1731036009 최성준

빅데이터를 지배하는 통계의 힘 : 통계학이 최강의 학문이다.

우리들의 삶 속에서 통계는 많은 곳에서 영향력을 발휘하고 있으며, 그것은 우리의 삶의 질을 올려주기도 하고, 우리의 효율을 올려주기도 하고, 서로 연관이 없을 것 같던 것도 연관성을 가질 수 있음을 보여주기도 한다. 우리는 평소에도 간단하게 통계를 통해서 사고를 하고 그것을 기준으로 결정을 내리기도 한다. 대표적으로 우리는 여러 선택지 중에서 하나를 선택하고자 할 때 그 기준으로 과거의 경험, 여러 매체를 통해 얻은 지식 등을 총합하고 그것을 기반으로 어떤 선택지가 가장 좋은 선택이 될 지를 고르고 그 선택을 따라가게 된다. 이러한 통계는 의학, 과학, 사회과학 등의 여러 연구 분야에서도 사용되고 있다.

대표적으로 들어 볼 수 있는 예는 책에서도 나온 산업 혁명 시기의 영국에서 런던은 시골에서 돈을 벌기 위해 수 많은 사람들이 모여들었지만, 도시는 그 인구수를 받아들이는 속도가 느렸고 배수 등의 시설도 매우 열악했고 위생은 매우 나쁜 상태였다. 그리고 위생이 나쁜 환경에서 생활한 런던의 빈민 노동자들 사이에서 콜레라가 돌기 시작했고 수 많은 사람들이 목숨을 잃었다. 이 때 존 스노우라는 의사는 콜레라의 원인을 파악하기 위해 콜레라가 발생하는 지역에 대해 조사를 하여 특정 급수대를 사용한 곳에서 수 많은 콜레라 환자가 발생했다는 것을 알게 되었다. 이러한 결과를 도출하기 위해 존 스노우는 통계를이용하여 그 지역의 사람들이 물을 어디서 얻어오는 가를 확인하고 확인한 내용을 바탕으로 통계를 내어 지하수를 끌어올리는 급수대가 문제라는 것을 알 수 있었고, 그 결과로 많은 사람들의 목숨을 지킬 수 있었다. 콜레라와 급수대 보기에는 연관이 없을 것 같아 보이는 두 개이나, 통계를 통해 콜레라 환자가 물을 어디서 얻는 가를 확인하여 콜레라와 급수대의 물이 연관성을 지닌다는 것을 도출해낼 수 있게 된 것이다.



그림 . 존 스노우가 콜레라에 대해 조사할 때 사용한 지도 (지도에 표시된 막대는 콜라라로 사망한 사람의 수를 의미함)

하지만, 언제나 통계가 좋은 것은 아니다. 통계는 어떻게 하느냐에 따라 그 결과를 조작할 수 도 있기 때문이다. 예를 들어 말하면, 교통 사고가 나는 원인을 조사하기 위해 교통 사고자가 운전을 할 때 무엇을 가장 많이 마셨는가를 조사한다고 가정해볼 때, 표본을 조작하여 10만건의 사건에서 술을 마시고 사고를 내는 경우가 5만건, 커피를 마시고 사고를 내는 경우가 1천건, 나머지를 기타라고 할 때, 술을 마시고 사고를 내는 경우를 70건 커피를 마시고 사고를 내는 경울 500건, 기타 요소를 430건으로 넣는다면 그 통계 결과는 커피를 마시면 운전 중 교통 사고 발생량이 매우 높게 나온다 라는 식의 결과를 도출할 수 있다. 물론 이는 표본을 조작하여 낸 것이기에 문제가 있다. 방금 든 예시와 같이 통계는 어떻게 사용하느냐에 따라 이로울 수도 있으나 악용이 될 소지도 있다.

통계에 대해 이야기를 하면서 생각해볼 수 있는 것이 있다. 예시로 든 존 스노우의 콜레라 조사는 그 지역에 사는 몇 천, 몇 만 명의 데이터를 가지고 그들이 어떤 물을 마셨는가에 대하여 결과를 도출한 것이다. 이를 다시 생각해보면, 데이터에서 유의미한 가치를 뽑아내기 위해 통계를 사용했다는 것을 생각해볼 수 있다. 한 발 나아가 정보화 시대에서 하루에도 수 많은 정보가 나오는 이 시대에 통계는 그러한 데이터들에서 유의미한 가치를 뽑아낼 수 있는 매우 중요한 도구가 될 수 있다는 것이다.

데이터, 우리가 사용하는 컴퓨터는 수 없이 많은 데이터들을 가지고 있다. 그러나 데이터 그 자체가 무엇을 의미하는 지는 잘 모르는 경우가 많다.

