본 데이터들이 정규분포 하지 않음으로 보다 적합한 비모수적인 분석 방법인 Wilcoxon 부호순위 검정 방법을 추천한 결과 화면이다. 그림 6(b) 에서는 T=15.000 값을 얻을 수 있는데 자유도 9일때, 유의수준 0.05에서 기각치에 해당하는 최대 T값은 10.0 이므로 "안정제와 가짜약은 차이가 없다"라는 결론을 얻을 수 있다.

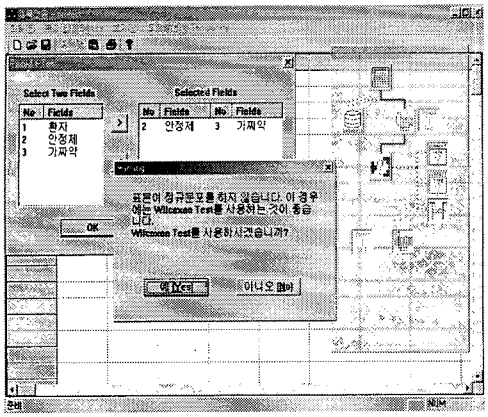
3.2 인터랙티브 시각화 결과

일반적으로 통계 분석에는 2차원 그래프가 사용되는데[15], 3차원 그래프를 사용하면 더 이해하기 쉬운 경우가 있다. 그림 7은 다중선형 회귀분석의 3

