韓國數學敎育學會誌 시리즈 E
<數學敎育 論文集>
제 21집 제 2호, 2007. 5. 227-270

J. Korea Soc. Math. Ed. Ser. E:
Communications of Mathematical Education
Vol. 21, No. 2, May. 2007. 227-270

초·중·고등학교 확률 및 통계영역 교육에서의 R 통계패키지의 활용(II)

장 대 홍 (부경대학교)

장대홍(2007)에서는 R 패키지에 대한 전반적인 설명을 행하였다. 본 연구에서는 제 7차 수학과 교육과정 내의 확률 및 통계영역 목표와 내용을 중심으로 하고 제 7차 수학과 교육과정에 따라 집필된 중·고등학교 수학교과서들의 확률 및 통계단원을 참고로 하여 R 패키지를 구체적으로 수업에 어떻게 적용할 수 있는 지를 제안하여 보고자 한다.

I. 서 론

초·중·고등학교 확률 및 통계교육에서의 R 통계패키지의 활용(I)에서는 R 패키지에 대한 전반적인 설명을 행하였고 초·중·고등학교 확률 및 통계교육 현장에서 학생들이 표준 통계패키지로서 R을 배운다면 고등학교 졸업 후 대학에 가거나 사회에 진출해서도 표준 통계패키지로서 R을 계속 사용할 수 있기 때문에 통계패키지 사용의 연속성이라는 측면에서도 지금 시점이 초·중·고등학교 확률 및 통계교육 표준 통계패키지로서 R을 사용할 것인지를 심각하게 고려해야 할 시점이라고 언급하였다. 본 연구에서는 제 7차 수학과 교육과정 내의 확률 및 통계영역 목표와 내용을 중심으로 하고 제 7차 수학과 교육과정에 따라 집필된 1-10단계 수학교과서들(1-6단계: 교육인적자원부(2002), 7-9단계: 16종(2003), 10단계: 16종(2003))의 확률 및 통계단원을 참고로 하여 R 패키지를 구체적으로 수업에 어떻게 적용할 수 있는 지를 예제들을 통하여 살펴보고자 한다. 이 연구를 위하여 장대홍(1995), 장대홍외 2인(2000), 장대홍(2005)을 참조하였다.

II. 제 7차 수학과 교육과정에 따른 R의 활용

1. 제 7차 수학과 교육과정 1-10단계 확률과 통계영역의 목표와 내용

제 7차 수학과 교육과정 내의 1-10단계 확률과 통계영역 목표 체계표와 내용 체계표는 다음 <표 1>과 <표 2>와 같다.

* ZDM분류 : N80
* MSC2000분류 : 97U70
* 주제어 : 확률과 통계 교육, R 통계패키지

<표 1> 확률과 통계영역 목표 체계표

단계	단계별 목표
1-가	사물을 간단한 기준에 따라 분류할 수 있다.
2-나	간단한 자료의 크기를 표나 그래프로 나타낼 수 있다.
3-나	실생활에서 찾을 수 있는 간단한 자료의 크기를 표와 그래프로 정리할 수 있다.
4-나	꺾은선그래프를 알고, 이를 이용하여 자료를 정리하고 표현할 수 있다.
5-나	자료를 정리하여 이를 줄기와 잎 그림으로 나타낼 수 있고, 주어진 자료의 평균을 구 할 수 있다.
6-가	생활 속의 자료를 적절한 비율그래프로 표현할 수 있다.
6-나	경우의 수를 이해하고, 확률의 의미를 안다.
7-나	간단한 통계 자료를 조사, 정리하여 표나 그래프로 나타내고, 평균을 구할 수 있다.
8-나	확률의 뜻과 기본 성질을 이해하고 간단한 확률을 구할 수 있다.
9-나	상관도와 상관표를 알고, 두 변량 사이의 상관관계를 알 수 있다.
10-가	산포도와 표준편차를 구할 수 있다.

<표 2> 확률과 통계영역 내용 체계표

단계	단계별 내용
1-가	한 가지 기준으로 사물을 분류하기
2-나	표와 그래프 만들기
3-나	자료의 수집, 정리, 막대그래프로 나타내기
4-나	꺽은선그래프, 여러 가지 그래프로 나타내기
5-나	줄기와 잎 그림, 평균
6-가	비율그래프(띠그래프, 원그래프)
6-나	경우의 수와 확률
7-나	도수분포표, 히스토그램, 도수분포다각형, 도수분포표에서의 평균, 상대도수, 누적도수
8-나	확률의 뜻과 기본 성질, 확률의 계산
9-나	상관도, 상관표, 상관관계
10-가	산포도와 표준편차

1-10단계 중 6-나 단계와 8-나 단계에서 다루는 경우의 수와 확률을 제외하면 나머지 단계들은 모두 '자료의 정리와 요약'이라는 기술통계학 영역에 해당한다.

2. R의 활용

제 7차 수학과 교육과정 1-10단계 확률 및 통계영역 목표와 내용을 중심으로 하고 제 7차 수학과 교육과정에 따라 집필된 초·중·고등학교 수학교과서들의 확률 및 통계단원을 참고로 하여 R 패키지를 활용하는 방법을 예제들을 통하여 살펴보자. 예제 뒤에 따라 나오는 단계는 이 예제와 관련된 단계를 나타낸다.

<예제 1> (2-나, 3-나, 6-가 단계) M&M milk chocolate는 6가지 색깔의 단추모양의 밀크쵸코캔디의 상표명이다. 한 봉지 안에 들어있는 이 여섯 가지 색깔의 밀크쵸코캔디의 개수와 비율은 봉지마다 다름을 알 수 있다. 학생들이 구입한 M&M milk chocolate를 개봉하게 하여 6가지 색깔을 갖는 밀

크쵸코캔디 각각의 개수와 총 개수를 세어 기록하게 하고 학생 전체의 결과를 수집하여 분석하여 본다. 총 개수의 분포는 어떤 구조를 갖는 지, 6가지 색깔을 갖는 밀크쵸코캔디 각각의 개수는 어떠한 비율로 나뉘는 지를 각 종 그래프를 통하여 확인한다. 이러한 현상을 우리는 통계적 시뮬레이션으로 확인하여 볼 수 있다. 미국 M&M 홈페이지(http://us.mms.com/us/about/products/milkchocolate/)에 가보면 여섯 가지 색깔의 밀크쵸코캔디의 비율은 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=0.24: 0.20: 0.16: 0.14: 0.13: 0.13이라고 나와 있다. 이 비율로 여섯 가지 색깔의 밀크쵸코캔디를 포장한다고 가정하고 편의상 한 봉지 안의 밀크쵸코캔디의 개수는 100개라 하자. 그러면 이 분포는 다항분포가 된다. 우리는 다음과 같이 통계적 시뮬레이션을 행하여 볼 수 있다. 통계적 시뮬레이션 결과 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색의 비율이 0.24: 0.23: 0.18: 0.12: 0.11: 0.12가 나왔다. 즉 여섯 가지 색깔의 밀크쵸코캔디의 비율을 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=0.24: 0.20: 0.16: 0.14: 0.13: 0.13을 지키며 랜덤하게 하나의 봉지에 100개의 밀크쵸코캔디을 담은 결과 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=24개: 23개: 18개: 12개: 11개: 12개가 나왔다는 것이다. 100개의 밀크쵸코캔디를 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=24개: 20개: 16개: 14개: 13개: 13개가 되도록 한 봉지에 담으려고 노력했는 데 왜 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=24개: 23개: 18개: 12개: 11개: 12개가 나왔을까? 이것을 이해하는 것이 '변동(variation)'을 이해하는 것이 된다. 통계학은 크게 보면 다음과 같은 3가지 주제를 연구하는 학문이라 할 수 있다.

