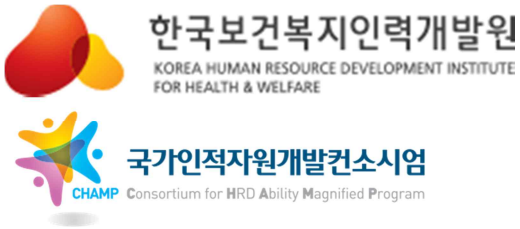


의약품 빅데이터 분석 과정

- 5 차시 -

빅데이터 시각화

5차시. 빅데이터 시각화

• 학습목표

1. 빅데이터와 인공지능을 설명할 수 있다.
2. 빅데이터 시각화 기술을 설명할 수 있다.

• 학습하기

1. 빅데이터와 인공지능

■ 인공지능의 개념

인공 지능이라는 개념은 언제 처음 사용되었을까요? 그 처음은 존 매카시라는 교수가 주최한 다트머스 회의에서였다. 1956년 미국 다트머스 대학의 매카시 교수는 인공지능이라는 분야를 확립하고 설립한 학술회의인 다트머스 회의를 개최했는데요. 이 회의에서 인공지능이라는 용어가 처음 등장했다고 한다. 용어의 등장 이후 인공지능 분야는 점차 발전되어 오기 시작한다.

■ 인공지능의 발전

인공지능의 발전은 점점 가속화되는데요. 크게 영향을 끼친 것으로 3가지 정도를 꼽을 수 있다. GPU의 도입이다. 2015년 이후 개발된 GPU는 데이터를 강력하고 신속하고 병렬로 처리하는 기능을 제공한다. 또, 데이터를 저장하는 공간의 용량이 갈수록 폭발적으로 증가한 것도 그 원인이 된다. GPU의 개발과 데이터 저장용량의 증가는 이미지와 텍스트, 매핑 데이터 등 모든 영역의 데이터를 범람하게 만들었는데요. 이와 같은 빅데이터 시대의 도래도 인공지능의 발전에 영향을 미치고 있다.

■ 인공지능

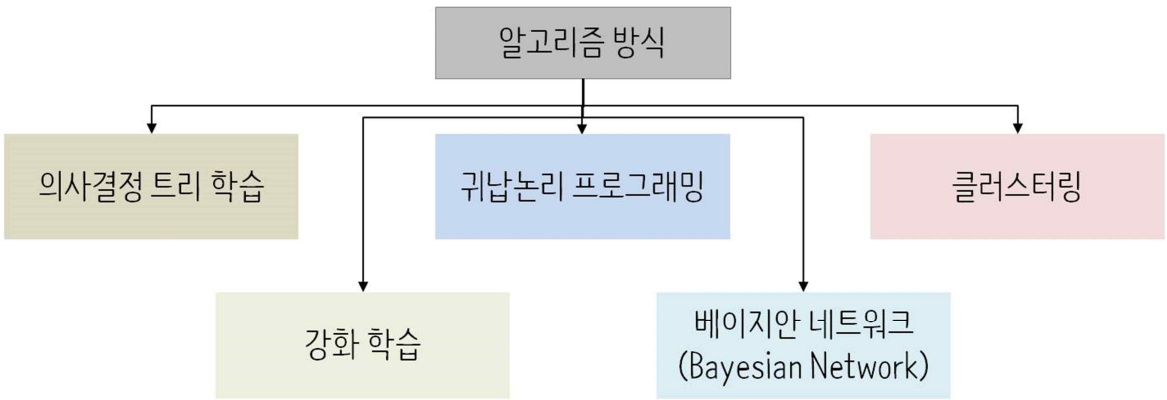
1956년 당시 다트머스 회의에 참석했던 과학자들은 인공지능에 대해 이렇게 꿈꿨습니다. 바로 인간의 지능과 유사한 특성을 가진 복잡한 컴퓨터를 제작하는 것으로 말이다. 이와 같이 인간의 감각과 사고력을 지닌 채 마치 인간처럼 생각하고 행동하는 인공지능을 일반 인공지능, 제너럴 AI라고 한다.

그런데 이와 같은 일반 인공지능은 현재의 기술 수준으로 구현하기는 아직 어려움이 있다. 그래서 현재의 기술 발전 수준에서 만들 수 있는 인공지능을 따로 '좁은 AI(Narrow AI)'의 개념에 포함시켰다. 좁은 AI는 일반 AI처럼 고기능을 할 수는 없지만 소셜 미디어

의 이미지 분류 서비스나 얼굴 인식 기능 등과 같이 특정 작업을 인간 이상의 능력으로 해낼 수 있는 것이 특징이다.

■ 기계학습(Machine Learning)

다음은 기계학습에 대해 알아보겠습니다. 머신 러닝, 즉 기계 학습은 기본적으로 알고리즘을 이용해 데이터를 분석하고, 분석을 통해 학습하며, 학습한 내용을 기반으로 판단이나 예측을 한다. 따라서 궁극적으로는 의사 결정 기준에 대한 구체적인 지침을 소프트웨어에 직접 코딩해 넣는 것이 아닌, 빅데이터와 알고리즘을 통해 컴퓨터 그 자체를 '학습' 시켜 작업 수행 방법을 익히는 것을 목표로 하고 있다.



기계학습은 초기 인공 지능 연구자들이 직접 제창한 개념에서 나온 것이며, 알고리즘 방식에는 의사 결정 트리 학습, 귀납 논리 프로그래밍, 클러스터링, 강화 학습, 베이지안(Bayesian) 네트워크 등이 포함된다. 그러나 이 중 어느 것도 최종 목표라 할 수 있는 일반 AI를 달성하진 못했으며, 초기의 기계학습 접근 방식으로는 좁은 AI조차 완성하기 어려운 경우도 많다.

■ 기계학습의 한계

현재 기계학습은 컴퓨터 비전 등의 분야에서 큰 성과를 이뤄내고 있으나 그 한계가 분명히 있다. 바로 구체적인 지침이 아니더라도 인공 지능을 구현하는 과정 전반에 일정량의 코딩 작업이 수반된다는 점이다.

