R프로그래밍및실습 연습_5

Jinseog Kim
Dongguk University
jskim1986@gmail.com

2018-06-07

1 2014년 건강검진자료-1백만명 대상

1. 건강검진_2014.csv
2. 건강검진_2014_일부추출.csv : 전체에서 1만명 추출

2 건강검진_2014_일부추출.csv자료를 읽어 시도별 검진대상자수를 구하라.

```
health_2014_s <- read.csv("건강검진_2014_일부추출.csv", fileEncoding = "CP949")
health_2014_s$시도 <- factor(health_2014_s$시도,
                          labels=c("서울","부산","대구", "인천", "광주","대전", "울산","세종",
                                   "경기", "강원", "충북", "충남", "전북", "전남", "경북", "경남", "제주"))
head(health_2014_s)
```

```
        X 성별 연령대_5세단위 시도 신장 체중 허리둘레 수축기혈압 이완기혈압 흡연상태 음주여부
1 462508    2            12 경기  155   75      104        140        90        1        0
2 732734    2             5 충북  155   45       68        120        80        1        1
3 274368    1             7 세종  175   90       94        120        70        2        0
4 861295    1             8 광주  180   65       77         90        60        3        0
5 983308    2            10 부산  165   70       79         92        62        1        1
6 128569    1            11 부산  180   80       90        125        88        2        1
```

```
a <- aggregate(성별~시도, health_2014_s, length)
names(a)[2] <- "검진대상자수"
a
```

```
   시도 검진대상자수
1  서울        331
2  부산       2243
3  대구        714
4  인천        584
5  광주        289
6  대전        503
```

```
7  울산          312
8  세종          724
9  경기         1799
10 강원           18
11 충북          250
12 충남          599
13 전북          378
14 전남          377
15 경북           99
16 경남          414
17 제주          366
```

```
#aggregate(health_2014$성별코드, list(health_2014$시도코드), length)
```

3 "성별" 변수를 범주형 변수로, 각 수준(level)값이 1은 "남", 2는 "여"로 수준이름(label)을 바꾸시오.

```
health_2014_s$성별 <- factor(health_2014_s$성별, labels=c("남", "여"))
head(health_2014_s)
```

```
        X 성별 연령대_5세단위 시도 신장 체중 허리둘레 수축기혈압 이완기혈압 흡연상태 음주여부
1 462508   여            12 경기  155   75      104        140         90        1        0
2 732734   여             5 충북  155   45       68        120         80        1        1
3 274368   남             7 세종  175   90       94        120         70        2        0
4 861295   남             8 광주  180   65       77         90         60        3        0
5 983308   여            10 부산  165   70       79         92         62        1        1
6 128569   남            11 부산  180   80       90        125         88        2        1
```

4 연령별 변수인 "연령대_5세단위"를 10세단위로 바꾸시오.

