Ch11 기술통계분석

1.1 빈도분석

빈도분석(Frequency Analysis)
명목척도 또는 서열척도 같은 범주형 데이터를 대상으로 비율을 측정하는데 주로 이용
명목척도: 명목상 의미 없는 수치로 표현. 예) 거주지역, 성별
서열척도: 계급 순위를 수치로 표현. 예) 직급, 학력 수준

빈도수, 비율 등으로 나타냄

1.2 기술통계분석

등간척도나 비율척도와 같은 연속적 데이터를 분석할 때 이용
등간척도: 속성의 간격이 일정한 값을 갖는 변수. 예) 만족도 조사의 보기
비율척도: 등간척도의 특성에 절대 원점이 존재하는 척도. 0을 기준으로 한 수치. 사칙연산이 가능.
예) 성적, 나이, 수량, 길이, 금액

2. 척도별 기술통계량 구하기

실습 (전체 데이터 셋의 특성 보기)

1단계: 데이터 셋 가져오기

```
setwd("C:/Rwork/ ")
data <- read.csv("descriptive.csv", header = TRUE)
head(data)
```

descriptive.csv데이터 셋 설명

2단계: 데이터 셋의 데이터 특성 보기

```
dim(data)
length(data)
length(data$survey)
str(data)
```

3단계: 데이터 특성(최소값, 최대값, 평균, 분위수, 결측치 등) 제공

summary(data)

2.1 명목척도 기술통계량

명목상 의미 없는 수치로 표현된 거주지역이나 성별과 같은 명목척도 변수를 대상
구성비율은 표본의 통계량으로 의미

실습 (성별 변수의 기수롱계량과 빈도수)
summary()함수, table()함수

```
length(data$gender)
summary(data$gender)
table(data$gender)
```

실습 (이상치(outlier) 제거)

```
data <- subset(data, gender == 1 | gender == 2)
x <- table(data$gender)
x
barplot(x)
```

실습 (구성 비율 계산)

```
prop.table(x)
y <- prop.table(x)
round(y * 100, 2)
```

prop.table(x)함수

형식: prop.table(x, margin=NULL)
Where margin=1(행), 2(열) 계산기준

https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/prop.table

Ex.

```
m <- matrix(1:4, 2)
m
```

prop.table(m, 1)
prop.table(m, 2)
결과 비교!

round()함수
형식: round(x, digits=0)
Where digits: 표시할 소수점 이하 자리

https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/Round

2.2 서열척도 기술통계량

계급 순위를 수치로 표현한 계급이나 학력 수준 등과 같은 서열척도 변수를 대상
table()함수로 구해진 빈도수를 통해서 표본의 통계량 산출

실습 (학력수준 변수를 대상으로 구성 비율 산출)

length(data$level)
summary(data$level)

table(data$level)

실습 (학력수준 변수의 빈도수 시각화)
막대차트, 파이차트 주로 사용

x1 <- table(data$level)
barplot(x1)

2.3 등간척도 기술통계량
등간척도: 속성의 간격이 일정한 값을 갖는 변수
예) 설문의 응답, 1점~5점 척도(매우 불만족, 불만족, 보통, 만족, 매우만족)

실습 (만족도 변수를 대상으로 요약통계량)

1단계: 등간척도 변수 추출

```
survey <- data$survey
survey
```

2단계: 등간척도 요약통계량

```
summary(survey)
```

실습 (등간척도 빈도분석)

```
x1 <- table(survey)
x1
```

실습 (등간척도 시각화)

```
hist(survey)
```

```
pie(x1)
```

2.4 비율척도 기술통계량

비율척도: 등간척도의 특성에 절대 원점(0)이 존재하는 척도
사칙연산 가능
빈도분석과 기술통계량 등 가장 많은 표본의 통계량을 얻을 수 있는 척도

실습 (생활비 변수 대상 요약통계량)

```
length(data$cost)
summary(data$cost)
```

실습 (데이터 정제(결측치 제거))

```
plot(data$cost)
data <- subset(data, data$cost >= 2 & data$cost <= 10)
x <- data$cost
mean(x)
```

plot()함수로 결측치 발견
subset()함수로 결측치 제거된 subset 산출

(1) 대표값 구하기
자료 전체를 대표하는 값으로 분포의 중심위치를 나타내는 평균, 중위수, 사분위수, 최빈수 등의 통계량

실습 (생활비 변수를 대상으로 대표값 산출)