예를 들어보겠습니다. 기계학습 시스템을 기반으로 정지 표지판의 이미지를 인식할 경우, 개발자는 물체의 시작과 끝 부분을 프로그램으로 식별하는 경계 감지 필터, 물체의 면을 확인하는 형상 감지, 'S-T-O-P(에스티오피)'와 같은 문자를 인식하는 분류기 등을 직접 코딩으로 제작해야 한다.

이처럼 기계학습은 '코딩'된 분류기로부터 이미지를 인식하고, 알고리즘을 통해 정지 표

지판을 '학습'하는 방식으로 작동된다.

■ 기계학습의 이미지 인식률

기계학습의 이미지 인식률은 상용화하기에 충분한 성능을 구현하지만, 안개가 끼거나 나무에 가려서 표지판이 잘 보이지 않는 특정 상황에서는 이미지 인식률이 떨어지기도 한다. 최근까지 컴퓨터 비전과 이미지 인식이 인간의 수준으로 올라오지 못한 이유는 이 같은 인식률 문제와 잦은 오류 때문이다.

■ 인공신경망(Artificial Neural Network)

초기 기계학습 연구자들이 만들어 낸 또 다른 알고리즘인 인공 신경망(artificial neural network)에 영감을 준 것은 인간의 뇌가 지닌 생물학적 특성, 특히 뉴런의 연결 구조였다. 그러나 물리적으로 근접한 어떤 뉴런이든 상호 연결이 가능한 뇌와는 달리, 인공 신경망은 레이어 연결 및 데이터 전파 방향이 일정하다. 예를 들어, 이미지를 수많은 타일로 잘라 신경망의 첫 번째 레이어에 입력하면, 그 뉴런들은 데이터를 다음 레이어로 전달하는 과정을 마지막 레이어에서 최종 출력이 생성될 때까지 반복한다. 그리고 각 뉴런에는 수행하는 작업을 기준으로 입력의 정확도를 나타내는 가중치가 할당되며, 그 후 가중치를 모두 합산해 최종 출력이 결정된다.

정지 표지판의 경우, 팔각형 모양, 붉은 색상, 표시 문자, 크기, 움직임 여부 등 그 이미지의 특성이 잘게 잘려 뉴런에서 '검사'되며, 신경망의 임무는 이것이 정지 표지판인지 여부를 식별하는 것이다. 여기서는 충분한 데이터를 바탕으로 가중치에 따라 결과를 예측하는 '확률 벡터(probability vector)'가 활용된다.

■ 딥러닝(Deep Learning)

딥러닝은 인공신경망에서 발전한 형태의 인공 지능으로, 뇌의 뉴런과 유사한 정보 입출력 계층을 활용해 데이터를 학습한다. 그러나 기본적인 신경망조차 굉장한 양의 연산을

필요로 하는 탓에 딥러닝의 상용화는 초기부터 난관에 부딪혔다.

■ 딥러닝의 발전

그럼에도 토론토대의 제프리 힌튼(Geoffrey Hinton) 교수 연구팀과 같은 일부 기관에서는 연구를 지속했고, 슈퍼컴퓨터를 기반으로 딥러닝 개념을 증명하는 알고리즘을 병렬화하는데 성공했다. 그리고 병렬 연산에 최적화된 GPU의 등장은 신경망의 연산 속도를 획기적으로 가속하며 진정한 딥러닝 기반 인공 지능의 등장을 불러왔다.

■ 신경망 네트워크의 학습 과정

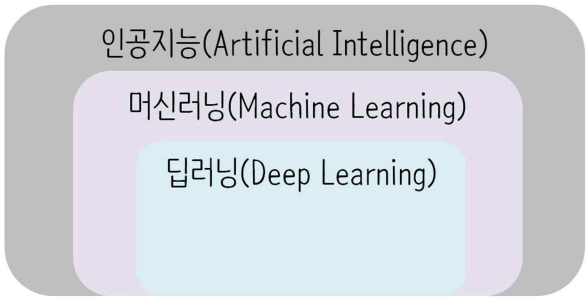
신경망 네트워크는 '학습' 과정에서 수많은 오답을 낼 가능성이 크다. 정지 표지판의 예로 돌아가서, 기상 상태, 밤낮의 변화에 관계없이 항상 정답을 낼 수 있을 정도로 정밀하게 뉴런 입력의 가중치를 조정하려면 수백, 수천, 어쩌면 수백만 개의 이미지를 학습해야 할지도 모른다. 이 정도 수준의 정확도에 이르러서야 신경망이 정지 표지판을 제대로 학습했다고 볼 수 있다.

2012년, 구글과 스탠퍼드대 앤드류 응(Andrew NG) 교수는 1만6,000개의 컴퓨터로 약 10억 개 이상의 신경망으로 이뤄진 '심층신경망(Deep Neural Network)'을 구현했다. 이를 통해 유튜브에서 이미지 1,000만 개를 뽑아 분석한 뒤, 컴퓨터가 사람과 고양이 사진을 분류하도록 하는데 성공했습니다. 컴퓨터가 영상에 나온 고양이의 형태와 생김새를 인식하고 판단하는 과정을 스스로 학습하게 한 것이다.

■ 딥러닝의 현재

딥러닝으로 훈련된 시스템의 이미지 인식 능력은 이미 인간을 앞서고 있다. 이 밖에도 딥러닝의 영역에는 혈액의 암세포, MRI 스캔에서의 종양 식별 능력 등이 포함된다. 구글의 알파고는 바둑의 기초를 배우고, 자신과 같은 AI를 상대로 반복적으로 대국을 벌이는 과정에서 그 신경망을 더욱 강화해 나갔다.