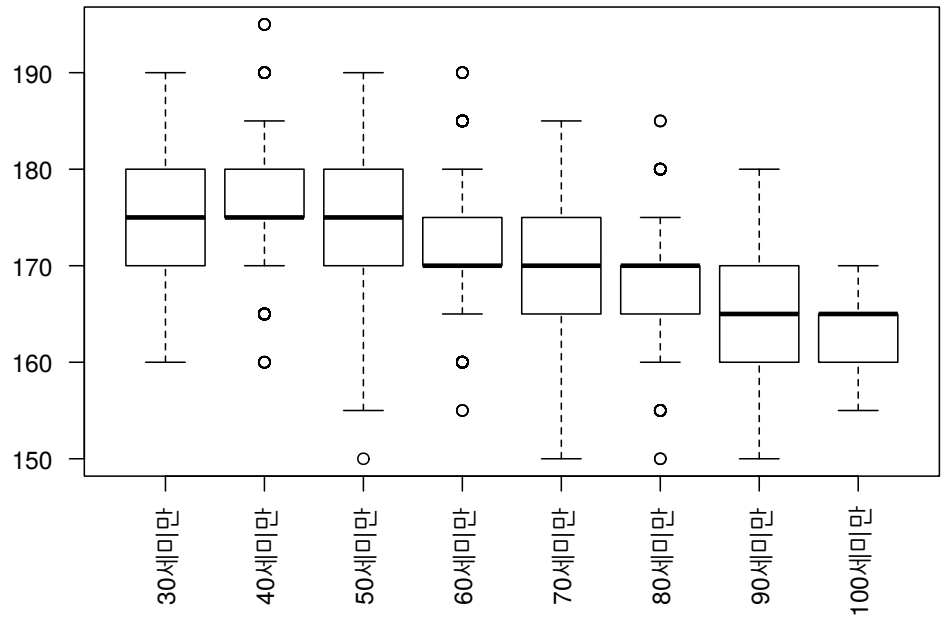
```
health_2014_s$연령 <- cut(health_2014_s$연령대_5세단위, c(2:10)*2, right=FALSE,
                        labels=paste0(c(3:10)*10, "세미만") )
#health_2014_s$연령대_5세단위 <- NULL
health_2014_s$연령2 <- (health_2014_s$연령대_5세단위+ 1)%/%2
head(health_2014_s[, -c(1:4)])
```

```
  신장 체중 허리둘레 수축기혈압 이완기혈압 흡연상태 음주여부     연령 연령2
1  155   75      104        140         90        1        0 70세미만     6
2  155   45       68        120         80        1        1 30세미만     3
3  175   90       94        120         70        2        0 40세미만     4
4  180   65       77         90         60        3        0 50세미만     4
```

```
  5  165   70       79         92         62        1        1 60세미만      5
  6  180   80       90        125         88        2        1 60세미만      6
```

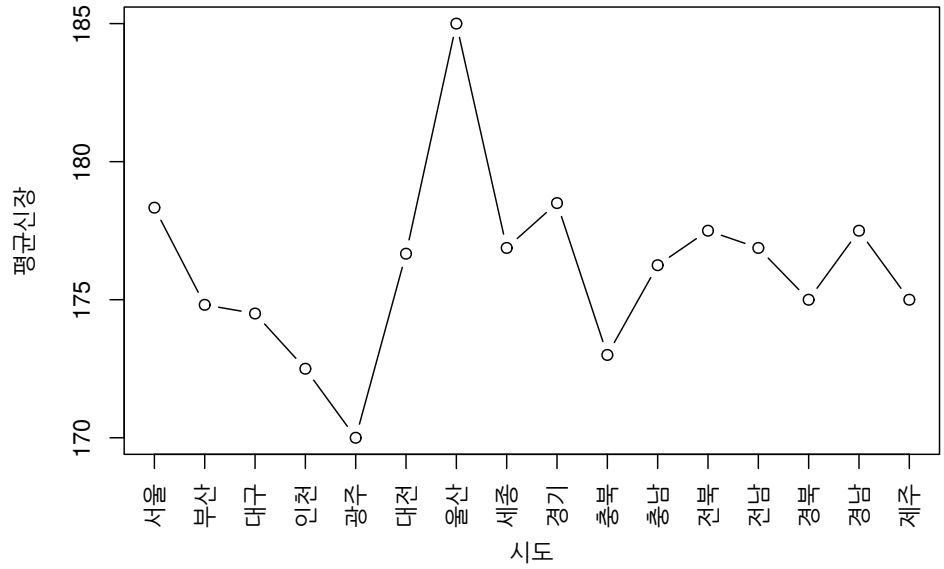
5 남자/여자의 “연령”에 따른 “신장”의 분포를 상자그림으로 나타내시오.

```
boxplot(신장~연령, subset(health_2014_s, 성별=="남"), las=2)
```