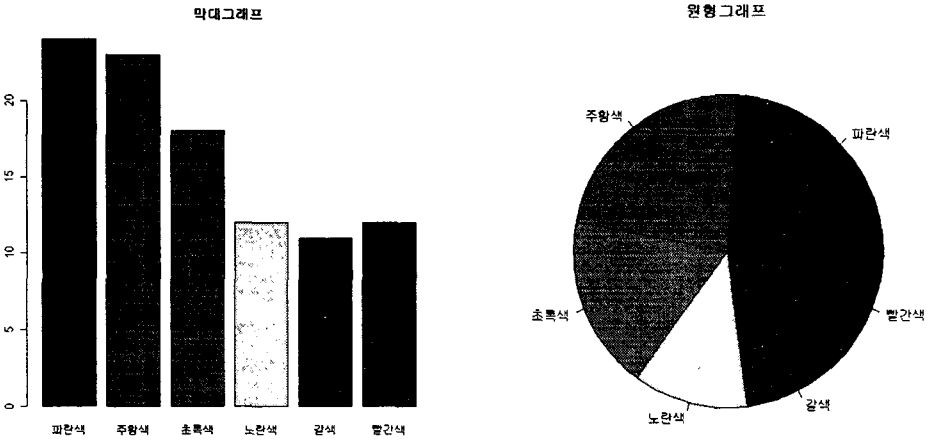
1. 모집단(population)
2. 자료의 축약방법(methods of data reduction)
3. 변동(variation)

기업체들이 자주 언급하는 식스시그마(6σ) 운동도 변동에 초점을 맞춘 개념이다. 키가 모두 큰 부모 하에서 주로 키 큰 자녀들이 나오지만 키가 작은 자녀도 나타나고 키가 모두 작은 부모 밑에서 주로 키가 작은 자녀들이 나오지만 키가 큰 자녀도 나타나는 현상도 변동의 문제인 것이다.

```
> # 다항분포에서 100개의 난수를 뽑기(M&M 밀크쵸코캔디)
> m.m=rmultinom(1,100,prob=c(24,20,16,14,13,13))
> mm=c(m.m[1],m.m[2],m.m[3],m.m[4],m.m[5],m.m[6])
> mm
[1] 24 23 18 12 11 12
> names(mm)=c("파란색", "주황색", "초록색", "노란색", "갈색", "빨간색")
> mm
파란색 주황색 초록색 노란색   갈색 빨간색
    24     23     18     12     11     12
> prop.table(mm)
파란색 주황색 초록색 노란색   갈색 빨간색
  0.24   0.23   0.18   0.12   0.11   0.12
```

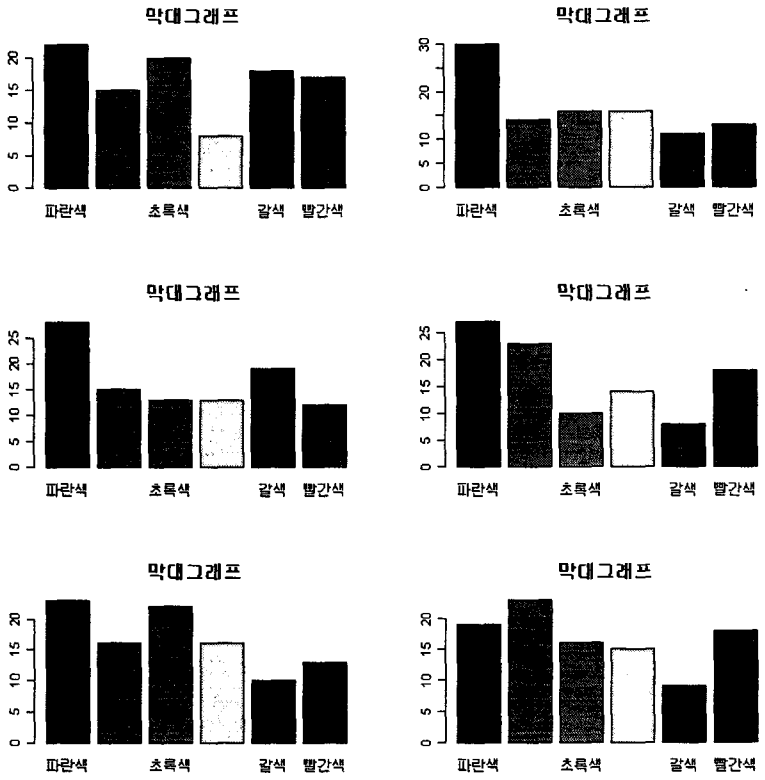
위의 시뮬레이션 결과를 막대그래프와 원형그래프로 그리면 다음과 같다.

```
> # 막대그래프
> barplot(mm, col=c("blue", "#FF4C00","green","yellow","brown","red"),main="막대그래프")
> # 원형그래프
> pie(mm, col=c("blue", "#FF4C00","green","yellow","brown","red"), main="원형그래프")
```

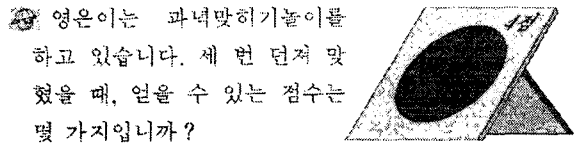


'변동'이라는 현상을 더 자세히 보기 위하여 위와 같은 시뮬레이션을 다음과 같이 6번 반복시행하여 보자. 우리는 6번 반복시행이 조금씩 다 다른 모습을 이루는 것을 볼 수 있다. 이러한 현상이 바로 '변동'이라는 현상이다. 이 시뮬레이션을 통하여서도 변동을 확인할 수 있다. 6번 반복시행 모두 여섯 가지 색깔의 밀크쵸코캔디의 비율을 파란색: 주황색: 초록색: 노란색: 갈색: 빨간색=0.24: 0.20: 0.16: 0.14: 0.13: 0.13을 지키며 시뮬레이션을 행하였는데도 결과는 다 다르다.

```
> par(mfrow=c(3,2))
> for(i in 1:6)
+ {
+ m.m=rmultinom(1,100,prob=c(24,20,16,14,13,13))
+ mm=c(m.m[1],m.m[2],m.m[3],m.m[4],m.m[5],m.m[6])
+ names(mm)=c("파란색", "주황색", "초록색", "노란색", "갈색", "빨간색")
+ barplot(mm, col=c("blue", "#FF4C00","green","yellow","brown","red"),main="막대그래프")
+ }
>
```



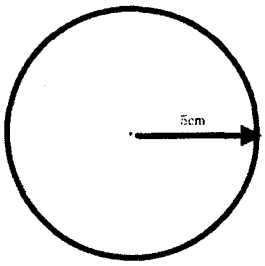
<예제 2> (2-나, 5-나, 6-나, 7-나, 10-가 단계) 2-나 교과서 7단원 '문제 푸는 방법 찾기' 중 '문제를 해결하여 봅시다'(115p)에 보면 다음과 같은 과녁맞히기놀이가 나온다. 화살을 던져 화살이 빨간 원 안에 꽂히면 5점, 빨간 원 밖으로 나가면 4점이다.



