1

정규변환

1. 개념

확률변수(데이터)가 정규분포를 따르지 않는 경우, 변수변환을 통하여 정규분포를 따르도록하는 것을 정규변환이라 한다.

1) 필요이유

- 양적 데이터의 중앙 위치를 나타내는 통계량은 중위수(순서의 중앙), 평균(크기의 중앙)
- 중위수가 중앙 위치의 최적 값이지만 중심극한정리(모평균 추론에 필요)에 의해 평균이 추론의 중심
- 평균은 치우침과 이상치에 민감하게 반응, not resistant, 그러므로 평균에 대한 추론을 위하여 먼저 치우침과 이상치 해결해랴 함
- 이를 해결하지 않으면 신뢰구간의 폭이 넓어져 귀무가설을 기각할 가능성이 낮아짐 -? 이는 연구가설이 채택될 가능성이 낮아짐

2) 변수변환

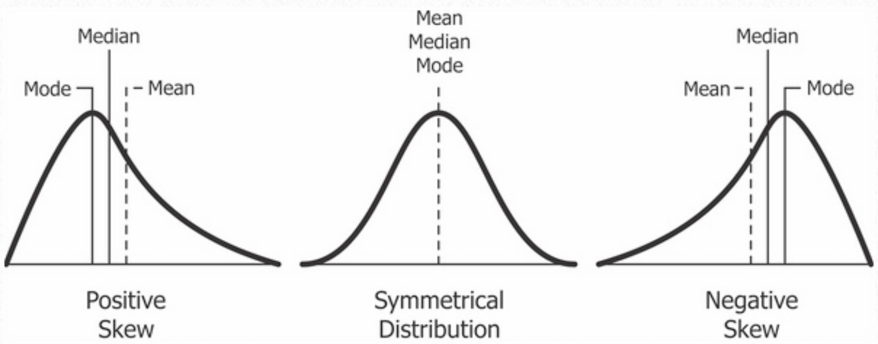
확률변수 X의 확률밀도함수가 $f(x)$를 따른다. 그렇다면 X의 함수 $g(X)$가 갖는 확률밀도함수는?

(예1) 만약 $U \sim U(0,1)$, $-\frac{1}{\lambda}\ln(1-U) \sim \exp(\lambda)$

(예2) 만약 $Z \sim SN(0,1)$, $Z^2 \sim \chi^2(1)$

2. 진단

1) 치우침 진단



(1) 통계량 활용

- 수리 왜도 skewness : $\frac{E(X-\mu)^3}{\sigma^2}$
- EDA 왜도 : $\frac{(Q_3+Q_1-2Median)}{(Q_3-Q_1)}$
- Pearson Median 왜도 : $PS = \frac{3(Mean-Md)}{sd}$

정규분포=0, 우로 치우침 +, 좌로 치우침 -

(2) 적합성 검정

<u>귀무가설 : 데이터는 정규분포를 따른다.</u>

대립가설 : 정규분포를 따르지 않는다.

