R Studio를 이용한 데이터 분석

하이미디어

전주환

INDEX

One Sample T-test (단일표본 T검정)

서울시의 월별 미세먼지 수치가 (WHO기준) 보통(40) 수준에 달하는지 검사하고자 한다.

서울 열린 데이터 광장 - 서울시 대기오염 통계
2015.01 ~ 2019.12 (단위: μg/m³/년)

데이터 불러오기

```
> # csv 파일 불러오기
> ddata = read.csv('미세먼지.csv', header = T, na.strings = '.')
> # 전처리(날짜 데이터 제거)
> ddata = ddata[,2]
> str(ddata)
 int [1:60] 49 84 71 45 45 35 30 34 28 44 ...
```

기본 통계치 확인

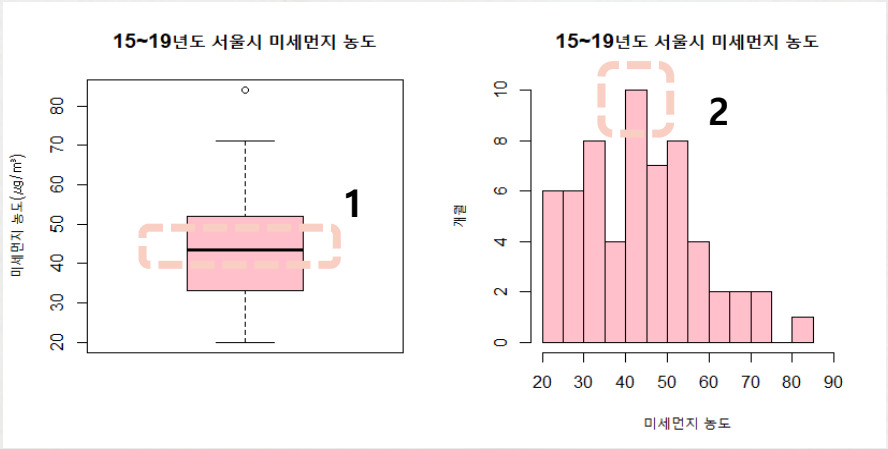
```
> # 기본 통계 확인
> library(psych)
> describe(ddata)
   vars  n mean    sd median trimmed   mad min max range skew kurtosis
X1    1 60 43.7 14.15   43.5      43 14.08  20  84    64 0.42    -0.21
    se
X1 1.83
```

통계분석(t-test)

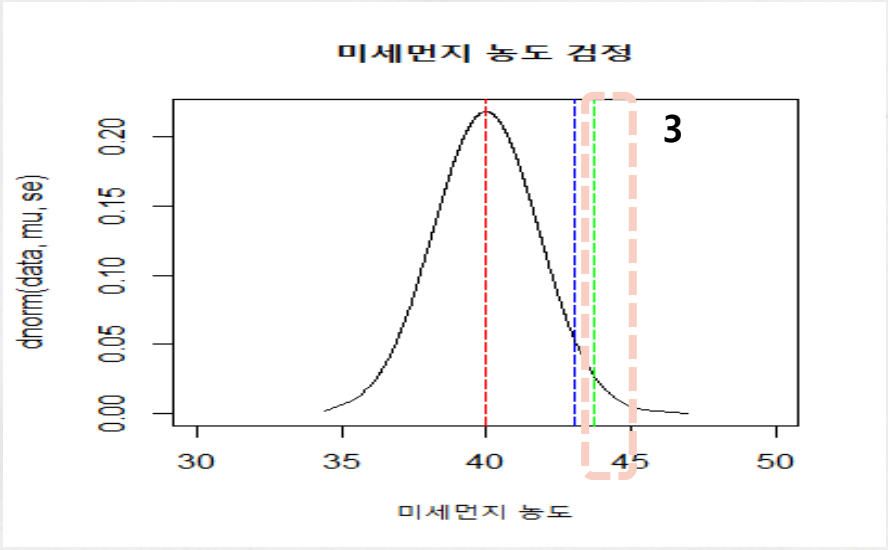
```
        One Sample t-test

data:  ddata
t = 2.0248, df = 59, p-value = 0.02371
alternative hypothesis: true mean is greater than 40
95 percent confidence interval:
 40.64635      Inf
sample estimates:
mean of x
     43.7
```

4

Boxplot, Histogram을 통한 데이터 분석



통계 분석 결과 그래프



= C(1 : 4)

1. 표본의 15~ 19년도 미세먼지 농도의 월별 평균은 40~50정도임을 알 수 있다.

