ISSN 2005-1379

Window of Statistics

통계의 창

2020
SUMMER
Vol.25

이슈

데이터 3법, 데이터 경제의 전망

데이터 3법 개정에 따른 공공데이터 추진 방향
데이터 3법에 따른 헬스케어 데이터 경제 활성화를 위한 도전과 과제
「데이터 거래」 현대판 김선달일까 시장의 혁신일까

통계광장

데이터 리터러시 : 텍스트 마이닝으로 뉴스를 분석하기
통계의 오류와 진실 : 코로나 사태 이후 되짚어보는 빅데이터
통계로 들여다보는 바이러스와 인간의 전쟁
올해 인구주택총조사는 무엇이 바뀌고 어떻게 진행될까

통계탐방

아시아 · 태평양 범죄통계 협력센터

PEOPLE

코로나 이후, 데이터가 바꾸는 세상
데이터 이야기꾼 신현호

목차

통계의 창 2020 Summer Vol.25

발행일 | 2020년 5월 29일
발행인 | 임병권
발행처 | 통계교육원
기 획 | 김정란, 김경환

주 소 | 대전광역시 서구 한밭대로 713(월평동) 통계센터 통계교육원
전 화 | 042-366-6151
팩 스 | 042-366-6498
이메일 | duli007@korea.kr

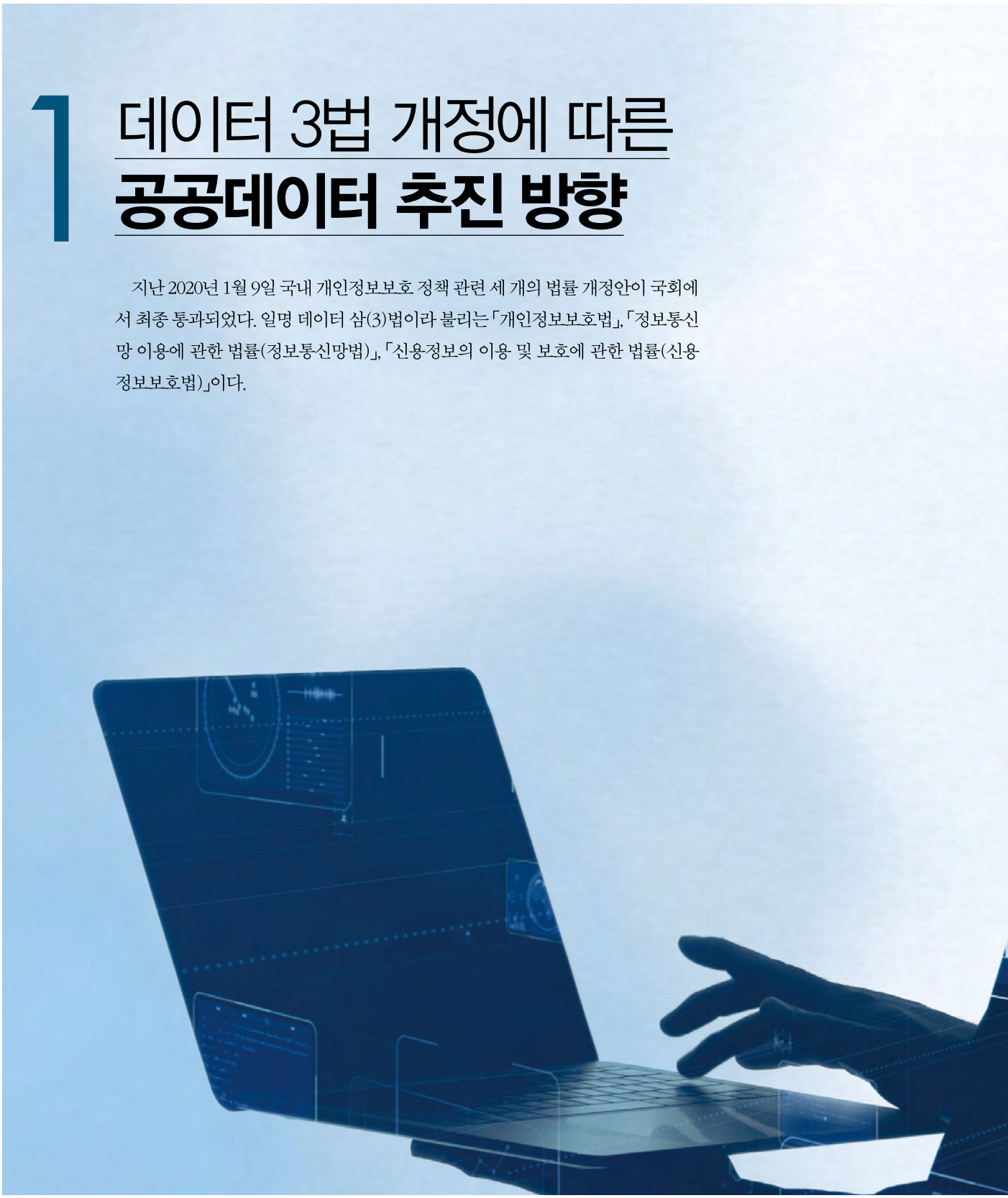
디자인 | ALL contents group
인 쇄 | 올인피앤비(031-955-9497)
ISSN 2005-1379

※ 『통계의 창』에 실린 내용은 필자 개인의 의견이므로 필자의 소속기관이나 본지의 공식적인 견해를 대변하는 것은 아닙니다.

신신애 ◆ 한국정보화진흥원 공공데이터기획팀장

1 데이터 3법 개정에 따른 공공데이터 추진 방향

지난 2020년 1월 9일 국내 개인정보보호 정책 관련 세 개의 법률 개정안이 국회에서 최종 통과되었다. 일명 데이터 삼(3)법이라 불리는「개인정보보호법」,「정보통신망 이용에 관한 법률(정보통신망법)」,「신용정보의 이용 및 보호에 관한 법률(신용정보보호법)」이다.



이번 개정의 목적은 4차 산업혁명 시대를 맞아 핵심 자원인 데이터의 이용 활성화를 통해 신산업을 육성하기 위한 것으로, 인공지능(AI), 클라우드, 사물인터넷(IoT) 등 신기술을 활용한 데이터 이용이 중요해짐에 따라 안전한 데이터 이용을 위한 사회 규범을 정립하기 위한 것이다. 또한 개인정보 개념의 모호성 등으로 인한 기존의 한계를 보완하고, 그동안 행정안전부·방송통신위원회·개인정보보호위원회 등으로 분산된 개인정보보호 감독 기능과 개인정보보호 법령을 「개인정보보호법」과 개인정보보호위원회로 일원화하는 것이다. 데이터 3법 개정으로 우리나라의 공공데이터 개방 및 이용 활성화 정책에도 변화가 필요할 수밖에 없다.

데이터 활용 관련 법의 주요 개정 내용을 살펴보면

데이터 3법 개정 사항 중 가명화 등 데이터 활용 관점에서의 관련 사항을 담고 있는 개인정보보호법과 신용정보보호법의 주요 개정 내용은 다음과 같다.

① 개인정보보호법

○개인정보의 개념과 법 적용 범위의 명확화

먼저 개인정보를 성명, 주민등록번호 및 영상 등 특정 개인을 알아볼 수 있는 정보나 정보집합물, 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보로 명확화하였다. 이 경우, 쉽게 결합할 수 있는지 여부는 다른 정보의 입수 가능성 등 개인을 알아보는 데 소요되는 시간, 비용, 기술 등을 합리적으로 고려하여야 함을 제시하였다

또한 시간·비용·기술 등 개인정보 처리자가 활용할 수 있는 모든 수단을 합리적으로 고려할 때 다른 정보를 사용해도 더 이상 개인을 알아볼 수 없는 정보를 익명정보로 규정하고, 익명정보는 개인정보보호법의 적용 대상이 아님을 명시하였다.

○가명정보 개념 도입 및 가명정보와 개인정보의 이용범위 확대

대통령령으로 정하는 바[1]에 따라 추가정보 없이 특정 개인을 알아볼 수 없게 조치한 정보를 가명정보로 규정하고, 가명정보는 정보주체의 동의 없이 통계작성, 연구, 공익적 기록보존의 목적으로 이용·제공이 가능함을 규정하였다.

또한 개인정보 처리자가 정보 주체의 동의 없이 가명으로 처리된 개인정보를 추가로 이용하거나 제공할 수 있게 했다. 다만 추가 개인정보 수집은 기존 개인정보 수집 목적과 관련성이 크고 정보 주체나 제3자의 이익을 침해하지 않아야 한다는 조건이 있다.

---

1) 개인정보처리자는 가명정보와 추가정보(가명정보를 원래 식별가능한 정보로 복원시킬 수 있는 정보)에 대한 안전성 확보 조치를 하도록 규정. 추가 정보는 별도 분리 보관하고 추가 정보에 대한 접근 권한을 분리하여야 함

서로 다른 기업/기관 간에 가명정보들의 결합은 개인정보 처리자 간에 임의로 수행할 수 없고, 법상의 기준에 따라 지정받은 결합전문기관을 통해서 결합할 수 있으며, 결합된 데이터의 활용은 결합전문기관의 분석공간에서 활용하도록 제한하였다. 다만 결합전문기관의 분석공간에서 결합 목적을 달성하기 어렵고 분석공간을 이용하기 힘든 경우 결합전문기관의 평가를 거친 후 승인을 받아 반출할 수 있도록 규정하였다.

참고 | 개정 개인정보보호법상의 개인정보 · 가명정보 · 익명정보 개념

1. 개인정보 : 해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 정보도 포함하되, 쉽게 결합할 수 있는지 여부는 다른 정보의 입수 가능성 등을 합리적으로 고려해야 함을 명시
2. 가명정보 : 특정 개인을 알아볼 수 없게 조치(가명조치)한 정보로, 가명 조치의 기준과 방법 등 세부내용은 대통령령에 위임
3. 익명정보 : 활용 가능한 모든 수단을 합리적으로 고려할 때 더 이상 개인을 알아볼 수 없는 정보로, 개인정보 보호법의 적용 대상이 아님

② 신용정보보호법

○금융 분야 가명정보 활용 명확화

개인정보보호법과 같은 가명정보는 통계작성(상업적 목적 포함), 연구(산업적 목적 포함), 공익적 기록보존 목적으로 동의 없이 활용 가능함을 규정하였고 결합전문기관을 통해 서로 다른 기업/기관 간에 가명정보들을 결합할 수 있도록 하였다. 다만 신용정보보호법에서는 결합된 가명정보를 결합전문기관의 분석공간에서 사용하도록 규정하지 않고 결합전문기관이 결합된 데이터를 결합 요청 기관에 제공할 수 있게 하였다.

○금융 분야 마이데이터 산업 도입

신용정보보호법은 추가로 정보 주체의 권리행사에 따라 본인정보 통합조회, 신용·자산관리 등 서비스를 제공하는 마이데이터(MyData) 산업 도입 근거 규정을 신설하였다. 정보주체는 개인신용정보 전송요구권에 따라 금융회사, 상거래기업, 공공기관 등이 보유한 각종 개인정보를 금융회사, 개인신용평가회사, 마이데이터 사업자 등에게 보낼 수 있게 하였다.

공공데이터 정책 방향은 어떻게 바뀌나

공공데이터는「공공데이터 제공 및 이용 활성화에 관한 법(이하 '공공데이터법'이라 함)」제17조(제공대상 공공데이터의 범위)에 따라 공공기관이 보유·관리하는 공공데이터를 국민에게 제공해야 함을 규정하고 있다. 다만 법 제17조 1항 1호에 따라 정보공개법상 제공 제외 대상을 공공데이터법에서도 제공 제외 대상으로 규정하고 있다. 정보공개법은 비공개대상정보에 성명·주민등록번호 등 개인에 관한 사항으로, 공개될 경우 사생활의 비밀 또는 자유를 침해할 우려가 있다고 인정되는 정보를 규정(정보공개법 제9조 제1항 제6호)하고 있다. 단, 공익적 차원에서 필요성이 인정되는 경우는 예외로 한다. 따라서 공공데이터법은 사생활 침해 중심으로 규정하고 있는 정보공개법을 준용하고 있어 별도의 법 개정은 필요하지 않을 것으로 판단하고 있다. 또한 개인정보보호법에서의 가명정보 활용은 통계목적, 연구, 공익적 기록보존 등 특정 목적을 위해서만 이용될 수 있으므로 가명정보를 제공할 경우 조건부 제공이 필수적이다. 그러나 공공데이터법은 데이터의 상업적 활용을 포함한 자유로운 활용을 전제로 개방을 하고 있어 공공데이터법의 정책 취지와는 다소 차이가 있다.

그러나 지난 2018년 공공부문 전체의 보유 데이터에 대한 전수조사 결과 개인정보 사유로 비공개 유형으로 분류된 데이터가 약 32%에 해당하여 개인정보 분리 또는 익명화·가명화를 통한 제공이 가능한지 검토해 제공으로 전환을 유도할 필요가 있다.

이에 행정안전부는 5월 6일 공공데이터법에 의한 4기 공공데이터전략위원회 회의에서 '공공데이터 이용활성화 지원 전략'을 통해 데이터 3법 개정 시행에 대응하기 위한 지원 정책을 발표하였다.

① 공공데이터의 익명화 · 가명화 가이드라인 개발 및 배포

데이터 3법의 하위법령과 고시 등의 정부 정책과 세부 기준에 맞추어 개인정보를 포함한 공공데이터의 익명화·가명화를 위한 절차와 방법, 가명 정보의 제공 및 이용방법 등을 담은 가이드라인을 개발 및 마련할 계획이다. 또한 가명처리 된 공공데이터의 제공 절차 등은 명확하게 '공공데이터 관리 지침(행안부 고시)'에 개정 반영할 예정이다.

②「공공부문 개인정보 가명화 지원센터」설치 · 운영

익명화 및 가명화를 데이터 보유 공공기관이 수행하여야 하나 공공기관의 관련 전문 역량이 부족하고 제공 요청 데이터의 양이 많아 가명화 비용이 많이 수반되는 경우도 발생하게 된다. 특히 공공데이터는 법 제26조 3항에 의거, 제공을 위한 가공의 의무를 부여하고 있지 않은 데 반해, 익명화·가명화는 가공을 통해 가능하므로 기관들의 적극적 가공을 통한 제공이 어려울 수 있다. 이에 따라 익명화·가명화 방법에 대한 교육, 컨설팅 등을 제공하고 익명화·가명화에 필요한 행정적 지원 등을 위한 전문 지원체계를 설치·운영할 필요가 있다.

---

2) 개방이 어려운 데이터를 활용하여 분석 후 결과만 반출 가능한 폐쇄적 분석공간

③ 가명정보의 연계 · 융합을 촉진하기 위한 '공공부문 데이터 결합 전문기관' 지정

공공부문 가명정보의 결합수요에 대응하기 위해 법에 따른 결합전문기관 중 공공부문의 데이터 결합을 지원하는 전문 기관을 신속히 지정할 계획이다.

④ 공공데이터의 안전한 분석을 위한 공간(안심구역) 구축

개방이 어려운 데이터도 분석을 위해 활용할 수 있는 안전한 분석공간(안심구역[2])을 공공부문에 구축하여, 공공데이터를 활용하는 민·관의 다양한 연구와 비즈니스 모델 개발을 지원한다.

법 개정이 데이터 경제에 주는 의의와 한계는 무엇인가

8월 5일 개정된 데이터 3법이 시행된다. 풍부한 데이터 활용에 대한 기대가 큰 반면에 기관 및 기업의 가명화 방법과 절차가 성숙되기 전까지 가명 정보의 제공에 소극적이고 조심스러울 수 있다. 가명 정보의 재식별 우려가 있는 것도 사실이다.

가명정보를 연구 목적으로 활용할 수 있어 신제품 등의 개발을 위한 연구, 과학적 행정과 정책 수립을 위한 모델 연구 등에 데이터 활용이 가능해진다. 그러나 모델을 수립한 후 서비스 등 운영으로 들어갔을 때에는 가명정보를 적용, 활용할 수 없는 한계를 가진다.

그러나 기대와 우려, 한계에도 불구하고 데이터 3법 개정안은 많은 기회를 제공할 수 있기에 가명정보의 활용이 목적에 맞게 성공할 수 있도록 모두의 노력이 필요하다.

한현욱 ◆ 차의과학대학교 의학전문대학원 교수

2 데이터 3법에 따른 헬스케어 데이터 경제 활성화를 위한 **도전과 과제**

4차 산업혁명 성공의 핵심은 데이터에 있다. 다양한 수준의 데이터를 연결하고 가공하여 이를 가장 잘 사용할 수 있는 자가 미래사회의 최종 승자가 될 것이다. 우리는 지금 천문학적인 데이터의 홍수 속에 살고 있다. 그리고 미래사회에서 데이터의 양은 광활한 우주의 크기만큼 훨씬 더 방대해질 것이다.

이러한 데이터 중에 가장 방대하게 생산되는 데이터가 바로 인체에서 생산되는 헬스케어 데이터이다. 혹자는 한 개인이 평생 동안 생산하는 헬스케어 데이터가 1,100 테라바이트(TB) 이상이라 말한다. 하지만 이 또한 현재의 인체 데이터를 생산하는 기술을 고려한 추측일 뿐, 미래에는 더 다양한 방법으로 인체의 신호를 측정할 수 있을 것이기에 헬스케어 데이터의 종류와 양은 우리가 지금 추측하는 것보다 훨씬 더 크고 다양할 수 있다.

전 세계는 지금 헬스케어 데이터와 이와 관련된 산업 변화에 이미 주목하고 있다. IBM은 이미 20여 년 전부터 헬스케어 데이터가 미래 의료를 변화시킬 혁신의 키워드가 될 것이라 예측하고 인공지능 의사 왓슨(Watson) 프로젝트에 착수했다. 구글(Google)은 빅데이터와 인공지능을 이용해 독감 유행 예측의 가능성을 타진하기도 했다.

현재 전 세계 다양한 의학 저널에서는 헬스케어 빅데이터와 디지털 헬스케어를 주제로 엄청난 양의 연구논문을 쏟아내고 있다. 심지어 기존의 가장 보수적인 성격의 전통적인 의학저널에서 조차 데이터 기반의 '디지털 의학'으로 변신을 꾀하고 있다. 의료 분야에서 그것이 연구가 되었건 산업이 되었건 결국 헬스케어 데이터가 중심이 될 것이다. 이러한 시대에 우리는 지금 이러한 생태계의 주도권을 잡기 위해 보다 더 전략적 접근이 필요하다.

산업 발전과 개인정보보호라는 두 마리 토끼를 쫓는 데이터 3법

이러한 시대적 흐름에 발맞추어 우리나라도 지난 2020년 1월 9일 데이터 3법 개정이 국회를 통과했다. 데이터 3법이란 「개인정보 보호법」, 「정보통신망 이용 촉진 및 정보보호 등에 관한 법률」 그리고 「신용정보의 이용 및 보호에 관한 법률」이다. 그동안 개인정보보호에 관한 법령이 여러 소관 부처로 나뉘어 발생한 중복 규제를 없애, 향후에는 개인과 기업이 정보를 보다 더 효과적으로 활용할 수 있게 하겠다는 취지로 개정된 것이다. 이를 통해 빅데이터의 활용을 촉진하고 부작용을 방지하여 빅데이터 분야 산업 발전과 함께 개인의 정보보호도 함께 강화하여 두 마리 토끼를 잡아보겠다는 의도이다.

그동안 우리나라 법령에서는 개인정보의 개념이 다소 모호하고, 각 소관부처마다 개인정보에 관한 해석 방법과 관리감독의 체계가 다르며, 데이터를 다루는 기업의 업무 범위가 불확실해 기업의 비즈니스 또한 합법과 불법 사이에서 줄타기를 하고 있었다. 이러한 시점에 데이터 3법 개정 법령이 국회 본회의에서 통과된 것은 매우 고무적이라 평가할 만하다. 그렇다면 이러한 데이터 3법 개정이 의료 분야에도 합리적으로 적용되어 헬스케어 데이터 산업 분야 혁신을 과연 촉진할 수 있을까?

이번에 개정된 데이터 3법의 핵심은 가명정보 개념을 도입했다는 것이다. 가명정보란 '추가적인 정보의 사용 없이는 개인을 식별할 수 없는 정보'를 말한다. 그리고 이러한 가명정보는 통계작성, 과학적 연구 및 공익적 기록 보존 등에 개인의 동의 없이 자유롭게 데이터를 사용할 수 있게 한다. 또한 어떤 방식이던 상관없이 더 이상 개인을 알아볼 수 없게 조치한 익명정보의 경우, 목적에 상관없이 자유롭게 활용하게 하였다. 한편, 개인정보처리자에게 데이터 활용에 대한 책임을 강화하여, 안전조치의 의무를 소홀히 하거나 위반할 경우 형사적인 처벌과 함께 과징금 부과에 관해 구체적으로 명시했다. 뿐만 아니라 별도의 개인정보 감독기구를 두어 유럽의 GDRP에 대응해나갈 수 있도록 하였다.

데이터 3법이 헬스케어 데이터 산업에 주는 의미와 한계점

그렇다면 이러한 데이터 3법 개정이 헬스케어 분야에도 낙관적인 미래를 만들어낼 수 있을까? 현행 「생명윤리 및 안전에 관한 법률」에 의하면 개인정보, 개인식별정보, 유전정보 및 개인 건강에 관한 정보 등은 「개인정보보호법」에 생명윤리 및 안전에 관한 특별한 규정이 없는 한 「생명윤리 및 안전에 관한 법률」의 적용에 따르게 되어 있다. 따라서 인체 유래물에 기반한 연구는 「생명윤리 및 안전에 관한 법률」이 우선 적용되므로 데이터 3법의 개정되었다 하더라도 헬스케어 분야는 「생명윤리 및 안전에 관한 법률」에 영향을 받는다.

또한 「생명윤리 및 안전에 관한 법률」에 의하면, 인간 대상 및 인체 유래물 등에 관한 연구를 하려면 그 연구를 수행하기 전에 대상 기증자로부터 반드시 연구목적에 따라 서면 동의를 받아야 한다. 동의서에는 '연구의 목적'을 초반부터 명확히 해야 하는데, 이럴 경우 초반에 예상하지 못한 연구를 진행하는 데 어려움을 겪을 수 있다. 즉, 기존의 데이터를 이용

해 새롭게 계획한 연구를 위해 기증자에게 다시 연구에 관한 서면동의를 일일이 받아야 하는데, 익명화가 이루어진 기증자를 찾아가 동의를 받는다는 것은 현실적으로 불가능에 가깝다.

한편, 「생명윤리 및 안전에 관한 법률」에 의하면, 모든 인체대상 연구는 결국 기관생명윤리위원회의 심의 승인이 필요하다. 이 경우, 개인정보의 가명처리나 익명처리와는 전혀 무관하게 필수적으로 이루어진다.

새롭게 개정될 데이터 3법에도 모호한 면이 많다. '과학적 연구 및 공익적 기록 보존'이라 함은 과연 어디까지일까? 특히 의료 분야에서는 연구와 산업이 혼재하는 경우가 많다. 많은 경우, 디지털 헬스케어 기업에서는 새로운 비즈니스 서비스를 개발하고 나서 특허출원만 고민하는 것이 아니라 이를 저명한 학술지에 게재하기 위해 노력하는 경우가 많다.

최근에는 '디지털 치료제'라는 개념이 등장하여, 기업이 개발하는 제품도 데이터에 근거한 임상적 유효성 확보에 고민하고 있다. 만약 기업이 병원에서 수집한 다양한 수준의 데이터를 디지털 치료제 개발, 즉 상업적 목적으로 이용하려고 한다면 이는 과학적 연구나 공익적 목적이라고 주장할 수 있을까? 만일 기업에서 의료기관에 가명화 혹은 익명화된 의료정보를 요청할 경우 의료기관이 이를 자연스럽게 수용해 제공해줄 것인가?

의료 분야의 빅데이터는 현재에도 연구와 비즈니스의 중간쯤 어딘가, 모호한 경계에 위치해 있다. 새로운 혁신적인 치료제를 만들어 인류의 건강 문제를 해결해줄 것이라는 점에서는 공익적 성격이 강하지만, 기업이 이를 통해 비즈니스를 확장해나가면서 금전적 가치를 만든다는 점에서는 사익을 추구하는 것이기 때문이다. 따라서 보다 구체적인 시행방안이 수립되어야 한다. 단순히 연구의 차원이 아닌 산업적인 면에서 접근이 필요한데, 이를 반영할 경우 시민사회의 극심한 반발이 예상된다.

「보건의료 빅데이터 플랫폼」 서비스와 의료법의 해결 과제

현재 보건복지부를 중심으로 '보건의료 빅데이터 플랫폼' 시범사업이 실시되고 있다. 2018년 초기 시범사업자 선정이 이루어지고 나서 근 1년간 첨예한 대립 속에 플랫폼이 구축되지 못하고 있다가 최근에 들어서야 시스템이 구축되어 이제 막 한발을 내딛게 되었다.

'보건의료 빅데이터 플랫폼'의 핵심은 보건복지부 산하 건강보험심사평가원, 국민건강보험공단, 질병관리본부 및 국립암센터, 4개 기관이 보유한 빅데이터를 개인을 중심으로 연결하여 하나의 데이터 세트로 만들고 비식별화여 익명화 처리한 후 연구를 목적으로 누구나 이용할 수 있게 하는 것이다.

4개 기관에 분산된 데이터를 하나의 풀로 묶는 것은 기술적으로는 어렵지 않다. 그런데 이러한 단순한 기술이 현재의 법령에서는 원칙적으로 불법으로 규정되고 있다. 큰 울타리 내에서는 보건복지부 산하 기관이라면 보건복지부가 하나의 커다란 조직이라고 생각할 수도 있겠으나, 시스템이 상이하고, 관리주체가 다르기 때문에 서로 다른 데이터라고 할 수 있다. 같은 상위기관을 공유하고 있음에도 불구하고, 이와 같이 데이터와 관련된 사항은 개인정보보호법 아래 그 어떤 것도 할 수 없는 처지였다.

현재의 의료법 또한 개선이 필요하다. 현재의 의료법은 개인의료정보에 대해 정의하고 있지 않기 때문에 개인정보보호법의 적용을 받고 있다. 따라서 가명화를 통한 개인의료정보의 활용에 대해 적합한 근거를 마련해야 한다. 개인의료정보에 관한 문제가 발생할 경우 개인정보보호법과 의료법 중 어떤 것을 우선 따라야 하는지에 관해 논란이 발생할 수 있기 때문이다. 이러한 이유로 정부는 의료 분야의 특수성을 인정하고, 의료 분야의 개별 사안에 대해 개인정보위원회를 통해 복지부와 논의 중이다.

개인정보위원회에서는 의료 분야에 관련된 민감한 개인 의료정보에 관해 복지부에 권고하고, 보건복지부에서 최종적인 판단을 담당한다. 결국 개인정보보호법(일반법)인 데이터 3법이 국회를 통과하였지만, 의료 분야의 경우 특별법인 의료법에 의해서 특정한 의료정보의 경우 가명화 처리가 불가능할 수 있다.

따라서 의료 분야에서는 데이터 3법에 의한 완벽한 의미의 데이터 활용은 쉽지 않을 전망이다. 행정안전부에서는 "원칙적으로 개인정보보호법을 따라야 하지만 의료법등 개별법에서 보호의 취지가 있는 경우 특별법을 따르는 것이 맞다"는 의견이다. 결국 의료분야에서 '보호'의 수준이 어디까지인지가 관건이 될 것이다. 의료 분야에서 데이터 보호를 주장하는 시민사회단체나 의료기관을 통해 수렴된 의견이 과하게 반영될 경우, 이번 데이터 3법의 개정은 의료 분야에서는 영향력을 발휘하기 어려울 수 있다.

바이오 헬스 분야 발전을 위해 주어진 과제

데이터 경제의 미래가치 면에서 최고의 분야는 아마도 바이오 헬스 분야일 것이다. 이러한 이유로 이번 데이터 3법 개정에 대해 의료계와 관련된 산업체의 관심이 뜨겁다. 그동안 산업체에서는 헬스케어 데이터를 합법적으로 확보하는 것이 매우 어려웠을 뿐만 아니라, 확보된 이후에도 생명윤리법 및 의료법에 의해 확보된 데이터를 제대로 활용하는 데 걸림돌이 많았다. 데이터 3법 통과 후에도 민감 정보로 분류된 유전정보의 경우 결국 개별적으로 환자 동의를 받아야하기 때문에 이를 적극적으로 활용하는 데는 여러 한계점이 있을 것으로 예상된다.

어떤 이유에서건 개인이 건강에 대한 차별을 받아서는 안 된다. 시민사회단체에서 우려하는 것 중에 하나가 바로 헬스케어 분야와 관련된 데이터 개방으로 인해 개인이 차별받게 될지도 모른다는 것이다. 그리고 이로 인해 발생하게 될 사회적 불평등과 소외 등의 문제를 어떻게 해결할 것이냐에 대해도 의문을 갖고 있다. 하지만 시민사회단체에서도 시대적 흐름에 맞게 보다 전향적인 생각의 전환이 필요하다. 데이터 개방으로 인해 개발되는 새로운 현대의료 기술의 최대 수혜자는 결국 질병을 앓고 있는 환자 자신이 될 것이기 때문이다. 이것이 개인정보위원회에서 개별 사안에 대해 복지부가 시민사회단체 및 의료계의지지 하에 어떻게 협의할지 궁금한 이유이다.

민경영 ◆ MBN 기자

3 「데이터 거래」 현대판 김선달일까 시장의 혁신일까

지난 1994년, 정부는 그동안 금지됐던 생수의 국내 판매를 공식 허용했다. 1988년 국내에서 첫 생수 제품이 개발된 지 6년 만이었다. '식수=끓인 물'이라는 공식이 상식이었던 시절이었다. 곳곳에서 "사회적 위화감을 조성한다"거나 "수돗물에 대한 불신을 조장할 것"이라는 반발과 "과연 생수 산업이 성공할 수 있을까"라는 우려가 함께 터졌다.

지금은 어떨까? 1994년 1,000억 원이었던 국내 생수시장 규모는 2018년 약 1조 3,600억 원까지 성장했다. 13배 넘게 급성장한 건데, 2023년에는 2조 원을 넘길 것이란 예측도 조심스레 나오고 있다.

이뿐인가. 산소도 캔에 넣어 판매되기 시작했고, 개인 동영상을 온라인에 올려 떼돈을 버는 유튜버도 불과 몇 년 전까진 상상할 수 없는 일이었다. 이런 획기적인 산업이 등장할 때면, 누군가는 항상 '현대판 봉이 김선달'이라며 비웃거나 비난하기도 한다. 그러나 확실한 건, 이 김선달들이 시장을 혁신하고, 새로운 부(富)를 창출한다는 것이다.

데이터를 사고 파는 시대 도래… 국내 민간 데이터거래소 출범

데이터 3법이 통과되면서, 이제 데이터를 사고 파는 시대가 열렸다. 데이터 선진국에서는 일찌감치 데이터 거래 산업이 활성화됐다.

전 세계에는 현재 4천 곳이 넘는 데이터거래소가 있다고 하는데 세계에서 데이터 시장 규모가 가장 큰 미국의 경우 650곳 이상의 민간 데이터거래소가 있으며, 유럽에서는 총 917곳의 데이터거래소가 활발하게 운영되고 있다.

아시아에서는 중국이 지난 2014년 구이양 데이터거래소가 설립된 데 이어, 상하이 등 중국 전역에서 20개 이상의 데이터거래소가 문을 열었다. 일본 역시 59개 기업과 단체로 구성된 민간 데이터거래소가 지난 2018년 10월 출범하였고, 2021년까지 거래 총액 300억 원이 목표라고 밝히고 있다.

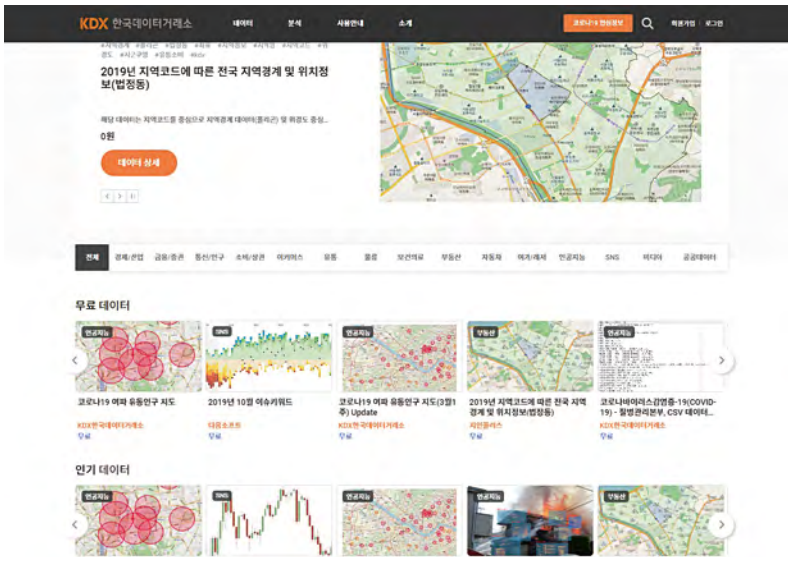
한국은 지난 2019년 12월 2일, 국내 최초의 민간 데이터거래소인 KDX 한국데이터거래소(이하 KDX)가 설립되었다.

KDX에는 주관사인 종합편성채널 MBN과 함께 거래소 구축을 담당한 플랫폼사(삼성카드, SK텔레콤, SK플래닛, CJ올리브네트웍스, 웰컴에프앤디, GS리테일), 데이터 공급을 담당

한 센터사(한국우편사업진흥원, KCB, 다음소프트, 지인플러스, 식신, 로플랫, 빌트온, NICE 디앤알, 데이블, 온누리H&C) 등 총 17개 기업이 참여했다.

이는 과학기술정보통신부가 추진하고 있는 10대 빅데이터 플랫폼 구축 사업의 한 부분으로, KDX는 현재 총 97억 원에 달하는 2,910건의 데이터 상품이 등록하고 '유통·소비 분야'를 담당하고 있다. KDX는 출범식에서 "유통·소비 분야에 한정되지 않고 종합 데이터 거래 플랫폼으로 발전해 국내 데이터 거래 생태계 조성을 이끌겠다"고 밝혔다. 따라서 향후 금융, 통신, 물류, 상권, 부동산, 소셜, 미디어, 의료 등 분야로 확대해 나갈 계획이다.

KDX 웹사이트



출처 : https://kdx.kr/main

누가 어떤 데이터를 구매할까

데이터거래소의 첫 거래 데이터는 바로 MBN이 그동안 보도했던 화재 관련 뉴스 동영상이었다. 이를 구매한 딥러닝 전문 업체 씨이랩은 "아무리 AI가 발달해도 빅데이터가 없으면 무용지물이다. 고성능 장비가 동원된 사건사고 영상을 구하기가 어려운 탓에 CCTV에 인공지능을 적용하는 데 어려움이 많았는데, KDX를 비롯해 영상 데이터를 구할 수 있는 채널이 많아지면서 CCTV와 같은 스마트 보안 시스템도 고도화할 수 있게 됐다"고 밝혔다.

2019년 12월 2일 〈MBN 종합뉴스〉 보도

최근 들어 코로나19에 대한 연구의 필요성이 높아지면서 이와 관련된 데이터를 많이 찾고 있다. 한국은행과 중소기업중앙회도 KDX를 통해 로플랫이 수집, 가공한 〈전국 유동인구 데이터〉 상품을 구매했다. 이들은 이 데이터 상품을 활용해 코로나19가 전국 실물 경제에 끼친 영향을 분석할 예정이다.

데이터 3법 통과로 고품질의 데이터 상품이 제공되면서 앞으로 더 활발한 데이터 거래가 예상된다.

2020년 5월 7일 매일경제신문 보도

MK 매일경제 MBN

뉴스 오피니언 프리미엄 연예 스포츠 증권 부동산 교육 비즈&

뉴스홈 전체기사 경제 기업 사회 국제 부동산 증권 정치 IT·과학 문화 기획·연재 Special Edition 인기뉴스

韓銀, KDX `코로나 데이터` 샀다

로플랫 유동인구 데이터 구매

이동인 기자 | 입력 : 2020.05.07 09:34:23 수정 : 2020.05.07 20:38:09 0

△KDX한국데이터거래소가 로플랫 위치 기반 데이터를 가공해 만든 4월 4째주 코로나19 유동인구 지도

한국은행이 코로나19가 실물경제에 끼친 영향을 분석하기 위해 KDX한국데이터거래소가 판매한 빅데이터를 활용한다.

