데이터 통계 분석 기초

환경생태데이터사이언스 실습 December 3, 2019

지난 주 과제 풀이

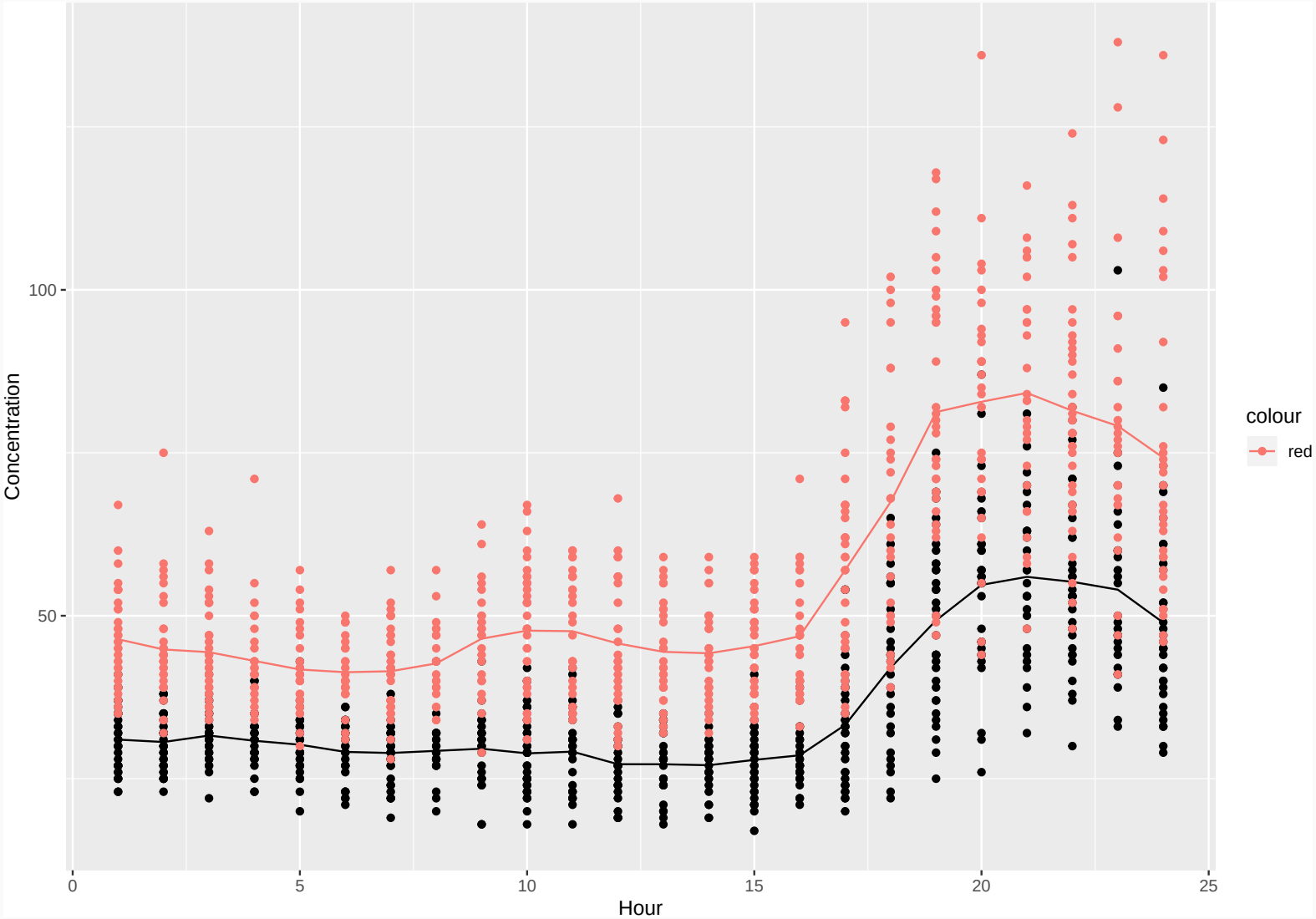
```
library(readxl)
library(dplyr)
library(reshape2)
PM25 <- read_xls("./Data/PM25_20190930.xls")
PM100 <- read_xls("./Data/PM100_20190930.xls")
PM25_melt <- melt(PM25, id=c("□□□", "□□□□")) %>%
    mutate(variable = as.integer(gsub("□", "", variable)) ) %>%
    na.omit(.)
PM100_melt <- melt(PM100, id=c("□□□", "□□□□")) %>%
    mutate(variable = as.integer(gsub("□", "", variable)) ) %>%
    na.omit(.)
```

