세계적인 선도 IT기업 Google은 기존 직관에만 의존하여 관리하던 직원들의 행동 역학에서 빅데이터 기술을 이용한 체계적이고 구체적인 행동 역학을 구축하면서 선진적인 기업 문화를 만든 대표적인 기업이다.
Google이 선진 기업 문화를 구축하면서 직원들의 행복감과 친밀도가 높아지고 자연스럽게 이직률도 낮출 수 있게 되면서 빅데이터는 기업의 중요한 기반이 되었다.
위의 예시처럼 세계적으로 빅데이터 기술의 효과가 나타나면서 현재는 빅데이터라는 기술의 필요성이 요구되었다. 이처럼 오늘날 빅데이터는 크게 붐을 일으키게 되었다.
나는 이러한 빅데이터가 통계와 관련이 많고 통계는 R 언어와 관련이 많은 것을 이번 통계학개론 수업 OT에서 확인할 수 있었다.
R 언어는 수리 연산과 분석 속도는 약하지만 통계에 최적화되어 있는 언어로 성공적인 데이터 분석을 실현할 수 있는 중요한 대안 중에 하나라고 한다. 그래서 주로 시스템 개발보다는 분석 방법론 테스트에 활용을 한다고 한다. 하지만 나는 이 내용을 봤을 때 약간의 의문이 생겼다. 왜냐하면 [도구로 푸는 통계, R 2강]에서 예제를 공부하면서 pi나 1/3 같은 무한소수와 순환소수의 소수점이 많은 자릿수로 표현되지 않는다는 점이었다. 나는 “통계가 정확하려면 우선 계산이 정확해야 한다”라고 생각했다. 그렇기에 “pi와 1/3을 3.141592와 0.3333333까지만 표현한다면 계산에서 평소보다 더 큰 오차가 발생하지 않겠나”라는 생각을 했다.
그래서 어떻게 한정된 소수점을 가지고도 분석 정확도가 높을 수 있는지 궁금했다. 계속해서 이유를 생각해보고 탐색을 해보았다. 나는 수학에서의 통계의 의미부터 다시 짚어보게 되었는데 “집단적 현상이나 수집된 자료의 내용에 관한 수량적인 기술”이라고 한다. 즉, 집단적 현상 자료와 수집된 자료들을 가지고 사회나 자연 현상을 정리하고 분석하는 것이라고 한다. 따라서 통계에 있어서 정확한 수치도 필요하지만 통계의 쓰임은 사회나 자연 현상을 분석하기 위함이니 수치의 양도 중요할 것이라는 생각이 들었다. 그런 생각을 하면서 동영상 강의를 쭉 보았는데 [도구로 푸는 통계, R] 5강에서 score=c(94,95,92,100,90,88) 명령어를 쓰고 deviation=score-mean(score)을 해서 각 요소들의 deviation을 한꺼번에 구하는 것을 학습하게 되었다.
여기서 나는 “R 언어가 pi나 1/3에서 소수점의 자리가 한정돼 있는 것처럼 연산에서는 약점이 있지만 전문가들에게 자주 쓰이는 이유는 위와 같이 방대한 수치를 한꺼번에 정리하고 계산할 수 있는 장점이 R언어에게 있어서 전문가들이 분석 방법론 테스트로 많이 활용한다는 것”을 간접적으로 느낄 수 있었다. 또한 이게 곧 R 프로그래밍 시에 명시적인 반복법 피하고 루프 코드 대신 내부에서 반복을 수행하는 R의 함수 기능 이용하여 큰 데이터의 처리시간을 단축한다는 의미라는 생각이 들었다.
나는 방학 동안 ‘이미테이션 게임’이라는 영화를 접하면서 컴퓨터 공학의 아버지이자 통계학을 이용한 ‘에니그마 해독 기계’를 발명하여 제2차 세계대전에서 영국을 구해낸 앨런 튜링을 알아보면서 통계학과 빅데이터에 호기심이 생겼었다. 그러던 도중 운 좋게 대학에서 통계학개론 수업을 수강할 수 있게 되었고 이번 R을 맛보기 하면서 단순 호기심에 들었던 수업이지만 평소 분석과 추리를 좋아하는 나랑 굉장히 잘 맞을 거 같은 학문이라는 생각이 들었다. 또한 객체지향 언어와 함수형 언어의 특징을 모두 포함했다고 하니 방학 동안 살짝 맛보기 했던 C#에서 흥미를 느꼈기 때문에 R 언어에서도 마찬가지로 흥미를 느낄 수 있을 것 같다. 더군다나 R 언어는 오픈소스라 사용자 커뮤니티가 활성화되어 있다고 하니 더 폭넓게 공부할 수 있는 언어가 될 수 있겠다는 생각이 들었고 더욱 기대가 생기게 되었다.