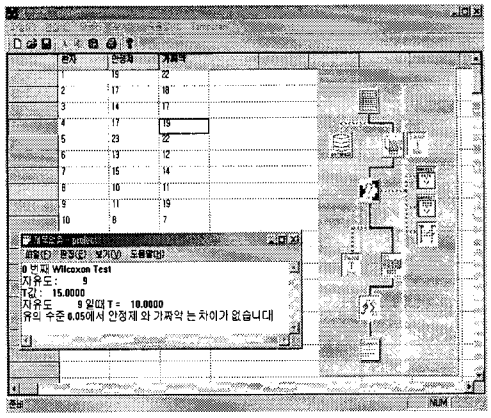


(a) 데이터의 정규분포 여부 검사 (b) 분석 결과

그림 5. 쌍을 이룬 자료의 t-검정 결과



(a) 적합한 통계 분석 방법 가이드



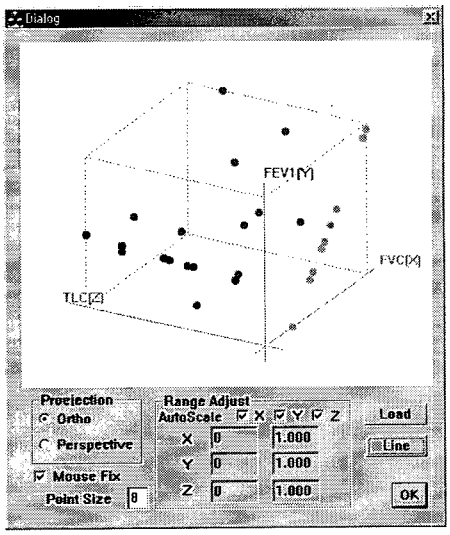
(b) Wilcoxon 부호순위 검정 결과

그림 6. 표본 데이터 분포에 따른 적합한 방법 가이드

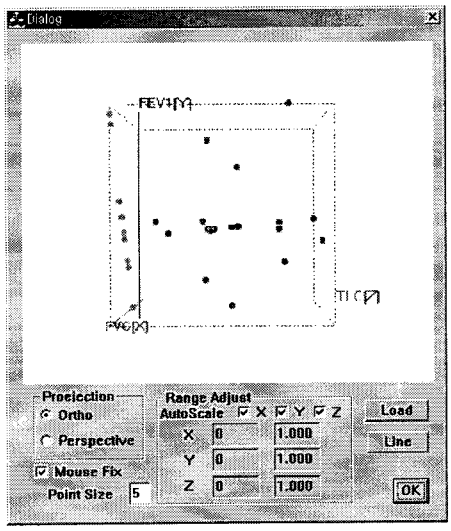
차원 산점도 그래프를 보여주는데, 데이터 여러 연관 관계를 직관적으로 파악할 수 있게 해준다. 다중선형 회귀분석은 한 개체에서 측정된 세 가지 이상의 데이터에 대하여 종속변수에 영향을 주는 독립변수가 2개 이상 일 때 상관관계를 분석하는 방법이다. 10명의 환자들을 대상으로 폐기능 검사를 하여 초당강제호기량 (FEV1), 폐활량(FVC), 총폐용량(TLC)를 얻었다. 초당 강제호기량은 폐질환을 검사하는데 중요한 측도중의 하나이다. 초당 강제호기량을 종속변수로 놓고, 폐활량과 총폐용량을 독립변수로 두어 어떤 상관관계가 있는지 그림 7과 같은 3차원 산점도 그래프를 통하여 여러 가지 뷰에서 관찰할 수 있다. 불확실성 시각화는 통계분석 모듈과 연동되어 수행된다. 다음 예에 대하여 단순선형 회귀분석을 수행한 그래프와 텍스트 결과가 그림 8에서 보인다.

실험 예>

신생아의 황달치료에 요구되는 혈액용적(blood volume)은 무해한 염색을 혈액내로 투여하여 염색 투여량을 혈액내 염색농도로 나누어 계산한다. 한편,

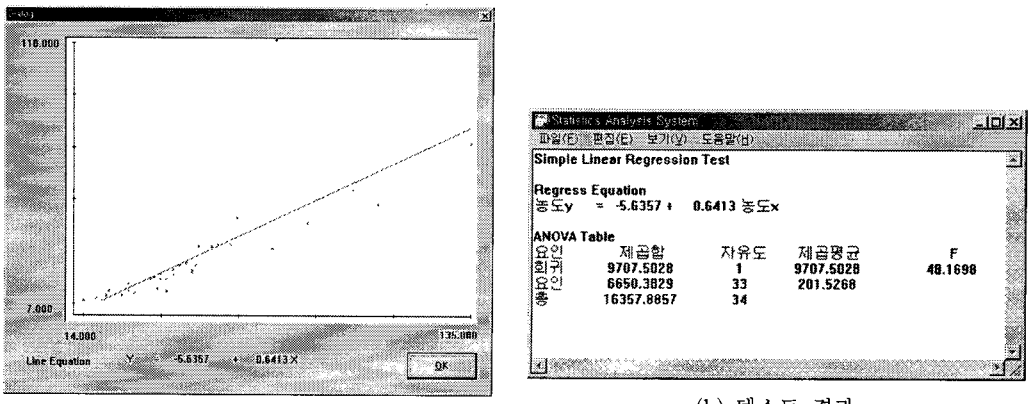


(a) 3차원 산점도 가시화 인터페이스



(b) 다른 뷰에서의 가시화

그림 7. 3차원 산점도 가시화



(a) 회귀직선 그래프 (b) 텍스트 결과

그림 8. 단순선형 회귀분석 결과

이 혈액내 염색농도는 광학기의 파장이 620인 지점(y)에서 측정되는데 다른 색깔도 같은 파장을 겹쳐 가질 수 있으므로 기술적으로는 이 염색과 전혀 무관한, 파장이 740인 지점(x)의 염색농도로부터 추정하게 된다. 35 명의 신생아부터 자료가 측정되었다. 여기서는 파장 620인 지점의 염색농도를 종속변수로 정하고 파장 720인 지점의 염색농도를 독립변수로 취하여 두변수의 관련성을 예측하게 될 것이다. 분석의 목적은 720 지점의 염색농도를 사용하여도 620 지점의 것을 사용하는 것과 같은 효과를 본다는 것을 증명하는데 있다.