(1) Shapiro Wilk W-통계량

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

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

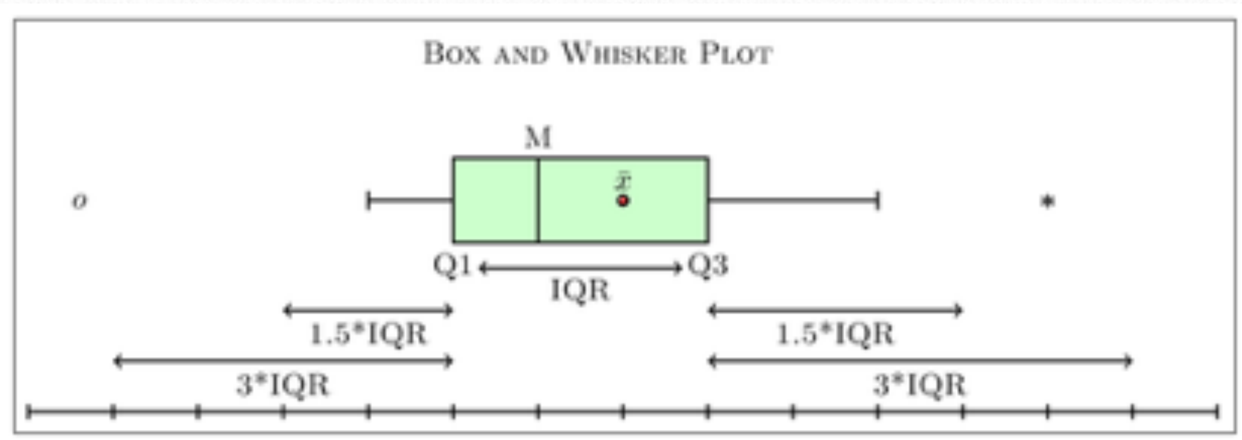
(2) Kolmogorov D-통계량

- $D = \max_x \left| F_n(x) - \Phi(x) \right|$
- $\Phi(x)$ 누적정규분포, $Fn(x)$ 데이터누적분포함수

(3) Anderson-Darling AD 통계량

$$A^2 = n \int (Fn(x) - \Phi(x))^2 \left| \Phi(x)\Phi(1-x) \right|^{-1} d\Phi(x)$$

2) 이상치 진단



- Box-whisker 상자-
- 수염 그림
- IQR (Inter Quartile Range 사분위 범위) = $Q_3 - Q_1$
- $\pm 1.5 * IQR$ - mild 약한 이상치
- $\pm 3 * IQR$ - severe 강한 이상치

3. 해결

이상치와 치우침 중 어느 것을 먼저 해결해야 하나? 치우침을 해결하면 이상치도 정상적인 관측치에 포함될 가능성이 있으므로 (극심한 이상치가 아니라면) 치우침을 먼저 해결하는 것이 적절하다.

1) 치우침 해결 : Power 변환

치우침을 해결하는 방법은 다음의 변환을 활용한다. 이 방법은 Box-Cox Power 변환을 근거하고 있음

$$Y* = \frac{Y^\lambda - 1}{\lambda}$$

정규변환 : $Y* = \begin{cases} Y^3, \ left \\ Y^2, \ mild\ left \\ \sqrt{Y}, \ mild\ right \quad < Y^{1/3} \\ \ln(Y), \ right \quad \text{< LOG10(Y)} \\ -1/Y, \ severe\ right \end{cases}$

2) 이상치 해결

이상치는 제거 후 분석을 실시한다.

4. 사례연구

| LPGA2008 데이터 : 우승상금 | - 우로 치우친 변수(일반적인 데이터 치우침), (좌로 치우친 데이터는 베타분포 이외에는 없음)