2. 표본에서 미세먼지 농도의 월별 평균이 40이상인 달이 많은 것을 알 수 있다.

3. WHO기준 보통 수준인 미세먼지 농도 40μg/m³의 신뢰구간 밖에 표본 집단의 평균이 포함되어 있음.

4. P값(0.02371)이 0.05보다 작으므로 WHO기준 미세먼지 농도 보통 수준보다 좋지 않다는 연구가설을 채택.

결론

15~19 년도 서울시의 대기오염 정도는 통계적으로 WHO기준 미세먼지 보통(40 μg/m³) 수준보다 좋지 않다고 볼 수 있다.

2020년 12월 강남구의 유동인구 차이를 성별로 검사하고자 한다.

SKT 데이터 허브 – 서울시 유동인구 데이터
2020. 12

데이터 불러오기

```
> # 불러오기
> fp <- read.csv('fp202012.csv', header = T, na.strings = '.')
> # 전처리
> fp$성별 <- factor(fp$성별, levels = c(1, 2), labels = c('남성', '여성'))
> str(fp)
'data.frame':   62 obs. of  2 variables:
 $ 성별       : Factor w/ 2 levels "남성","여성": 1 1 1 1 1 1 1 1 1 1 ...
 $ 유동인구수: int  7379190 7359180 7310400 7257340 5773890 5361690 7275850 7227400 7187060 7185450 ...
```

기본 통계치 확인

```
> describeBy(유동인구수, 성별, mat = T)
    item group1 vars  n    mean       sd  median trimmed      mad
X11    1   남성    1 31 6605180 794845.3 7047700 6682002 310812.3
X12    2   여성    1 31 6940110 717330.9 7301530 7011442 304348.1
        min     max   range       skew  kurtosis       se
X11 5156950 7379190 2222240 -0.8211900 -1.148466 142758.4
X12 5587790 7656920 2069130 -0.8315933 -1.052090 128836.4
```

등분산 검정(var.test)

```
> var.test(유동인구수 ~ 성별, fp)

        F test to compare two variances

data:  유동인구수 by 성별
F = 1.2278, num df = 30, denom df = 30, p-value = 0.5776
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5920102 2.5463796
```

2

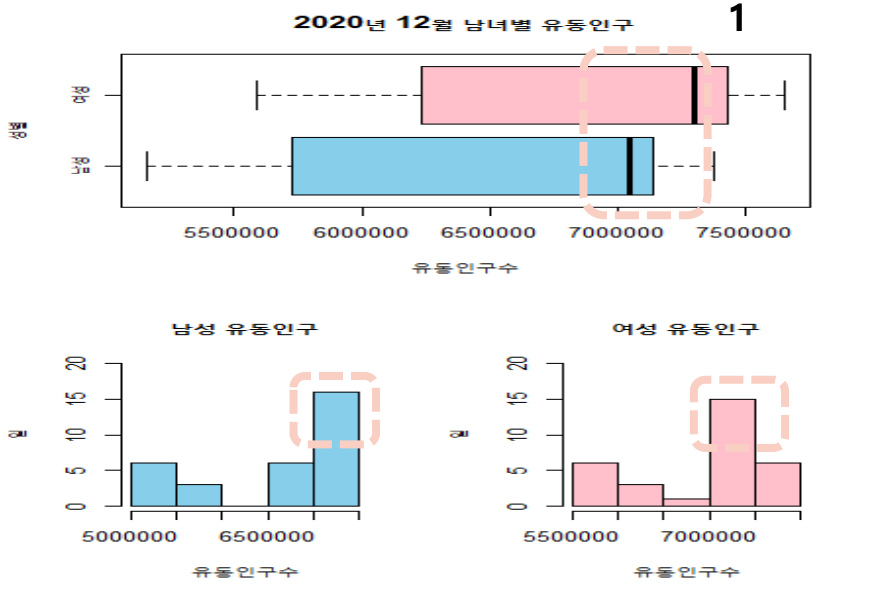
T검정(t.test)

```
        Two Sample t-test

data:  유동인구수 by 성별
t = -1.7417, df = 60, p-value = 0.08668
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -719585.03   49724.39
```