기본적으로 나타는 회귀식의 결과는 720과 640 파장의 데이터가 일정한 연관성을 가지고 있으며, 720 파장의 것으로 대체해서 사용해도 무리가 없다는 결론을 내릴 수 있을 것이다. 하지만 이것만으로는 회귀모형의 적절성을 평가하기에는 부족하다. 분석결

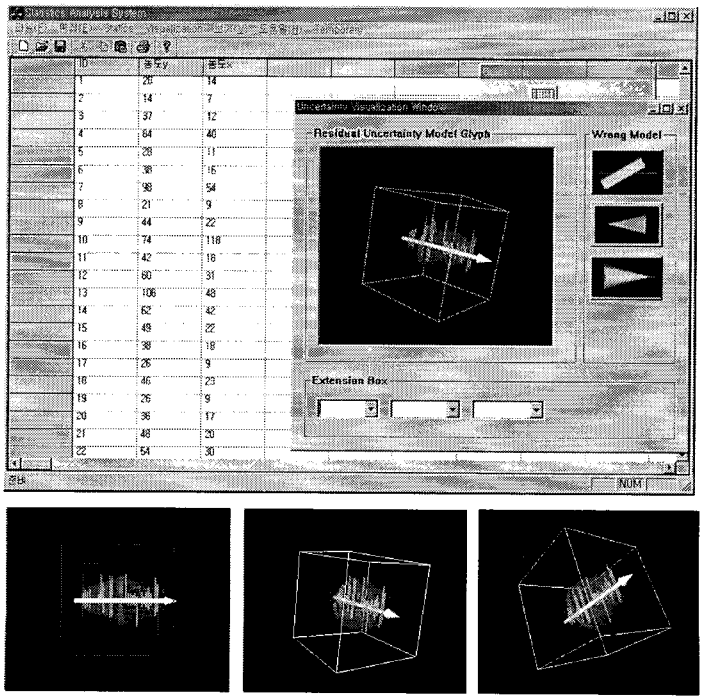


그림 9. 잔차(residual) 시각화 결과

과가 과연 올바르게 나왔는지에 대한 확신과 신뢰도를 높이기 위해 분석 후 적절성 평가를 위한 불확실성 시각화 윈도우가 사용자에게 보인다. 그림 9는 분석이 끝난 후 회귀모형의 잔차(residual)를 계산하여 3차원 글리프로 보여주는 화면이다. 오른쪽에는 3가지의 자주 발생되게 되는 오류 모델을 보여주고, 왼쪽에 현재 분석법과 데이터 사이의 오차 모델이 나타난다. 사용자가 왼쪽에 나타나는 모델과 오른쪽의 모델을 비교하여 분석법의 적절성을 평가할 수 있도록 하였다. 그림 9에서 보이는 3차원 잔차(residual) 모델로 오차 등분산의 가정을 위배하지 않는 것을 직관적으로 알 수 있다. 따라서 이 실험에서 회귀직선의 적절성도 좋다고 평가할 수 있다. 이러한 접근법은 사용자가 인지하기 쉬어 적절한 의사결정에 도움을 줄 수 있다.

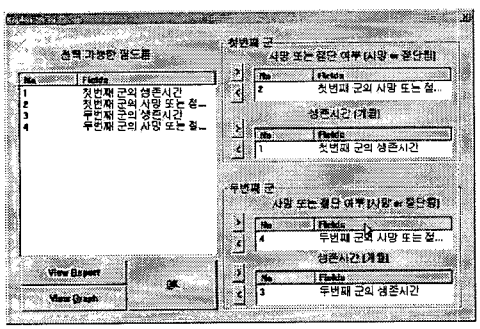
3.3 통계 분석 결과

다양한 통계 분석 모듈 중에서 본 절에서는 그룹간 생존 분석과 일원 배치법 분석 결과를 제시하고자 한다. 통계 분석 결과는 그래프 또는 텍스트로 제공되며 SAS™ 코드 또한 자동 생성되었다.

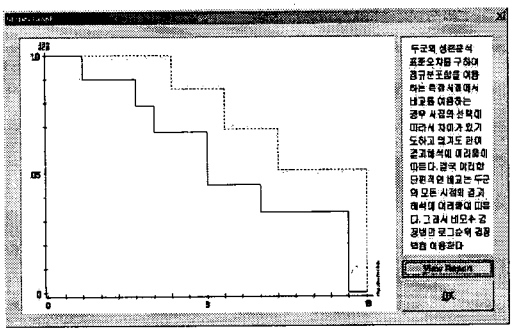
생존 분석[13]이란 어떤 정해진 시작점으로부터 사건의 발생시점까지의 기간을 생존기간 또는 생존시간(survival time)이라 하며 이때 생존시간을 분석하여 생존경험을 적절히 요약하는 방법이다. Kaplan-Meier 생존곡선은 생존율을 Y축으로 누적 생존시간을 X축으로 갖으며 계단곡선으로 만들어 진다. 다음 실험에 대하여 그룹간 생존 분석을 수행한 결과를 제시한다.

