1

Box Whisker Plot

1. 정의

나무 상자 그림은 데이터의 시각적 표현으로 5개의 순서 통계량 값을 이차원 그래프에 표현함

- x-축 : 범주형 변수명 혹은 범주 값
- Y-축 : 데이터 값

순서통계량

데이터 값 $x_1, x_2, \ldots x_n$ 을 크기 순서대로 정렬한 통계량 $x_{(1)}, x_{(2)}, \ldots, , x_{(n)}$

1) 최소값 min $x_{(1)}$
2) 최대값 max $x_{(n)}$
3) 중앙값 median $x_{(MD)}$, Median Depth 중위값 깊이 = $\frac{n+1}{2}$ - 제 2사분위 (정의) 데이터 순서 개념의 중앙 척도, 분석 데이터에는 중위값보다 적은 관측치가 (적어도) 50%, 큰 관측치가 (적어도) 50% 있음
4) 제일 사분위 first Quartile: $Q_1 = x_{(QD)}$, 사분위 깊이 = $\frac{(MD)+1}{2}$: 분석 데이터에는 Q1 보다 적은 관측치가 (적어도) 25%, 큰 관측치가 (적어도) 75% 있음 *) (MD)는 MD 값을 넘지 않은 최대 정수
5) 제삼 사분위 $Q_3 = x_{(n-QD)}$
6) 범위 Range : $R = x_{(n)} - x_{(1)}$
7) 사분위 범위 Mid-range, Inter-Quartile Range : $IQR = Q_3 - Q_1$