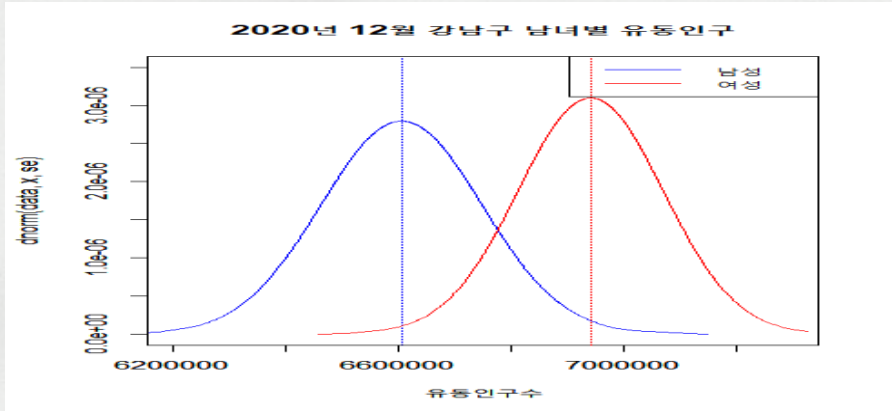
3

Boxplot, Histogram을 통한 데이터 분석

1



통계 분석 결과 그래프



= C(1 : 3)

1. 12월의 강남구의 유동인구 수는 여성이 남성보다 높다는 것을 알 수 있다.

2. 등분산 검정 결과 p값이 0.5776으로 0.05 보다 높아 등분산임을 알 수 있다.

3. t검정을 한 결과 p값이 0.08668로 0.05보다 크므로 남녀별 유동인구 수의 차이는 없다는 귀무가설을 택함.

결론

2020년 12월 한달 간 남녀별 유동인구 수는 통계적으로 유의한 차이가 없었으며, 같은 비율로 이동하였다는 것을 알 수 있다.

2019년 12월의 강남역 승하차 인원수와 2020년 12월의 강남역 지하철 승하차 인원수가 다른지 검사 하고자 한다.

서울 열린 데이터 광장 - 서울시 지하철 호선별 역별 승하차 인원 정보
2019.12, 2020.12

데이터 불러오기

```
> # 불러오기
> sw <- read.csv('swdata_pst.csv', header = T, na.strings = '.')
> str(sw)
'data.frame':   31 obs. of  2 variables:
 $ X201912: int  102073 210942 215884 216803 224464 244040 187587 103797 209880 217355 ...
 $ X202012: int  141760 138849 136637 145035 81997 43695 136959 126968 123615 123527 ...
>
```

기본 통계치 확인

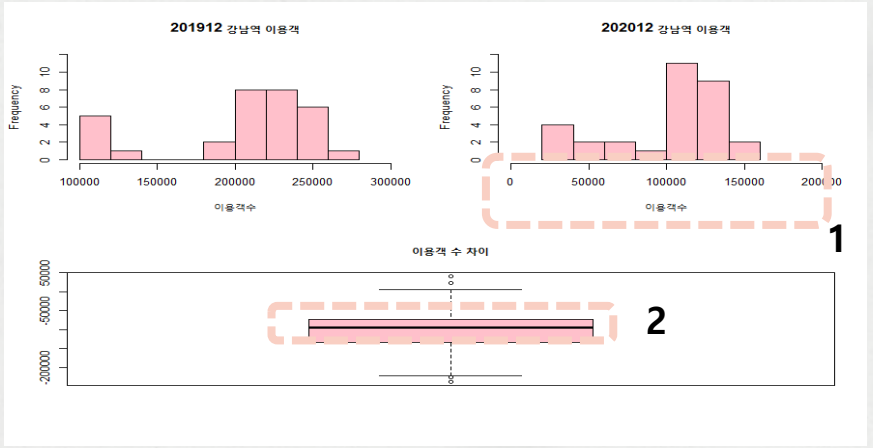
```
> #기본 통계치 확인
> library(psych)
> describe(sw)
        vars  n      mean       sd median  trimmed      mad    min    max  range skew kurtosi      se
X201912    1 31 204780.9 49467.05 217355 210003.0 33122.77 102073 269685 167612 -1.03    -0.3 8884.54
X202012    2 31 102189.8 36660.72 116924 105930.5 14891.23  29199 145035 115836 -0.89    -0.7 6584.46
> dif = c(X202012 - X201912)    # 2020년과 2019년의 승하차 인원수 차이
> describe(dif)
   vars  n      mean       sd median trimmed      mad     min   max  range skew kurtosis      se
X1    1 31 -102591.1 71542.74 -95602 -102456 44181.48 -237079 39687 276766 0.01    -0.54 12849.46
```

통계분석(t-test)