6일 한은은 국내 위치 인식 데이터 전문 기업 로플랫이 KDX한국데이터거래소를 통해 판매한 코로나19 유동인구 데이터를 구매했다. 전국 와이파이 기반 위치 데이터 20억여 건을 분석한

개인도 데이터 거래를 할 수 있을까

현시점 KDX의 총 회원 수 1,941명(5월 12일 기준)을 달성했다. 2020년 한 해 목표치였던 회원 수 1,000명을 2배나 빠르게, 2배나 많이 달성하였다. 이는 많은 시민들과 기업들이 데이터 거래에 대해 큰 관심을 보이고 있다는 증거다.

그렇다면 KDX의 데이터 거래 어떻게 하면 될까? 우선 KDX의 회원은 '개인 회원'과 '법인 회원'으로 구분된다. 데이터의 품질 보증 문제 때문에 개인 회원은 데이터 판매는 불가능하고, 무료로 가입해 언제든지 데이터 구매를 할 수 있다. 구매는 신용카드와 가상계좌를 통해 온라인으로 간편하게 이뤄진다.

법인 회원은 연회비를 내고 가입한 뒤, 판매자 등록을 신청하면 내부 검토를 거쳐 판매자 자격이 주어진다. 판매자로 등록되면 KDX가 데이터 표준 지침서를 발송하고, 판매기업은 이를 검토해 자신의 데이터 상품에 적절하게 반영한 뒤 상품으로 등록하면 된다.

＊법인 회원

※현재는 법인 회원만 별도 판매자 승인 후 데이터 판매가 가능합니다.

분류		연회비	이용 가능 서비스 및 ID 할당 수				
			무료 데이터 이용	유료 데이터 구매	데이터 판매	분석 환경	ID 할당 수
법인	법인 고급	1천만 원	○	○	○	○	관리자 1/판매자 9
	법인 일반	2백만 원	○	○	○	○	관리자 1/ 판매자 1

＊개인 회원

분류		연회비	이용 가능 서비스 및 ID 할당 수			
			무료 데이터 이용	유료 데이터 구매	데이터 판매	분석 환경
개인	유료	1백만 원	○	○	×	○
	무료	0원	○	○	×	×

또한 모든 법인 회원과 연회비를 낸 개인 회원은 KDX가 제공하는 분석 툴 제플린, R Studio, Jupyter, Neo4j을 활용할 수 있다. 이 분석 툴을 통해, 데이터 구매자는 따로 대용량 저장 장치를 보유하지 않고도, KDX 내부에서 데이터를 가공·분석할 수 있다.

시작은 늦었지만, 다양한 서비스 개발을 통해 세계로 도약

KDX는 향후, 현재 정부가 추진하고 있는 마이데이터 사업과 인공지능 학습용 데이터 구축 사업에도 참여하겠다고 밝혔다. 또한 시민들의 데이터 인식 재고를 위한 뉴스레터 서비스를 출범하고, 데이터AI 인력 양성을 위한 각종 프로그램을 준비하고 있다.

대한민국의 데이터 시장은 아직 선진국들에 비해 많이 뒤쳐져 있다는 평가를 받는다. 아직 규모도 작거니와 경제 주체들의 데이터 거래에 대한 인식 역시 많이 떨어진다. 그러나 지금이라도 데이터 거래가 싹을 틔우고 있는 것은 고무적인 일이다. IT 강국답게 머지 않아 세계적인 데이터 시장을 구축할 것으로 기대하고 있다.

구자룡 ◆ 밸류바인 대표

데이터 리터러시 : 텍스트 마이닝으로 뉴스 분석하기

시장 환경이나 트렌드를 알고 싶을 때, 신제품 아이디어를 얻고 싶을 때 가장 쉽게 접할 수 있는 데이터는 무엇일까? 아마도 시장에 나와 있는 2차 자료를 먼저 찾아볼 것이다. 또한 각종 설문조사 결과나 전문기관의 분석 자료를 살펴볼 것이다. 이런 자료에서 나에게 딱 맞는 자료가 있다면 다행이지만 대체로 뭔가 부족한 부분 때문에 소비자들에게 직접 물어보고 싶은 욕구가 생긴다. 그래서 내가 목적으로 하는 결과를 얻기 위해 조사 기획을 하고 직접 설문조사를 하면 원하는 결과를 얻을 수 있다고 생각한다.

하지만 트렌드나 신제품 아이디어는 소비자에게 질문한다고 해서 답을 찾을 수 있는 내용이 아니다. 설문조사는 신제품 콘셉트를 제시하고 수용도 조사를 할 때 어느 정도 답변을 들을 수 있다. 하지만 어떤 제품이 필요한지, 어떤 콘셉트로 제품을 만들어야 하는지는 답하기 어렵다. 즉, 설문조사로 기술적인 조사는 가능하지만, 탐색적인 조사에 성공할 가능성은 적다. 마케터나 제품기획자가 신제품에 대한 아이디어를 얻고자 한다면, 설문조사로는 어느 정도 한계가 있는 것이다. 앞에서 언급한 2차 자료에서도 아이디어가 없다면 어떻게 해야 할까?

뉴스 빅데이터 속에 있는 진주를 어떻게 찾을까

이럴 때 소비자에게 직접 물어보지 않고, 시장에서 원하는 뭔가를 찾아내거나 아이디어가 될 만한 키워드라도 얻고 싶은 욕구가 생긴다. 우리 주변에서 매일 접하는 가장 대표적인 데이터로 뉴스, 즉 언론사의 기사가 있다. 뉴스는 그 자체로 빅데이터이며, 대표적인 비정형 데이터다. 데이터의 특성이 비정형 데이터이기 때문에 텍스트 형태로 되어 있다. 텍스트를 분석할 수 있는 텍스트 마이닝으로 키워드 빈도 분석과 연관 키워드 분석, 워드 클라우드 분석을 통해 원하는 결과, 즉 진주를 찾을 수 있다.

예를 들어, 만약 내가 건강기능식품의 마케터나 제품기획자라고 생각해보자. 소비자들의 건강 추구 경향은 오래전부터 있었고, 지금도 지속되고 있는 트렌드로 알고 있다. 소득의 증가, 수명의 연장, 삶의 질 추구 등 소비자의 건강 추구 욕구를 충족시키기 위한 다양한 제품이 이미 시장에 많이 나와 있다. 성숙 시장에 접어든 건강기능식품을 담당하고 있는 마케터라면 혹은 제품기획자라면 어떤 신제품으로 시장을 확대할 수

있을까? 상사에게서 히트할 수 있는 신제품을 개발하라는 주문을 받았다면 이제 어떻게 해야 할까?

나는 마케터로서 혹은 기획자로서 통계분석을 전문적으로 해본 적이 없다. 설문조사를 통해 통계분석의 결과를 활용해왔을 뿐이다. 조사 전문기관에 의뢰하여 설문조사를 수행했고, 그 결과를 받아서 처리했기 때문에 딱히 통계분석에 대해 고민해보지 않았다. 분석 결과를 해석하는 것은 학교에서 배운 확률과 통계 정도로 어느 정도 이해하고 있다. 기술적인 조사라면 조사 전문기관에 의뢰하여 처리하면 쉽게 해결할 수 있다.

그런데 신제품 개발을 위한 탐색적인 조사를 외주로 처리하면 왠지 무능한 사원으로 보일 것 같다. 데이터가 풍부한 빅데이터 시대에 데이터로 증거를 제시해야 하는 상황이 되면서 더욱 스트레스를 받고 있다. 직접 빅데이터를 분석해서 그 결과로 멋진 아이디어를 내고 싶다. 데이터 리터러시가 필요한 순간이다. 문제를 해결할 수 있는 필요한 데이터를 수집하고, 적합한 방법으로 분석하여, 적절하게 활용할 수 있을 때 나의 리터러시 역량은 높아진다.

문제를 정의하고 필요한 뉴스 데이터를 수집하자

데이터 분석을 위해 가장 먼저 해야 할 일은 문제를 정의하는 것이다. 문제는 바로 해결해야 할 과제다. 여기서는 히트할 신제품 아이디어 도출이라고 하자. 예를 들어, 건강기능식품 시장이라고 하고, 이 시장에서 소비자들에게 물어보지 않은 상태에서 물어보는 것보다 더 정확한 신제품 욕구를 찾아보자. 산업 분야의 뉴스에 담긴 텍스트는 주로 기업에서 홍보용으로 제공한 보도 자료를 기반으로 작성된 내용이 많다. 또한 전문가들의 견해나 연구 결과물들이 기사화되기도 한다. 그리고 인터넷과 소셜미디어가 발달하면서 뉴스와 같은 텍스트 데이터가 기하급수적으로 증가하고 있다.

여기에서는 전문적으로 빅데이터를 분석하는 데이터 과학자가 아닌 마케팅 기획자의 관점으로 접근해보고자 한다. 따라서 문과생으로서 학습한 지식과 마케팅 업무 역량 안에서 가능한 방법을 찾아야 한다. 여기서 마케팅 기획자는 일반적인 사무직으로 대체해도 된다. 문제 정의를 현재 담당하고 있는 업무에서 찾고, 관련 검색 키워드만 선정하면 된다. 나머지는 동일하다.

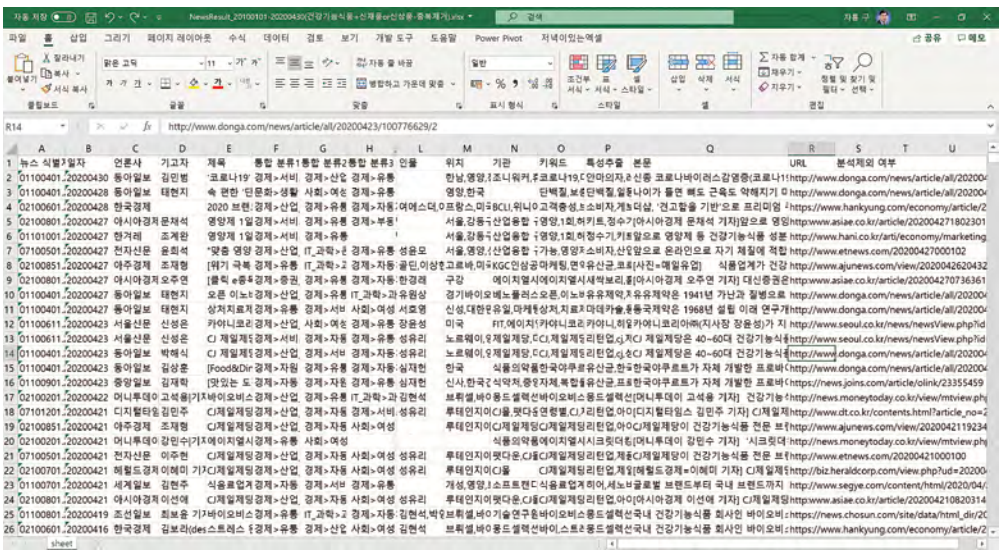
데이터 분석에 앞서 정의한 문제에 대한 데이터를 수집해야 한다. 비전문가가 텍스트 데이터를 수집할 방법으로 빅카인즈(BIG KINDS)를 이용한 웹 크롤링 방법과 MS 파워 쿼리를 이용한 웹 크롤링 방법 등이 있다. 여기서는 빅카인즈에서 텍스트 데이터를 수집하고자 한다.

빅카인즈(www.kinds.or.kr)는 한국언론진흥재단이 운영하는 뉴스 빅데이터 분석 시스템으로 뉴스 속 키워드 관계망, 주요 이슈, 정보원, 이슈 트렌드 분석 정보를 제공하고 있다. 1990년부터 현재까지 54개 언론매체에서 발행한 약 6 천

만 건의 뉴스 콘텐츠를 검색하고 활용할 수 있다. 검색 방법은 간단하다. 빅카인즈 사이트에 들어가서 네이버나 구글에서 검색하듯이 키워드 검색만 하면 대체로 기본적인 분석이 이루어진다. 분석 과정은 3단계로 Step 01. 뉴스 검색, Step 02. 검색 결과, Step 03. 분석 결과 및 시각화로 구성되어 있다.

좀 더 구체적인 맞춤형으로 분석하기 위해서는 무료로 회원가입을 하고 로그인 후 이용하면 수집한 데이터를 엑셀 파일로 다운로드할 수 있다. 엑셀 파일을 다운로드하면 수집한 데이터가 어떻게 되어 있는지 알 수 있다.

빅카인즈에서 수집한 텍스트 데이터의 엑셀 파일 내용

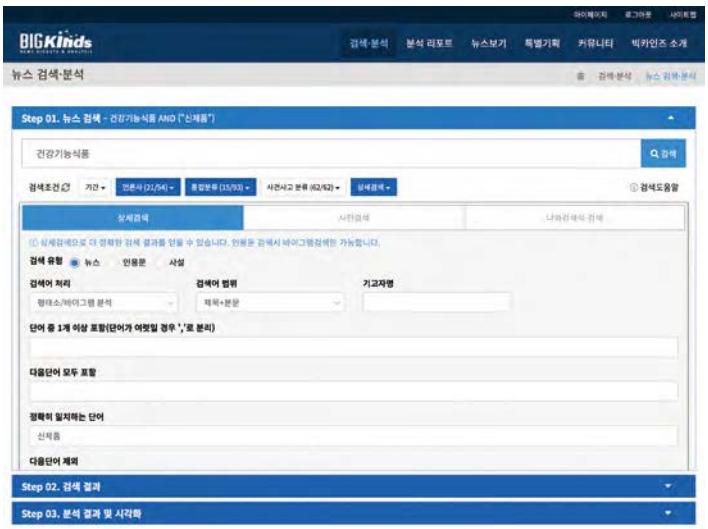


검색한 결과 데이터의 문서(기사)가 많으면 좋긴 하지만 무조건 좋은 것은 아니다. '선택한 키워드가 있는지'로만 검색되기 때문에 문제 정의에 부합하는 문서를 찾아내는 것이 문서의 양보다 더 중요하다. 분석자의 선행지식이나 키워드 선정 등 탐색 활동에 따라 수집 데이터의 품질에 차이가 발생할 수 있다. 좀 더 정확한 데이터를 수집하기 위해서는 관계없는 문서를 사전에 검색되지 않도록 검색조건을 설정할 필요가 있다.

예제를 수행하기 위해 기본적인 검색 키워드로 '건강기능식품'을 선정하고, 기간을 2010년 1월 1일부터 2020년 4월 30일까지로 설정했다. 54개 언론사 중에서 중앙지와 경제지, 전문지로 한정했다. 기업에서 신제품에 대한 보도 자료를 주로 제공하는 언론사로 한정한 것이다. 그리고 정치, 경제, 사회 등 8개의 통합분류 중에서 '경제' 하나만 선택했다. 상세검색에서 제목과 본문에서 '형태소 분석'으로, 그리고 단어 중 1개 이상

포함에 '신제품, 신상품'을 설정했다. 즉, '건강기능식품'과 '신제품' 혹은 '신상품' 키워드가 들어간 문서를 검색한 결과 최종 수집된 문서는 1,822건이다. 검색 결과에서 분석에 사용한 기사는 1,658건이며, 중복, 예외 등으로 분석에서 제외한 기사는 164건이다.

빅카인즈 뉴스 빅데이터 분석시스템의 뉴스 검색



수집된 데이터로 텍스트 마이닝을 하고 시각화하여 제시하자

텍스트 마이닝(text mining)은 자연어로 구성된 비정형 텍스트 데이터에서 특정한 패턴 또는 관계를 추출하여 의미 있는 정보를 찾아내는 기법이다. 즉, 문서 중에 특정 단어가 얼마나 많이 출현하는지 단어 빈도(Term Frequency)를 찾아낸다. 이때 분석에 사용한 데이터는 뉴스인데 문장, 즉 자연어로 되어 있어서 문장 그대로 분석할 수 없다. 하나의 단어로 분리해야 하는데, 이를 형태소 분석이라고 한다.

빅카인즈에서는 R, 파이썬 등 전문적인 빅데이터 분석 언어에서 사용하는 형태소 분석을 시스템에 포함해놓았기 때문에 분석자가 별도로 형태소를 분석하지 않아도 된다. 비전문가도 쉽게 텍스트 마이닝을 할 수 있게 되어 있다.

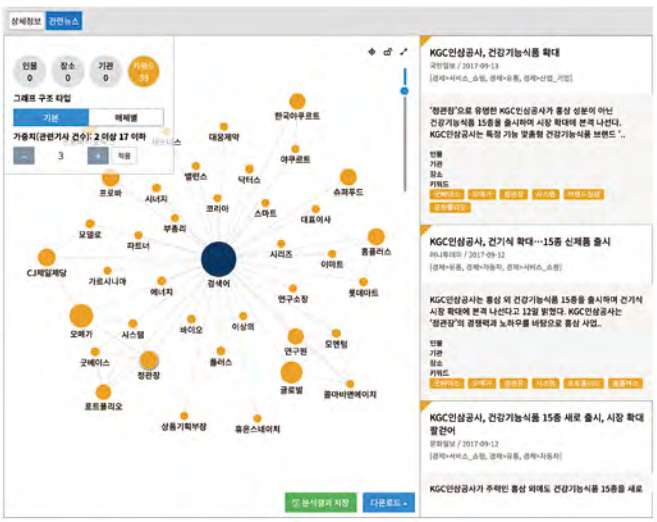
빅카인즈에서는 텍스트 마이닝의 분석 결과 및 시각화의 결과물을 '관계도 분석', '키워드 트렌드', '연관어 분석' 등으로 제공해준다. 관계도 분석을 통해 데이터의 전체적인 특성을 살펴볼 수 있다. 관계도 분석은 검색 결과 중 정확도 상위 100건의 분석 뉴스에서 추출된 개체명(인물, 장소, 기관, 키워드) 사이의 연결 관계를 네트워크 형태로 시각화하여 보여준다. 함께 제공되는 '관련 뉴스'는 '검색 결과 중 정확도 상위 100건의 뉴스'를 최신순으로 정렬한 결과를 보여준다. 관계도 분석 결과를 보면서 '관련 뉴스'

를 함께 살펴보면, 어떤 기사에서 어떤 내용이 언급되고 있는지를 한눈에 볼 수 있다.

예를 들어, 정관장 브랜드에 대해 구체적인 정보를 알고 싶다면, 그림에서 정관장을 클릭하면 관련 기사가 우측에 나타나고 키워드도 표시된다. 여기서 정관장은 오메가, 한국야쿠르트는 프로바이오틱스와 밀접한 관계가 있을 것 같다는 추측을 할 수 있다. 즉, 정관장은 오메가를, 한국야쿠르트는 프로바이오틱스를 원료로 하는 신제품을 출시하고 있는 것을 본문 기사를 통해 확인할 수 있다. 오메가를 기준으로 보면 정관장과 CJ 제일제당도 관계가 깊은 것으로 추측된다.

다만 데이터 전처리 과정을 생략했기 때문에 주제와 관련되어 있지 않은 이상한 키워드가 있을 수 있다. 관계도 분석의 결과를 살펴보면, 관련 기사 건수를 3건으로 했을 때 '부총리, 이상의, 연구소장, 대표이사, 상품기획부장, 연구원' 등의 단어가 불필요하게 느껴진다. 이런 단어로 인해 결과의 내용 타당성을 저해할 수 있다. 이럴 때는 관련 기사 건수를 4건, 5건 등으로 높이면서 불필요한 단어가 나타나지 않는 그림을 최종적으로 선택할 수 있다. 때에 따라서는 검색 결과를 다운로드한 엑셀 파일에서 데이터 전처리를 한 후 빅카인즈가 아닌 다른 분석 방법으로 추가 분석을 하면 해결할 수 있다.

빅카인즈의 분석결과 중 관계도 분석과 관련 뉴스



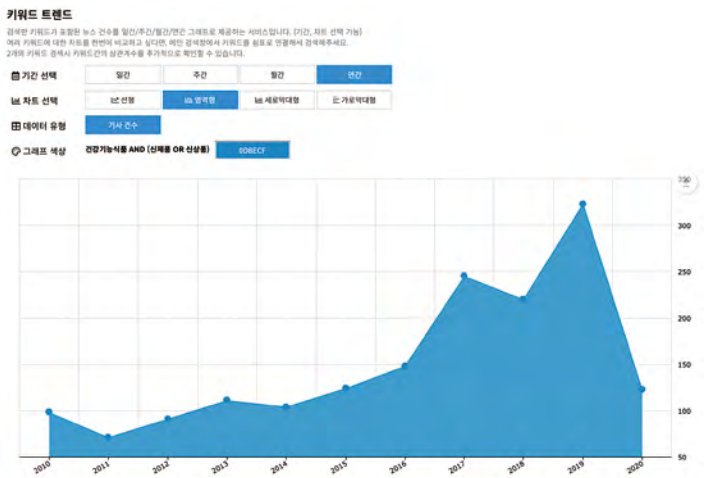
그리고 키워드 트렌드를 연간기준으로 살펴보면, 건강기능식품의 신제품에 대한 기사는 지속해서 상승하는 추세임을 알 수 있다. 특히 2019년도에 가장 많이 언급된 것을 알 수 있다. 2020년은 4월 30일까지 언급 양으로 4개월 동안의 언급 양이 2015년 전체 언급 양과 비슷한 정도로 많다는 것을 알 수 있다. 이는 코로나19와 관련이 있는 것으로 파악된다.

연관어 분석은 검색 결과 중 분석 뉴스와 연관성(가중치, 키워드 빈도수)이 높은 키워드를 시각화하여 보여준다. 텍스트 시각화 방법 중 대표적인 방법으로 워드 클라우드가 있다. 최소의 의미를 지니는 문장 구성 성분인 형태소를 분석하고 그 빈도에 따라 문자의 크기를 나타내는 방법이다.

빅카인즈에서는 별도의 워드 클라우드 분석을 하지 않고도 시각화 결과를 바로 확인할 수 있다. 키워드 중에서 주제와 관련이 없는 단어는 제외해야 하는데, 선택항목으로 분석제외를 할 수 있다.

또한 막대그래프로도 볼 수 있다. 그리고 필요하다면 엑셀 테이블로 연관어 분석 결과를 다운로드할 수 있다. 즉, 단어 빈도(TF)를 쉽게 분석 알 수 있다. 여기서는 '기능성, 소비자, 화장품, 중국, 의약품, 유산균, 프로바이오틱스, 그린알로에' 등의 단어가 많은 빈도수를 나타내고 있다.

빅카인즈의 분석결과 중 키워드 트렌드



빅카인즈의 분석결과 중 연관어 분석

텍스트 마이닝의 결과를 바탕으로 신제품 아이디어를 도출하자

텍스트 마이닝으로 분석한 결과에서 의미 있는 뭔가를 찾아야 분석의 의미가 있다. 예제 분석은 건강기능식품의 신제품 트렌드를 파악하고자 뉴스 1,600여 건의 기사를 분석했다.

그 결과 첫째, 건강기능식품과 신제품 단어가 들어간 뉴스가 지속해서 상승하고 있다는 것을 알 수 있다. 특히 2020년 4개월간의 기사 건수가 2015년 1년의 기사 건수와 유사한 정도로 많다. 기업들이 신제품을 만들어 내는 만큼 소비자들도 관심이 많다는 것을 연관어 분석의 결과에서도 확인할 수 있다. 둘째, 건강기능식품의 주성분으로 오메가, 유산균, 프로바이오틱스 등이 주목을 받고 있다는 것을 알 수 있다. 셋째, 건강기능식품을 생산하는 기업으로 그린알로에, 오리온, CJ제일제당, KGC인삼공사, 에이치엘사이언스 등을 확인할 수 있다.

이상의 결과를 바탕으로 성장세를 보이는 건강기능식품 시장에서 차별화할 수 있는 신제품 아이디어를 찾아야 한다. 새로운 원료나, 새로운 콘셉트나 새로운 가치를 제안하면 새로운 시장을 만들 수 있을 것이다. 마케터나 기획자의 전문성과 통찰력이 요구되는 순간이다.

한때 워드 클라우드 분석이 빅데이터 분석으로 잘못 알려진 적이 있다. 빅데이터를 이용하여 글자의 크기와 색이 다른 멋진 구름을 보고 번뜩이는 통찰을 한 사람도 있지만, 대다수의 사람은 통찰을 하지 못했다. 대부분 텍스트 데이터를 시각적으로 보여주기 위한 하나의 방법으로 워드 클라우드를 사용할 뿐이며, 이것만으로는 빅데이터 분석이라고 할 수 없다.

어쩌면 워드 클라우드는 제대로 된 빅데이터 분석을 위한 탐색적 분석의 하나로 보는 것이 타당할 것이다. 텍스트 마이닝의 결과를 바탕으로 예측분석까지 할 수 있다면 제대로 된 빅데이터 분석을 활용하는 단계까지 접근한 것이다. 텍스트 빅데이터로 주가를 예측하거나 트렌드를 예측하거나 질병을 예측한다면 충분한 가치를 발휘하는 경우들이다. 예측 분석은 빅데이터 분석 전문가의 영역으로 별도의 분석 방법을 사용해야만 가능하다.

지금까지 건강기능식품과 관련하여 신제품(신상품)에 대해 뉴스 빅데이터를 빅카인즈로 수집하고, 텍스트를 분석하고, 그 결과를 시각화해봤다. 데이터 분석 전문가가 아닌 일반적인 마케팅 기획자 수준에서도 비정형 빅데이터를 텍스트 마이닝으로 분석하고, 그 결과를 탐색적 결과물로 활용할 수 있다는 자신감을 가질 필요가 있다. 데이터가 풍부한 시대를 앞서가는 방법의 하나는 내가 직접 데이터를 수집하고 분석하고 시각화하여 직접 사용하는 것이다. 나의 데이터 리터러시를 높이는 것이 최선이다.

조재근 ◆ 경성대학교 수학응용통계학부 교수

통계의 오류와 진실 : 코로나 사태 이후 되짚어보는 빅데이터

우리는 지난 4월 15일에 있었던 국회의원 선거를 앞두고 언론을 통해 많은 여론조사 결과를 접할 수 있었다. 보도가 거의 하루도 빠짐없이 이어지다 보니 여론조사라는 것이 단순하고 쉬운 것처럼 보였는데, 사실 자동응답 방법을 이용하더라도 조사에는 적지 않은 비용이 든다. 그렇다고 예측이 늘 실제 결과와 맞아떨어지는 것도 아니다. 그럼에도 불구하고 선거 때가 되면 전국 또는 지역 단위로 많은 조사가 이루어진다. 심지어 정당들은 선거에 내보낼 후보를 정할 때부터 여론조사를 널리 이용한다고 한다. 국회의원이 되고 싶은 사람들은 유권자들의 선택으로 판가름 나는 본선 투표에 앞서 정당 내부에서부터 여론조사라는 예선전을 무사히 통과해야 하는 셈이다. 물론 그런 후보 선출 방식은 다들 여론조사 결과에 승복하기 때문에 가능할 테다. 그만큼 여론조사의 위력이 상당하다는 의미이다.

예측에 실패한 선거 여론조사들

선거를 앞두고 이루어지는 여러 여론조사들의 대미를 장식하는 것은 역시 투표가 끝난 직후 방송사들이 발표하는 출구조사일 것이다. 선거 전에 이루어지는 조사와 달리 실제 투표를 마치고 나온 사람을 대상으로 하는 출구조사는 선거 결과를 가장 정확하게 예측할 수 있는 방법으로 알려져 있다. 비록 몇 시간 뒤에 정확한 개표 결과를 알 수 있지만 후보자들은 물론 유권자들도 출구조사 결과 발표를 손꼽아 기다리곤 한다. 당연히 방송사들도 많은 공을 들여 출구조사 발표를 준비한다.

사실 지난번 2016년 총선 때 방송사들이 발표한 출구조사는 성공적이지 못했다. 각 지역구의 당선자는 물론이고 제1당이 될 정당을 거꾸로 예측할 정도였으니 당시의 조사는 여론조사의 역사에서 실패로 기록될 만했다. 2020년 선거를 앞두고 조사기관들과 방송사들은 그런 실패를 거듭하지 않기 위해 엄청난 노력을 기울였을 것이다.

그럼 2020년 봄에는 출구조사가 개표 결과와 잘 들어맞았을까? 아쉽게도 그렇지 못했다. 지난번처럼 가장 의석을 많이 차지할 정당을 잘못 예측한 정도로 틀린 것은 아니었지만 출구조사에서 예측한 정당별 당선자 수는 거의 모두 실제 결과와 동떨어진 것들이었다. 제1당이 차지한 180석이라는 의석수는 선거 전에 나온 조사들은 물론이고 선거 당일에 나온 출구조사의 예측 범위까지 훌쩍 벗어나는 것이었다. 많은 조사원의 노력과 수십억 원에 이르는 조사비용에도 불구하고 최근 두 차례 총선 결과를 예측한 출구조사는 모두 실패였다.

혹시 우리나라 조사기관들의 실력이 부족해서 자꾸 이런 결과가 나오는 것일까? 세계 최고 수준의 실력과 오랜 경험을 자랑하는 미국의 조사기관들이라면 그런 실패를 피할 수 있었을까? 물론 그렇지 않다. 가장 최근에 있었던 2016년 가을 미국 대통령 선거를 돌아보자.

공화당 후보는 트럼프, 민주당 후보는 클린턴이었고 당선자는 트럼프였다. 그런데 거의 백년에 가까운 역사를 자랑하는 갤럽은 물론 직전 선거까지 놀라운 예측 성공률을 자랑하던 네이트 실버까지도 당선자를 잘못 짚었다. 선거구가 수백 개에 이르는 의회 의원 선거에 비해 전국이 단일 선거구인 대통령 선거는 훨씬 정확하게 예측할 수 있어야 마땅한데도 다들 실패한 것이다.

한국, 미국의 여론조사만 그랬던 것은 아니다. 브렉시트, 즉 영국의 유럽연합 탈퇴 여부를 묻는 국민투표를 앞두고 실시된 여론조사들 역시 거의 모두 투표결과와 반대되는 예측을 내놓은 바 있다.

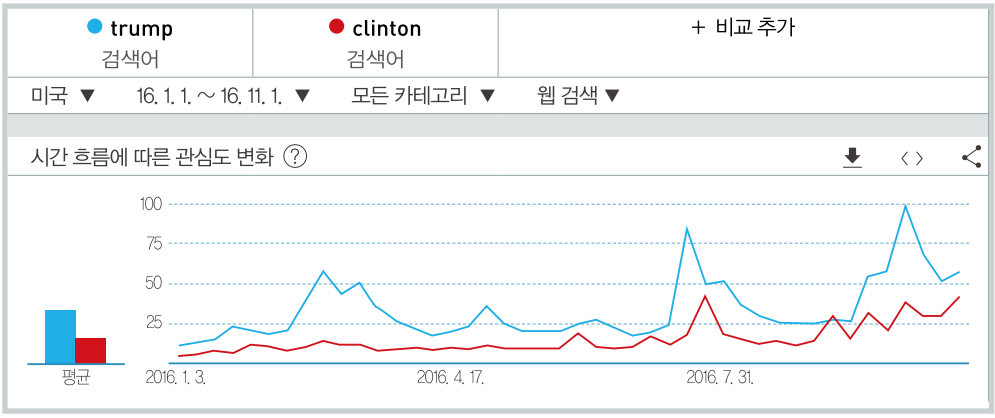
빅데이터로 선거결과를 예측하면?

미국 대선이 끝난 뒤 여론조사기관의 예측이 거의 다 틀린 것으로 드러나자 통계학에 바탕을 둔 전통적인 여론조사를 비판하는 목소리도 높아졌다. 사람들은 특히 표본의 크기를 문제 삼았다. 겨우 1,000명 정도의 표본만으로 수천만 명에서 수억 명에 이르는 사람들의 생각을 파악하는 것이 가능할 리 없다는 것이다. 사람들은 그렇게 적은 표본을 근거로 예측을 한다면 점쟁이의 예언과 다를 바가 없으므로 여론조사가 자꾸 틀리는 게 당연하다고들 했다.

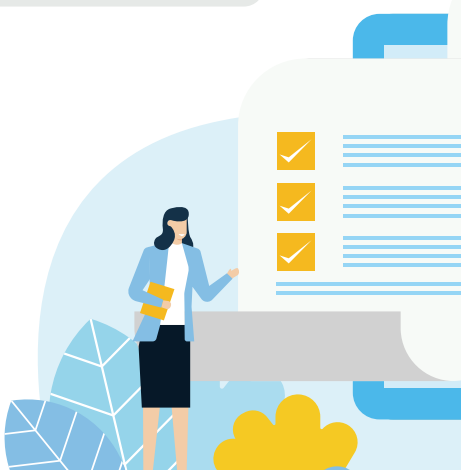
국내 대학의 어느 교수 역시 미국 대선 직후에 그런 주장을 담은 책을 냈다. 그는 미국 유권자수가 2억 2천만 명이 넘는데 여론조사에서 이용하는 표본은 1,000명 남짓으로 전체 유권자의 0.00001%도 안 된다고 지적했다. 그런 표본만 가지고는 평균적으로 미국 각 주에서 겨우 20명 정도만 조사하는 셈이니 제대로 된 예측이 나올 수가 없다고 했다. 그렇게 여론조사의 한계를 지적한 다음 그 교수는 자신이 여론조사가 아닌 새로운 방법으로 '국내에서 유일하게' 트럼프의 당선을 예측했다고 밝혔다.

책에서 공개한 그만의 비법이란 바로 구글의 검색어 빅데이터를 분석하는 방법이다. 선거를 앞두고 사람들이 어느 후보의 이름을 많이 검색하는지 비교해보면 당선자를 예측할 수 있다는 것이다. 그가 제시한 근거는 다음 그래프와 같다.

2016년 미국 대선후보에 대한 구글의 검색어 빅데이터



그래프는 선거가 있던 해 1월부터 선거 직전까지 미국인들이 트럼프와 클린턴을 검색한 빈도를 비교해서 보여주고 있다. 트럼프를 검색한 빈도를 나타내는 푸른색 선이 시종일관 클린턴을 검색한 빈도를 나타내는 붉은색 선보다 위에 있다. 그 교수의 분석에 따르면 높은 검색 빈도는 트럼프에 대한 높은 관심 때문이고 그런 관심이 표로 연결되었다고 한다. 아니, 이렇게 단순 명쾌할 수가! 빅데이터 만세!



겨우 1,000명 남짓 조사하는 여론조사에 비하면 구글 검색어 데이터는 그 규모부터가 어마어마하다. 게다가 구글 검색어 통계는 '구글트렌드(https://trends.google.co.kr/trends/?geo=KR)'를 통해 공짜로 제공되기 때문에 조사비용이 한 푼도 안 든다.

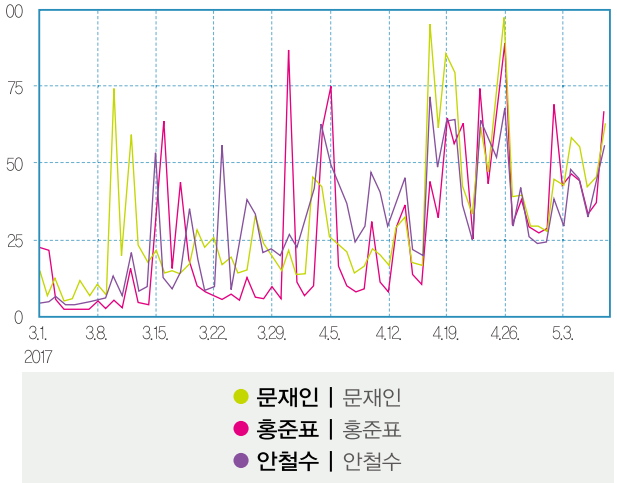
이런 주장을 담은 그의 책은 당연히 베스트셀러가 되었고 미국 대선보다 겨우 몇 달 뒤에 치러진 한국의 대선 때 각 당 후보 진영에서 구글트렌드 순위에 각별히 신경을 쓰도록 만들기까지 했다. 그해 5월 초 선거 직전에 우리나라 언론의 기사 제목을 몇 개 보자.

