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 *2*사분위 *(*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 :* $IQR = Q_3 - Q_1$

상자의 높이 = IQR 사분위 범우;

상자의 넓이 - 의미 없음

2. 활용

1) 이상치 outlier 진단

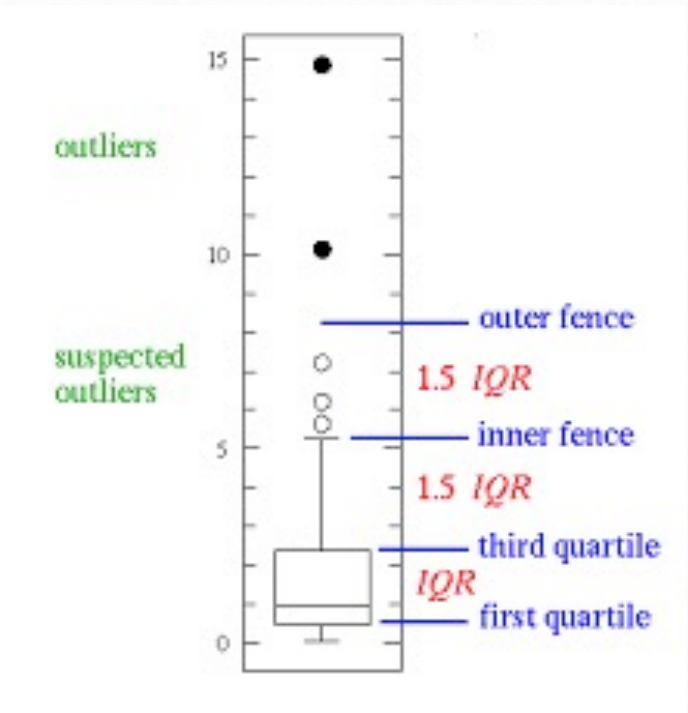
(1) 진단 방법

순한 이상치 mild outlier

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

심한 severe 이상치

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

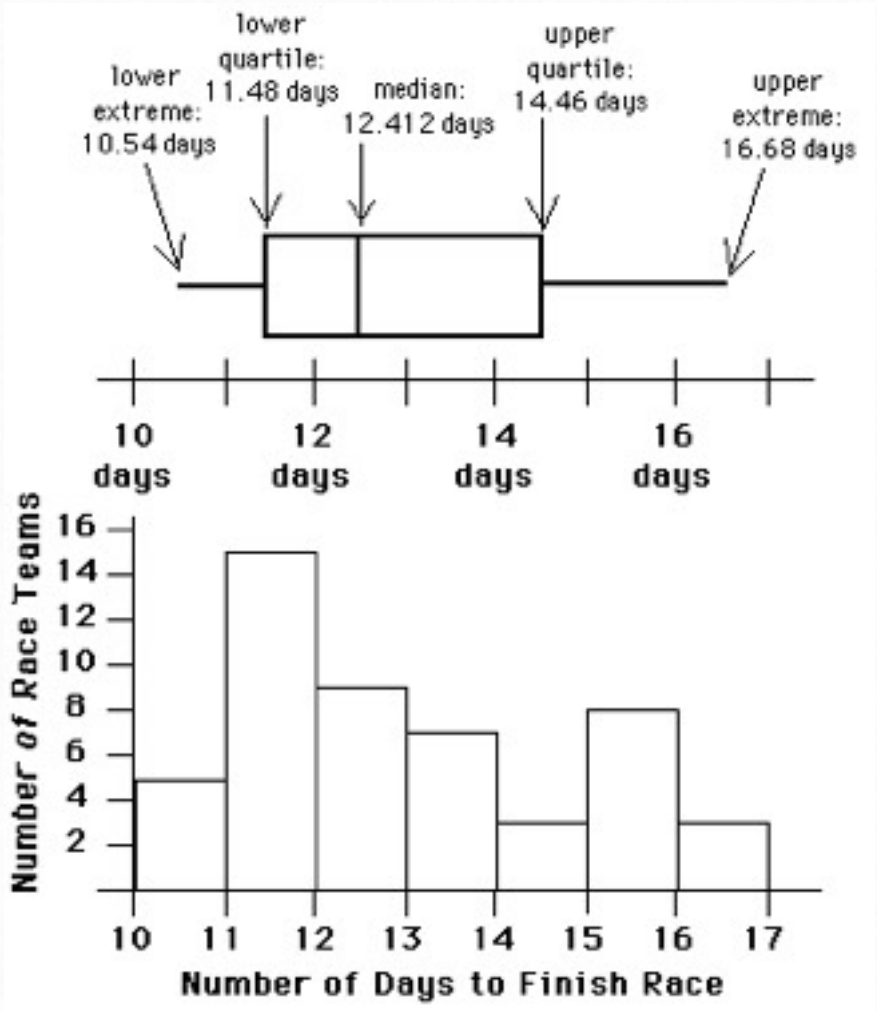


(2) 처리

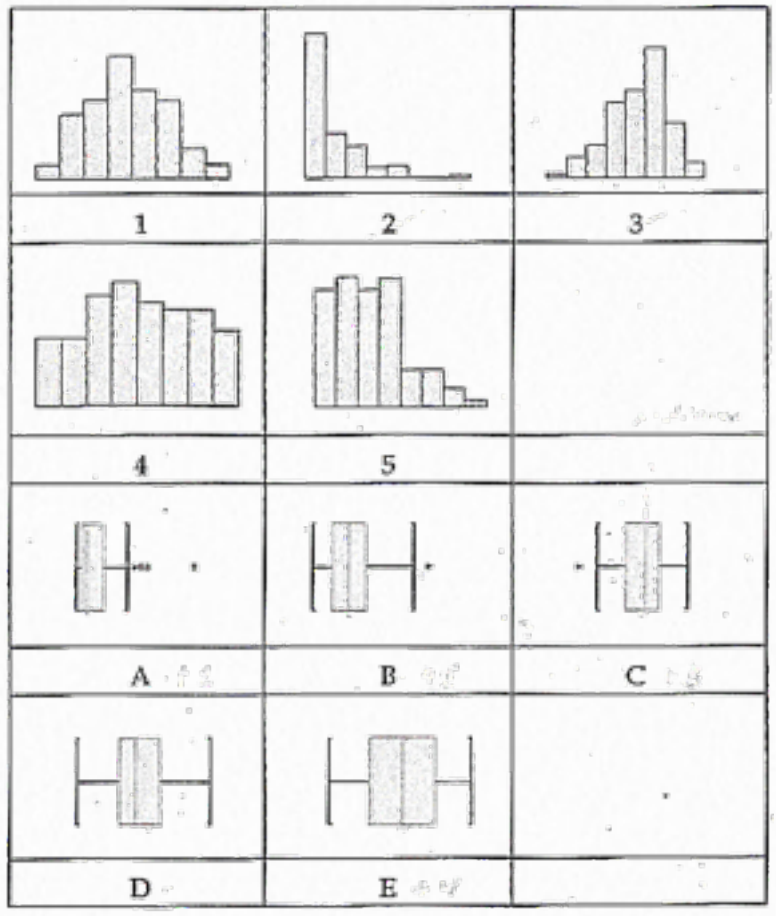
- 제거 : 관측치 입력 오류 확인 후
- 정규변환 : 추론, 모형을 위한 2차 분석이 필요할 때 (대표본이 아닌 경우)