알아두기

상자의 높이 = IQR 사분위 범위

상자의 넓이 - 의미 없음

2. 활용

1) 이상치 outlier 진단

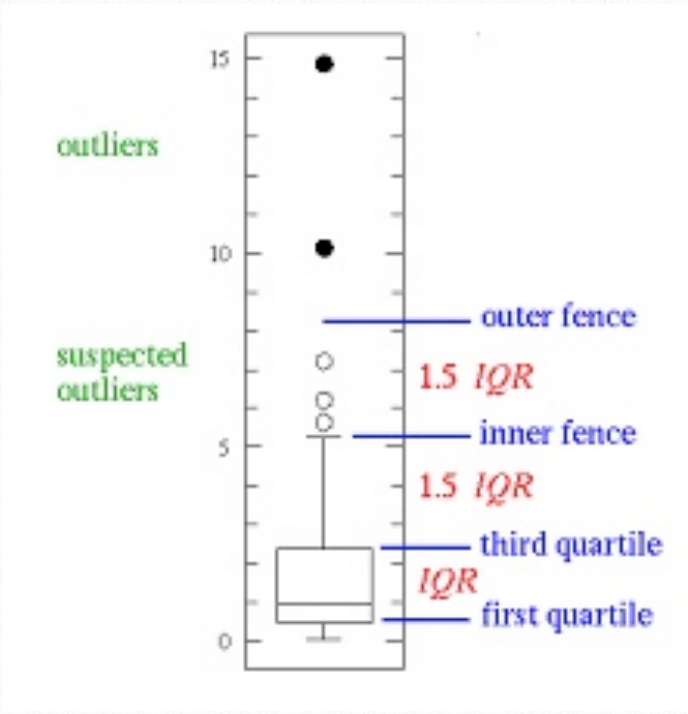
(1) 진단 방법

순한 이상치 mild outlier

2) $Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR$

심한 severe 이상치

3) $Q_1 - 3 * IQR, Q_3 + 3 * IQR$

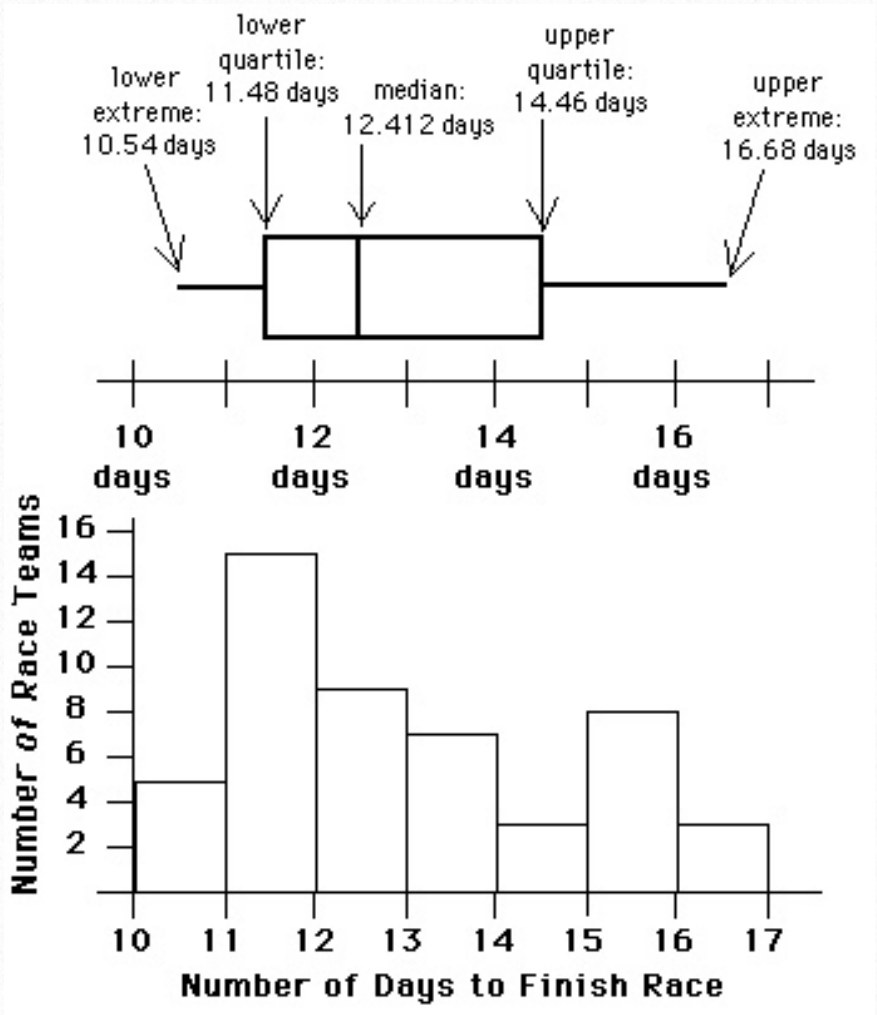


(4) 처리

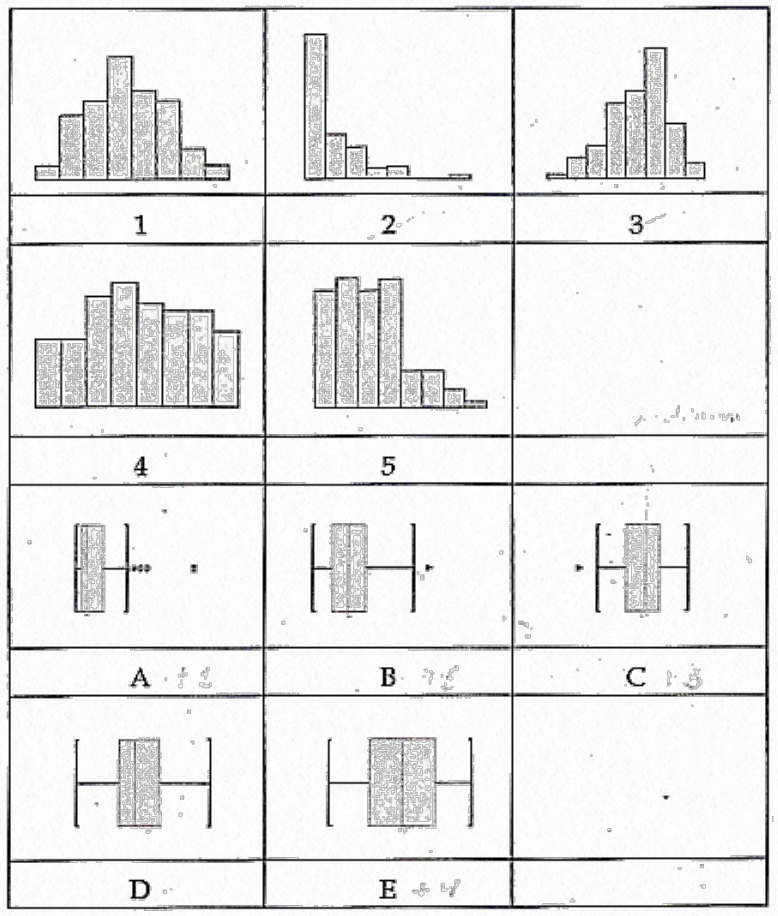
- 제거 : 관측치 입력 오류 확인 후
- 정규변환 : 추론, 모형을 위한 2차 분석이 필요할 때 (대표본이 아닌 경우)