1단계: 평균과 중위수

```
mean(x)
median(x)

sort(x)
sort(x, decreasing = T)
```

mean()함수, median()함수,

sort()함수
형식: sort(x, decreasing = FALSE, ...)
decreasing = TRUE : 내림차순
decreasing = FALSE 또는 option 제외 : 오름차순

2단계: 사분위수

quantile(x, 1/4)
quantile(x, 2/4)
quantile(x, 3/4)
quantile(x, 4/4)

실습 (생활비 변수의 최빈수)

1단계: 최빈수: 빈도수가 가장 많은 변량

length(x)
x.t <- table(x)
max(x.t)

table()함수

2단계: 2개의 행을 묶어서 matrix생성

x.m <- rbind(x.t)
class(x.m)
str(x.m)
#which(x.m[1,] == 18)
which(x.m == 18)

which()함수

3단계: 데이터프레임으로 변경

x.df <- as.data.frame(x.m)
#which(x.df[1,] == 18)

which(x.df == 18)

4단계: 최빈수와 변량 확인

x.df[1, 19]
attributes(x.df)
names(x.df[19])

attributes()함수: 특정 객체의 속성 정보를 확인할 수 있는 기능

(2) 산포도 구하기

산포도: 자료가 대표값으로부터 얼마나 흩어져 분포하고 있는가의 정도를 나타내는 척도
분산(Variance), 표준편차(Standard Deviation)

실습 (생활비 변수를 대상으로 산포도 구하기)

var(x)
sd(x)
sqrt(var(data$cost, na.rm = T))

var()함수, sd()함수, sqrt()함수

(3) 표본분산과 표본표준편차

표본분산
표본표준편차
표준오차: 표본과실제 모집단 간의 차이를 나타내는 값

변동계수(Coefficient of variation)
변동계수 = 표준편차/평균
측정단위가 다른 아이템의 편차를 비교하기 위해 사용
Ex. 평균값이 100일 때 표준편차가 10이면 변동성이 큼
평균값이 1000일 때 표준편차가 10이면 변동성이 작음

(4) 빈도분석

비율척도를 대상으로 직접 빈도분석을 수행한 결과는 의미가 없음.
일정한 간격으로 범주화 하여 빈도분석을 해야 의미가 있음

실습 (생활비 변수의 빈도분석과 시각화)

1단계: 연속형 변수의 빈도분석

```
table(data$cost)
```

2단계: 연속형 변수의 히스토그램 시각화

```
hist(data$cost)
```

3단계: 연속형 변수의 산점도 시각화

```
plot(data$cost)
```

4단계: 연속형 변수 범주화

```
data$cost2[data$cost >= 1 & data$cost <= 3] <- 1
data$cost2[data$cost >= 4 & data$cost <= 6] <- 2
data$cost2[data$cost >= 7] <- 3
```

5단계: 범주형 데이터 시각화

```
table(data$cost2)
par(mfrow = c(1, 2))
barplot(table(data$cost2))
pie(table(data$cost2))
```

2.5 비대칭도 구하기

왜도, 첨도

왜도: 평균을 중심으로 하는 확률분포의 비대칭 정도를 나타낸는 지표. 분포의 기울어진 방향과 정도를 나타내는 양
>0 : 분포의 오른쪽 방향으로 비대칭 꼬리가 치우침
<0 : 분포의 왼쪽 방향으로 비대칭 꼬리가 치우친다.
=0 : 평균을 중심으로 좌우대칭

첨도: 표준정규분포와 비교하여 얼마나 뾰족한가를 측정하는 지표
=0 (또는 3) : 정규분포 곡선
>0 : 정규분포보다 뽀족한 형태
<0 : 정규분포보다 완만한 곡선 형태
* 첨도식에서 -3을 적용하지 않으면 정규분포의 첨도는 3

실습 (패키지를 이용한 비대칭도 구하기)

