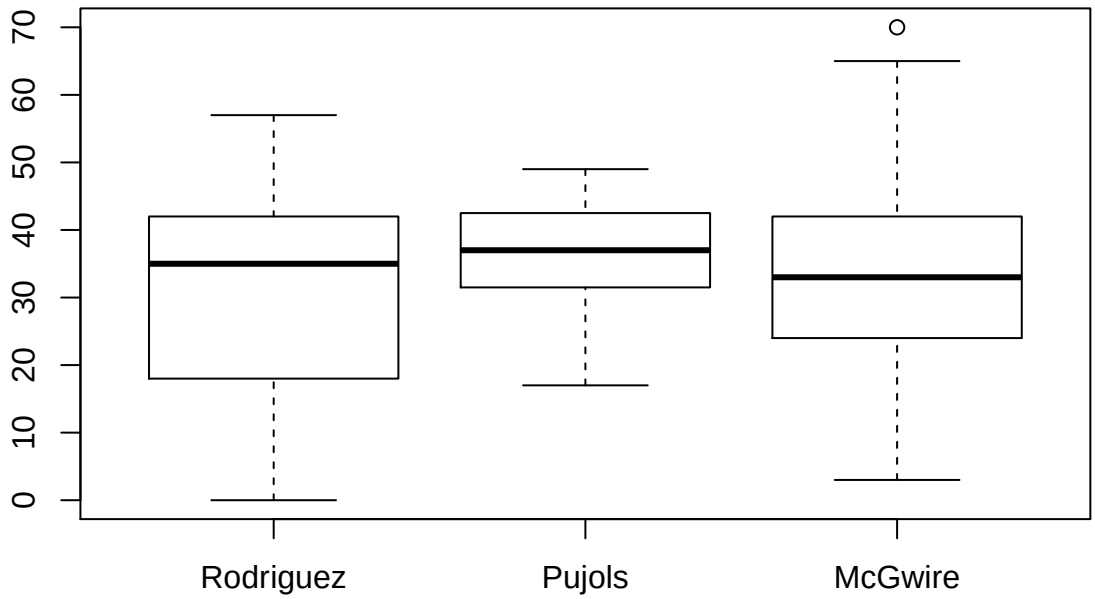
chap05_VisualComparison

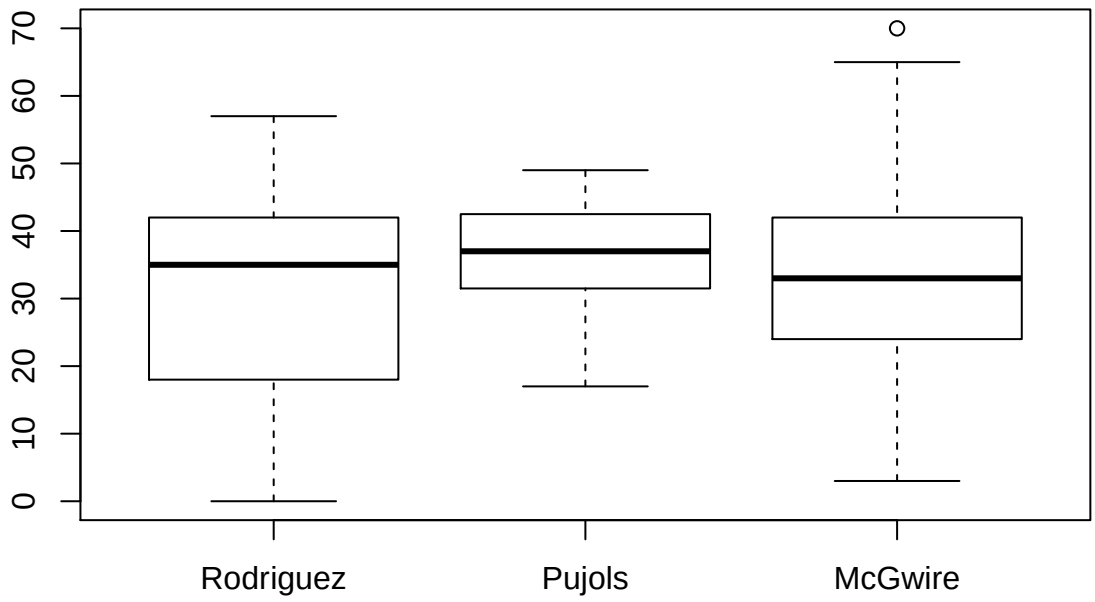
1. Mark Mcgwire, Albert Pujols, Alex Rodriguez Homerun Comparison

```
library(Lahman)
a = subset(Batting, playerID=='rodrial01' | playerID=='pujolal01' | playerID =='mcgwima01')
a$name = factor(a$playerID, levels=c('rodrial01', 'pujolal01', 'mcgwima01'),
                labels=c('Rodriguez','Pujols', 'McGwire'))
boxplot(a$HR~a$name)
```