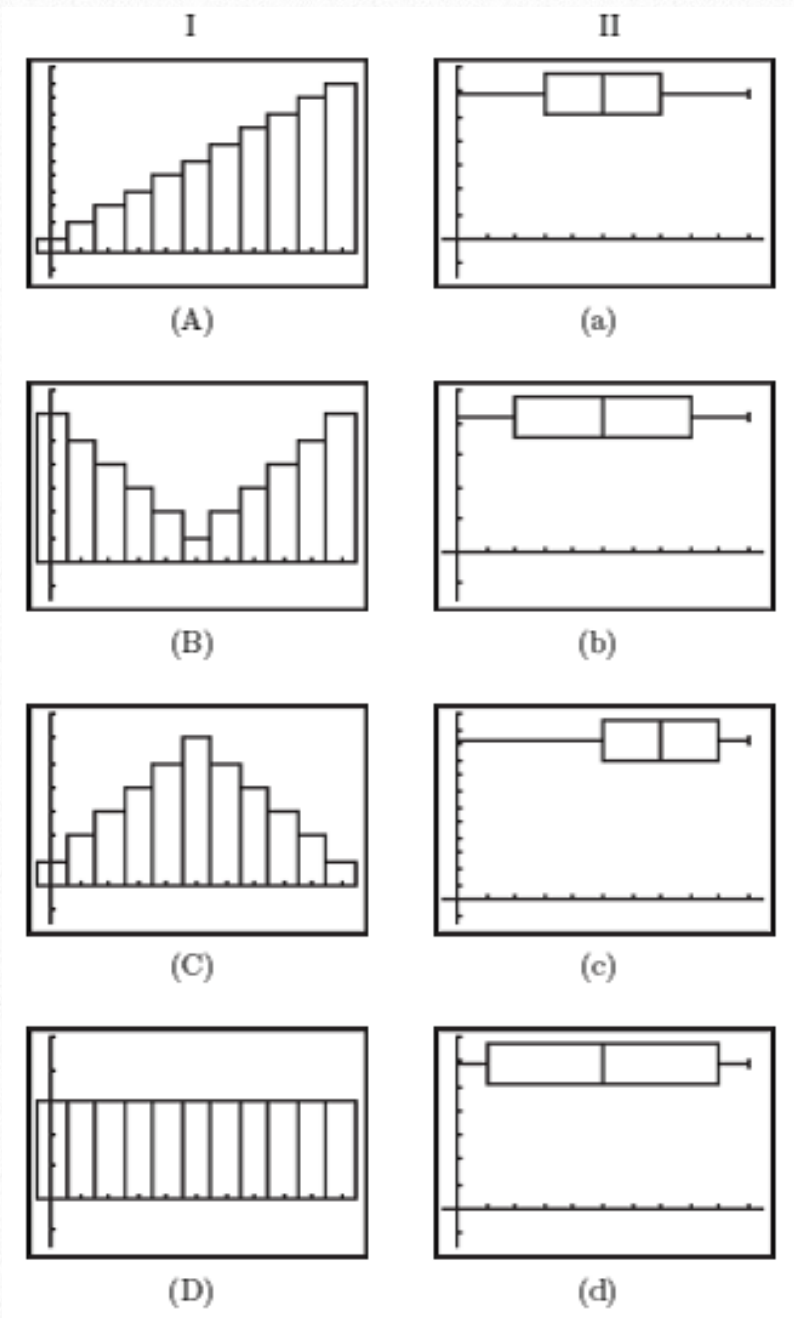
5) 분포함수 진단

최소값, 제1사분위, 중위값, 제3사분위, 최대값, 범위, 사분위 범위를 이용하여 데이터 확률분포함수를 예상할 수 있음



http://www.wisfaq.nl/show3archiveict.asp?id=47406&j=2006 (참고 사이트)