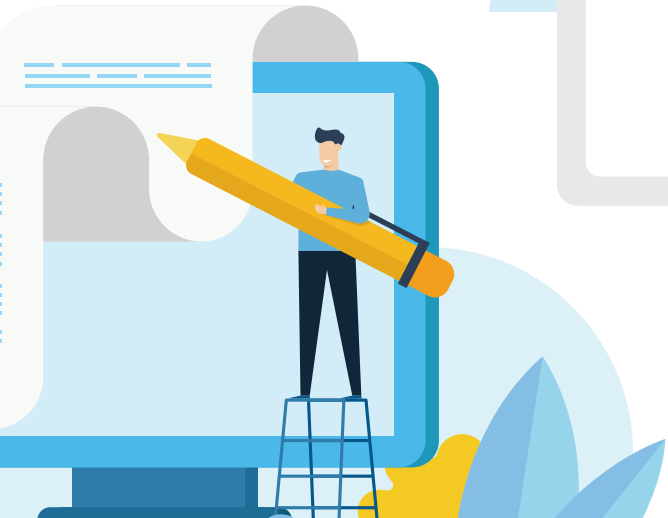
'3일 남긴 대선, 구글트렌드에선 문재인·홍준표 혈투'(서울경제), '서로 1위 주장 구글트렌드, 진짜 1위는 누구'(TV조선), '홍준표 구글트렌드 1위지만 검색 단어 1위가 '홍준표 돼지'(미디어오늘).

그런데 구글트렌드에서 볼 수 있었다는 문-홍 두 후보 간의 '혈투'가 선거 결과에서도 그대로 드러났을까? 우리가 기억하다시피 전혀 그렇지 않았다. 문재인, 홍준표, 안철수 세 후보의 실제 득표율은 41: 24: 21로서 혈투와는 거리가 먼 일방적인 결과였다. 혹시 구글 트렌드와 실제 결과 사이의 엄청난 불일치는 한국인들이 검색할 때 구글을 많이 쓰지 않는다는 이유 때문인지도 모른다.

그렇다면 네이버가 제공하는 검색어 서비스를 살펴보면 어떨까? 네이버 역시 구글트렌드와 닮은 네이버트렌드 검색어 통계 서비스를 제공하고 있다. 다음 그래프는 선거 전날까지 2개월 남짓 동안의 결과이다.

네이버트렌드의 검색어 통계





그래프에서 녹색은 문재인 후보, 붉은색은 홍준표 후보, 보라색은 안철수 후보의 검색어 빈도다. 네이버 트렌드에 따르면 선거 전날인 5월 8일 검색 순위 1위는 홍 후보였으므로 검색어 빈도를 이용한 예측 방식이 맞았다면 지금 청와대의 주인은 다른 사람이 되었을 테다!

트럼프의 당선을 유일하게 성공적으로 알아맞혔다는 교수의 예측을 과학적인 것이라고 보기 어려운 이유 중 하나로 책에 실린 예상득표율을 들 수 있다. 저자는 트럼프가 선거인단 수에서 클린턴을 앞서고 득표율에서도 역시 52~54%로 앞설 것으로 예측했다. 그런데 우리가 기억하다시피 실제 선거 결과 선거인단 수에서는 트럼프가 앞섰지만 득표수에서는 반대였다. 힐러리 클린턴은 상대 후보보다 수백만 표를 더 얻고도 선거에서는 패배한 것이다. 이렇게 볼 때 검색어 트렌드를 이용하여 당선자를 예측해보는 것은 무척 흥미롭고 누구나 시도해볼 만하다. 단, 예측에 성공할 것이라는 기대를 하지 않을 때 그렇다는 말이다.

제한된 데이터만 얻을 수 있었던 과거에 비해 오늘날 우리는 거의 무제한적인 정보를 활용할 수 있게 되었고, 덕분에 과거에는 알기 어려웠던 많은 것들을 신속하게 파악할 수 있게 되었다. 그렇다면 빅데이터의 시대에 접어들면서 과거의 통계학은 수명을 다한 것일까?

알다시피 통계학의 핵심적인 과제 중 하나는 불확실성을 측정하고 적절히 통제하는 것이다. 하지만 불확실성을 완전히 없애지는 못하기 때문에 통계학에서는 언제나 분석이나 예측이 어긋날 가능성을 함께 알려준다. 그런데 맞을 수도 있고 틀릴 수도 있다는 결론은 늘 찜찜하고 거추장스럽다. 여론조사 대신 검색어를 이용해서 선거 결과를 예측할 수 있다는 주장 뒤에는 빅데이터 덕분에 우중충한 불확실성으로부터 해방될 수 있으리라는 희망이 숨어 있다. 그런 희망에서부터 빅데이터에 대한 맹목적인 신앙까지의 거리는 그리 멀지 않아 보인다.

빅데이터가 놓친 코로나

2019년 말부터 시작된 코로나 사태는 전 세계 사람들의 삶에 엄청난 영향을 미쳤다. 심지어 인류의 역사를 코로나 전과 코로나 이후를 뜻하는 BC, AC로 나누어야 할 것이라는 주장까지 나왔다고 한다. 근본적인 질문을 해야 할 시기라는 것이다.

한편 어떤 사람들은 이런 상황 속에서 빅데이터의 역할이 무엇인가라는 질문을 던지기도 한다. 새로운 세계를 열어줄 것이라던 빅데이터가 코로나 사태를 해결하는 데 어떤 도움이 되고 있는지 궁금하다는 것이다. 첨단과학과 기술을 이용해서 코로나19와 같은 감염병이 언제 어디서 어떤 규모로 유행할지 미리 예측할 수 있다면 온 세계가 겪는 이런 재난도 피할 수 있을 것이기 때문이다.

더군다나 빅데이터를 설명하는 책에서 구글이 검색어 빅데이터를 가지고 미국의 독감 유행을 정확하게 예측했다는 사실을 아는 사람이라면 더욱 그럴 것이다. 이미 10여 년 전부터 구글은 검색어를 이용한 구글 플루 트렌드(Google Flu Trends)를 통해 독감

발생을 예측하기 시작했다. 놀랍게도 구글의 예측은 2009년과 2010년의 실제 독감 발생 데이터와 아주 잘 들어맞았다. 우리나라의 질병관리본부에 해당하는 미국 CDC에 있는 많은 전문가들도 하지 못한 예측을 공중보건과 아무 상관없는 IT 기업이 검색어를 이용해서 해냈던 것이다. 이러한 성과는 당연히 빅데이터의 위력을 유감없이 보여주는 사례로 널리 알려졌다.

그런 눈부신 성공에 힘입어 구글은 그 이후 알고리즘을 더욱 개량하여 점점 완벽에 가깝게 독감 유행을 예측했을까? 구글의 성공담에 비해 그 뒤의 이야기는 조금 덜 알려진 것 같다. 뜻밖에도 구글은 얼마 뒤에 구글 플루 트렌드 사업을 접고 만다. 짐작컨대 그 이유 중 하나는 2012년, 특히 2013년 구글의 예측이 실제 결과와 아주 많이 어긋나 버렸기 때문일 것이다. 당시 구글은 독감이 유행할 규모를 실제보다 거의 두 배나 될 정도로 과장해서 예측했다. 아마 다양한 원인이 작용해서 그런 잘못된 예측이 나왔겠지만 구글의 독감 예측 사례는 빅데이터의 눈부신 성공담인 한편, 빅데이터에 대한 지나친 믿음을 경고하는 대표적인 사례가 될 만하다.

사실 독감은 거의 일정한 시기에 주기적으로 찾아오기 때문에 이미 상당한 데이터가 축적되어 있다. 그런데도 독감 발생을 제대로 예측하는 것이 그처럼 어렵다면 선례가 전혀 없는 코로나와 같은 사태를 데이터를 통해 파악하기란 더욱 어려울 것이다. 그래도 어쩌겠는가, 우리는 코로나와 같은 대규모 재난을 만나고 나서야 비로소 이전까지의 사고방식과 삶의 방식을 되돌아보기 시작한다. 데이터 역시 마찬가지일 것이다.

빅데이터의 놀라운 성과에 경탄하면서 모든 것이 데이터가 되는 빅데이터의 시대를 살고 있다고 말했지만 정작 중요한 것들은 데이터로 측정조차 하지 않고 있었던 것은 아닐까? 지금은 감염병뿐 아니라 인간과 생태계를 그동안 어떤 방향에서 보고 어떤 데이터들을 모아왔는지 돌아보고 평가해볼 시점이 아닐까? 그렇다면 지금 우리가 겪는 어려움은 당장의 감염병 사태뿐 아니라 우리가 살아온 방식, 사회와 국가의 역할, 그리고 세계의 미래까지 두루 되돌아보고 새로운 방향을 모색할 기회일지도 모른다. 또한 장밋빛 환상의 대상이었던 빅데이터에 대해서도 새로운 사고방식으로 성찰해볼 기회일지 모른다.

김준래 ◆ 통계의 창 객원기자

통계로 들여다보는 바이러스와 인간의 전쟁

전염병 공포가 다시 세계를 강타하고 있다. 중증급성호흡기증후군(SARS)와 중동호흡기증후군(MERS)에 이어, 이번에는 코로나바이러스감염증(COVID-19)까지 등장하여 전 세계를 공포의 도가니로 몰아넣고 있다.

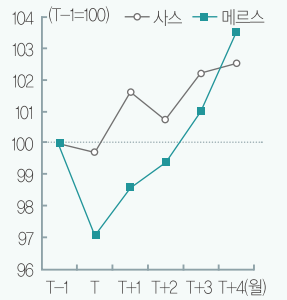
물론 희망도 보인다. 이글을 쓰는 시점에서 우리나라를 비롯한 전 세계 의료진들의 헌신적인 노력으로 급증하던 환자 수는 조금씩 줄어들고 있고, 사망자도 감소 추세를 보이고 있다. 문제는 이 상황이 끝난다 하더라도 전염병에 대한 사람들의 생각이 이전과 같지 않으리라는 점이다. 전염병은 단순히 건강에 위협을 가하는 병원체가 아니라, 언제든지 우리의 생활을 근본적으로 뒤흔들 수 있는 공포의 대상으로 여겨질 것이기 때문이다.

이런 전염병이 주는 공포를 이겨내려면 치료에 집중하는 것도 중요하지만, 앞으로 이런 전염병이 다시 창궐하지 않도록 다양한 대처 방안을 마련하는 것이 무엇보다 필요하다고 전문가들은 입을 모은다.

그렇다면 전염병이 유행하지 않도록 하려면 인류는 어디에서 답을 찾아야 할까? 이에 대해 전문가들은 역사와 통계가 답을 제시해줄 것이라고 조언한다. 역사적으로 유행했던 전염병들의 정체를 되짚어보고, 그 과정에서 파악한 통계로 전염병에 대처해 나가야 한다는 것이다.

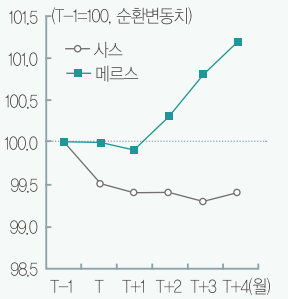
전염병 유행 이후 경기 및 체감지수

전염병 전후 소매판매



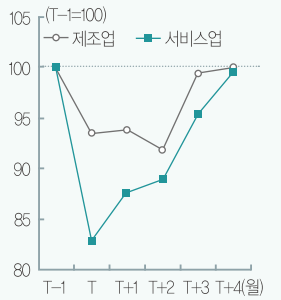
주 : 사스 T=03.4월, 메르스 T=15.6월
자료 : 통계청

전염병 전후 경기동행지수



주 : 사스 T=03.4월, 메르스 T=15.6월
자료 : 통계청

메르스 전후 중소기업 체감경기



주 : T=15.5월
자료 : 중소기업중앙회(SBHI 기준)

(출처 : 통계청, 중소기업중앙회)

전염병 역사는 인류의 역사

전염병의 역사는 가히 인류의 역사라 할 수 있을 만큼 오래되었다. 사료를 살펴보면 기원전 430년경에 그리스의 도시 국가인 아테네에서 전염병이 발생했다는 기록이 남아 있다. 그로부터 20년 후인 410년경에는 '히포크라테스 선서'로 유명한 그리스의 의사 히포크라테스가 독감의 증상을 처음 기록에 남겼다.

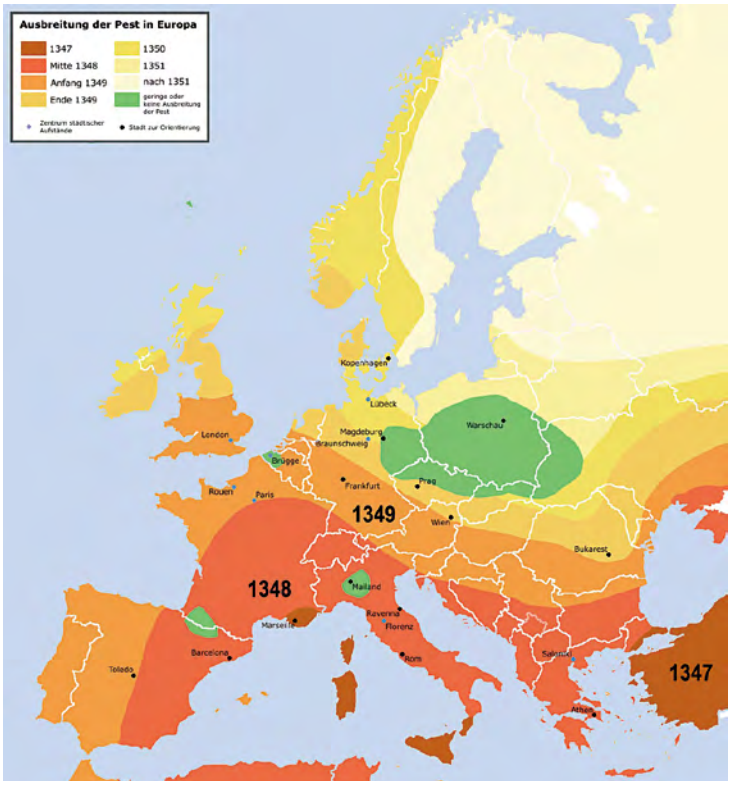
이처럼 오랜 역사를 가진 전염병 중에서도 인류의 생존에 치명상을 가한 전염병으로는 흑사병과 스페인독감 그리고 콜레라 및 결핵 등이 대표적 사례로 꼽힌다. 이들 전염병은 적게는 수천만 명에서 많게는 수억 명까지 사람의 생명을 빼앗아 지금까지도 최악의 전염병으로 기록되어 있다. 흑사병은 몸이 새까맣게 변하면서 서서히 죽어간다고 해서 붙여진 이름이다. 들쥐가 갖고 있는 페스트균에 의해 발생하기 때문에 '페스트(Pest)'라고도 불리는 이 전염병은 14세기 중반에 등장하여 순식간에 유럽 전역으로

퍼져나가면서 당시 유럽 인구의 3분의 1인 1억 명 정도를 죽음으로 몰아넣었다.

흑사병은 체액이나 분뇨에 의해 전염되는데, 흑사병에 걸리면 초기에는 두통, 발열, 오한 등의 증상을 보인다. 이후 피부가 검은색으로 변하며 몸 전체에서 괴사가 일어나게 되는데, 전염성이 매우 강해 감염되면 대부분의 사람들이 끔찍한 모습으로 죽게 된다.

흑사병이 중세 유럽에서 퍼진 당시에는 별다른 치료방법이 없었으나, 항생제가 탄생하면서 본격적인 치료가 가능하게 되었다. 다만 흑사병은 발병 이후 증상이 빠르게 진행되기 때문에 조기에 치료할수록 완치 효과가 높다.

흑사병이 창궐했던 14세기의 유럽 지형



(출처 : wikipedia)

흑사병의 대유행 이후 한동안 잠잠하던 전염병은 19세기 들어 콜레라와 결핵이 등장하면서 연이어 전 세계를 공포의 도가니로 몰아넣었다.

콜레라는 콜레라균(Vibrio Cholerae)의 감염에 의해 나타나는 급성 전염병으로, 대

표적 증상은 설사다. 급성 설사가 자주 유발되어 탈수가 매우 빠른 속도로 진행되다가 결국에는 사망에 이르게 된다.

콜레라는 원래 인도 갠지스 강 유역에서 나타난 풍토병이었다. 하지만 19세기 들어 이 지역을 식민지화했던 영국으로 전염되기 시작했고, 결국에는 러시아와 아프리카 등지로 퍼져나가면서 약 1,500만 명의 사망자가 발생했다.

콜레라가 우리에게 낯설지 않은 이유는 조선 순조 시절에 역시 한반도 전역을 휩쓸었기 때문이다. 당시 괴질(怪疾)이라 불렸던 이 전염병은 순조 21년인 1821년부터 유행하기 시작하여 수백 년간 백성들을 괴롭혔다.

당시 평안도 감사였던 김이교가 작성했던 기록을 살펴보면 "최근 들어 갑자기 발생한 괴질로 인해 백성들이 설사와 구토를 하면서 순식간에 죽어버렸다"라고 언급하고 있다. 그러면서 "불과 10일 만에 일천여 명이 죽었으나, 치료할 약과 방법이 없어 절망적이다"라고 적은 것을 보면 당시 상황이 얼마나 어려웠는지를 조금이나마 느낄 수 있다.

이처럼 전 세계를 초토화시켰던 콜레라는 당시의 과학 수준에서 볼 때 병원균이 공기를 떠돌아다니면서 사람들을 전염시키는 것이라고 생각했다. 하지만 영국의 한 의사는 공기가 아닌 오염된 물에 의해 콜레라가 전염된다고 주장했다.

그의 주장은 당시 학계에서 많은 논란을 불러일으켰지만, 수많은 실험 끝에 결국 이 의사의 주장이 사실이라는 점이 밝혀졌다. 많은 희생을 치렀지만 콜레라의 유행으로 인해 상하수도 시설 보급 및 공중위생 확립을 제공하는 계기가 되었다.

19세기의 최대 재앙이 콜레라 감염이라면 20세기의 재앙은 스페인독감(Spanish flu)의 창궐이다. 이 전염병은 1918년부터 1920년까지 2년 동안 전 세계에서 5천만 명 이상의 사람들을 죽음으로 몰아넣었다.

의료 기술의 수준이 과거와는 비교가 되지 않을 정도로 발전한 시기였지만, 이토록 많은 사람들이 목숨을 잃은 이유는 스페인독감만이 가진 무서운 특징 때문이었다. 딱

히 이렇다 할 증상이 나타나지 않는 무증상 감염이 대부분이라 치료 시기를 놓치는 경우가 대부분이었던 것이다.

스페인 독감에 걸리면 처음에는 감기에 걸린 듯한 증상을 보이다가 시간이 흐를수록 몸에서 급격하게 열이 나면서 보랏빛으로 변해 죽어갔다. 이런 증상 때문에 스페인 독감은 '3일 열병'이라고도 불렸는데, 그 이유는 3일 내외로 짧은 열병 증상을 보이다가 급격히 상태가 악화되는 경우가 많았기 때문이다.

스페인 독감도 콜레라처럼 조선시대에 한반도까지 전염된 기록이 남아 있다. 일제강점기 시절인 1918년에 발간된 조선총독부 통계연감에 따르면, 조선의 총인구 1,670만 명 중 44%인 742만 명이 독감에 걸렸고 그중 14만 명이 사망한 것으로 기록되어 있다.

무섭게 전 세계로 번지던 스페인 독감도 1920년에 접어들면서 자연스럽게 잦아들기 시작했고, 사람들은 전염병의 공포에서 벗어날 수 있게 되었다. 과거의 전염병들과 다른 점이라면 치료제보다 예방접종을 강화하여 처음부터 질병에 걸리지 않도록 하는 예방 문화가 생겨났다는 점이다.

마지막으로 결핵(Tuberculosis)은 역대 가장 많은 사망자를 낸 전염병으로 유명하다. 처음 발병이 시작된 19세기 초반부터 지금까지 200여 년 동안 약 10억 명의 사망자가 발생한 것으로 알려져 있다. 결핵은 사람의 입에서 미세한 침방울에 들어 있던 결핵균이 상대방의 코와 입을 통해서 전염되는 감염병이다. 지금의 신종 코로나 바이러스와 유사한 감염 경로를 갖고 있다고 볼 수 있다.

결핵균은 지난 1882년 세균학자였던 로베르트 코흐(Robert Koch)에 의해 발견되었고, 그해 3월 24일 베를린에서 개최된 학회에서 발표됐다. 그 이후 3월 24일은 '세계 결핵의 날'로 지정되어 지금까지 매년 관련 행사가 열리면서 결핵에 대한 경각심을 일깨워주고 있다.

결핵균에 감염되면 한 달 이상 기침이 지속되고, 가슴에 통증이 나타나며, 기침 후 피가 섞여 나오는 증상을 보인다. 또한 추위를 타거나, 식욕감퇴 및 체중감소 등의 증상이 동반되기도 한다.

결핵이 본격적으로 유행하기 시작한 19세기 후반만 하더라도 환자들은 맑은 공기를 마시면 병이 치유된다고 믿었다. 이 때문에 공기가 좋기로 유명한 스위스의 다보스 지역은 결핵 환자를 위한 요양지로 유명해졌고, 그 결과 경제적 풍요로움을 누리는 혜택을 보면서 다보스 포럼 같은 국제적 행사를 많이 유치하고 있다.

전염병 퇴치에 활용된 통계

전염병 퇴치에 통계가 활용되기 시작한 것은 콜레라가 유럽에서 창궐하던 19세기부터다. 특히 피해가 컸던 영국에서는 콜레라가 주기적으로 몇 번씩 발생하면서 수많은 사람이 목숨을 잃었다.

영국의 의사들과 과학자들은 하루에도 수천 명씩 사망하는 피해를 방지하기 위해 콜레라의 발병 원인을 찾기 위해 애를 썼다. 하지만 당시의 과학기술 수준으로는 세균의 존재를 정확히 알 수 없었기 때문에 콜레라가 정확히 어떻게 전염되는지를 알 길이 없었다.

콜레라 발병의 원인과 관련하여 당시 유럽에서 가장 인정받았던 가설은 콜레라가 길거리에 버려진 쓰레기에서 발생하는 악취에 의해 전염되므로 악취를 제거해야 한다는 것이었다. 그래서 사람들은 거리에 쌓인 쓰레기들을 모두 강물에 흘려보내는 캠페인을 벌였는데, 그런 노력에도 불구하고 콜레라 환자는 더욱 늘기만 했다.

이때 존 스노(John Snow)라는 젊은 의사는 콜레라의 전염 원인이 악취로 가득한 공기라는 주장에 대해 의문을 품었다. 만일 콜레라가 공기에 의해 감염되는 것이라면 같은 공기를 마시는 사람들은 모두 병에 걸려야 하지만, 그렇지 않다는 점이 스노를 의심하게 만들었다.

그때부터 스노는 콜레라의 발생 원인에 대해 조사하기 시작했는데, 이를 위해 도입한 방법이 바로 통계였다. 그는 우선 콜레라로 생명을 잃은 사람들의 주변을 철저히 조사했다. 그리고 동일한 환경에서도 콜레라에 걸리는 사람과 걸리지 않는 사람들 사이에 어떤 차이가 있는지를 수치로 비교했다. 이렇게 분석한 결과를 그는 지도 위에 일일이 기록했다. 콜레라를 앓고 있는 환자와 앓다가 죽은 사망자의 숫자를 집계하여 기록하자, 지역 별로 일목요연한 통계치가 생성되었다.

스노는 지역과 질병 관련 통계치의 상관관계를 분석하다가 놀라운 사실을 발견했다. 바로 런던에 위치한 여러 개의 급수 시설 중 특정한 급수 시설을 이용하고 있던 사

람들의 사망률이 다른 급수 시설을 사용하는 사람들의 사망률보다 무려 여덟 배나 많다는 점을 발견한 것이다.

그는 수차례 지도와 통계치를 비교해 보면서 콜레라 전염의 원인이 공기가 아닌 급수 시설이라는 점을 확신했다. 이후 스노는 자신이 정리한 통계 자료를 근거로 런던시청에 특정한 급수 시설의 폐쇄를 요청했다.

처음에는 반신반의하던 시청 공무원들이었지만 눈앞에 놓인 통계 자료를 보자 믿지 않을 수 없었다. 이후 특정 급수시설이 폐쇄되었고 며칠이 지나자 거짓말처럼 환자수가 줄기 시작했다. 살기등등하던 콜레라의 기세를 꺾고 이를 퇴치할 수 있는 반전의 기회를 한 젊은 의사의 집념이 만들어낸 순간이었다.

그로부터 30년 후 독일의 세균학자인 코흐가 콜레라 세균을 발견하고, 그 세균이 물과 밀접한 관계를 맺고 있다는 것을 발견하여 학계에 보고했다. 이로써 그때까지만 해도 가설 수준에 머물렀던 스노의 통계를 기반으로 한 주장은 과학적 증명을 통해 옳다는 점을 인정받았다.

놀라운 점은 스노가 의사였지만 세균이나 병리학에 대한 지식이 거의 없었다는 점이다. 그럼에도 불구하고 스노가 자신의 주장이 옳다고 믿었던 이유는, 스스로 산출한 통계 결과가 콜레라의 발병 원인이 급수 시설에 있다는 점을 명확히 보여주었기 때문이다.

당시만 해도 전염병의 치료는 소위 전문가라고 하는 사람들의 경험과 생각에 의해 좌지우지 되던 시대였다. 하지만 콜레라를 퇴치하는데 있어 통계 시스템이 결정적 영향을 끼침에 따라, 이후부터는 전염병이 유행하게 되면 스노의 통계 방법을 따라 하는 사례가 늘어났다.

콜레라의 감염 경로를 통계적으로 규명한 존 스노와 당시 지도

(출처 : geoind.wordpress)

전염병 예방 및 치료는 오픈 데이터 시스템에 달려 있다

콜레라의 사례에서 보듯이 19세기부터 전염병 퇴치에 활용되기 시작한 통계 시스템은 오늘날에 와서는 '오픈데이터 시스템(open data system)'으로 확대되어 의료 현장에서 활용되고 있다.

대표적으로는 코로나19 사태로 인해 국내에서 탄생하게 된 '전염병 관련 공공데이터 시스템'을 꼽을 수 있다. 여러 사회적 협동조합이 힘을 합쳐 만든 공동대응팀이 제안한 이 시스템은 감염병 관련 종합통계 정보를 다양한 시각데이터를 통해 제공하도록 설계되었다.

공동대응팀의 관계자는 이 같은 시스템을 제안한 이유에 대해 "정부가 그동안 투명하게 전염병 관련 데이터를 공개하고 있었으나, 코로나19가 확산되면서 산발적으로 생산되는 데이터를 일반 시민이 일일이 확인하기 어렵다는 문제점이 발생했기 때문"이라고 지적했다.

그러면서 "정부와 공공기관은 보유하고 있는 데이터와 통계치를 제공하고, 민간에서는 다양한 아이디어를 통해 이를 시각화하여 재난이나 전염병 같은 긴급 상황에 대처하고자 한다"라고 밝혔다.

시스템 제안서를 살펴보면 감염병 대응 상황을 종합적으로 파악할 수 있는 '종합 통계'를 비롯하여 '확진자 및 의심 환자 목록'과 '방역대상 장소', '선별진료소 위치' 같은 데이터가 포함되어 있다. 이들 데이터는 시각화된 정보를 제공할 뿐만 아니라 공유 프로그램을 통해 지도와 결합한 새로운 서비스가 가능하다는 것이 공동대응팀의 설명이다.

이 같은 제안에 대해 건강보험심사평가원 관계자는 "기획재정부를 중심으로 질병관리본부와 건강보험심사평가원, 그리고 한국보건산업진흥원 등 다양한 기관이 모여 데이터 개방을 논의하고 있다"라고 전하면서 "추후에도 지속적으로 민간이 활용할 수 있는 데이터와 통계치를 공개할 것"이라고 말했다.

실제로 공공데이터를 모아 전염병 퇴치에 활용하고 있는 사례로는 최근 들어 국내에서도 다양하게 등장하고 있다. 데이터를 활용하여 마스크 판매점의 위치와 재고 정보를 알려주는 어플리케이션을 비롯하여 지리정보시스템(GIS)을 기반으로 환자의 동선이나 선별진료소 정보 등을 제공하고 있다.

이 중에서도 코로나19 환자 경로 시각화와 관련된 데이터셋(data set)을 오픈소스로 공개한 마인즈랩은 가장 돋보이는 성과를 거두고 있다. 이 회사가 구축한 코로나19 동선 추적 데이터셋은 전염병 퇴치를 위한 거의 모든 데이터와 통계치가 저장된 시스템이라 할 수 있다.

데이터셋에는 확진자의 경로와 연령, 성별 및 진단 날짜 등 기초적인 환자 경로 데이터는 물론, 22가지의 주요 전염병과 16개의 백신에 대한 정보가 들어 있다. 또한 의료시설 등을 포함한 의료 통계 데이터와 다양한 변수에 따른 시각화된 데이터도 포함되어 있다.

이에 대해 마인즈랩의 관계자는 "해당 데이터셋은 기존 코로나19 관련 오픈 데이터셋들에 비해 데이터의 양과 품질, 그리고 데이터 시각화 부분에서 뛰어난 성능을 자랑한다"라고 소개하며 "이런 차별성으로 인해 미국의 커뮤니티 사이트 중 딥러닝 부문 1위에 오르는 등, 해외 개발자 사이에서도 주목을 끌고 있다"라고 말했다.

이뿐만이 아니다. 과학기술정보통신부와 국토교통부, 질병관리본부는 코로나19 확산 방지를 위해 역학 조사에 필요한 통신 및 카드 사용 정보 등을 수집하는 '스마트시티 시스템'을 활용하는 방안을 추진하고 있어 관심이 모아지고 있다.

코로나가 숙제로 던져준 데이터를 통한 감염병 예측

국내에서 오픈 데이터를 이용한 통계 시스템을 통해 전염병 확산 방지에 주력하고 있다면 캐나다의 스타트업인 블루닷(BlueDot)은 확보한 데이터를 이용한 감염병 예측 시스템을 운영하고 있어 눈길을 끌고 있다. 특히 이 회사는 자사 연구진이 개발한 인공지능 시스템을 활용하여 세계보건기구(WHO)보다 열흘 먼저 코로나19 확산 위험성을 미리 경고하여 유명세를 탄 바 있다.

조그만 스타트업에서 만든 인공지능 시스템이 어떻게 신종 전염병의 위험을 더 일찍 발견할 수 있었을까? 그 비결은 바로 이 회사가 개발한 데이터 분석 기술에 숨어 있다. 블루닷은 질병을 추적하는 데 필요한 정보를 SNS가 아닌 항공 탑승권 결제 데이터에서 수집했다. 우한에 체류했던 사람들의 일부가 아시아 각국별로 출국한 것을 확인하고 코로나19가 우한에서 아시아 전역으로 확산될 것을 예측한 것이다.

세계보건기구보다 빨리 코로나19 전염을 경고했던 블루닷

(출처 : CBC.CA)

미국의 하버드대와 MIT 공대 연구진이 빅데이터 기반 전염병 분석 웹사이트인 헬스맵(Health Map)도 있다. 헬스맵 역시 블루닷처럼 에볼라 사태나 메르스 사태 때 WHO보다도 먼저 위험성을 경고해 화제가 되었다. 헬스맵은 지난 2006년 수만 개의 소셜미디어 사이트와 지역뉴스, 그리고 의료진 네트워크에서 질병 발생 정보를 모은 뒤 이 중에서 믿을 만한 내용을 선별하여 웹상에서 지도 형태로 보여주는 시스템이다.

헬스맵의 가장 큰 장점은 역시 신속성이다. 소셜미디어 등을 활용하기 때문에 신뢰도는 떨어질 수 있지만 질병 발생정보가 올라오는 속도는 WHO보다도 훨씬 빠르다. 이런 활약상으로 헬스맵은 현재 코로나19 감염증 발생 지역 등에 대한 정보를 제공하고 있다.

이 밖에도 캐나다 온타리오공과대 연구진이 개발한 병원 감염 예측 프로그램은 산·학·연 협력의 좋은 사례로 꼽힌다. 병원이 환자 데이터를 제공하면 대학이 데이터를 분석하고, IBM 같은 기업이 통계분석 시스템을 제공하는 형태다. 이렇게 분석된 감염병 예측 데이터는 질병에 상대적으로 더 취약한 신생아와 미숙아의 질병 발병 예측 등에 활용되고 있다.

정남수 ◆ 통계청 인구총조사과장

2020년은 인구주택총조사의 해

올해 인구주택총조사는 무엇이 바뀌고 어떻게 진행될까

우리나라에서 통계를 하는 사람들에게 0과 5로 끝나는 해는 매우 의미가 있는 해인데 바로 인구주택총조사가 실시되기 때문이다. 2015년부터 2024년 중에 실시되는 2020 라운드 센서스에서는 238개 UN회원국 중 95%에 이르는 227개국에서 인구총조사를 실시할 예정으로, 2020년은 미국, 일본, 중국을 비롯한 세계 57개국에서 센서스가 계획되어 있다. 우리나라 역시 11월 1일에 인구주택총조사를 실시한다.

그런데 지난 1월 중국을 시작으로 아시아, 유럽, 아메리카 등 전 대륙으로 코로나19가 확산되면서 세계 여러나라가 그 충격을 극복하고 안정을 찾기 위하여 고군분투하고 있다.

통계 분야도 예외는 아니어서 미국, 필리핀을 비롯해 상반기 중에 센서스를 실시할 예정이었던 국가들은 코로나 19의 여파로 총조사 관련 주요 일정들을 연기하고 있다.

미국은 당초 4월 1일기준으로 센서스를 시행하여 지난 3월 12일부터 온라인, 전화, 우편으로 자기응답(self-responses)을 시작하여 6월 말까지 완료할 예정이었으나 그 기간을 10월 말까지로 연기하였다.

우리나라는 외국으로 통하는 관문을 폐쇄하지 않고 집단감염을 잘 관리하면서 슬기롭게 코로나를 이겨낸 경험으로 전 세계의 감탄과 부러움을 받고 있다. 통계청에서도 '사회적 거리두기' 기간에 전자조사 등 비대면조사를 활용하고 면접조사 규칙을 엄격히 실행하여 조사자와 응답자의 안전을 확보하고 비상상황의 정책 대응에 필요한 산업활동동향, 고용동향 등의 통계자료들을 정상적으로 제공함으로써 코로나의 충격을 극복하고자 하는 세계 여러 나라의 귀감이 되고 있다.

통계청은 11월에 실시되는 2020 인구주택총조사에서도 이러한 자산과 경험을 바탕으로 코로나19 이후 비대면을 더욱 선호하는 사회변화를 반영하여 인터넷조사를 모바일까지 확대하고, 인터넷 취약계층에 대해서는 전화조사를 도입하여 대응할 방침이다. 한편 과거 조사항목 중 10개 항목을 행정자료로 대체하여 국민 응답부담을 경감하면서도 총 조사 공표항목을 늘릴 예정이다.

그럼 지금부터 국민들의 적극적인 관심과 참여가 필요한 2020 인구주택총조사의 특징과 진행방법에 대하여 자세히 살펴보자.

인구주택총조사의 역사를 살펴보면

인구주택총조사의 기원은 고대 바빌로니아(B.C. 3,600년경)까지 거슬러 올라간다. 로마시대 인구조사는 재정과 징병을 목적으로 시민의 수와 재산을 조사하였는데, BC 435년부터는 로마제국의 시민등록과 시세조사를 센소(Censor)라는 관리가 담당하였다. '총조사'의 영어명인 센서스(Census)는 여기에서 유래하고, 미국의 통계청은 센서스 뷰로(Census Bureau)라고 한다.

우리나라 인구총조사에 대한 기록은 삼한시대까지 거슬러 올라가지만 최초의 근대적 의미의 인구총조사는 1925년에 시작되었고 1960년에 UN의 권고에 의하여 세계적으로 실시되는 인구, 주택 및 농업총조사 프로그램(World Census Program)을 계기로 조사기획 단계부터 자료처리 및 평가에 이르기까지 현대적인 총조사의 면모를 갖춰 오늘날 센서스의 발전을 가져오는 밑바탕을 마련했다. 이때 처음으로 주택부문이 병행조사되었고 이후 매 5년마다 실시하여 2020년 인구총조사는 제20차, 주택총조사는 제12차에 해당된다.

인구주택총조사 추진체계

최근 한국의 인구주택총조사는 외국에서 벤치마킹할 정도로 혁신하고 있다. 2010년에 본격적

으로 인터넷조사를 도입하여 국민의 50%가 인터넷으로 응답하였다. 이를 통해 국민의 편리성이 제고되고 개인정보보호도 강화되었으며, 예산도 절감되었다.

지난 2015년에는 공공데이터를 활용하는 등록센서스 방식을 도입하여 전수조사를 대체하고, 행정자료로 수집할 수 없는 항목은 전 국민의 20%를 표본으로 조사를 실시하였다. 이를 통해 국민 응답부담이 80% 절감되고, 예산도 약 1,500억 원 절감되었다.