■ 인공지능, 머신러닝, 딥러닝의 관계



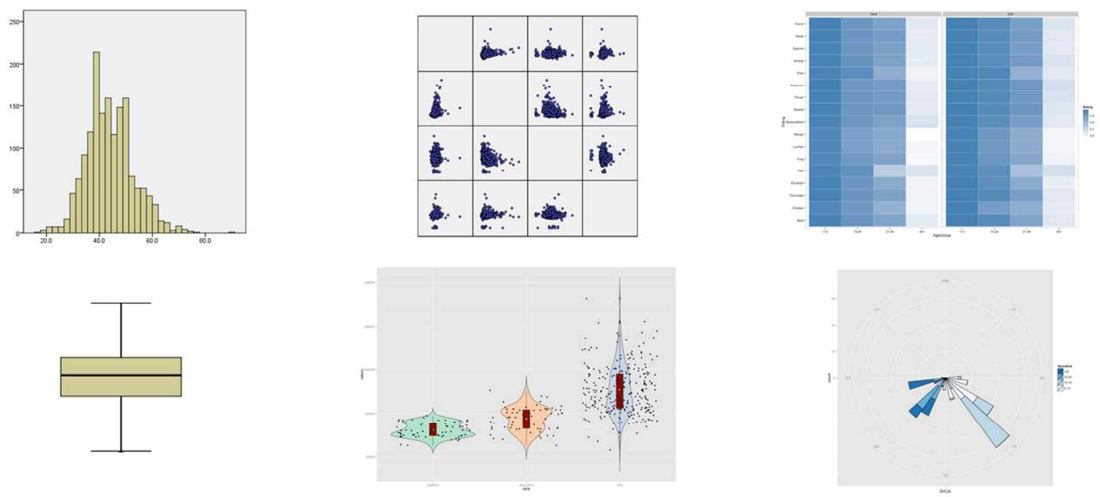
빅데이터를 기반으로 인공지능과 머신러닝, 딥러닝은 모두 밀접한 관련이 있는데요. 인공지능은 머신러닝과 딥러닝을 모두 포함하는 상위 개념이다. 인공지능은 인간의 학습 능력과 추론, 지각, 자연언어 이해 능력 등을 컴퓨터 프로그램으로 실현한 기술로 컴퓨터가 인간의 지능적 행동을 모방할 수 있도록 한 것이다. 인공지능 안에 포함되는 개념이 바로 머신러닝이다. 머신러닝은 인공지능을 가능하게 하는 방법 가운데 하나로 컴퓨터가 데이터를 분석하고 스스로 학습하는 과정을 거치면서 입력하지 않은 정보에 대해서도 판단하고 결정할 수 있게 된다. 이것이 바로 머신러닝이다. 마지막으로 딥러닝은 머신러닝의 하위개념이다. 머신러닝은 스스로 학습하는 컴퓨터이다. 컴퓨터가 사람처럼 생각하고 배울 수 있도록 하는 기술이다.

2. 빅데이터 시각화

■ 그래프를 통한 시각적 표현

빅데이터를 가장 잘 표현하는 것 중의 하나는, 바로 그래프이다. 아무리 방대한 정보도 그래프를 통해 시각화 하면 정보를 보다 빠르고 효율적으로 파악할 수 있다.

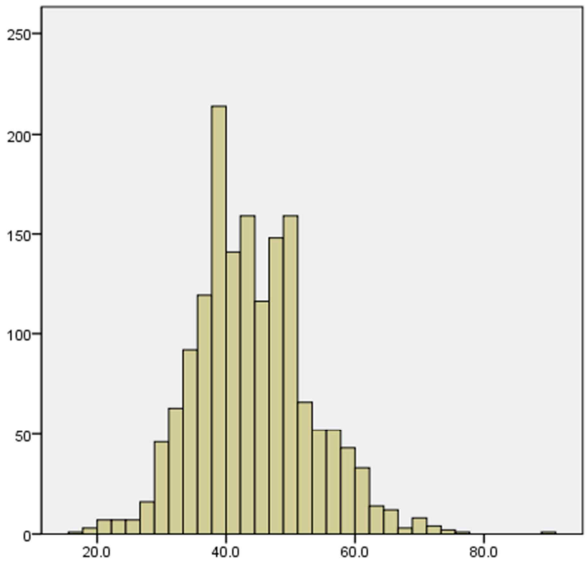
■ 그래프의 종류



그래프의 종류는 매우 많다.

가장 많이 보는 막대그래프, 원그래프, 산점도, 방사형 그래프뿐만 아니라 상자도표, 바이올린 플랏, 히트맵, 포울러 플랏 등 많은 종류의 그래프 등이 있다.

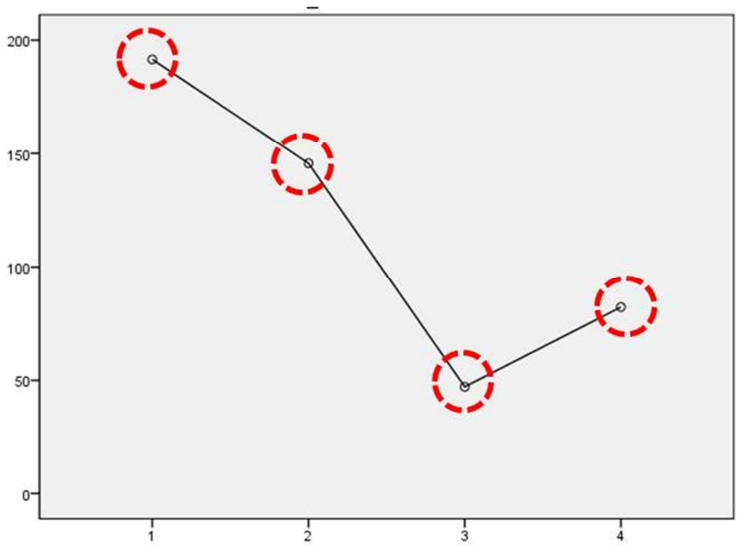
■ 막대 그래프



한 변수의 특성을 나타내는 가장 대표적인 그래프로 막대그래프가 있다. 막대그래프는 데이터의 분포가 전체적으로 어떤 형태인지를 나타내는데 아주 유용하다.

이 그래프를 보시면, 데이터가 대략 좌우대칭 형태 모양을 띄고 있고 가운데에 특히 많은 데이터가 몰려 있다는 것으로 알 수 있다.