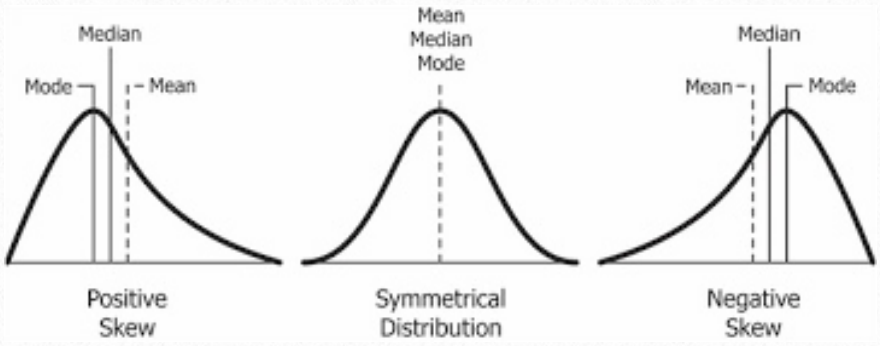




3. 치우침 skewness과 통계량

1) 정의

- 좌우 대칭 : 평균=중위값
- 우로 치우침, 양의 치우침 : 평균 > 중위값
- 좌로 치우침, 음의 치우침 : 평균 < 중위값



2) 왜도 척도

(1) 크기 통계량

$$E(\frac{X-\mu}{\sigma})^3 = \frac{\mu^3}{\sigma^3}$$

(성질) $skewness[\sum X_i] = Skew[X]/\sqrt{n}$

(성질) 정규분포, 좌우 대칭 분포 왜도 = 0

(2) 순서통계량

(Galton 왜도) $G.Skew = \frac{Q_1 + Q_3 - 2*Q_2}{Q_3 - Q_1}$

(Pearson 왜도) $P.Skew = \frac{3(Mean - MD)}{SD}$

4. 문제 해결

1) 이상치 outlier, extreme value

- 이상치 코딩 오류 여부 확인 후 제거
- 만약 분포의 치우침이 있는 경우 정규변환 후 여전히 이상치이면 제거
- 실제 이상치의 경우 정보를 가진 관측 값이므로 이에 대한 정성적 분석이 필요함

2) 치우침

정규성 검정을 통하여 정규성 검정을 실시하고 치우침이 있는 경우 정규 변환을 활용하여 변환 후 분석함

5. 예제 데이터

SASHELP Baseball Data [내용] [데이터]

1986 메이저 리스 타자들의 연봉의 95% 신뢰구간을 구한다고 하자.

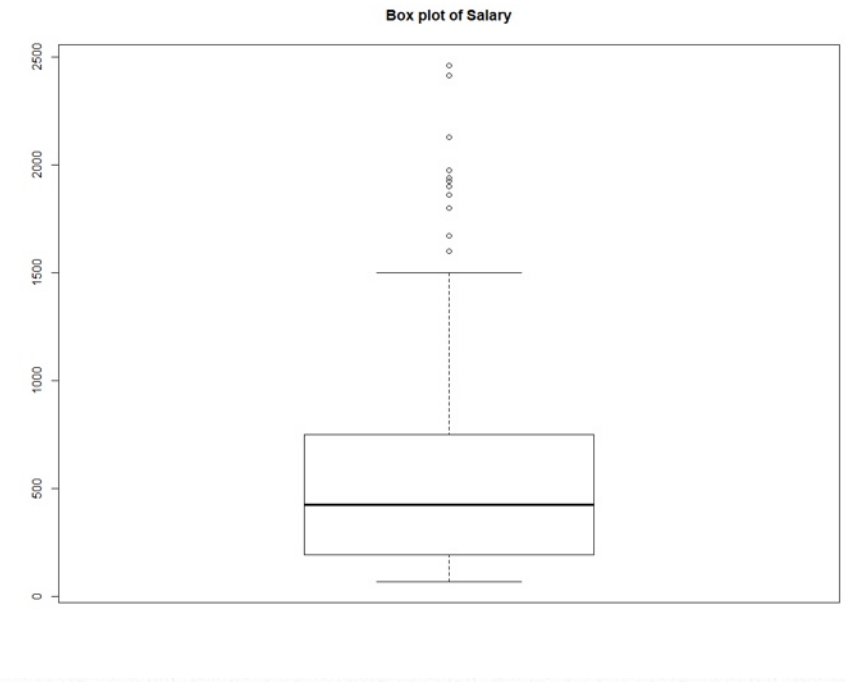
```
library(sas7bdat)
ds<-read.sas7bdat('http://wolfpack.hnu.ac.kr/Spring2018/baseball.sas7bdat') #read sas data into R
t.test(ds$Salary) #t-test H0: mu=0
```