맥과이어 선수의 이상치 기준선 위에 존재하는 둥근점은 중심 부분인 사분범위를 고려했을 때 지나치게 벗어난 상태로 판명되어, 이상치로 분류된 홈런기록이다.

```
boxplot(HR~name, data=a)
```



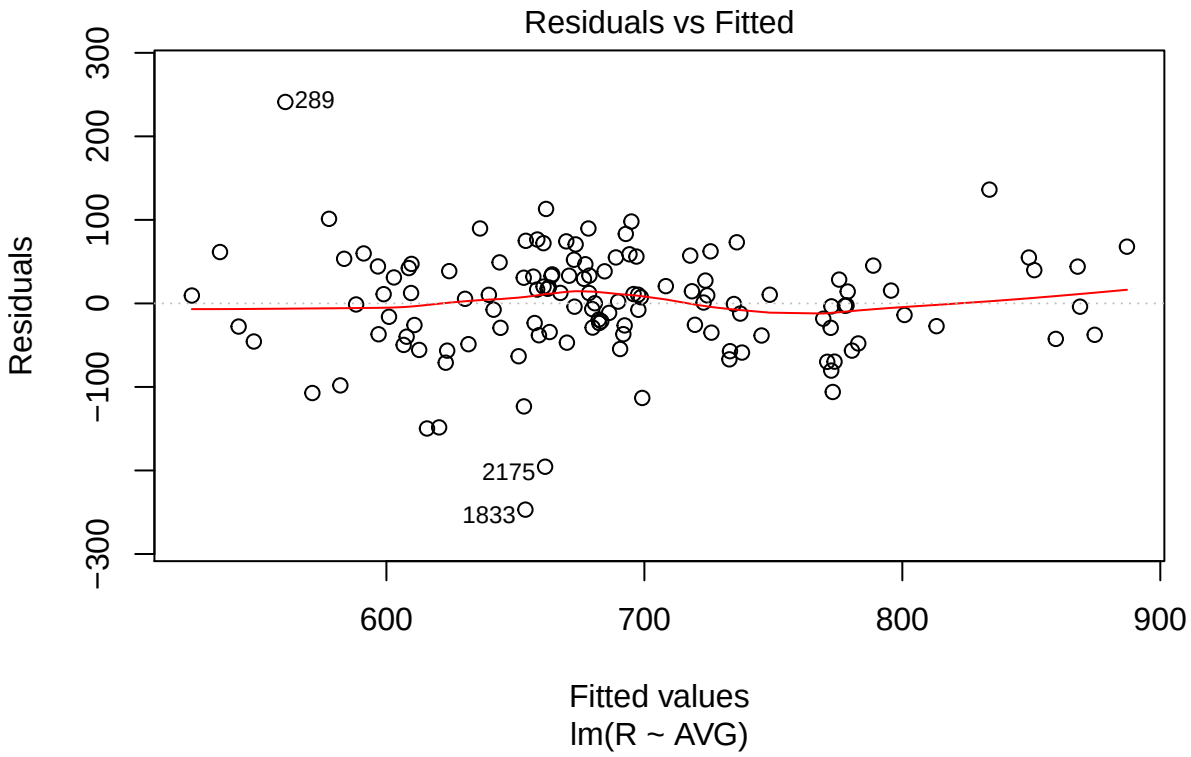
푸홀스 선수 홈런 변수의 주요 값과 사분범위

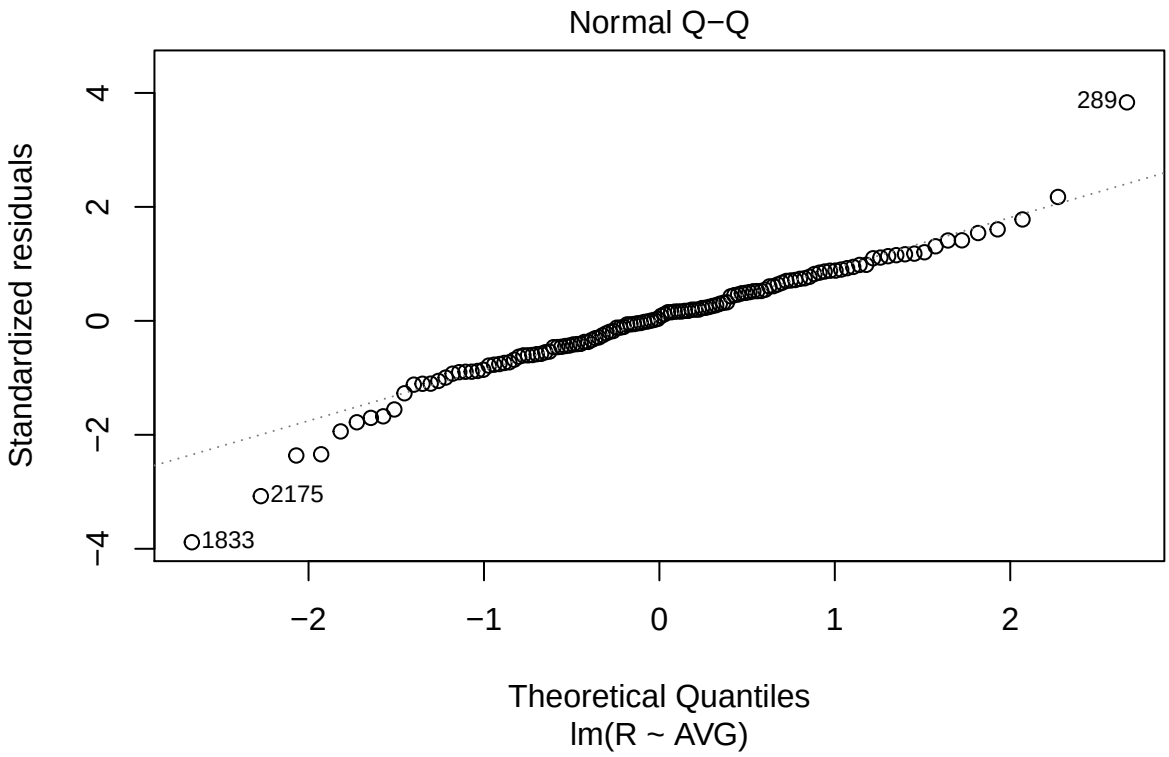
```
fivenum(a$HR[a$playerID=='pujolal01'])
```

