

국가통계란 정부 기관 같은 국가기관에서 목적에 따른 정보를 얻고자 수치적 정보를 정기적으로 작성하는 것을 일컫는다. 이러한 국가통계는 개인, 기업, 정부 측면에서 그 목적이 다르게 이용된다.

첫째, 개인은 합리적인 사고와 이에 대한 추구를 위한 자료로 국가통계를 이용한다. 예를 들어 오늘 날씨에 따른 옷차림이나 우산 등을 챙길지 판단할 때 기상청의 통계 자료를 참고하여 결정할 때를 생각해볼 수 있다.

둘째, 기업은 이윤 극대화를 위한 활동의 일환으로 시장분석이나 전략 수립, 기획 등의 전반적인 경영활동상 기본자료로 국가통계를 이용한다. 가령, 마케팅을 할 때 시장분석은 필수이다. 이때, 다양한 자료를 활용할 수 있지만, 그중에서도 KOSIS 내 주제별 통계자료나 공공데이터 같은 국가통계가 빈번하게 활용된다. 이를 통해 어떤 활동에 대한 근거를 확보하고 어떤 방향성을 가져야 하는 지 등 전반적인 경영활동의 기반을 정한다고 할 수 있다.

셋째, 정부는 정책 등 국가적인 차원의 업무를 진행할 때 국가통계를 활용한다. 예를 들어, 저출산 사회가 지속됨에 따라 이에 대한 정책을 수립할 때 저출산의 원인이 무엇인지, 어떤 정책이 효과적인지 등을 고려하는 근거로 국가통계를 활용할 수 있을 것이다. 이처럼 국가통계는 사용자 별로 그 목적에 따른 이용이 다양함을 알 수 있다.



통계학은 숫자정보를 요약, 정리하고 이와 관련해 판단하는 방법론에 대한 학문이다. 이러한 통계학의 크게 숫자정보의 수집, 요약/설명, 결론 도출이라는 세 가지 역할을 가진다.

먼저 숫자정보에 대한 수집이 필요하다. 다시 말해 어떤 활동을 위한 기초적인 자료가 필요한데, 통계학은 이러한 자료를 수집하는 방법에 대한 역할을 가진다. 가령, 통계학은 수집은 어떤 방법으로 진행할지, 어떤 기준으로 수집할 지 등을 알려주는 역할을 가진다.

그 후, 이렇게 수집한 데이터를 어떻게 요약하고 설명할 것인지에 대한 방법을 제공하는 것이 통계학의 또 다른 역할이다. 수집된 자료를 바탕으로 데이터의 특성이나 유형에 맞게 어떻게 요약하는 것이 효율적인지, 어떻게 설명하는 것이 효과적인지 등을 제공하는 것이다.

마지막으로 앞선 과정을 거친 자료를 바탕으로 어떤 결론을 도출할 수 있는지, 결론을 도출하는 방법은 무엇인지 과학적인 방법론을 제시하는 것이 통계학의 역할이라고 할 수 있다.



