2018년 07월 05일 9일차

수치를 이용한 자료의 요약

1) 데이터의 비대칭도 : 왜도와 첨도

데이터의 비대칭도 : 왜도

> 분포의 모양이 대표값을 중심으로 대칭인지를 판단, 얼마나 기울어져 있는지 측정한 값
> 0보다 작으면 왼쪽, 0보다 크면 오른쪽으로 기울어지며, 0이면 대칭적인 분포를 의미한다.

데이터의 비대칭도 : 첨도

> 분포가 대표값을 중심으로 얼마나 모여있는가를 나타내는 척도
> 정규분포의 첨도는 3이며, 3보다 크면 뾰족해지고, 3보다 작으면 뭉툭해진다.

```
library(moments)
```
> moments 라이브러리를 이용한 첨도와 왜도 측정

```
x <- rnorm(3000, 30, 5)
skewness(x)
kurtosis(x)
```
> 정규분포의 왜도와 첨도. 표본이 많아질수록 0 과 3에 가까워진다

```
x2 <- rchisq(100, 2)
skewness(x2)
kurtosis(x2)
```
> 카이분포의 왜도와 첨도. 왜도가 1 이상, 첨도는 7에 가까운 것을 볼 수 있다

```
library(psych)
```
> psych 라이브러리를 이용한 첨도와 왜도 측정

```
describe(Sat.Data[, 4:6])
describe(x)
```
> skew 와 kurtosis 를 비롯해 다양한 요약값이 출력된다.
> 이 때, kurtosis 는 기준값이 3이 아닌 0 으로 연산되어 나온다는 점에 주의하자.

2) 도수분포표의 작성

> 계급의 수와 간격을 결정하고, 계급별로 자료의 빈도를 계산하여 표현한 도표

수량화 자료의 구간 범주형 자료로의 변환
cut (x , breaks, labels = NULL, include . lowest = F, right = T)
> breaks 는 계급의 구분점을 벡터형태로 표현한 것이다
> right 는 상한에 대한 포함 여부이다. T 일때 상한을 포함, F 일때 하한을 포함한다.
> 즉, right = T 이면 (a , b] 의 형태이고, right = F 이면 [a , b) 형태이다.
> include . lowest = T 이면 계급의 최초값을 포함하는 것이다.

```
Inter2 <- cut(Sat.Data$S, breaks = 7, include.lowest = T, right = T)
```
> breaks 를 계급의 수로 정하면 동일한 간격으로 계급이 자동생성된다.

3) 그 외 요약값 반환 함수

```
table ( x1, x2 )
```

```
apply( X, MARGIN, FUN = function )
```

lapply(X, FUN = function)

sapply(X, FUN = function , simplify = T)

for (i in 4:7){ print(mean(Sat.Data[,i])) }
> 이 문장은 apply(Sat.Data[, 4:7], MARGIN = 2, FUN = mean) 과 동일
> 하지만 apply 가 for문 보다 더 빠른 속도를 낸다

aggregate(Sat.Data[, 4:7], by = list(Sat.Data$City), FUN = mean)
> City 별 통계량을 구하는 함수이다. 이는 apply 를 이용할 수도 있다

apply(Sat.Data[Sat.Data$City == '서울시', 4:7], MARGIN = 2, FUN = mean)
> 위와 같은 City별 통계량의 함수이다. 효과적인 방법을 선택하여 사용할 수 있다

그래프를 이용한 자료의 요약

1) Windows OS의 화면 출력 방법

windows(width = 7, height = 7, rescale = 'R', bg = 'transparent')
win.graph(width = 7, height = 7, pointsize = 6)
win.metafile(filename = 'mygraph.wmf', width = 7, height = 7, pointsize = 12)

dev.list()
> 현재 사용중인 그래픽 장치 번호의 출력

graphics.off()
> 모든 그래픽 장치 닫기

dev.off(wihch = dev.cur()) / dev.set(which = dev.cur())
> 하나의 그래픽 장치 닫기 / 열기

dev.next(which = dev.cur()) / dev.prev(which = dev.cur())
> 현재 다음 / 이전의 그래픽 장치의 번호 출력

R Studio 에서는 Plot 이라는 기본 그래픽 장치가 존재한다
Export 기능을 이용해 밖으로 내보내기 또한 가능하다.

2) Mac OS 에서의 화면 출력 방법

Mac 에서는 windows 를 사용할 수 없고, quartz 라는 함수를 사용하여 창을 만든다.

quartz(width = 7, height = 7, bg = 'transparent')
> 가로 - 세로 길이와 배경색을 지정하여 quartz 그래픽 창을 생성한다.

quartz.save(file = 'abc', type = 'png', device = dev.cur())
> file 이름과 포맷을 정하여 특정 device 의 내용을 wd 에 저장한다.