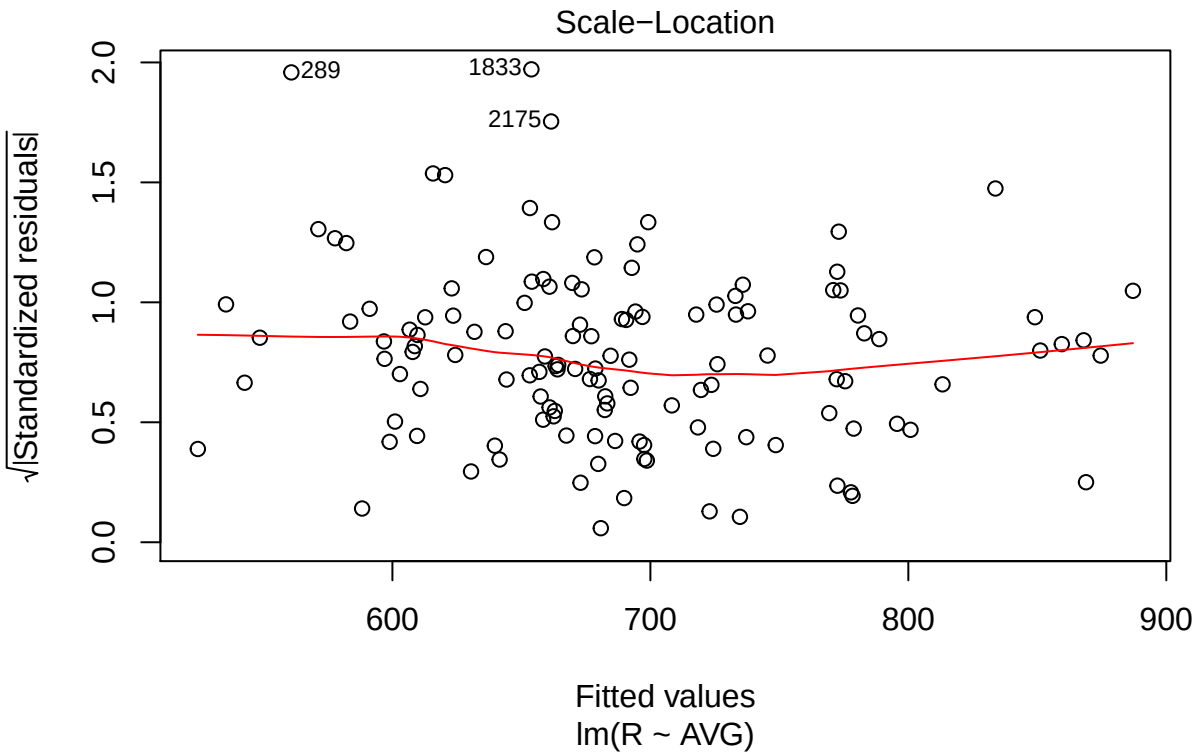
```
## [1] 17.0 31.5 37.0 42.5 49.0
```

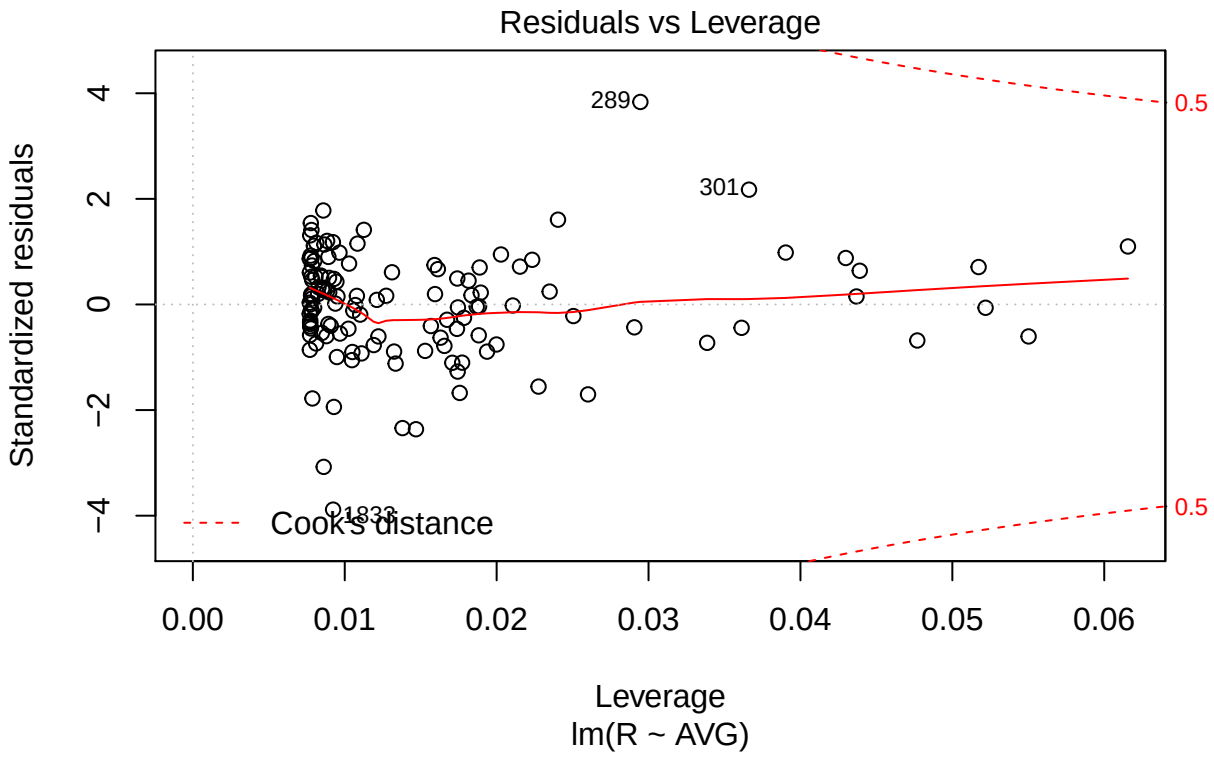
2. Finding outliers (Rule of Thumb)

```
library(Lahman)
a = subset(Teams, teamID=='PIT')
a$AVG = a$H / a$AB
b = lm(R~AVG, data=a)
plot(b)
```


