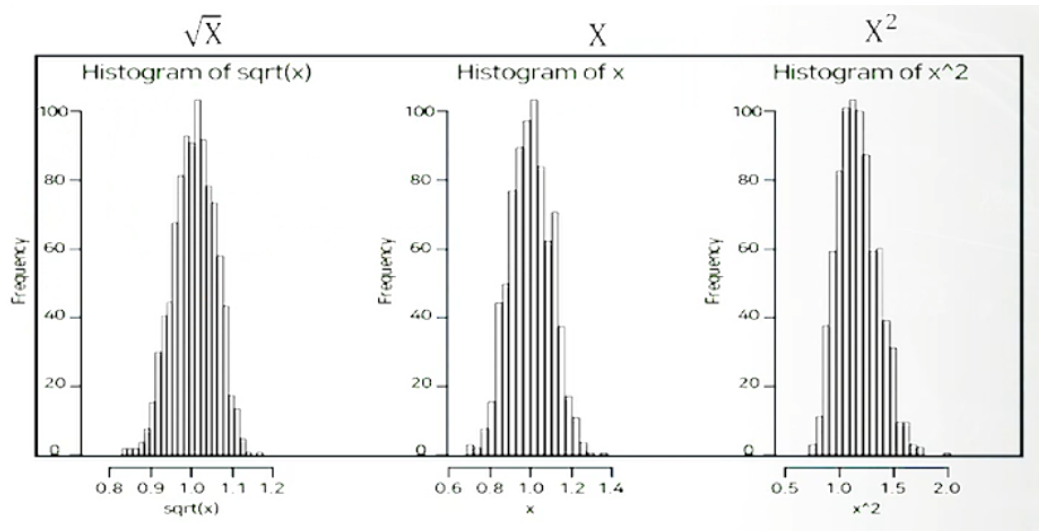


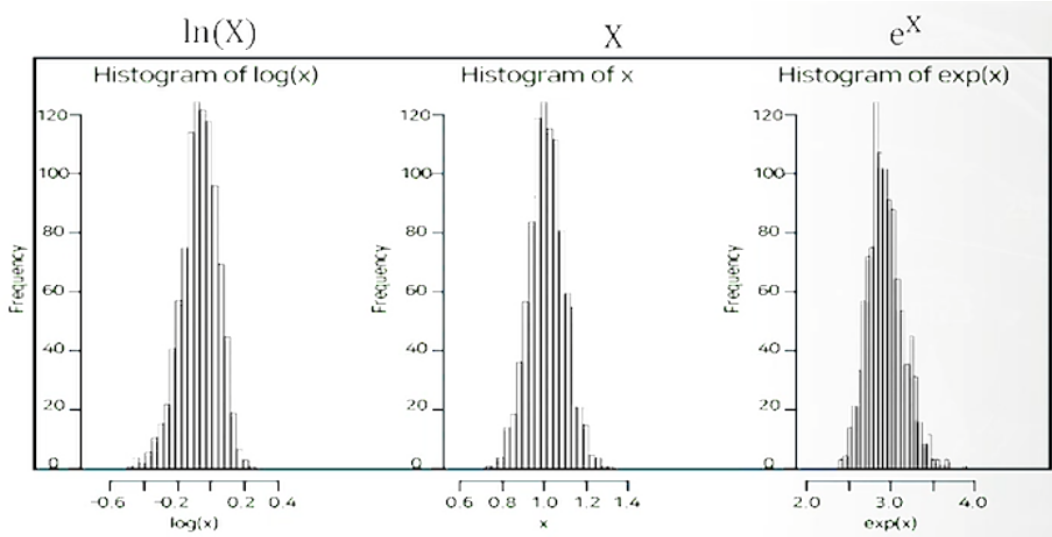
01.데이터 변환, 데이터 결합

- 데이터 변환
 - 자료 변환을 통해 자료의 해석을 쉽고 풍부하게 하기 위한 과정
ex) 로그변환, 지수변환, 제곱근변환
- 데이터 변환 목적
 - 분포의 대칭화
 - 산포를 비슷하게 하기 위함
 - 변수간 관계를 단순하게 하기 위하여
- 데이터 변환
 - 변환 유형 1 : 제곱근 변환 vs 제곱 변환
x값을 기준으로 제곱근 변환시 왼쪽꼬리가 길어진 형태의 분포를 땀
x값을 기준으로 제곱 변환시 오른쪽 꼬리가 길어진 형태의 분포를 땀
따라서, 데이터의 분포에 따라 변환의 형태를 바꿀 수 있음
(단, 둘중 하나만 채택, 병행 불가)



