Recipes 데이터 분석

강사 : 문성민

What is Statistic?

What is Statistic?

- 통계학이란?
  - 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

- 추론통계학(Statistical Inference)
  - 관심의 대상이 되는 집단에서 모집단을 추출하고 이를 분석하여 전체집단의 특성을 과학적으로 추론하는 방법
  - 모수의 추정, 가설검증, 이론추정, 확률론, 통계량의 확률분포 등을 거쳐 모집단의 특성을 예측한다.

What is Statistic?

- 수치형 자료(Quantitative Data)
  - 숫자로 표현되어 있는 자료
  - 계량적 자료, 정량적 자료
  - Ex) 등간척도, 서열척도, 비율척도
- 명목형 자료(Qualititative Data)
  - 숫자로 표시될 수 없는 자료.
  - 범주형 자료, 정성적 자료
  - Ex) 명목척도
- 명목척도(Nominal Scale)
  - 상하관계는 없고 구분만 있는 척도
  - Ex) 성별, 지역, 날씨 등
- 서열척도(Ordinal Scale)
  - 순서에 의해 부여되는 척도
  - Ex) 직위, 학력, 등수 등
- 등간척도(Interval Scale)
  - 간격이 일정하여 가감승제가 가능하지만 절대0점이 존재하지 않는 척도
  - Ex) 시간, 온도 등
- 비율척도(Ratio Scale)
  - 등간척도와 비슷하지만 절대 0점이 존재하는 척도
  - Ex) 성적, 키, 무게, 금액 등

T-test and Scatter plot

What is Independence 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 T 기준 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.
  - Ex) 95%신뢰구간 : 신뢰구간 내에 0 이 포함될 경우 귀무가설을 채택한다.

What is Independence T-test?

- 가설(Hypothesis)

  - H0 (귀무가설) : 두 그룹간 유의한 차이가 없다.
  - H1 (대립가설) : 두 그룹간 유의한 차이가 있다.

- 분산 검정에서 검정 통계량(F값)

$$F = \frac{{S_1}^2}{{S_2}^2} \qquad S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

- T-test에서 검정 통계량(T값)

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}} \qquad S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

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

Independence T-test with R

- 예제(Example)
  - 학교(초,중,고)에 다니는 학생들의 키와 몸무게에 대한 데이터

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

```
> getwd()
[1] "/Users/Seongmin_M/Downloads"
> setwd("/Users/Seongmin_M/Downloads")
> 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     : Factor w/ 2 levels "f","m": 1 1 1 1 1 1 1 1 1 1 ...
 $ 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
```

Independence T-test with R

- 데이터 확인

```
> getwd()
[1] "/Users/Seongmin_M/Downloads"
> setwd("/Users/Seongmin_M/Downloads")
> 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     : Factor w/ 2 levels "f","m": 1 1 1 1 1 1 1 1 1 1 ...
 $ 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
```

- Read.csv함수를 사용하여 데이터를 불러오고 요약한다.

- 데이터 가공

```
> 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
```

- subset함수를 사용하여 전체 데이터를 성별에 따라 분리하고 분리된 내용을 확인한다.

Independence T-test with R

- 가설설정
  - H0 : 성별에 따른 몸무게는 차이가 없다.
  - H1 : 성별에 따른 몸무게는 차이가 있다.
- 성별에 따른 몸무게의 평균 차 분석

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

    F test to compare two variances

data:  group.1$weightLb and group.2$weightLb
F = 0.9513, 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
```

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

- F값을 구하시오.

```
> var(group.1$weightLb)
[1] 346.5555
> var(group.2$weightLb)
[1] 364.2931
> var(group.1$weightLb)/var(group.2$weightLb)
[1] 0.9513095
```

Independence T-test with R

- 가설설정
  - H0 : 성별에 따른 몸무게는 차이가 없다.
  - H1 : 성별에 따른 몸무게는 차이가 있다.
- 성별에 따른 몸무게의 평균 차 분석

```
> 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
```

- 평균 차 비교: t값이 -1.636이고 자유도가 234(n-1), p값이 0.1032이므로 두 그룹간 몸무게는 차이가 나지 않는다. 대립가설(H1: 성별에 따른 몸무게에 차이가 있다.)을 기각 , 귀무가설(H0: 성별에 따른 몸무게에 차이가 없다.)을 채택한다.
- 또한 신뢰구간이 -8.8733370 < α < 0.8220938로 구간에 0을 포함하므로 귀무가설을 채택한다.

- T값과 신뢰구간 값을 구하시오.