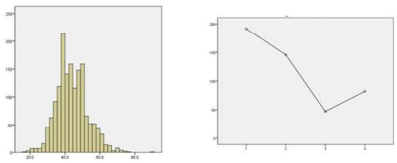
■ 꺾은선 그래프



꺾은선 그래프 역시 연속형 변수에서 매우 자주 사용하는 그래프이다. 막대 그래프가 한

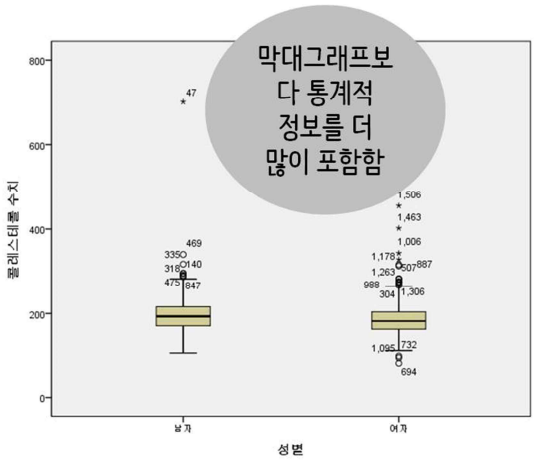
변수에 대한 전반적인 정보, 즉 분포 형태에 대해 알려준다면 꺾은선 그래프는 여러 변수의 비교, 또는 한 변수의 시간의 변화에 대한 추이 등을 표현하는데 매우 유용하다.
이때 꺾은선 그래프는 변수의 평균, 중위수 등의 대푯값만으로 표시를 하게 된다.
따라서 꺾은선 그래프를 어떤 변수가 높고 낮은지에 대한 정보를 파악하는데 도움이 된다.

■ 상자 도표 (Box Plot)



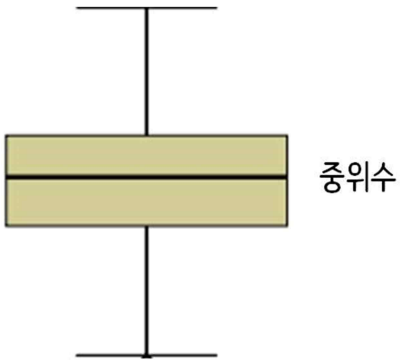
전체 데이터의 형태를 표현

변수의 대푯값을 표시



콜레스테롤, 혈당, 혈압 등과 같이 수치로 측정된 연속형 변수에서 막대그래프, 꺾은선 그래프와 더불어 매우 자주 사용하는 그래프로 상자도표인 Box plot이 있다. 하지만 의외로 상자도표에 대한 이해를 제대로 알지 못하는 경우가 많이 있다.
막대 그래프는 전체 데이터의 형태를 표현한 것이다. 꺾은선 그래프는 변수의 대푯값을 표시하는 그래프이다. 이에 비하여 상자도표는 막대그래프에서 여러 가지 통계적인 정보를 더 많이 포함하고 있는 그래프라고 할 수 있다.

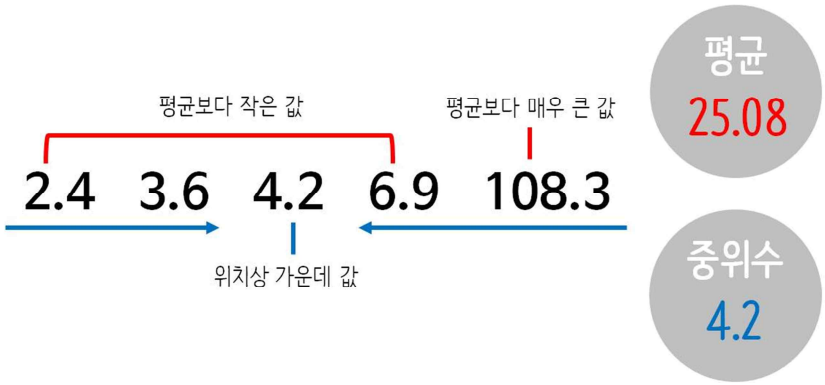
■ 상자 도표 (Box Plot)의 5가지 순서통계량



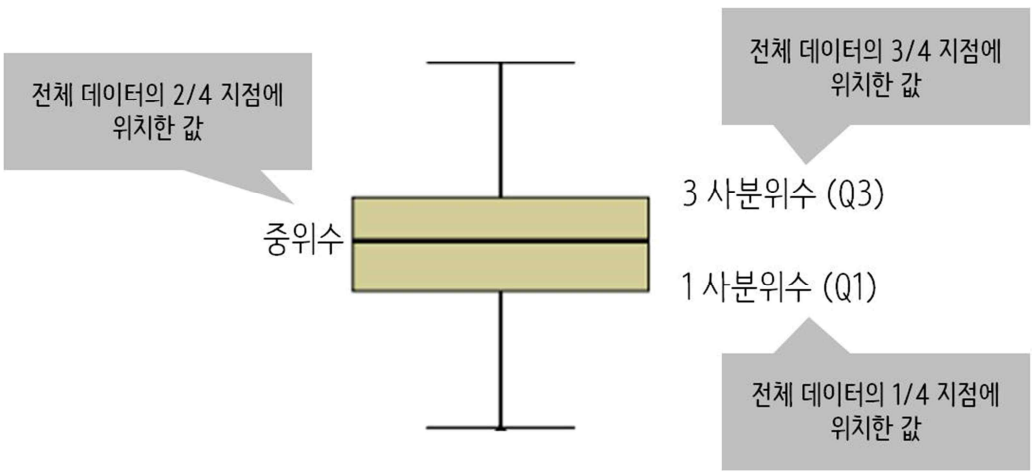
상자 도표를 좀 더 세분해 보면 다음과 같다. 먼저 가운데에 상자가 있고 가운데에 상자를 관통하는 선이 있다. 그리고 상자의 위와 아래에 줄기가 나와 있고, 그 줄기의 끝은

막혀 있는 것으로 알 수 있다. 정 가운데에 상자를 관통하는 선은 중위수이다. 중위수는 평균과 더블어 한 변수의 대푯값이다.

평균은 자료가 좌우 대칭 형태일 때 그 집단을 대표하는 값이다. 하지만 이상값이 있거나 데이터가 한쪽으로 치우친 경우에 평균은 그 집단을 대표하는 것은 무리가 있다. 예를 들어 소득과 같은 경우 평균을 구하게 되면 평균소득이 되는데 소득이 매우 높은 몇몇 사람에 의해서 평균은 실제보다 훨씬 더 높게 나오게 된다. 이런 경우 평균보다는 중위수가 훨씬 더 자료를 잘 표현하는 대푯값이 될 수 있다. 중위수를 구하는 것은 전체 자료를 작은값에서 큰 값으로 크기순으로 정렬한 다음에 그 자료들의 정 가운데 위치하고 있는 값이다.