2020년에는 표본조사 항목의 등록센서스 전환을 확대하고, 사회·경제 변화와 정책 수요를 반영한 안전과 환경, 반려동물 등을 신규 항목으로 조사한다. 또한 현장조사에는 총 조사 최초로 조사원이 태블릿을 활용한 면접조사를 실시하고, 국민은 모바일로 인터넷조사에 참여하고 콜센터를 통해 전화조사가 가능하다. 이를 통해 전면적인 전자조사(Paperless Census)를 추진한다. 종이조사표의 입력내검과정이 거의 없어짐에 따라 조사의 정확성을 제고하면서 자료처리기간을 전주기 대비 3개월 단축하여 자료의 시의성을 제고할 예정이다.

인구주택총조사는 왜 하는가

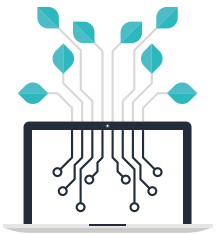
UN의 정의에 따르면 인구주택총조사는 국가가 주관이 되어 특정한 시점에 한 국가 또는 일정한 지역의 모든 사람, 가구, 거처와 관련된 인구·경제학적 및 사회학적 자료를 수집, 평가, 분석, 제공하는 전 과정을 의미한다. 인구주택총조사는 국가 영토 내의 모든 사람과 거처를 개별적으로 기준시점에서 일정 기간에 일정한 주기로 파악하는 국가 기본 통계조사로, 중앙정부와 지방정부의 정책입안 및 각종 중장기 계획수립, 대학·연구소, 기업의 연구 및 평가 등 각종 분야의 기초자료로 활용된다.

* 중앙행정기관 및 지방자치단체의 주요 활용 사례

주요정책 및 사업	활용된 자료	활용기관
• 보통교부세 산정(매년)	가구 수	행정안전부
• 지역경제활성화 계획수립(매년)	인구, 가구, 주택현황	17개 시도
• 주거종합계획(2019)	가구 및 주택항목	국토교통부
• 주택시장 안정대책(2018)	인구 · 가구 구조	기획재정부
• 제1차 아동정책 기본계획(2015) • 독거노인 종합지원대책(2018)	가구, 노인인구, 혼인상태	보건복지부
• 평생교육진흥 기본계획(2018)	연령별 교육정도	교육부
• 제2차 에너지 기본계획(2014) • 제8차 전력수급 기본계획(2017)	인구, 가구 · 가구원, 주택	산업통상자원부
• 양성평등정책 기본계획(2016)	가구원, 가구주의 성 · 연령	여성가족부
• 중장기 인력수급전망(2016)	인구, 가구원	고용노동부
• 제3차 관광개발기본계획(2012)	인구, 가구, 주택현황 등	문화체육관광부

또한 읍면동 단위까지 통계가 작성·제공되는 유일한 조사로서, 지방자치단체 등 지역 정책의 수립 및 개발을 위한 기초 자료를 제공하고 지리정보시스템(GIS)과 연계한 소지역별 정보를 제공하여 다양한 공간 분석에 활용이 가능하다.

한편 인구주택총조사는 모든 사회분야 통계의 기준(benchmark) 통계로서, 각종 가구 표본조사의 모집단 및 표본 틀을 제공하는데, 2020년 2월 기준 가구대상 조사통계 292종 중 75.3%인 220종의 표본조사가 인구주택총조사 자료를 모집단으로 활용하고 있다.



조사 환경 변화에 따라 전자조사 도입

이렇듯 국가의 중장기적 계획수립 및 각종 연구, 경영의 기초자료가 되는 인구주택총조사는 전 국민의 적극적인 관심과 참여가 무엇보다 중요하다. 그러나 최근의 조사환경은 그리 녹록하지 않다.

1인 가구와 맞벌이 가구가 급격히 증가하여 지난 2018년 기준으로 1인가구 비율은 29%를 상회하고 맞벌이 가구는 46.3%에 이른다. 그만큼 낮 동안 면접이 어려운 가구가 늘어나고 있고 주택의 형태도 아파트, 원룸과 같이 정문에서부터 외부인 출입이 통제되는 경우가 증가하고 있다. 무엇보다 개인의 프라이버시를 중요시하는 사회인식의 변화로 개인과 가족의 정보 제공을 기피하므로 총조사의 불응이 지속적으로 증가하여 2015년 총조사 기준 부재 및 불응률이 2.4%에 이르고 있다.

인건비의 상승으로 급격하게 증가하고 있는 총조사 비용도 문제이다. 2020년에도 기존의 방법대로 읍면동에 조사상황실을 설치하고 종이조사표로 조사할 경우 2015년 대비 30억 원 증가한 1,370억 원의 조사비용이 소요될 것으로 추정되었다.

한편, 국민들의 스마트폰 사용이 일반화되어, '18년 PC 보유율은 67.9%이나 스마트폰은 89.4%으로 높아, 스마트폰으로 인터넷조사 참여요구가 많았다.

이러한 조사환경의 변화 속에서 신기술을 도입하여 전면적인 전자조사를 추진하였다. '18.1월 UN이 2020 인구주택총조사 라운드에 태블릿 등 전자조사 도입을 권고하였고, '18년 상반기에 정보전략계획을(ISP)을 세워 전자조사 체계를 만들었다. '18.5월에 이를 바탕으로 청내 회의에서 2019년 가구주택기초조사 및 2020년 인구주택총조사에 태블릿 도입을 결정하였다. '18.11월에는 인구주택총조사 조사방법 혁신을 위한 국제워크숍을 개최하여 미국, 캐나다, 싱가포르 등의 전자조사 경험을 공유하였다.

'19.2월에는 전면적인 전자조사를 위해 인터넷조사(스마트폰), 태블릿조사(CAPI), 전자지도(GIS), 전화조사(CATI), 콜센터(채팅, 문자상담) 등을 위한 시스템을 구축하였다. '19.6월에는 태블릿을 2년에 걸쳐 28천대('19년 11천대, '20년 17천대)를 구매하였다.

이런 시스템 및 기기의 현장조사 적합성과 타당성 검증을 위한 시험 및 시범예행조사를 실시하였다. '17.10.~11월 1차 시험조사에서 스마트폰으로 인터넷조사가 가능한지 테스트하였고, '18.10.~11월 2차 시험조사에서는 태블릿 면접조사, 전화조사, 시군

구 상황실, 신규 조사항목 등을 테스트하였다. '19.5.~6월 3차 시험조사에서는 전자지도, 실시간 현장관리, 업무분장, 참여안내문 우편 배부 등을 테스트하였다. '19.10.~11월 시범예행조사에서는 분야별 최종 종합점검을 하였다. '19.7월에는 인구주택총조사 규칙을 개정하였다.

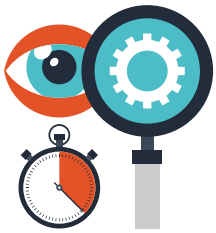
2020 인구주택총조사는 어떻게 시행될까

2020 인구주택총조사는 11월 1일을 기준으로 전 국민의 20%를 대상으로 10.15.부터 인터넷조사와 전화조사등 비대면 조사를 우선 실시한다. 인터넷 참여 안내문은 우편으로 2번에 걸쳐 가구에 배송된다. 인터넷조사기간에 응답하지 않은 가구는 11.1.부터 조사원이 가구를 방문하여 태블릿으로 전자조사를 실시하게 된다.

조사요원은 약 2만 7천 명이 동원되는데, 시설조사구는 조사지원담당자가 조사한다. 조사관리자는 조사원 10명당 1명이 배정되며, 조사원은 전자조사표로 입력 및 자동내검을 하고 서버로 실시간 전송함으로 2015년과 같이 종이조사표 내검을 위해 상황실을 방문할 필요가 없다.

* 2020 인구주택총조사의 개요

2020 인구주택총조사 개요	
연혁	• 인구총조사는 1925년, 주택총조사는 1960년 이후 매 5년마다 실시 – 2015년 인구총조사는 제20차, 주택총조사는 제12차에 해당
법적근거	• 총조사 실시(통계법 제5조의 3), 지정통계(동법 제17조 제1항) – 인구총조사 : 지정통계 제101001호, 주택총조사 : 지정통계 제101002호 • 인구주택총조사 규칙(기획재정부령 제738호, '19. 7. 11. 일부개정)
조사기준 시점 및 대상	• 기준시점 : 2020. 11. 1. 0시 현재 • 조사대상 : 대한민국 영토 내에 상주하는 모든 내 · 외국인과 이들이 살고 있는 거처
조사항목	• 전수 16개, 표본 55개(현장조사 45, 행정자료 대체 10)
조사방법	• 전수조사 : 등록센서스 • 표본조사 : 현장조사 실시
조사기간	• 인터넷 및 전화조사 : 2020. 10. 15. ~ 10. 31.(17일 간) • 방문 면접조사 : 2020. 11. 1. ~ 11. 18.(18일 간) ※ 준비조사 : 2020. 10. 31.(1일)
현장조사 실시체계	• 통계청(주관기관), 지방자치단체(실시기관)
현장조사 동원인력 (30천 명)	• 공 무 원 : 1,280명(통계청 459명, 지자체 821명) • 조사요원 : 27,047천 명 ※ 총관리자 250명, 조사관리자 2,319명, 조사원 23,054명, 조사지원담당자 1,424명
소요예산 (2020년)	949억 원



8대 중점 추진 과제를 살펴보면

2020 인구주택총조사는 조사방법을 다양화하여 국민들의 총조사 참여기회를 대폭 확대하였다. 먼저 국민들이 보다 쉽고 편리하게 조사에 참여할 수 있도록 인터넷 조사를 모바일까지 확대하고 인터넷취약계층이 비대면 조사에 참여할 수 있도록 총 조사 최초로 콜센터를 통한 전화조사를 도입할 계획이다.

한편 면접조사에 있어서는 그동안의 종이조사표 틀에서 벗어나 태블릿을 활용한 전자조사를 실시하게 된다. 이렇게 되면 조사와 동시에 기초 내검이 가능하여 자료의 정확성이 높아지고 자료처리기간의 단축이 가능하여 시의성 있는 자료를 제공할 수 있다. 자세한 내용을 2020년 인구주택총조사 8대 중점추진과제에서 살펴보자.

* 2020 인구주택총조사 중점 추진방향

추진 사항	추진 내용
① 태블릿 PC로 면접조사(CAPI) 실시	• 전자지도가 장착된 태블릿으로 실시간 조사 및 자료 전송 • 태블릿의 보안 강화 및 전용 통신망 구축 • 태블릿 사용이 가능한 조사원 및 교관 확보, 교육 강화
② 전자조사에 따른 현장조사 체계 혁신	• 상황실을 읍면동에서 시군구로 변경 • 지리정보기술(GIS, GPS)을 활용한 현장관리 기능 강화 • 태블릿을 활용하여 조사원간 공정한 업무분장 실시
③ 국민의 조사 참여 방법 다양화	• 국민의 인터넷 및 전화조사 참여 확대로 방문조사 최소화 • 스마트폰으로 인터넷조사가 가능하도록 국민편의성 제고 • 인터넷이나 대면 조사를 선호하지 않는 국민은 전화조사 가능
④ 매체 다변화를 반영한 국민 중심 홍보	• 조사참여가 국민의 삶에 도움이 된다는 메시지 강조 • 국민의 매체 이용 변화를 반영한 디지털 홍보 강화 • 인터넷 응답률 제고를 위한 홍보 강화
⑤ 자료처리 고도화 및 기간 단축	• 전자조사 도입에 따른 효율적인 입력 내검 체계 구축 • 내검단계를 3단계로 체계화하여 효율적인 내검 수행 • 행정자료 활용 확대를 통한 무응답 대체 및 E&I 시스템 고도화
⑥ 등록센서스 등 행정자료 활용 확대	• 기존 표본조사의 일부 항목을 등록센서스로 전환 • 등록센서스 자료를 가구주택기초조사로 보완하여 조사모집단 생성 • 신규 공동주택은 국토교통부자료의 입주예정자료로 보완
⑦ 공표 항목 증가 및 시기 단축	• 사회 · 경제 변화 및 정책수요 반영한 신규 조사항목 발굴 • 공표항목 증가 및 공표시기 단축으로 자료 활용성 및 시의성 제고
⑧ 통계서비스 다각화	• 마이크로데이터 활용 활성화를 통한 부가가치 제고 • 시각화된 콘텐츠(GIS기반, 인포그래픽) 제공

2015년 대비 달라진 점… 스마트폰 · 전화조사 도입

2020년 인구주택총조사를 2015년과 비교해보면 예산이 24억 절감이 되고, 조사항목은 45개로 4개가 감소하였다. 조사방법으로는 스마트폰으로 인터넷조사가 가능하고, 전화조사가 처음으로 도입되었다. 상황실이 읍면동에서 시군구로 전환되었고, 조사원은 종이 조사표 및 지도 대신 태블릿에 탑재된 전자 조사표 및 지도를 사용하게 된다. 조사진척률은 실시간으로 파악가능하며, 공평한 업무분장을 위해 인터넷조사율을 고려하고, 조사구를 나눌 수 있다. 전자조사표 사용으로 오류건수가 줄고, 공표를 3개월 앞당길 것이다. 공표항목은 행정자료 확대로 3개가 늘어난 56개가 되며, 홍보에서 디지털 매체의 비중이 크게 높아질 것이다.

* 2020 인구주택총조사의 전주기(2015) 비교

구분		2015년	2020년
조사 환경	• 조사예산 • 조사대상 • 표본항목	973억 원 360만 가구 49개	949억 원 4백만 가구 45개
조사 방법	• 인터넷 • 면접 • 효과	PC 종이조사표 –	PC, 모바일기기(스마트폰 등) 태블릿(CAPI), 전화(CATI) 국민 및 조사원 편의성 제고
현장 조사	• 상황실 • 지도 • 조사관리 • 응답시간 • 조사원 • 조사관리자 • 업무분장 • 효과	읍면동 단위(3,500개) 종이지도 대면 관리 20분(4인 가구 기준) 상황실 정기방문 상황실에서 조사표 내검 인터넷 조사 전 업무분장 –	시군구 단위(250개) 전자지도 실시간 관리 15분(4인 가구 기준) 상황실 방문불필요 현장에서 조사원 지도 및 지원 인터넷조사 응답률 및 가구수를 고려하여 업무분장 비용 절감 및 현장조사 관리 개선
입력 내검	• 입력방법 • 오류건수 • 처리기간 • 내검방법 • 효과	종이조사표 작성 후 스캐닝 2.8건(종이조사표) 8개월 아이체킹, 기계내검, 수동내검 –	조사와 동시 입력 · 전송 1.1건(전자조사표) 6개월 기계내검, 수동내검 자료처리 기간 단축, 정확성 향상
결과 공표	• 공표항목[1] • 공표시기 • 서비스유형 • 효과	53개 전수('16.9월), 표본('16.12월) 통계표 중심 –	56개 전수('21.7월), 표본('21.9월) 시각화(GIS, 인포그래픽) 자료 활용성 및 시의성 제고
홍보	• 메 시 지 • 광고매체 • 효과	정책 등 정부 활용 강조 전통매체(TV, 신문) 중심 –	국민에게 도움되는 측면 강조 전통매체와 디지털매체 균형 집행 국민의 관심과 참여 유도

1) 등록센서스 방식의 전수조사항목 포함

조사항목… 무엇에 대해서 조사를 할까

조사항목 선정시 기본방향은 기본항목에 대한 시계열을 유지하고, 사회경제 변화상 및 정책수요를 반영하며, 조사항목을 행정자료로 대체하여 국민의 응답부담을 경감하는 것이다. 조사항목은 중앙 및 지방정부, 연구기관 및 학계 등에서 18회의 의견수렴을 걸쳐 선정되었으며, 시험조사 3회 및 시범예행조사 1회를 통해 검증하였다.

2015년 대비 2020년 조사항목의 주요변화는 조사항목의 행정자료 대체를 3개에서 10개로 늘리고, 신규항목은 국적취득연도, 활동제약돌봄, 1인가구 사유, 혼자산기간, 안전·환경, 마시는 물, 반려(애완)동물 등 7개를 추가하였다. 조사항목은 4개가 줄어 45개이며, 공표항목은 총 56개로 3개가 늘었으며, 등록센서스(전수)는 4개가 늘어 16개가 되었다.

2020 조사항목의 특징

* 연도별 조사항목 수

구분	2000년	2005년	2010년	2015년	2020년				
					1차시험	2차시험	3차시험	시범예행	최종(안)
계	50	44(3)[1]	50(3)	53	51	58	60	56	56
인구	29	24	31	35	33	37	37	33	34
가구	16	11	13	12	12	15	17	17	16
주택	5	6	6	6	6	6	6	6	6
전수	20	21	19	12	11	21	21	13	16[2]
표본	50	44	50	52	48	43	45	50	55

1) ( )안은 시도 항목 수
2) 전수는 국민전체에 대해 등록센서스로 수집되는 항목
2020년 표본 항목 : 55개(현장조사 45개, 행정자료 대체 10개)
2015년 표본 항목 : 52개(현장조사 49개, 행정자료 대체 3개)

* 2020 공표항목

	전수 항목(16)	표본 항목(55)		
		현장조사(45)		행정자료(10)*
인구 (34)	① 성명 ② 성별 ③ 나이 ④ 가구주와의관계 ⑤ 국적 ⑥ 입국연월 ⑦ 1년전거주지 ⑧ 국적취득연도	① 성명 ② 성별 ③ 생년월일(나이) ④ 가구주와의관계 ⑤ 국적 ⑥ 입국연월 ⑦ 교육정도 ⑧ 교육영역 ⑨ 출생지 ⑩ 아동보육 ⑪ 활동제약 ⑫ 활동제약 돌봄 ⑬ 통근 · 통학여부 ⑭ 통근 · 통학장소 ⑮ 이용교통수단 ⑯ 통근 · 통학 소요시간	⑰ 경제활동상태 ⑱ 종사상지위 ⑲ 산업 ⑳ 직업 ㉑ 현직업근무연수 ㉒ 근로장소 ㉓ 혼인상태 ㉔ 혼인연월 ㉕ 출산자녀수 ㉖ 자녀출산시기 ㉗ 추가계획자녀수 ㉘ 결혼전취업여부 ㉙ 경력단절 ㉚ 사회활동 ㉛ 생활비원천	1년전거주지 5년전거주지
가구 (16)	① 가구구분 ② 주거시설형태	① 가구구분 ② 1인가구사유 ③ 혼자산기간 ④ 반려(애완)동물 보유여부 ⑤ 마시는물 ⑥ 주택용소방시설 ⑦ 사용방수	⑧ 난방시설 ⑨ 주차장소 ⑩ 건물및거주층 ⑪ 거주기간 ⑫ 주거전용 · 영업 겸용여부 ⑬ 점유형태 ⑭ 임차료	주거시설형태 타지주택소유여부
주택 (6)	① 거처의종류 ② 총방수 ③ 주거시설수 ④ 주거용연면적 ⑤ 대지면적 ⑥ 건축연도			거처의종류 총방수 주거시설수 주거용연면적 대지면적 건축연도

* 행정자료 대체 항목 10개
** 음영표시: 전수 및 표본 공통항목 15개
*** 조사항목 56개: 전수 16개+표본 55개−공통 15개

코로나 극복의 역량이 인구주택총조사까지 이어지기를

통계청은 급속히 변화하는 환경에 대비하기 위해 지난 2018년 인구주택총조사에 전자조사 도입을 결정하고 그동안 3번의 시험조사와 1번의 시범예행조사를 통하여 현장조사방법과 체계를 점검하였으며, 지난해에는 가구주택기초조사를 실시하여 가구와 주택의 변화를 반영하는 등 차근차근 2020 인구주택총조사를 준비해왔다.

한편 국민이 안심하고 참여할 수 있도록 보안을 강화하기 위해 다양한 조치를 취했다. 태블릿에 MDM(보안솔루션)을 설치하여 조사에만 사용하고, 분실 시 자료유출을 방지하고자 한다. 자료 전송은 기업전용통신망을 활용하여 외부침입을 차단한다.

우리나라는 지난 4월 '강력한 사회적 거리두기' 기간에 실시한 21대 총선에서 66%가 넘는 높은 참여율로 안전하게 성공적으로 선거를 치루는 등 모든 국민이 서로를 믿고 배려하며 코로나19를 슬기롭게 극복하고 있다.

이런 성숙한 시민의식으로 오는 10~11월 인구주택총조사에서도 유감없이 발휘되어 세계의 모범이 될 수 있도록 국민 여러분의 많은 관심과 적극적인 참여를 부탁드린다.

이용희 ◆ 서울시립대 통계학과 교수

데이터 정보보호… 누구나 원하는 통계를 얻을 수 있는 **정보 평등 사회의 조건**

내가 2005년 한국의 통계 작성기관들과 학문적 교류를 시작하면서 전공분야도 아닌 통계적 정보보호 방법에 관심을 가지게 된 이유는 불평등에 대한 안타까움 때문이었다. 당시 통계 작성기관들은 보유하고 있는 마이크로데이터를 원하는 일반 사람들에게 제공하고 싶어도 정보 유출에 대한 우려를 해소할 수 있는 적절한 방법이 없어서 제공할 수 없었다. 반면에 작성기관과 특별한 관계를 가진 사람들은 원시자료를 쉽게 얻을 수 있는 기회도 있었다. 조금 애매한 표현이지만 이러한 불평등에 대하여 안타까움을 느끼면서 이를 해결할 수 있는 방법이 꼭 있어야 한다고 느낀 것이다. 누구는 자료를 사용하여 원하는 정보를 얻을 수 있고 누구는 얻을 수 없다는 환경이 매우 안타까웠다.

"누구나 원하는 통계를 얻을 수 있는 사회가 되었으면 좋겠다."

이것이 15년 전에 내가 상상한 한국 사회이다. 여기서 주목할 점은 내가 상상한 사회는 '누구나 데이터를 얻을 수 있는 사회'가 아니라 '누구나 통계를 얻을 수 있는 사회'라는 점이다.

정보 공개의 조건 … 마이크로데이터 그리고 빅데이터

2000년대 초반 통계청에서 통계적 정보보호 방법에 관심을 가지고 있었던 소수의 전문가들이 기초적인 연구를 시작하면서 마이크로데이터 공개에 대한 연구를 시작하였다. 그 후 학계와 공동으로 정보보호 방법에 대한 다양한 연구들을 수행했고, 각종 관련 매뉴얼도 만들어졌으며, 교육 과정도 통계교육원에 만들어졌다. 현재 이러한 노력의 결과로 많은 통계작성 기관들은 마이크로데이터 공개 서비스를 제공하고 있으며 공개의 범위도 넓혀가고 있는 추세이다.

마이크로데이터를 일반에게 공개하기 위해서는 많은 노력이 수반된다. 노력이 많이 드는 이유는 자료가 공개되었을 때 발생할 수 있는 개인정보의 노출에 대한 위험을 줄이기 위해서다. 공개된 자료에서 몇 가지 항목을 조합하여 다른 외부 자료와 결합하면 개인의 노출이 쉽게 일어날 수 있다.

이러한 위험을 줄이기 위해서 다양한 통계적 정보보호 방법을 이용한다. 대표적인 방법이 구간화, 감추기, 바꾸기, 잡음 첨가 등이며 이러한 방법들은 흔히 식별변수라고 부르는 중요한 항목들의 실제 값을 변형하는 대가로 노출의 위험을 감소시킨다. 현재 서비스되고 있는 마이크로데이터는 이러한 전통적인 통계적인 방법들이 적용된 후에 공개되고 있다.

이러한 항목의 값을 변형하는 방법들이 최전선에서 개인 정보를 방어하고 있을 때 최근 '빅데이터'라는 복병을 만나게 되었다. 빅데이터의 개념을 이해할 때 흔히 사람들은 빅(big), 즉 '크다'는 개념을 단위(unit)의 수가 많다는 것으로 생각한다. 예를 들어, 표본조사에서 얻은 자료는 많아야 몇 만 명의 자료이지만 빅데이터는 수백만 명, 수천만 명의 자료로 구성되어 있다고 생각한다.

이것이 틀린 생각은 아니지만 빅데이터의 또 하나의 중요한 요소는 항목(item)의 수가 많다는 것이다. 표본조사에서 개인이 응답한 항목이 100개를 넘는 경우는 매우 드물다. 하지만 빅데이터에서는 개인에 대한 항목의 수가 상상 이상으로 많다. 예를 들어, 통신회사에서 수집하는 개인의 이동 경로는 항목의 개수로는 가늠할 수 없는 거의 무한대의 차원이다. 지구에서 한 위치의 점이 1차원이라면 이를 연결한 선, 즉 경로는 무한차원이다. 결론적으로 빅데이터가 가진 개인 정보의 양은 마이크로데이터가 가지고 있는 개인 정보의 양과 비교할 수 없을 정도로 크다.

마이크로 자료에서 중요한 항목들에 대한 실제 값을 변형하는 목적은 몇 개의 식별변수의 값을 조합해서 유일한 존재가 나오는 것을 최대한 방지하기 위해서다. 예를 들어, 어떤 조사에서 '성별은 남자, 나이는 53세, 직업은 교수, 직장은 서울시 동대문구 전농동에 위치한다'는 정보가 공개되면 다른 항목 또는 외부 자료를 이용하여 항목들의 조합에 유일하게 해당하는 개인이 필자라는 것을 알아내는 것은 그리 어려운 일이 아니다. 이러한 노출의 위험을 낮추기 위하여 공개된 정보에서는 '나이는 50대, 직업은 전문직, 직장은 서울시 동

대문구'와 같이 구간화하고 감추기를 하면 노출의 위험은 줄어든다.

이러한 단순한 방법들을 빅데이터에 그대로 적용하는 것은 쉬운 일이 아니다. 개인의 이동경로는 정보보호를 위하여 어떻게 변형해야 하는지 전통적인 방법이 해결책을 제시하기 어렵다. 더 나아가 빅데이터는 중요한 식별변수를 다른 변수와 구별해서 지정하는 것이 쉽지 않다. 빅데이터에서는 대부분의 항목이 중요한 식별변수이다.

더 나아가 현재는 연결의 시대이다. 사회망으로 개인들이 촘촘히 엮여 있을 뿐만 아니라 데이터도 밀접하게 연결되어 있다. 이러한 연결은 우리에게 편리함을 주는 축복과 같은 존재이다. 데이터도 마찬가지이다. 빅데이터는 한 개인에 대한 정보의 양이 많아서 서로 다른 데이터를 연결하기 쉽다. 특히 연결된 서로 다른 형태의 데이터는 정보의 양이 수십 배로 증폭될 수 있다. 데이터를 상업적으로 이용하려는 사람들에게는 축복이다. 데이터의 사용과 연결을 위한 데이터거래소 등이 생기는 이유이다. 하지만 빅데이터의 연결 때문에 개인 정보보호를 지키는 전쟁에서 전통적인 무기들이 무력화되는 위험에 처한 것이 현실이다.

정보보호의 새로운 접근…차등정보보호, 재현자료

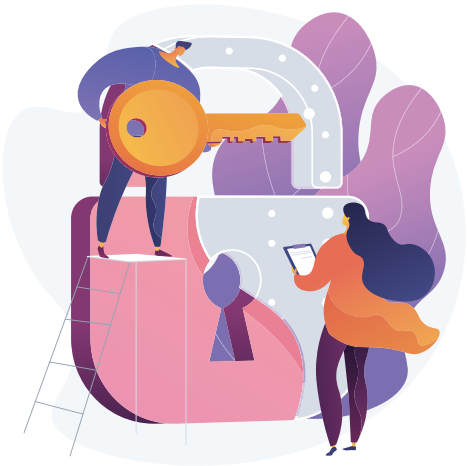
IT가 고도로 발달된 시대에 반드시 마이크로 자료를 공개해야 원하는 정보를 얻을 수 있을까? 다른 방법으로 정보를 주고 새로운 방법으로 노출 위험을 조절할 수 있는 방법은 없을까? 내가 상상하던 평등한 사회를 2000년대 초반에 컴퓨터 과학자들이 먼저 생각하고 새로운 정보보호의 개념을 제시하였다.

새로운 정보보호의 개념은 차등정보보호(differential privacy)라고 불린다. 차등정보보호는 개인이 마이크로 자료를 얻어서 직접 정보를 생성하는 환경이 아니라, 데이터 분석을 제공하는 서버에 원격으로 접속하여 원하는 정보를 얻는 환경에서 적용되는 개념이다. 누구나 자료를 얻을 수 있는 사회가 아니라 누구나 정보를 얻을 수 있는 사회를 고려한 것이다.

한 번이라도 자료 분석을 해본 경험이 있는 사람이라면 통상적인 분석 절차를 상상해보자. 자료를 파일 형태로 받아서 컴퓨터에 저장하고 데이터베이스 또는 통계 프로그램을 이용하여 자료를 변형하고 분석한다. 예를 들어, 통계청 마이크로데이터 통합서비스에서 가계금융복지조사의 마이크로데이터를 받아서 저장하고, 통계 프로그램을 써서 자료를 변형하고 통계 수식을 적용하여 우리 나라 소득 하위 70%선을 추정할 수 있다.

차등정보보호를 제안한 사람들은 이러한 전통적인 절차가 아니라, 자료의 획득과 처리가 원격으로 분리되어 처리되는 환경을 고려한 것이다. 개인이 마이크로데이터를 보거나 직접 처리하지 않고 원하는 정보를 얻을 수 있는, 원격으로 접속하는 대화형 분석 환경을 생각한 것이다.

이러한 환경에서는 이용자가 마이크로데이터를 직접 볼 수가 없기 때문에 안전한 것 같지만 새로운 노출의 위험이 발생하여 정보보호의 개념을 새롭게 정의하였다. 새로운 노출의 위험은 대화형 시스템에 접속하여 조금씩 다른 형태의 통계를 계속해서 반복 생산하면 직접 볼 수 없는 자료를 어느 정도 복원할 수 있고, 그로 인해 개인정보가 노출될 수 있는 위험이다. 예를 들어, 대화형 통계분석 시스템에 접속하여 소득 하위 0.000001%부터 소득 하위 99.999999% 까지 0.000001%씩 증가시키면서 십만 번의 통계를 얻으면 가계금융복지조사 표본 20000가구의 소득을 거의 복원할 수 있을 것이다.



차등정보보호는 이러한 새로운 개인 정보의 노출 위험성을 계량화할 수 있는 개념이다. 전통적인 통계적 기법은 항목의 값을 한 번만 변형하지만 차등정보보호를 고려하면 정보의 생성 과정에서 미세한 변형을 연속적으로 적용하는 방법들이 이용된다.

최근에 주목받고 있는 다른 형태의 새로운 정보보호 방법은 재현자료(synthetic data)이다. 기계학습과 인공지능의 발달로 실제 마이크로데이터와 유사한 구조와 특성을 가진 인공으로 생성된 자료를 재현자료라고 한다. 마이크로데이터를 공개하지 않고 재현자료를 공개하면 보안의 안정성이 크게 향상된 자료를 일반에 공개할 수 있다. 재현자료는 안전성이 많이 향상된 대신 유용성이 떨어진다는 단점이 있지만 선행 연구를 위한 분석이나 교육 자료용으로 사용될 수 있다.

누구나 필요한 정보를 얻을 수 있는 세상을 꿈꾸며

데이터에 대한 모든 환경이 너무 빠르게 변하고 있다. 이에 대응할 수 있는 새로운 정보보호에 대한 개념들과 방법들도 함께 연구되고 실제로 적용되고 있다. 하지만 새로운 개념과 방법이라고 해서 전통적인 방법들이 가지고 있는 본질에서 벗어난 것은 아니다. 새로운 정보보호 방법도 심하게 적용하면 유용성이 감소하는 것이 사실이다. 또한 하나의 방법이 모든 경우에 보편적으로 적용되는 것도 아니다.

지금까지 마이크로 자료의 공개를 위하여 노력한 것보다 더 노력해야 빅데이터 환경에서 정보보호에 대한 성과가 날 수 있을 것이다. 하지만 우리나라는 IT 분야 강국으로서 대화형 원격 통계 시스템에 대한 새로운 기술을 다른 나라보다 빠르게 개발할 수 있는 역량이 있다고 본다. 누구나 필요한 정보를 얻을 수 있는 환경이 우리나라에 먼저 생기는 상상이 현실이 될 수 있기를 바란다.

손은락 ◆ 통계청 통계기준과장

통계청–UNODC 공동 「아·태 범죄통계 협력센터」 설립

(2020. 5. 11., 통계기준과)

지금으로부터 약 20년 전 개봉되었던 스티븐 스필버그 감독의 〈마이너리티 리포트〉라는 영화가 있다. 영화 속에서 국가는 미래에 발생할 범죄와 범인을 예측하는 시스템을 구축하여 범죄가 실제로 발생하기 전에 미리 범인을 체포함으로써 범죄로부터 자유로운 사회를 지향한다.

영화 속의 이야기는 최근 미국 캘리포니아 LA 경찰청에서 범죄예측서비스(PredPol)를 통해 현실이 되고 있다. 범죄예측은 범죄와 관련된 모든 정보를 사전에 축적하고 활용함으로써 가능하다. 최근 들어 IT 기술이 엄청나게 발전하면서 범죄 데이터의 축적·관리·분석 시스템에 관한 관심도 높아지고 있다. IT 기기의 보급과 데이터 처리능력의 혁명적 발전은 비정형적 형태의 빅데이터까지도 시스템화하는 가능성을 열어주었고, 국제사회에서는 인공지능 기술을 활용하여 각종 범죄데이터의 결과를 분석하고 범죄를 예측하는 논의가 한창이다.

특히, 국제공조 수사에서 범죄데이터를 활용한 치안정책의 효과가 인정되면서 효과적인 범죄 데이터 생산과 활용을 위한 국가간 협력이 강화되고 있다. 우리나라에서도 통계청을 중심으로 범죄분류체계 개발, 통계의 수집 및 활용방안에 대하여 UN 등 국제사회와 긴밀한 협력관계를 높여 나가는 중이다.

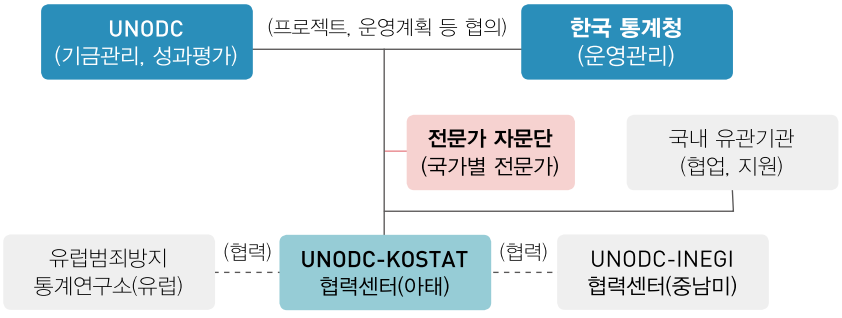
범죄통계 국제협력센터의 설립

통계청은 2016년부터 우리나라 범죄통계[1]를 UN에 제공하고 있으며 범죄 데이터의 편리한 활용을 위해 데이터 정형화에 필요한 국제표준[2] 도입을 추진하고 있다. 이와 더불어 아시아·태평양 지역의 범죄통계 허브 구축을 위해 UNODC[3]와 공동으로 2019년 5월에 '아·태 범죄통계 협력센터'(이하, 협력센터)를 설립하였다.

멕시코에 이어 세계에서 두 번째로 설립된 '아·태 범죄통계 협력센터'는 유엔마약범죄사무국이 한국 통계청에 설립을 제안하면서 사업추진이 본격화되었다. UNODC는 국제적 차원의 범죄예방 정책 지원을 위한 국제표준범죄분류(ICCS)의 보급, 아·태지역 범죄통계 인프라 구축을 목표로 아·태지역 내에 협력센터 설립을 희망하였다. 이에 범죄통계 작성·분석·보급 방법론에 관한 선진기법을 도입하고 범죄통계를 한 단계 발전시키는 기회로 활용코자 협력센터 유치가 본격화되었다.

이러한 배경 속에 UNODC가 통계청에 협력센터 설립을 2018년 2월 제안하였고, 3월 뉴욕에서 개최된 UN통계위원회에서 통계청과 UNODC는 범죄통계 발전과 통계발전 인프라 거점을 강화하기 위하여 협력센터를 대전에 설치하기로 합의했다. 이후 1년여의 준비 기간을 거쳐 지난해 5월 대전에 협력센터를 개소하였다.