* 데이터는 수, 영상, 단어 등의 형태로 된 의미 단위이다. 보통 연구나 조사 등의 바탕이 되는 재료를 말하며, 자료를 의미 있게 정리하면 정보가 된다. (위키백과)
* 데이터는 의미 있는 정보를 가진 모든 값, 사람이다 자동 기기가 생성 또는 처리하는 형태로 표시된 것을 뜻한다. (두산백과)

두 개의 인터넷 백과사전에서 본 결과이다. 데이터는 무엇인가 발생한 일, 현상 등을 모아두면서 이를 수, 영상, 단어 등 이후에 사용할 수 있는 형태로 보관한 자료이며 이는 정보와는 다른 것이다. 두산 백과를 보다 자세히 보면, [데이터 자체는 단순한 사실에 불과하지만, 일련의 처리과정에 따라 특정한 목적에 소용되는 정보를 만들기 위한 재료로 사용되는 것이다.]라고 서술하였다. 예를 들어 설명하면, 설문조사를 통해 우리나라의 자동차 종류에 따른 분포를 조사하면 수 백 만개가 넘는 데이터를 얻을 수 있을 것이다. 하지만 이는 단순히 사실을 보관한 자료이다. 여기에 이를 통계적으로 분석하여 어떤 차종에 얼마나 많은 차량이 있는 지를 도출해서 나온 “우리나라는 세단이 전체의 33.6%의 비율로 가장 많은 대수를 자랑한다.”가 실제적으로 사용될 수 있는 정보 가 되는 것이다. 예시를 다시 생각해보면, 데이터에서 정보를 뽑아내기 위해 사용한 것이 통계였다는 것을 생각해볼 수 있다. 이만큼이나 데이터와 통계는 매우 중요한 관계를 가지고 있다.

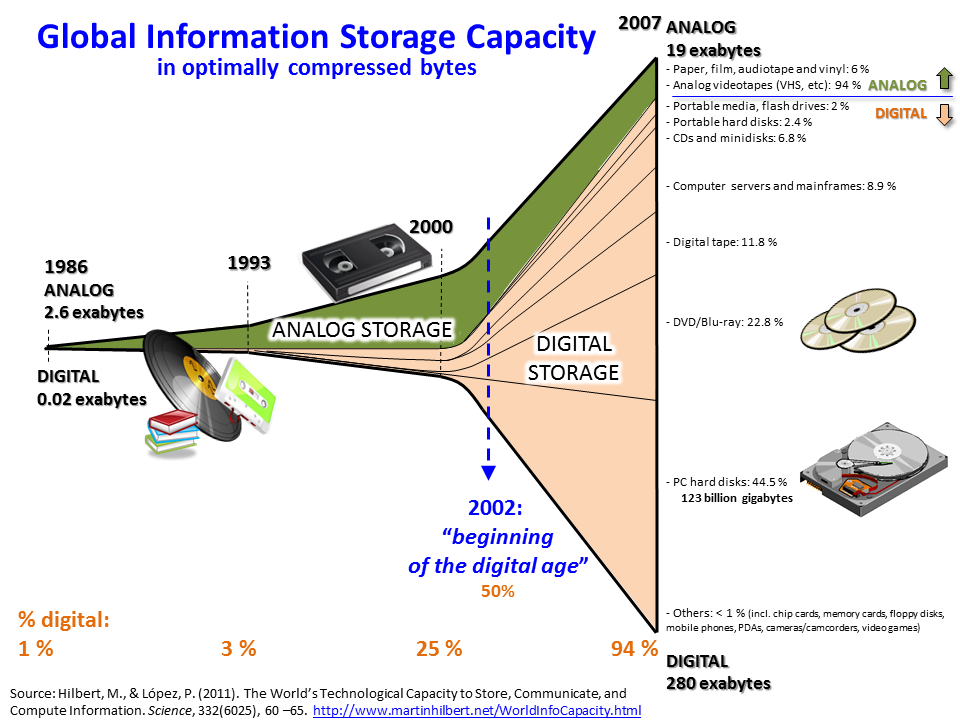


그림 . 시대에 따른 데이터의 총량

데이터는 하루에도 끊임없이 생산되고 있다. 그리고 그 데이터의 양은 매우 커지고 있어서 그 규모가 테라바이트를 넘는 거대한 데이터가 하루에도 쏟아지고 있고 이 시대의 데이터의 총량은 엑사바이트를 넘어섰다. 우리가 단순히 핸드폰을 들고 걷고 있어도 우리는 수 많은 데이터를 만들고 있다. 그것은 GPS 정보이며, 이 정보들을 모아보면 우리가 어디를 다녔는지, 얼마나 다녔는지를 알아볼 수 있다. 그리고 이런 정보는 어느 누군가는 매우 좋은 가치를 창출해낼 수 있는 데이터가 될 것이다.