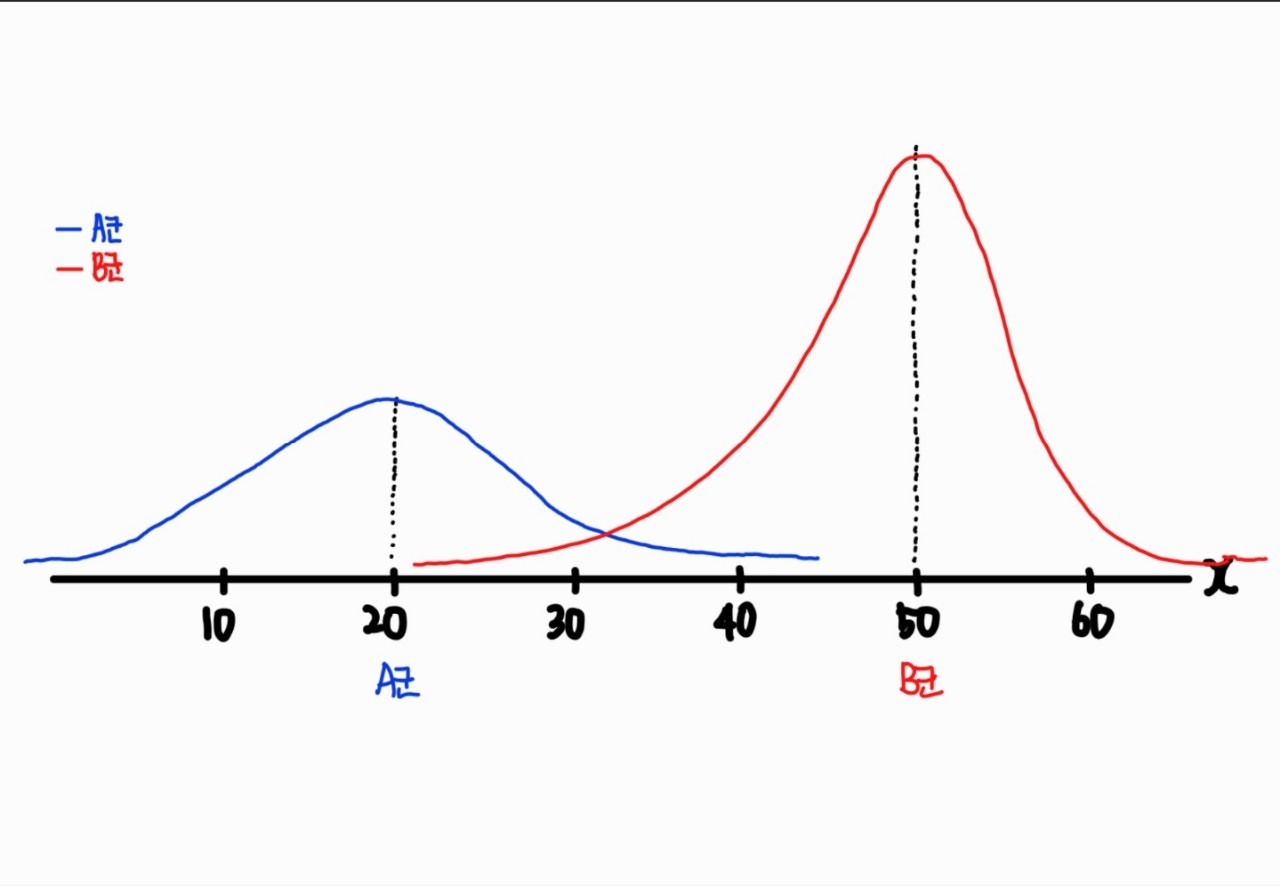


판매기대수 E(X)는 각 판매량 X \* 확률 P(X)를 모두 더한 값으로 이해할 수 있다. 따라서, (0\*0.2)+(1\*0.4)+(2\*0.2)+(3\*0.2) 이다. 이를 정리하면 1.4대가 평균 판매기대수, 즉 E(X) = 1.4이다.





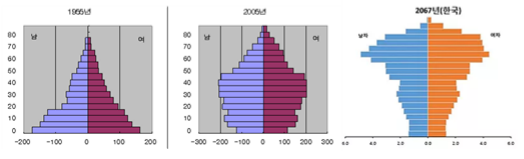
A군과 B군 모수에 의한 변화를 X측 상 분포도로 나타내면 다음과 같다.



먼저, A군의 경우 평균은 20이며, 분산은 15이다. 이는 B군의 평균인 50보다 수직선 위 좌측에 위치하므로 분포도는 상대적으로 좌측에 위치한다. 또한, 분산의 경우도 A군이 15, B군이 5로 A군이 분산이 크다는 것을 알 수 있다. 분산은 퍼진 정도를 나타내는 것으로 분산이 크다는 것은 퍼진 정도가 크다, 즉 흩어진 정도가 크다는 것을 의미한다. 반대로 말하면 분산이 작을수록 값들의 집중도가 커져 첨도가 비교적 높아짐을 의미한다.

따라서 B군과 비교했을 때, A군의 모수 평균은 낮고 분산은 크기 때문에 분포도는 좌측에, 첨도는 낮게 그려진다. 반면 A군과 비교했을 때, B군의 모수 평균은 높고 분산은 작기 때문에 분포도는 우측에, 첨도는 높게 그려진다. 이처럼 정규분포의 모수에 의한 변화는 관측치와 분포도를 통해 직관적으로 파악이 가능하다.