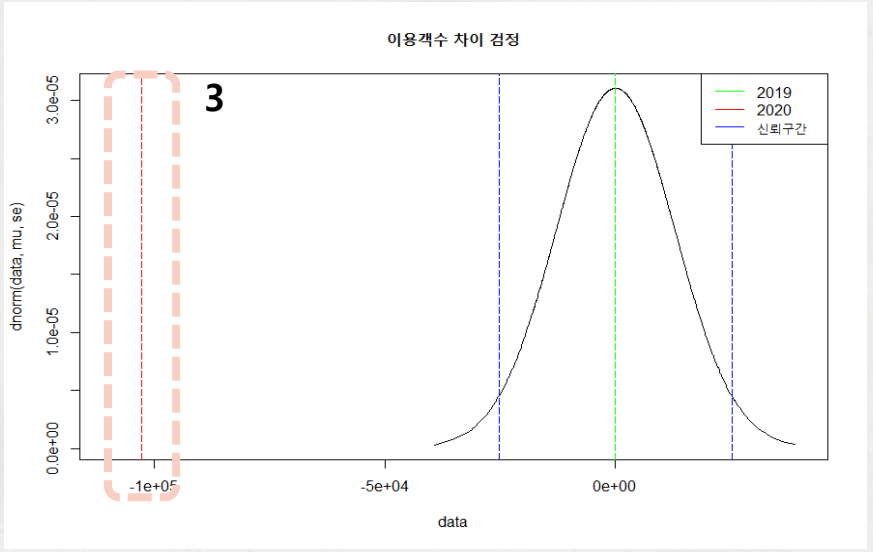
```
        Paired t-test

data:  X202012 and X201912
t = -7.9841, df = 30, p-value = 6.53e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -128833.22  -76349.04
sample estimates:
mean of the differences
              -102591.1
```

4

Boxplot, Histogram을 통한 데이터 분석



통계 분석 결과 그래프



= C(1 : 4)

1. 2019년 12월에 비해 2020년 12월의 강남역 지하철 이용객이 급감하였음을 알 수 있다.

2. 지하철 이용객 수의 차이 정도를 상자 그림으로 알 수 있다.

3. 2020년 12월

4. 다른 집단에서의 시간에 따른 평균 차이 검정이므.
P값이 0.05보다 작으므로 2019년과 2020년의 강남역 이용객 수는 차이가 있다.

결론
2020년 12월은 2019년 12월에 비해 강남역 이용객 수가 줄었음을 알 수 있다. P값이 0.05보다 작기 때문에 연구가설을 채택.

2~50대의 연령별 가해 사고 건수에 차이가 있는지 검사해 보고자 한다.