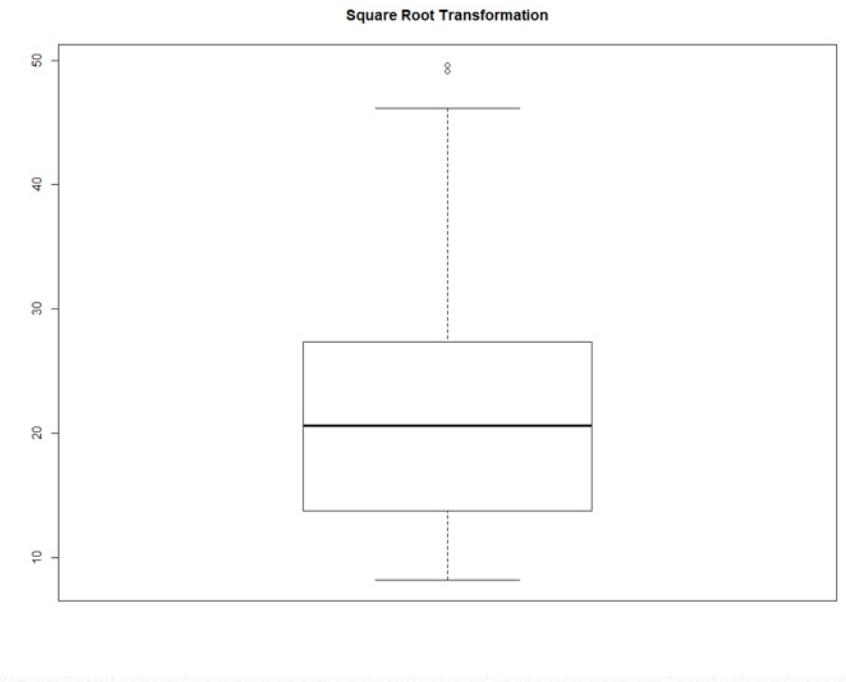
```
        One Sample t-test

data:  ds$Salary
t = 19.266, df = 262, p-value < 2.2e-16
alternative hypothesis: true mean is not
95 percent confidence interval:
 481.1522 590.6996
sample estimates:
mean of x
 535.9259
```

메이저 리그 타자 연봉 95% 신뢰구간 (481.15, 590.7) 신뢰구간 폭은 109.6

(신뢰성?) 상자 나무 수염 그림을 그려보면 선수들의 연봉은 우로 치우침

```
boxplot(ds$Salary,main='Box plot of Salary')$out #graph box plot
library(nortest)
ad.test(ds$Salary) #check Normality
```





