

R 데이터(텍스트) 마이닝 및 시각화 분석

Seongmin Mun

24th January 2022



아주대학교
AJOU UNIVERSITY

Outline

통계학이란?

평균분석(T-test)

산점도(Scatter plot)

통계학이란?

통계학의 종류

- ▶ 관심대상이 되는 자료를 수집하고 정리, 요약 하여 불확실한 사실에 대하여 과학적인 판단의 기준을 제시해주는 학문
- ▶ **기술통계학(Descriptive statistics)**
 - ▶ 관심의 대상이 되는 자료에 대해 그림 및 수치를 사용하여 정리하고 요약하는 방법
 - ▶ 추론통계를 위한 사전단계로 수집된 자료의 분석에 초점을 둔다.
 - ▶ 수치를 활용한 방법 : 비율, 지수, 평균, 분산 등
 - ▶ 그림을 활용한 방법 : 막대그래프, 히스토그램, 상자그림 등

자료의 종류

- ▶ 수치형 자료(Quantitative Data)
 - ▶ 숫자로 표현되어 있는 자료
 - ▶ 계량적 자료, 정량적 자료
 - ▶ (e.g., 등간척도, 서열척도, 비율척도)
- ▶ 명목형 자료(Qualitative Data)
 - ▶ 숫자로 표시될 수 없는 자료
 - ▶ 범주형 자료, 정성적 자료
 - ▶ (e.g., 명목척도)

척도의 종류

- ▶ **명목척도(Nominal Scale)**
 - ▶ 상하관계는 없고 구분만 있는 척도
 - ▶ (e.g., 성별, 지역, 날씨 등)
- ▶ **서열척도(Ordinal Scale)**
 - ▶ 순서에 의해 부여되는 척도
 - ▶ (e.g., 직위, 학력, 등수 등)

척도의 종류

- ▶ 등간척도(Interval Scale)
 - ▶ 간격이 일정하여 가감승제가 가능하지만 절대0점이 존재하지 않는 척도
 - ▶ (e.g., 시간, 온도 등)
- ▶ 비율척도(Ratio Scale)
 - ▶ 등간척도와 비슷하지만 절대 0점이 존재하는 척도
 - ▶ (e.g., 성적, 금액 등)

평균분석(T-test)

개념(Concept)

- ▶ 두 집단간 모평균에 차이가 있는지 없는지를 검증하고자 할 때 사용한다.
- ▶ 분산이 같은 경우와 다른 경우의 검정 기각역이 다르다.
- ▶ 가설을 세운 후 검증하는 연역적인 접근방법이다.
- ▶ T-test에 사용되는 T값(검정 통계량)이 T분포에서 문제 상황에 해당하는 T기준 값보다 크면 대립가설을 채택한다.
- ▶ 분산 분석의 F값(검정 통계량)이 F분포의 임계값보다 크면 대립가설을 채택한다.

신뢰구간(Confidence interval)

- ▶ 실제 모수가 존재할 것으로 예측되는 구간으로 90%, 95%, 99% 정도의 구간 추정이 가능하다.
- ▶ 실제로는 95%신뢰 구간 추정이 통상적으로 사용된다.
- ▶ 95%신뢰구간 : 신뢰구간 내에 0 이 포함될 경우 귀무가설을 채택한다.

가설(Hypothesis)

- ▶ H_0 (귀무가설) : 두 그룹간 유의한 차이가 없다.
- ▶ H_1 (대립가설) : 두 그룹간 유의한 차이가 있다.

F값 & T값

분산 검정에서 검정 통계량(F값) & T-test에서 검정 통계량(T값)

$$F = \frac{s_1^2}{s_2^2}$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

\bar{X} =표본평균, μ =모집단의 평균, S =표본 표준편차, n =표본의 수

데이터 설명

학교(초,중,고)에 다니는 학생들의 키와 몸무게에 대한 데이터

- ▶ X = 번호
- ▶ Sex = 성별
- ▶ ageYear = 연도별 나이
- ▶ ageMonth = 월별 나이
- ▶ heightIn = 키
- ▶ weightLb = 몸무게

데이터 불러오기

```
> #경로확인하기
> getwd()
[1] "/Users/seongminmun/Desktop/2022/Ajou/Data"
> setwd("/Users/seongminmun/Desktop/2022/Ajou/Data")
> heightweight<-read.csv("heightweight.csv",head=T)
> str(heightweight)
'data.frame': 236 obs. of 6 variables:
 $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex    : chr  "f" "f" "f" "f" ...
 $ ageYear: num  11.9 12.9 12.8 13.4 15.9 ...
 $ ageMonth: int  143 155 153 161 191 171 185 142 160 140 ...
 $ heightIn: num  56.3 62.3 63.3 59 62.5 62.5 59 56.5 62 53.8 ...
 $ weightLb: num  85 105 108 92 112 ...
> head(heightweight)
  X sex ageYear ageMonth heightIn weightLb
1 1  f   11.92     143     56.3      85.0
2 2  f   12.92     155     62.3     105.0
3 3  f   12.75     153     63.3     108.0
4 4  f   13.42     161     59.0      92.0
5 5  f   15.92     191     62.5     112.5
6 6  f   14.25     171     62.5     112.0
```