공공 데이터 포털- 도로교통공단_가해운전자 연령층별 월별 교통사고 통계

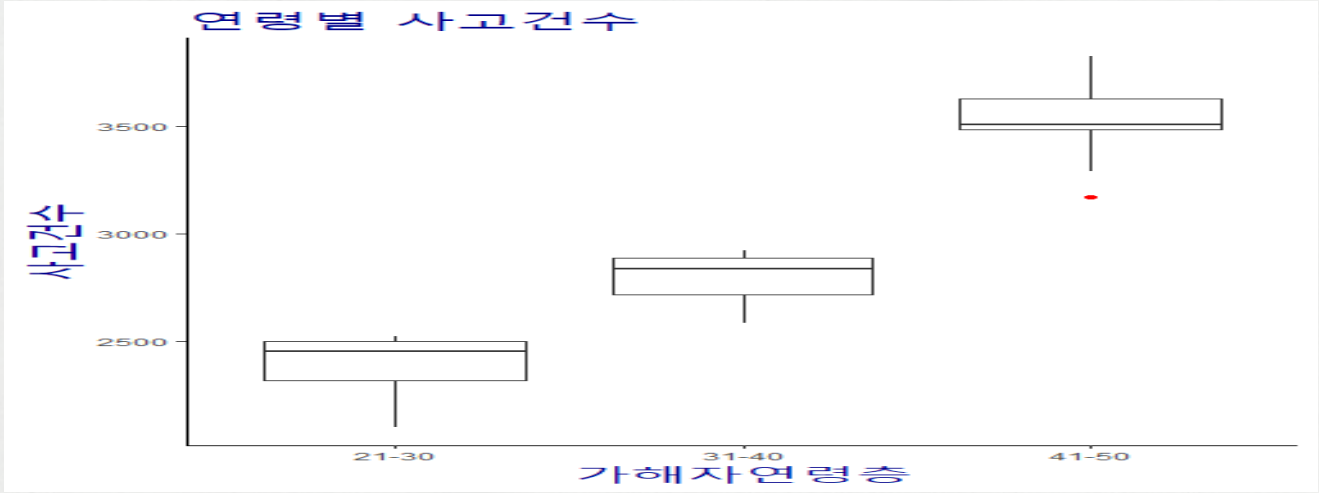
데이터 불러오기

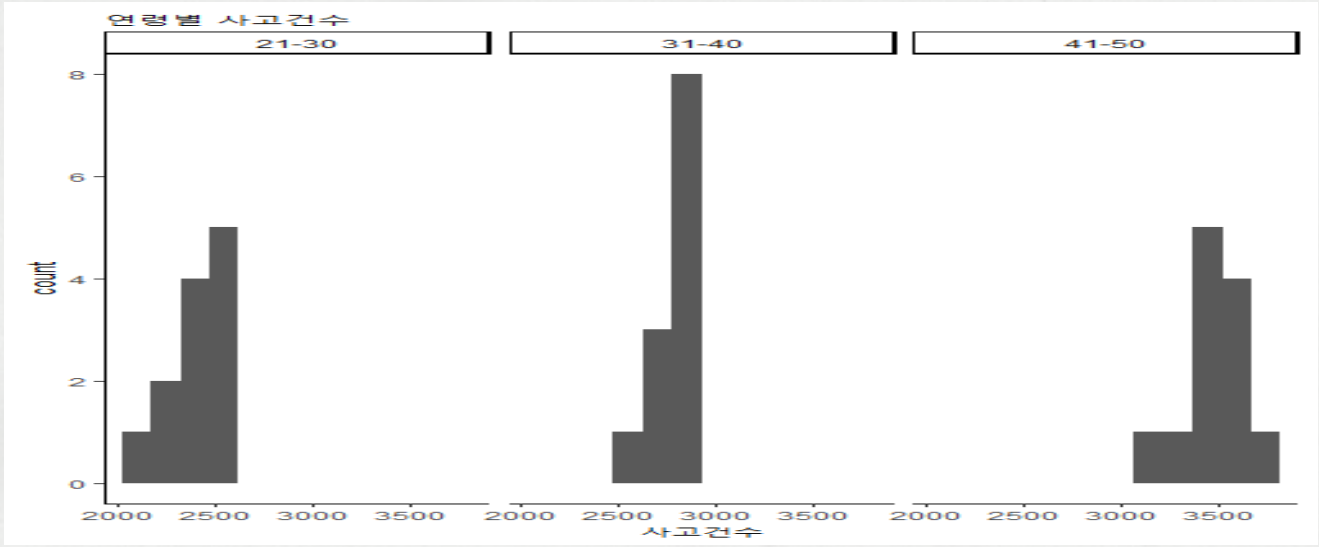
```
> # 불러오기
> df <-  read.csv('연령별사고건수.csv', header = T, na.strings = '.')
> # 전처리
> df$가해자연령층 <- factor(df$가해자연령층, levels = c('21-30세', '31-40
세', '41-50세'), labels = c('21-30', '31-40', '41-50'))
> str(df)
'data.frame':   36 obs. of  3 variables:
 $ 가해자연령층: Factor w/ 3 levels "21-30","31-40",..: 1 1 1 1 1 1 1 1 1
 1 ...
 $ 월          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ 사고건수    : int  2200 2101 2202 2354 2509 2448 2481 2523 2441 2513
 ...
```

기본 통계치 확인

```
> describeBy(사고건수, 가해자연령층, mat = T)
    item group1 vars  n     mean       sd median trimmed      mad
X11    1  21-30    1 12 2393.917 145.6106 2454.0  2410.3  84.5082
X12    2  31-40    1 12 2801.500 109.0017 2834.5  2811.2  97.1103
X13    3  41-50    1 12 3517.000 169.0352 3507.0  3521.0 137.8818
     min  max range       skew   kurtosis       se
X11 2101 2523   422 -0.8623609 -0.9639654 42.03416
X12 2585 2921   336 -0.6523137 -1.1083409 31.46607
X13 3171 3823   652 -0.3546004 -0.3104110 48.79627
```

Boxplot, Histogram을 통한 데이터 분석





2~50대의 연령별 가해 사고 건수에 차이가 있는지 검사해 보고자 한다.