1) 데이터 읽기

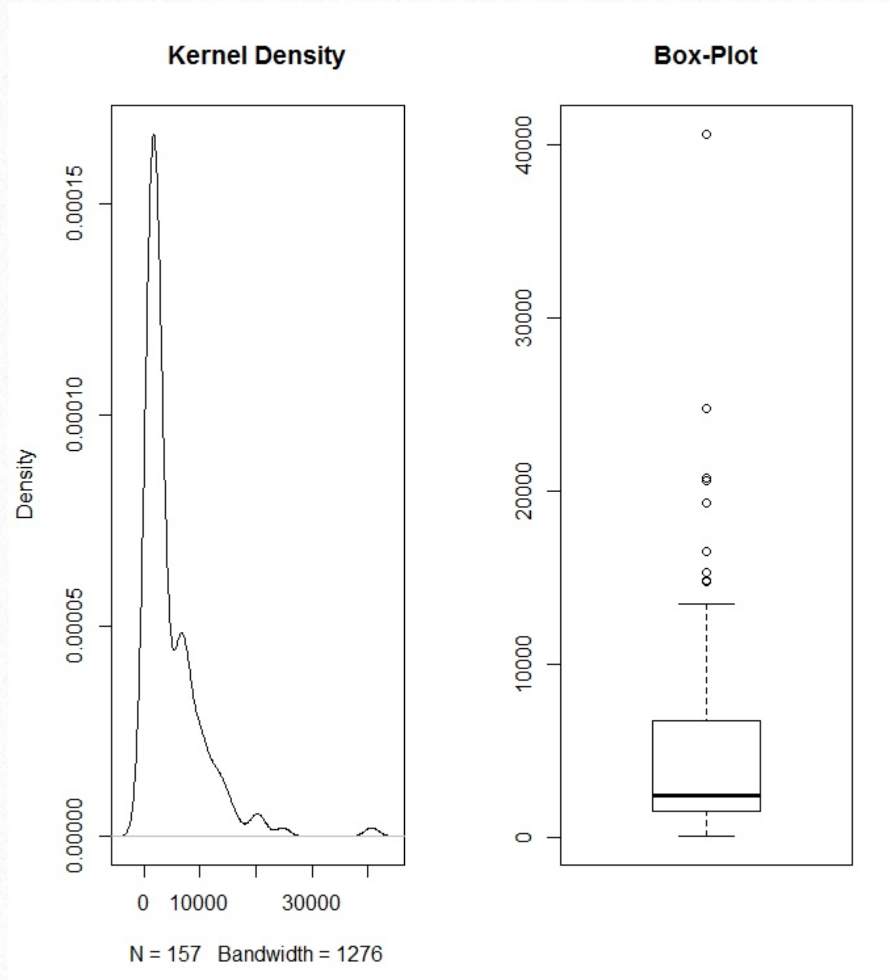
- 데이터 홈페이지 url 활용 불러오기

2) 함수설명

- par(mfrow=(c(a,b))) : 행 a개 분할, 열 b개 분할하여 그래프 그리기
- plot(density()) : kernel 확률밀도함수 그리기
- boxplot() : 상자 수염 그리기
- shapiro.test() : 정규성 검정
- ad.test() : 정규성 검정

```
DS=read.csv("http://wolfpack.hnu.ac.kr/iBooks/example_data/lpga2008.csv")
names(DS)

par(mfrow=(c(1,2))) #원 데이터
plot(density(DS$상금), main="Kernel Density")
boxplot(DS$상금, main="Box-Plot")

library(nortest)
shapiro.test(DS$상금)
ad.test(DS$상금)

par(mfrow=(c(1,2)))  #제곱근 변환
plot(density(sqrt(DS$상금)), main="Kernel Density(Sqrt)")
boxplot(sqrt(DS$상금), main="Box-Plot(Sqrt)")
shapiro.test(sqrt(DS$상금))
ad.test(sqrt(DS$상금))

par(mfrow=(c(1,2))) # 로그 변환
plot(density(log(DS$상금)), main="Kernel Density(Log)")
boxplot(log(DS$상금), main="Box-Plot(Log)")
shapiro.test(log(DS$상금))
ad.test(log(DS$상금))
```

```
> names(DS)
[1] "골퍼"           "평균_비거리"     "페어웨이_안착율"
[4] "그린_적중률"     "평균_버팅수"     "샌드_회수"
[7] "샌드_세이브"     "상금"            "참가_라운드수"
```