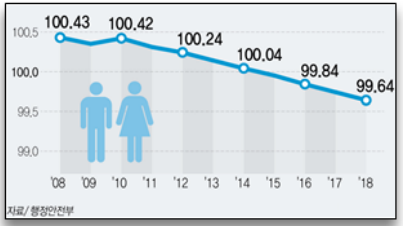
해당 인구피라미드는 1966년, 2005년, 2067년의 대한민국의 인구 구성과 예측을 나타내고 있다. 먼저, 1966년 인구피라미드를 보면, 연령별 구조가 흔히 생각하는 피라미드의 형태를 띠고 있다. 이는 후진국형 인구 구조라고도 말하는데, 영유아와 청소년 연령이 전체 인구의 대부분을 차지하는 모습을 가진다.

인구 통계(예측)는 일명 코호트 예측이라고도 하는 만큼 가령 몇 년 전의 10대 인구가 몇 년 후에는 2~30대가 되는 등 비교적 정확한 예측을 나타낸다. 이에 2005년 인구피라미드를 보면 1966년의 영유아 및 청소년 인구들이 4~50대로 변화하여 해당 연령대에 집중적으로 분포한 일명 항아리형의 분포를 보인다. 2005년에는 비교적 생산가능인구에 인구가 집중된 모습을 보이는데 이는 베이비붐 세대, 국가의 가족 정책 등이 영향을 미쳤다고 볼 수 있다.

하지만 여기서 한 가지 더 주목해야 할 점은 영유아와 청소년 인구의 감소에 있다. 2005년 인구피라미드에서 이러한 모습을 보인다는 것은 결국 이후의 인구 구조는 점차 고령층이 늘어나고 출생아 수는 감소한다는 예측을 가능하게 한다. 또한 생산가능인구가 줄어들 것도 예측할 수 있다. 이러한 점을 반영한 것이 2067년의 예측 인구피라미드(종형)라고 할 수 있다. 이 시기에는 앞선 두 시기보다도 출생인구나 영유아, 청소년 인구가 확연히 줄어든 모습을 확인할 수 있으며, 생산가능인구도 확연히 줄어 초고령사회 이상으로 진입할 수 있음을 알 수 있다. 이는 2005년의 생산가능인구가 시간이 지남에 따라 연령은 높아졌지만, 출생아 수는 줄어들어 나타난 결과라고 할 수 있다.

만약 이러한 예측대로 인구피라미드가 구성된다면 출산율 반등 정책을 제대로 시행하더라도 가임 인구가 적기 때문에 단기간에 저출산-고령화 구조를 벗어나긴 어려울 것으로 생각한다. 또한, 고령 지원에 대한 정책이 늘어날 것으로 예상하며 이와 관련하여 생산가능인구의 부양 금액도 이전보다 부담스러워질 것으로 생각한다.





출생성비는 여아 100명당 남아의 비율을 나타낸 지표로, 여아 100명당 남아의 수에 100을 곱하여 표현한다. 따라서 2008년의 성비는 100.43으로 여아 100명당 남아는 100.43명이라는 의미이다. 즉, 남아의 수가 여아보다 더 많이 출생했다고 해석할 수 있다. 제시된 통계자료와 같이 짝수 연도의 출생성비를 보면, 점차 출생성비가 내려가고 있음을 알 수 있다. 이는 추후 점차 남성보다는 여성의 인구수가 더 많을 것임을 의미하는데, 이러한 통계자료가 나오는 데에는 여러 가지 원인이 있을 것이다.

첫째, 남아선호에 대한 사회적 분위기의 변화이다. 1990년대까지만 해도 대를 잇는다는 의미에서 남아선호사상이 사회적으로 퍼지면서 남아가 출생할 때까지 아이를 낳는 경우도 많았다. 그러다 보니 출생성비는 높아질 수밖에 없었다. 하지만 현대 사회에 들면서 남아선호에 대한 분위기는 이전에 비해 찾아보기 힘들고, 여아선호에 대한 분위기가 형성되기도 하였다. 이런 점이 출생성비를 낮추는 데 한 원인으로 작용했을 것이다.