이 문제를 변동의 문제로 보기 위하여 관심의 초점을 빨간 원의 중심(앞으로 '원중심'이라 하자.)과 꽂힌 화살촉 사이의 거리에 두고 보자. 화살을 100번 던져 보면 100개의 화살촉 자국이 원중심을 중심으로 흩어져 있을 것이다. 우리는 화살을 던질 때 원중심을 향하여 던지는 데 던진 결과는 원중심을 중심으로 흩어지게 되는 것이다. 이러한 현상이 바로 '변동'이라는 현상이다. 학생들에게 원중심과 꽂힌 화살촉 사이의 거리를 재어 기록 한 후 크기순으로 늘어놓게 한다. 5학년 이상의 학생들을 대상으로 할 때는 줄기와 잎 그림을 그려보게 한다. 오른 쪽으로 치우친 비대칭분포를 이루는 것을 확

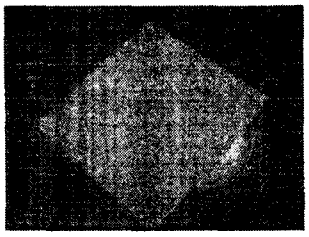
인하게 될 것이다. 제조업체에서 제품을 만들 때 제품의 규격(specification)에서 목표값(target value)을 설정하고 제품이 이 목표값을 갖도록(예로 나사를 제조하는 제조업체에서 나사의 지름의 목표값이 1 mm라고 하면 나사의 지름이 1 mm보다 커도 안 되고 나사의 지름이 1 mm보다 작아도 안 된다.) 제품을 만드는 데 제품을 만들어 보면 제품이 목표값을 갖도록 노력을 하여도 목표값을 중심으로 흩어지게 된다. 제조업체에서는 이 제품의 변동을 어떻게 하면 최대한 줄일 것인가를 강구하게 되는 것이다. 서비스산업에서도 서비스요원들의 서비스품질을 어떻게 균질화할 것인가, 즉 서비스품질의 변동을 어떻게 하면 줄일 것인가가 기업의 사활이 걸린 문제가 되곤 한다.

위의 예와 같은 과녁맞히기놀이보다 더 손쉬운 방법이 동전을 가지고 하는 놀이일 것이다. 다음과 같은 반지름이 5cm인 원을 도화지에 그리고 이 원의 중심으로부터 1 m 떨어진 거리에서 동전을 100번 던졌을 때 동전의 중심과 원의 중심과의 거리를 재어 히스토그램과 줄기와 잎그림을 학생들에게 그려보게 하였다. 또한 동전 100개를 던진 것 중 원 안에 들어가는 경우의 비율을 구하게 하였다.

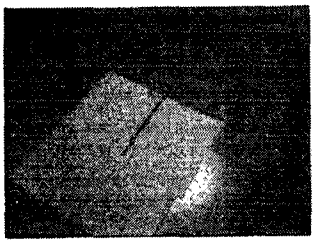


다음 사진들은 위의 실험 과정을 사진으로 찍은 것들이다.

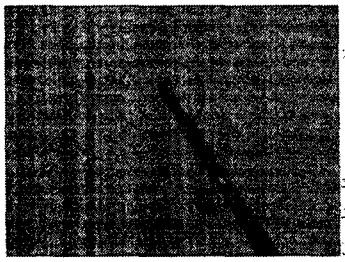
숙제 1) 100원짜리 동전을 100번 던져 각각의 동전의 중심과 원의 중심과의 거리를 재어 기록하기 --> 이 거리는 어떤 분포를 이룰까?



(먼저 큰 종이위에 지름 5cm인 원을 그려 넣었다.)



(그리고 1m 떨어진 지점을 표시하여 그곳에서 동전을 던졌다.)



(동전을 던지고 나면 줄자로 원의 중심과 동전의 중심과의 거리를 재어서 기록하였다.)

** 이런 숙제를 약간 복잡하다면 복잡한 장판이 깔린 방에서 해서 그런지 동전을 던질때 동전이 그리 많이 튀지 않고 착 달라붙어서 어느 한 경우도 저 원통이 밖을 벗어나는 경우는 없었다.

어느 한 학생의 실험 결과(동전을 100번 던졌을 때 동전의 중심과 원의 중심과의 거리)를 보면 다음과 같았다.

1.0, 1.4, 1.2, 1.4, 1.7, 2.2, 3.8, 7.9, 5.6, 5.1, 3.2, 2.5, 1.7, 1.4, 1.0, 0.7, 1.6, 2.2, 2.5, 9.0,
9.3, 3.9, 2.3, 2.4, 1.0, 1.1, 1.1, 1.4, 2.0, 0.7, 3.8, 2.6, 1.6, 1.2, 0.3, 0.4, 2.1, 1.9, 8.4, 5.5,
1.2, 1.1, 1.0, 1.5, 1.7, 2.9, 3.0, 0.6, 0.9, 11.9, 1.1, 1.5, 2.7, 2.9, 0.9, 1.3, 5.7, 3.6, 0.6, 0.4,
2.1, 2.5, 2.0, 1.6, 2.7, 3.6, 2.6, 0.9, 2.3, 3.0, 1.2, 1.0, 0.3, 1.5, 2.5, 3.2, 3.4, 2.6, 4.7, 8.8,
1.2, 1.0, 0.4, 3.0, 2.9, 2.6, 2.8, 1.0, 1.7, 1.4, 1.2, 0.7, 1.0, 1.9, 2.4, 1.2, 0.9, 0.4, 12.5, 0.2