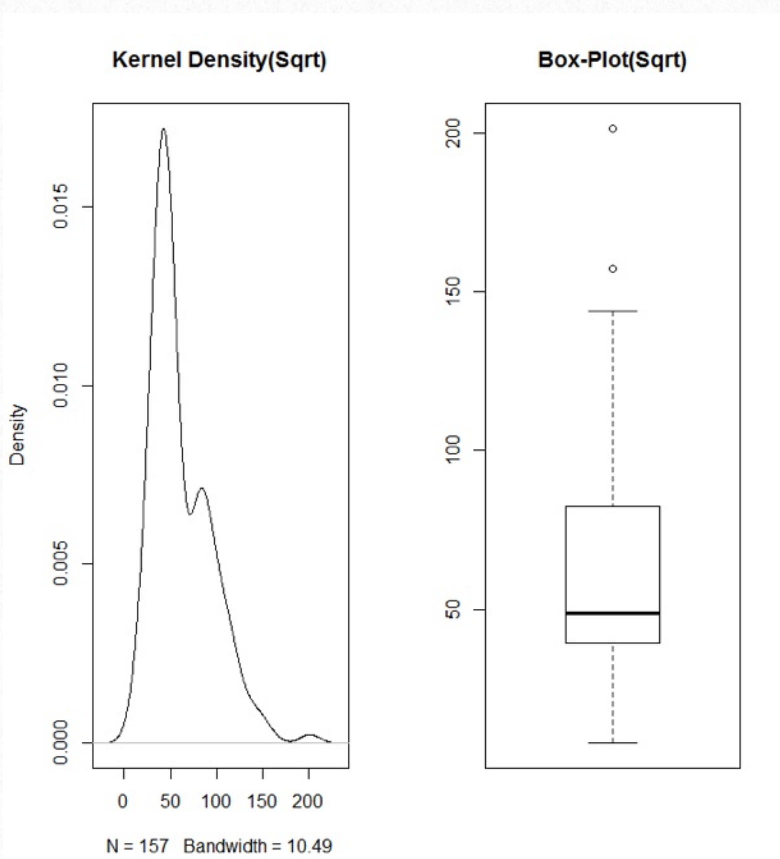
3) 치우침 해결

우승 상금 원 데이터는 우로 치우침, 이상치 존재

```
> shapiro.test(DS$상금)

        Shapiro-Wilk normality test

data:  DS$상금
W = 0.71659, p-value = 4.757e-16
```

우선 치우침을 해결하자 ->
해결방법 : 제곱근, 로그변환

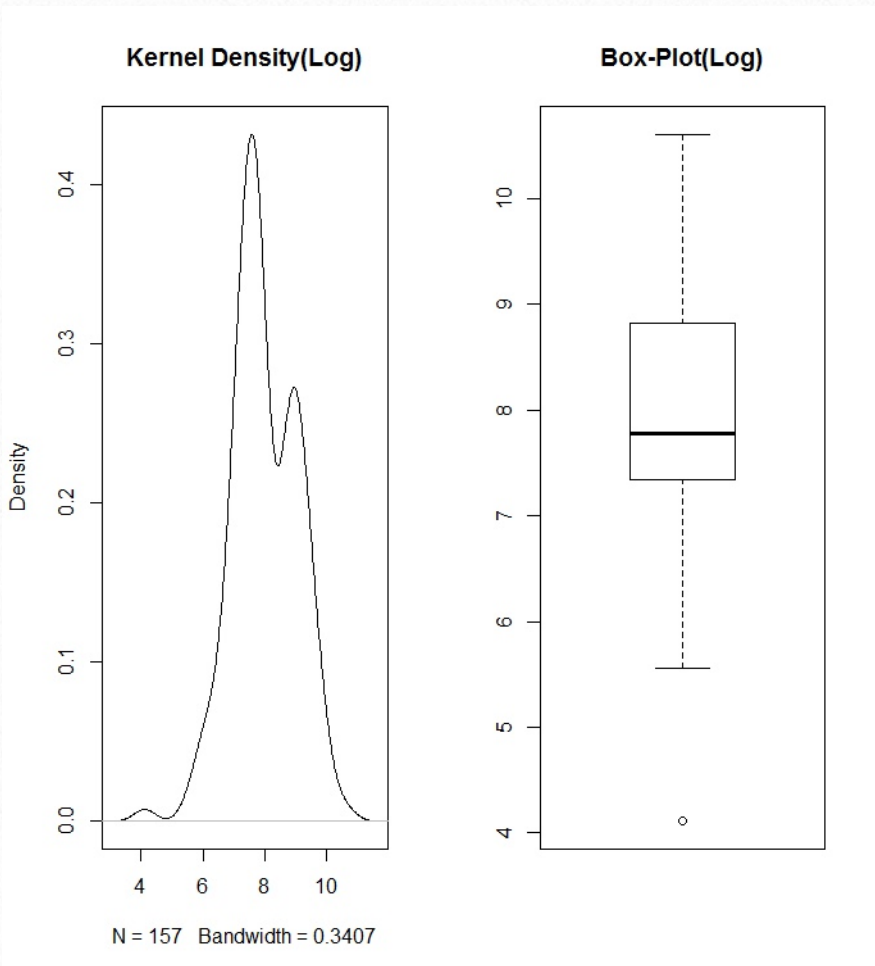
(1) 제곱근 변환




```
> shapiro.test(sqrt(DS$상금))

        Shapiro-Wilk normality test

data:  sqrt(DS$상금)
W = 0.90147, p-value = 8.899e-09

> ad.test(sqrt(DS$상금))

        Anderson-Darling normality test

data:  sqrt(DS$상금)
A = 5.2031, p-value = 6.597e-13
```

