

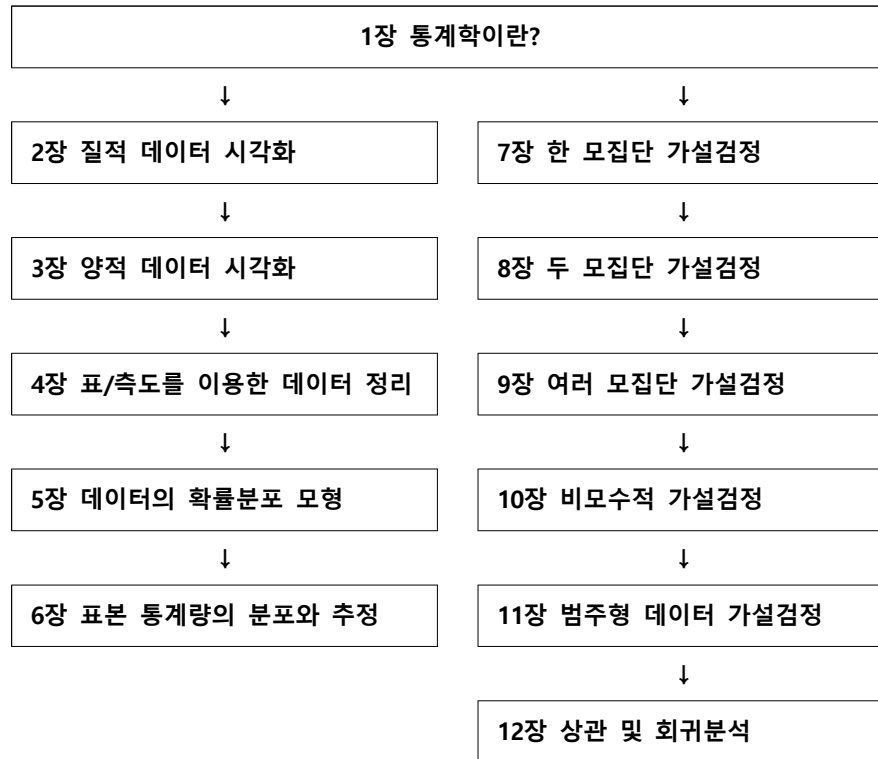
1장 통계학이란?

1.1 통계학과 데이터과학

- 1946년 미국 펜실베이니아대학의 존 에커트와 존 모클리에 의해 처음 개발되었던 현대 디지털 컴퓨터는 1960년대 이후 현실에 응용되기 시작하여 지난 반세기 동안 엄청난 발전을 이룩하고 우리 사회의 많은 변화를 가져왔다. 특히 1980년대 이후 컴퓨터와 컴퓨터의 연결이 시작되고, 개인용 컴퓨터가 활성화되고, 유무선 정보통신 기술이 발전되면서 최근에는 전 세계의 거의 모든 컴퓨터가 유무선 인터넷을 통하여 연결되어 있다. 2000년대 이후에는 성능이 우수한 컴퓨터가 소형화 되면서 전화기와 연결한 스마트폰이 탄생되어 우리 사회에 많은 변화를 가져왔다.
- 이와 같은 컴퓨터와 정보통신 기술의 발전은 최근에 더욱 심화되어 알파고와 같은 인간의 지능을 능가하는 **인공지능**(artificial intelligence; AI)을 만들어내고 있다. 또한 모든 전자기기를 인터넷으로 컴퓨터에 연결시키는 **사물인터넷**(internet of things; IoT) 시대를 준비하고 있다. 자동운행 차, 로봇 의사, 로봇 선생님 등 현재와는 획기적으로 다른 사회가 예견되는데 이를 4차 산업혁명 기술사회로 부른다.
- 이러한 기술의 발전은 과거에는 상상도 할 수 없었던 크기의 빅데이터(big data)를 생성하였다. 빅데이터의 대표적인 예로는 전 세계인이 많이 사용하고 있는 구글의 검색기록 데이터, 스마트폰의 소셜미디어 데이터, 인터넷의 웹로그(web log) 데이터, 글로벌 통신회사의 통화기록 데이터 등이 있다. 향후 4차 산업혁명이 진행되면서 빅데이터는 점점 더 커지고 많아질 전망이다. 이 빅데이터를 효율적으로 활용하여 과거에는 불가능했던 미래에 대한 초예측(hyper-forecasting)이 가능할 전망이다. 4차 산업혁명 사회에서는 어떻게 빅데이터를 유효적절하게 만들고 이를 사용하느냐에 따라 각 개인, 단체, 기업, 나아가 국가의 성패가 달려 있다.
- 문자 및 숫자 등으로 이루어지는 데이터는 인류가 문자를 발명하여 역사를 기록하면서 생겨났다고 볼 수 있다. 고대의 이집트, 그리스 로마 등에서는 인구수, 농지 면적 등의 데이터를 만들어 국가 경영에 사용한 기록이 있다. 이러한 단순한 데이터 활용은 17세기 이후 수학의 확률론 발전에 힘입어 **통계학**(statistics)이란 학문으로 발전하였다. **현대통계학은 데이터를 효율적으로 수집하고, 이를 정리, 요약한 후 분석을 하여 불확실한 상황의 의사결정에 대해 여러 가지 확률적 모형을 이용하여 과학적인 판단을 내릴 수 있도록 도움을 주는 학문이다.**
- 4차 산업혁명 사회에서도 현실의 불확실한 상황에 대한 의사결정을 할 때 전통적인 통계학의 기법이 주를 이룬다. 하지만 금세기에 출현한 빅데이터의 분석은 데이터의 양도 엄청나고 다양해 단지 통계학적인 접근만으로 그 활용을 모두 할 수는 없다. 이러한 빅데이터의 분석을 위해서는 전통적인 통계학의 이론과 수학의 최근 이론, 컴퓨터 과학, 그리고 분석된 결과를 효율적으로 활용하기 위해서는 경영학 등 관련 학문도 같이 적용되어야 한다. 이와 같이 여러 학문 분야가 융합하여 금세기에 출현한 빅데이터를 분석해 현실에 응용하는 학문을 **데이터과학**(data science)이라 부른다.