이 자료에 대하여 동전 100개를 던진 것 중 원 안에 들어가는 경우의 비율을 구하고 줄기와 잎 그림을 그리면 다음과 같다. 동전 100개를 던진 것 중 원 안에 들어가는 경우의 비율은 0.89이고 동전의 중심과 원의 중심과의 거리는 오른 쪽으로 치우친 비대칭분포를 이루는 것을 확인할 수 있다.

```
> # 원 안에 동전 던지기
> coin=c(1.0, 1.4, 1.2, 1.4, 1.7, 2.2, 3.8, 7.9, 5.6,  5.1, 3.2, 2.5, 1.7, 1.4, 1.0, 0.7, 1.6, 2.2,  2.5, 9.0,
+ 9.3, 3.9, 2.3, 2.4, 1.0, 1.1, 1.1, 1.4, 2.0,  0.7, 3.8, 2.6, 1.6, 1.2, 0.3, 0.4, 2.1, 1.9,  8.4, 5.5,
+ 1.2, 1.1, 1.0, 1.5, 1.7, 2.9, 3.0, 0.6, 0.9, 11.9, 1.1, 1.5, 2.7, 2.9, 0.9, 1.3, 5.7, 3.6,  0.6, 0.4,
+ 2.1, 2.5, 2.0, 1.6, 2.7, 3.6, 2.6, 0.9, 2.3,  3.0, 1.2, 1.0, 0.3, 1.5, 2.5, 3.2, 3.4, 2.6,  4.7, 8.8,
+ 1.2, 1.0, 0.4, 3.0, 2.9, 2.6, 2.8, 1.0, 1.7,  1.4, 1.2, 0.7, 1.0, 1.9, 2.4, 1.2, 0.9, 0.4, 12.5, 0.2)
> coin
  [1]  1.0  1.4  1.2  1.4  1.7  2.2  3.8  7.9  5.6  5.1  3.2  2.5  1.7  1.4  1.0
 [16]  0.7  1.6  2.2  2.5  9.0  9.3  3.9  2.3  2.4  1.0  1.1  1.1  1.4  2.0  0.7
 [31]  3.8  2.6  1.6  1.2  0.3  0.4  2.1  1.9  8.4  5.5  1.2  1.1  1.0  1.5  1.7
 [46]  2.9  3.0  0.6  0.9 11.9  1.1  1.5  2.7  2.9  0.9  1.3  5.7  3.6  0.6  0.4
 [61]  2.1  2.5  2.0  1.6  2.7  3.6  2.6  0.9  2.3  3.0  1.2  1.0  0.3  1.5  2.5
 [76]  3.2  3.4  2.6  4.7  8.8  1.2  1.0  0.4  3.0  2.9  2.6  2.8  1.0  1.7  1.4
 [91]  1.2  0.7  1.0  1.9  2.4  1.2  0.9  0.4 12.5  0.2
> sort(coin)
  [1]  0.2  0.3  0.3  0.4  0.4  0.4  0.4  0.6  0.6  0.7  0.7  0.7  0.9  0.9  0.9
 [16]  0.9  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.1  1.1  1.1  1.1  1.2  1.2
 [31]  1.2  1.2  1.2  1.2  1.2  1.3  1.4  1.4  1.4  1.4  1.4  1.5  1.5  1.5  1.6
 [46]  1.6  1.6  1.7  1.7  1.7  1.7  1.9  1.9  2.0  2.0  2.1  2.1  2.2  2.2  2.3
 [61]  2.3  2.4  2.4  2.5  2.5  2.5  2.5  2.6  2.6  2.6  2.6  2.7  2.7  2.8  2.9
 [76]  2.9  2.9  3.0  3.0  3.0  3.2  3.2  3.4  3.6  3.6  3.8  3.8  3.9  4.7  5.1
 [91]  5.5  5.6  5.7  7.9  8.4  8.8  9.0  9.3 11.9 12.5

> # 동전을 던진 횟수 중 원 안에 드는 횟수의 비율
> length(coin[coin<=5])/length(coin)
[1] 0.89
```

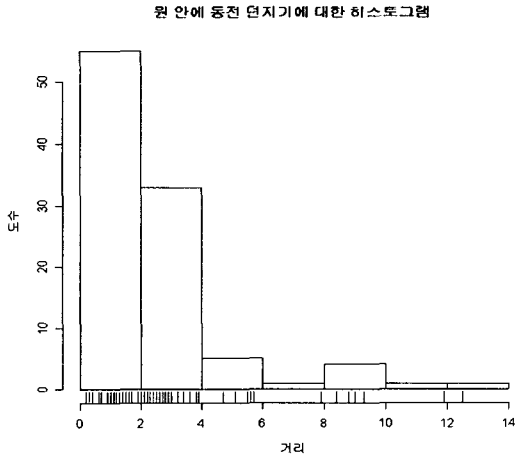
```
> # 줄기와 잎 그림
> stem(coin)

  The decimal point is at the |

   0 | 233444466777999900000000111122222223444445555666777799
   2 | 00112233445555666677899900022466889
   4 | 71567
   6 | 9
   8 | 4803
  10 | 9
  12 | 5
```

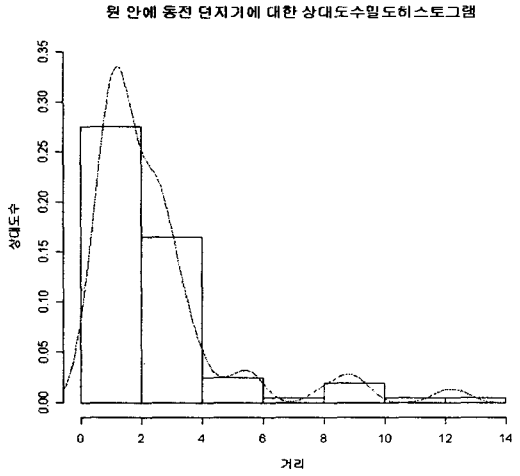
도수분포표와 히스토그램을 그리면 다음과 같다. 오른쪽으로 치우친 비대칭분포를 이루는 것을 확인할 수 있다.

```
> # 계급의 폭을 2로 하는 도수분포표
> class1=cut(coin, breaks=c(0, 2, 4, 6, 8, 10, 12, 14))
> table1=table(class1)
> table1
class1
  (0,2]   (2,4]   (4,6]   (6,8]  (8,10] (10,12] (12,14]
     55      33       5       1       4       1       1
> # 자료의 값을 표시한 히스토그램
> par(mfrow=c(1,1))
> hist(coin, xlab="거리", ylab="도수", main="원 안에
동전 던지기에 대한 히스토그램")
> rug(coin)
>
```



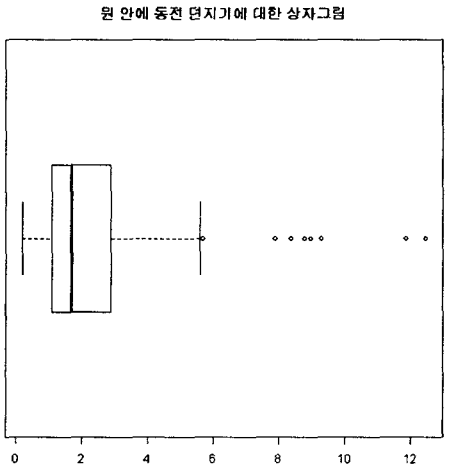
상대도수밀도히스토그램을 밀도함수추정량과 함께 그리면 다음과 같다.

```
> # 상대도수밀도히스토그램
> hist(coin, prob=T, ylim=c(0,0.35), xlab="거리", ylab="상대도수", main="원 안에 동전 던지기에 대한 상대도수
밀도히스토그램")
> lines(density(coin), col="red")
>
```



참고로 상자그림을 그리면 다음과 같다. 특이값(이상값, outlier)이 8개(5.7, 7.9, 8.4, 8.8, 9.0, 9.3, 11.9, 12.5)가 있음을 알 수 있다.

```
> # 상자그림
> boxplot(coin, horizontal=T, main="원 안에 동전 던지기에 대한 상자그림")
>
```



이 자료에 대하여 중심위치와 퍼진 정도를 나타내는 수치적 측도들을 구하면 다음과 같다.

```
> # 중심위치와 퍼진 정도를 나타내는 수치적 측도들을 구하기
> # 산술평균
> mean(coin)
[1] 2.495
```

```
> # 중앙값
> median(coin)
[1] 1.7
> # 분산
> var(coin)
[1] 5.522298
> # 표준편차
> sd(coin)
[1] 2.349957
> # 범위(R에서는 범위를 호출하면 최소값과 최대값을 출력함.)
> range(coin)
[1]  0.2 12.5
> # 수치적 측도 요약(다섯숫자 요약(최소값, 제 1사분위수, 중앙값, 제 3사분위수, 최대값)+산술평균)
> summary(coin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.200   1.100   1.700   2.495   2.900  12.500
> # 사분위수간범위=제 3사분위수-제 1사분위수
> IQR(coin)
[1] 1.8
```

<예제 3> (3-나, 6-가, 7-나 단계) 동양계 어느 민족의 지문을 조사하니 조사대상 1,000명 중 궁상문(arch)이 27명, 제상문(loop)이 480명, 와상문(whorl)이 493명이었다. 이 자료를 막대그래프와 원형그래프로 그려라.