지난 주 과제 풀이

```
PM25_ggplot <- ggplot(data = PM25_melt,
                      aes(x=variable, y=as.numeric(value))) +
    geom_point() +
    geom_line(data = PM25_melt %>% group_by(variable) %>%
              summarize(avg = mean(as.numeric(value), na.rm=T)),
              aes(x=variable, y=avg) ) +
    geom_point(data = PM100_melt,
               aes(x=variable, y=as.numeric(value), col="red") ) +
    geom_line(data = PM100_melt %>% group_by(variable) %>%
              summarize(avg = mean(as.numeric(value), na.rm=T)),
              aes(x=variable, y=avg, col="red") ) +
    xlab("Hour") + ylab("Concentration")
```

지난 주 과제 풀이

PM25_ggplot



오늘의 학습 목표

1. 통계란 무엇이고 왜 필요할까?
2. 분포 함수
3. 기초 기술 통계
4. 자주 쓰이는 추론 통계 방법

통계란?

통계란?

1

[1]출처: 나무위키, 계란

통계란?

- 통계 [2]
  - 자료를 수집, 분석.
  - 자료를 해석하여 관심이 되는 현상을 과학적으로 설명
  - 사람들이 이해하기 쉽게 효과적인 자료 표현.

[2]출처: Wikipedia, Statistics

통계의 종류

기본 적인 통계 방법 분류 [3]

- 기술 통계 (Descriptive statistics)
  - 수집한 데이터를 요약 묘사 설명하는 통계 기법
  - 집중화 경향 (평균, 중앙값, 최빈값 등)
  - 분산도 (표준 편차, 사분위 값)
- 추론 통계 (Inferential statistics)
  - 수집한 데이터를 바탕으로 추론 및 예측을 하는 통계 기법.
  - 단순히 자료를 보여주는데 그치지 않고 자료의 비교, 분석, 해석을 통해 의미있는 결과를 끌어내는 방법.

[3]출처: 홍박사의 데이터 노트, 기술통계와 추리통계란 무엇인가?

통계의 목적

통계의 목적 [4]

- 현재 상황 설명 및 묘사 $\rightarrow$ 기술 통계
- 상황의 특성 설명 및 예측 $\rightarrow$ 추론 통계

[4]출처: 홍박사의 데이터 노트, 기술통계와 추리통계란 무엇인가?

통계의 목적

자주 쓰이는 통계의 목적

- 집단 간 비교를 통한,
  - 경험적 (선험적) 지식의 검증.
  - 예) 현재 고등학생의 키는 한 세대 (30년) 전보다 크다.
  - 예) 담배를 많이 피우면 건강에 좋지 않다 (암 발병률이 높다).
- 처리 전 후의 차이 비교를 통한,
  - 특정 현상에 중대한 영향을 주는 요소 확인.
  - 예) 사람의 키에 가장 영향을 많이 주는 요소는 무엇일까?
  - (영양분, 유전자, 운동 강도 등)
  - 특정 현상이 다른 요소에 반응하는 정도 확인.
- 요소 간 관계 확인 및 예측
  - 예) 도시 숲의 비율과 미세먼지 농도는 어떤 관계가 있는가?
  - 예) 도시에 나무를 현재보다 2배 식재 했을 때 미세먼지 저감 효과는?
- 그 밖에 다양한 목적에 쓰이고 있음.

