통계학 전공자와 데이터베이스

1 통계학

"통계학은 **자료**(데이터, data)를 수집하고 관리하고 분석하는 학문" 으로 과학 연구의 가장 중요한 도구이다. 과학은 통계적 방법으로 연구 대상의 전체 집합인 **모집단**(population)에 대한 사실을 알아내는 것을 목적으로 한다. 이 때 알아내려는 것은 모집단 자체가 아니라 관찰 가능한 모집단의 어떤 측면에 대한 것임을 인식하는 것은 상당히 중요하다. 관찰하는 모집단의 측면을 **변수**(variable)라고 한다. 모집단의 구성원에서 측정한 변수의 값이 모두 같지 않기 때문에 알아내고자 하는 모집단에 대한 사실은 변수의 **분포**(distribution)이다. 결국 통계적 방법을 사용하여 과학이 알아내고자 하는 것은 모집단에서 변수가 어떤 분포를 하는가 하는 것이다.

변수의 분포를 알아내기 위해 전통적으로 사용하는 방법은 모집단의 일부분을 표본으로 추출하고, 표본에서 변수의 값을 측정하여 자료를 수집한 다음, 자료를 분석하여 변수의 분포에 관해 추론한다. 하지만 표본과 모집단이 동일하지 않으므로 표본 자료를 기반으로 추정한 결과는 일반화 오류를 피할 수 없다. 일반화 오류가 어느 정도 되는지, 이 오류를 고려할 때 변수의 분포는 어떻게 추정되는지 그리고 추정 결과를 어떻게 해석하는지를 연구하는 것이 전통적인 통계학이다.

2 통계학과 데이터베이스

통계학의 중심은 자료이다. 자료는 모집단 구성원에 대한 경험이며, 자료를 분석하여 얻은 분포에 대한 결과는 경험으로부터 도출한 모집단에 대한 지식(knowledge)이다. 통계학은 경험으로부터 지식을 도출하는 방법이므로 지식을 얻는 방법으로 **지능**(intelligence)의 구체적 모습이라고 할 수 있다. 이런 맥락에서 보면 **인공 지능**(artificial intelligence, AI)을 구현하는 핵심적인 방법이 통계학일 수밖에 없다.

경험이 많을수록 경험에서 도출한 지식은 견고할 것이기 때문에 자료의 질과 양은 참으로 중요하다. 정보화가 진행되면서 엄청난 양의 자료가 수집되어 보관되어 있으며 앞으로 더 많은 자료가 수집될 것이 분명하다. 이 자료는 대부분 데이터베이스 형태로 저장되어 있으며 앞으로 다양하게 분석될 것이 분명하다. 이런 대규모의 자료를 관리하고 분석하기 위해 요구되는 IT 기술과 통계적 방법을 융합한 분야가 바로 **빅 데이터**(Big Data), **기계 학습**, **심층 학습**과 같은 최근 화두가 되고 있는 분야이다.

이런 상황에서 통계학 전공자는 다음과 같은 지식을 갖추는 것이 바람직하다.

- 데이터베이스에 대한 이해
- 데이터베이스를 구축하는 방법
- 데이터베이스에 저장된 자료에 접근하는 방법
- 데이터베이스에서 필요한 자료를 추출하는 방법
- 데이터베이스에 저장된 자료를 분석하는 방법
- 데이터베이스에 새로운 자료를 저장하거나 수정하는 방법