SEOUL highschool weekend study time

▼ Problem Description

▼ Search for What?

서울 소재 3개의고등학교(영일고등학교, 강서고등학교, 신목고등학교) 주말공부 시간을 조사했을때, 공부 시간에 차이가 있는가를 확인해보겠다.

H0 : 세 고등학교 학생들의 평균 주말 학습 시간은 같다.

H1 : 세 고등학교 학생들의 평균 주말 학습 시간이 같지 않다.

▼ 참고사항

영일고등학교, 강서고등학교는 남자고등학교이며, 신목고등학교는 남녀공학 고등학교이다.

▼ 유의수준

유의수준은 0.05로 설정한다.

▼ Dataset

▼ Column

고등학교 이름, 주말 평균 공부시간, 학생번호, 성별이 열을 구성한다.

평균 공부시간은 3자리에서 반올림하여 저장한다.

 고등학교 이름은 강서고, 신목고, 영일고 순으로 순서를 매겨 factor화 시킨다.

▼ row

총 65명의 학생의 정보가 들어간다.

▼ Structure

```
'data.frame':   65 obs. of  4 variables:
 $ highschool : Ord.factor w/ 3 levels "강서고"<"신목고"<..: 1 1 1 1 1 1 1 1 1 1 ...
 $ avg.weekend: num  12.53 9.35 12.66 12.54 10.83 ...
 $ student.id : int  1 2 3 4 5 6 7 8 9 10 ...
 $ gender     : Ord.factor w/ 2 levels "M"<"F": 1 1 1 1 1 1 1 1 1 1 ...
```

▼ ANOVA analysis

▼ 정규성 검사

```
> shapiro.test(highschool$avg.weekend[highschool == "영일고"])

        Shapiro-Wilk normality test

data:  highschool$avg.weekend[highschool == "영일고"]
W = 0.97323, p-value = 0.6309

> shapiro.test(highschool$avg.weekend[highschool == "신목고"])

        Shapiro-Wilk normality test

data:  highschool$avg.weekend[highschool == "신목고"]
W = 0.92906, p-value = 0.2642

> shapiro.test(highschool$avg.weekend[highschool == "강서고"])

        Shapiro-Wilk normality test

data:  highschool$avg.weekend[highschool == "강서고"]
W = 0.95453, p-value = 0.4412
```

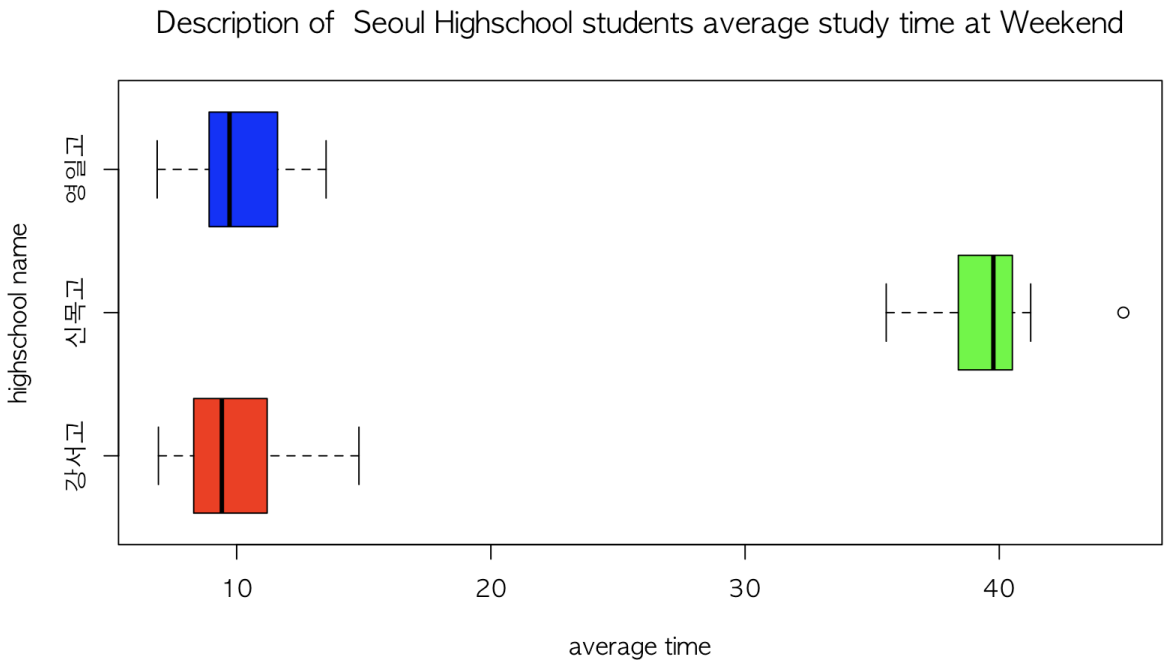
→ 세 고등학교의 shapiro-wilk 검사를 실시했고 , 세개의 검사 모두 p-value > 0.05

즉, 귀무가설을 채택하며, 이는 모두 정규분포를 따른다는 것을 알 수 있다.

▼ 기술 통계

```
# A tibble: 3 × 7
  highschool count  mean median   max   min    sd
  <ord>      <int> <dbl>  <dbl> <dbl> <dbl> <dbl>
1 강서고        20  10.0   9.42  14.8  6.92  2.04
2 신목고        15  39.6  39.8   44.9 35.6   2.07
3 영일고        30  10.1   9.72  13.5  6.87  1.71
```

▼ Boxplot



▼

▼ Variance testing.

