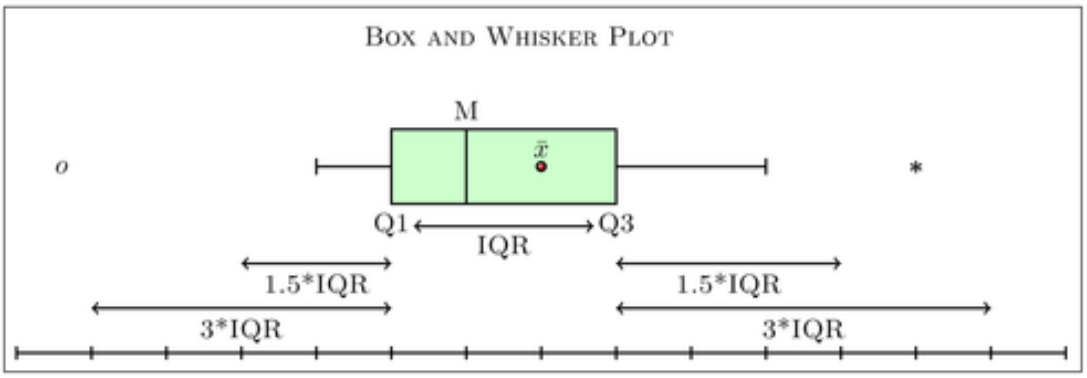
정규변환 Normal Transformation

개념

- 일변량 분석의 기초는 평균에 대한 추론 – 평균은 치우침이나 이상치에 영향을 많이 받으로 므로 평균, 평균 차이 방법론을 적용하기 전에 반드시 치우침과 이상치 진단을 함
- 치우침은 정규성 검정, 이상치 진단은 상자-수염 그림에 의함
- 선형모형에서 확률변수들에 의한 치우침이나 이상치 (회귀모형 관계 속에서는 스튜던트 잔차) 진단을 실시함

이상치 진단

- 상자 수염 그림 BOX- whisker plot
- 5개 기초 순서 통계량 : 최소값-제일사분위 Q1 – 중위값 – Q3 – 최대값



데이터 SMSA.csv

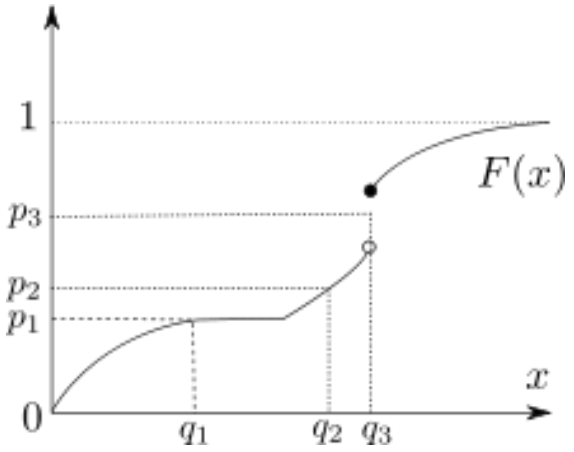
Mortality 사망률지수, Non-White 비백인 비율

city	Mortality	JanTemp	JulyTemp	RelHum	Rain	Educatior	PopDensi	NonWhite
Akron, O	921.87	27	71	59	36	11.4	3243	8.8
Albany-S	997.87	23	72	57	35	11	4281	3.5
Allentowr	962.35	29	74	54	44	9.8	4260	0.8
Atlanta, (	982.29	45	79	56	47	11.1	3125	27.1