분포 함수와 난수 생성

분포 함수

통계 분석을 위해서는 모집단과 표본 집단의 분포를 알아두는 것이 유용하다.

1. 정규 분포 (Normal distribution)
    - 오차나 자연적인 변이와 관계된 분포 (사람들의 키, 실험 오차 등)
2. 포아송 분포 (Poisson distribution)
    - 특정 시간동안 발생하는 이벤트의 횟수 (하루에 치는 천둥의 횟수, 강우 일수)
3. 이항 분포 (Binomial distribution)
    - n 번의 베르누이 시행을 할 때 특정 이벤트가 나오는 횟수
4. 연속 균등 분포 (Uniform distribution)
    - 연속적인 변수
5. Student's t 분포 (Student's t distribution)
    - 정규 분포에서 추출된 집단이 가지는 분포

난수를 이용한 분포 함수 작성

[5]

[5]출처: 서민구 (2014). R을 이용한 데이터 처리 & 분석 실무. 길벗

난수를 이용한 분포 함수 작성

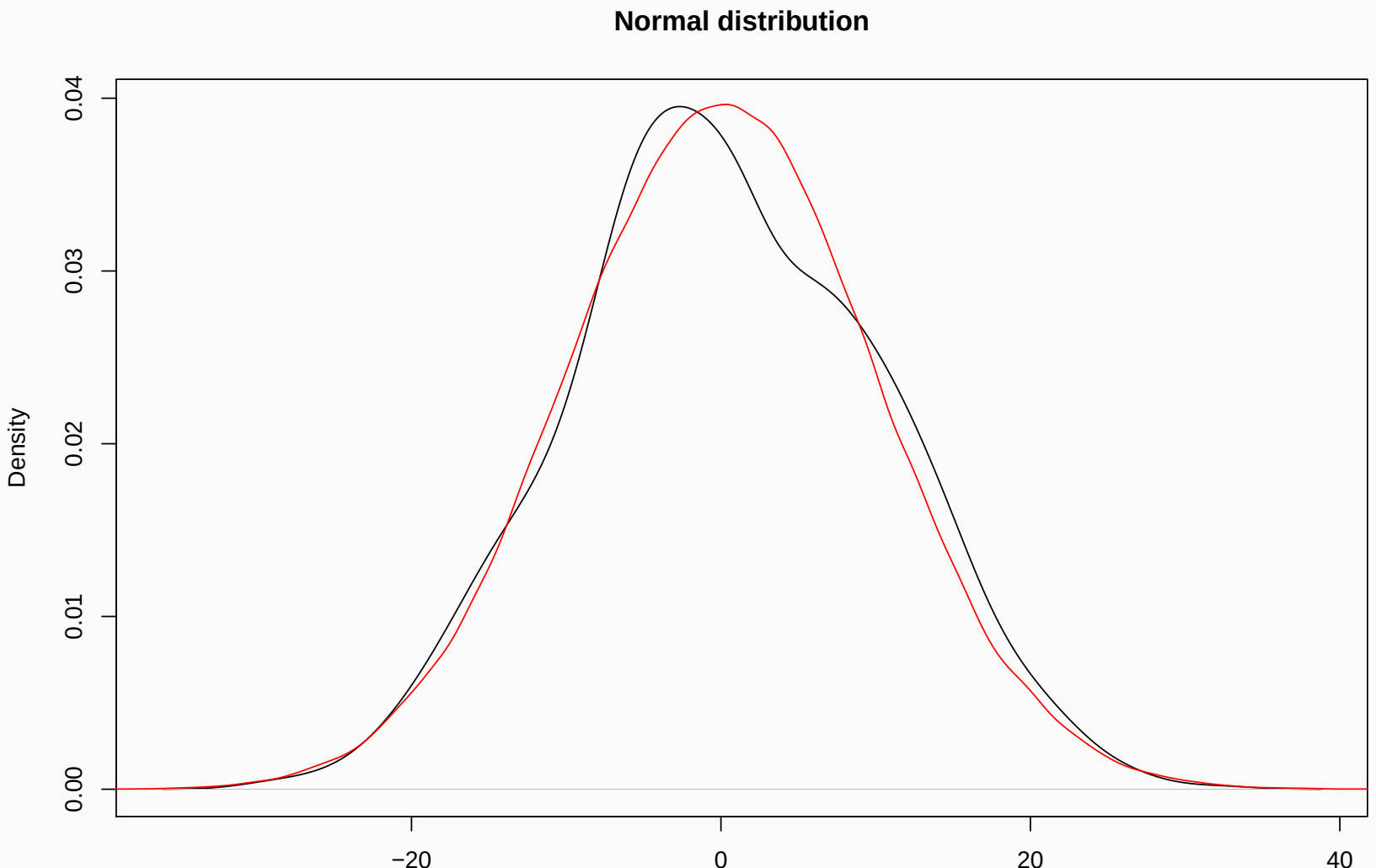
예) 정규 분포 (Normal distribution)

```
set.seed(2017464)
# Select 1000 random numbers with in the normal distribution
# whose mean is 0 and standard distribution is 10.
head(rnorm(1000, 0, 10))
```

```
## [1] -9.169712 -7.683962  9.632437 -7.857349  2.969502
## [6] -2.834711
```