공공 데이터 포털- 도로교통공단_가해운전자 연령층별 월별 교통사고 통계

등분산 검정(Bartlett.test, LeveneTest)

```
> # 통계분석
> bartlett.test(사고건수 ~ 가해자연령층, df)

        Bartlett test of homogeneity of variances

data:  사고건수 by 가해자연령층
Bartlett's K-squared = 1.981, df = 2, p-value = 0.3714

> leveneTest(사고건수 ~ 가해자연령층, df)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.2999 0.7429
      33
```

ANOVA 분석

```
> owa_result <- aov(사고건수 ~ 가해자연령층, df)
> summary(owa_result)
              Df   Sum Sq Mean Sq F value Pr(>F)
가해자연령층   2 7757522 3878761   188.7 <2e-16
Residuals     33  678224   20552

가해자연령층 ***
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

사후검정(Duncan /Scheffe/TukeyHSD)

```
Duncan's new multiple range test
for 사고건수

Mean Square Error:  20552.24

가해자연령층,  means

      사고건수       std  r  Min  Max
21-30 2393.917 145.6106 12 2101 2523
31-40 2801.500 109.0017 12 2585 2921
41-50 3517.000 169.0352 12 3171 3823

Alpha: 0.05 ; DF Error: 33

Critical Range
       2        3
119.0734 125.1587

Means with the same letter are not significantly different.

      사고건수 groups
41-50 3517.00      a
31-40 2801.50      b
21-30 2393.91      c
```

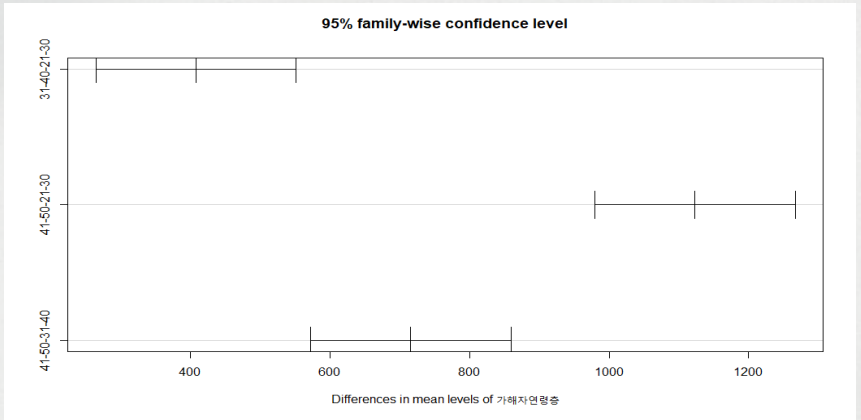
```
Scheffe Test for 사고건수

Mean Square Error  : 20552.24

가해자연령층,  means

      사고건수       std  r  Min  Max
21-30 2393.917 145.6106 12 2101 2523
31-40 2801.500 109.0017 12 2585 2921
41-50 3517.000 169.0352 12 3171 3823

Alpha: 0.05 ; DF Error: 33
Critical Value of F: 3.284918

Comparison between treatments means

              Difference pvalue sig      LCL      UCL
21-30 - 31-40    -407.5833      0 *** -557.5971 -257.5695
21-30 - 41-50   -1123.0833      0 *** -1273.0971 -973.0695
31-40 - 41-50    -715.5000      0 *** -865.5138 -565.4862
```

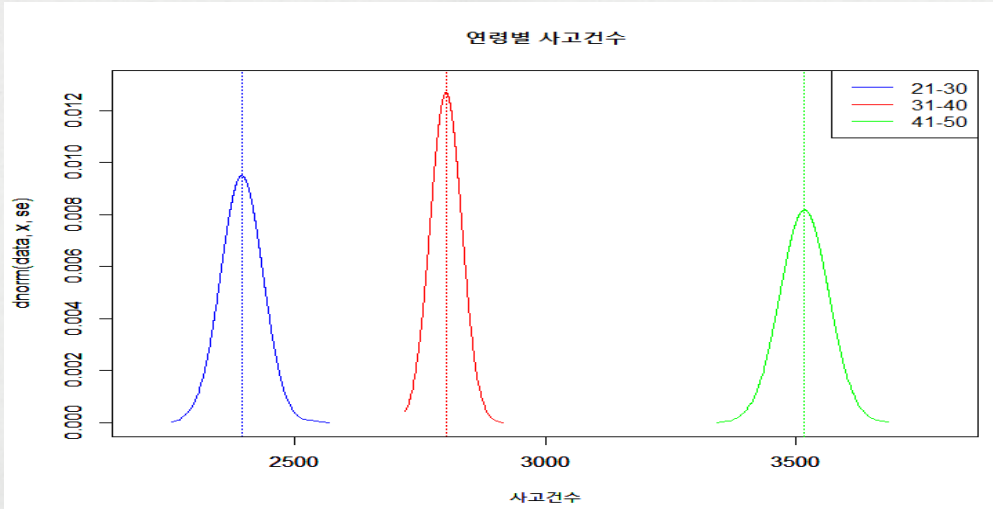
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = 사고건수 ~ 가해자연령층, data = df)

$가해자연령층
                diff       lwr       upr p adj
31-40-21-30  407.5833  263.9709  551.1958 2e-07
41-50-21-30 1123.0833  979.4709 1266.6958 0e+00
41-50-31-40  715.5000  571.8875  859.1125 0e+00
```