아 · 태 범죄통계 협력센터 운영 체계



1) UN-CTS(United Nations Surveys on Crime Trends and the Operations of Criminal Justice Systems): UN에서 여러 나라의 범죄사건 통계자료와 형사사법제도 운영에 관련된 자료들을 수집하는 조사

2) 국제범죄분류(ICCS, International Classification of Crime for Statistical Purposes): 범죄 통계의 일관성과 국제 비교성을 높이고, 국내 및 국제적 차원의 분류 해석능력을 향상시키기 위해 국제적으로 합의된 개념, 정의 및 원칙에 근거한 범죄 분류체계

3) 유엔마약범죄사무국(UNODC, United Nations Office on Drugs and Crime): 약물 규제와 마약 범죄 예방을 목적으로 1997년 설립된 유엔 산하기관으로 범죄통계와 제도협력을 담당

협력센터의 주요 활동과 향후 발전모습

협력센터는 지난해 출범 이후 아시아 태평양 지역의 범죄통계 역량제고, 국내·외 범죄통계 허브기관, 범죄·안전·정의와 관련된 정책연구와 지표개발 등을 주요 활동목표로 정하고 다양한 행사와 국제활동을 전개하였다.

그림 1 | 아 · 태 범죄통계 협력센터 개소 기념식(2019.8.)

2019년 8월에는 협력센터 개소를 기념하여 국제 심포지엄을 개최하였다. 심포지움에는 UNODC뿐만 아니라 UNESCAP, UNSIAP 등 국제기구, 해외 연구기관 및 형사사법·통계작성 기관 고위급 인사를 포함하여 80여명이 참가하였고, 1박 2일간 범죄통계 개선을 위한 지역거점 강화방안 및 이를 위한 협력센터의 역할과 운영전략에 대하여 심도 있게 논의하였다.

협력센터는 같은 해 10월 UNODC, 중국 법무부와 공동으로 '제3차 아·태 지역 범죄 및 형사사법 회의'를 중국 청도에서 개최하였다. 이 회의를 통해 한국의 범죄분류체계 개발 현황, 개발 경험과 선도국 지원을 활용한 아·태 지역 범죄통계 개발 계획을 알리고 참여국가 간의 협력 네트워크를 공고하게 구축하였다.

또한, 해외 우수사례를 벤치마킹하기 위하여 2010년에 설립된 '중·남미 지역 범죄통계 협력센터'[4] 방문하여 협력센터의 중·장기 발전전략을 수립하고 운영 노하우를 습득하는 연수교육을 실시하였다.

그림 2 | 아 · 태 범죄통계 협력센터 개소 기념 심포지엄(2019.8.)

4) 중 · 남미 지역 범죄통계 협력센터(The Center of Excellence for Statistical Information on Government, Crime, Victimization and Justice), 중남미지역의 범죄통계 작성, 분석 및 모니터링 역량 강화를 위한 UNODC와 멕시코통계청 간 기술협력 활동 추진을 위해 2010년 설립된 기관을 2020년 현재 지역 범죄통계 분석, 교육, 국제협력 관련 3개팀에서 16명의 인원이 근무 중임

협력센터는 올해 설립 2년차를 맞이하고 있다. 아직은 초보단계에 있다고 할 수 있겠으나 통계청-UNODC 사업수행을 위한 국제기구 협력기관으로서의 소임과 아시아·태평양 지역 범죄통계 발전의 허브역할을 수행할 수 있도록 노력할 계획이다.

이를 위하여 우선 각종 범죄통계 관련 교육자료 개발 등 협력사업을 충실히 마련하여 지역 범죄통계 거점으로서의 활동을 준비하고 있다. 범죄통계 관련 기술협력을 원하는 아·태 지역 국가의 형사사법 통계기관 관계자를 대상으로 국제범죄분류, 범죄통계 데이터 수집방법 등을 중심으로 질 높은 국제 워크숍과 아·태지역의 범죄통계 활용 우수사례를 발굴하는 다양한 추진할 예정이다. 또한 국내에서 그동안 추진해왔던 한국범죄분류 개발성과와 향후 활용계획을 범죄통계 데이터 활용의 선도사례로 아·태 지역 여러 나라에 소개할 계획이다

'아·태 범죄통계 협력센터'는 아직 걸음마 단계에 있다. 그러나 향후 아·태 지역 범죄통계의 표준화, 범죄정보의 수집·분석방법의 공유, 국가간 협력체제 정비 등을 적극 추진해 나간다면 머지않아 국제사회에서 범죄통계의 중심기관으로 성장할 것으로 믿어 의심치 않는다.

＊협력센터-범죄통계 발전 방안

(범죄통계 조기 확충) 국제기구의 공신력 · 기술 · 경험을 활용하여 국제기준에 부합하고 범죄 관련 정책추진을 뒷받침할 국가통계의 효율적 조기 확충 및 관리체계 구축

(안전 · 청렴사회 구현) 고품질 범죄통계에 기반한 안전 및 반부패 정책 추진으로 안전 · 청렴사회 실현 및 국민 불안 · 불신 해소

(국가 위상 강화) 국제기구와의 강력한 파트너십 형성으로 아 · 태 지역 범죄통계발전을 주도함으로써 국제사회에서의 위상 및 영향력 강화

아·태 범죄통계 협력센터의 활동과 국제범죄분류에 관한 자세한 정보는 협력센터 홈페이지(http://unodc-kostatcoe.or.kr)를 통해 확인할 수 있다.

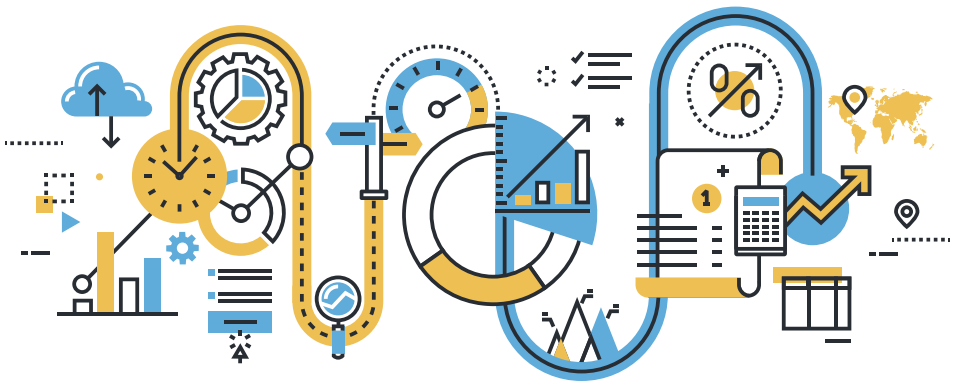
참고 | 추진 경과

○ UNODC, 한국 내 관련기관에 협력센터 설립 의사 타진	('13.~)
○ UNODC, 통계청에 협력센터 설립 제안	('18.2.)
○ 설립 합의	('18.3., 제49차 통계위원회)
○ 법무부, 대검찰청, 경찰청, 형사정책연구원 등 관계기관 협의	('18.4.)
○ UNODC와 통계청 간 협력센터 설립 약정서 체결	('19.2.)
○ 협력센터 사무소 현판식	('19.5.)
○ 협력센터 개소 기념 국제 심포지엄 개최	('19.8.)
○ 제3차 아 · 태 지역 범죄 및 형사사법 회의 공동 개최	('19.10.)

심송용 ◆ 한림대학교 데이터과학스쿨 교수

나도 R유저 ⑦

R에 도전하자… 따라가다 보면, **나도 R유저 ⑦**



'R에 도전하자' ①부터 ⑥에서 R의 설치부터 시작하여 몇 가지 기능을 따라 해보고 R에서의 색깔처리, R의 기초적인 통계함수 및 R의 벡터 및 데이터 프레임을 사용한 자료처리, 함수작성, 그래픽 장치 등에 대해 알아보았다. 이번 호에서 R 그래픽의 기초를 제공하는 plot 함수와 그에 관련된 몇 가지 함수에 대해서 알아보기로 하자.

1 | plot 함수

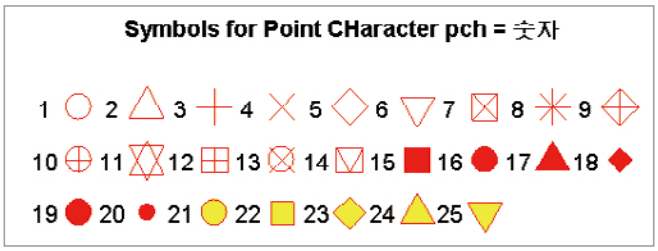
plot 함수는 n개의 짝으로 얻은 자료 (x_1, y_1), (x_2, y_2), …, (x_n, y_n)의 산점도나 함수의 그래프를 얻기 위해 사용하는 함수로 이 함수는

```
plot(x, y, type=, main=, sub=, pch=, lty=, ...)
```

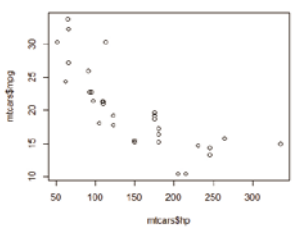
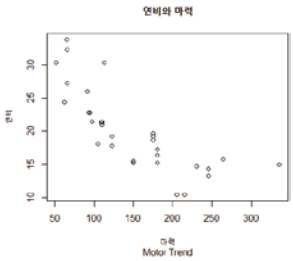
로 사용하며, 각각의 매개변수는

- x는 x-좌표값 x_1, x_2, …, x_n으로 이루어진 벡터를 설정하거나 행렬, data.frame 또는 plot 함수를 적용할 수 있는 객체를 설정한다. x에 행렬을 설정하면 첫 번째 열이 가로축, 두 번째 열이 세로축인 산점도를 그린다. data.frame인 경우 pairs 함수를 사용한 모든 가능한 산점도를 그린다.

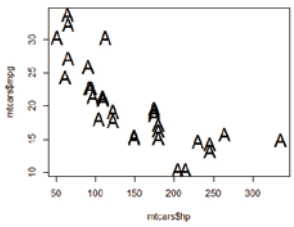
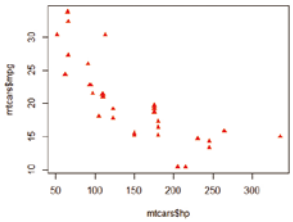
- y: $y_1, y_2, \cdots, y_n$을 원소로 갖는 벡터를 지정한다. x가 행렬 또는 plot 함수를 적용할 수 있는 객체인 경우 생략이 가능하다.
- main : 주 제목에 해당하는 문자열을 지정한다.
- sub : 보조제목에 해당하는 문자열을 지정한다.
- xlab, ylab : 각각 x축 및 y축 이름을 설정한다.
- pch : 산점도에서 점에 사용할 문자를 따옴표 안에 설정한다. "."(마침표)는 작은 한 픽셀짜리 점을 사용하며 그 외의 경우는 주어진 문자를 산점도의 점으로 사용한다. pch에는 숫자로 설정이 가능하며 숫자로 설정한 경우 각 숫자에 따른 점의 모양은 다음과 같다.



다음의 예는 R에 내장된 데이터 프레임인 mtcars 의 자료를 사용하여 마력(hp)과 연비(mpg)의 산점도를 얻은 것이다. mtcars는 32개 자동차 모델에 대한 모델명, 연비(mpg=mile per gallon), 마력(hp), 실린더수(cyl), 배기량(disp), 무게(wt) 등에 대한 정보가 저장된 데이터 프레임이다. mtcars에 대한 자세한 정보는 R에서 help(mtcars)로 볼 수 있다.

> plot(mtcars$hp, mtcars$mpg)	> plot(mtcars$hp, mtcars$mpg, main="연비와 마력", xlab="마력", ylab="연비", sub="Motor Trend")
	

위의 왼쪽 그림은 plot 함수에서 사용할 수 있는 가장 간단한 형태로 x축과 y축에 사용할 변수만 설정한 것이다. 이 경우 x축 및 y축의 이름은 plot 함수에서 설정한 변수의 이름과 같다. 오른쪽 그림은 같은 자료에 주제목, 부제목(그림의 아래쪽), x축 이름, y축 이름을 설정한 결과이다.

> plot(mtcars$hp, mtcars$mpg, pch="A", cex=2)	> plot(mtcars$hp, mtcars$mpg, pch=17, col=rgb(1,0,0))
	

위의 그림은 산점도에서 점에 사용할 문자를 설정한 것으로 왼쪽 그림은 점에 문자 A를 사용하라고 설정한 것이다. 이때 사용한 cex는 글자 크기를 설정하는 값으로 1이 기본값이며 1보다 크면 기본값보다 큰 글자를, 1보다 작으면 작게 만들어준다. 오른쪽 그림은 pch에서 숫자 17을 사용하여 17번에 해당하는 문자를 점으로 사용하였으며, col에 색을 설정하여 빨간색으로 점을 만들었다. 여기서 사용한 rgb 함수는 Red, Green, Blue의 세 개의 색의 농도를 0부터 1사이의 값으로 설정하여 임의의 색을 만드는 함수로 Red에 해당하는 값이 1로 가장 농도가 높고 다른 색은 0이라서 결과적으로 빨간색이 된다.

plot 함수는 산점도 뿐 아니라 함수도 그릴 수 있다. 함수의 형태를 그리기 위해서는 대체로 type에 기본값 "p"가 아닌 값을 설정하는 경우가 많다.

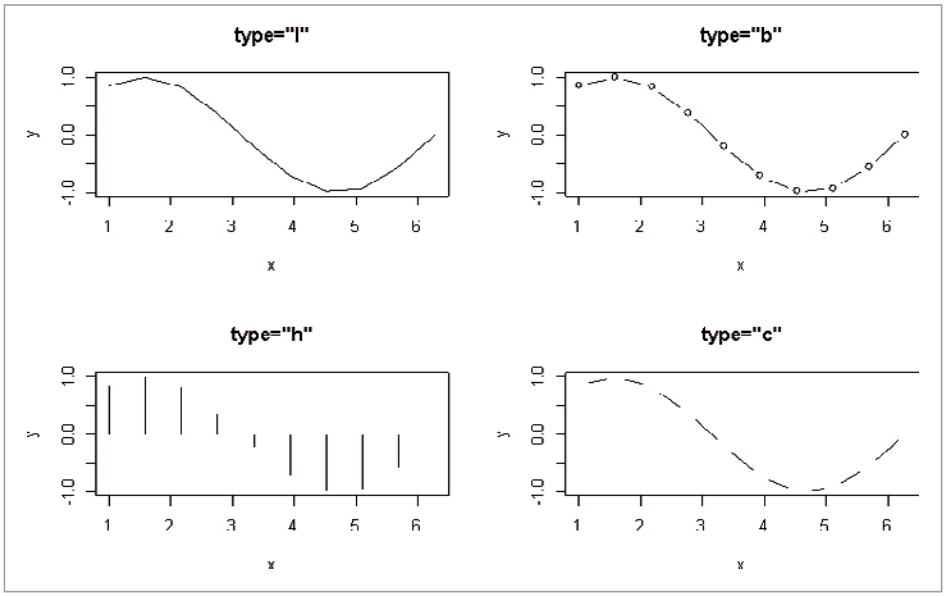
• type : 산점도의 종류를 선택한다. type에 설정 가능한 값은
  – "p" 좌표 (x_1, y_1)에 점을 그린다. 기본값이다.
  – "l" (x_1, y_1)와 (x_{i+1}, y_{i+1})을 연결하는 선을 그린다. 두 점의 연결은 직선(선분)으로 이루어지므로 점의 수가 작으면 꺾은선 그림으로 보이기도 한다.
  – "b" 점과 선을 모두 사용한다. 이때 점과 선은 연결되지 않고 점과 점 사이에 선분을 그린다.
  – "c" "b"를 사용한 경우에서 점만 뺀다.
  – "o"는 "b"와 마찬가지로 점과 선을 모두 사용하나 점과 선이 겹치게 그린다.
  – "h" 점이나 점들끼리 연결하지 않고 각 점에서 중앙을 기준으로 아래위로 수직선을 만들고자 할 때 설정한다.

다음 프로그램은 type에 따른 산점도의 모양을 알아보기 위해 $y=\sin(x)$함수를 그려본 것이다.

```
> par(mfrow=c(2,2))
> x <- seq(1, 2*pi, length=10)
> y <- sin(x)
> plot(x,y, type="l", main="type=₩"l₩"")
> plot(x,y, type="b", main="type=₩"b₩"")
```

```
> plot(x,y, type="h", main="type=₩"h₩"")
> plot(x,y, type="c", main="type=₩"c₩"")
```

이 명령의 결과는 다음과 같다.



첫 줄의 par 함수는 한 그래픽 영역을 2개의 행과 두 개의 열로 구분하여(네 개의 그림영역), 이 명령 이후의 그림은 구분된 영역에 하나씩 채우라는 명령이다. 이 명령은 한 번 명령하면 그 효과가 계속 남아 있으므로 한 영역에 그림 하나씩 그리려면

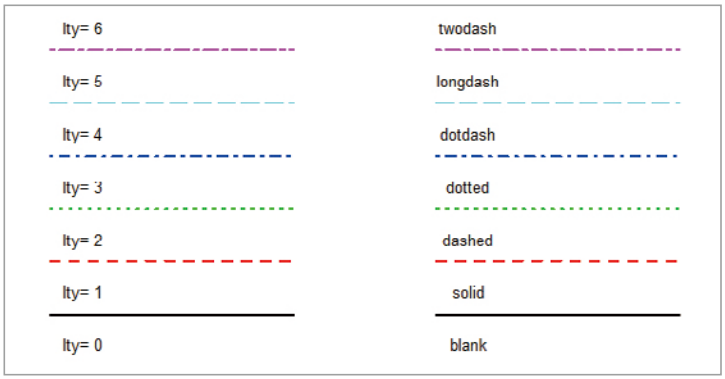
```
par(mfrow=c(1,1))
```

를 명령하여 한 개의 그림영역에 한 개의 그림을 그리도록 해주면 된다.

위의 명령들은 x를 1부터 2π까지 9등분한 경곗값 10개를 얻고, y를 x의 sin 함숫값을 얻은 후 이 (x, y)의 산점도를 네 개의 type 값에 대해 산점도를 그리는 명령이다. 첫 번째 그림은 선으로(이 경우 x의 길이를 늘이면, 즉 위의 seq 함수의 length를 100 정도로 크게 하면 좀더 부드러운 선을 그릴 수 있다), 두 번째 그림은 선과 점 모두를, 세 번째 그림은 세로로 수직선을 그렸으며, 마지막 그림은 "b"에서 점만 제외한 선을 그렸다.

- "s" 계단 그림(step)을 그린다.
- "S", "s"와 마찬가지로 계단 그림을 그리는데 (x_i, y_i)에서 (x_{i+1}, y_{i+1})로 이동할 때 "s"는 먼저 수평으로 이동한 후 수직으로 이동하고 "S"는 먼저 수직으로 이동한 후 수평으로 이동한다.
- "n" 점을 그리지 않고 그림 영역만 확보한다. 주로 그림영역을 확보한 후 추가로 그림이나 문자 등을 넣고자 할 때 사용한다.

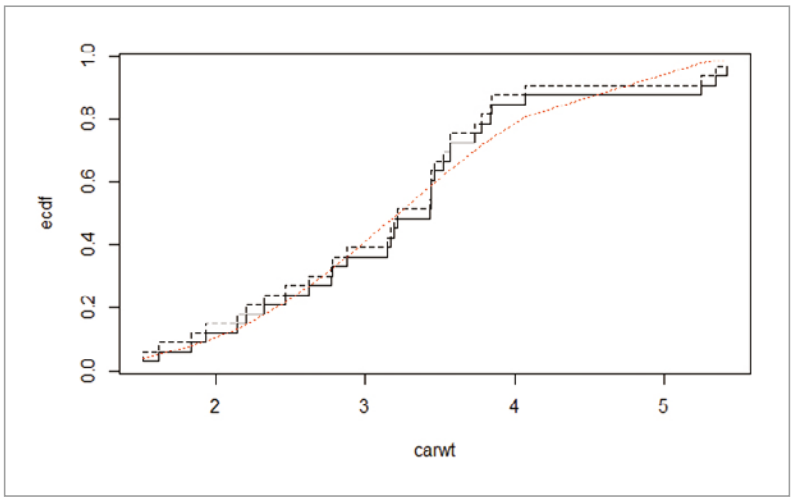
• lty : 산점도에서 선을 사용할 경우(type이 "l" 또는 "b" 등일 때) 사용할 선의 종류를 0부터 6까지 설정한다. 각 숫자에 따른 선의 모양은 다음과 같다. 0은 선이 없으며 1은 실선이고 나머지 2-6은 서로 다른 종류의 파선이다.



다음 보기는 mtcars 자료에서 자동차의 무게(wt)를 사용하여 누적확률을 계단 형태의 그림을 그리는 보기이다.

```
> nn <- dim(mtcars)[1]; ecdf <- seq(1,nn)/(nn+1)
> carwt <- sort(mtcars$wt)
> plot (carwt, ecdf, type="s")
> lines (carwt, ecdf, type="S", lty=2)
> lines(carwt, pnorm(carwt, mean(carwt), sd(carwt)), lty=3, col=2)
```

이 명령의 결과는 다음 그림과 같다.



위의 명령은 먼저 nn을 자료의 개수로 잡은 후(dim 함수는 행렬이나 데이터 프레임 등에서 행과 열의 개수를 벡터로 얻는데 첫 번째 원소는 행의 개수임) 크기순으로 했을 때 번째 자료의 누적확률밀도 함수를 $\frac{i}{n+1}$로 계산하여 이를 ecdf라고 하였다. 자동차 무게 wt도 오름차순으로 정리한 후 이 둘의 산점도를 그리되 type을 "s", "S"를 사용하여 둘의 차이를 살펴보았다. 마지막은 변수 wt의 평균과 표준편차를 갖는 정규분포의 누적확률밀도함수를 추가로 그렸다. 여기서 사용한 lines 함수는 기존 그림에 추가로 그림을 그리는 함수로 다음 절에서 자세히 소개한다.

2 abline 함수, lines 함수, points 함수를 사용한 겹쳐 그리기

abline, lines, points 등의 함수는 low-level 그래픽 함수로 기존에 생성된 그래픽 장치 위에 그림을 겹쳐 그리는 함수이다. abline 함수는 직선 $y=a+bx$에서 절편 a와 기울기 b를 지정하면 해당 직선을 기존의 그래픽 위에 겹쳐 그린다. 이 함수는

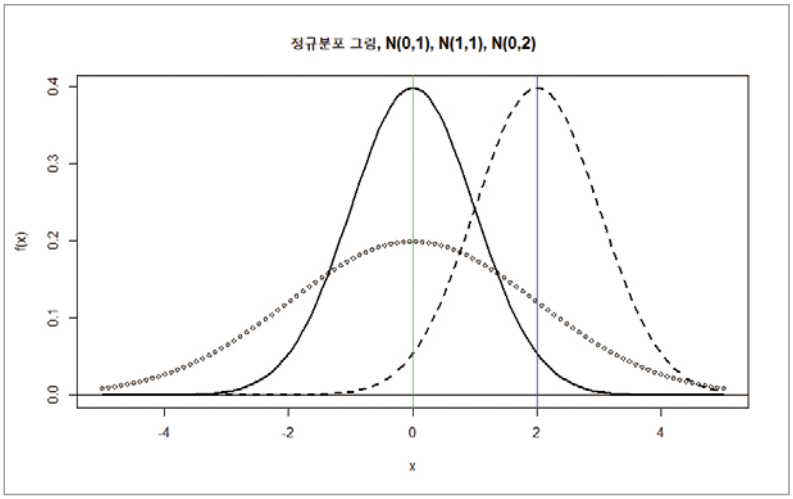
```
abline(a = NULL, b = NULL, h = NULL, v = NULL, reg = NULL,
                                    coef = NULL, untf = FALSE, ...)
```

로 사용하며 a에는 절편, b에는 기울기를 설정한다. 만일 수평이나 수직선을 그리려면 a, b 값을 설정하지 않고 수평선의 경우 h에 y-절편 값을, 수직선을 그리려면 v에 x-절편값을 설정한다. 절편과 기울기를 벡터로 설정하고자 하는 경우 coef에 두 값이 저장된 벡터를 설정한다.

lines 함수와 points 함수는 앞에서 본 plot 함수와 사용법이 같으나 main, xlab, ylab 등은 설정하지 않는다. lines 함수는 plot 함수에서 type에 "l"을, points 함수는 plot 함수의 type에 "p"를 설정한 것과 같은 결과를 얻으며 기존의 그래픽 장치에 그림을 그린다. 다음의 보기는 이들 함수의 사용방법 및 그 결과이다.

```
> x <- seq(-5,5, length=100)      # x의 범위를 -5에서 5사이의 등간격 100개의 점으로 잡음
> y1 <- dnorm(x)                  # y1: 기댓값 0, 표준편차 1인 정규분포의 함숫값
> y2 <- dnorm(x, 2, 1)            # y2: 기댓값 2, 표준편차 1인 정규분포의 함숫값
> y3 <- dnorm(x, 0, 2)            # y3: 기댓값 0, 표준편차 2인 정규분포의 함숫값
> plot(x, y1, col=1, lwd=2, type="l", main="정규분포 그림, N(0,1), N(1,1), N(0,2)",
    ylab="f(x)")                  # (x, y1)의 산점도 그림. 제목 등 추가.
> lines(x, y2, lwd=2, lty=2)      # (x, y2)의 산점도를 기존 그림에 추가
> points(x, y3, pch=5, cex=0.5)   # (x, y3)의 산점도를 기존 그림에 추가. 글자크기 축소
> abline(h=0)                     # 수평선( 절편값 0)을 그림
> abline(v=0, col=3)              # 수직선( 절편값 0)을 green 색으로 그림.
> abline(v=2, col=4)              # 수직선( 절편값 2)을 blue 색으로 그림.
```

결과는 다음 그림과 같다.



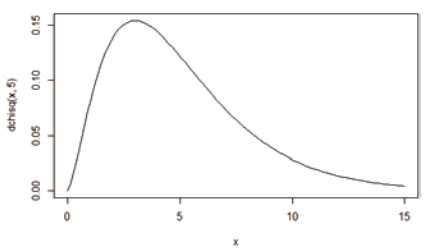
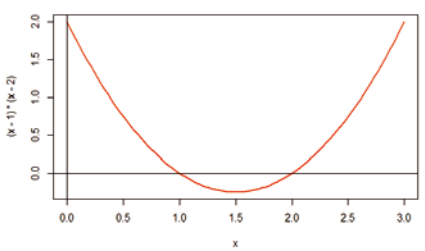
3 curve 함수를 사용한 함수 그리기

plot 함수를 사용하여 함수를 그리려면 x축과 y축에 사용할 변수를 설정하고 type에는 주로 "l"을 설정하는 것을 위에서 봤는데, 함수를 그리기 위한 좀 더 편한 방법으로 curve 함수를 사용할 수 있다. curve 함수는 그리고자 하는 함수의 식이나 함수의 이름을 설정하면 특정영역의 함수를 그려준다. 사용법은

```
curve(expr, from = NULL, to = NULL, n = 101, ...)
```

이며 expr에 사용할 함수의 이름이나 함수의 식을 설정하고 그림을 그릴 x의 범위는 from과 to에 설정한다. from과 to를 n에 주어진 값(기본값 101)의 개수만큼 등간격으로 나누어 함수를 그린다.

다음 두 가지 예에서 왼쪽은 자유도 5인 카이제곱 그림을 x의 범위 0부터 15까지 그린 것이며(from은 기본값이 0이므로 따로 설정하지 않음) 오른쪽은 함수 $f(x)=(x-1)(x-2)$의 그래프를 0부터 3 사이에 그린 후 x축과 y축을 abline 함수를 사용하여 추가하였다. col이나 lwd는 plot 함수에서와 마찬가지로 색깔과 선의 굵기를 설정한다.

> curve(dchisq(x,5), to=15)	> curve((x-1)*(x-2), from=0, to=3, col="red", lwd=2) > abline(v=0) ; abline(h=0)
	

4 그룹별 점 구분하기

R에 내장된 mtcars 자료에는 cyl이라는 변수가 있는데 이 변수는 해당 자동차 엔진의 실린더 개수를 저장하고 있다. 실린더수는 4기통, 6기통 및 8기통 중의 하나이며, 실린더수가 많으면 연료소비량이 많은 반면 마력수는 올라갈 것으로 짐작할 수 있다. 따라서 앞에서 본 마력과 연비에 대한 산점도를 얻더라도 실린더수에 따라 표시를 다르게 하는 그림이 요구될 수 있다. 이를 위해서는 실린더수가 각각 4, 6, 8일 때의 자료를 필요로 할 수 있으며 이 목적으로 subset이라는 R의 내장함수를 사용할 수 있다. subset 함수는 데이터 프레임에서 특정한 조건을 만족하는 일부의 자료만 추출하는 함수이다.

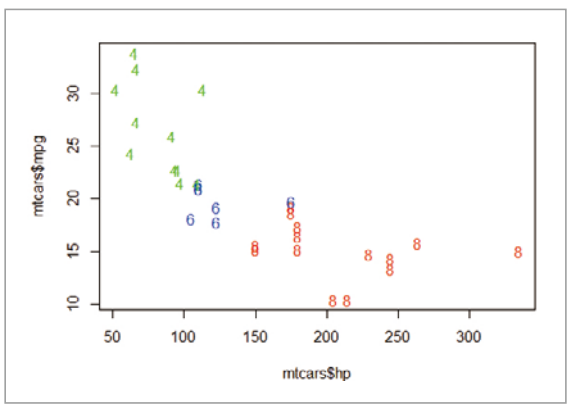
예를 들어

```
subset(mtcars, cyl==8)
```

은 데이터 프레임 mtcars에서 cyl의 값이 8인 자료만 얻게 한다. 또한 with 함수를 사용하여 subset의 결과로 얻은 데이터 프레임의 이름을 $를 사용하지 않고 바로 변수 이름을 사용할 수 있다. 즉,

```
> plot(mtcars$hp, mtcars$mpg, type="n")
> with(subset(mtcars, cyl==8), points(hp, mpg, pch="8", col="red"))
> with(subset(mtcars, cyl==6), points(hp, mpg, pch="6", col="blue"))
> with(subset(mtcars, cyl==4), points(hp, mpg, pch="4", col="green"))
```

로 첫 번째 plot 함수는 실제로는 그림을 그리지 않고(type="n") 그림 영역만 확보하며 나머지 세 줄의 명령에서 각각 실린더수가 8, 6, 4일 때의 산점도를 그리게 된다. 이 명령의 결과는



이다. 이 그림에서 예상대로 실린더수가 작은 자동차는 연비는 높지만 마력이 낮으며 실린더수가 많은 자동차는 마력은 높고 연비는 낮음을 알 수 있다.

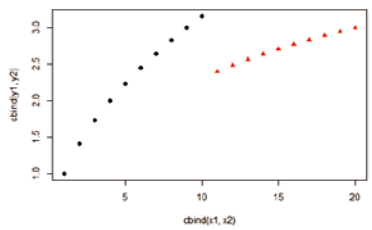
5 matplot 함수를 사용한 여러 그림 동시에 그리기

한 개 또는 여러 개의 x값에 대해서 여러 개의 y값이 있는 경우 이를 하나의 그림으로 그릴 수 있는 함수가 matplot 함수이다. 이 함수는

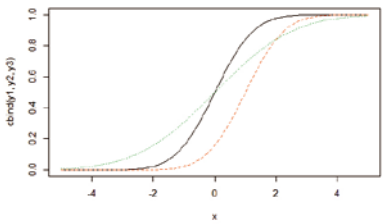
```
matplot(x, y, type = “p”, lty = 1:5, lwd = 1, pch = NULL, col = 1:6, ...)
```

로 사용하며 x와 y는 행렬, 벡터 또는 데이터 프레임으로 x와 y의 행의 개수는 같아야 한다. matplot 함수로 얻는 산점도는 한 개지만 이 산점도에 x의 첫 번째 열과 와 y 첫 번째 열, x의 두 번째 열과 와 y 두 번째 열 등의 산점도가 얻어진다. 만일 x가 벡터(열이 1개)이면 x와 y의 첫 번째 열, x와 y의 두 번째 열 등의 산점도가 얻어진다.

```
> x1 <- seq(1,10)
> x2 <- seq(11, 20)
> y1 <- sqrt(x1)
> y2 <- log(x2)
> matplot(cbind(x1,x2), cbind(y1,y2), pch=c(16,17))
```




```
> x <- seq(-5,5, length=100)
> y1 <- pnorm(x)
> y2 <- pnorm(x, 1, 1)
> y3 <- pnorm(x, 0, 2)
> matplot(x, cbind(y1,y2, y3), type="l")
```



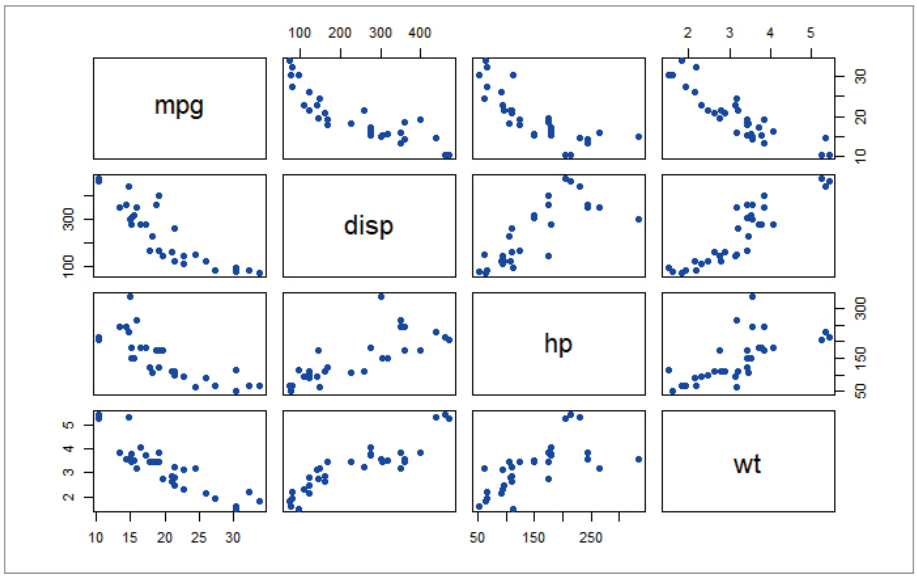
위의 왼쪽 그림은 x1은 1부터 10까지 10개의 자연수, x2는 11부터 20까지 10개의 자연수이고 y1는 $\sqrt{x1}$, y2는 log(x2)이다. 이 두 개를 각각 cbind하면 행렬이 되는데 이 행렬로 matplot 함수를 호출하여 x축의 값이 1부터 10사이에는 x1와 y1의 산점도 $y=\sqrt{x}$를 만들고, x축 값이 11부터 20 사이에는 $y=\log(x)$의 산점도를 그린 것이다. 오른쪽의 함수는 R의 내장함수인 pnorm 함수를 사용하여 정규분포 $N(0, 1)$, $N(1, 1)$, $N(0, 2^2)$의 누적확률을 순서대로 y1, y2, y3에 저장한 후 matplot 함수를 사용하였다. 이때 x에는 벡터인 x만 사용하였으므로 x와 y1, x와 y2 및 x와 y3의 산점도(이 경우 type="l"이므로 선이 그려짐)를 얻었다. 색을 따로 지정하지 않아도 기본값인 색번호 1, 2, 3이 자동으로 적용되어 검정, 빨강 및 녹색이 얻어진다.

참고로 plot 함수에서 추가로 점이나 선을 그리기 위한 함수로 lines와 points 함수가 있었는데 matplot 함수에도 lines와 같은 기능의 matlines 함수와 points 함수에 대응하는 matpoints 함수가 있다.

6 모든 가능한 산점도 그리기와 pairs 함수

데이터 프레임에는 일반적으로 여러 개의 변수가 포함되어 있고, 자료를 회귀분석 등의 방법으로 구체적인 분석하기 전에 모든 가능한 산점도를 보고자 하는 경우가 있다. 예를 들어 데이터 프레임에 p개의 변수 x1, x2, xp가 있다면 x1과 x2, x1, x3, x1과 x4, x1과 xp 등 $p\times(p-1)/2$ 개의 산점도를 확인하고자 하는 경우이다. 이 경우 각각의 산점도를 그리기 위해 $p\times(p-1)/2$번 plot 함수를 사용하지 않고 데이터 프레임 내의 모든 변수들에 대한 가능한 산점도를 얻는 방법으로 pairs 함수를 사용할 수 있다. 앞의 mtcars에서 1, 3, 4, 6번째 변수들의 모든 가능한 산점도는 다음과 같이 얻을 수 있다.

```
> pairs(mtcars[,c(1,3,4,6)], pch=16, col="blue")
```



이 경우 pairs(mtcars) 명령으로 모든 변수에 대한 산점도를 얻을 수 있으나, 변수가 조금 많은 편이라 연속인 일부 변수(1, 3, 4, 6번째 열인 mpg, disp, hp, wt)만 선택하여 산점도를 얻었다. R의 최근 버전은 plot 함수의 매개변수가 데이터 프레임인 경우 위의 pairs 함수와 같은 결과를 얻는다. 즉,

```
plot(mtcars[,c(1,3,4,6)])
```

으로도 위의 결과를 얻을 수 있다. 이 산점도를 보면 배기량(disp)이 높을수록 연비(mpg)는 낮아지고 마력(hp), 무게(wt)는 높아지는 관련성을 시사한다.