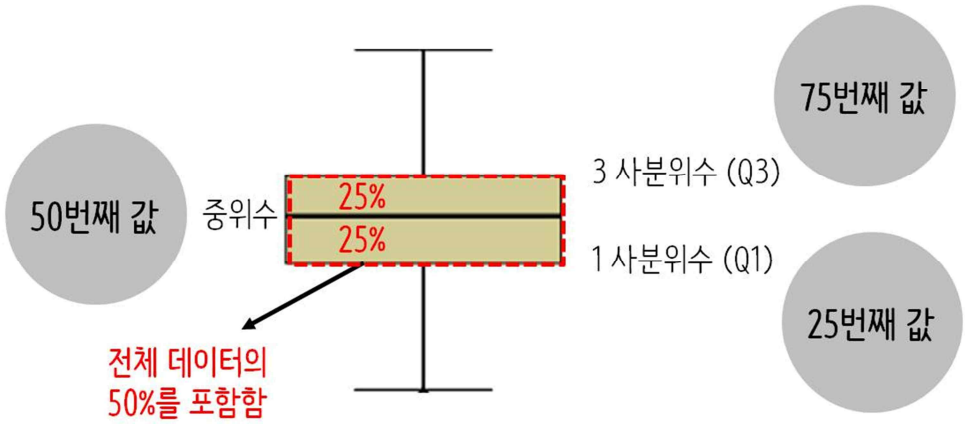
예를 들어 5개의 데이터가 있다고 했을 경우, 이 자료의 평균은 25.08이다. 4개의 데이터는 10보다도 작은 값을 가지는데 비하여 하나의 데이터가 108.3으로 매우 큰 값을 가질 때 평균은 실제보다 훨씬 더 크게 나타납니다. 중위수는 이 5개의 데이터에서 위치상 정 가운데 있는 데이터, 즉 3번째 데이터인 4.2가 된다. 이렇듯 중위수는 데이터가 치우쳐 있다 하더라도 그 집단 또는 변수의 대푯값으로 유용하다.



상자의 위, 아래 경계 부분은 1 사분위수와 3 사분위수라고 하며 Q1, Q3 로 표시한다. 사분위수는 전체 데이터를 4 등분해서 4분의 1 지점에 위치하고 있는 값이 1 사분위수

이고, 4분의 3지점에 위치하고 있는 값이 3 사분위수 있다.

중위수는 정 가운데 위치하므로 4분의 2지점에 위치하고 있으므로 2 사분위수라고 할 수 있다.

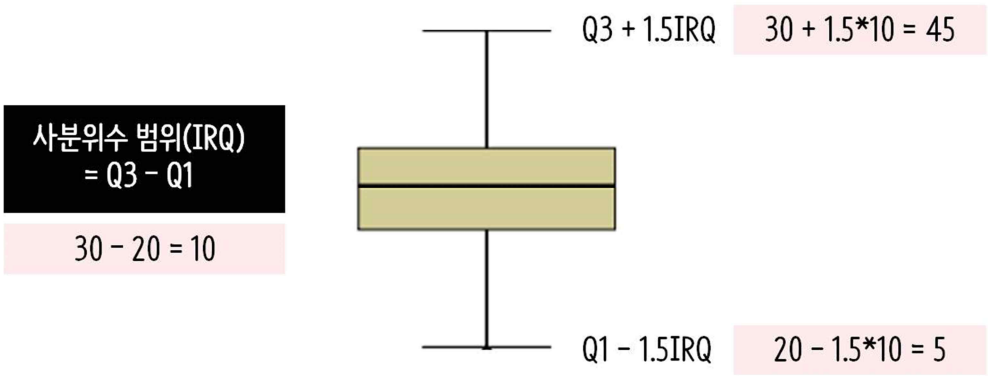


만약 데이터의 수가 99개라면 50번째 위치하고 있는 값은 중위수,

25번째 위치하고 있는 값은 1 사분위수,

75번째 위치하고 있는 값은 3 사분위수가 된다.

그럼 상자도표에서 상자 안에는 전체 데이터의 50%가 포함될 것이다. 또한 상자의 절반에는 각각 25%씩의 데이터가 포함된다.

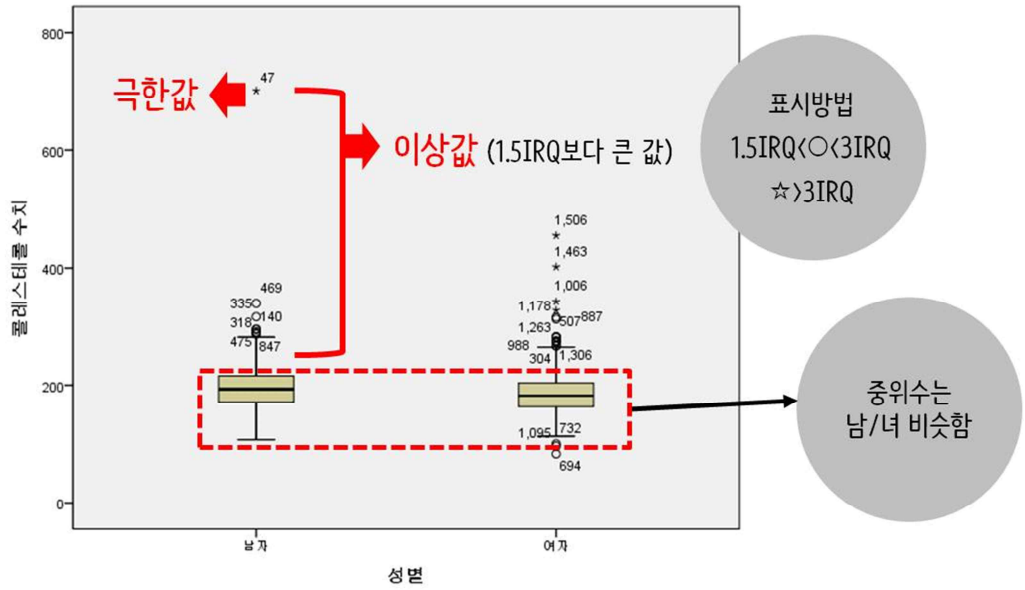


3 사분위수와 1 사분위수의 차이를 사분위수 범위라 하며 IRQ 또는 IQR로 표시한다.