실험 예>

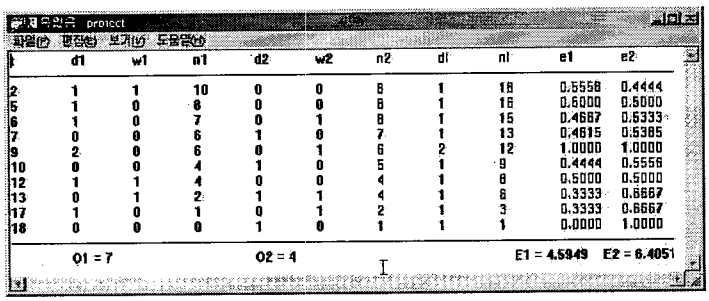
심장이식 수술을 받은 환자를 대상으로 18개월간 독립된 두 개의 군에서 환자들의 생존을 추적하였다고 하자. 그림 10 (a)는 집단 비교분석의 메인메뉴, (b)는 View Graph를 클릭한 후 생존곡선 그래프, (c)는 View Report를 클릭한 후 보고서가 출력된 화면이다. 그림 10 (c)는 로그순위 검정법에 의해 작성된 결과표를 보여준 것으로, 카이제곱 분포표에 의해 이 생존자료로는 두 군의 생존경험이 유의하게 다르다는 결론을 내릴 수가 없다.



(a) 집단 비교분석 메뉴



(b) 생존곡선 그래프



	d1	w1	n1	d2	w2	n2	dl	nl	e1	e2
2	1	1	10	0	0	8	1	18	0.5556	0.4444
5	1	0	8	0	0	8	1	16	0.6000	0.5000
6	1	0	7	0	1	8	1	15	0.4667	0.5333
7	0	0	6	1	0	7	1	13	0.4615	0.5385
9	2	0	6	0	1	6	2	12	1.0000	1.0000
10	0	0	4	1	0	5	1	9	0.4444	0.5556
12	1	1	4	0	0	4	1	8	0.5000	0.5000
13	0	1	2	1	1	4	1	6	0.3333	0.6667
17	1	0	1	0	1	2	1	3	0.3333	0.6667
18	0	0	0	1	0	1	1	1	0.0000	1.0000

Q1 = 7 Q2 = 4 E1 = 4.5949 E2 = 6.4051

(c) 레포트 출력

그림 10. 그룹간 생존분석 결과

그림 11 (a)는 한 모집단 한 반복 요인 분석을 메뉴에서 선택한 화면을 보여주고, (b)는 필드를 선택한 화면이다. (c)는 분석된 텍스트 결과 화면이고, (d)와 (e)는 생성된 SAS 코드를 보여준다.

3.4 다른 통계분석 시스템과의 비교

개발된 시스템을 기존의 통계분석 시스템과 통계분석 가이드, 분석과정 시각화, 불확실성 시각화, 그래프 형식 측면에서 비교 하면 표 1과 같다. 의료 통계분석에 특화된 시스템으로 의학 분야 실험에서 자주 사용되는 분석법을 중점적으로 만들어진 MedCalc™은 텍스트 형식으로 분석과정을 보여주고, 2차원 형식의 다양한 의료용 그래프를 지원하고, 통계분석 가이드를 위해 데이터 형 검사와 같은 기능을 일부 지원하고 있다. Vista™는 분석과정을 보여주기 위한 WorkMap과 통계분석 가이드를 위한 GuideMap을 제공하고 있다. 하지만 상용 통계 시스템과 마찬가지로 초보자가 배우는 데에는 복잡하다.