난수를 이용한 분포 함수 작성

```
# Create plot with the extracted points
plot(density(rnorm(1000, 0, 10)), main="Normal distribution")
lines(density(rnorm(100000, 0, 10)), col="red", cex=1.5)
```



기술 통계

기초적인 기술 통계량 계산

- 표본 평균 (`mean(x)`)

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- 표본 분산 (`var(x)`)

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- 표본 표준 편차 (`sd(x)`)

$$\sqrt{S^2}$$

- 최솟값, 제1사분위수, 중앙값, 제3사분위수, 최댓값 표시 (`fivenum(x)`)
- `fivenum`에 평균을 덧붙인 요약 통계 값 (`summary(x)`)

자주 쓰이는 추론 통계 방법

가설 검정

가설 검정의 예

1. 30년 평년 기온 분포 자료를 확인했을 때 2월 평균 기온은 섭씨 5°이다. 올해 2월 기온이 8도라고 할 때 올 2월 기온은 이상 온난화라고 할 수 있을까?
2. 팀팀 클래스를 듣는 산림환경시스템학과 학생들의 신발 사이즈와 정보보안암호수학과 학생들의 신발 사이즈는 다르다고 할 수 있을까?
3. 풍향에 따라 미세먼지 농도의 차이가 있는가?

Student's t-test (t 검정)

t 검정 [6]

- 모집단이 정규 분포인 집단에서 추출한 표본이 알려진 모평균과 다른 값을 갖는지 확인.
- 모집단이 정규 분포인 집단에서 추출한 두 표본 집합이 같다고 할 수 있는지 확인.
- 귀무가설과 대립가설을 설정한 후 확인.
- t-검정에 필요한 가정
  - 검정 대상이 되는 표본 집단은 정규 분포를 갖어야 함 (정규성 검정)
  - → `shapiro.test()`를 통해 검정.

[6]출처: 이태권 (2018). R을 활용한 기초 환경자료 분석 및 시각화

하나 혹은 두 집단 비교

7

[7]출처: 이태권 (2018). R을 활용한 기초 환경자료 분석 및 시각화

하나 혹은 두 집단 비교

```r
library(dplyr)
# T-test example
KMA_Data <- read.csv("./Data/KMA_20190916_0922.csv",
                     fileEncoding="euc-kr")
str(KMA_Data)
```

```
## 'data.frame':    13536 obs. of  6 variables:
##  $ ID           : int  90 90 90 90 90 90 90 90 90 90 ...
##  $ Time         : Factor w/ 144 levels "2019-09-16 1:00",..: 1 12
##  $ Temperature  : num  16.5 16.2 16.2 16.8 16.9 17.3 17.6 18.6 20.
##  $ Precipitation: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WindVelocity : num  1.8 0.9 1.3 1.6 1.1 1.2 1.2 1.3 1.1 1.9 ...
##  $ WindDirection: int  290 270 230 270 270 270 270 290 340 50 ...
```

```r
KMA_Data <- KMA_Data %>% mutate(FID = as.factor(ID))
```