1단계: 왜도와 첨도 사용을 위한 패키지 설치

```
install.packages("moments")
library(moments)
cost <- data$cost
```

moments 패키지

2단계: 왜도 구하기

```
skewness(cost)
```

skew()함수

3단계: 첨도 구하기

```
kurtosis(cost)
```

kurtosis()함수
*kurtosis()함수에서 첨도 식에서 -3을 미적용 → 정규분포첨도는 3

4단계: 히스토그램으로 왜도와 첨도 확인

```
hist(cost)
par(mfrow = c(1, 1))
```

hist()함수

실습 (히스토그램과 정규분포 곡선 그리기)

```
hist(cost, freq = F)
lines(density(cost), col = 'blue')
x <- seq(0, 8, 0.1)
curve(dnorm(x, mean(cost), sd(cost)), col = 'red', add = T)
```

line()함수: 분포선 추가
curve()함수: 정규분포 확률밀도 구함

실습 (attach()/detach()함수로 기술통계량 구하기)

```
attach(data)
length(cost)
summary(cost)
mean(cost)
min(cost)
max(cost)
range(cost)
sd <- sd(cost, na.rm = T)
sqrt(var(cost, na.rm = T))
sd(cost, na.rm = T)
detach(data)
```

attach()함수: database가 R search path에 추가됨. 데이터셋을 추가하면 이후부터는 'data$' 생략 가능
detach()함수: attach()함수 해제

실습 (NA제거 후 기술통계량)

1단계: NA가 있으면 error발생 함수

```
test <- c(1:5, NA, 10:20)
min(test)
max(test)
range(test)
mean(test)
```

데이터에 NA가 있는 경우 min(), max(), range(), mean()함수는 결과로 NA 출력

2단계: NA제거 후 통계량 구하기

```
min(test, na.rm = T)
max(test, na.rm = T)
range(test, na.rm = T)
mean(test, na.rm = T)
```

na.rm=T 로 NA제거

3. 기술통계량 보고서 작성

설문 조사 결과를 토대로 논문이나 보고서를 작성하는 경우 응답자의 인구통계학적 특성을 반드시 제시해야 함.

빈도분석과 기술통계량의 분석 결과를 토대로 표본의 인구통계학적 특성을 제시하는 방법

3.1 기술통계량 구하기

실습 (변수 리코딩과 빈도분석 하기)

1단계: 거주지역 변수의 리코딩과 비율계산

```
data$resident2[data$resident == 1] <- "특별시"
data$resident2[data$resident >= 2 & data$resident <= 4] <- "광역시"
data$resident2[data$resident == 5] <- "시구군"

x <- table(data$resident2)
x

prop.table(x)

y <- prop.table(x)
round(y * 100, 2)
```

2단계: 성별 변수의 리코딩과 비율계산

```
data$gender2[data$gender == 1] <- "남자"
data$gender2[data$gender == 2] <- "여자"
x <- table(data$gender2)
prop.table(x)
y <- prop.table(x)
round(y * 100, 2)
```

3단계: 나이 변수의 리코딩과 비율계산

```
data$age2[data$age <= 45] <- "중년층"
```

```
data$age2[data$age >= 46 & data$age <= 59] <- "장년층"
data$age2[data$age >= 60] <- "노년층"
x <- table(data$age2)
x

prop.table(x)
y <- prop.table(x)
round(y * 100, 2)
```

4단계: 학력수준 변수의 리코딩과 비율계산

```
data$level2[data$level == 1] <- "고졸"
data$level2[data$level == 2] <- "대졸"
data$level2[data$level == 3] <- "대학원졸"
x <- table(data$level2)
x

prop.table(x)
y <- prop.table(x)
round(y * 100, 2)
```

5단계: 합격여부 변수의 리코딩 및 비율계산

```
data$pass2[data$pass == 1] <- "합격"
data$pass2[data$pass == 2] <- "실패"
x <- table(data$pass)

x
prop.table(x)

y <- prop.table(x)
round(y * 100, 2)

head(data)
```

3.2 기술통계량 보고서 작성

[표 11.2] 표본의 인구통계적 특성 결과

연습문제 풀기