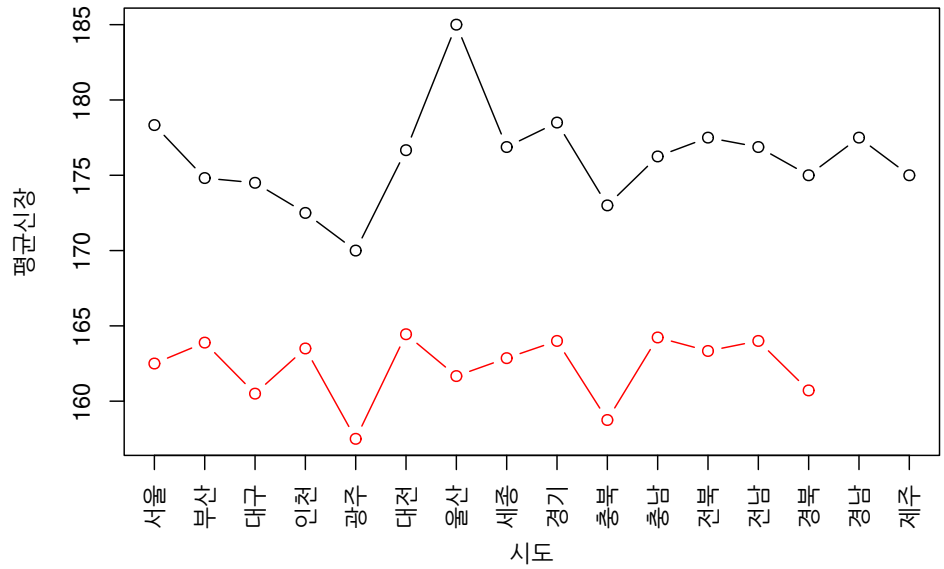
6 시도별로 남자 “30대미만”의 평균 “신장”을 “선그래프”로 나타내시오.

```
sx <- subset(health_2014_s, 성별=="남" & 연령=="30세미만")
o1 <- aggregate(신장~시도, sx, mean)
plot(o1$신장, type="b", ylab="평균신장", xlab="시도", xaxt="n")
axis(side=1, at=seq_along(o1$시도), labels = o1$시도, las=2)
```



7 위의 그래프에 시도별로 여자 "30대미만"의 평균 "신장"을 겹쳐 나타내시오.

```
sx <- subset(health_2014_s, 성별=="남" & 연령=="30세미만")
o1 <- aggregate(신장~시도, sx, mean)
sy <- subset(health_2014_s, 성별=="여" & 연령=="30세미만")
o2 <- aggregate(신장~시도, sy, mean)
plot(o1$신장, type="b", ylab="평균신장", xlab="시도", xaxt="n", ylim=c(min(o2$신장), max(o1$신장)))
lines(o2$신장, type="b", col="red")
axis(side=1, at=seq_along(o1$시도), labels = o1$시도, las=2)
```