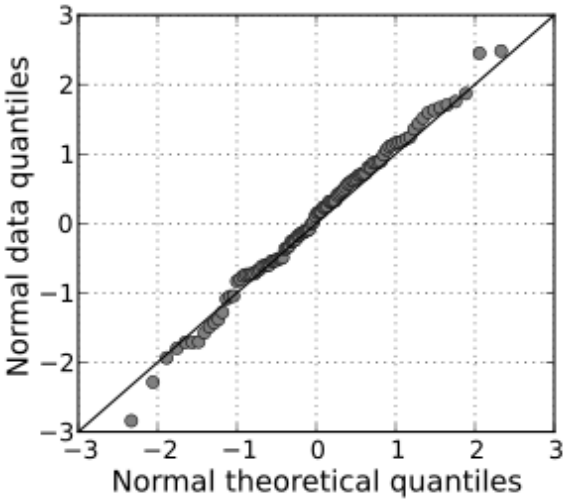
정규성 진단 - 그래프 활용

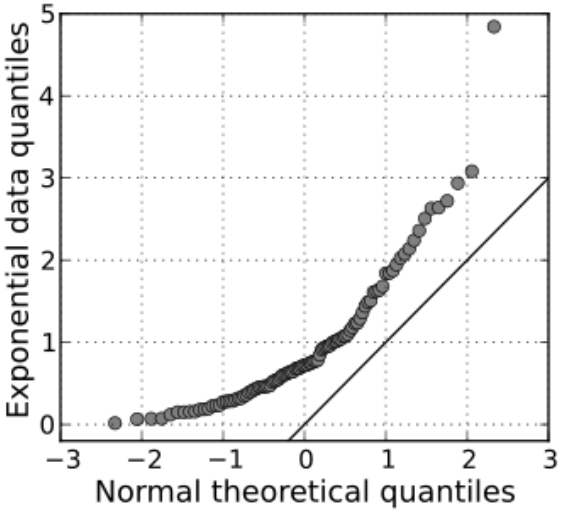
(1) Q-Q plot

- 이론 확률분포 Quantile (x-축)과 데이터 Quantile(y-축) 값을 산점도에 표현
- Quantile 값은 Decile, Quartile, Percentile 값이다. $Q = F^{-1}$



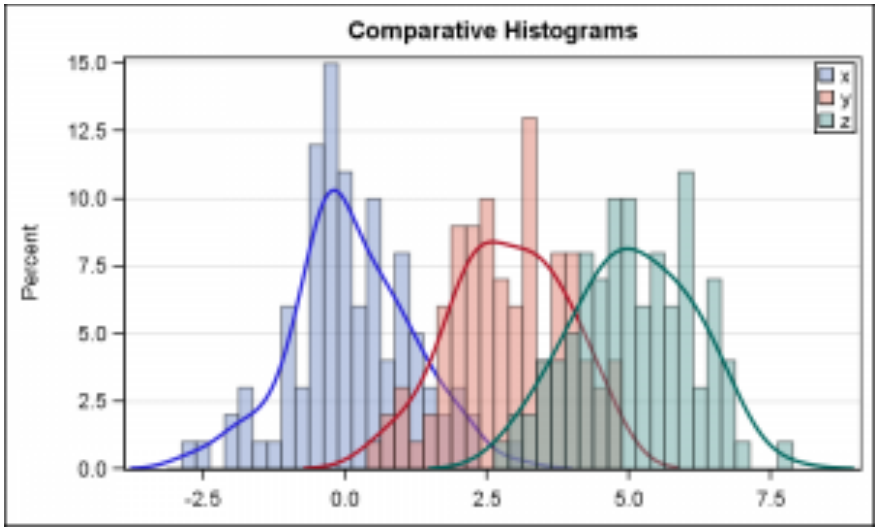
- 이론 분포와 실증적(데이터) 분포가 동일하면 Q-Q 그래프는 직선
- 데이터가 25개인 경우 F(x)=1/25, 2/25, … (roughly) 이것을 만족하는 x값이 Quantile





(2) 히스토그램

• 히스트그램에서 중위값과 평균이 일치 – 시각적 판단



정규성 진단 - 통계량 활용 (적합성 검정 Goodness of Fits Test)

(1) 통계량 활용

• 수리 왜도 skewness : $\frac{E(X-\mu)^3}{\sigma^2}$

• EDA 왜도 : $\frac{(Q_3-M)-(M-Q_1)}{(Q_3-M)+(M-Q_1)}$

(적합성 검정)

귀무가설 : 데이터는 정규분포를 따른다.

대립가설 : 정규분포를 따르지 않는다.