```
> (98.87838-102.90400)/(sqrt(346.5555/111+364.2931/125))
[1] -1.638481

> 98.87838-102.90400
[1] -4.02562
```

```
> -4.02562-1.96*(sqrt(346.5555/111+364.2931/125))
[1] -8.841187

> -4.02562+1.96*(sqrt(346.5555/111+364.2931/125))
[1] 0.7899474
```

Independence T-test with R

- 가설설정
  - H0 : 성별에 따른 키에 차이가 없다.
  - H1 : 성별에 따른 키에 차이가 있다.
- 성별에 따른 키의 평균 차 분석

```
> var.test(group.1$heightIn,group.2$heightIn)

    F test to compare two variances

data:  group.1$heightIn and group.2$heightIn
F = 0.6196, num df = 110, denom df = 124, p-value = 0.01076
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4310984 0.8942666
sample estimates:
ratio of variances
           0.61965
```

- 등분산성 검증: F값은 0.6196이고 p값은 0.01076이므로 두 성별간 분산 값에 차이가 난다. 등분산성이 성립하지 않는다.

- F값을 구하시오.

```
> var(group.1$heightIn)
[1] 11.27813
> var(group.2$heightIn)
[1] 18.20081
> var(group.1$heightIn)/var(group.2$heightIn)
[1] 0.61965
```

Independence T-test with R

- 가설설정
  - H0 : 성별에 따른 키에 차이가 없다.
  - H1 : 성별에 따른 키에 차이가 있다.
- 성별에 따른 키의 평균 차 분석

```
> t.test(group.1$heightIn,group.2$heightIn,var.equal=FALSE)

    Welch Two Sample t-test

data:  group.1$heightIn and group.2$heightIn
t = -3.085, df = 230.766, p-value = 0.002284
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.5135123 -0.5542354
sample estimates:
mean of x mean of y
 60.52613  62.06000
```

- 평균 차 비교: t값이 -3.085이고 자유도가 230.766(분산이 다르므로 welch의 자유도를 사용, 자유도 값은 실수가 된다.), p값이 0.002284이므로 두 그룹간 키는 차이가 난다. 대립가설(H1: 성별에 따른 키에 차이가 있다.)을 채택 , 귀무가설(H0: 성별에 따른 키에 차이가 없다.)을 기각한다. 하지만 두 데이터의 분산의 크기가 다르므로 검증결과를 뒷받침하기 어렵다.
- 또한 신뢰구간이 -2.5135123 < α < -0.5542354로 구간에 0을 포함하지 않으므로 대립가설을 채택한다.
- T값과 신뢰구간 값을 구하시오.

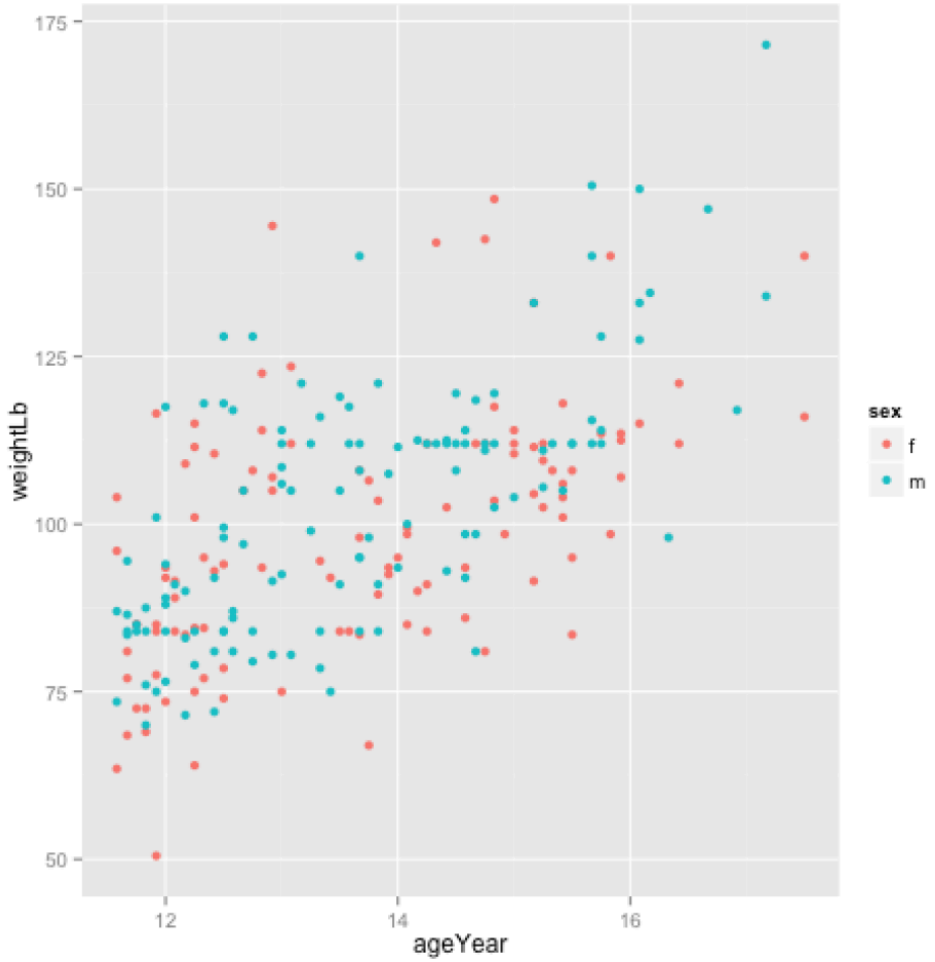
```
> (60.52613-62.06000)/(sqrt(11.27813/111+18.20081/125))
[1] -3.084995
> 60.52613-62.06000
[1] -1.53387
```