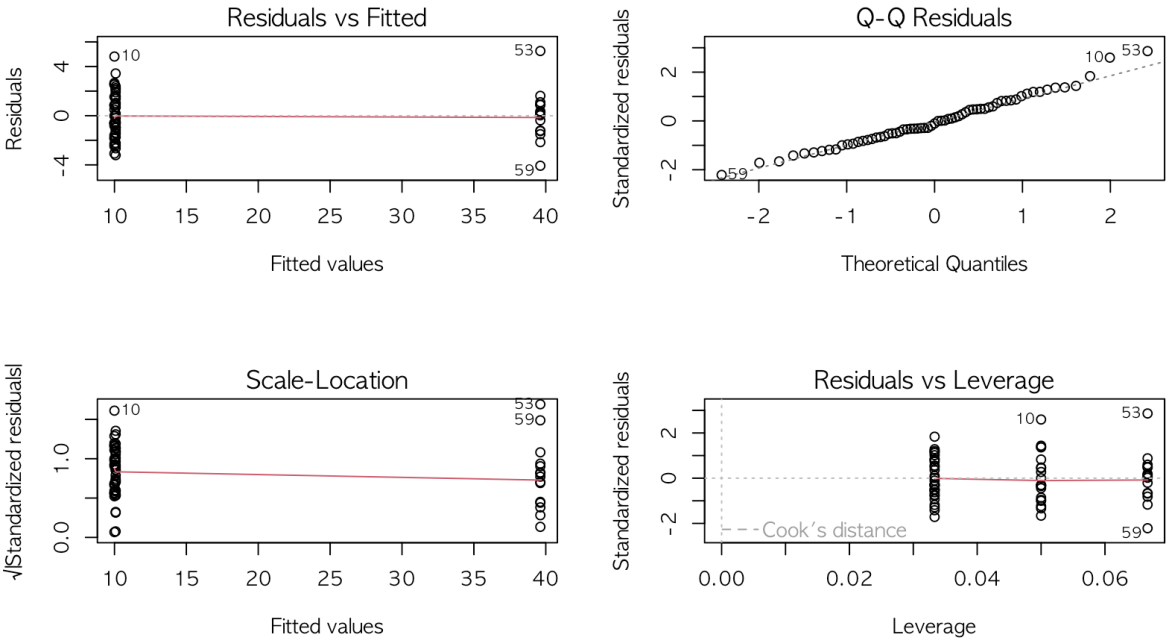
AOV 검정한 result를 토대로, 좀 더 디테일한 분석을 실시한다.

총 네가지 그래프를 순서대로 분석

1. 관측값과 예측값의 차이를 그린 첫번째 그래프에서 양 끝에 극단치를 제외하고는 평이하게 점선은 고르게 분포되어있다는 점을 알 수 있다.
2. 정규성을 나타낸 q-q plot. 직선 근처에 데이터들이 밀집되어있다. 정규성을 확인할 수 있다.
3. 관측갑과 예측값의 차이를 양수값으로 나타내준 그래프이다. 첫번째 그래프와 비슷한 양상을 그린다.
4. Cook's distance을 넘어간 값들이 상당히 보인다.

→ 결론적으로 이 데이터는 등분산성을 만족한다. 하지만 이상치가 상당히 많다. 이는 신목고의 평균시간이 다른 고등학교보다 유의미하게 높다는것을 증명한다.

허나 세가지 모든 고등학교 모두 정규분포를 따른다는 점 또한 중요하다.



▼ Conclusion.

▼ result

```
> result= aov(avg.weekend~highschool, data = highschool)
> summary(result)
            Df Sum Sq Mean Sq F value Pr(>F)
highschool   2  10092    5046    1394 <2e-16 ***
Residuals   62    224       4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

▼ F-value & P-value

F-value는 1394로 엄청나게 큰 값이 나왔으며, P-value는 2e-16로 0.05보다 한참 작은 값이 나왔다.

즉, 설정했던 귀무가설의 대한 근거가 부족해도 한참 부족하다.

귀무가설을 기각하고 대립가설을 채택한다.

결론적으로 서울 소재 세가지 고등학교의 주말 평균 시간은 같지 않다.

▼ Finding causes.

Tukey를 활용하여 평균이 달라지게 된(대립가설을 채택하게 된) 원인을 파악한다.

```
> TukeyHSD(result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = avg.weekend ~ highschool, data = highschool)

$highschool
                   diff        lwr        upr     p adj
신목고-강서고  29.62548  28.065119  31.185848 0.0000000
영일고-강서고   0.08565  -1.233099   1.404399 0.9866834
영일고-신목고 -29.53983 -30.984450 -28.095217 0.0000000
```

▼ 영일고 vs 강서고

p-value가 월등히 높다. 즉 이 둘은 매우 비슷한 평균치를 지니고 있다.

▼ 신목고 vs 강서고

신목고 - 강서고 에서 29.62로 꽤 나 큰 양수값이 나왔다. 또한 p-value는 0에 수렴한다. 즉, 이 결과는 유의미한 차이가 있으며, 신목고의 주말 평균 시간이 강서고의 평균 시간보다 월등히 높다는 사실을 알 수 있다.

▼ 영일고 vs 신목고

영일고 - 신목고 에서는 -29.53이 나온 상당히 작은 음수값이 나왔다. 여기도 마찬가지로 p-value는 0에 수렴한다. 이 결과 또한 유의미한 차이가 있으며, 신목고 평균 주마라 평균 시간이 영일고의 평균 시간보다 월등히 높다는 사실을 알 수 있다.

written by Choi Wonbin (MJU 60203042)