둘째, 출산에 대한 인식 변화이다. 최근 합계출산율이 OECD 국가 중 최저라는 기사를 접했다. 이처럼 점차 출산하지 않으려는 기조가 유지되는 가운데, 경제적 요인부터 다양하게 출산에 대한 인식이 점차 변화하고 있다. 그중 여러 아이를 낳기보다는 낳지 않거나 낳아도 한 명만 낳겠다는 인식이 강해지고 있다. 이런 점은 성을 구분하지 않는다는 점에서 출생성비에 영향을 미쳤을 것이다.

셋째, 그래프처럼 성비만 봤을 때, 여성의 기대수명이 남성보다 길다는 등의 이유로 전체적인 성비는 갈수록 줄어들고 있음을 알 수 있다.

이 외에도 여성의 경제활동 증가, 2030 혼인율 감소 등의 다양한 원인으로 인해 출산성비와 성비는 감소하는 상황이 만들어졌다고 생각한다. 통계청 자료에 따르면, 이러한 상황이 지속되어 2029년에는 여초 사회가 시작되어 꾸준히 성비는 떨어질 것으로 예상하기도 한다 (중앙일보, 2021)

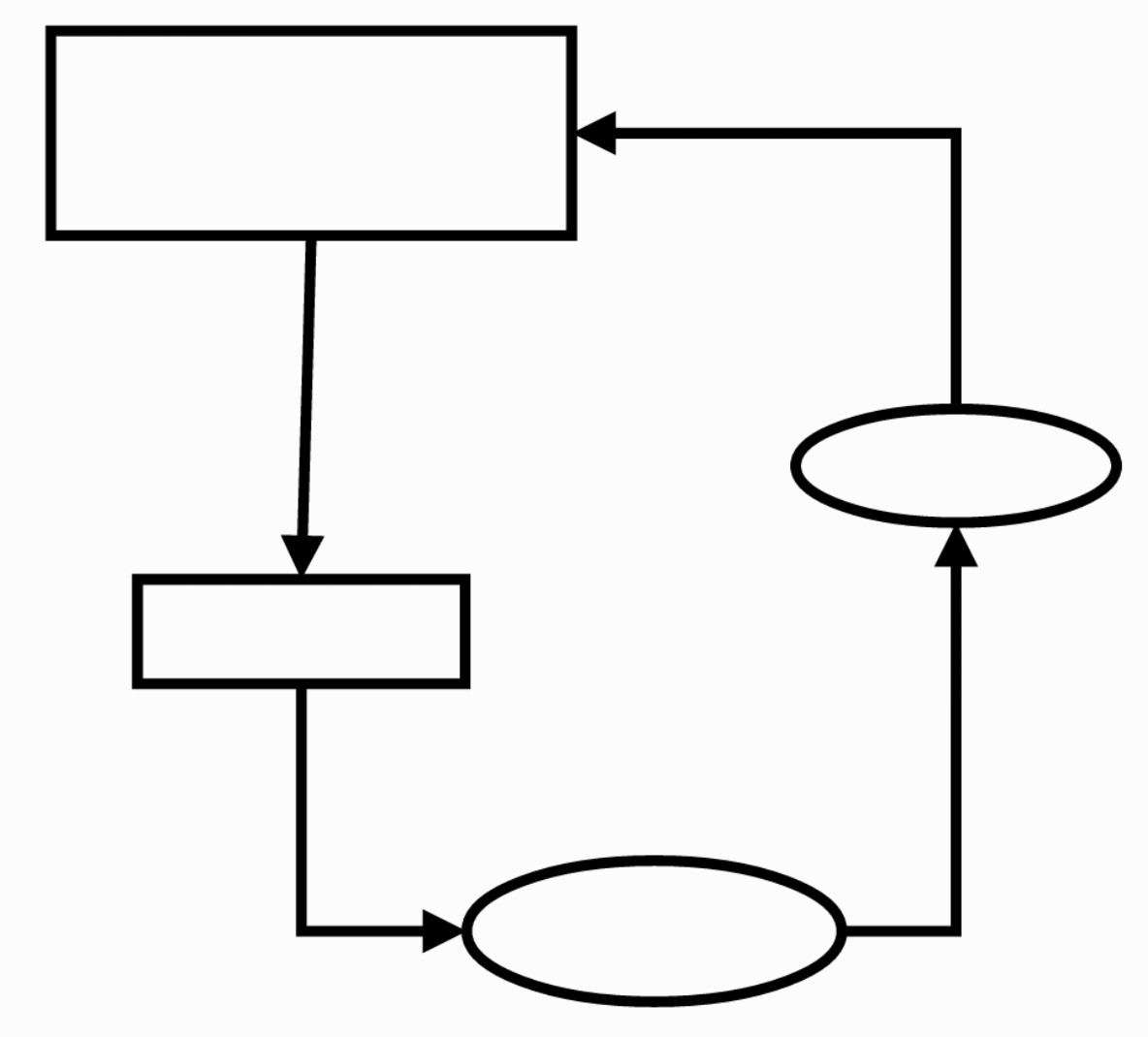
이상치는 자료의 전처리 과정에서 발생할 수 있는 문제로 자료의 전체적인 패턴에서 벗어난 값을 말한다. 이러한 이상치가 있으면 정확한 전처리가 이뤄지지 않아 통계 분석에 있어 잘못된 의사 결정을 할 수 있는 문제가 생긴다. 이러한 이상치를 탐색하기 위해서는 시각화, Z-Score, DBSCAN, 고립 의사나무 등의 방법을 사용할 수 있다(나홍석 외. 2022).