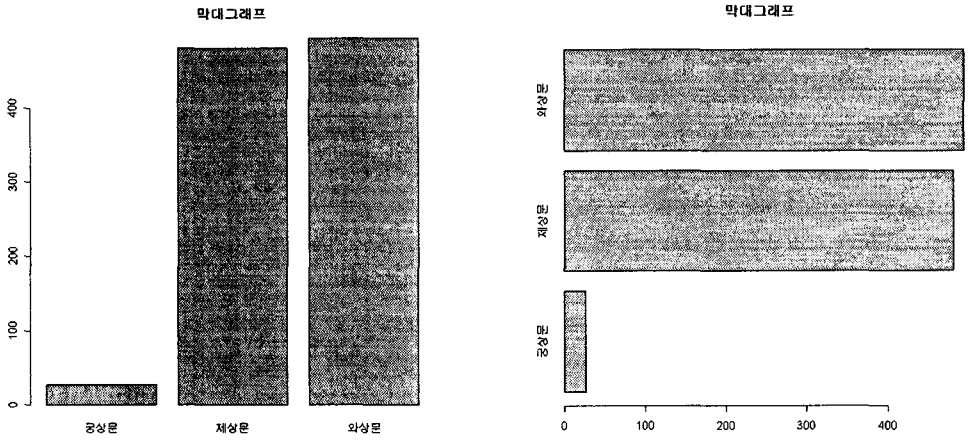
(풀이)

● 우선 도수분포표와 상대도수분포표를 작성하여 보자.

```
> # 도수분포표
> 지문=c(27, 480, 493)
> names(지문)=c("궁상문", "제상문", "와상문")
> 지문
궁상문 제상문 와상문
   27    480    493
> # 상대도수분포표
> prop.table(지문)
궁상문 제상문 와상문
0.027  0.480  0.493
```

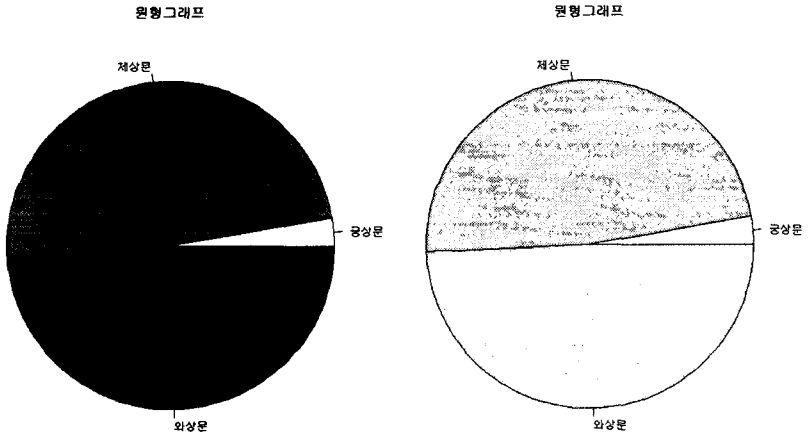
● 막대그래프를 그려보자.

```
> # 수직막대그래프
> barplot(지문, col="cyan", main="막대그래프")
> # 수평막대그래프
> barplot(지문, horiz=T, col="cyan", main="막대그래프")
```



● 원형그래프를 그려보자.

```
> # 원형그래프
> pie(지문, main="원형그래프")
> # 원형그래프(색깔 지정)
> slices=c("white","grey50","black")
> pie(지문, col=slices, main="원형그래프")
```



<예제 4> (3-나, 6-가, 9-나 단계) 다음 표는 어느 대학교의 4개의 학부(이 대학교는 4개의 학부로 구성되어 있음.)에 지원한 학생 수와 합격한 학생 수를 남녀별로 정리한 표이다. 이런 표를 분할표(contingency table)라고 한다.

(괄호안은 지원자수)

	합격한 남학생	합격한 여학생
학부 A	511(825)	89(108)
B	352(560)	17(25)
C	137(407)	132(375)
D	22(373)	24(341)
합계	1,022(2,165)	262(849)

(a) 전체 합격률, 남자합격률과 여자합격률을 막대그래프로 그려라. 무엇을 알 수 있는가?
(b) 각 학부별 합격률, 남자합격률과 여자합격률을 막대그래프로 그려라. 무엇을 알 수 있는가?
(c) (a)와 (b)를 종합하여 결론을 내려 보아라.
(d) 합격한 남학생과 합격한 여학생의 학부 분포율을 원형그래프로 그려 비교하여라.

(풀이) (a)

● 우선 분할표를 다음과 같이 작성하여 보자.

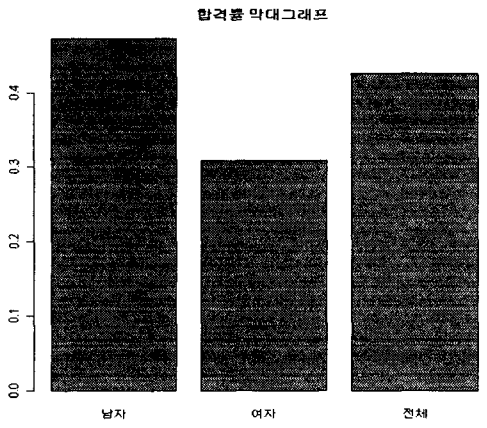
```
> # 분할표
> entrance=matrix(c(511, 825, 89, 108, 352, 560, 17, 25, 137, 407, 132, 375, 22, 373, 24, 341, 1022, 2165, 262, 849), nrow=5, byrow=T)
> entrance
     [,1] [,2] [,3] [,4]
[1,]  511  825   89  108
[2,]  352  560   17   25
[3,]  137  407  132  375
[4,]   22  373   24  341
[5,] 1022 2165  262  849
> colnames(entrance)=c("합격 남학생", "지원 남학생", "합격 여학생", "지원 여학생")
> rownames(entrance)=c("학부 A", "학부 B", "학부 C", "학부 D", "전체")
> entrance
       합격 남학생 지원 남학생 합격 여학생 지원 여학생
학부 A         511         825          89         108
학부 B         352         560          17          25
학부 C         137         407         132         375
학부 D          22         373          24         341
전체          1022        2165         262         849
```