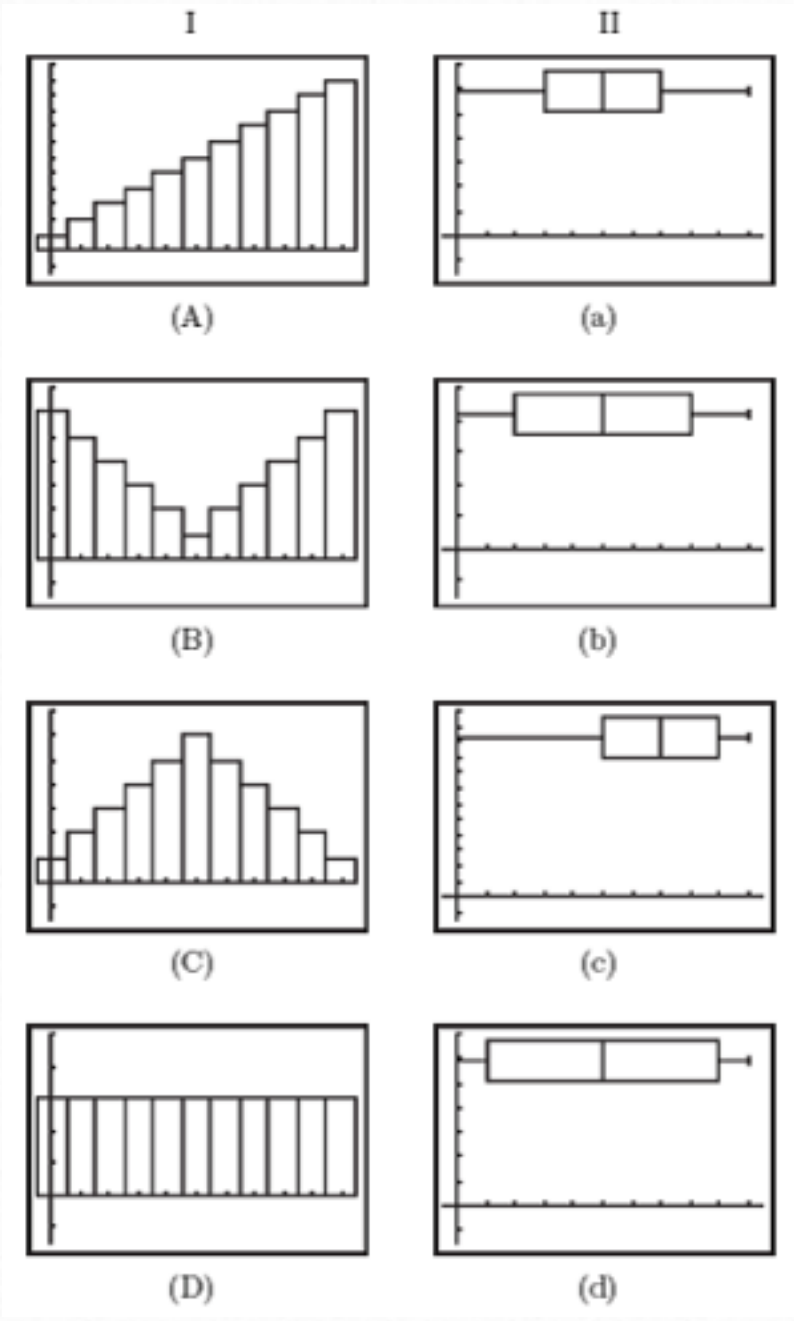
3) 분포함수 진단

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)

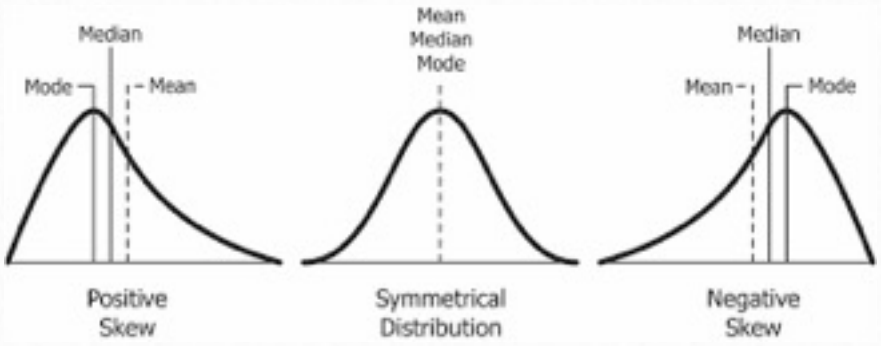




3. 치우침 skewness과 통계량

1) 정의

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



2) 왜도 척도

3)

(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

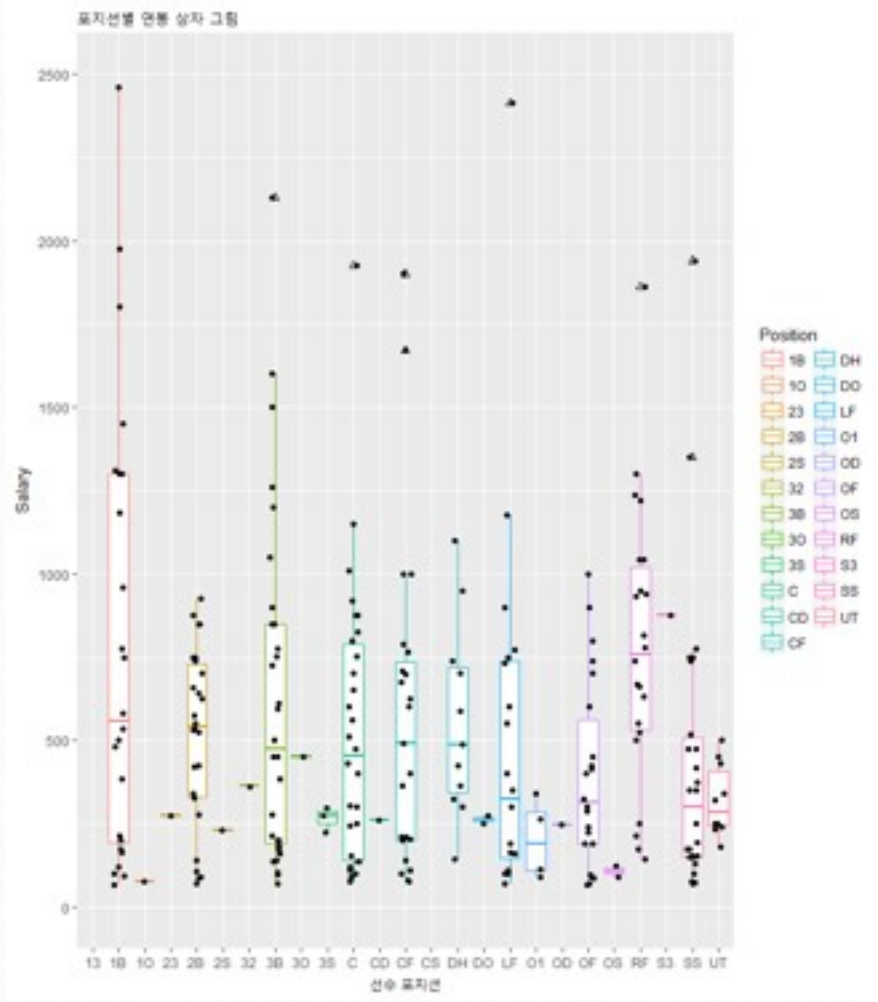
(2) 순서통계량

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

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

```
library(sas7bdat)
ds<-read.sas7bdat('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')
```



4.예제 데이터

SASHELP Baseball Data [내용] [데이터]

범주형 변수가 2개인 경우

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