- 빅데이터를 분석하여 현실에 응용하는 데이터과학이 활용된 예는 많이 있다.
 - 구글의 검색 엔진에 자동차 구입에 관한 질문을 조사하여 다음 달 미국서 판매되는 자동차 모델의 수를 예측하였다.
 - 구글 검색 엔진에 감기약을 검색한 결과를 분석하여 올해 미국서 유행하는 감기의 전파 경로를 지도에 표시하였다. 이를 구글 플루라 부르는데 미국 정부의 질병관리본부보다 앞서서 감기의 전파경로를 예측하여 세상을 놀라게 하였다.
 - 베네수엘라의 한 식품체인 회사는 분산되었던 각 지점의 데이터를 통합 분석하여 재고관리 개선과 이에 맞는 상품 판매 전략을 수립하여 매출이 30%나 증가하는 성과를 이루었다.
 - 한 온라인 쇼핑몰은 웹로그를 분석하여, 회원 고객이 어떤 취향을 가지고 어떤 제품에 관심이 있는지 파악하여 고객 개개인에 맞는 맞춤형 광고를 하여 매출이 증가하였다.
 - 한 원유 탐사회사에서 테라바이트 규모의 지질학 데이터를 분석해 원유 시추의 성공률을 높였다.
 - 남아프리카의 어느 보험회사에서 기존 보험금 청구 빅데이터를 분석하여 보험사기 가능성이 있는 사건을 찾을 수 있는 알고리즘을 구현하였다. 이를 활용하여 많은 보험사기를 적발하였고 심지어 대형 보험사기 조직을 적발하기도 하였다.
 - 미국의 한 대학에서 온라인 수업에서 학생들이 시스템에 클릭하는 정보를 분석하여 학생 개개인의 학습 성과를 모니터링하고 학생의 이해도에 맞춘 수준별 수업 내용을 제안하고, 향후 수강할 과목 등을 학생별로 제안하였다. 이 결과 전공별 학위 취득률이 많이 향상되었다.
 - 덴마크의 한 풍력발전 회사는 기존 발전기에서 축적된 페타바이트 규모의 데이터를 분석하여 풍력발전기에 대한 날씨와 위치의 영향을 정확히 파악하고 이를 바탕으로 풍력발전기의 부지 선정 및 운영을 효율적으로 할 수 있게 되었다.
- 데이터과학은 여러 학문의 융합이어서 데이터과학을 연구하기 이해서는 여러 학문 분야를 두루 많이 알아야 한다. 구체적으로 최근 빅데이터의 분석에 많이 사용되는 기법은 통계학의 가설검정, 다변량분석, 선형모형 등의 전통적인 이론과 함께 수학에서 발전한 신경망(neural network), 지지벡터기계(support vector machine), 컴퓨터 과학의 데이터베이스(database), 분산컴퓨팅(distributed computing), 기계학습(machine learning), 인공지능(artificial intelligence) 등이다.
- 여러 학문의 융합인 데이터과학을 공부하는 것은 쉽지 않다. 잘못하면 이 분야도 많이 알지 못하고 저 분야도 제대로 많이 모를 위험이 있다. 그러나 데이터과학을 잘 공부한 사람은 21세기가 필요로 하는 인재가 될 것임이 틀림없다.
- 이 책에서는 통계학에 입문하는 초보자를 위해 데이터과학의 기초인 데이터 시각화와 데이터 정리 방법을 소개하고, 확률 및 확률분포함수, 표본을 이용한 모집단의 특성을 추론하는 추측통계의 여러 가지 통계적 분석방법을 소개한다. 표 1.1은 이 책의 구성을 보여준다.