먼저, 시각화를 이용하여 이상치를 파악하는 방법에는 박스플롯, 줄기-잎 그림, 산점도, 히스토그램 등을 활용할 수 있다. 이러한 방법은 시각화를 진행하였을 때 직관적으로 어떤 부분에 이상치가 있는지 파악할 수 있기 때문에 가장 활용하기 좋은 방법이라고 생각한다.

또한, Z-Score를 활용할 수 있다. 정규화를 통해 Z-Score를 산출하고 threshold를 1 표준편차, 2 표준편차, 3 표준편차 등 각 표준편차 간 범위를 설정하여 이를 벗어나는 값은 이상치로 판단하여 검토할 수 있다.

이 외에도 비모수적 다변량의 경우 군집간 밀도를 이용하여 특정 거리 내 데이터 수가 지정 개수 이상이면 군집으로 정의하는 DBSCAN(밀도기반 클러스터링), 의사결정나무 기반의 정상치의 단말 노드(Terminal Node)보다 이상치의 노드에 이르는 길이가 더 짧은 성질을 이용하는 방법인 고립 의사나무 방법 등을 이용하여 이상치를 파악할 수 있다.



**모집단**

**표본**

**임의추출**

**표본자료분석**

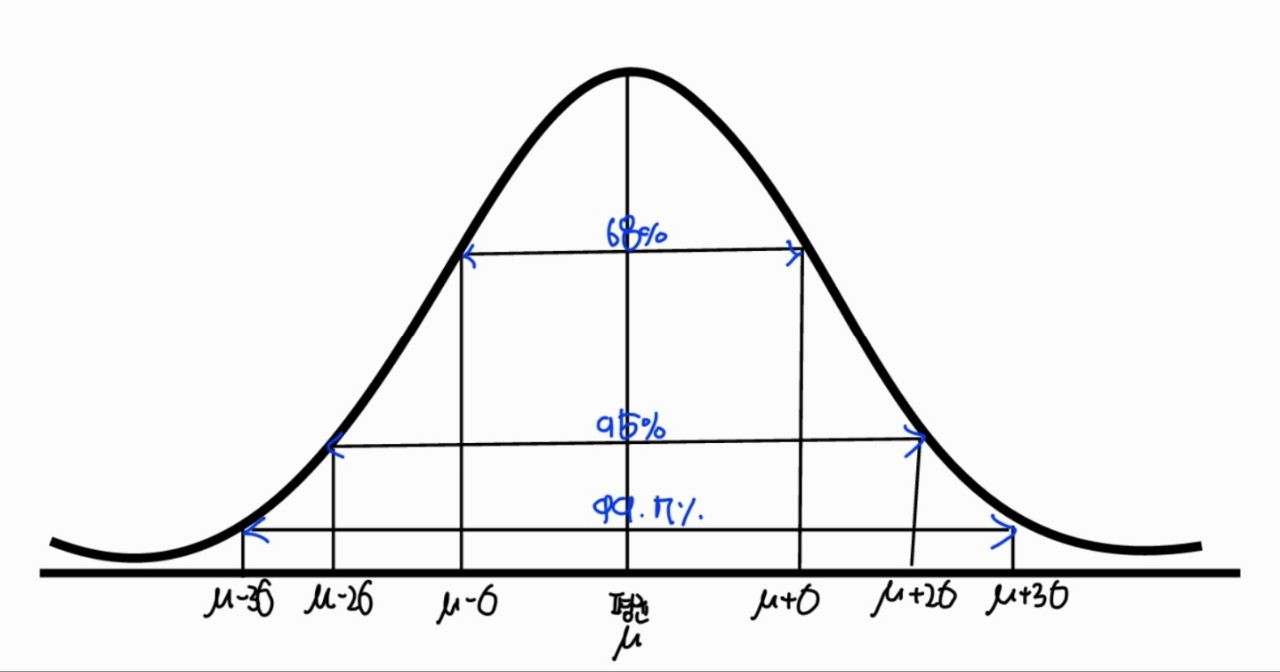
**모집단특성추정**

모집단과 표본의 관계를 도표로 나타내면 위와 같다. 모집단은 ‘조사 대상이 되는 전체’를 의미하고, 표본은 ‘모집단의 정보를 얻기 위한 조사 대상으로 선정된 모집단 일부’를 의미한다. 이러한 정의에서 표본은 모집단에 포함되는 것을 알 수 있으며, 임의추출을 통해 뽑힌 표본은 표본자료분석을 통해 모집단의 특성을 추정하는 단계를 거쳐 모집단을 대략 가늠할 수 있게 한다.

이때, 모집단의 유형에 따라 표본의 크기도 다양해지는 특징을 가진다. 가령, 대통령 선거의 예측조사는 표본을 1,500명으로 설정하지만, 인구총조사는 대한민국 인구 전체를 조사한다든지, 신제품 과자의 맛은 하나만 맛봐도 모집단의 특성을 알 수 있는 등 목적에 따라 표본의 크기는 달라진다.