● 분할표를 이용하여 합격률표를 다음과 같이 작성한다.

```
> prop.man.entrance=entrance[,1]/entrance[,2]
> prop.woman.entrance=entrance[,3]/entrance[,4]
> prop.total.entrance=(entrance[,1]+entrance[,3])/(entrance[,2]+entrance[,4])
> prop.entrance=cbind(prop.man.entrance, prop.woman.entrance, prop.total.entrance)
> prop.entrance
```

	prop.man.entrance	prop.woman.entrance	prop.total.entrance
학부 A	0.61939394	0.82407407	0.64308682
학부 B	0.62857143	0.68000000	0.63076923
학부 C	0.33660934	0.35200000	0.34398977
학부 D	0.05898123	0.07038123	0.06442577
전체	0.47205543	0.30859835	0.42601194

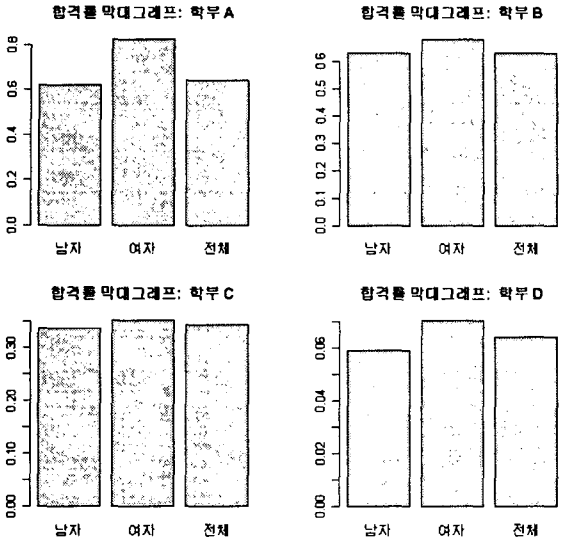
● 전체 합격률, 남자합격률과 여자합격률을 막대그래프로 그려보면 다음과 같다. 전체 남자의 합격률이 여자의 합격률보다 큼을 알 수 있다.




```
> # 막대그래프
> prop.total=prop.entrance[5,]
> names(prop.total)=c("남자", "여자", "전체")
> prop.total
      남자       여자       전체
 0.4720554  0.3085984  0.4260119
> barplot(prop.total, col="green", main="합격률 막대그래프")
```

(b) 각 학부별 합격률, 남자합격률과 여자합격률을 막대그래프로 그려보면 다음과 같다. 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 큼을 알 수 있다.

```
> # 학부별 막대그래프
> par(mfrow=c(2,2))
> for (i in 1:4)
+ {
+ prop=prop.entrance[i,]
+ names(prop)=c("남자", "여자", "전체")
+ barplot(prop, col="yellow", main=paste("합격률 막대그래프: ", rownames(entrance)[i]))
+ }
>
```

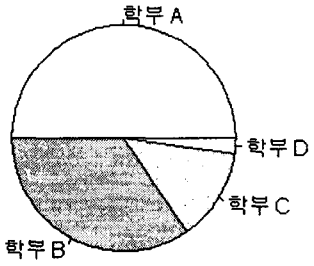


(c) 전체적으로는 남자의 합격률이 여자의 합격률보다 크나 각 학부별로는 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 큼을 알 수 있다. 이러한 현상을 심슨의 역설(Simpson's paradox)이라고 한다.

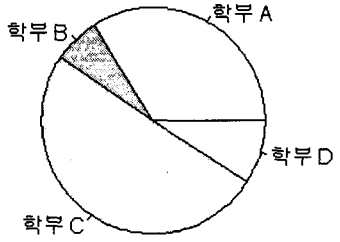
(d) 합격한 남학생과 합격한 여학생의 학부 분포율을 원형그래프로 그리면 다음과 같다. 합격한 남학생과 합격한 여학생의 학부 분포율이 아주 다름을 알 수 있다. 합격한 남학생은 A(50.0%) -> B(34.4%) -> C(13.4%) -> D(2.2%) 순이나 합격한 여학생은 C(50.0%) -> A(34.0%) -> D(9.1%) -> B(6.5%) 순이다.

```
> # 원형그래프
> par(mfrow=c(1,2))
> 합격남학생=c(511, 352, 137, 22)
> names(합격남학생)=c("학부 A", "학부 B", "학부 C", "학부 D")
> 합격여학생=c(89, 17, 132, 24)
> names(합격여학생)=c("학부 A", "학부 B", "학부 C", "학부 D")
> pie(합격남학생, main="합격남학생 원형그래프")
> pie(합격여학생, main="합격여학생 원형그래프")
>
```

합격남학생 원형그래프



합격여학생 원형그래프



<예제 5> (4-나 단계) 다음 자료는 2007년 부산지역 1월 한 달 동안의 일평균기온(°C) 자료이다.

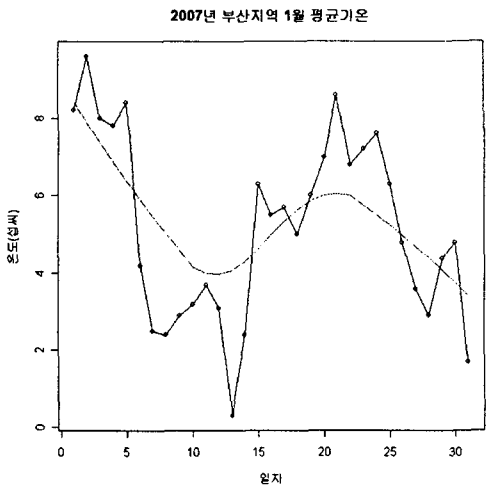
일자	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
온도	8.2	9.6	8.0	7.8	8.4	4.2	2.5	2.4	2.9	3.2	3.7	3.1	0.3	2.4	6.3

일자	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
온도	5.5	5.7	5.0	6.0	7.0	8.6	6.8	7.2	7.6	6.3	4.8	3.6	2.9	4.4	4.8	1.7

꺽은선그래프를 작성하여라.

(풀이) 꺽은선그래프를 다음과 같이 작성할 수 있다. 영도 이하로 떨어진 날이 없음을 알 수 있고 2주 동안은 온도가 내려가다 나머지 2주 동안은 온도가 다시 오르다가 내려갔음을 알 수 있다. 첨가한 곡선은 평활모수 α=0.75인 LOWESS이다.

```
> # 2007년 부산지역 1월 평균기온(단위: 섭씨온도)
> 온도=c(8.2, 9.6, 8.0, 7.8, 8.4, 4.2, 2.5, 2.4, 2.9, 3.2, 3.7, 3.1, 0.3, 2.4, 6.3
+ ,5.5, 5.7, 5.0, 6.0, 7.0, 8.6, 6.8, 7.2, 7.6, 6.3, 4.8, 3.6, 2.9, 4.4, 4.8, 1.7)
> 온도
 [1] 8.2 9.6 8.0 7.8 8.4 4.2 2.5 2.4 2.9 3.2 3.7 3.1 0.3 2.4 6.3 5.5 5.7 5.0 6.0
[20] 7.0 8.6 6.8 7.2 7.6 6.3 4.8 3.6 2.9 4.4 4.8 1.7
> 일자=seq(1:31)
> 일자
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[26] 26 27 28 29 30 31
> plot(일자, 온도, type="o", ylab="온도(섭씨)", main="2007년 부산지역 1월 평균기온")
> lines(lowess(일자, 온도), col="red")
>
```



<예제 6> (5-나, 7-나, 10-가 단계) 다음 자료는 미국 Yellowstone 국립공원에 있는 간헐온천(Old Faithful geyser)의 온천물의 지속시간을 분 단위로 잰 자료(107개)이다.