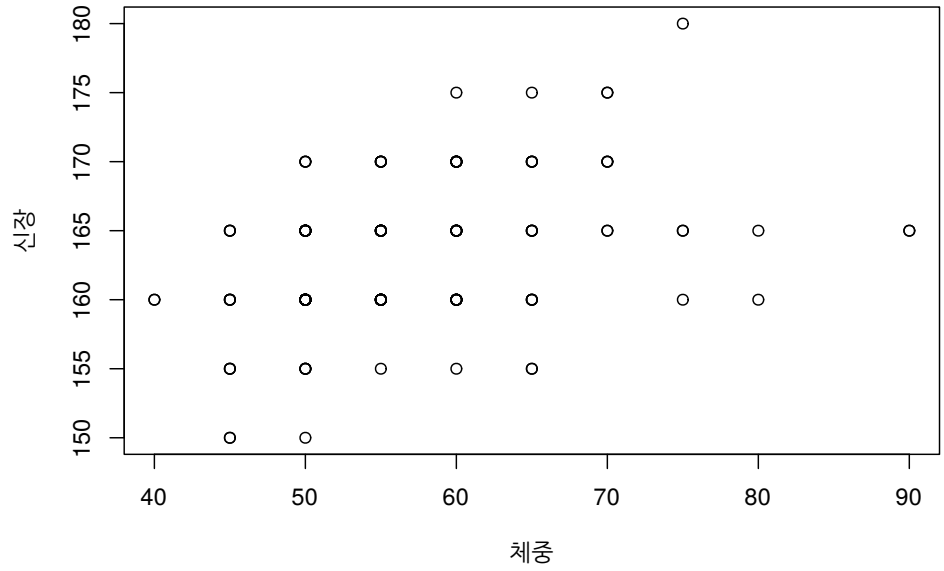
8 “성별”, “연령”에 따라 “신장”의 평균 및 표준편차를 구하시오.

```
height_m <- aggregate(신장~성별+연령, health_2014_s, mean)
names(height_m)[3] <- "평균"
height_s <- aggregate(신장~성별+연령, health_2014_s, sd)
names(height_s)[3] <- "표준편차"
out <- merge(height_m, height_s, by=c("성별", "연령"))
out[order(out$연령),]
```

```
   성별      연령     평균 표준편차
2    남  30세미만 175.8879 6.135771
10   여  30세미만 163.1206 5.262367
3    남  40세미만 176.3641 6.230781
11   여  40세미만 163.6472 5.533118
4    남  50세미만 174.7801 5.811544
12   여  50세미만 161.5167 5.470596
5    남  60세미만 172.0868 5.884567
13   여  60세미만 159.3197 4.942717
6    남  70세미만 169.6763 5.824334
14   여  70세미만 156.9471 5.467290
7    남  80세미만 167.9198 5.841449
15   여  80세미만 154.5601 5.400712
8    남  90세미만 165.4861 6.154180
16   여  90세미만 150.5412 5.744188
1    남 100세미만 163.1818 5.134553
9    여 100세미만 146.9048 5.804350
```

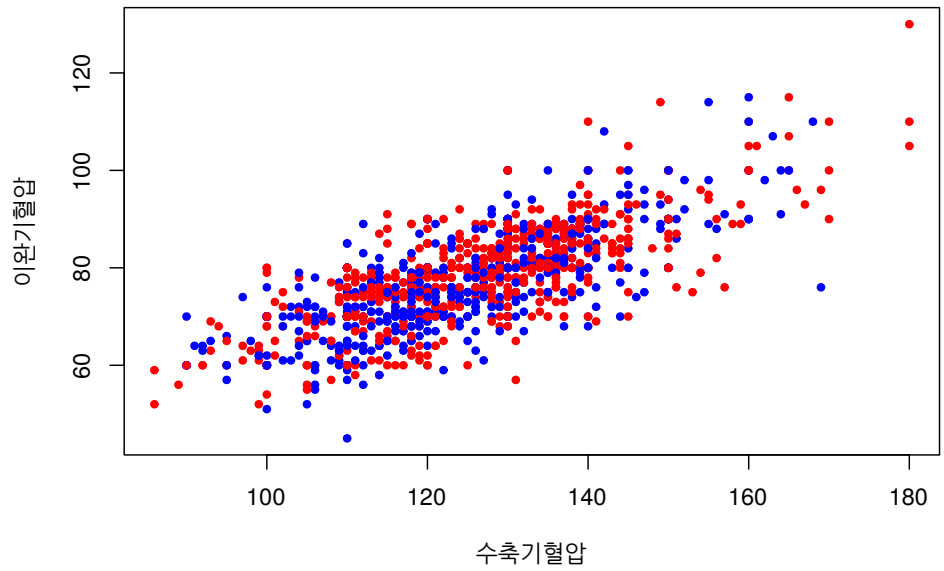
9 30세미만 여성의 신장과 체중의 관계를 산점도로 나타내시오.

```
plot(신장~체중, subset(health_2014_s, 성별=="여" & 연령=="30세미만"))
```