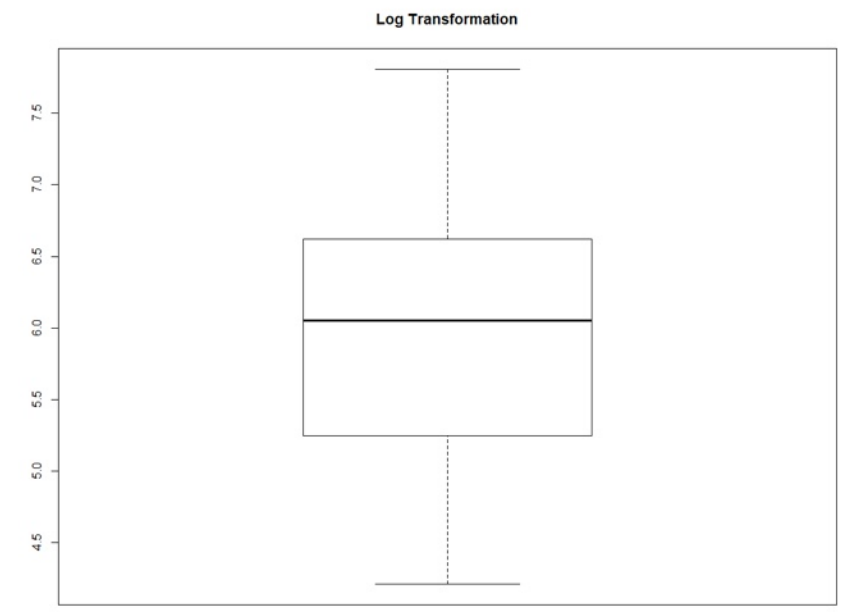
이상치가 11개 존재하고 정규성 검정결과 유의확률이 매우 낮아 데이터가 정규분포를 따르지 않음

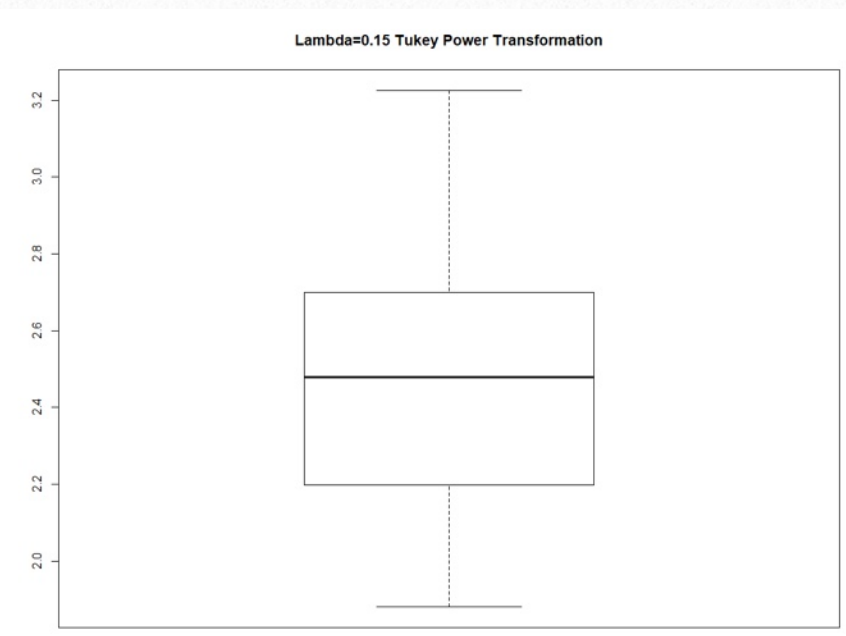
```
> boxplot(ds$Salary,main='Box plot of Salary')$out
 [1] 1975.000 1900.000 1861.460 2460.000 1925.571 2
[10] 1670.000 1600.000
```

```
        Anderson-Darling normality test

data:  ds$Salary
A = 9.3077, p-value < 2.2e-16
```

```
ds0<-na.omit(ds) #delete missing value
boxplot(sqrt(ds0$Salary),main='Square
Root Transformation')
boxplot(log(ds0$Salary),main='Log
Transformation')
library(rcompanion)
tukey.ds<-transformTukey(ds0$Salary)
boxplot(ds0$Salary^0.15,main='Lambda=0.
15 Tukey Power Transformation')
```

- 정규변환은 $log(Salary)$,
- Tukey Power 변환 $Salary^{0.15}$가 가장 적절





```
t.test(log(ds0$Salary))
t.test(ds0$Salary^0.15)
```

```
data:  log(ds0$Salary)
t = 108.1, df = 262, p-value < 2.2e-16
alternative hypothesis: true mean is not
95 percent confidence interval:
 5.819258 6.035185
```

로그 연봉 95% 신뢰구간은 (5.819, 6.035)-exp(x)