참고 |

plot 및 이와 관련된 그림 함수는 ggplot2 패키지를 사용하면 편리하면서도 더 나은 그림을 얻는 경우도 종종 있어 ggplot2 패키지를 많이 사용한다. 하지만 그래픽 매개변수를 직접 변경하면서 그림을 그릴 경우도 많아 R의 기본 그래픽 함수의 중요성은 여전하다.

이수동 ◆ 한국인포그래픽협회 대표(samyicamper@gmail.com)

데이터 인포그래픽 강좌 series 10

실전에서 자주 사용하는 「정책연구 데이터」 시각화 방법

같은 통계 자료라도 편집 방향에 따라 해석과 의미가 크게 달라지는 경우가 있다. 예를 들어, 시간 순으로 자료를 나열하는 경우, 데이터 비율 순으로 나열하는 경우, 데이터 간 차이를 구한 후 큰 순서대로 부각하는 경우 등 메시지 목적에 따라 편집 방법이 변경된다. 정부자료는 대부분 국민에게 공개하는 것이 원칙이므로 메시지 목적을 먼저 수립하고 자료를 가공하는 것이 중요하다. 이에 동일 데이터도 메시지를 다르게 하여 표현하는 방법을 알아보자.

1994년과 2008년의 두 자료를 비교한 데이터를 소개한다. 2008년과 1994년 사이에는 굉장히 긴 시간 차가 있다. 이처럼 전년대비, 5년간 대비가 아닌 14년이란 긴 시간을 비교한다는 것은, 사회의 통념이 그만큼 크게 변화했다는 것을 강조하려는 의도가 반영됐다는 뜻이기도 하다.

해당 데이터는 시간 순으로 데이터 항목을 단순히 나열하는 방식도 있지만 차이 값을 계산하여 큰 데이터를 순서대로 배열하는 방법도 고려해볼 수 있다. 데이터 결과는 하나처럼 보이나 데이터분석가는 데이터 재편집을 통해 강조하고자 하는 메시지를 다르게 전달할 수 있다. 이 부분을 대부분 간과하고 통계를 분석하는 컴퓨터 프로그램을 배우는 데만 치중하는 것이 안타깝다. 인사이트, 통찰력을 얻기 위해서는 데이터를 바라보는 관점을 몸으로 자연스럽게 체화하는 훈련이 필요하다.

Infographics

❶ 최근 연도, 가구형태 비율이 큰 순서를 도해로 나타내는 경우

한국보건사회연구원에 의하면 우리나라 노인의 가구 형태가 빠르게 변화하는 것으로 조사됐다. 1994년 자녀동거 가구가 54.7%를 차지해 가장 높은 비율을 차지했다. 다음으로는 노인부부 가구 26.8%, 노인독신 가구 13.6%, 기타 4.9% 순으로 나타났다. 반면 2008년은 노인부부 가구가 47.1%로 가장 높았으며, 자녀동거가 27.6%, 노인독신 가구 19.7%, 기타 5.6% 순으로 나타났다.

데이터를 독해한 후 바로 계산하는 것보다 먼저 간단한 표로 요약한 후 '데이터 표현 방법'을 결정하는 것이 좋다. 표는 일반적으로 가로에 많은 항목을 배열하는 것보다 세로, 즉 1열에 많은 항목을 넣어 세로표로 자료를 모아 보는 것이 보기에 편하다.(*모바일 가독성에 좋음) 표는 표제목, 강조 숫자 등을 컬러로 처리하거나 글자 크기, 제목 옆 아이콘으로 표현방법을 선택할 수 있다.

가구형태/연도	2008년	1994년
노인부부	47.1	26.8
자녀동거	27.6	54.7
노인독신	19.7	13.6
기타	5.6	4.9
합계	100.0	100.0

노인 가구 형태 변화 자료

1994년과 비교해 2008년에는 노인들이 어떤 가구 형태를 가장 많이 선호하는지가 중요하다.
이 경우 시간 순으로 배열하기보다는 2008년이란 최근 연도를 바로 옆 열에 넣어 (기준) 가구 형태 비율을 순서대로 보여주는 것이 필요하다.

▶ 도해로 시각화하는 과정

표를 보는 사람에게 좀 더 쉽게 이해시키려면 도해를 사용하는 방법도 선택할 수 있다. 이때 메시지를 전달하려는 사람이 보는 사람에게 자신과 같은 방식으로 메시지를 해독하도록 유도하는 것이 좋다.

여기서는 표에 있는 구분선은(세로선) 최대한 없애고, 2008년과 1994년을 쉽게 비교할 수 있도록 했다. 14년 만에 노인부부, 자녀동거 등 두 가지 항목 변화가 가장 중요하므로 해당 항목만 다른 색으로 처리하였다. 이 밖에 가장 큰 비율을 차지하는 데이터는 가장 위에 표시하였으며, 화살표로 아래쪽으로 내려가면서 보도록 시선 처리하였다. 제목 역시 강조하고자 하는 내용을 구체적으로 적어주는 것이 좋다. 표를 도해로 처리하는 연습은 매우 중요하며 도해는 평소 손으로 자주 만들어보는 습관을 갖는 것이 필요하다.

2008년 노인부부 가구 가장 높은 비율 차지

단위: 비율(%)

가구형태	2008	1994
노인부부	47.1	26.8
자녀동거	27.6	54.7
독신노인	19.7	13.6
기타순	5.6	4.9

(출처 : 한국보건사회연구원)

❷ 변화율 흐름과 차이를 그래프로 나타내는 경우

변화율을 나타내는 경우 일단 시간 순으로 자료를 나열하는 것이 중요하다. 특히 변화율에서 큰 차이를 나타냄을 강조할 때는 '100% 2중 막대그래프' 또는 '선 그래프'로 표현하는 것을 우선 생각해볼 수 있다. 두 개의 그래프는 연도 차이를 동시에 살펴볼 수 있다는 장점이 있다. 다만 수직으로 그릴 때와 수평으로 그릴 때 연도순서는 다르다. 특히 수평으로 표현하는 경우 최근 연도, 즉 2008년은 맨 위에 그리고 강조해야 하는 막대는 명도를 낮게 처리한다.

▶ 100% 이중 막대그래프로 표현하는 경우

수평 막대그래프로 표현 시 가장 상단에 높은 비율의 데이터가 위치하도록 하며, 최근 연도가 중요하므로 2008년 막대그래프를 위에 올려놓는다. 제목은 최상위 항목의 차이를 계산한 내용을 언급하고, 2008년 그래프 컬러는 명도를 낮게 하여 주목도를 높일 필요가 있다.

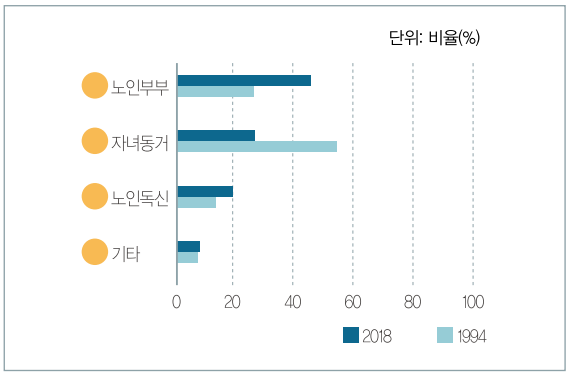


그림 1 2008년 노인부부 가구비율 가장 높아, 자녀동거 가구는 가장 큰 폭 감소

▶ 선(라인) 그래프로 표현하는 경우

선그래프의 장점은 연속해서 두 개의 선 변화를 한눈에 살펴볼 수 있다는 점이다. 변화, 흐름을 보겠다고 만든 이의 목적이 주라면 선 그래프로 나타내는 것도 고려해본다. 선 그래프는 무엇보다 변곡점 부분의 숫자가 중요하다. 선 그래프는 크게 3가지 의미를 갖는다. 우선 두 선의 간격이 가장 큰 부분, 가장 작은 부분, 마지막으로 교차하는 부분에 어떤 중요한 의미가 내포돼 있다.

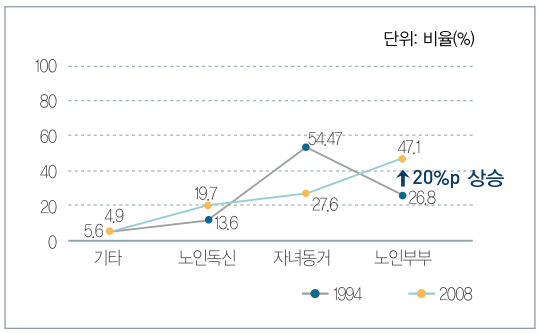


그림 2 2008년 노인부부 가구비율 가장 높아, 자녀동거 가구는 가장 큰 폭 감소

키노트(Key Note)

1. 표는 열과 행으로 정보를 분류하며, 1열에 가급적 많은 정보를 모아서 세로로 표시하면 해독이 편하다.
2. 표 디자인에서는 항목 영역 컬러, 글자/숫자의 굵기 또는 크기, 음영 등으로 강조한다.
3. 표는 가급적 도해 형태로 전환하여 메시지 흐름을 제작자가 의도하는 방법으로 볼 수 있도록 처리한다.
4. 선 그래프는 간격의 차이를 한눈에 보는 데 적합하다. 변곡점에 숫자표시, 강조할 선을 명도를 낮게 표시한다.
5. 100% 누적 수평 막대그래프에서는 최근 연도를 위에 표시한다. 큰 데이터부터 아래 방향으로 순으로 배열한다.

데이터 분석 및 편집 과정을 거친 후 도해(圖解) 처리 과정까지 끝냈다면 정책연구 통계 데이터를 다양한 기준으로 시각화하는 실전 제작에 들어가야 한다. 최종 결과물 사례는 다음과 같다.

정책연구 자료를 기획 파트에서 다양한 방법으로 그래프를 사용했듯이 디자인의 방법도 다양하다. 자료를 어떻게 해석할지에 따라 표현을 다르게 할 수 있다. 정확한 정보 전달이 가능하다면 정해진 방식은 없다. 시각적인 도해 자료, 사용할 컬러의 의미와 표현 등을 고려하여 보다 전달력을 높일 수 있도록 디자인한다.

인포그래픽 프리뷰 1

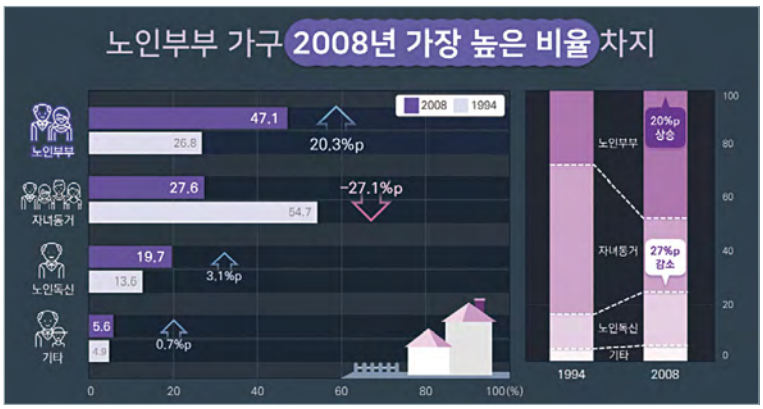


그림 3 (출처: 한국보건사회연구원)

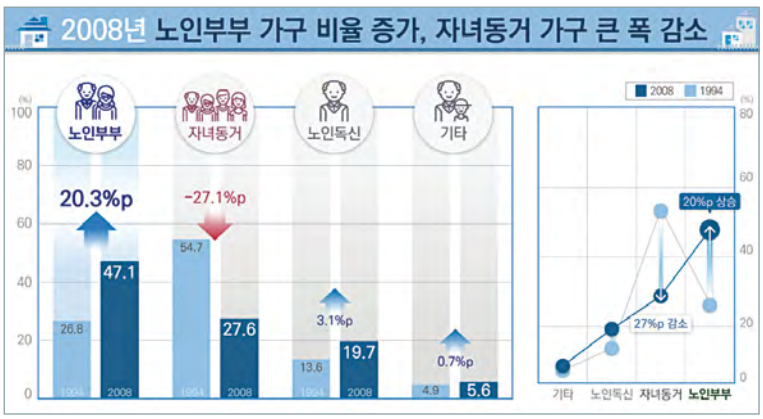


그림 4 (출처: 한국보건사회연구원)

디자인 P·O·I·N·T 다양한 비율 비교 통계 자료 시각화

① 전달하고자 하는 의도에 따라 큰 디자인의 변화 없이도 색을 다르게 사용하여 전체적인 분위기 전환이 가능하다. 부정적인 내용 또는 사회 문제의식을 다루는 부분에 중점을 둔다면 어둡고 차가운 색을 활용하여 경각심을 부각시키거나 반대로 밝고 차분한 색을 선정하여 어두운 내용을 긍정적인 메시지로 전달할 수도 있다. 여기서 중요한 부분은 웹에서 선명하게 보일 수 있는 색을 사용하는 것이다.

② 비율의 증감율은 Preview 1의 오른쪽 그래프처럼 누적 막대그래프를 활용해 데이터의 크기를 쉽게 비교할 수 있도록 하거나 Preview 2의 그래프처럼 점과 선 그래프를 활용하여 집단의 기간 사이에 생략되어 있는 중간 값과 추세, 흐름 등을 알 수 있도록 표현한다.

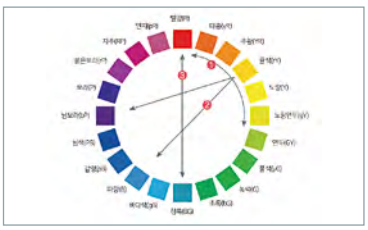
③ 강조하고자 하는 데이터의 아이콘이나 그래프에 사용된 컬러는 약간 다르게 선택하거나, 설명 없이 상징적인 이미지나 아이콘을 사용하면 전달력을 높여주고 공감을 이끌어 낼 수 있다.

컬러 T·I·P

20개 컬러로 이루어진 색상환을 알고 있으면 그래프 컬러 사용에 도움이 된다. 20개 컬러 색상환은 시계방향 순으로 무지개 색으로 구성되며, 따뜻한 색은 '난색', 차가운 색은 '한색'으로 불린다. 추후 기회가 되면 '컬러와 통계'의 관계에 대해 자세히 소개할 예정이며 우선 일반인들이 알아두면 유용한 '색상표 기초 지식' 만 간단히 소개한다.

그림 5 20개 컬러 색상

그림 6 유사대비, 반대대비, 보색대비로 나타낸 모습(출처: 데이터인포그래픽 디자인제작실무(이수동, 김선주 예문사)



① 유사대비 : 색상환에 가깝게 근접해 있는 색상. 즉, 유사 조화란 같거나 비슷한 성격을 가진 색들이 배색 되었을 때 얻어지는 조화.

② 반대대비 : 대비 조화란 서로 다른 색이나 성격이 반대되는 색들이 배색되었을 때 얻어지는 조화.

③ 보색대비 : 색상환에 반대되는 색상끼리 배색되었을 때 얻어지는 조화 색을 선정할 때는 데이터 구분이 확실히 되도록 대비가 뚜렷한 색을 선택하는 것이 좋다. 명도의 차이를 두거나 보색을 사용할 수 있다. 불가피한 경우가 아니라면 색각 이상자를 고려하여 빨간색과 녹색, 파란색과 노란색 등을 함께 쓰는 것은 피해야 한다.

※ 데이터시각화를 위해서는 특히 '보색대비'에 대한 이해도를 높이는 것이 중요하다.

정승호 ◆ 영남중학교 교사

통그라미,

통그라미로 자료 분석 해보기

클릭 한 번이면 나도 통계 전문가! ❸

통그라미의 자료 분석은 초등학생도 클릭 한 번으로 자료를 분석할 수 있는 교육용으로 개발되었기 때문에 분석기법에서는 SPSS, SAS, R과 같은 통계 프로그램 패키지보다는 제한적이다. 하지만 기초통계량, 표, 그래프를 통해 자료를 분석할 수 있으며 일상적인 통계 처리에는 큰 무리가 없다.

자료 분석창

통그라미의 자료 분석창은 [그림 1]과 같이 크게 3개로 구성되어 있다. ①의 난이도 설정 창은 사용하는 학생들의 소속급에 따라 기능을 선택할 수 있다. 고등학교 학생 이상인 경우 고급을 선택하여 사용한다. ②의 메뉴 창을 통해 분석기법을 선택할 수 있으며, ③의 바로가기 창은 어린 학생들이 사용할 수 있도록 메뉴 창의 목록을 아이콘화한 것이다. ④는 변수창은 자료를 정제한 후에는 닫고 사용하는 경우가 많다.

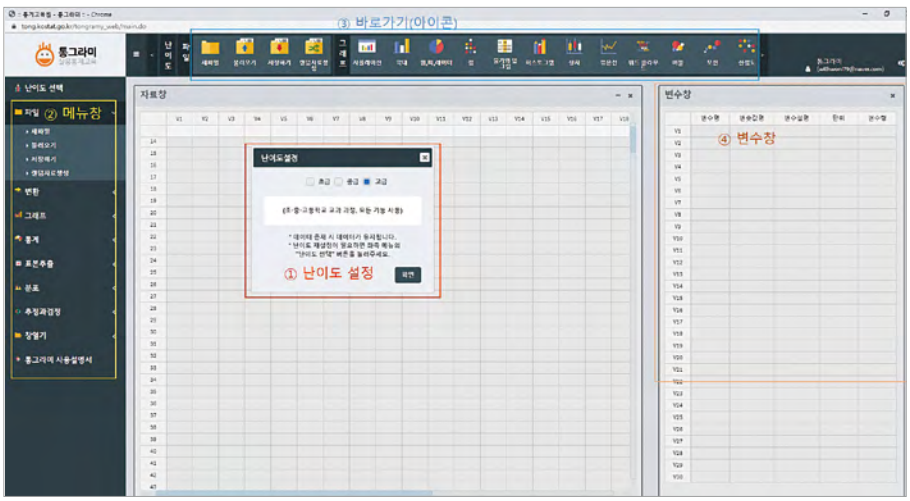


그림 1 | 통그라미의 자료분석 창

통그라미를 이용한 자료분석 1 기초통계량

기초통계량은 [그림 2]의 과정을 통해 구할 수 있다.

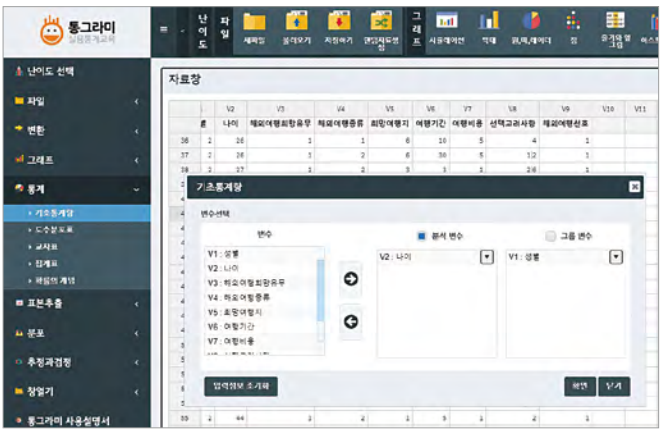


그림 2 | 기초통계량 구하기

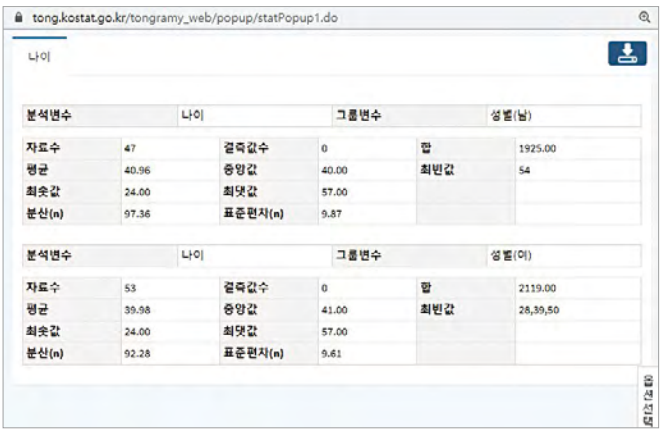


그림 3 | 기초통계량 결과

① 메뉴창의 '통계-기초통계량'을 선택한다.

② '분석변수'와 '그룹변수'를 설정한 후 '확인'을 선택한다.

만약, 성별에 따른 나이의 기초통계량을 구하고자 할 경우 나이는 '분석변수'로, 성별은 '그룹변수'로 설정한다. 이 변수 설정 방법은 통그라미의 전 메뉴에서 동일하다. 변수 설정 시 해당 변수를 선택한 후 화살표를 이용하여 이동할 수도 있지만, 해당 변수를 선택 후 드래그하여 이동할 수도 있다.

③ [그림 3]의 기초통계량을 확인하고 '옵션선택'을 통해 범위, 표본의 분산과 표준편차, 사분위수를 추가로 구할 수 있으며 소수점표시를 통해 통계량의 자리수를 통일할 수 있다.

통그라미를 이용한 자료분석 2 표(예) : 도수분포표)

통그라미에서는 메뉴창의 '통계'에서 '도수분포표, 교차표, 집계표'를 선택하여 자료를 표로 정리할 수 있다.

도수분포표는 [그림 4]의 과정을 통해 구할 수 있다.

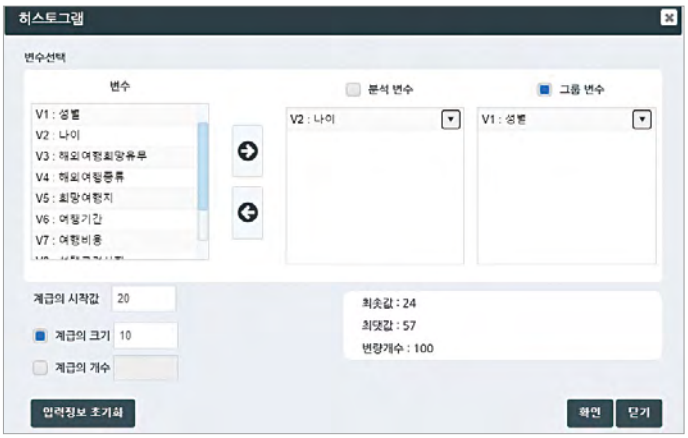


그림 4 | 도수분포표 만들기

① 메뉴창의 '통계–도수분포표'를 선택한다.

② 최솟값, 최댓값, 변량개수를 확인하고 '계급의 시작값'과 '계급의 크기'를 결정한다. 계급의 개수를 통해서도 계급을 결정할 수 있다.

③ 그룹에 따라 비교할 경우, 그룹변수를 설정한다.

④ '확인'을 선택한다.

⑤ [그림 5]와 같이 옵션선택을 통해 '계급값', '계급값 × 도수', '상대도수' 등을 나타낼 수 있다. 이는 중학교 1학년 과정의 수업에 최적화되어 있다.

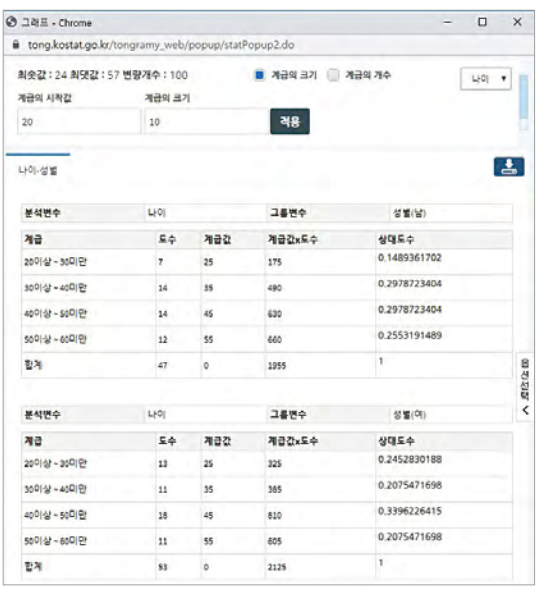


그림 5 | 도수분포표 결과

통그라미를 이용한 자료분석 3 그래프(예 : 막대그래프)

통그라미는 막대그래프, 비율그래프(원, 띠, 레이더), 점그래프, 줄기와 잎 그림, 히스토그램, 상자그림, 꺾은선그래프, 산점도와 같은 기존의 그래프와 워드클라우드, 모션(세 변수를 가로와 세로, 버블의 크기로 표시)과 같은 빅데이터에 사용하는 그래프를 제공하고 있다.

도수분포표는 [그림 6]의 과정을 통해 구할 수 있다.

그림 6 | 막대그래프 만들기

그림 7 | 막대그래프 결과

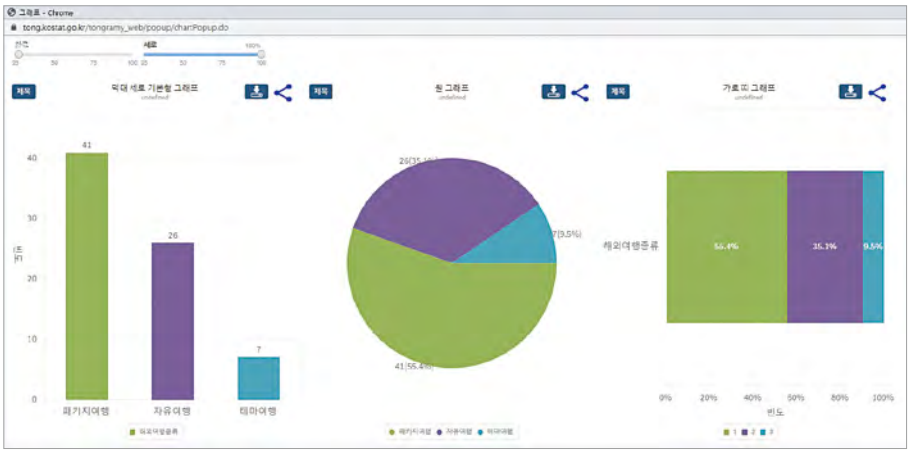


그림 8 | 그래프 추가를 통한 적절한 그래프의 선택

① 메뉴창의 '그래프–막대그래프'를 선택한다.

② 분석변수를 설정하고 '확인'을 선택한다.

③ [그림 7]과 같이 옵션선택을 통해 범례, 빈도 등을 표시한다.

④ 옵션선택에서 그래프 추가를 통해 [그림 8]과 같이 여러 가지 그래프를 동시에 나타낼 수 있다. 이는 자료의 특성에 적합한 그래프를 선택할 수 있도록 하기 위해서이다. 통그라미에서 분석 결과를 저장하고 싶을 때는 그래프나 표 위에 있는 저장하기 아이콘을 이용하여 결과를 저장할 수 있다.

통그라미 이용한 보고서 만들기

통계보고서는 결과를 분석한 후에 작성할 수 있으므로 분석 결과를 미리 그림으로 저장해두어야 한다.

통그라미에서의 보고서는 다음과 같은 과정을 통해 작성할 수 있다.

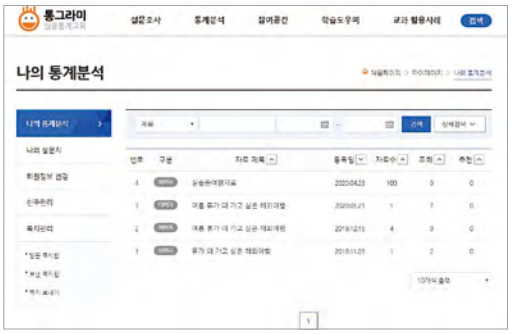


그림 9–1 | 나의 통계분석에서 데이터 선택

그림 9–2 | 나의 통계분석에서 보고서 작성


통그라미에서 보기
통계보고서 보기
데이터 상세보기
목록
제목
실습용여행자료
등록일
2020.04.28
등록자
통그리미
▪ 템플릿 배경 *
▪ 통계보고서 제목 *
데이터산업시장의 전망
▪ 조사동기 및 목적 *
앞으로 4차 산업혁명이 시작되고 기장 전망이 좋다고 인식되는 데이디 시장에 관심이 생겨서, 관련 조사를 하게되었습니다. 이 조사를 통해서 앞으로의 정확한 전망을 알고 싶습니다.
▪ 조사방법 *

그림 9-3 | 보고서 작성

① 메인화면의 우측 상단의 '마이페이지'를 클릭한 후 [그림 9-1]과 같이 '나의 통계분석'이 나타나면 보고서로 작성하고 싶은 데이터를 선택한다.

② [그림 9-2]의 화면이 나타나면 '보고서 작성'을 클릭한다. 이 화면에서 보고서 배경, 보고서 항목 생략은 보고서 작성 중에도 할 수 있기 때문에 미리 선택하기 보다는 보고서를 작성하는 과정에서 생략하는 것이 더 효과적이다.

③ 템플릿 배경을 선택한다.

④ 통계보고서 제목, 조사동기 및 목적을 입력한다.

⑤ 조사방법을 확인한다.

⑥ 조사결과에서 항목별로 응답분포를 '막대그래프', '원그래프', '꺾은선그래프'중 하나를 선택하여 나타내고 검토의견란에 분석 결과를 적는다.

⑦ 조사자료분석은 저장된 결과를 가져와 해석을 적을 때 사용하며 제외할 수 있다.

⑧ 종합결론 및 느낀점, 참고문헌을 적는다. 불필요시 제외할 수 있다.

⑨ 작성된 보고서는 [그림 9-2]의 화면이 '보고서 내려받기'를 통해 다운받을 수 있다.

기타 기능

통그라미는 처음에 초중고등학생을 대상으로 개발되었기 때문에 [그림 10]의 시뮬레이션 기능과 같은 통계를 이해하는 데 도움이 되는 기능들이 있다.

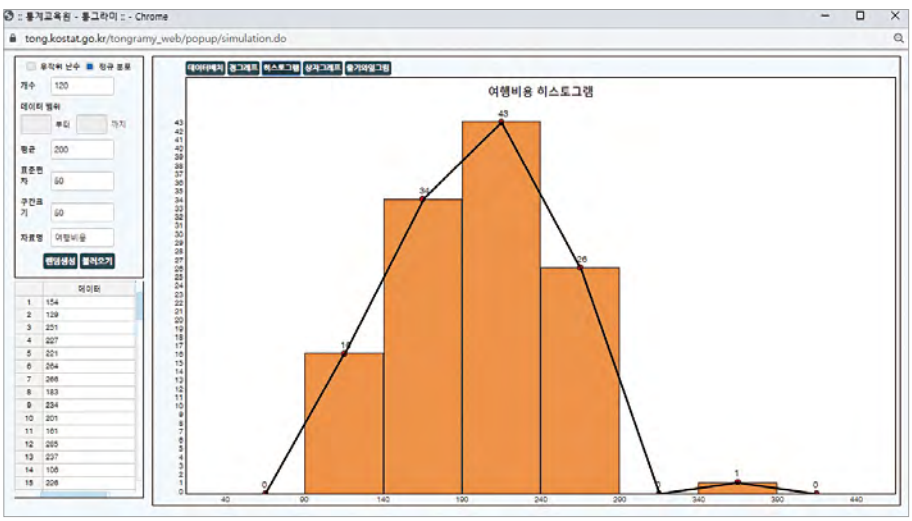


그림 10 | 그래프의 시뮬레이션 기능

시뮬레이션 기능은 '무작위난수' 또는 '정규분포'를 통해 데이터를 생성한 후 이 자료를 이용하여 점그래프, 히스토그램, 상자그래프, 줄기와 잎 그림으로 변화는 과정을 보여준다. 이를 통해 학생들은 자료의 수집과 정리, 그래프의 변환 관계 등을 이해할 수 있다.

또, 통그라미에서 제공하는 표본추출, 분포, 추정과 검정 기능을 통해 복원추출 및 비복원 추출, 이항분포의 정규분포 근사, 모평균의 추정과 표본검정을 할 수 있다.

마무리하며

통그라미를 개발하던 초기만 하더라도 통그라미가 교육 현장에서 잠시 나타났다 사라지리라고 생각하는 사람들이 대부분이었다. 하지만 사용의 편리함과 교육과정에 맞는 메뉴 및 배치 등으로 인해 통그라미를 사용하는 교사와 학생의 수가 증가하고 있다. 특히 중학교 수학교사 중 80% 이상이 통그라미를 접해보았다는 교육부 조사 자료도 있었다.

이와 같은 흐름으로 인해 처음 통그라미 사용법을 연재할 때보다 지금이 많은 기능을 제공하고 있으며 현재도 새로운 기능이 추가되고 있다. 통그라미가 우리나라뿐만 아니라 다른 나라 학생들도 사용하는 프로그램이 되길 기대한다.

코로나 이후, 데이터가 바꾸는 세상

데이터 이야기꾼
신 현 호

한동안 빅데이터라는 말의 홍수 속에서 살아온 듯합니다. SF영화를 보듯이 신비롭고 낯선 무언가가 곧 다가올 듯하지만, 현실에서 이런 변화를 느끼는 것은 쉽지 않습니다. 현실과 다른 특별한 기대감이 더 멀게 만든 것일 수 있다는 생각도 듭니다.

「나는 감이 아니라 데이터로 말한다」라는 책은 2019년에 출판되었습니다. 그 당시 이 책을 접하면서 든 생각은 이 하나였습니다. '이런 주제의 데이터를 일부러 모은 것일까 아니면 이런 쪽 일에 종사하는 분일까.'

다루는 주제들이 연구적이거나 학문적이지 않고 오히려 살아가면서 한번쯤 던져봄직한 질문들이기 때문입니다. 대부분 이런 질문들에 대한 답은 자신의 경험이나 가치관에서 얻은 직감으로 쉽게 결론을 내어버리고 곧 잊어버립니다. 하지만 이 책에서는 관련된 데이터 통해 진실에 한걸음 더 다가서는 모습을 보여주고 있습니다.

작년에 인터뷰를 하려고 준비했었지만 갑작스런 사정으로 저자분과의 인터뷰가 이루어지지 못했습니다. 그렇게 일 년의 시간이 지났지만 이분을 만나고 싶다는 생각은 여전히 내 머리 속을 떠나지 않았습니다. 지금 서점을 둘러보아도 우리 삶에 대한 질문들을 던지고 이와 관련된 데이터를 통해 접근하는 책을 아직 만나지 못했기 때문입니다.

언론사 홈페이지에서 기사를 본다면
언론사들은 경쟁적으로 좋은 정보를 발굴하고 이것을 효과적으로
전달할 인포그래픽와 같은 서비스가 발달할 것으로 여겨집니다.

● **작년에 내신 책에서 다룬 주제가 우리 생활에서 흔히 접하는 질문들이 많았습니다. 이런 주제를 연구하는 분야에서 일을 하고 계시는 건가요.**

일부는 일하면서 얻은 자료도 있지만, 제 취미가 이런 것을 찾아보고 고민하는 것입니다. 취미 생활의 결과들을 책으로 엮은 것이지요. 남들 보기에는 괴상한 취미일지 모르지만 데이터를 통해 세상을 바라보는 것이 저는 너무 재미있습니다. 특히 그래프는 어떤 정보보다 많은 인사이트를 포함하고 있습니다. 그래서 '한 장의 그림은 천 개의 단어만큼 가치가 있다'라는 말도 있습니다. 그래프에서 정보를 찾는 작업이 저에게는 마치 보물 지도를 보는 것처럼 흥분되고 짜릿한 일입니다.

● **이렇게 데이터를 통해 들으니 막연한 주장에 비해 이해하기가 더 쉬운 측면이 많았습니다. 우리나라의 데이터 리터러시 수준은 어느 정도라고 생각하시는지요.**

일반 국민들이 정보를 다루고 판단하는 수준은 다른 선진국에 비해서도 높다고 생각합니다. 오히려 지식인이나 언론의 수준이 상대적으로 낮다고 생각해요. 데이터에서 정보를 얻고 이것을 독자들에게 전달하는 능력이나 관심은 선진국에 비해 떨어진다고 생각됩니다. 그리고 포털에서 이런 기사들을 접하는 방식도 문제라고 생각합니다. 언론사 홈페이지에서 기사를 본다면 언론사들은 경쟁적으로 좋은 정보를 발굴하고 이것을 효과적으로 전달할 인포그래픽와 같은 서비스가 발달할 것으로 여겨집니다.