```
> -1.53387+1.96*(sqrt(11.27813/111+18.20081/125))
[1] -0.5593513
> -1.53387-1.96*(sqrt(11.27813/111+18.20081/125))
[1] -2.508389
```

Scatter plot with R

- 나이에 따른 몸무게에 대한 산점도 그리기

```
> library(ggplot2)
>
> ggplot(heightweight,aes(x=ageYear,y=weightLb,colour=sex))+geom_point()
```

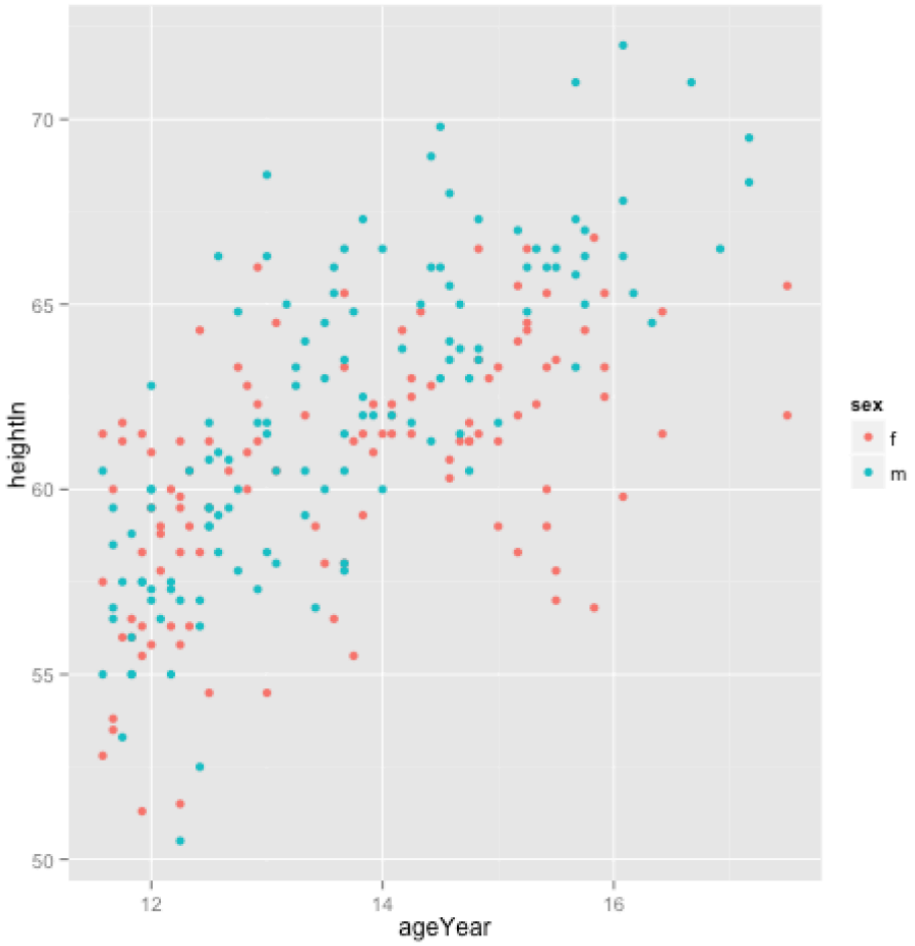