예를 들어 1 사분위수가 20이고 3 사분위수가 30이라면 사분위수 범위 IRQ는 30-20=10이 된다. 상자 밖의 줄기인 위, 아래쪽의 수염은 바로 이 사분위수 범위와 관계가 있다. 아래쪽 수염은 1 사분위수 값에서 IRQ의 1.5 배만큼을 빼준다. 즉 20-1.5(10)=5가 된다. 위쪽 수염은 3 사분위수 값에서 1.5배의 IRQ 만큼 더해주면 30+1.5(10)=45가 된다. 즉 위, 아래 꼬리의 끝에 해당하는 부분의 값은 45와 5가 되죠.

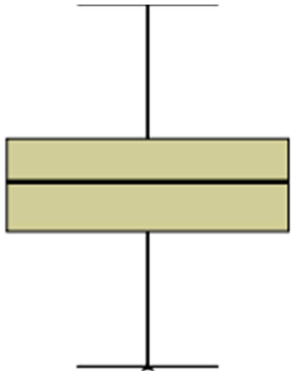
■ 상자도표의 이상값 표현

만약 실제 값이 45에 해당하는 값이 없을 경우에는 그것보다 작은 값에 꼬리 부분의 가로 선이 그려지게 된다. 따라서 이 그래프의 위, 아래의 경계 부분은 최대값과 최소값과는 약간 차이가 있다는 것으로 알 수 있다. 실제값이 44와 50이 있었더다면 이 자료의 최대값은 50 이지만 상자도표에서는 45보다 작은 44가 위 꼬리의 경계선에 그려지는 것이다.

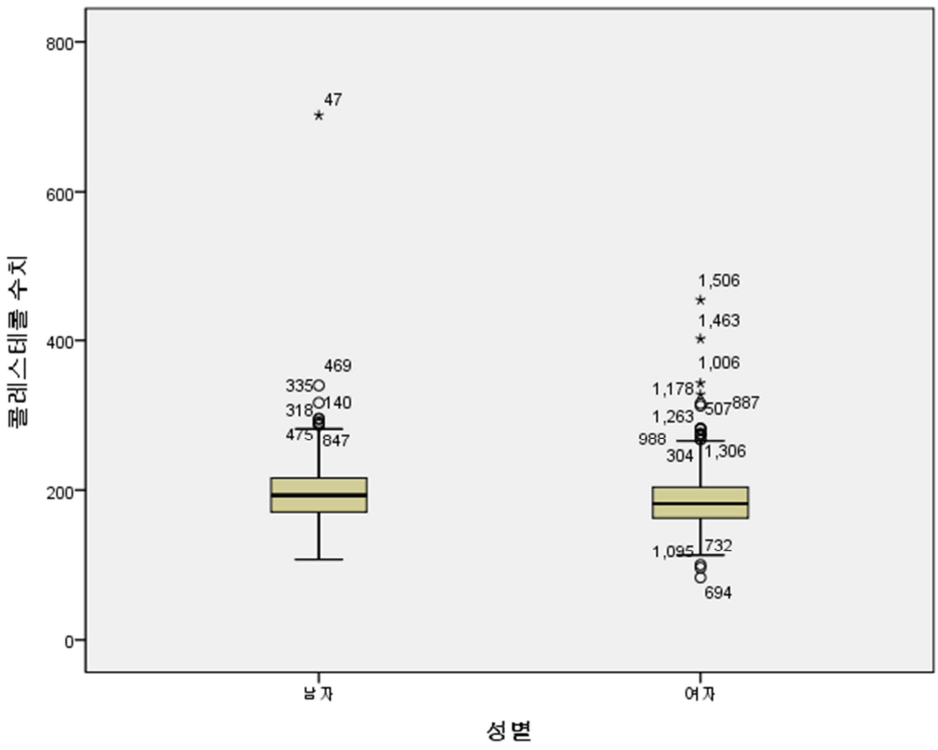


남자와 여자의 콜레스테롤 수치에 대한 상자도표를 그린 그래프이다. 중위수는 남자와 여자가 비슷하다는 것으로 알 수 있다. 그리고 남자의 경우 1.5 IRQ 보다 큰 값들이 표시되어 있으며 원과 별표로 표시된 것으로 알 수 있다. 1.5IRQ 보다 크고 3 IRQ 보다 작으면 원으로, 3IRQ 보다도 크면 별표로 표시한다. 상자도표에서는 이 원과 별을 이상값이라 하며, 별표로 표시된 것으로 특히 극한값이라 한다. 실제 이 극한값은 outlier인 이상값인 경우가 많습니다.

■ 상자도표 보는 법

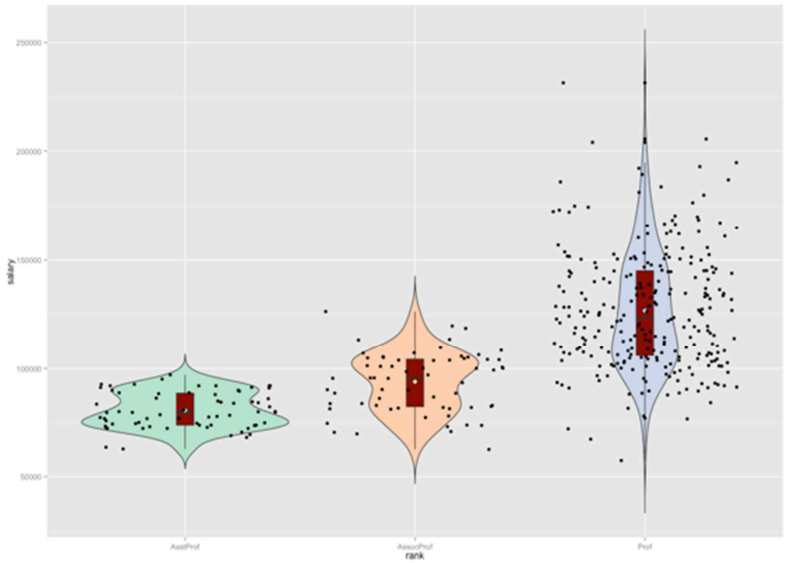


남자와 여자의 콜레스테롤 수치에 대한 상자도표를 그린 그래프이다. 중위수는 남자와 여자가 비슷하다는 것으로 알 수 있다. 그리고 남자의 경우 1.5 IRQ 보다 큰 값들이 표시되어 있으며 원과 별표로 표시된 것으로 알 수 있다. 1.5IRQ 보다 크고 3 IRQ 보다 작으면 원으로, 3IRQ 보다도 크면 별표로 표시한다. 상자도표에서는 이 원과 별을 이상값이라 하며, 별표로 표시된 것으로 특히 극한값이라 한다. 실제 이 극한값은 outlier인 이상값인 경우가 많습니다.