3) 그래픽스 모수를 이용한 옵션 조정

par()
par(no.readonly = T)
> 모든 그래픽스 모수 / 옵션 조정 가능한 그래픽스 모수의 목록을 보여준다.

par(pch = 20)
pch.old <- par(pch = 1)
pch.old2 <- par(pch = 3, cex = 1.5)

> 옵션값을 특정 변수 (pch.old / old2) 에 저장하여 필요할 때 불러올 수 있다

plot(1:10, pch=19, cex=0.5)
> 그래프 생성 단계에서 그래픽스 모수의 변경 또한 가능하다. 이는 해당 그래프에 대해서만 적용하는 것이다.

par (oma = c(a, b, c, d)) / par (mar = c (a, b, c, d))
> 외부 / 내부 마진 조정

4) 그래프 함수 활용

범주 그래프 : 파이 / 막대 그래프

pie(c(1, 1, 1), labels = c('a', 'b', 'c'), main = 'my_pie_graph')
> 파이 그래프 생성 함수. 내용, 이름, 제목 순서로 입력한다.

barplot(c(2, 5, 3, 1.5, 0.5, 4), main = 'my_barplot_graph', xlab = 'x label', ylab = 'y label')
> 막대 그래프 생성 함수. 내용, 제목, x축, y축 이름 순서로 입력한다.

줄기 잎 그래프 (stem and leaf plot)

> 데이터에 대해 줄기와 잎 부분으로 구분함. 도수분포표 + 히스토그램의 형태임

stem(stem_data) 형태로 생성한다.

상자 그래프

boxplot(x1, x2, … , xn, horizontal = T)

> 중심과 퍼짐의 요약을 얻는 도표
> 최소값, 제1사분위수, 중앙값, 제3사분위수, 최대값 요소로 표현하게 된다.

X <- boxplot (~)
> 이렇게 변수 형태로 boxplot 을 저장하면, boxplot 내의 정보들이 X에 저장된다.
> out : 이상값 / group : 이상값이 속한 그룹 / 즉, out 과 group 을 함께봐야 한다.

히스토그램

hist(stem_data, breaks = 5, include.lowest = T, right = T, labels = F)

> 자료의 중심위치, 분포상태, 치우침, 산포 정도를 파악하는 도표
> 막대도표와의 큰 차이점은 계급경계에서 막대가 연결된다는 점

산점도

plot (x, y, type = ‘ p ‘)

> 두 변수 사이의 관계를 파악하기 위한 점의 분포 도표
> type 에는 p, l, b 가 들어가며, 각각 점, 선, 점과 선으로 표시하는 것이다.

화면 분할을 이용한 다중 그래프 생성

1) par 의 mfrow 속성을 이용한 분할

mfrow = c (a, b) / mfcol = c (a, b)
> 매트릭스 형태로 분할하여 순차적으로 그래프를 삽입

2) par 의 fig 설정을 이용한 분할

par(fig = c (a, b, c, d), new = T)
> 특정 위치의 특정 범위를 정하여 그래프를 삽입
> 이 때, 다음 그래프의 생성에 따라 이전 그래프가 자동삭제되지 않도록 new = T 로 맞춤

3) split . screen 을 이용한 분할

split . screen (fig = c (a, b) , screen = n)
> fig = c (a, b) 로 매트릭스 형태로 화면을 분할하며, screen 을 지정하여 다시 나눌 수 있다.
> 나누어진 화면에 대해, screen(n = x) 의 형태로 특정 화면을 선택하여 그래프 생성이 가능하다.
> erase . screen (n = x) 로 특정 화면을 삭제하는 것이 가능하다.

4) layout 을 이용한 분할

layout (mat, widths = rep(1, ncol (mat)), heights = rep(1, nrow(mat)))
> 분할하려는 형태를 행렬로 표현하여, 가장 첫 번째 인자로 입력한다.
> 해당 분할에 대한 비율을 widths 와 heights 로 지정하여 자유로운 크기의 화면분할이 가능하다.

5) 차트 제목 / 축 이름의 설정 / 범례의 설정 / 문자의 입력

title
> 차트의 주 제목, 보조 제목, 제목의 크기, 문자 색, x축, y축 이름을 정하여줄 수 있다.

legend
> 차트에 대한 정보를 표시하는 범례를 생성하는 함수이다.
> 범례의 위치, 범례에 대한 설명, 조각색, 외곽선 등등을 모두 지정할 수 있다.

text
> 그래프가 아닌, 문자만 나타나도록 만들어주는 함수이다.
> 문자에 대한 좌표, 문자의 내용, 문자의 색상, 효과, 크기 등을 모두 지정할 수 있다.