하나 혹은 두 집단 비교

```r
# Null hypothesis: the average temperature
# between ID-90 and ID-295 are same.
# Alternative hypothesis: The two values are different.
ID90Temp <- KMA_Data %>% filter(FID==90) %>% select(Temperature)
ID295Temp <- KMA_Data %>% filter(FID==295) %>% select(Temperature)
# Normality test (if the p-value is larger than 0.05),
# then we can use t-test
# shapiro.test(ID90Temp$Temperature)  # Run without comment
# shapiro.test(ID295Temp$Temperature) # Run without comment
# Equality variance test (if P is less than 0.05, then not equal)
# var.test(ID90Temp$Temperature, ID295Temp$Temperature) # Run without
```

하나 혹은 두 집단 비교

```r
# t-test
t.test(ID90Temp$Temperature, ID295Temp$Temperature)
```

```
##
##  Welch Two Sample t-test
##
## data:  ID90Temp$Temperature and ID295Temp$Temperature
## t = -5.1982, df = 267.8, p-value = 3.997e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.440595 -1.099683
## sample estimates:
## mean of x mean of y
##  19.65347  21.42361
```

```r
# We can conclude that
# the average temperature of two places are different.
```

하나 혹은 두 집단 비교

```
# Find the exact location of ID 90 and 295, using station info file
# in 10 minutes
```

셋 이상의 집단 비교

셋 이상의 모집단 간 평균의 차이를 검정할 때 사용하는 분산 기반 분석방법

- ANOVA (Analysis of variance): 분산분석
- 필요한 가정

1. 독립성 (`dwtest()`)
2. 정규성 (`shapiro.test()`)
3. 등분산성 (`bartlett.test()`)

셋 이상의 집단 비교

```r
# Filter KMA_Data that having ID 90, 295, and 100
KMA_Data_aov <- KMA_Data %>% filter(ID %in% c(90, 100, 295))

# install.packages("lmtest")
# library(lmtest)
 # if p is less than 0.05, they are not indep.
# dwtest(Temperature ~ ID, data = KMA_Data_aov)
# aov_result <- aov(Temperature ~ ID, data = KMA_Data_aov)
 # p value is less than 0.05, they are not normally distributed
# shapiro.test(resid(aov_result))
# P values is less than 0.05, variances are diff.
# bartlett.test(Temperature ~ ID, data = KMA_Data_aov)
# kruskal_result <- kruskal.test(Temperature ~ ID,
#                                data = KMA_Data_aov)
# At least one group is different from others.
```

셋 이상의 집단 비교

다중 비교 (TukeyHSD, mctp)

- aov의 경우 TukeyHSD
- kruskal.test의 경우 mctp

```
# install.packages("nparcomp")
# library(nparcomp)
# summary(mctp(Temperature ~ ID, data = KMA_Data_aov))
# detach("package:nparcomp", unload=TRUE)
```

시각화

```
Result_ggplot <- ggplot(data=KMA_Data_aov,
                        aes(x=as.factor(ID),
                            y = Temperature,
                            fill = as.factor(ID)))
Final_plots <-
    Result_ggplot +
    geom_boxplot() +
    geom_text(x = 1, y = 30, label="*", size=15) +
    geom_text(x = 2, y = 30, label="*", size=15) +
    geom_text(x = 3, y = 30, label="*", size=15) +
    geom_hline(yintercept = mean(KMA_Data_aov$Temperature, na.rm=T),
               col = "red", lty="dashed")
```

시각화

Final_plots