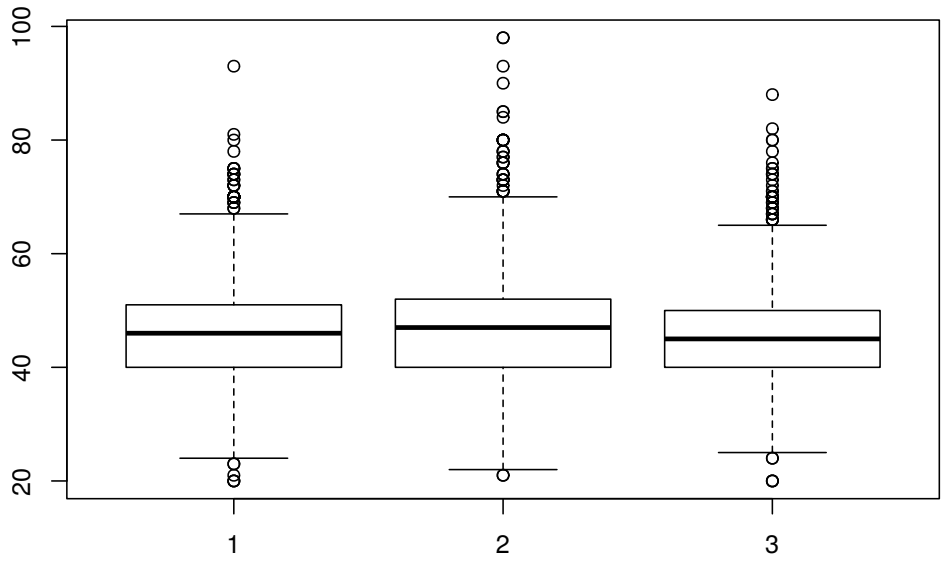
10 흡연상태가 1인 남성의 “수축기혈압”과 “이완기혈압”을 산점도 나타내시오. 단 음주여부에 따라 점의 색을 다르게 표현하시오.

```
dd <- subset(health_2014_s, 성별=="남" & 흡연상태==1)
col <- c("red", "blue")[dd$음주여부+1]
plot(이완기혈압~수축기혈압, dd, col=col, pch=20)
```



11 흡연상태에 따라 남성의 혈압차이(“수축기혈압”과 “이완기혈압”의 차)를 그래프로 나타내시오.

```
boxplot(abs(이완기혈압-수축기혈압)~흡연상태, subset(health_2014_s, 성별=="남"))
```



12 흡연상태가 “1”인 남성의 “수축기혈압”과 “이완기혈압”의 상관계수를 구하시오.

```
dd <- subset(health_2014_s, 성별=="남" & 흡연상태==1)
cor(dd$이완기혈압, dd$수축기혈압)
```

```
[1] 0.7473961
```

```
cor.test(dd$이완기혈압, dd$수축기혈압)
```

```
 Pearson's product-moment correlation

data:  dd$이완기혈압 and dd$수축기혈압
t = 46.287, df = 1693, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7256046 0.7676899
sample estimates:
      cor
0.7473961
```

13 서울시에 거주하며 남성인 검진대상자에 대하여 흡연상태가 “1”의 혈압의 차이(“수축기혈압” - “이완기혈압”)가 흡연상태가 “3”인 검진대상자의 혈압의 차이보다 크다는 가설에 대하여 가설검정을 하시오.

```
dd <- subset(health_2014_s, 성별=="남" & 시도=="서울")
bp_diff <- dd$수축기혈압 - dd$이완기혈압
bp_diff_1 <- bp_diff[dd$흡연상태==1]
bp_diff_3 <- bp_diff[dd$흡연상태==3]
t.test(bp_diff_1, bp_diff_3, alternative = "greater")
```

```
    Welch Two Sample t-test

   data:  bp_diff_1 and bp_diff_3
   t = 1.4393, df = 140.47, p-value = 0.07614
   alternative hypothesis: true difference in means is greater than 0
   95 percent confidence interval:
    -0.3675175        Inf
   sample estimates:
   mean of x mean of y
    48.05634  45.61250
```