그룹별 표본수가 같으므로(Duncan Test)

- 주어진 자료에서 표본의 수는 월별 데이터로 동일하므로 채택함.
- (표본의 수가 다르면 Scheffe Test)

통계 분석 결과 그래프





결론

ANOVA 분석에서 p값이 2e-16으로 0.05보다 작으므로 21~50대의 연령별 사고 건수가 같을 것이라는 귀무가설을 기각하고, 연령대 별로 차이가 있다는 연구가설을 채택한다.

2019년 12월, 2020년 7월, 2020년 12월의 강남역 지하철 승하차 인원수가 다른지 검사 하고자 한다.

서울 열린 데이터 광장 - 서울시 지하철 호선별 역별 승하차 인원 정보
2019.12, 2020.06, 2020.12

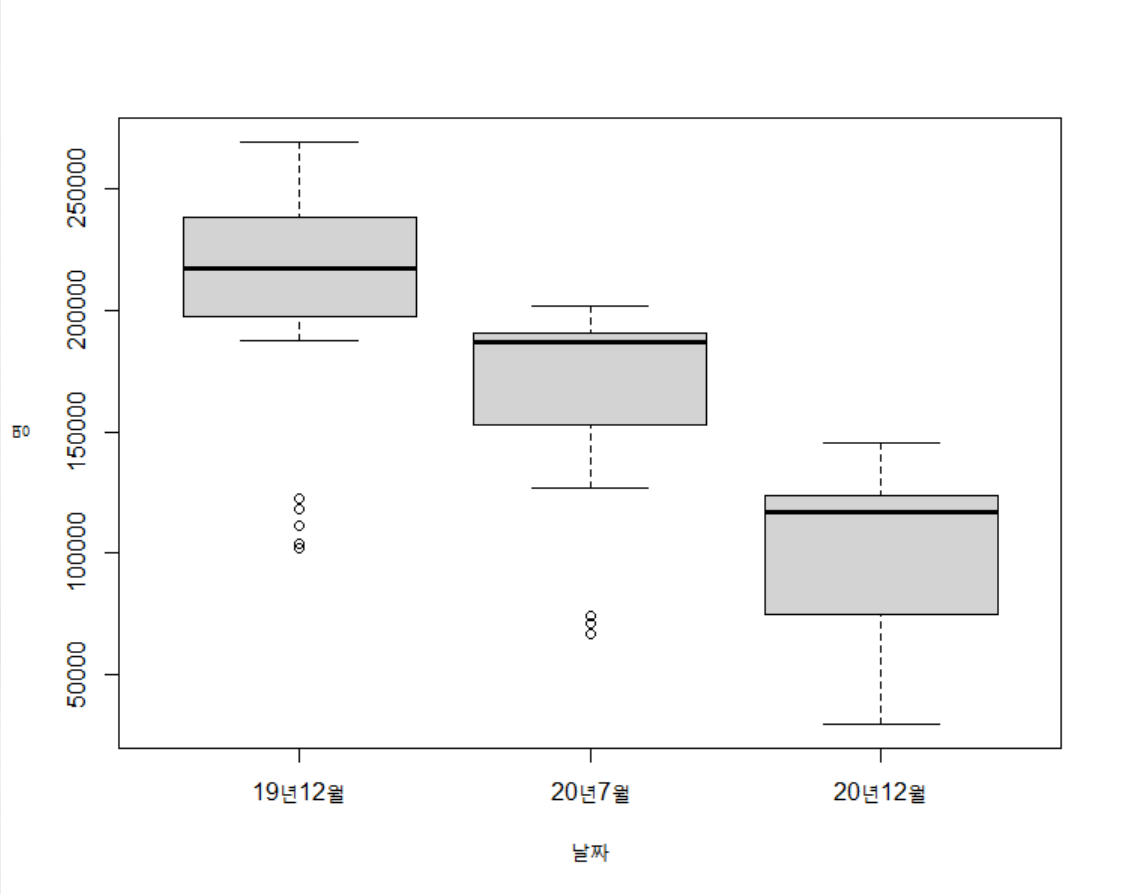
기본 통계치 확인

```
> describeBy(이용객수, 날짜, mat = T)
   item   group1 vars  n     mean      sd median trimmed     mad   min   max range     skew kurtosis      se
X11    1 19년12월    1 31 204780.9 49467.05 217355 210003.0 33122.77 102073 269685 167612 -1.0309560 -0.321876 8884.544
X12    2  20년7월    1 31 166368.5 42773.30 186678 173808.4  9973.45  66611 201751 135140 -1.3637150 0.348936 7682.311
X13    3 20년12월    1 31 102189.8 36660.72 116924 105930.5 14891.23  29199 145035 115836 -0.8949553 -0.754865 6584.460
```