Residuals vs Fitted
Residuals
Fitted values
lm(R ~ AVG)
289
2175
1833
300
200
100
0
−100
−300
600
700
800
900

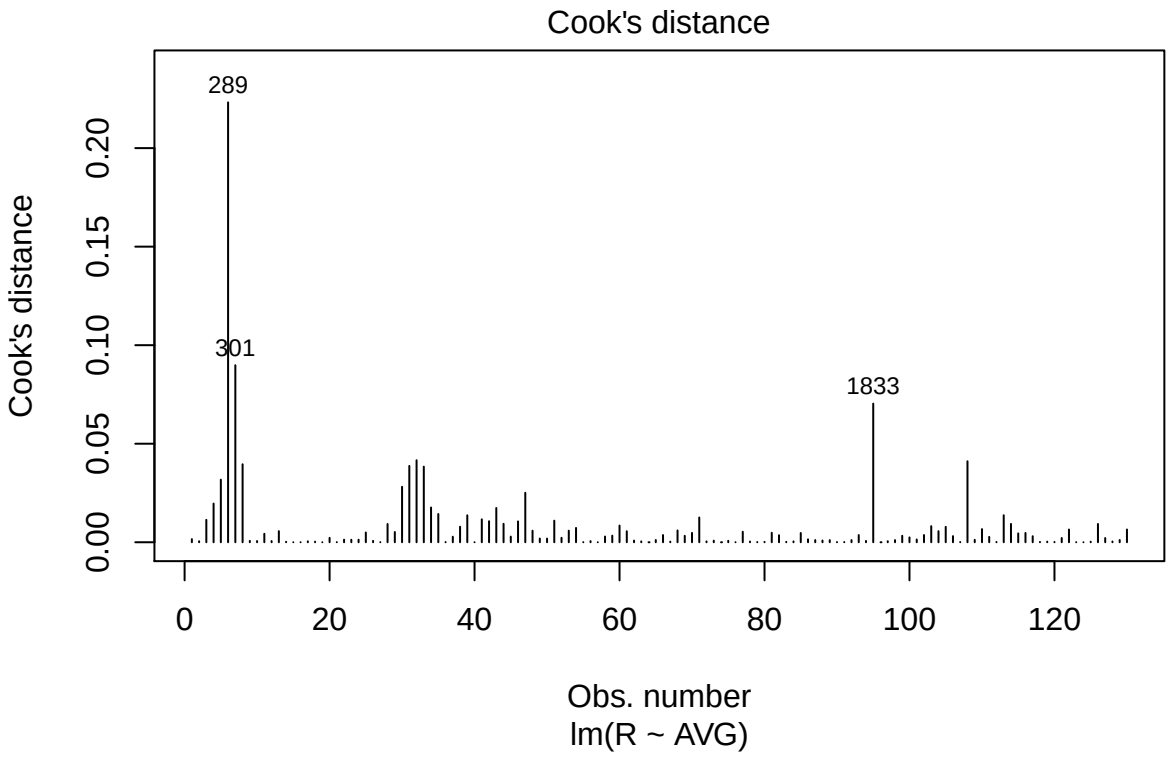

Normal Q−Q
Standardized residuals
4
2
0
−2
−4
289
2175
1833
−2
−1
0
1
2
Theoretical Quantiles
lm(R ~ AVG)


Scale−Location
289
1833
2175
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
600
700
800
900
Fitted values
lm(R ~ AVG)



1. 등분산성을 확인할 수 있는 적합성 그래프
2. 첨도와 왜도를 확인할 수 있는 정규확률도(또는 Q-Q 플롯)
3.
4. 영향력을 파악하는 표준잔차도표

```
plot(b, which=4)
```