- 몸무게가 많이 나가는 그룹에 남성이 많고 나이와 몸무게는 비례하는 것을 확인 할 수 있다.

Scatter plot with R

- 나이에 따른 키에 대한 산점도 그리기

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

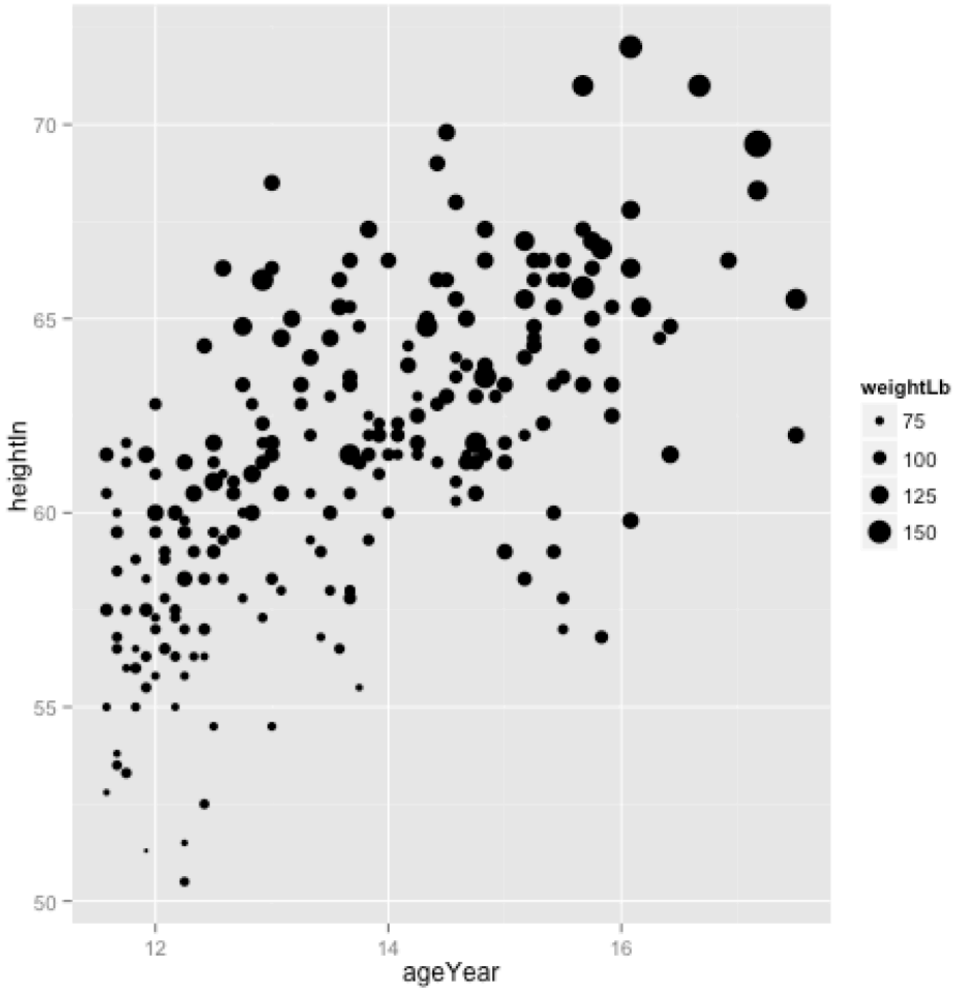


- 키가 큰 그룹에 남성이 많다. 여성의 경우 14세를 기준으로 키의 성장이 멈춘 듯하고 이에 비해 남성의 키는 시간의 흐름에 따라 꾸준히 성장함을 확인 할 수 있다.

Scatter plot with R

- 나이에 따른 키,몸무게에 대한 산점도 그리기

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