표 1.1 이 책의 구성



- 2장은 막대, 원, 띠, 꺾은선 그래프 등의 범주형 데이터 시각화를 다룬다. 3장은 히스토그램, 줄기와 잎 그림, 산점도 등의 연속형 데이터 시각화를 다룬다. 4장은 표/측도를 이용한 데이터 정리를 소개한다.
- 5장은 데이터에 대한 확률분포 모형을 소개하고, 6장은 표본과 모집단의 관계에 대해서 살펴보고 표본통계량에 대한 분포와 이를 바탕으로 모집단 모수에 대한 추정을 설명한다.
- 7장에서 9장까지는 연속형 변량에 대한 모수적 가설검정을 설명하고, 10장에서는 연속형 변량의 비모수적 가설검정, 11장은 범주형 변량에 대한 가설검정을 설명한다. 12장은 두 변량에 대한 상관 및 회귀분석을 설명한다.

1.2 데이터의 구분

- 데이터는 관심의 대상이 되는 사물이나 사건의 속성을 일정한 규칙에 의해 관찰하거나 측정된 값들이다. 이러한 사물이나 사건의 속성을 **변수** 또는 **변량**(variable)이라고 한다. 예를 들어, 어느 대학 재학생의 성별과 신장을 측정하였다면 여기에는 두개의 변량(성별, 신장)이 있다. 성별에 대한 측정값은 ‘남’, ‘여’, ‘여’, ‘남’, 과 같은 형태이고, 신장에 대한 측정값은 180cm, 165cm, 158cm, 175cm, ... 와 같은 형태일 것이다.
- ‘성별’과 같은 변량의 데이터를 **질적 데이터**(qualitative data), 신장과 같은 변량의 데이터를 **양적 데이터**(quantitative data)로 구분한다. 질적 데이터의 경우 범

주 형태를 갖는 경우가 많아 이를 **범주형 데이터**(categorical data)라 부르기도 한다. 양적 데이터의 경우 여러 개의 동전을 던졌을 때 나타나는 결면의 수와 같이 가능한 값이 유한개 또는 셀 수 있는 변량의 데이터를 **이산형 데이터**(discrete data)라고 하고, 신장 및 체중과 같이 무한개의 가능한 값(실수이기 때문)을 갖는 변량의 데이터를 **연속형 데이터**(continuous data)라 부른다.

- 데이터를 구분하는 이유는 데이터의 종류의 따라 처리하는 방법과 분석 방법이 다르기 때문이다. 이 책의 2장은 범주형 데이터의 시각화를 다루고, 3장은 연속형 데이터의 시각화를 다룬다. 4장에서는 범주형 데이터의 요약인 도수분포표와 교차표를 다루고, 표 및 측도를 이용한 연속형 데이터 정리를 설명한다. 5장에서 10장 그리고 12장은 연속형 데이터의 통계 분석 이론을 설명한다. 11장은 범주형 데이터의 분석 이론을 설명한다.
- 소프트웨어를 이용한 데이터 분석을 위해 범주형 데이터는 **원시 데이터**(raw data)와 **도수분포 데이터**(frequency table data)로 구분한다. 예를 들어, 어느 초등학교 한 학급 학생 10명의 성별을 남, 여, 남, ... 등으로 조사하여 다음과 같이 엑셀 시트에 정리하였다면 이를 원시 데이터라 한다. 여기서 변량의 이름 ‘성별’을 **변량명**(variable name), ‘남’ 또는 ‘여’와 같은 값을 **변량값**(variable value)이라 부른다.

표 1.2 성별을 조사하여 엑셀에 정리한 원시 데이터

성별
남
여
남
여
남
남
남
여
여
남

- 표 1.2의 한 학급 성별 데이터는 ‘남’이 6명이고 ‘여’가 4명이다. 이렇게 빈도수를 정리한 데이터를 도수분포 데이터라고 부른다. 엑셀에서는 이러한 도수분포 데이터를 이용하여 그래프를 그린다.