(2) 로그변환

제곱근 변환도 우로 치우침을 해결하지 못했다. 보다 심각한 우로 치우침을 해결하기 위하여 로그 변환

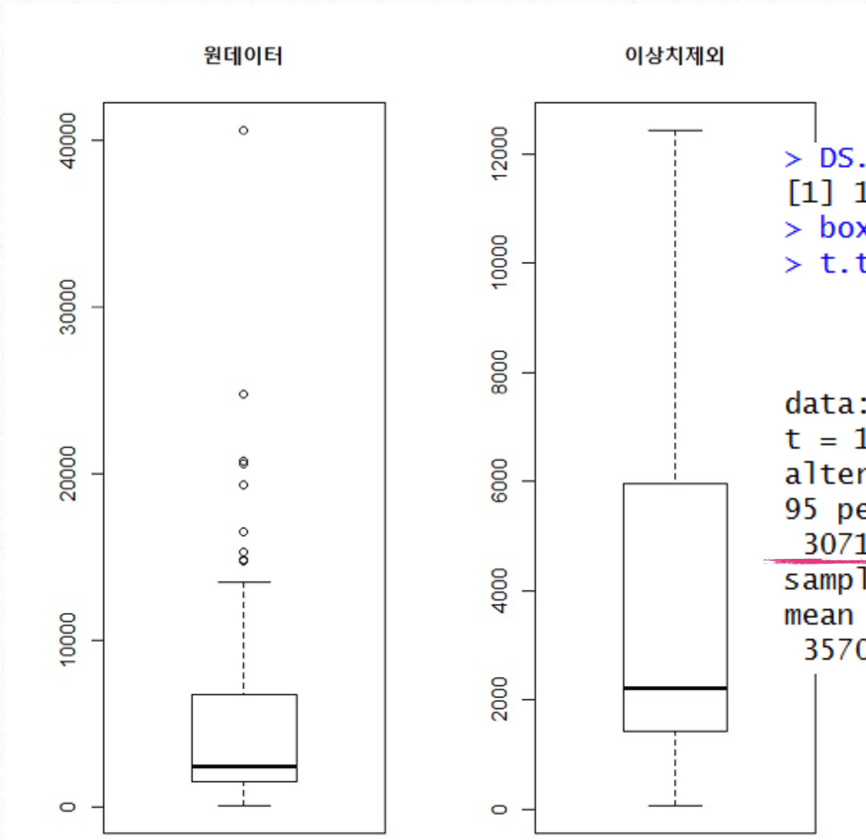



```
> shapiro.test(log(DS$상금))

        Shapiro-Wilk normality test

data:  log(DS$상금)
W = 0.97856, p-value = 0.01512

> ad.test(log(DS$상금))

        Anderson-Darling normality test

data:  log(DS$상금)
A = 1.3187, p-value = 0.001952
```

우로 치우침은 어느 정도 해결됨, - 그러나 중앙 50% 부분(상자) 위쪽(50%~75% 부분이 25%~50%보다 길고) 폭이 길므로 중앙에서 우로 치우침 경향 보임 - 그리고 원 데이터와 달리 작은 부분에 이상치가 존재함.