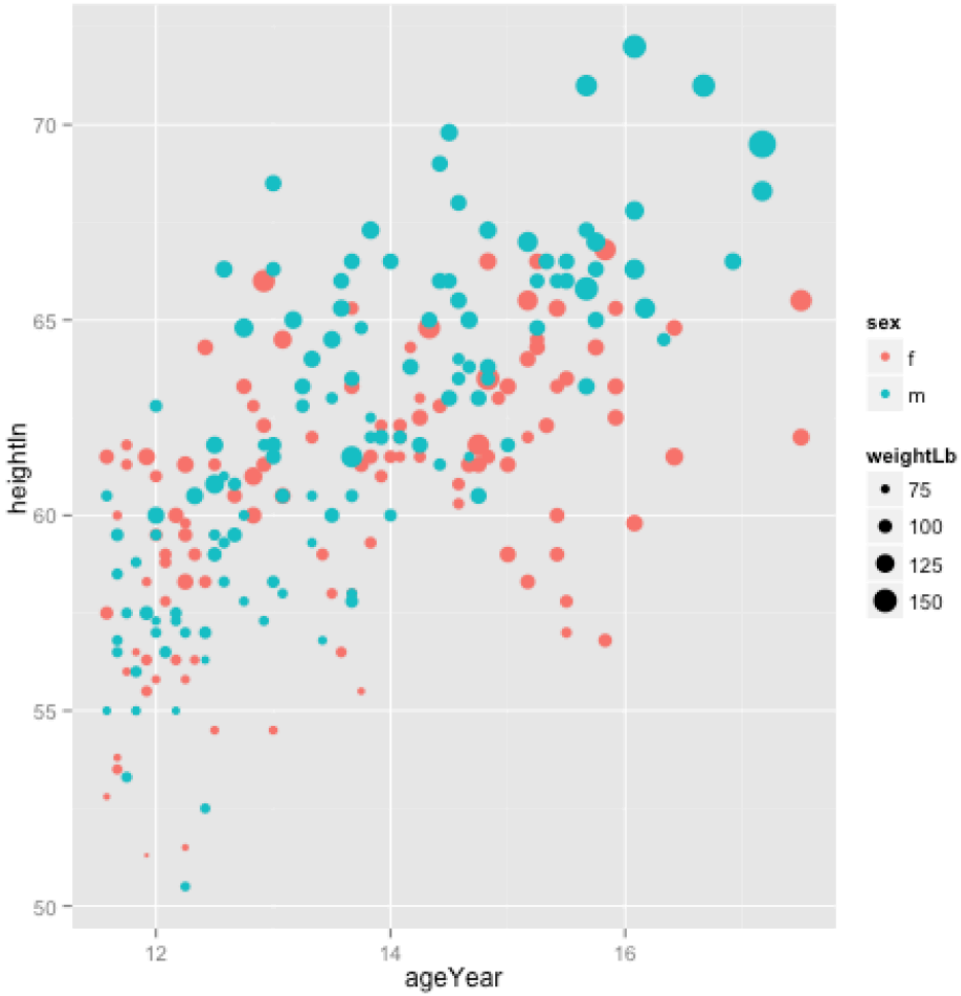


- 키가 큰 아이들이 몸무게도 많이 나간다는 것을 확인 할 수 있다.

Scatter plot with R

- 나이와 성별에 따른 키,몸무게에 대한 산점도 그리기

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

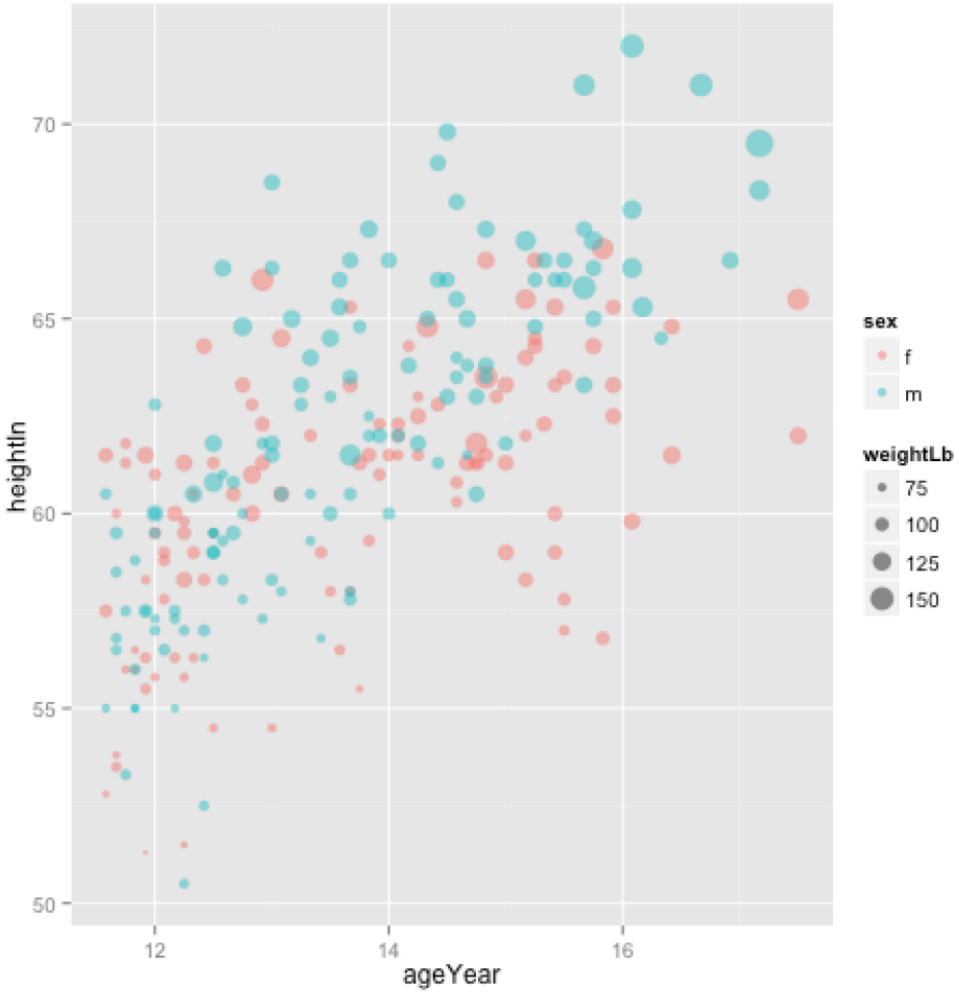


- 나이와 성별에 따라 여성과 남성의 몸무게,키의 차이가 확연히 보인다.

Scatter plot with R

- 나이와 성별에 따른 키,몸무게에 대한 산점도의 투명도를 달리하여 그리기