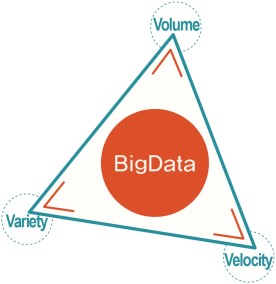


그림 . 빅데이터의 3V

또한 우리는 소셜 네트워크 서비스를 이용하여 내가 무엇을 했고 무엇을 선호하며 무엇에 관심이 있는 지에 대한 정보를 매일 매일 쏟아내고 있다. 이러한 거대한 데이터들 기존의 컴퓨터에서 사용되던 데이터 베이스라는 데이터를 관리하고 필요한 정보만을 추출하던 것에서 사용되는 정형 데이터가 아닌 비정형 (사진, 영상, 소셜 네트워크 서비스 로그, 쿠키 등)의 데이터가 많아지고 있다. 이러한 시대에서 사람들은 기존의 데이터 베이스보다 더 큰 테라 바이트 이상의 데이터를 이용할 수 있으며, 속도는 대용량 데이터를 빠르게 처리하고 분석할 수 있어야 하면서도 다양한 종류의 데이터를 수용할 수 있는 것을 생각했고 요구했다. 그리고 그 결과로 나온 것이 빅 데이터이다. 현대 시대는 빅 데이터가 매우 주목을 받고 있으며, 많은 곳에서 사용되고 있다.

빅 데이터는 단순하게 거대한 데이터를 말하는 것뿐만이 아니라 그러한 거대 데이터를 어떻게 빠르게 처리하며, 단순한 사실을 담고 있는 데이터를 가지고 가치를 창출할 수 있는 가까지에 대한 것으로 빅 데이터 처리를 위한 기법으로 분석 기술과 표현 기술이 존재한다. 그 중에서 분석 기술은 통계학과 전산학에서 사용되던 데이터 마이닝, 기계 학습, 자연 언어 처리, 패턴 인식 등이 사용되고, 더불어 소셜 미디어 같은 비정형 데이터의 증가로 기존 방법으로는 한계가 있어 새로운 기법들이 제시되었다. 데이터 마이닝은 기존 통계학 분야, 전산학 분야에서 오랫동안 존재한 것으로 통계라는 학문에 매우 많이 중요한 학문이다. 데이터 마이닝에서 사용되는 것으로 선형 회귀분석, 로지스틱 회귀분석, 판별분석, 주성분 분석 등이 있다. 데이터 마이닝은 매우 다양한 분야에서 사용되고 활용되고 응용되는 데, 우리가 가장 근처에서 볼 수 있는 것으로는 인터넷 쇼핑몰에서 나에게 맞춰 상품을 추천해주는 시스템, 소비 패턴에 맞춰 성향을 분석하여 적합한 상품을 소개해주는 것이 있다. 또한 선거 운동을 수치화하고 유권자들의 분류와 그에 따른 접근 방식을 결정하여 오바마가 재선에 성공한 사례가 있다. 즉, 데이터 마이닝을 우리가 얻고자 하는 결과를 얻을 수 있다. 이러한 여러 이유로 저자는 이러한 빅 데이터와 빅 데이터에서 사용되는 기술들이 사용되고 있는 이 시대에 통계는 빅 데이터를 이루기 위한 도구이고 통계학은 빅 데이터를 지배하는 학문이라고 말하고 있다.

책을 읽으며, 많은 것을 느끼게 되었다. 특히 빅 데이터는 지금 연구하고자 하는 분야이기에 이 책을 골랐고 그 선택에 매우 만족할 수 있었다. 책에서는 통계, 그 중에서도 통계학이라는 학문이 빅 데이터에서 얼마나 중요한 것인가에 대해서 생각해볼 수 있게 되었다. 더불어서 통계는 논문에서 자신의 주장을 신뢰받을 수 있도록 하기 위해 사용되는 중요한 증거로서도 사용됨을 이 책을 읽고 기존에 읽던 논문들을 다시 보면서 눈을 뜰 수 있게 되었다. 이 책은 아직 대학원생으로서 많은 것이 부족한 나에게 수학, 그 중에서도 통계가 얼마나 내가 연구하고자 하는 분야에서 자주 사용되고 있는 가, 통계가 어떻게 빅 데이터를 지배할 수 있는 가에 대해 수 많은 예시를 들면서 설명하고 있다. 특히, 데이터 마이닝이라는 기술에서 통계를 무시하고 연구를 하는 것이 얼마나 잘못된 행동인 것인지에 대해 많은 생각을 해볼 수 있게 되었다.

최근 논문을 읽으면서 논문에서 제시한 알고리즘에 대하여, 알고리즘이 얼마나 정확성을 지니는 가에 대해 증명을 하기 위해 논문에서 통계적인 방법을 사용하여 증명을 했으나, 그것을 알지 못하여 논문을 이해하는 것에 있어서 많은 시간을 할애한 적이 있다. 하지만, 그 통계적인 방법을 단순히 그 논문에서만 사용하는 것이라 생각하고 넘겨버렸으나, 이후에 다시 그 통계적인 방법을 만났을 때, 다시 그 방법을 이해하느라 많은 시간을 할애하여 불필요하게 시간을 낭비한 경험을 이 책을 읽으며 내 행동이 얼마나 어리석었으며, 내가 얼마나 무지 했는지에 대하여 많은 반성을 하는 계기를 가지며, 책을 통해 나의 부족함을 깨달으며 많은 교훈을 얻을 수 있었다.



그림 . 1분동안 인터넷에서 생성되는 데이터의 양