(2) Shapiro Wilk W-통계량

• $W=\frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$, 상수 a_i 는 분산행렬을 이용하여 구함

$$Z_n=\begin{cases}(-\log(\gamma-\log(1-W_n))-\mu)/\sigma & \text{if } 4\le n\le 11\\ (\log(1-W_n)-\mu)/\sigma & \text{if } 12\le n\le 2000\end{cases}$$

(3) Kolmogorov D-통계량

- $D = \sup_x |F_n(x) - F(x)|$

(4) Anderson-Darling AD 통계량

- $A^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 [F(x)(1 - F(x))]^{-1} dF(x)$

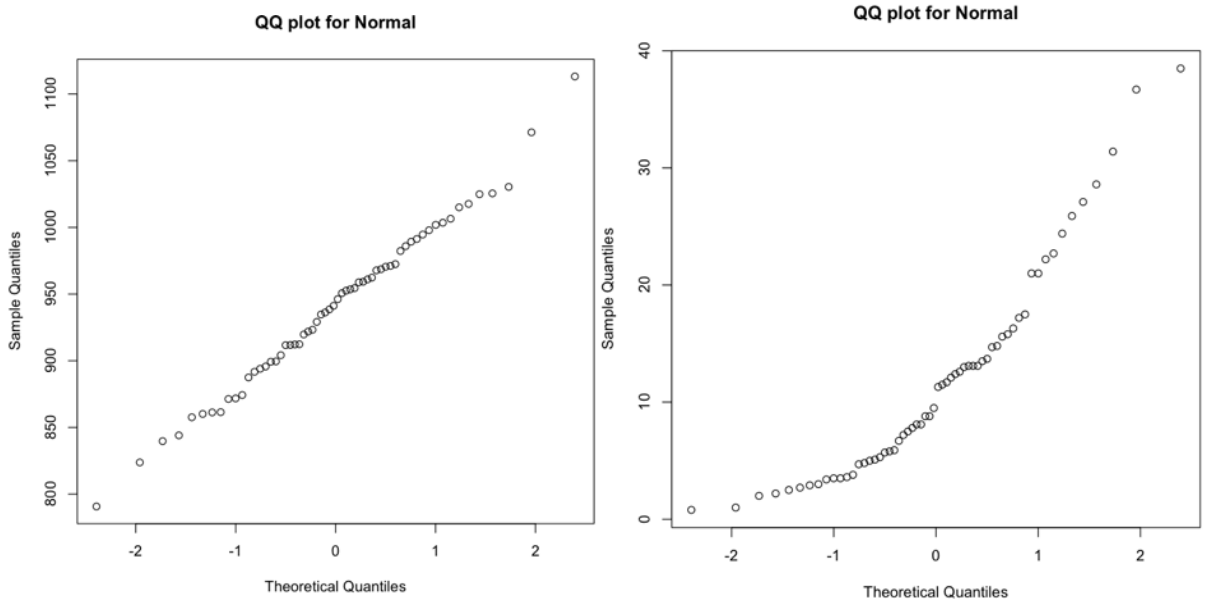
정규 변환

$$\text{Power Transformation } Y^* = \begin{cases} Y^3, \textit{ left} \\ Y^2, \textit{ mild left} \\ \sqrt{Y}, \textit{ mild right} \\ \ln(Y), \textit{ right} \\ 1/Y, \textit{ severe right} \end{cases}$$

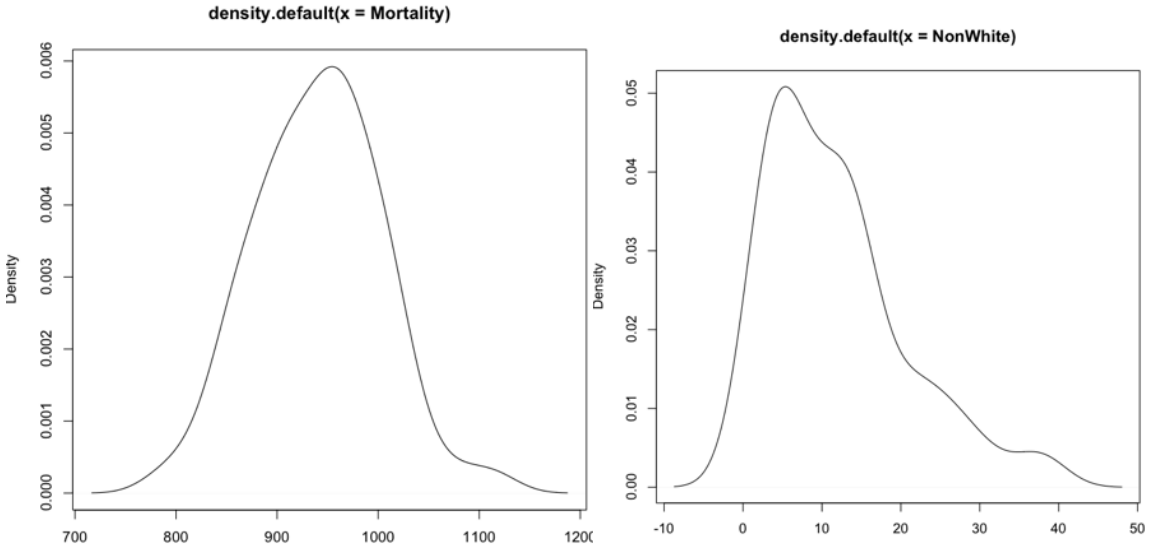
In R

```
ds=read.csv("SMSA.csv")
names(ds)
attach(ds)

qqnorm(Mortality, main="QQ plot for Normal") #QQ plot
plot(density(Mortality))  #확률분포함수

qqnorm(NonWhite, main="QQ plot for Normal")
plot(density(NonWhite))
plot(density(sqrt(NonWhite))) #우로 치우침 해결
plot(density(log(NonWhite))) #우로 치우침 해결

library(nortest)
ad.test(Mortality) #정규성 검정
ad.test(NonWhite)
ad.test(sqrt(NonWhite))
ad.test(log(NonWhite))
```