4.37, 3.87, 4.00, 4.03, 3.50, 4.08, 2.25, 4.70, 1.73, 4.93, 1.73, 4.62, 3.43, 4.25, 1.68, 3.92, 3.68, 3.10, 4.03, 1.77, 4.08, 1.75, 3.20, 1.85, 4.62, 1.97, 4.50, 3.92, 4.35, 2.33, 3.83, 1.88, 4.60, 1.80, 4.73, 1.77, 4.57, 1.85, 3.52, 4.00, 3.70, 3.72, 4.25, 3.58, 3.80, 3.77, 3.75, 2.50, 4.50, 4.10, 3.70, 3.80, 3.43, 4.00, 2.27, 4.40, 4.05, 4.25, 3.33, 2.00, 4.33, 2.93, 4.58, 1.90, 3.58, 3.73, 3.73, 1.82, 4.63, 3.50, 4.00, 3.67, 1.67, 4.60, 1.67, 4.00, 1.80, 4.42, 1.90, 4.63, 2.93, 3.50, 1.97, 4.28, 1.83, 4.13, 1.83, 4.65, 4.20, 3.93, 4.33, 1.83, 4.53, 2.03, 4.18, 4.43, 4.07, 4.13, 3.95, 4.10, 2.72, 4.58, 1.90, 4.50, 1.95, 4.83, 4.12

(a) 계급의 폭을 0.5로 하고 제 1계급의 하한치(원점이라고도 함.)를 1.35로 하는 히스토그램과 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.50으로 하는 히스토그램을 그려 서로 비교하여라.
(b) (a)에서의 구한 두 개의 히스토그램에서 계급의 폭을 0.5, 0.5/3, 0.5/5로 줄여나간 3개의 상대도수 밀도히스토그램(히스토그램의 높이 = (상대도수/계급의 폭)인 히스토그램)을 각각 그리고 비교하여라.
(c) (a)에서의 구한 두 개의 히스토그램을 이용하여 평균과 분산을 각각 구하고 원 자료를 이용한 평균 및 분산과 비교하여라.
(d) 제 1계급의 하한치를 1.50으로 하고 계급의 폭이 0.1, 0.2, 0.3, 0.4, 0.5인 히스토그램을 이용하여 평균과 분산을 각각 구하고 원 자료를 이용한 평균 및 분산과 비교하여라.

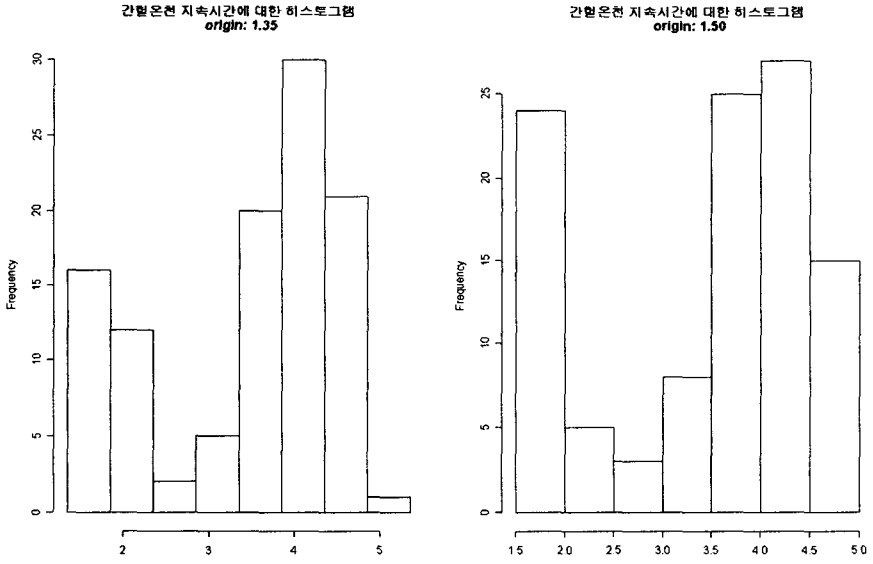
(풀이) (a)
우선 두 종류의 히스토그램을 그리기 위한 각각의 도수분포표를 작성하면 다음과 같다.

```
> # eruption lengths(in minutes) of 107 eruptions of Old Faithful geyser
> eruption.length=c(4.37,3.87,4.00,4.03,3.50,4.08,2.25,4.70,1.73,4.93,1.73,4.62,3.43,4.25
+ ,1.68,3.92,3.68,3.10,4.03,1.77,4.08,1.75,3.20,1.85,4.62,1.97,4.50,3.92
+ ,4.35,2.33,3.83,1.88,4.60,1.80,4.73,1.77,4.57,1.85,3.52,4.00,3.70,3.72
+ ,4.25,3.58,3.80,3.77,3.75,2.50,4.50,4.10,3.70,3.80,3.43,4.00,2.27,4.40
+ ,4.05,4.25,3.33,2.00,4.33,2.93,4.58,1.90,3.58,3.73,3.73,1.82,4.63,3.50
+ ,4.00,3.67,1.67,4.60,1.67,4.00,1.80,4.42,1.90,4.63,2.93,3.50,1.97,4.28
+ ,1.83,4.13,1.83,4.65,4.20,3.93,4.33,1.83,4.53,2.03,4.18,4.43,4.07,4.13
+ ,3.95,4.10,2.72,4.58,1.90,4.50,1.95,4.83,4.12)
> eruption.length
  [1] 4.37 3.87 4.00 4.03 3.50 4.08 2.25 4.70 1.73 4.93 1.73 4.62 3.43 4.25 1.68
 [16] 3.92 3.68 3.10 4.03 1.77 4.08 1.75 3.20 1.85 4.62 1.97 4.50 3.92 4.35 2.33
 [31] 3.83 1.88 4.60 1.80 4.73 1.77 4.57 1.85 3.52 4.00 3.70 3.72 4.25 3.58 3.80
 [46] 3.77 3.75 2.50 4.50 4.10 3.70 3.80 3.43 4.00 2.27 4.40 4.05 4.25 3.33 2.00
 [61] 4.33 2.93 4.58 1.90 3.58 3.73 3.73 1.82 4.63 3.50 4.00 3.67 1.67 4.60 1.67
 [76] 4.00 1.80 4.42 1.90 4.63 2.93 3.50 1.97 4.28 1.83 4.13 1.83 4.65 4.20 3.93
 [91] 4.33 1.83 4.53 2.03 4.18 4.43 4.07 4.13 3.95 4.10 2.72 4.58 1.90 4.50 1.95
[106] 4.83 4.12
# 표본의 크기
> n=length(eruption.length)
> n
[1] 107
# 범위
> range(eruption.length)
[1] 1.67 4.93
> class1=seq(1.35,5.35,by=0.5)
> class1
[1] 1.35 1.85 2.35 2.85 3.35 3.85 4.35 4.85 5.35
> class2=seq(1.5,5.0,by=0.5)
> class2
[1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
# 계급의 폭을 0.5로 하고 제 1계급의 하한치(원점이라고도 함.)를 1.35로 하는 도수분포표
> cat.class1=cut(eruption.length,breaks=class1)
> table(cat.class1)
cat.class1
(1.35,1.85] (1.85,2.35] (2.35,2.85] (2.85,3.35] (3.35,3.85] (3.85,4.35]
         16          12           2           5          20          30
(4.35,4.85] (4.85,5.35]
         21           1
# 계급의 폭을 0.5로 하고 제 1계급의 하한치(원점이라고도 함.)를 1.50으로 하는 도수분포표
```

```
> cat.class2=cut(eruption.length,breaks=class2)
> table(cat.class2)
cat.class2
(1.5,2] (2,2.5] (2.5,3] (3,3.5] (3.5,4] (4,4.5] (4.5,5]
   24       5       3       8      25      27      15
```