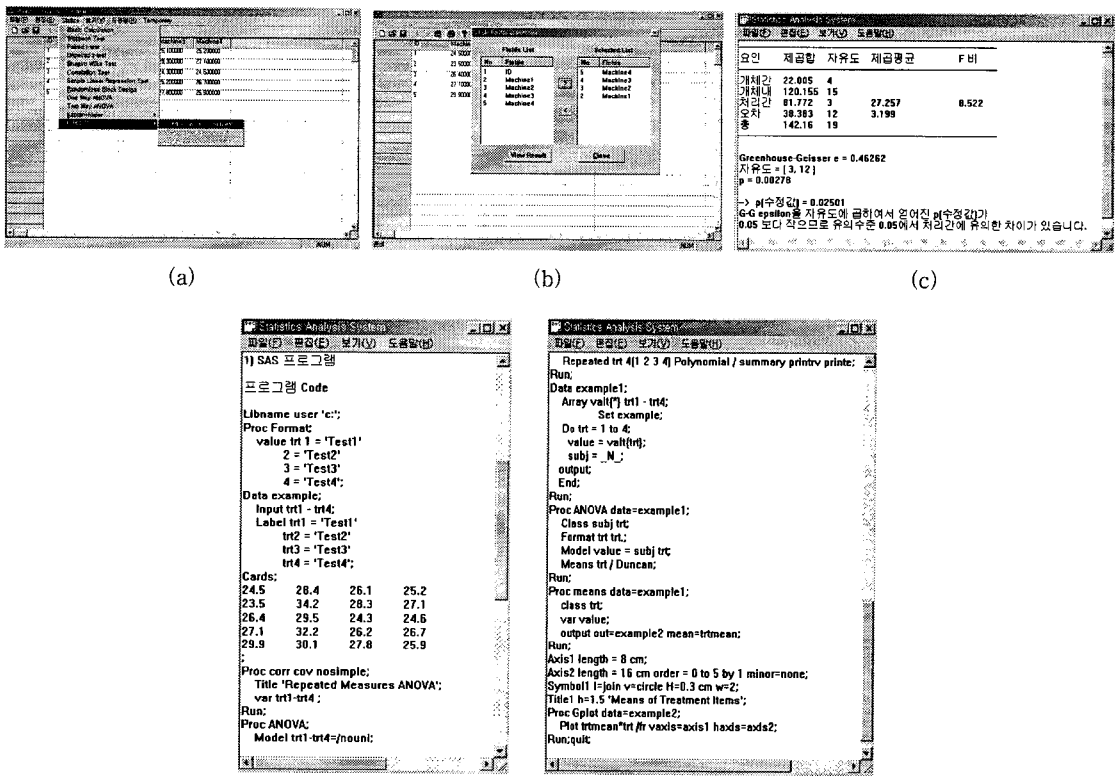


그림 11. 한 모집단 한 반복 요인 분석 결과

표 1. 통계분석 시스템의 특징 비교

특징 \ 소프트웨어	MedCalc™	ViSta™	dBSTAT™	제안 시스템
통계분석 가이드	데이터형 검사	Guidemap	통계 마법사	비쥬얼 가이드 경고 메세지
분석과정 시각화	텍스트	Workmap	없음	아이콘 트리맵
불확실성 시각화	없음	없음	없음	Glyph 시각화
그래프 형식	2차원	2차원, 3차원	2차원	2차원, 3차원, 인터랙티브 형식

dBSTAT™[16]는 통계지식이 부족한 일반인을 위해 설계된 통계 소프트웨어로서 2차원 형식의 그래프를 지원하고 있다.

제안 시스템은 의료 통계에 자주 사용되는 분석법들을 쉽게 사용할 수 있도록 아이콘들의 트리로 구성한 가이드 맵을 제공하고, 표본 데이터의 수 및 분포에 따라 적절한 분석 방법을 가이드 해주고 있다. 또한 2차원 그래프 이외에 3차원 글리프 (glyphs)[5,6]를 이용하여 데이터 결과를 인터랙티브하게 볼 수 있도록 하였다. 뿐만 아니라 통계 방법에 내재한 불확실성 또한 3차원 글리프를 이용하여 쉽게 알 수 있도록 하였다.