표본의 크기만큼이나 중요한 요소는 표본을 뽑는 방법인데, 이중 대표적인 것은 임의추출이 있다. 임의추출은 추정의 타당성과 신뢰성을 높여주는 요소라고 생각한다. 임의추출은 표본추출 과정에서 주관이 개입되지 않고 객관적인 특성을 유지하게 해주며, 이로써 편향 없이 전체를 대표할 수 있는 표본을 만들어 준다. 만약 임의추출이 아닌 방법으로 추출한다면, 한쪽으로 치우진 결과와 분석으로 타당성과 신뢰성 없는 잘못된 결과를 초래할 것이다.





위와 같은 분포를 정규분포라고 하며, 정규분포는 수리적 함수를 의미한다. 표본의 수가 증가할수록 자료의 그래프가 점차 평활화되어 정규분포에 가까워진다. 정규분포는 가우스가 평균을 중심으로 좌우 대칭인 오차의 확률 정규분포를 알아냈다는 점에서 가우스 분포라고도 한다. 어떤 사건에 대한 확률분포를 정규분포라고 가정하고 분석하는 경우가 많은 만큼 정규분포는 통계 분석에 있어 기초적인 사항이라고 인지할 수 있다.

이러한 정규분포의 특성은 그래프의 모양, 평균과 표준편차의 관계와 의미에서 알 수 있다. 먼저, 정규분포의 그래프는 평균을 중심으로 좌우 대칭인 종 모양을 가진다. 이는 정규분포는 평균에 가까울수록 관측치 빈도가 높음을 의미한다. 또한 정규분포는 평균이 어떤 값을 갖는지에 따라 수직선 위 그래프의 중심이 변한다. 그뿐만 아니라 퍼진 정도를 나타내는 표준편차에 따라서 첨도가 결정되는 특징을 가진다. 가령, 평균에서 1 표준편차 이내에는 전체의 68%를 포함하는 자료가 존재하고, 2 표준편차 이내에는 95%를 포함하는 자료가 존재한다.

정규분포는 연속적 확률분포의 한 예로 이산형 확률분포와 달리 그 확률을 구할 때 정규분포 그래프의 넓이 즉, 적분값이 곧 확률값이라는 특징을 가지기도 한다.