데이터 추출하기

```
> # 데이터 추출하기
>
> group.1=subset(heightweight,sex=="f")
>
> group.2=subset(heightweight,sex=="m")
>
> head(group.1)
  X sex ageYear ageMonth heightIn weightLb
1 1  f   11.92      143      56.3      85.0
2 2  f   12.92      155      62.3     105.0
3 3  f   12.75      153      63.3     108.0
4 4  f   13.42      161      59.0      92.0
5 5  f   15.92      191      62.5     112.5
6 6  f   14.25      171      62.5     112.0
>
> head(group.2)
  X sex ageYear ageMonth heightIn weightLb
112 112  m   13.75      165      64.8      98.0
113 113  m   13.08      157      60.5     105.0
114 114  m   12.00      144      57.3      76.5
115 115  m   12.50      150      59.5      84.0
116 116  m   12.50      150      60.8     128.0
117 117  m   11.58      139      60.5      87.0
```

등분산 검증

등분산성 검증: F값은 0.9513이고 p값은 0.791이므로 두 성별간 분산 값에 차이가 나지 않는다. 등분산성이 성립한다.

```
> # 등분산 검증
> var.test(group.1$weightLb, group.2$weightLb)

      F test to compare two variances

data:  group.1$weightLb and group.2$weightLb
F = 0.95131, num df = 110, denom df = 124, p-value = 0.791
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6618382 1.3729111
sample estimates:
ratio of variances
 0.9513095
```


t-test

평균 차 비교: t값이 -1.636이고 자유도가 234(n-1), p값이 0.1032
이므로 두 그룹간 몸무게는 차이가 나지 않는다. 대립가설(H1:
성별에 따른 몸무게에 차이가 있다.)을 기각, 귀무가설(H0: 성별에
따른 몸무게에 차이가 없다.)을 채택한다.

```
> t.test(group.1$weightLb,group.2$weightLb,var.equal=TRUE)
```

Two Sample t-test

```
data: group.1$weightLb and group.2$weightLb  
t = -1.636, df = 234, p-value = 0.1032  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 -8.8733370  0.8220938  
sample estimates:  
mean of x mean of y  
 98.87838 102.90400
```

t-test

또한 신뢰구간이 $-8.8733370 < \alpha < 0.8220938$ 로 구간에 0을 포함하므로 귀무가설을 채택한다.

```
> t.test(group.1$weightLb, group.2$weightLb, var.equal=TRUE)
```

Two Sample t-test

data: group.1\$weightLb and group.2\$weightLb

t = -1.636, df = 234, p-value = 0.1032

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-8.8733370 0.8220938

sample estimates:

mean of x mean of y

98.87838 102.90400

생각하기

남학생 & 여학생 사이의 키의 차이를 **t-test**로 분석하자

산점도(Scatter plot)

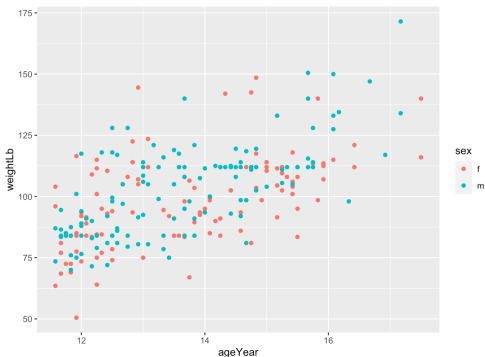
ggplot2설치

```
> install.packages("ggplot2")
URL 'https://cran.rstudio.com/bin/macosx/contrib/4.1/ggplot2_3.3.5.tgz'을 시도합니다
Content type 'application/x-gzip' length 4125542 bytes (3.9 MB)
=====
downloaded 3.9 MB

The downloaded binary packages are in
  /var/folders/f9/tcyhkyhx32q6nxcypnp1rl2c0000gn/T//RtmpGGQtIC/downloaded_packages
> library(ggplot2)
```

나이에 따른 몸무게에 대한 산점도 그리기

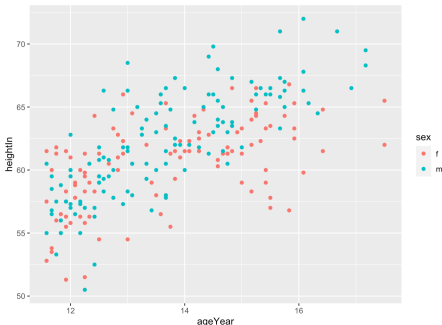
몸무게가 많이 나가는 그룹에 남성이 많고 나이와 몸무게는 비례하는 것을 확인 할 수 있다.



```
> ggplot(heightweight,aes(x=ageYear,y=weightLb,colour=sex))+geom_point()
```

나이에 따른 키에 대한 산점도 그리기

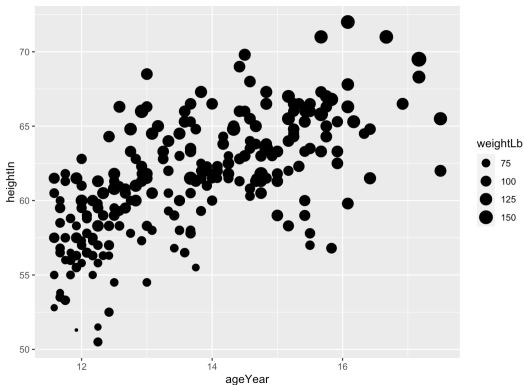
키가 큰 그룹에 남성이 많다. 여성의 경우 14세를 기준으로 키의 성장이 멈춘듯하고 이에 비해 남성의 키는 시간의 흐름에 따라 꾸준히 성장함을 확인 할 수 있다.



```
> ggplot(heightweight,aes(x=ageYear,y=heightIn,colour=sex))+geom_point()
```

나이에 따른 키,몸무게에 대한 산점도 그리기

키가 큰 아이들이 몸무게도 많이 나간다는 것을 확인 할 수 있다.



```
> ggplot(heightweight,aes(x=ageYear,y=heightIn,size=weightLb))+geom_point()
```