연봉 95% 신뢰구간은 (336.72, 417.87). - 폭 81.2

```
data:  ds0$Salary^0.15
t = 123.08, df = 262, p-value < 2.2e-16
alternative hypothesis: true mean is not
95 percent confidence interval:
 2.415085 2.493615
```

(연봉)^0.15 95% 신뢰구간 (2.415, 2.493) - (x)^(1/0.15)

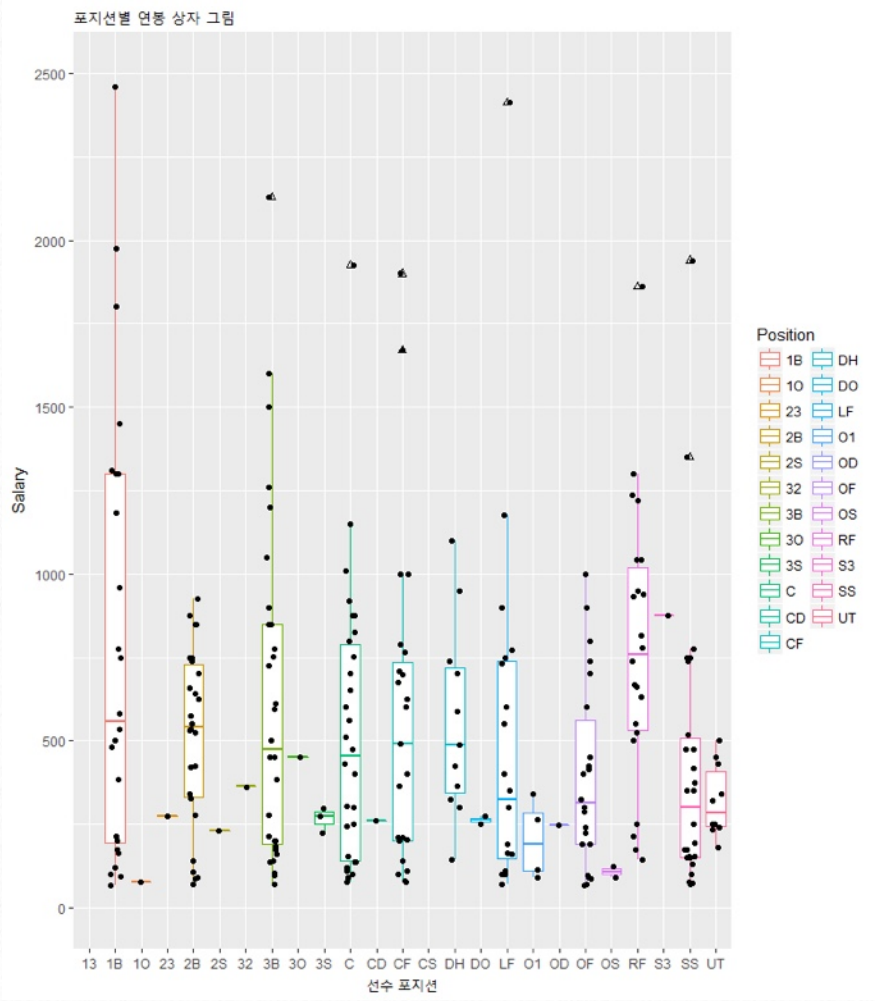
연봉 95% 신뢰구간은 (357.1, 441.4). - 폭 84.3

결론

데이터의 정규 변환 결과 이상치도 제거되고(우로 치우침 해결로 인하여 우측 극단치가 정상 데이터로 흡수) 동일 신뢰수준의 신뢰구간 폭도 좁아짐(신뢰수준이 동일한 경우 구간 폭이 작을수록 정보의 질은 높음, 정보의 신뢰성 높아짐

6. 나무상자 그리기 (범주형 변수별)

```
library(sas7bdat)
ds<-read.sas7bdat('http://
wolfpack.hnu.ac.kr/Spring2018/
baseball.sas7bdat') #read sas data into R

library(ggplot2)
ggplot(ds, aes(x=Position, y=Salary)) +
  geom_boxplot(aes(colour =
Position),outlier.colour=1, na.rm=T,
outlier.shape = 2)+
  geom_jitter(width = 0.2)+
  ggtitle('포지션별 연봉 상자 그림')+
  ylim(c(0, 2500)) +
  xlab('선수 포지션')
```