● **감이 아니라 데이터로 말해야한다고 말씀하셨는데, 사실 많은 사람들이 데이터로 말하는 것이 먹힐까 하는 의문을 품고 있는 듯합니다. 아직 우리나라에서는 목소리 크고 힘 센 사람 말이 통한다는 생각이 많은 것도 같습니다.**

예전에는 지도자에 있는 사람이 다른 사람에 비해 압도적으로 많은 정보나 경험을 가질 수 있었습니다. 하지만 지금은 많은 사람들이 대학을 나오고 인터넷을 통해 수많은 정보를 접할 수 있습니다. 그리고 세상이 너무 빨리 변해 어제의 지식이 오늘 쓸모없게 되는 경우도 많습니다. 이런 상황에 빠르고 효율적인 의사결정을 내리기 위해서는 데이터에 의존할 수밖에 없습니다. 실제 데이터 자료가 없는 문서의 경우 인정받기가 어렵습니다.

● **이번 선거에서도 많은 가짜 뉴스가 범람했습니다. 정보의 홍수가 오히려 사람들의 판단을 흐리게 만드는 요인이 되고 있는 건 아닐까요?**

가짜 뉴스는 진위 여부를 판단하기에는 오랜 시간이 걸리는 반면에 그럴듯한 논리로 포장되어 있고 흥미로운 내용이라 빨리 퍼집니다. 심지어 팩트 체크라는 이름을 쓰는 가짜 뉴스도 많습니다. 나중 진실이 밝혀지더라도 당사자들은 많은 상처를 입고 난 뒤죠.

제도적으로 가짜 뉴스를 방지할 방안을 만들기에 앞서 먼저 필요한 부분이 언론에 대한 신뢰 회복입니다. 언론에 대한 불신이 이런 가짜 뉴스를 양산하게 만들고 이를 믿는 사람이 늘어나는 이유가 됩니다. 영국 로이터저널리즘 연구소에서 발표한 '뉴스 신뢰도 국제비교'에 의하면 한국의 언론 신뢰도가 조사대상국 중 최하위 그룹에 속해 있습니다. 언론이 권위 있고 신뢰가 쌓이면 가짜 뉴스는 크게 힘을 못 쓸 겁니다.

● **얼마 전에 데이터 3법이 통과되었습니다. 그동안 산업계에서는 우리나라가 데이터 활용에 대한 규제 장벽을 낮추어야 한다는 말이 많았습니다. 데이터 3법이 데이터 활용이나 산업에 있어 변곡점이 될 수 있을까요?**

분명 도움이 되리라고 여겨집니다. 세계적으로 데이터가 미래 산업에 있어 원유나 쌀처럼 핵심적인 역할을 할 것으로 보고 있습니다. 우려되는 부분은 개인정보보호 문제인데, 분명 앞으로 몇 번 개인정보에 대한 문제가 불거질 것이라 여겨집니다. 하지만 모든 것을 근절할 수는 없습니다. 이것은 교통사고를 근절할 수 없는 이치하고 같습니다. 교통사고 피해를 최소화하기 위해 과속단속 카메라를 설치하고 자동차의 에어백을 강화하는 것과 같이 다양한 보완책을 만들어 나가는 과정이 필요합니다. 교통사고가 무서워 도로통행을 막아버리는 것은 누가 봐도 문제가 있는 해결법인 거죠.

● **데이터 3법 이후 통계청이 어떤 역할을 가져야 한다고 생각하시는지요.**

아마 통계청이 우리나라에서 데이터에 대한 가장 풍부한 경험이나 지식을 가지고 있다고 생각합니다. 그런데 아쉬운 점은 통계청이 세상 사람들과 커뮤니케이션이 부족해 보인다는 겁니다. 물론 지금도 적극적으로 하신다는 것은 잘 알고 있습니다. 제 개인적 의견으로는 이제 통계청이 적극적으로 정책이나 산업에 대해서 주장하고 제안에 나서야 한다고 생각합니다. 신뢰성 높은 국가통계를 생산하는 기본적인 업무에서 더 나아가 데이터 3법의 시행령 작성에도 적극적으로 참여하고, 관련 정책 개발에도 적극적으로 목소리를 내야할 때라고 봅니다. 전문성과 객관성을 지닌 통계청을 사람들은 절대 무시할 수 없어요. 적극적으로 사람을 확충하고 법률가나 전문가를 영입하여 정책 개발에 앞서 나가셔야 한다고 생각합니다.

● **사실 근래 정책에 맞추는 통계라는 논란도 있었습니다. 적극적으로 나서다보면 그런 오해를 받지 않을까 염려스럽기도 합니다.**

그동안 논란 좀 있었죠. 제가 국회에서 보니 끝난 다음에 설명이나 해명하려고 애쓰기 보다는 사전에 어떤 기준이 바뀌었고 어떤 이유에서 이런 결과가 나왔다는 것을 설명할 필요가 있습니다. 공표하고 문제가 제기된 다음에 그것을 설명하려고하니 모든 것이 변명처럼 보이는 측면도 있어요. 사전에 오해할 수 있는 부분이나 해석하는 관점에 대해서 설명해주는 것이 필요하다고 봅니다.

그리고 가장 중요한 것이 바로 통계 교육입니다. 일반인들뿐 만 아니라 국회 보좌관이나 언론인들에게도 적극적인 교육을 해야 합니다. 아직도 많은 사람이 통계청 서비스 사용 방법이나 지표가 어떻게 쓰이는지 잘 몰라요. 예산을 확보해서 적극적으로 통계교육을 확대해 나가야합니다. 저는 이런 교육이 통계청에서 가장 우선적으로 추진해야 할 일이라고 생각합니다.

● **그럼 우리나라 데이터 산업 측면에서 아쉬운 점이나 개선해 나가야 할 사항이 있다면 무엇이 있을까요?**

데이터 제공 속도입니다. 데이터 기반 의사결정이 가능하려면 아무리 좋은 데이터라도 적시에 공급되지 않으면 곤란합니다. 물론 데이터 활용과 개인정보보호가 서로 상충되는 부분이 있듯이 속도와 정확도도 상충되는 면이 있습니다. 하지만 빠른 의사결정이 필요할 경우 정확성이 다소 떨어지더라도 한 템포 빠른 정보가 효과적인 경우가 많습니다. 이런 속도를 위해 전체적인 프로세스나 시스템, 인력 확보, 인식의 변화가 필요합니다.

● **코로나19 이후 전혀 다른 세상이 될 것이라는 얘기가 많습니다. 코로나 이후 세상을 어떻게 전망하고 계시는 지요.**

제가 학생이었을 때 성적 우수상보다 더 인정해주는 상이 바로 '개근상'이었어요. 몸이 아파도 학교에 갈 수 있는 근면 성실이 가장 인정받는 덕목이었습니다. 하지만 코로나 이후 아픈데 학교에 가면 당장 제정신이냐는 소리 들을 겁니다. 바로 비대면과 온라인 중심 생활에 익숙해지는 계기가 된 것입니다. 그리고 이번 코로나 사태를 통해 발전된 IT 인프라와 질서 있는 한국 국민들의 모습이 전 세계에 알려졌습니다. 당연히 우리나라를 보는 시선이 많이 달라졌습니다.

또 필수적인 제조업 기반은 각국이 갖추는 방향으로 돌아갈 것으로 여겨집니다. 바로 사회 복원력 확보가 필요해진 것입니다. 예를 들어 선진국들이 마스크를 구하지 못해 애를 먹었습니다. 마스크가 값싼 노동력 국가에서만 생산된 결과였죠. 우리나라도 처음 마스크 문제를 겪다가 바로 안정화 되었습니다.

이런 요소들은 앞으로 우리나라에 새로운 도약의 계기가 될 것이라고 보고 있습니다. 개인적으로 코로나 이후의 삶에 대해서 현재 많은 관심을 가지고 연구를 하고 있습니다.

● **코로나가 바꾼 우리 삶의 모습에 대해서 얘기해보는 것도 재미있을 것 같습니다. 다음에 코로나에 대해서 집중적으로 얘기를 할 수 있는 기회를 만들어봐야 할 것 같습니다. 국회 일을 일단 그만두신 다고 알고 있는데 향후 계획을 듣고 싶습니다.**

현재 출판사와 다음 책에 대해서 상의를 하고 있는 중입니다. 코로나에 대해서 좀 더 연구를 하고 싶고 이후로도 계속 증거기반정책 도입을 위해 노력하고 싶어요.

신현호 저자가 걸어온 길 …

- 서울대학교 경제학과에서 학사와 석사 학위를 받았고, 서울대 경제연구소 근무를 마친 뒤, 삼정KPMG 파트너로 근무하면서 기업 경영 컨설팅을 수행했다. 2012년 이후로는 국회에서 경제정책 분석과 입안을 담당했다. 최근에는 더불어민주당 원내대표실 정책조정실장으로 근무했다.
- 경제와 데이터분석 관련해서 각종 신문에 기고해왔고, 〈나는 감이 아니라 데이터로 말한다(2019)〉를 집필했고 〈IMF, 불평등에 맞서다(2020)〉를 번역했다.

김여환 ◆ 의학박사, 가정의학과 전문의

코로나 19와의 싸움, 이렇게 이겨냈다

"미래는 예측하는 것이
아니라 함께 꿈꾸고
만들어나가는 것이다."

– 짐 데이터(미래학자, 하와이대 명예교수)

나는 대체로 겁이 없다. 아무리 가냘파보여도 일단 의사가운을 걸치면, 굵은 팔뚝을 가진 남자가 눈앞에서 술을 마시고 난동을 부려도 눈썹 하나 까딱하지 않는다. 의대 공부와 응급실 수련을 하다 보면 '바이탈 사인(활력징후. 사람이 살아 있음을 알려주는 혈압, 호흡, 체온 등의 측정치)'이 흔들리는 것을 제외하고는 호들갑 떨며 살 일이 별로 없다는 것을 터득하기 때문일 것이다. 거기에다가 나는 천 명 이상의 사망선언을 한 의사다.

그러나 이번 일은 달랐다. 31번 환자가 입원했던 한방병원이 우리 집 코앞에 있었고, 코로나 청정, 대구에서 31번을 확진하는 바람에 졸지에 밀접접촉자가 된 의사도 잘 아는 동네 사람이었다. 또 아들이 근무하는 병원의 간호사가 신천지교인인 확진자임이 밝혀졌고, 그녀와 함께 근무한 호흡기내과 전공의도 확진자가 되었다는 소식을 알게 되었을 때 솔직히 무서웠다. 대구는 의과대학이 네 곳이나 될 정도로 의료의 질이 높다. 그런 곳에서 의료진과 병실이 턱없이 부족해서 수천 명이 입원을 대기하고 있다고 하니 믿을 수가 없었다.

데이터와 통계를 바탕으로 하는 현대의학이 데이터가 없는 '코로나19'라는 신종 바이러스를 어떻게 통제해나갈지도 두려웠다. 속수무책으로 가만히 앉아 있는 것보다는 소위 '몸빵'이라도 하는 것이 낫겠다 싶었다. 그래서 시작한 의료봉사였다.

코로나와의 사투

코로나19 의료봉사를 하려면 레벨 D 보호복을 입는 것부터 숙지해야 했다. 스마트폰으로 유튜브 영상을 받았다. 봉사를 하면서 감염이 될까 두려워서 시끌벅적한 코로나 거점 사무실 복도에 쭈그리고 앉아 그 영상을 세 번이나 봤다. 원리는 간단했다. 수술복은 입을 때는 입는 것을 잘해야 하지만 보호복은 벗는 것을 잘해야 한다. 우주복같이 생긴 것을 안쪽에서 돌돌 말아 벗어놓고 곧장 샤워실로 달려가면, 코로나가 설사 조금 묻어 있더라도 몸속으로 들어올 확률은 거의 없었다. 하지만 봉사하는 한 달 내내 코로나 증후군에 시달려야만 했다. 콧물이 조금 나거나 목만 칼칼해도 '코로나가 아닐까?'라는 걱정이 되었다.

내가 봉사한 동산병원은 세련된 호텔 분위기가 나는 요즘의 대형병원하고는 거리가 멀었다. 120년이나 된 곳이라 무선인터넷이나 샤워실 등의 편의시설은 턱없이 부족했고, 병실의 페인트도 너덜너덜 벗겨지고 전등마저 희미했다. 그러나 질병을 치료할 수 있을 의료장비만은 완벽했다. 침대마다 산소를 쓸 수 있었고, 중환자실이나 CT실도 있었다. 돌이켜보면, 대구시 한복판에 세워진 낡고 허름한 골동품 같은 이 병원이 절망에 빠진 우리를 구한 것 같다.

비밀번호가 걸려 있는 2개의 자동문을 통과하면 좀비영화 촬영소 같이 썰렁한 코로나 병실이 나온다. 폐렴환자는 수시로 가슴 청진도 하고 가습기도 틀어주는 것이 정석인데 코로나19 폐렴은 그렇게 할 수가 없었다. 아직은 특별한 치료약이 없고 가족들의 면회도 당연히 금지다. 어둡고 칙칙한 병동 한 구석에서 홀로 코로나와 싸워야 한다.

마음의 면역 시스템도 작동해야 한다

그것도 하루 이틀이지 시간이 지나면 급속도로 우울해지거나 울화통이 터질 것이 분명했다. 폐쇄병동에 갇힌 코로나 환자들은 담당 의사가 3평 남짓 작은 공간에 응급으로 마련한 열악한 의료 환경에서 각자의 전공을 내팽개친 채 밤낮으로 코로나 치료에만 집중한다는 사실을 몰랐다. 내가 그들에게 해줄 수 있는 것은 '그 누구도 당신들을 버리지 않았다'는 사랑을 전달하는 일이었다. 주치의에게 환자의

상태를 알려주는 의학적 회진뿐만 아니라, 손톱깎이나 빨래비누가 없다고 하면 사다주고 기침을 많이 해서 가슴에 통증을 호소하면 스트레칭 동작을 알려줬다. 베트남 출신의 젊은 엄마가 먼저 퇴원한 여덟 살 딸아이를 걱정하고 있으면 "내가 애들 키워봐서 아는데 아이들은 엄마 없으면 더 잘 먹고 더 잘 지낸다"라고 안심시켰다.

장숙자(66세, 가명)님은 목이 따끔거려 입원한 환자였다. 환자복 위에 목수건을 두르고 침상 식판을 책상삼아 유리창 너머로 보이는 까치를 섬세하게 그려냈다. 그림공부를 한 번도 해본 적은 없다고 했지만 선이 살아 있었다. "좀 어떠세요?"라고 물으면, "저는 선생님만 믿습니다" 하면서 보호복을 입은 내 모습도 그려주었다. 의사인 나도 이렇게 두려운데 환자는 오죽하겠는가? 하지만 갑자기 맞닥뜨린 코로나 사태를 침착하게 받아들이는 그녀의 순박한 그림 속에서 한 가닥의 희망이 보이기 시작했다.

환자가 그려준 내 모습

박분례(69세, 가명)님은 이번 일로 할아버지를 떠나보냈다. 요양병원에서 간병하다가 할아버지한테서 옮았다. 할아버지는 콧줄을 달고 산소도 6리터나 쓰고 있을 만큼 위중했지만 연명치료를 거부했다. 그래서 부부가 2인실을 썼다. 코로나 환자인 박여사가 코로나 환자인 할아버지를 돌봤다. 어느 날 할아버지 침대가 비워졌다. 박여사는 할아버지가 어제 새벽에 떠났고, 화장을 해서 안방 경대 위에 모셔놓았다고 했다. 떠나기 전에 아들에게 화상전화를 걸어 함께 임종을 지켰으니 여한도 없었다. "남편과 함께 코로나에 걸리게 된 것이 너무도 감사하다. 내가 안 걸렸으면 혼자서 쓸쓸히 임종을 맞이했을 텐데"라며 그녀는 흐느껴 울었다.

내 경험에 의하면, 정성스럽게 떠나보낸 사람은 사랑하는 사람이 떠난 후에도 잘 살아간다. 코로나 환자가 줄어 병실을 정리하면서 박여사는 다른 젊은 환자와 함께 쓰는 방으로 배정받았다. 그녀는 어디서 가져왔는지 쥐포 세 마리를 들고 있었다. 옆방에서 회진하고 있는 N95 마스

함께 고생한 의료진들

크를 한 나에게도 냄새가 솔솔 나는 것을 보면, 코로나 병동 전체에 콤콤한 쥐포 냄새 그리고 사람 살아가는 냄새가 진동했을 것이다. 코로나 19를 이긴다는 것은 몸뿐만 아니라, 마음의 면역 시스템도 작동해야 한다는 것이 분명했다.

'코로나19 바이러스'는 어떤 사람은 아무 증상 없이 지나고, 어떤 사람한테는 사망에 이르게 하는 등 천의 얼굴을 가졌다. 문제는 의사도 아직 그 정체를 잘 모른다는 것이다. 48세, 김간호사는 환자의 가래를 뽑아주다가 코로나에 감염됐다. 젊고 건강해서 쉽게 나을 것이라고 생각했지만, 폐렴이 급속도로 진행되어 중환자실에 일주일 있었다. "건강했던 내가 코로나 때문에 인공호흡기까지 사용할 줄은 꿈에도 몰랐어요"라는 그녀의 음성은 꽉 잠겨 있었다. 기관 삽관을 하다가 긁힌 상처 때문이라고 했다. 아무 병도 없었던 그녀가 위험했다면 열 살 위인 나는 더 위험할 수도 있고, 또 아닐 수도 있다.

코로나19를 이기는 핵심은 의학적 접근

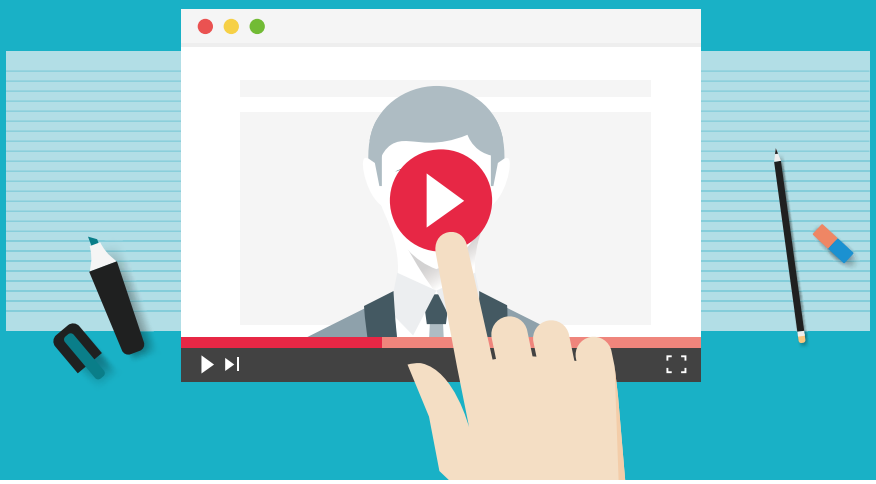
지구촌 곳곳에 퍼져있는 코로나19를 이기기 위해서는 대체 어떻게 하면 좋을까? 장수에 좋은 음식으로 토마토와 올리브유를 많이 먹는 지중해 식단이 유명하다. 하지만 이번 코로나 19로 처참하게 무너지는 이탈리아 사태를 보면 그것이 다는 아닌 것 같다. 음식으로 되는 병이 있고, 음식으로 안 되는 병이 있다. 코로나 틈새시장을 노린 식품 영양제나 근거 없는 약물에 더 이상은 현혹되지 말아야 한다.

코로나19를 이기는 핵심은 의학적 접근이다. 백신이나 치료약이 나오기까지는 불편하더라도 사회적 거리두기와 마스크 착용 등의 개인위생을 해야 한다. 코로나가 의심되면 주저하지 말고 검사를 받고 치료도 받아야 한다. '잘 넘어가겠지' 하고 방심하면 한순간에 무너진다. 의사는 치료제가 없어도 김간호사처럼 상태가 나빠지면 도와줄 수 있는 방법을 많이 알고 있다. 그러니 아프면 병원부터 가야 한다. 그 다음이 좋은 음식이고, 적당한 운동이다.

정영국 ◆ 디자이너

내 PC로 동영상 편집하기

이것만 알아도 나도 유튜버 II



"핸드폰으로 영상편집하려니까 눈 빠지겠네, 더 큰 화면으로 쉽게 영상편집할 수 없을까?"

지난 호에 연재한 "이것만 알아도 나도 유튜버" 모바일을 이용한 영상편집방법을 알려드렸습니다. 그런데 영상을 핸드폰으로 찍는 것은 쉽지만 그 작은 화면으로 편집을 하려면 좀 답답한 게 사실입니다. 필자의 경우도 정말 급한 일이 아니면 노트북이나 PC에서 작업을 하지 모바일에서 영상편집을 하게 되는 경우는 흔치않습니다. 그래서 이번에는 누구나 쉽게 PC화면에서 무료로 영상편집을 할 수 있는 방법을 준비했습니다.

초보자를 위한 PC용 영상편집 프로그램

영상 전문가들은 영상편집을 할 때 주로 큰 화면의 PC에서 작업을 합니다. PC용 영상편집 툴도 여러 가지가 있는데, 전문가들이 주로 사용하는 툴은 ADOBE사의 프리미어CC와 APPLE사의 파이널컷(FinalCut Pro)가 있습니다. 이 중 파이널컷은 전문가도 사용하지만 쉬운 사용자환경을 제공하여 영상편집 입문자에게도 많은 인기가 있습니다.

〈어도비 프리미어CC 와 파이널컷 프로〉

그러나 맥OS라는 애플사의 운영체제 환경에서만 사용할 수 있기 때문에 많은 비용이 들게 됩니다. 그래서 앞서 언급한 두 개의 영상편집 프로그램은 초보자들에게 다소 진입장벽이 높습니다. 그러면 쉽게 누구나 영상편집을 시작할 수 없을까요? 아닙니다! 있습니다!

그래서 제가 소개해드릴 영상편집 프로그램은 바로 '곰믹스'입니다. 이 프로그램은 누구나 쉽게 영상편집을 할 수 있는 사용자화면과 전문가처럼 영상을 만들 수 있는 무료 자막과 효과를 제공합니다. 물론 유료 소스를 포함하고 있지만 무료로 제공하는 소스로도 충분히 영상을 제작할 수 있습니다. 가장 먼저 내 PC에 프로그램을 설치해야겠죠? 곰랩(Gomlab.com)홈페이지에 접속하면 쉽게 '곰믹스' 프로그램을 다운로드하고 설치할 수 있습니다. (*'곰믹스 프로'는 유료이니 '곰믹스'를 다운받아 설치합니다)

〈'곰랩'홈페이지 Gomlab.com〉

곰믹스
화면구성

〈곰믹스 실행화면〉

처음 곰믹스를 실행한 화면입니다. ❶ 영상 미리보기 ❷ 기능탭 ❸ 타임라인 이렇게 크게 나누어져 있는데요. 각각의 영역을 설명하자면

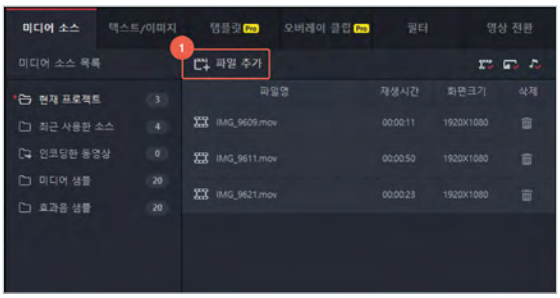
❶ 영상 미리보기 : 편집된 영상을 미리 보는 화면입니다. 사용자가 영상을 편집하고 효과를 넣은 결과화면을 미리 볼 수 있습니다.

❷ 기능탭 : 영상편집에 필요한 다양한 기능들이 있습니다. 기본기능은 무료로 제공되지만 (Pro)라고 붙은 기능은 유료입니다.

❸ 타임라인 : 영상을 불러와서 자르고 붙이고 길이를 조정하여 실제로 컷편집 작업을 작업 하는 영역입니다.

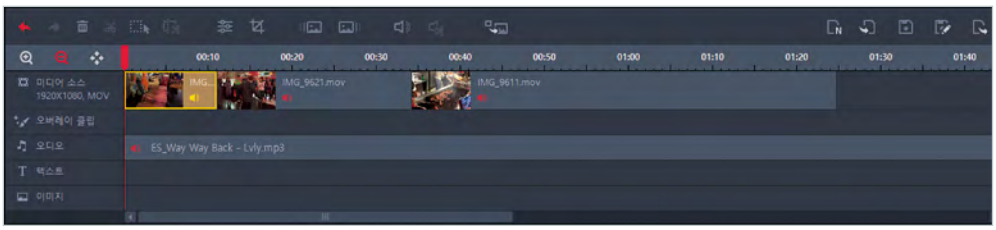
곰믹스를 이용한 영상편집

❶ 영상 불러오기



- 기능탭의 〈미디어 소스〉에서 (1)파일추가하여 편집할 영상 클립들을 불러옵니다. 또는 드래그하여 파일을 추가합니다.(이때 배경음악도 있다면 함께 불러옵니다)
- 불러온 영상클립을 〈타임라인〉위에 올려놓습니다.

❷ 타임라인에서 배경음악과 컷 편집하기



- 타임라인에 영상클립과 오디오(배경음악)를 올리고 상단의 툴을 이용해 편집을 합니다.

TIP 배경음악의 박자에 따라 컷을 잘라 장면전환을 하면 더 멋진 영상을 만들 수 있어요.

툴(tool)의 기능보기

❶ 실행취소 : 적용 및 실행한 명령을 취소합니다.

❷ 실행취소 되돌리기 : 취소한 명령을 다시 되돌립니다.

❸ 선택한 컷 삭제하기 : 선택된 영상클립을 삭제합니다.

❹ 컷 자르기 : 영상클립의 컷을 잘라냅니다.

❺ 영역 선택하기 : 영상클립 영역을 선택한 후 영역 제거, 선택 영역만 유지 혹은 분할할 수 있습니다.
❻ 선택 영역 제거 : 영상클립의 선택 영역만 제거하거나 분할 또는 유지합니다.
❼ 비디오 조정 : (반전/회전/배속) *유료포함
❽ 화면 크롭 : 선택한 영역만큼의 화면을 남겨두고 잘라냅니다.
❾ 영상 페이드 인 : 영상이 점점 나타납니다.
❿ 영상 페이드 아웃 : 영상이 점점 사라집니다.
⓫ 음량조절 : 오디오의 음량을 조절합니다.
⓬ 오디오 편집 : 오디오 편집기로 이동하여 오디오를 편집합니다.
⓭ 영상전환 : 화면전환(트랜지션) 효과를 적용합니다.

영상전환(트랜지션)

- 영상전환을 원하는 영상클립을 선택합니다.(영상전환은 클립의 앞부분에 생성됩니다)
- 기능탭의 〈영상전환〉에서 트랜지션 효과를 선택하여 적용할 수 있습니다.

TIP 1 너무 많은 종류의 트랜지션을 과하게 사용하는 것보다, 1–2가지 종류로 강조할 부분 또는 시간의 흐름을 표현하고 싶은 부분 등에 사용하는 것이 좋습니다.

TIP 2 조금 더 색다른 영상전환을 원한다면 〈오버레이클립〉을 이용해보세요! 하지만 유료기능을 포함하고 있으니 신중히 선택하세요.

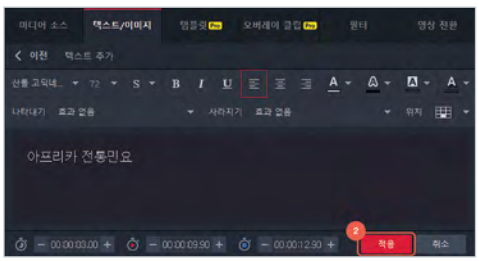
필터(컬러그레이딩)

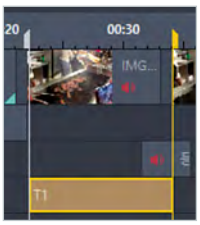
- 필터를 입힐 영상클립을 선택합니다.
- 기능탭의 〈필터〉에서 원하는 색감의 필터를 적용합니다.

TIP 1 다양한 종류의 필터가 있으니 하나씩 선택하면서 영상과 어울리는 색감을 찾아보세요.

TIP 2 나만의 영상 색감을 만들고 싶다면 '색상보정'의 메뉴에서 색상/채도/명도 등을 직접 조절하여 적용할 수 있어요.

자막 넣기

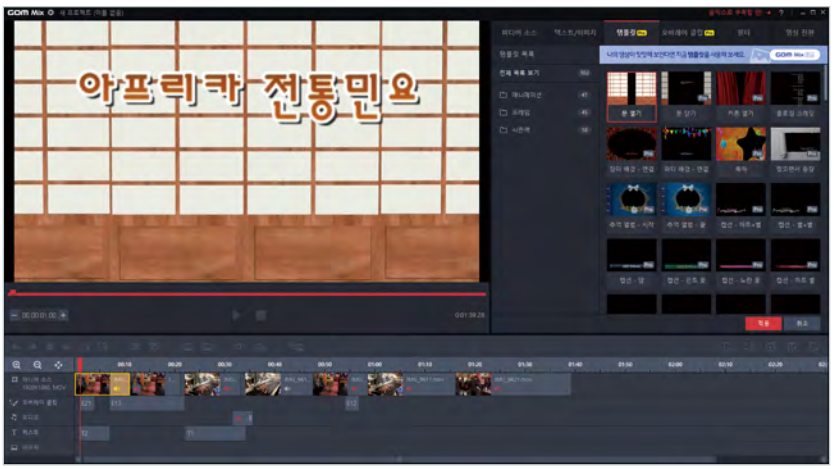




- 기능탭의 〈텍스트/이미지〉에서 ❶ 텍스트 추가를 선택합니다.
- 텍스트 편집 도구에서 자막을 입력합니다.
- 타임라인에서 텍스트 클립의 길이를 조절한 뒤 ❷ 적용을 선택합니다.
- 타임라인에 텍스트 클립이 생성됩니다.

곰믹스에서는 〈템플릿〉탭에서 애니메이션 효과를 제공하고 있습니다. 시작 장면과 마지막 장면에 템플릿을 활용하면 풍성한 느낌의 영상을 만들 수 있어요. 하지만 유료로 제공하는 기능이니 신중하게 선택하세요!

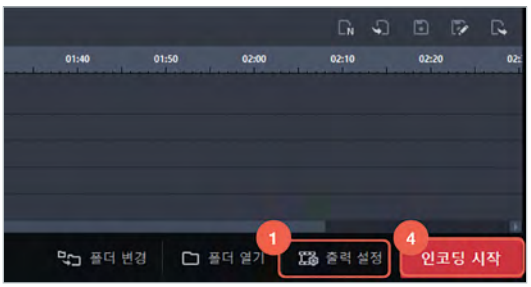
템플릿



곰믹스에서는 〈템플릿〉탭에서 애니메이션 효과를 제공하고 있습니다. 시작 장면과 마지막장면에 템플릿을 활용하면 풍성한 느낌의 영상을 만들 수 있어요. 하지만 유료로 제공하는 기능이니 신중하게 선택하세요!

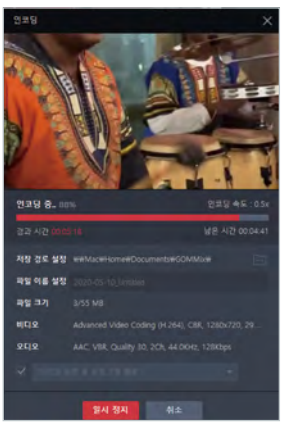
인코딩

영상의 편집이 모두 끝났다면 마지막으로 인코딩을 합니다. 인코딩은 쉽게 말해 영상작업파일을 압축하는 것을 말하는데, 누구나 쉽게 볼 수 있도록 확장자를 MP4로 바꿔줄 수 있습니다. 하지만 이렇게 인코딩된 파일은 다시 작업할 수 없으니 모든 작업이 끝난 후 마지막에 실행합니다.



- 인코딩 시작 전 ❶ 출력설정을 확인해봅니다.
- 출력형식을 ❷ MP4로 합니다.
- 영상 크기를 ❸ 1920x1080 으로 합니다

TIP 1 1920x1080은 FullHD(FHD) 크기로 고품질의 영상 크기입니다. 이보다 아래는 영상 화질이 안 좋게 보일 수 있습니다. 그리고 곰믹스는 아쉽게도 4K는 지원하지 않습니다.



- ❹ 인코딩 시작을 선택합니다.
- 이제 인코딩이 끝날 때까지 커피 한 잔을 합니다.

폰 속에 잠들어 있는 사진 · 영상에 새로운 생명을 불어넣어보자

어떤가요? 모바일의 작은 화면에서 편집하는 것 보다 훨씬 쾌적한 환경에서 영상작업을 할 수 있죠? 그리고 생각보다 어렵지 않습니다. 영상을 불러오고 음악 넣고 컷 편집 그리고 장면전환과 색감 맞추고 자막 넣으면 끝.

사실 모든 영상편집 과정이 딱 이 정도입니다. 이렇게 영상편집의 큰 흐름을 익혀둔 다음, 내가 만들고 싶은 영상을 어떻게 만들지 구상해보세요. 막상 구상하려니 머릿속에서 안 떠오르죠? 그래서 다음 호부터는 위의 영상편집의 기본과정을 바탕으로 '내 폰 속에 잠들어있는 사진으로 영상 만들기', '가족과 함께 보낸 시간' 등의 주제를 두고 쉽게 영상 만드는 법을 소개해드리려 합니다.

그럼 다음 호에서 만나요!

신동헌 ◆ 통계의 창 객원기자

통계로 바라보는 세상 이야기

코로나19가 바꿔놓은 대한민국의 변화

코로나19가 바꿔놓은 대한민국의 변화

글로벌 데이터 인사이트 컨설팅 기업 '칸타'의 '코로나19 여파에 따른 라이프스타일 조사'에 따르면, 응답자의 75%가 코로나19가 내 일상생활에 영향을 미치고 있다고 답변했고, 41%는 온라인 쇼핑을 주로 이용할 것이라고 했으며, 42%는 스트리밍 서비스 콘텐츠를 이용하겠다고 응답했다. 사람들이 외부 활동을 자제하고 집에서 머무는 시간이 늘면서 '온라인 쇼핑'의 비중이 날로 높아지고 있다. 또한 시장조사업체 '엠브레인 트렌드모니터'가 전국 만 19~59세 성인 남녀 2000명을 대상으로 '코로나19'와 관련한 인식을 조사한 결과, 소비자 10명 중 6명(62.1%)이 코로나19 발생 이후 인터넷 쇼핑몰 방문이 늘어날 것 같다고 응답했으며, 모든 연령대가 비슷한 것으로 확인됐다.

01

준연동형 비례대표제와 만 18세 선거 참여

대한민국의 국회의원선거는 국회의원 300석을 선출하기 위해 치러졌다. 지역구 253석과 비례대표 47석으로 구성되어 있는데, 기존에 연동형이 아닌 '병립형'은 비례대표 47석이 각 정당의 득표한 비율대로 배분됐다(단, 정당 득표율 3% 이상 정당에 한하여). 하지만 준연동형 비례대표제는 비례대표 47석 중에서 17석에 대해서는 기존의 병립형 방식을 그대로 적용하되, 나머지 30석은 준연동형 비례대표제에 따라 각 정당 득표율의 50%만 적용하여 배분했다. 리얼미터에서 실시한 2019년 선거권 연령 하양 조정에 대한 국민여론을 살펴보면 51.4%가 찬성, 46.2%가 반대를 보였고, 이번 선거에는 2002년 4월 16일 이전 출생한 만 18세까지 선거에 참여할 수 있었다.

02

대한민국 임시정부 수립부터 100년 뒤 미래까지!

"역사를 잊은 민족에게 나라는 없다"라고 한다. 임시정부 수립일을 기념하며, 우리가 지금 살고 있는 대한민국이 어떻게 시작되었는지 다시 생각해봐야 한다. 지금으로부터 101년 전인 1919년 4월 11일은 대한민국 임시정부 수립일이다. 중국 상해에 대한민국 임시정부가 설립되기까지 독립을 향한 우리 민족들의 간절한 노력이 있었다. 특히 1919년 3월 1일에 시작된 만세운동이 전국적으로 확산된 것이 임시정부 설립의 계기가 되었다. 문화체육관광부에서 2019년에 실시한 3 · 1운동 및 대한민국 임시정부 수립 100주년 국민인식조사에 따르면 '독립을 위한 외교활동 구심점 역할'이 29.0%로 1위를 차지하였다.