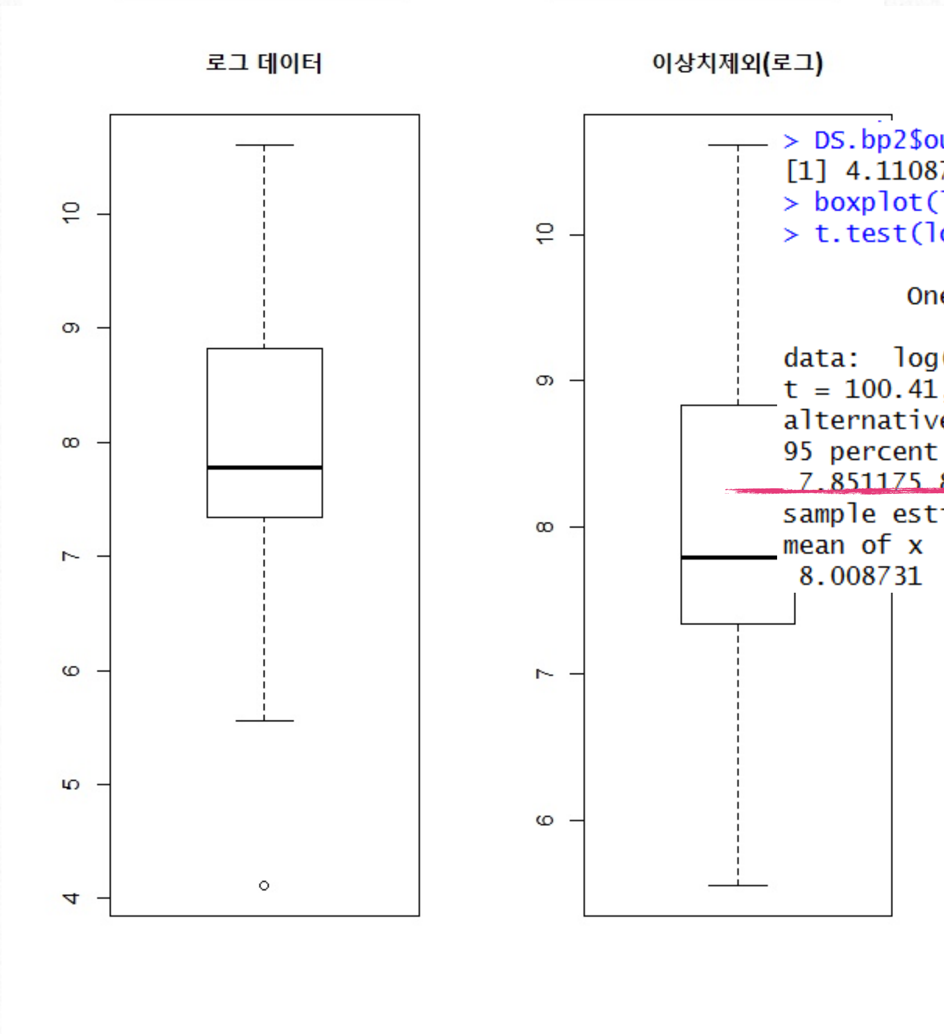
(3) 치우침 해결이 필요한 이유

```
par(mfrow=(c(1,2)))
DS.bp<-boxplot(DS$상금,main="원데이터")  #상자수염그림 결과 보기
DS.bp$out #이상치
boxplot(DS[which(DS$상금<=12500),]$상금, main="이상치제외")
t.test(DS[which(DS$상금<=12500),]$상금) #일변량 95% 신뢰구

DS.bp2<-boxplot(log(DS$상금),main="로그 데이터")
DS.bp2$out
boxplot(log(DS[which(log(DS$상금)>5),]$상금),main="이상치제외(로그)")
t.test(log(DS[which(log(DS$상금)>5),]$상금))
```




```
> DS.bp$out #이상치
[1] 19343 14808 20727 14862 40635 16498 15296 24799 20613
> boxplot(DS[which(DS$상금<=12500),]$상금, main="이상치제외")
> t.test(DS[which(DS$상금<=12500),]$상금) #일변량 95% 신뢰구

        One Sample t-test

data:  DS[which(DS$상금 <= 12500), ]$상금
t = 14.159, df = 142, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3071.837 4068.764
sample estimates:
mean of x
 3570.301
```

신뢰구간 폭 = 4068.8-3071.8=997




```
> DS.bp2$out
[1] 4.110874
> boxplot(log(DS[which(log(DS$상금)>5),]$상금),main="이상치제외(로그)")
> t.test(log(DS[which(log(DS$상금)>5),]$상금))

        One Sample t-test

data:  log(DS[which(log(DS$상금) > 5), ]$상금)
t = 100.41, df = 155, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.851175 8.166287
sample estimates:
mean of x
 8.008731
```

신뢰구간 폭 = exp(8.17-7.85)=967.6

```
> exp(8.17)-exp(7.85)
[1] 967.6096
```

- 치우침을 해결하면 동일 신뢰수준이라도 구간의 폭이 좁아져 정확도가 높아진다.
- 이상치는 치우침 해결 후에 최종적으로 삭제하여 해결하면 된다.
- 물론 치우침이 없는 경우에는 이성치 진단 및 해결만 하면 된다.