그래프는 시각화의 일종이라고 할 수 있는데, 시각화는 자료를 비교적 신속하게 정보화하고 정보 수용자에게 효과적으로 표현 및 전달하는 도구라고 할 수 있다(장영재 외, 2022). 또한 이를 통해 직관적이고 쉬운 이해가 가능하게 하는 장점이 있다. 우리가 데이터를 수집하고, 가공하고, 분석하는 등 데이터 분석에서의 전반적인 과정은 결국 이 데이터를 통해 의사소통하기 위함이라고 생각한다. 어떤 결정을 하거나 의사 표현을 할 때 근거로서 이러한 분석 과정과 결과를 제시하고 설득력을 높이려는 것이 목적이기 때문이다. 이런 측면에서 그래프에 의한 표현은 정보를 압축적이고 효율적으로 전달/표현하는 장점을 가진 최상의 도구라고 생각한다. 이러한 그래프 혹은 시각화 방법에는 크게 시간, 분포, 관계, 비교, 공간에 따라 나눌 수 있다.

먼저, 시간 시각화는 시간에 따른 변화를 나타낸 그래프이다. 즉, 시계열 데이터를 바탕으로 그래프를 그린 것으로 시간에 따른 추세 등을 알 수 있다. 대표적으로 막대 그래프, 누적 막대 그래프, (꺾은)선 그래프 등이 있다.

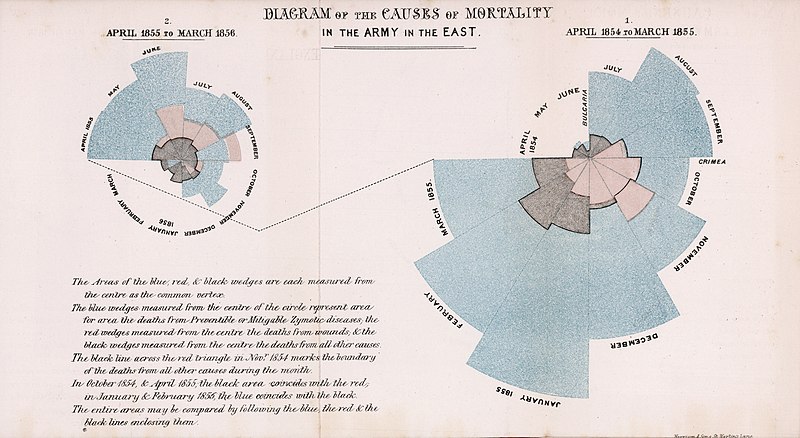
둘째, 분포는 속성별 비중을 나타내며, 그 합이 1이 되는 특징이 있다. 이를 나타낸 것이 분포 시각화인데 특징으로는 비중이나 중요도를 그래프로 그려 비교하기 쉬운 형태로 제공된다는 것이다. 대표적으로 파이차트, 트리맵, 도넛차트, 누적 연속 그래프 등이 있다.

셋째, 다른 두 변수 이상 간 관계를 표현한 관계 시각화가 있다. x축과 y축을 바탕으로 변수 지정 후 관측치들을 좌표평면에 표현 및 상관관계 혹은 관계성을 알고자 할 때 주로 사용하는 그래프이다. 대표적으로 산점도, 산점도 행렬, 버블차트 등이 있다.

넷째, 여러 변수를 비교할 때 세부적인 정보를 모두 추출하여 비교하기 보다 전반에 걸친 특징을 전체적으로 살피기 위한 방법인 비교 시각화가 있다. 대표적으로 히트맵, 체르노프 얼굴, 스타차트 등이 있다.

마지막으로 지도상에 관측치를 표시하여 위치와 변수 관계성을 나타낸 공간 시각화가 있다. 위치정보를 바탕으로 좌표를 특정하여 그 위에 관측치를 표시하는 등의 사례가 있다.

10.

[[1]](#footnote-1)

**<참고자료>**

1. 이태림 외. 2020. “생활과 통계”. 한국방송통신대학교출판문화원

2. 장영재 외. 2022. “데이터과학개론”. 한국방송통신대학교출판문화원

3. 나홍석 외. 2022. “빅데이터 분석기사 실기”. 영진닷컴. 46-47p

4. 김기환. 2021. “남아 성비 사상 최저...사라지는 ‘남아 선호’”. 중앙일보

1. Diagram of the causes of mortality in the army in the east, https://en.m.wikipedia.org/wiki/File:Nightingale-mortality.jpg [↑](#footnote-ref-1)