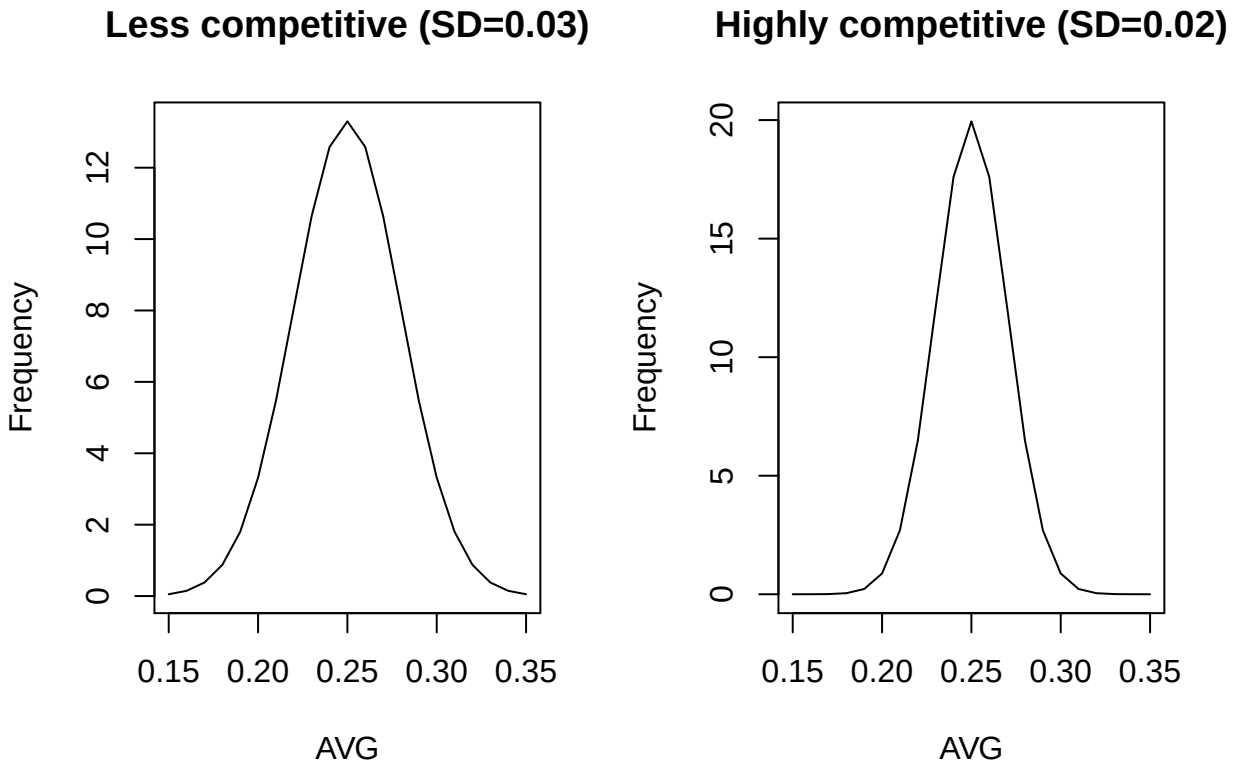


영향력 여부를 정확히 따져보기 위해서는 쿡의 거리만을 보여주는 도표를 활용하면 좀 더 정확한 확인이 가능하다. 특정 관측점이 0.5 이상이면 영향력이 있고, 1보다 크면 상단한 영향력이 있다고 본다.

3. The power of Standard deviation

선수층이 얇아 선수 타율의 분포가 넓을 경우, 어떤 선수의 기록이 평균보다 제법 높다고 해서 월등히 뛰어난 선수라고 주장하기는 힘들다. 반면에 선수층이 단단해서 선수들의 기록 분포가 매우 좁고 평균 부분이 매우 예리하게 모여 있는 타율 분포의 경우, 평균을 약간 상위한다고 해도 뛰어난 선수가 아니라고 자신있게 말할 수도 없다.

```
par(mfrow=c(1,2))
x = seq(0.15, 0.35, 0.01)
y = dnorm(x, 0.25, 0.03)
plot(x, y, xlab='AVG', ylab='Frequency',
     main='Less competitive (SD=0.03)', type='l')
t = seq(0.15, 0.35, 0.01)
u = dnorm(x, 0.25, 0.02)
plot(t, u, xlab='AVG', ylab='Frequency',
     main='Highly competitive (SD=0.02)', type='l')
```



표준점수를 이용한 비교분석을 위해서는 두 가지 조건이 충족돼야 한다. 1. 정규분포를 이뤄야 된다는 조건 2. 모집단의 평균과 표준편차가 공개돼야 한다는 조건 @ Teaching Statistics using baseball, Albert, J.(2003)