하지만 이와 같다면 좌우대칭 형태이기는 하지만 위로 치우친 값들인 이상값이 있는 분포라는 것이고, 특히 남자의 경우 47번째 환자의 경우 콜레스테롤이 700으로 매우 높은 이상값이라는 것으로 알 수 있다.

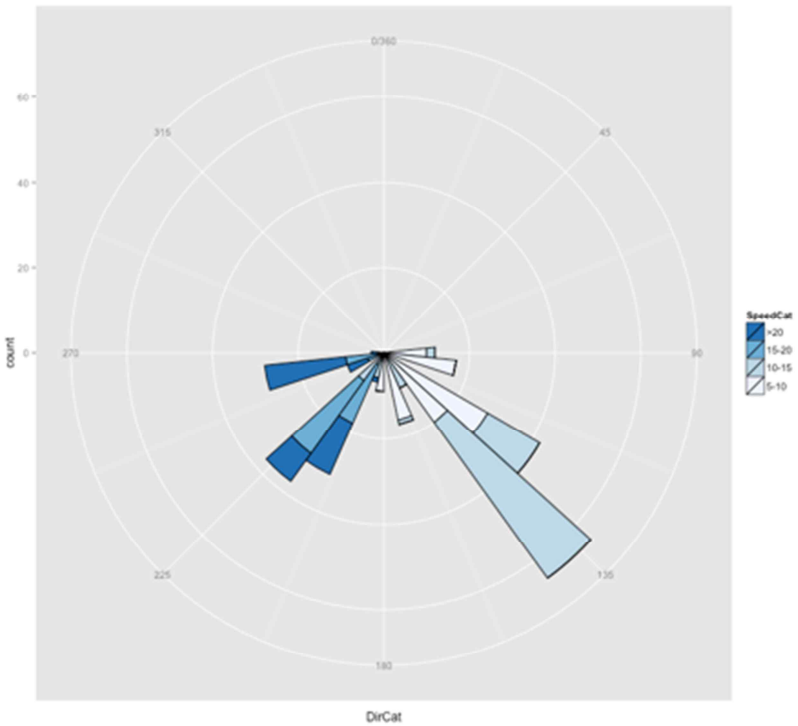
■ 바이올린 plot



요즘 부쩍 많이 보이고 있는 바이올린 플랏은 상자 도표에 밀도를 추가한 그래프로 마치 바이올린처럼 생겼다고 해서 붙인 이름이다.

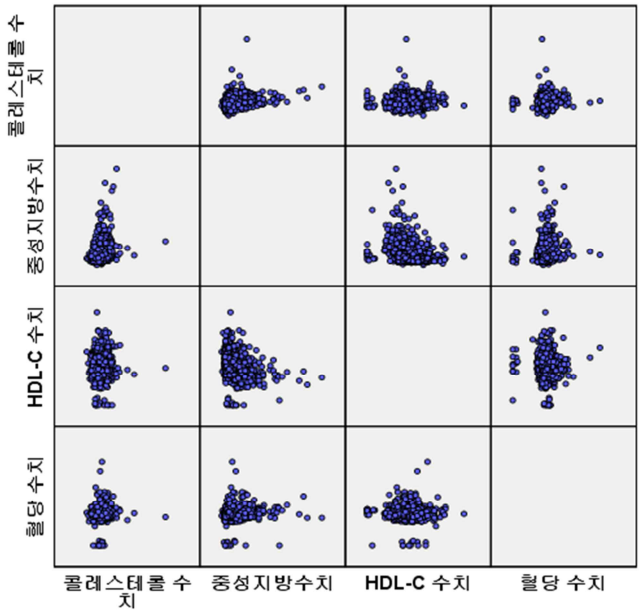
기본적으로 상자도표이며, 이 상자도표에서 산점도와 퍼져 있는 정도까지 같이 표현해주는 장점이 있다. 하지만 기본적으로는 상자도표에 대해서 이해하는 것이 더 중요하다.

■ Polar plot (극도표, 극좌표)

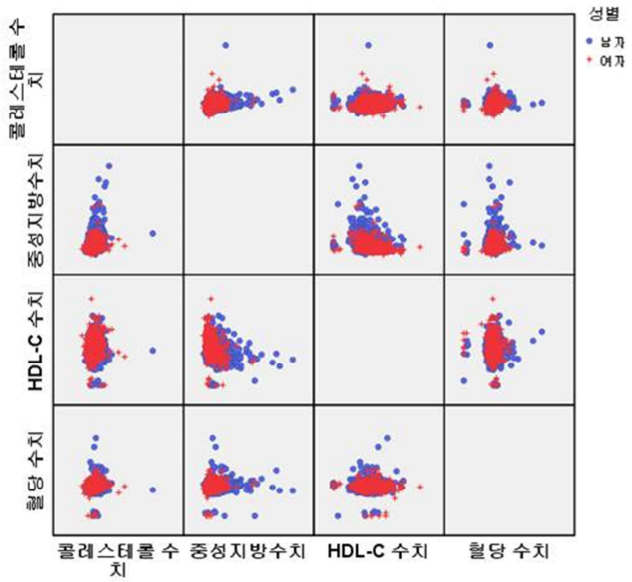


극도표, 극좌표라고 하는 polar plot은 방사형 그래프와 비슷한 그래프이다.
방사형 그래프는 주로 평균값으로 표시하는 데 비하여 polar plot은 막대도표와 결합해서 만들었다는 차이가 있어 좀 더 입체감 있게 표시한 것이다.

