파워쿼리를 통한 데이터 활용법

● 핵심내용 요약정리

1. 모집단, 전수 조사
2. 표본, 표본 조사
3. 유의수준 (오차한계): 5%
4. 신뢰수준: 95%
5. 가설 (귀무가설, 대립가설)
6. P-Value < 0.05, 대립가설 채택

1종 오류
귀무가설이 참인데 기각하는 경우.

2종 오류
귀무가설이 거짓인데 기각하지 않은 경우.

기술통계의 이해

- 중심을 의미하는 개념 이해하기

1, 2, 2, 3, 3, 3, 4

구분	산술평균(mean)	중앙값(median)	최빈값(mode)
개념	총합을 변수 n개로 나눈 값	변수들을 크기순으로 배열했을 때 중앙에 있는 수	가장 많이 등장한 수
활용Tip!	측정값의 분포가 비슷하거나 정상분포를 이룰 때 적절하게 활용한다.	측정값의 분포가 서열척도이거나 비대칭을 이루고 있을 때는 산술평균이나 최빈수보다 중앙값이 자료의 대표성을 높일 수 있다.	중앙값과 마찬가지로 극단적인 이상값에 영향을 받지 않으므로 데이터가 서열척도이거나 편향되어있을 때 적절하다. 같은 관측치를 나타내는 관찰대상의 규모 등을 파악할 때 사용한다.
활용예시	대한민국 여성 키의 평균	연예인의 평균 연봉 등 아주 극단적인 예외 값이 존재하는 경우	

기술통계의 이해

- 편차, 표준편차, 분산의 이해

1, 2, 3, 4, 5

편차	-2	-1	0	1	2
편차의 제곱	4	1	0	1	4
분산			(4+1+0+1+4) / 5 =2		
표준 편차			√2		

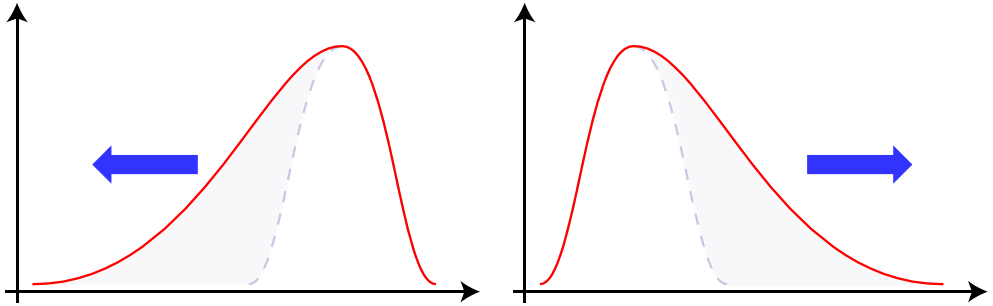
- **편차**는 하나의 데이터 값이 평균에서 얼마나 떨어져 있는지에 대한 값이다.
- **분산**은 '편차'의 제곱의 합을 계산한 값인데, 편차는 음수, 양수를 가질 수 있기 때문에 편차들의 합을 양수화하기 위해 제곱을 사용한 것이다.
- **표준편차**는 분산에서 제곱근(루트)를 씌워준 값으로, 제곱의 합으로 계산한 '분산'의 값이 너무 커서 실제 값과 근사한 오차의 값을 구하기 위해서 사용한 것이다.

시각적 패턴을 통한 이해

왜도	0.648866387

● 왜도, 데이터가 대칭이 아닌 정도로 분포의 특징 이해하기

- **데이터가 대칭을 이룰 수록 왜도 값은 0, 데이터가 한쪽으로 치우칠 수록 양수 또는 음수**
  - 데이터가 정규 분포를 이룰수록 왜도 값이 0에 가까워진다.
  - 분포의 '꼬리'가 왼쪽을 가리키고 왜도 값이 음수인 데이터는 대표적으로 고장률 데이터가 있다.
    _ 대부분의 전구는 오랫동안 켜져 있으며 적은 수의 전구만이 고장난다.
  - 분포의 '꼬리 ' 가 오른쪽을 가리키고 왜도 값이 양수인 데이터는 대표적으로 월급 데이터가 있다.
    - 대부분의 회사는 많은 사원이 비교적 적은 월급을 받지만 점점 더 적은 사람만 많은 월급을 받게 되기 때문이다.



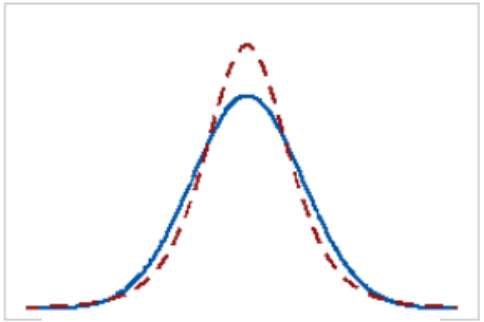
음수 또는 왼쪽으로 치우친 분포

양수 또는 오른쪽으로 치우친 분포

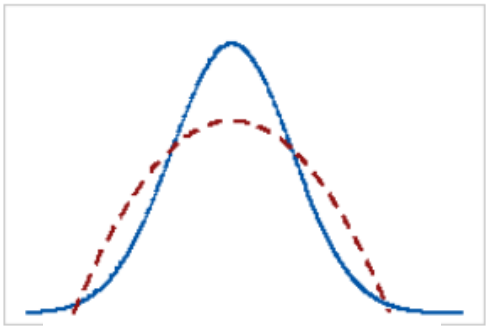
시각적 패턴을 통한 이해

첨도	-0.514163196

- 첨도, 데이터의 분포를 이해하는 또다른 특성
  - **완전히 정규 분포를 따르는 데이터의 첨도값은 0, 데이터의 꼬리의 모습에 따라 양수 또는 음수**
    - 데이터가 완전한 정규 분포를 따르는 첨도의 값은 0이다.
    - 0 이 아닌 첨도값은 데이터가 정규분포로부터 얼마나 벗어나있는지 알게 해준다.
    - 분포의 첨도값이 양수이면 분포의 꼬리가 정규분포보다 두껍다는 것을 나타낸다.
    - 분포의 첨도값이 음수이면 분포의 꼬리가 정규분포보다 얇다는 것을 나타낸다.



양의 첨도



음의 첨도

※ 이미지출처 : Minitab

● 상관계수 알아보기

-1 < 상관계수(p) < 1

-1.0: 완전한 음의 직선 상관관계

-0.8: 강한 음의 상관관계

-0.3: 약한 음의 상관관계

0: 직선 상관관계가 아니다

0.3: 약한 양의 상관관계

0.8: 강한 양의 상관관계

1.0: 완전한 양의 직선 상관관계