```
> ggplot(heightweight,aes(x=ageYear,y=heightIn,size=weightLb,colour=sex))+geom_point(alpha=.5)
```



- 겹쳐져서 보이지 않던 데이터까지 파악할 수 있다.

T-test and Scatter plot

- 과제

- (1) UCBAdmissions데이터를 사용하여 입학 학생수와 입학이 거부된 학생들 간의 평균 비교를 시행하고 결과를 해석하시오.

- (2) 산점도 시각화 분석을 사용하여 USArrests데이터를 분석하시오.

T-test and Scatter plot

- 데이터 설명

- (1) UCBAdmissions

Student Admissions at UC Berkeley

Description

Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

Usage

```
UCBAdmissions
```

Format

A 3-dimensional array resulting from cross-tabulating 4526 observations on 3 variables. The variables and their levels are as follows:

No	Name	Levels
1	Admit	Admitted, Rejected
2	Gender	Male, Female
3	Dept	A, B, C, D, E, F

설명 : 1973년 버클리 대학원에 지원한 학생에 대한 데이터.

변수:

Admit : 입학 , 거부
Gender : 남 , 여
Dept : 6개의 학과

- (2) USArrests

Violent Crime Rates by US State

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Usage

```
USArrests
```

Format

A data frame with 50 observations on 4 variables.

[,1]	Murder	numeric Murder arrests (per 100,000)
[,2]	Assault	numeric Assault arrests (per 100,000)
[,3]	UrbanPop	numeric Percent urban population
[,4]	Rape	numeric Rape arrests (per 100,000)

설명 : 1973년 미국 50개 주에서 일어난 사건 체포 데이터.

변수:

Murder : 살인사건 체포
Assault : 폭행 사건 체포
UrbanPop : 도시 인구 비율
Rape : 강간 사건 체포