- 변환 유형 2 : 로그 변환 vs 지수 변환
x값을 기준으로 로그 변환시 왼쪽꼬리가 길어진 형태의 분포를 땀

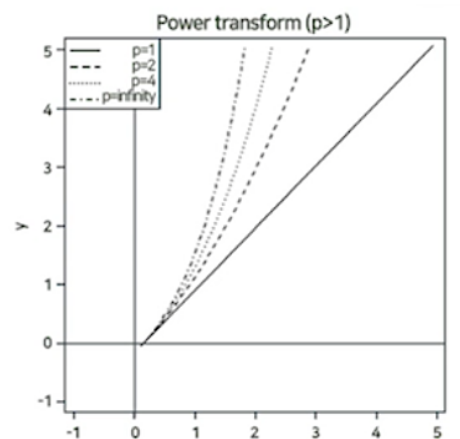
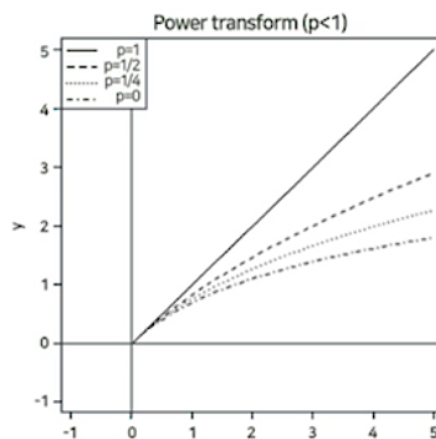
x값을 기준으로 지수 변환시 오른쪽 꼬리가 길어진 형태의 분포를 띠
따라서, 데이터의 분포에 따라 변환의 형태를 바꿀 수 있음
(단, 둘중 하나만 채택, 병행 불가)



- 로그, 지수 변환은 제곱, 제곱근 변환에 비해 더 큰 변화의 그래프가 만들어지므로 어느정도의 강도를 반영해서 설정해줄 필요가 있음.

○ 박스콕스 변환 (Box-Cox Transform)

- $y = \frac{1}{p} ((x + 1)^p - 1)$, $p = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$: 제곱근 유형의 변환을 일반화.
- $y = \ln(x + 1)$, $p = 0$
- $y = (1 + \frac{x}{n})^p - 1$, $p = 2, 4, 8, \dots$: 제곱 유형의 변환을 일반화.



- ($p < 1$) 일 경우, 제곱근 변환의 형태
- ($p > 1$) 일 경우, 제곱 변환의 형태
- ($p = 0$) 일 경우, 로그 변환의 형태
- ($p = \text{무한}$) 일 경우, 지수 변환의 형태

- 데이터 결합

X1	X2			X1	X3	X1 : 키(key) 변수
A	1	+		A	T	
B	2			B	F	
C	3			D	T	

- 이너조인(inner join)
 - 두 테이블에 키(key)가 공통으로 존재하는 레코드(record)만 결합.
 - (A, 1, T), (B, 2, F)
- 풀아우터조인(full outer join)
 - 두 테이블 중 어느 한쪽이라도 존재하는 키에 대한 레코드를 모두 결합
 - (A, 1, T), (B, 2, F), (C, 3, NA), (D, NA, T)
- 레프트 조인(left join)
 - 왼쪽 테이블에 존재하는 키에 대한 레코드를 결합
 - (A, 1, T), (B, 2, F), (C, 3, NA)
- 라이트 조인(right join)
 - 오른쪽 테이블에 존재하는 키에 대한 레코드를 결합
 - (A, 1, T), (B, 2, F), (D, NA, T)