(Q-Q plot) Mortality - 직선 형태, 정규분포 Non-White 직선에서 벗어남

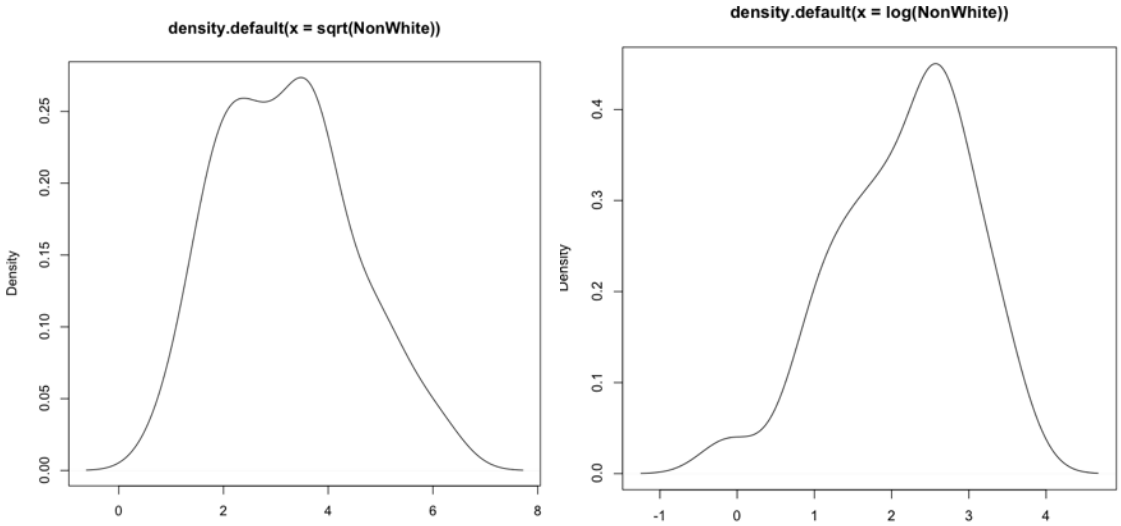


(히스토그램) Mortality 정규분포 Non-white 우로 치우침 - 제곱근 혹은 로그



(Non-white 정규변환)

(제곱근 변환) (로그변환)



(정규성 검정)

- Mortality 유의확률 = 0.97 – 정규분포를 따름
- Non White 유의확률 = 0.00018 – 정규분포 기각, 제곱근 변환 유의확률=0.32, 로그 변환 유의확률=0.18 두 변환 모두 정규분포 근사, 제곱근 변환 유의확률이 더 크므로 제곱근 변환이 가장 적절함

```
> ad.test(Mortality) #정규성 검정

    Anderson-Darling normality test

data:  Mortality
A = 0.138, p-value = 0.9745
```

```
> ad.test(NonWhite)

    Anderson-Darling normality test

data:  NonWhite
A = 1.7198, p-value = 0.0001847
```

```
> ad.test(sqrt(NonWhite))

    Anderson-Darling normality test

data:  sqrt(NonWhite)
A = 0.4128, p-value = 0.3286
```

```
> ad.test(log(NonWhite))

    Anderson-Darling normality test

data:  log(NonWhite)
A = 0.5151, p-value = 0.1843
```