■ scatter plot (산점도)



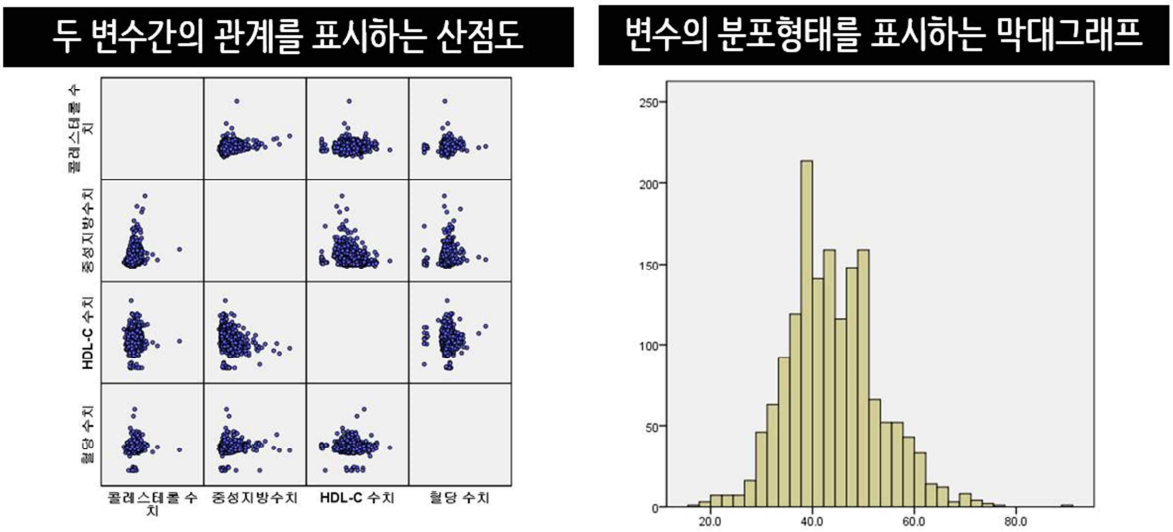
가장 많이 보는 그래프 중에서 산점도라는 scatter plot이 있다.
이 그래프는 연속형 변수들 간의 관계를 x, y 축으로 2차원에 표시하는 그래프로 두 변수간의 관계를 파악하는데 매우 유용하게 사용하는 그래프이다.



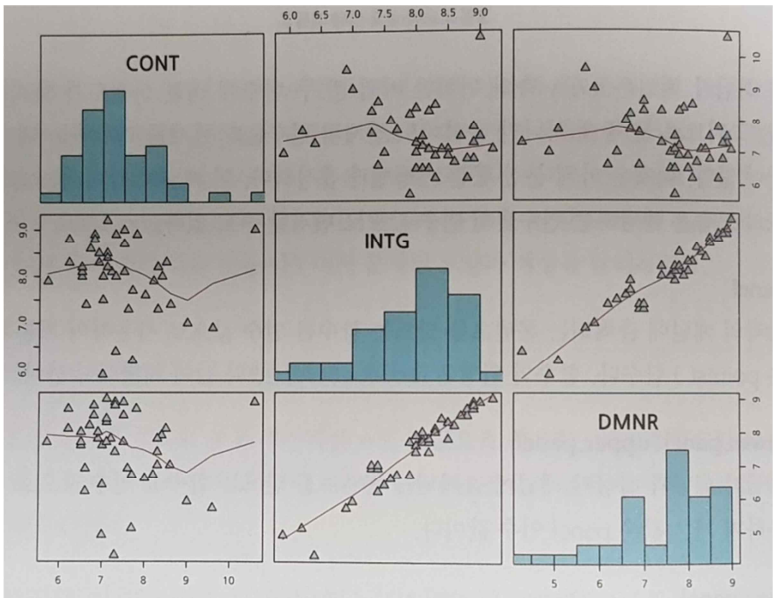
산점도 그래프에서 남자와 여자를 구분하여 표시할 수도 있다. 이렇듯 의미 있는 집단,

예를 들어 성별, 연령대, 음주 여부, 흡연 여부 등과 같이 집단별로 각각의 그래프를 그리기 보다는 하나의 그래프에 동시에 시각화하면 더 많은 그리고 의미 있는 정보를 보여줄 수 있는 그래프가 사용되기도 한다.

■ 두 가지 이상의 그래프를 동시에 표시하기



산점도 그래프는 두 변수 사이의 관계를 표시하는데 비하여 막대그래프는 변수의 분포형태를 파악하는데 도움이 된다. 이렇게 그래프는 한 종류의 그래프만 사용하기도 하지만



두 가지 이상의 형태 그래프를 동시에 사용하기도 한다. 이 그래프에서 정 가운데 있는 INTG 변수에 대한 막대그래프를 보면 오른쪽으로 약간 치우쳐 있다는 것으로 알 수 있

다. 또한 DMNR에 대한 막대그래프도 역시 오른쪽으로 약간 치우쳐 있다. 이제 이 두 변수의 관계인 2번째 줄 3번째 산점도를 보면 두 변수는 선형적인 관계가 매우 강하다는 것으로 알 수 있다. 또한 x 축과 y의 값을 보면 작은 값 보다는 큰 값에서 데이터가 많이 분포하는 것으로 알 수 있다. 이와 같이 막대그래프와 산점도를 동시에 표시하면 데이터만으로는 파악하기 힘든 정보를 파악하는데 많은 도움이 된다.

■ 시각화 도표의 사용

분포를 나타내는 막대그래프, 대푯값으로 비교 설명하는 꺾은선 그래프, 대푯값과 분포를 동시에 고려하는 상자도표, 상자도표에 밀도를 추가한 바이올린 도표, 방사형 도표에 입체감을 효과를 준 polor plot 등 여러 가지 형태의 그래프와 두 연속형 변수간의 관계를 표현하는 산점도 등에 대해서 살펴봤습니다. 그래프는 데이터로 이루어진 복잡한 구조를 단순화하는 효과가 있으며 시각적으로 분포의 형태 등을 직관적으로 표현해 준다는 것으로 알 수 있다.