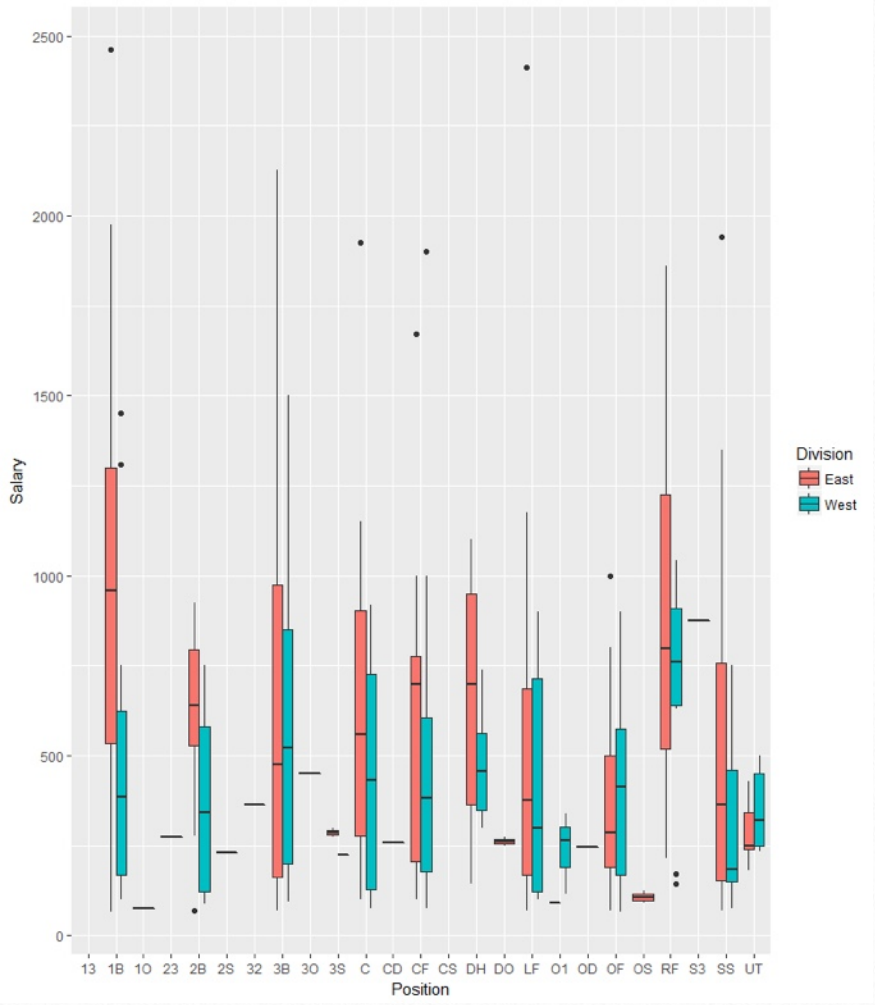
- ggplot() : 라이브러리 ggplot2에 있는 함수로 통계 관련 그래프를 그리는데 유용함

- 인수(argument) aes : aesthetics(미학)을 의미하는 것으로 x-y 축의 변수 이름과 표현 색, 심볼 등 미학적 요소를 설정함
- colour = 설정된 범주형 변수마다 색을 다르게 하여 표현함
- outlier.colour = 정수 (1~256) 색을 지정
- geom_ : 기하 geometry 약어 모든 그래프 내용 등 일종의 템플릿(예제 문서 형식) 개념
- ggtitle(), ylim(), xlim()

- ggplot 은 모든 설정은 geom_*** () 문장을 + 연결하여 사용하면 된다.
- 예를 들어 산점도를 그리고 싶다면 geom_point() 이용하면 된다.

범주형 변수가 2개인 경우

```
ggplot(ds,aes(x=Position,y=Salary,
fill=Division))+
geom_boxplot()
```



- 범주형 변수가 2개 인 경우 한 범주형은 x= 에 지정하고 fill= 다른 범주형 변수를 지정함