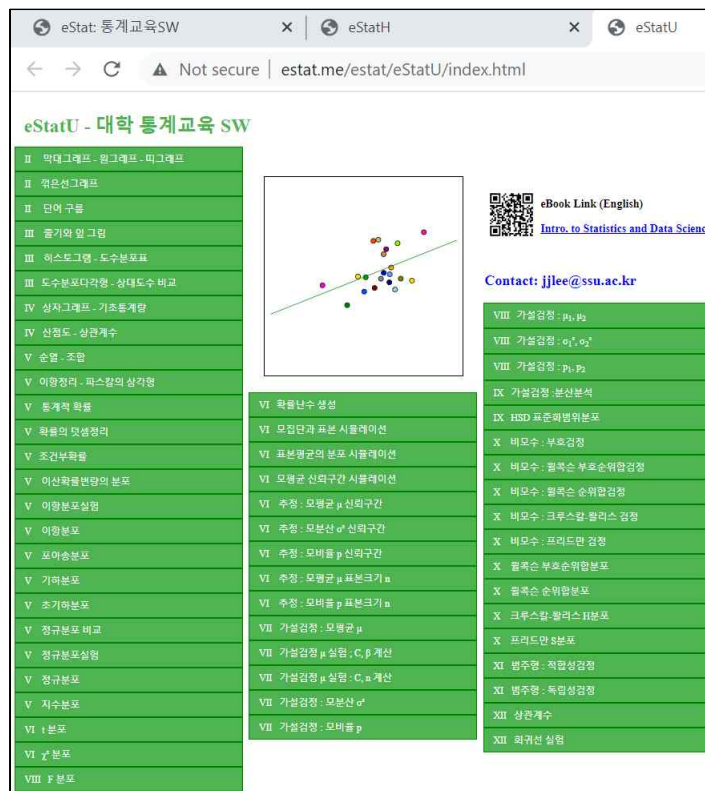
표 1.3 한 학급의 성별의 빈도수를 요약한 도수분포 데이터

성별	학생수
남	6
여	4

- 단어와 같은 질적 자료는 각 단어에 대한 빈도수를 조사하여 단어구름으로 만들어 분석한다(2장). 양적 자료는 평균 및 표준편차를 계산하고 줄기와 잎 그림, 히스토그램, 점그래프 등을 이용하여 시각화 한다(3장).


1.3 『eStat』 소프트웨어

- 데이터 분석을 위해서는 소프트웨어의 도움이 필수적이다. 특히 빅데이터 분석을 위해서는 전문적인 통계분석 모듈을 많이 가지고 있는 통계 패키지(statistical package)가 반드시 필요하다. 현재 빅데이터 분석을 위해서는 SAS, SPSS, R과 같은 통계패키지가 많이 사용되고 있다.
- 하지만 이들 통계패키지들은 초보자가 배우기는 쉽지 않고, SAS와 SPSS는 상업용이어서 엄청난 고가이다. 그리고 이러한 통계패키지는 빅데이터 분석의 핵심인 통계학 교육에 필요한 모듈의 기능은 거의 없다고 할 수 있다. 통계학 교육을 위해서는 일부 개인들이 부분적인 기능의 소프트웨어를 만들고 있으나 초·중·고·대·일반인들이 모두 사용할 수 있는 종합적인 통계교육용 소프트웨어는 아직 없었다.
- 『eStat』은 데이터과학을 초등생부터 대학 및 일반인까지 쉽게 교육하기 위하여 만든 통계패키지 + 교육용소프트웨어이다. 데이터가 주어지면 단지 마우스 클릭만으로 그래프를 그릴 수 있고, 동적인 데이터 시각화를 경험할 수 있으며, 데이터에 대한 통계 분석 및 처리 실습까지 가능하다.
- 『eStat』은 통계패키지와 같이 데이터 처리가 가능하며, 『eStatU』에는 통계학 이론에 대한 이해를 돕기 위한 다양한 시뮬레이션 모듈을 포함하고 있다. 이항분포와 정규분포가 무엇인지 보여주는 시뮬레이션, 대수의 법칙, 중심극한정리, 구간추정의 의미를 보여주는 시뮬레이션, 회귀분석의 이상값의 영향을 관찰할 수 있는 시뮬레이션 등이다. [그림 1.1]은 『eStatU』의 메뉴이다.



[그림 1.1] 『eStatU』 메뉴

- 『eStat』은 각급 교과서에 있는 많은 예를 포함하고 있으며, 웹 기반이어서 사용자들은 언제 어디서나 PC, 태블릿, 또는 스마트폰으로 이용할 수 있다. 『eStat』은 무료로 서비스하고 있고 다국적 언어를 지원하며 현재 한국어, 영어, 일본어, 중국어, 불어, 독어, 스페인어, 베트남어, 인도네시아 등 20개 언어로 번역되어 전 세계에서 사용하고 있다.
- 『eStat』에 대한 기본 운영은 다음 링크를 참조하라.

 <p>『eStat』 기본 운영</p>	<p>www.estat.me/estat/eLearning/kr/eStatBasicOperation.html</p>
--	---