구형성 검정 (Mauchly's test)

```
Mauchly Tests for Sphericity

        Test statistic  p-value
date.f         0.76221 0.019501


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

          GG eps Pr(>F[GG])
date.f 0.80789  4.987e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           HF eps   Pr(>F[HF])
date.f 0.8471101 2.080247e-10
```

Boxplot을 통한 데이터 분석



2019년 12월, 2020년 7월, 2020년 12월의 강남역 지하철 승하차 인원수가 다른지 검사 하고자 한다.

서울 열린 데이터 광장 - 서울시 지하철 호선별 역별 승하차 인원 정보
2019.12, 2020.06, 2020.12

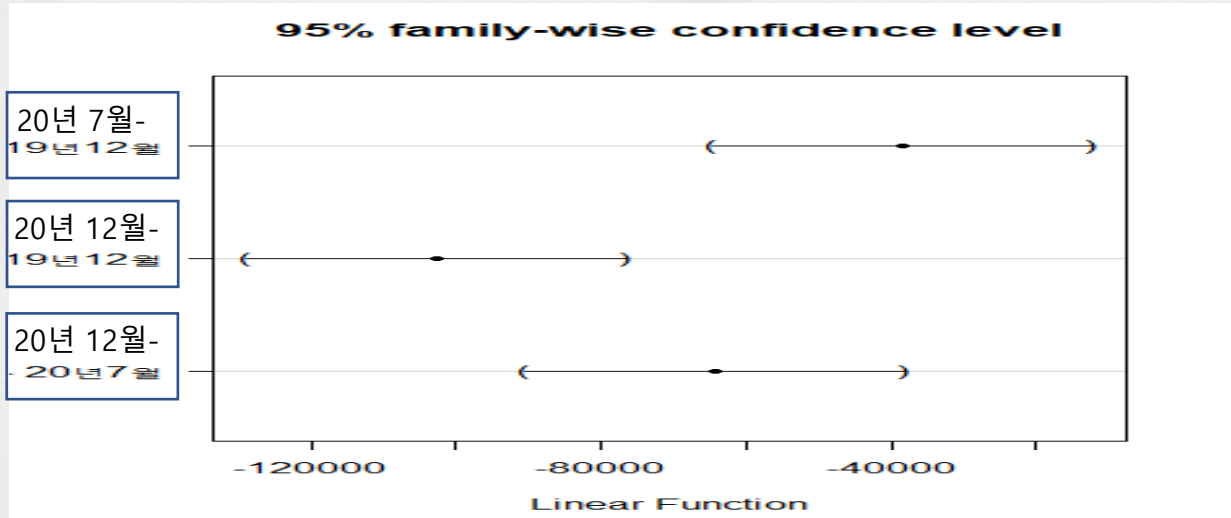
ANOVA 분석

```
> summary(rma.result)
            Df    Sum Sq   Mean Sq F value   Pr(>F)
날짜         2 1.666e+11 8.328e+10   44.45 3.74e-14 ***
Residuals   90 1.686e+11 1.874e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

사후 검정 (TukeyHSD)

```
> summary(tukey_result)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = 이용객수 ~ 날짜, data = sw)

Linear Hypotheses:
                          Estimate Std. Error t value Pr(>|t|)
20년7월 - 19년12월 == 0    -38412      10994  -3.494  0.00211
20년12월 - 19년12월 == 0   -102591      10994  -9.331  < 1e-04
20년12월 - 20년7월 == 0     -64179      10994  -5.838  < 1e-04

20년7월 - 19년12월 == 0  **
20년12월 - 19년12월 == 0 ***
20년12월 - 20년7월 == 0  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

통계 분석 결과 그래프(Tukeyplot)



결론

2019년 12월, 2020년 6월, 2021년 12월의 강남역 지하철 이용객 수는 차이가 없다는 귀무가설을 기각하고 차이가 있다는 연구가설을 채택한다.

사후 검정(TukeyHSD) 결과 19년 12월, 20년 7월, 20년 12월 시간이 갈수록 승하차 인원수가 줄어든 것을 확인할 수 있다.