4. 결 론

본 연구에서는 의료데이터의 통계분석을 위하여 의학통계에서 많이 사용되는 통계분석법을 구현하였고, 쉽고 정확한 사용을 위하여 비쥬얼 가이드 인터페이스를 개발 하였다. 그리고 3차원 글리프를 이용한 인터랙티브 시각화 방법을 도입하여 결과 및 데이터의 불확실성을 시각화하는 방법을 개발하였다. 개발 시스템은 임상 의학자들이 데이터를 분석하거나 결과를 쉽게 이해할 수 있도록 하는 의료용 통계 분석 및 교육 시스템으로 활용 가능하다. 향후 연구로는 잔차(residual) 모델과 오류 모델을 자동으로 비교분석 해주는 연구와 여러 분석법에 불확실성 시각화 방법을 적용하는 것이 필요하다.

참 고 문 헌

[1] F. W. Young and C. M. Bann, "ViSta: A Visual Statistic System," *Statistical Computing Environment for Social Research*, pp. 959-998, 1993.

[2] P. Sutherland, A. Rossini, and T. Lumley, "ORCA: A Visualization Toolkit for High-Dimensional Data," *NRCSE-TRS*, No. 046, 2000.

[3] F. Schoonjans, "MedCals," http://www.medcalc.be.

[4] H. Shiozawa, K. Okada, and Y. Matsushita, "3D Interactive Visualization for Inter-Cell Dependencies of Spreadsheets," *ACM Special Interest Group*, pp. 71-80, 1999.

[5] S. M. Choi and D. S. Lee, *et al.*, "Interactive Visualization of Diagnostic Data from Cardiac Images using 3D Glyphs," *Lecture Notes in Computer Science* 2868, pp. 83-90, 2003.

[6] D. Ebert, J. Kukla, and C. Shaw, "Perceptually-motivated Glyph-based Informtion Visualization" *CODATA Workshop : Visualization of Information and Data*, pp. 24-25, June 1997.

[7] A. T. Pang, "Approaches to Uncertainty visualization," 1996.

[8] V. Batagelj and A. Mrvar, "Visualization of Multivariate Data Using 3D and VR Presentation," http://vlado.fmf.uni-lj.si/vrml/paris.97, 1997.

[9] E. H. Chi, J. Riedl, E. Shoop, and J. V. Carlis, et al., "Flexible Information Visualization of Multivariate Data from Biological Sequence Similarity Searches," *IEEE Visualization 96*, ACM Press, pp.133-140, 1996.

[10] E. H. Chi, J. Riedl, P. Barry, and J. Konstan, "Principles for Information Visualization Spreadsheets," *IEEE Computer Graphics and Applications*, July 1998.

[11] L. Wilkinson, "Presentation Graphics," SPSS Ins. 233 South Wacker, Chicago, IL 60606.

[12] C. Ware, Information Visualization, Morgan Kaufmann Publishers, 2000.

[13] 김성권, *SAS와 통계*, 대선출판사, 2000.

[14] 송혜양, 김동재, *통계학*, 청문각, 2002.

[15] 유대근, 권영식, *통계분석을 위한 SPSS WIN 8.0*, 기한재, 1999.

[16] 김수녕, *윈도우용 통계소프트*, 탐진, 2000.

이 돈 수

2002년 세종대학교 전산과학 (학사)
2004년 세종대학교 컴퓨터공학부(석사)
2005년 현재 햄펙스(주) 분당연구소 주임연구원

관심분야 : 컴퓨터 그래픽스, 얼굴 애니메이션, 모바일, 임베디드 시스템

최 수 미

1993년 이화여자대학교 전자계산학과(학사)
1995년 이화여자대학교 전자계산학과(석사)
2001년 이화여자대학교 컴퓨터학과(박사)
2001년~2002년 이화여자대학교 정보통신연구소 연구전임강사
2002년~2004년 세종대학교 컴퓨터공학부 전임강사
2004년 3월~현재 세종대학교 컴퓨터공학부 조교수
관심분야 : 의료영상가시화, 그래픽스, 가상현실, HCI, 유비쿼터스 디스플레이 등