03

통계로 알아보는 멀고도 가까운 곳, 북한!

통계청은 통일 미래를 준비하고, 북한의 경제 · 사회 실상에 대한 정보를 국민에게 제공하고자 1995년부터 매년 북한 주요통계지표를 발표하고 있다. 지난 18일 발표한 '2019 북한의 주요통계지표'를 보면, 북한은 2019년 도시화율 약 62.1%를 달성하였고, 2010년부터 꾸준히 증가하고 있다. 2018년 무역 총액은 약 28억 달러로, 대부분의 교역이 중국과 이루어진다. 1인당 국민총소득은 2018년 142.8만 원이다. 국민의 33%가 서비스업에 종사하고 있고, 29.4%가 광공업, 23.3%가 농림어업, 8.9%가 건설업, 5.4%가 전기 · 가스 · 수도업에 종사하고 있다. 보다 자세한 내용은 통계청 홈페이지(kostat.go.kr)의 통계포털(KOSIS) 중 '북한통계'를 찾아볼 수 있다.

04

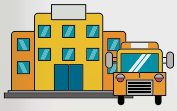
위기에 더욱 빛나는 대한민국 보건의료제도

2017년 기준 한국 국민 1인당 의사 외래 진료 횟수는 16.6회로 OECD 국가 중 1위로, OECD 평균(7.1회, 2017년 기준)보다 약 2.3배 높아 의료 접근성이 상당히 높다. 보건복지부, 2019년 의료서비스경험조사에서 우리국민의 보건의료제도에 대한 만족도를 묻는 질문에 '대체로 그렇다'가 58.5%, '매우 그렇다'가 7.7%로 대체적으로 만족하였다. 또한 보건의료제도에 대한 신뢰도를 묻는 질문 역시 '대체로 그렇다'가 58.2%, '매우 그렇다'가 7.7%로 우리나라 국민은 자국의 보건의료제도를 신뢰하고 있는 것으로 나타났다. 이렇듯 우리나라가 코로나19에 큰 혼란 없이 모범적으로 대응할 수 있었던 것은 우리 스스로 보건의료에 대한 믿음이 밑바탕에 있었기 때문일 것이다.

05

온라인 개학했으니까 학원 보내도 될까요?

코로나19 확산 방지로 연기되었던 초 · 중 · 고 개학이 지난 4월 9일 중 · 고등학생 3학년을 시작으로 오는 4월 20일 초등학교 1, 2, 3학년까지 모두 온라인으로 수업을 하고 있다. SM C&C의 4월 10일 설문조사에 따르면, 개학이 3월 23일로 연기된 직후 학부모를 대상으로 설문조사를 실시한 결과, 학원 개원에 대해 '개학할 때까지 휴원'이 42%로 높게 나타났다. 최근 발표된 서울시 교육청의 4월 14일 보도자료에 따르면 4월 13일 기준 서울 내 학원 · 교습소 휴원율은 18.3%이다. 특히 휴원율 추이 상 3월 16일과 23일에 휴원율이 각각 18.3%p, 15.5%p 씩 큰 폭으로 떨어졌다. 이는 한 달 넘게 이어진 장기 휴원에 부담을 느낀 학원들이 문을 연 것으로 보인다.



06

한국 아동 · 청소년, 학업 스트레스 심각

한국청소년정책연구원 '유민상 부연구위원'이 분석한 '아동 · 청소년 삶의 질 지표 분석 결과'를 통해 함께 살펴보자. 한국은 OECD 국제학업성취도평가(PISA)에서 수년째 상위권을 유지하고 있다. 2015년 PISA 결과에서는 순위가 조금 하락했으나, 여전히 OECD 35개국 중 읽기 4~9위 사이, 수학 6~9위 사이, 과학 9~14위 사이 등으로 상위권의 성적을 보인다. 하지만 익히 알려진 바와 같이 우리나라 아동 · 청소년의 학업으로 인한 스트레스는 매우 높은 편이다. 한 조사에서 "죽고 싶다는 생각을 가끔 하거나 자주 한다"고 응답한 아동 · 청소년들은 전체 응답자의 33.8%였는데, 그 중에서 37.2%가 학업 문제로 인해 죽고 싶다는 생각을 해보았다고 응답하였다.

07

1분기 경제성장률 2008년 금융위기 이후 최저

우리나라 1분기 경제 성장률이 −1.4%로 집계됐다. 2008년 금융위기 이후 최저치로 코로나19 쇼크가 현실로 나타났다. 민간소비는 1분기 6.4% 감소해 1998년 1분기 외환위기 이후 22년 만의 최저치를 기록했다. 한국은행은 실질 국내총생산(GDP) 속보치 통계에서 전기 대비 1분기 성장률이 −1.4%로 집계됐다고 밝혔다. 2008년 4분기 −3.3% 이후 11년 3개월 만에 가장 낮은 성장률이고, 2009년 3분기(0.9%) 이후 10년 반 만에 가장 낮은 수치다. 국내 감염병 확산이 2월부터 본격화하면서 충격을 받은 민간소비와 서비스업 생산이 성장률을 끌어내렸다. 국제통화기금(IMF)은 올해 한국 경제성장률을 −1.2%로 하향 조정했다.

08

슬기로운 직장생활, 세대차이 인정해요

통계청, 2019년 경제활동인구조사 청년층 부가조사에 따르면, 2019년 청년들이 첫 일자리를 그만둔 비율은 67.0%로 전년동월 대비 4.2%p 상승했고, 평균 근속기간은 1년 1.6개월로 0.3개월 감소했다. 청년들이 첫 일자리를 그만둔 사유는 보수, 근로시간 등 근로여건 불만족(49.7%)가 가장 높았다. 사람인, 2018년 직장 내 세대차이 조사에 따르면 응답자 중 세대차이를 줄이기 위한 노력을 하고 있는 사람은 48%다. 이들은 '서로 다름을 인정(75.7%)' 하고자 하는데, 이 중 가장 노력하고 있는 세대는 50대(72.2%)로 나타났다. 세대차이는 서로 살아온 시간과 환경이 다른 세대들이 어울려 생활하며 발생하는 것이기 때문에 특정 세대가 아닌 모두가 노력할 필요가 있다.

09

'집콕'해서 찐 살, 우리 같이 '홈트' 할래요?

'홈트'는 홈트레이닝의 줄임말로 헬스장이나 수영장 등 운동시설이 아닌 가정에서 할 수 있는 간단한 운동을 말한다. 네이버 데이터랩에 의하면 2020년 1월부터 3월까지 약 3개월간 '홈트, 홈트레이닝, 집운동, 집에서 하는 운동, 홈운동' 등 홈트레이닝 관련 키워드 검색량 변화가 많았다. 트렌드모니터의 2018년 홈트 관련 U&A조사에 따르면, 최근 홈트를 하는 사람들이 증가하는 이유는 '다른 곳으로 운동을 다닐 만큼 시간이 여유롭지 않아서(47.9%)'라고 한다. 이처럼 홈트를 사람들이 많이 찾는 이유는 '저렴하게 운동하고 싶어서(47.9%)', '전문가 못지않은 정보를 얻을 만큼 채널이 많아져서(34.6%)', '홈트용 운동 기구가 다양해져서(30.3%)' 순으로 나타났다.

10

SNS 피로증후군 디톡스가 필요!

과학기술정보통신부의 2019 스마트폰 과의존 실태조사에 따르면, 과의존위험군에 속하는 사람은 20.0%로 나타났다. 스마트폰 과의존위험군 비율은 2011년부터 지속적인 증가 추세를 보여준다. 이들은 스마트폰으로 주로 메신저를 이용했으며, SNS 사용률도 75.0%를 보이며 사람들과의 연결망을 중요시하는 콘텐츠를 꾸준히 소비하는 것을 알 수 있다. 트렌드모니터의 2017 SNS 이용 및 피로증후군 관련 인식 조사에 따르면 31.7%가 SNS피로증후군을 경험했다고 답했다. 'Disconnect to Reconnect(다시 연결하기 위한 단절)'는 디지털 디톡스를 전개하는 기관(digitaldetox.org)의 슬로건으로 더 나은 사용자가 되기 위해 노력하라고 강조한다.

11

한국인의 식습관, 어떻게 변했나

통계청, 2018 사회조사에 따르면 2018년 아침 식사를 먹지 않는 사람은 32.7%로 10년 전(23.8%)에 비해 8.9%p 증가하였다. 특히 20대의 경우, 57.4%로 2명 중 1명도 채 아침 식사를 먹지 않고 있다. 허벌라이프 뉴트리션, 2018 아태지역 건강한 아침식사 설문조사에 따르면 아침식사 섭취에 장애물이 되는 가장 큰 요인은 시간 부족(68%)으로 나타났다. 2018년 가족과 함께 저녁식사를 하는 비율은 66.0%이며, 이 중에서도 특히 19~29세는 47.9%로 가장 낮게 나타났다. 잡코리아와 알바몬이 실시한 2018 가족과 식사에 관한 설문조사에서는 가족과 저녁식사를 자주 못하는 이유로 '업무/과제가 너무 많아 가족과 식사할 시간이 없어서'라는 답변이 31.9%로 가장 높게 나타났다.

12

너도 나도, 지금은 유튜브 창업 열풍!

앱 데이터 분석업체 와이즈앱의 2019년 11월 한국인이 가장 오래 사용하는 앱 조사 결과에 따르면 유튜브가 442억분으로 가장 높은 사용 시간을 보였다. 특히, 226억분으로 2위를 기록한 카카오톡과 격차가 크게 벌어지면서, 유튜브가 대세임을 재확인하였다. 취업포털 MJ 플렉스의 2019년 자체설문조사에 따르면, 유튜브 시청 기기 1위는 스마트폰이며, 영상 시청 시간은 오후 7시부터 10시까지가 48.3%로 가장 높게 나타났다. 성공적인 유튜브 창업을 위해 유튜버로서 자신만의 콘텐츠에 대한 정확한 방향성을 잡아야 한다. 1인 크리에이터가 많이 등장한 만큼 남들과는 다른 독특한 아이템, 영상 촬영법과 디자인 등을 연출할 수 있도록 끊임없이 연구해야 한다.

13

창업 전 필수 통계청 서비스, 'My 통계로'

혹시 창업을 준비하고 있는가? 지난 2월 21일부터 진행된 'My 통계로(路)'는 통계청에서 통계지리정보 서비스(SGIS)를 이용할 수 있도록 만든 서비스로 총 1,147개의 공간통계 정보를 제공한다. 통계주제도, 대화형통계지도, 활용서비스 등의 콘텐츠가 있으며, 이는 ① 접속지역을 자동설정하거나 관심지역을 선택한 후에 ② 영유아/어린이, 청소년, 청년, 장년, 노년, 임신/출산/육아여성, 1인가구의 7개 나이를 선택하고 ③ 먹거리, 살거리, 일거리, 탈거리, 배울거리, 보고 놀거리, 건강거리, 안전거리의 8개 관심분야로 분류, 이를 토대로 키워드에 따른 공간통계정보를 추천하여 맞춤 정보를 도출함으로써 업종별 사업체와 밀집도 현황 등 다양한 자료를 동시에 조회할 수 있다.

14

나이팅게일과 세종대왕의 공통점은?

나이팅게일과 세종대왕의 공통점은 바로 데이터를 목적에 맞게 해석하는 능력인 '데이터 리터러시'라고 한다. 1854년, 크림전쟁에서 간호사로 활동했던 나이팅게일은 통계 데이터를 기반으로 '로즈 다이어그램'이라는 장미 모양의 그래프를 만들어 병원 위생의 중요성을 알렸고, 병원의 위생을 개선하여 5개월 만에 병원 내 군인 사망률이 42%에서 2%로 크게 감소하였다. 또한 세종은 국가재정의 안정을 위해 1430년 새로운 세법인 '공법' 실시 찬반에 대하여 각 도의 관리와 백성을 대상으로 여론조사를 실시한 결과, 찬성과 반대 비율이 커서 공법의 실시를 유보하였다. 이후 찬성이 높은 지역에서 우선 공법을 시행하고, 약 7년 후에 전국적으로 공법을 적용하였다.

15

드라마 〈스토브리그〉 속 통계 이야기

올해 초 16부작으로 종영한 드라마 〈스토브리그〉 속 통계가 흥미롭다. 〈스토브리그〉는 야구 시즌이 끝난 뒤부터 다음 시즌이 시작되기 전까지의 비시즌 기간을 말한다. 선수의 지난 플레이에 대한 통계값인 타율, 타자가 1루에 얼마나 많이 살아 나갔는지를 백분율로 나타낸 출루율, 단타를 1, 2루타를 2, 3루타를 3, 홈런을 4로 계산하여 합한 수를 타수로 나눈 값인 장타율, 앞의 출루율과 장타율을 합친 OPS, 한 투수가 9이닝(한 경기)을 던졌을 때 평균 몇 점을 주는가를 나타내는 ERA(평균자책점)를 비롯해 한 투수가 한 이닝에 볼넷과 안타를 얼마나 허용하는지를 보여주는 지표인 WHIP(이닝당 볼넷 안타 허용률)와 수비수의 수비율 등 통계지표로 기록되고 분석된다.

16

2020년도 통계교육원 교육훈련계획

[집합과정 – 100개 과정 163회 4,810명]

*서울교육장 교육과정

구분		과정명	교육 대상	교육 일수	기당 인원	교육 횟수	교육 일정
기본교육		신규자 기본교육	신규임용 예정자	15	60	1	8.10.~8.28.
		경력채용자 기본교육	신규임용 예정자	10	30	1	4. 6.~4.17.
		4급 승진후보자 역량향상	4급 승진후보자	5	20	1	7.20.~7.24.
		5급 승진후보자 역량향상	5급 승진후보자	5	20	2	5.11.~5.15, 9.21.~9.25.
		6급 승진자 역량향상 (신설)	6급 승진자	4	30	2	5.19.~5.22, 10.27.~10.30
		현장조사 역량강화 (신설)	조사담당자	3	30	2	3.25.~3.27, 4.22.~4.24.
		지방청 조사관리자 역량강화1 (신설)	지방청 총괄자	3	30	1	5.20.~5.22.
		지방청 조사관리자 역량강화2 (신설)	지방청 팀장	3	30	1	6.24.~6.26.
전문교육	국가통계정책	국가통계의 이해	제한없음	3	40	2	4.22.~4.24, 6.29.~7. 1
		통계와 정책	통계청	3	20	2	2.19.~2.21, 6. 3.~6. 5.
		정책과정과 통계의 역할	제한없음	3	20	2	5. 6.~5. 8, 10.21.~10.23
		국가승인통계관리	통계작성기관	3	30	2	3. 2.~3. 4, 9.16.~ 9.18
		통계품질관리	제한없음	3	30	2	2.12.~2.14, 7.6.~ 7. 8.
		정책지표 작성 방법론	제한없음	2	20	1	9.14.~9.15.
		북한통계의 이해	제한없음	3	30	1	10.12.~10.14.
		성인지통계의 이해	제한없음	2	40	1	8.27.~8.28.
		통계기초 및 활용	제한없음	5	30	3	3.23.~3.27, 6.15.~ 6.19, 1.9.~11.13
	국가통계기준	한국표준산업분류	제한없음	3	30	1	3.16.~3.18.
		한국표준직업분류	제한없음	3	30	1	6. 8.~6.10.
		한국표준질병 · 사인분류 이해	제한없음	1	30	1	3.30.*
		한국표준질병 · 사인분류 활용	제한없음	1	30	1	11. 2.
	경제분야	경제통계의 이해	제한없음	3	20	1	9.14.~9.16.
		국민계정	제한없음	3	20	1	8.26.~8.28.
		재무제표	제한없음	3	30	3	2.17.~2.19, 6.22, 6.24, 8.19.~8.21
	사회분야	농어업통계의 이해	제한없음	3	30	1	9.21.~9.23.
		사회통계의 이해	제한없음	3	20	1	8.26.~8.28.
		인구통계의 이해	제한없음	3	20	1	4.20.~4.22.
	행정통계	빅데이터와 행정자료의 이해	통계청	3	30	2	3.25.~3.27, 9. 2.~9. 4
	조사기획	국가통계실무1(조사설계 및 조사표설계 등)	통계청	4	40	2	2.24.~2.27, 9. 7.~9.10
		국가통계실무2(표본설계 및 추정)	통계청	4	40	2	3. 9.~3.12, 10.26.~10.29.
		조사설계 및 조사표설계	제한없음	3	20	1	5. 6.~5. 8.
		표본실무	제한없음	4	20	1	7. 7.~7.10.
	자료수집처리 및 분석	통계조사관 직무연수	통계청	3	30	1	8.24.~8.26.
		지역통계실무	제한없음	3	20	1	10.14.~10.16.
		국가통계실무3(자료수집 · 처리 및 분석)	통계청	4	40	2	4. 6.~4. 9, 11. 9.~11.12.
		자료수집 · 처리 및 분석	제한없음	3	20	1	7.20.~7.22.
		계절조정실무	통계청	2	30	1	1.13.~1.14.
		데이터 에디팅	제한없음	4	20	1	6. 1.~6. 4.
		시계열분석	제한없음	4	20	1	5.12.~5.15.
		지수이론	제한없음	3	20	1	7. 1.~7. 3.
	공표 및 관리	국가통계실무4(통계작성 · 공표 등)	통계청	4	40	2	6. 8.~6.11, 23~11. 26.
		국가통계정보의 활용	제한없음	3	20	1	4.27.~4.29.
		통계자료의 비밀보호	제한없음	3	20	1	7.13.~7.15.
		통계보고서 작성	제한없음	3	30	2	3.23.~3.25, 9. 2.~9. 4.
	R	R 초급 통계분석	제한없음	4	30	4	3. 3.~3. 6, 5.12.~5.15, 9. 1.~9. 4, 11. 3.~11. 6.
		R 데이터시각화	제한없음	3	30	2	6.17.~6.19, 10.21.~10.23
		R 중급 통계분석	제한없음	3	30	1	4. 1.~4. 3.
		R 고급 통계분석	제한없음	3	30	1	9. 9.~9.11.
전문교육	SAS	SAS 입문	제한없음	3	30	2	2.19.~2.21, 8.26.~ 8.28.
		SAS 중급 통계분석	제한없음	5	20	2	4. 6.~4.10, 10.19.~10.23.
		SAS 고급 통계분석	제한없음	5	20	2	6.22.~6.26, 11.23.~11.27.
	SPSS	SPSS 초급 통계분석	제한없음	3	30	2	3. 4.~3. 6, 7.15.~7.17.
		SPSS 중급 통계분석	제한없음	5	30	2	5.18.~5.22, 10.12.~10.16.
	엑셀	엑셀 초급 활용	제한없음	4	40	3	2.25.~2.28, 6.30.~7. 3, 10.27.~10.30
		엑셀 중급 통계분석	제한없음	3	30	2	4.27.~4.29, 9.14.~9.16.
	빅데이터	행정자료 통계작성	제한없음	3	20	1	6.10.~6.12.
		빅데이터 프로젝트 수행 (신설)	제한없음	4	20	2	4.21.~4.24, 7. 7.~7.10.
		하둡 기반 빅데이터 통계분석 (신설)	제한없음	3	20	2	5.13.~5.15, 10.28.~10.30*
		파이썬 통계분석 (신설)	제한없음	3	20	2	4.22.~4.24, 7. 1.~7. 3*
	전문관양성	경제시계열분석 및 지수이론	통계청	14	10		
		국민계정	통계청	12	10		
		무응답 자료처리 및 분석	통계청	12	10		
		인구통계분석	통계청	12	10		
		통계분류	통계청	12	10		
		표본설계 및 추정	통계청	13	10		
	맞춤형(기관)		통계작성기관	2	25	17	
	통계 세미나		제한없음	1	40	4	
	학생교육	사랑나눔 통계교실	초등학교 5~6학년	3	25	2	6. 1.~6. 3, 9.21.~ 9.23.
		어린이 통계캠프	초등학교 5~6학년	3	30	2	5.11.~5.13, 11. 9.~11.11.
		중학생 통계아카데미	중학생	3	30	2	7. 8.~7.10, 7.15.~7.17.
		고등학생 통계아카데미	고등학생	2	30	2	8. 3.~8. 4, 8. 6.~8. 7.
	교사교육	초등학교 교사 통계연수	초등학교 교사	2	30	2	1.20.~1.22, 7.27.~7.28.
		중학교 교사 통계연수	중학교 교사	2	30	2	1.15.~1.17, 7.30.~7.31.
		고등학교 교사 통계연수	고등학교 교사	2	30	2	1.13.~1.15, 8.10.~ 8.11
		중등 교사 통계연수(심화)	중등 교사	4	30	2	1.20.~1.23, 8. 3.~8. 6.
		실용통계 지도교사 통계연수 (신설)	고등학교 교사	4	20	1	7.27.~7.30
		시도교육청 교사연수	교사	2	30	6	미정
	외국인교육	KOICA 위탁연수	외국공무원	21	15	2	
		UNSIAP 공조 통계연수	외국공무원	5	25	2	
	정보화	스마트 문서편집	제한없음	4	40	4	3.24.~3.27, 5.19.~5.22*, 7.21.~7.24, 11.10.~11.13
		파워포인트 활용	제한없음	3	40	3	4. 1.~4. 3*, 9. 9.~9.11, 11.16.~11.18*
		액세스 활용	제한없음	5	30	1	7.13.~7.17.
		오피스를 활용한 데이터 시각화	제한없음	3	40	4	4.20.~4.22*, 7. 6.~ 7. 8, 9.21.~ 9.23, 11. 4.~11. 6.
		디지털 영상 및 이미지 활용	제한없음	3	40	1	5. 6.~5. 8
기타교육	시책	소셜미디어 활용	제한없음	2	20	1	6.22.~6.23.*
	일반소양	생활 속의 음악	제한없음	3	80	2	6. 3.~6. 5, 11. 4.~11. 6.
		현장조사 스트레스 관리	통계청	3	30	3	5.27.~5.29, 10.14.~10.16, 11.18.~11.20.
		공무직 은퇴설계 (신설)	통계청	3	30	1	

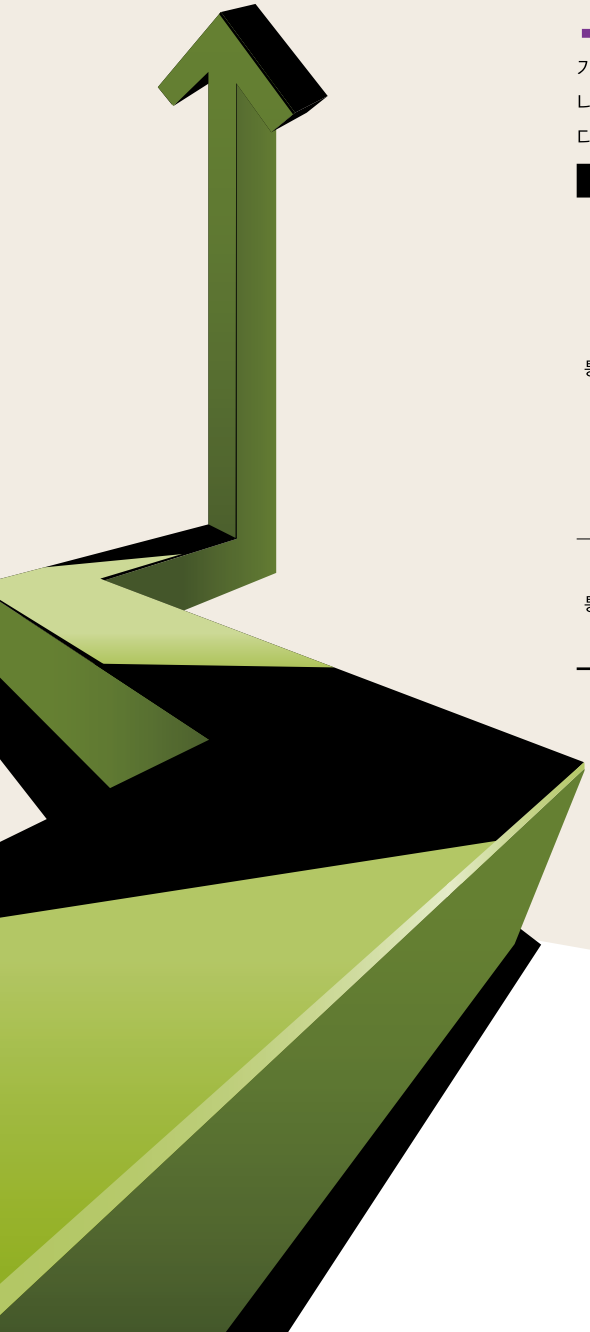
2020년도 통계교육원 교육훈련계획

[이러닝(e-Learning) 과정 – 103개 과정(2.1.~12.13.) 중 실시 운영]

구분		과정명	난이도	교육 대상	인정 시간
기본교육		5급 승진휴보자 역량평가의 이해(신설)	초급	통계청	5
전문교육	국가 통계 정책	자체통계 품질진단 관리	초급	제한없음	11
		지역사회지표 작성과 활용	초급	제한없음	11
		지역정책과 통계활용	초급	제한없음	23
		통계 맛보기	초급	제한없음	11
		통계기초 및 활용	중급	제한없음	25
		통계법	초급	제한없음	4
		통계업무 필수 지식	초급	제한없음	8
		통계작성기관을 위한 통계DB시스템사용법	초급	제한없음	5
		통계적으로 사고하기	중급	제한없음	13
		통계학의 이해	초급	제한없음	29
	국가통계 기준	한국표준산업분류	초급	제한없음	13
		한국표준직업분류	초급	제한없음	14
	경제통계	광업 · 제조업동향조사	초급	통계청	14
		서비스업동향조사	초급	통계청	10
		소비자물가조사	초급	통계청	10
		온라인쇼핑동향조사	초급	통계청	7
		재무제표이해	중급	제한없음	15
	사회통계	가계동향조사 (신설)	초급	통계청	6
		가축동향조사	초급	통계청	7
		경제활동인구조사	초급	통계청	6
		경제활동인구조사 사례집	초급	통계청	4
		농가경제조사	초급	통계청	10
		농가판매 및 구입가격 조사	초급	통계청	6
		농산물생산비조사	초급	통계청	9
		농업면적조사	초급	통계청	7
		농작물생산조사[생산량부문]	초급	통계청	8
		사회통계의 이해 (신설)	초급	제한없음	15
		산지쌀값조사	초급	통계청	5
		양곡소비량 조사	초급	통계청	4
		어가경제조사	초급	통계청	10
		어류양식동향조사	초급	통계청	5
		어업생산동향조사	초급	통계청	7
		이민자 체류실태 및 고용조사	초급	통계청	6
		인구동향조사	초급	통계청,공무원	6
		지역별 고용조사 (신설)	초급	제한없음	6
		축산물생산비조사	초급	통계청	8
	조사기획	조사방법 기초	초급	제한없음	10
		조사방법의 이해	초급	제한없음	20
	자료수집처리 및 분석	시계열자료의 분석과 실무	고급	제한없음	16
		표본이론 기초	중급	제한없음	14
		현장조사 인력양성	초급	제한없음	8
		회귀분석의 이해와 사례	고급	제한없음	18
	공표 및 관리	국가통계포털(KOSIS) 활용	초급	제한없음	6
		통계를 활용한 보고서 작성방법	중급	제한없음	12
		e-나라지표 업무시스템 이용방법 (신설)	초급	통계청 공무원	5
		NARA-PC 활용	초급	제한없음	10
		MDIS 활용	초급	제한없음	6
		SGIS 에듀	초급	제한없음	10
		SGIS 플러스 활용	초급	제한없음	7
	R	R 기초	초급	제한없음	25
		R 활용	중급	제한없음	20
	SAS	예제로 본 SAS	고급	제한없음	20
	SPSS	SPSS 초급 통계분석	초급	제한없음	14
		SPSS 중급 통계분석	중급	제한없음	13
	엑셀	엑셀을 이용한 통계분석	중급	제한없음	13
전문교육	빅데이터	공공데이터로 만들어가는 새로운 세상 (신설)	초급	제한없음	12
		공공빅데이터 업무적용 길라잡이	초급	통계청	16
		빅데이터와 통계	중급	제한없음	15
	기타 데이타 분석	통계패키지 학습을 위한 필수 통계지식	중급	제한없음	5
		통계 교육용 공학도구 툴그라미 활용	초급	제한없음	6
	학생교육	어린이 통계교실	초급	제한없음	20
		중학생 아카데미	초급	제한없음	20
		통계로 논리를 잡아라 (신설)	초급	제한없음	10
		통계로 보는 안전교육 (신설)	초급	제한없음	2
		통계를 보면 경제가 보여요 (신설)	초급	제한없음	2
	교사교육	통계포스터 만들기 지도방법 (신설)	초급	제한없음	6
	정보화 분야	엑셀 2010	초급	통계청	20
		한글 2010	초급	통계청	20
		한글, 엑셀, 파워포인트 활용 TIP	초급	제한없음	12
기타교육	시책 교육	4차 산업혁명의 이해와 미래대응전략	초급	통계청	11
		개인정보보호법 이해하기	초급	통계청	7
		내부 소통능력 및 국민소통 능력향상 (신설)	초급	통계청	8
		독립운동가를 통해 본 나라사랑과 국가관	초급	통계청	15
		사회적 경제(이해편) (신설)	초급	제한없음	3
		하동학대신고 의무자 교육	초급	통계청	4
		안전한 사회를 위한 폭력 예방교육	초급	통계청	5
		업무관리능력 향상 (신설)	초급	통계청	4
		역사속에서 찾은 청렴이야기	초급	통계청	10
		위기관리 커뮤니케이션	초급	통계청	7
		이순신장군의 청렴리더십	초급	통계청	8
		인권의 이해	초급	통계청	15
		장애인 동료와 함께 일하기	초급	제한없음	5
		적극행정의 이해	초급	통계청	6
		정보보안	초급	통계청	5
		청탁금지법의 이해	초급	통계청	5
		한반도 정책의 이해 (신설)	초급	제한없음	3
		현장에서 배우는규제개혁	초급	제한없음	5
		홍보업무의 달인되기	초급	통계청	7
	일반 소양	강의 운영 기법 및 전략 (신설)	초급	제한없음	10
		개인 역량 강화를 위한 개인리더십	초급	제한없음	2
		대국민지진교육 (신설)	초급	통계청	16
		마음을 움직이는 설득 전략 (신설)	초급	제한없음	10
		몸이 바로서야 그림자도 바로선다, 개인리더십	초급	제한없음	2
		문제해결능력 키우기	초급	제한없음	2
		부모교육(생애주기별) (신설)	초급	통계청	8
		설득의 심리학	초급	제한없음	3
		쉽게 배우는 서평쓰기	초급	통계청	12
		풍요로운 삶을 위해 자신을 코칭하기	초급	제한없음	2
		협상력을 높이는 협상력 증강공식	초급	제한없음	2
	공통 직무	공무원의 행복한 미래 설계	초급	통계청	8
		공무원 징계 및 소청제도의 이해	초급	통계청	12
		예산과정의 이해	초급	통계청	10
		공직자 영어	초급	통계청	15

통계청, 정부부처, 지방자치단체, 연구기관 등 모든 기관의 마이크로데이터를 한 곳으로...

보다 심도 있고 다양한 분석을 원한다면 지금 바로 MDIS를 클릭해보세요



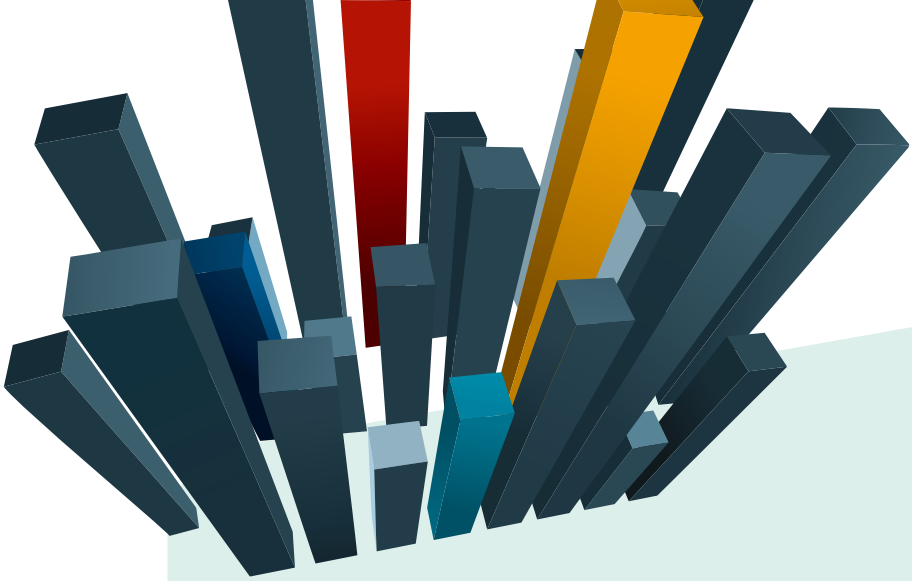
■ **서비스 소개** (2020년 3월 기준)

가. 서비스명 : 마이크로데이터통합서비스(MDIS, mdis.kostat.go.kr)

나. 제공통계수 : 총 202종(통계청 46종 및 통계작성기관 230종)

다. 제공형태 : 마이크로데이터(통계에 따라 사람, 사업체, 가구 기반 자료)

기준		주요 통계
통계청	인구 · 가구	경제활동인구조사, 가계동향조사, 국내인구이동통계, 사망원인통계, 가계금융복지조사, 지역별고용조사, 인구주택총조사, 인구동향조사, 생활시간조사, 사회조사 외 5종
	사업체 · 농어가	전국사업체조사, 광업 · 제조업조사, 농가경제조사, 기업활동조사, 농업총조사, 농산물생산비조사, 경제총조사, 어가경제조사, 운수업조사 외 14종
	행정통계 및 기타	귀농어귀촌인통계, 영리법인기업체행정통계, 신혼부부통계, 주택소유통계, 중장년층행정통계 퇴직연금통계, 일자리행정통계, 기업생멸행정통계
통계작성기관		전국다문화가족실태조사, 가족실태조사, 자동차주행거리통계, 직종별사업체노동력조사, 보육실태조사, 지상기상통계, 국민여가활동조사, 외래관광객실태조사, 한부모가족실태조사, 청소년종합실태조사 외 220종



■ **서비스 내용**

가. 구분 : 자료의 민감성 정도에 따라
공공용, 인가용, 특수목적용으로 구분 운영

나. 수수료
– 무료 : 공공용 자료
– 인가용/특수목적용 : 선택제 수수료 부과

다. 서비스 방법
– 추출·다운로드 : MDIS 포털에서
직접 무료 다운로드
– 원격접근서비스 : 승인 후 이용자가
집·사무실 등에서 통계청 서버 접속 후 활용
– 이용센터 : 승인 후 지정된 장소를 방문 · 활용

■ **문의**

– 연락처 : 재단법인 한국통계진흥원
– 전화 : (02) 512–0167 FAX : (02) 515–0240
– 주소 : (우) 06097 서울특별시 강남구 선릉로 612, 6층
– E–mail : MDIS@stat.or.kr

공공자료의 개방 및 공유 확대

MDIS
[mdis.kostat.go.kr]

통계청 46종 및 통계작성기관 230종의
통계자료로 제공확대

통계청에서 국가통계를 활용하세요!

통계청은 **통계개발·활용**에 필요한 모든 **정보와 도움**을 제공합니다.
다양한 국가통계정보 제공사이트를 활용하세요.

원하는 자료를 직접 분석 및 요청

MDIS
[mdis.kostat.go.kr]

온라인으로 추출/다운로드 선택 시
공공용 마이크로데이터를 무료로
분석 활용 가능

국가통계 쉽게 찾기

KOSIS
[kosis.kr]

국내, 국제, 북한의 주요 통계를
한 곳에 모아 알기 쉽게 분류해 제공

국가 발전 상황을 한눈에

국가지표체계
[www.index.go.kr]

국민의 관심이 크고 정책 수립에
활용 가능한 지표

지도 위 통계정보 살펴보기

SGIS
[sgis.kostat.go.kr]

인구, 가구, 주택, 사업체 통계 등 각종
통계를 지도(GIS) 위에서 한눈에 파악

국내 유일의 국가통계교육 전문기관

통계교육원
[sti.kostat.go.kr]

통계작성 및 활용 전문통계과정,
기관맞춤형과정, e-러닝 과정

통계청
통계교육원