두 개의 히스토그램을 그리면 다음과 같다.

```
> # 히스토그램
> par(mfrow=c(1,2))
> hist(eruption.length,breaks=class1,main="간헐온천 지속시간에 대한 히스토그램 \n origin: 1.35",xlab="간헐온천 지속시간")
> hist(eruption.length,breaks=class2,main="간헐온천 지속시간에 대한 히스토그램 \n origin: 1.50",xlab="간헐온천 지속시간")
```



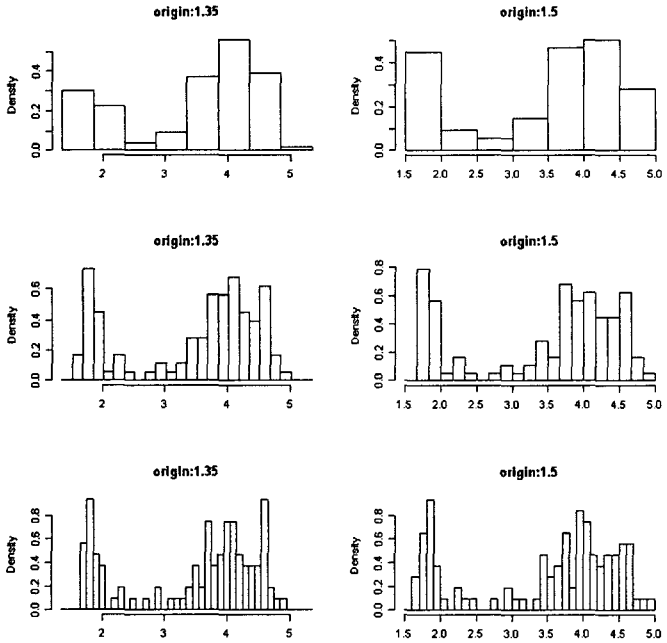
두 개의 히스토그램 모두 쌍봉분포(이봉분포, bimodal distribution)를 이루나 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.50으로 하는 히스토그램에서는 두 개의 봉우리가 거의 같은 높이이나 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.35로 하는 히스토그램에서는 두 개의 봉우리의 높이가 달라 두 개의 히스토그램에 대한 전체적인 인상이 다르다고 느끼게 된다. 이처럼 히스토그램에서는 계급의 폭을 얼마로 할 것인지, 제 1계급의 하한치를 무엇으로 할 것인지 하는 결정이 히스토그램의 모양을 결정하게 되므로 주의를 기울여야 한다. 같은 자료에 대하여 계급의 폭과 제 1계급의 하한치를 여러 가지 값으로 변경시켜가며 복수 개의 히스토그램을 그린 후 분포에 대하여 최종 결정을 내리는 것

이 좋은 전략이다.

(b) (a)에서의 구한 두 개의 히스토그램에서 계급의 폭을 0.5, 0.5/3, 0.5/5로 줄여나간 3개의 상대도수밀도히스토그램(히스토그램의 높이 = (상대도수/계급의 폭)인 히스토그램)을 각각 그리면 다음과 같다.

```
> # 계급의 폭 변화에 따른 상대도수밀도히스토그램의 변화
> hist.func2=function(n)
+ {par(mfrow=c(n,2))
+ for(i in 1:n)
+ {class1=seq(1.35,5.35,by=0.5/(2*i-1))
+ class2=seq(1.5,5.0,by=0.5/(2*i-1))
+ hist(eruption.length,breaks=class1,probability=T,main="origin:1.35",xlab=NULL)
+ hist(eruption.length,breaks=class2,probability=T,main="origin:1.5",xlab=NULL)
+ }
+ }
> hist.func2(3)
```

계급의 폭이 0.5일 때는 두 개의 히스토그램에 대한 인상이 다르게 느껴졌으나 계급의 폭이 0.5/3, 0.5/5로 작아짐에 따라 두 개의 히스토그램이 비슷한 구조를 갖게 됨을 알 수 있다.



(c) 원 자료를 이용하여 평균과 분산을 구하고 (a)에서의 구한 두 개의 히스토그램을 이용하여 평균과 분산을 각각 구하면 다음과 같다.

```
> # 원자료에 대한 정렬
> sort(eruption.length)
  [1] 1.67 1.67 1.68 1.73 1.73 1.75 1.77 1.77 1.80 1.80 1.82 1.83 1.83 1.83 1.85
 [16] 1.85 1.88 1.90 1.90 1.90 1.95 1.97 1.97 2.00 2.03 2.25 2.27 2.33 2.50 2.72
 [31] 2.93 2.93 3.10 3.20 3.33 3.43 3.43 3.50 3.50 3.50 3.52 3.58 3.58 3.67 3.68
 [46] 3.70 3.70 3.72 3.73 3.73 3.75 3.77 3.80 3.80 3.83 3.87 3.92 3.92 3.93 3.95
 [61] 4.00 4.00 4.00 4.00 4.00 4.03 4.03 4.05 4.07 4.08 4.08 4.10 4.10 4.12 4.13
 [76] 4.13 4.18 4.20 4.25 4.25 4.25 4.28 4.33 4.33 4.35 4.37 4.40 4.42 4.43 4.50
 [91] 4.50 4.50 4.53 4.57 4.58 4.58 4.60 4.60 4.62 4.62 4.63 4.63 4.65 4.70 4.73
[106] 4.83 4.93
> # 원자료에 대한 평균과 분산
> mean(eruption.length)
[1] 3.459907
> var(eruption.length)
[1] 1.082214
> class1=seq(1.35,5.35,by=0.5)
> class1
[1] 1.35 1.85 2.35 2.85 3.35 3.85 4.35 4.85 5.35
> class2=seq(1.5,5.0,by=0.5)
> class2
[1] 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> # 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.35로 하는 도수분포표
> cat.class1=cut(eruption.length,breaks=class1)
> t1=table(cat.class1)
> t1
cat.class1
(1.35,1.85] (1.85,2.35] (2.35,2.85] (2.85,3.35] (3.35,3.85] (3.85,4.35]
          16          12           2           5          20          30
(4.35,4.85] (4.85,5.35]
          21           1
> # 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.50으로 하는 도수분포표
> cat.class2=cut(eruption.length,breaks=class2)
> t2=table(cat.class2)
> t2
cat.class2
(1.5,2] (2,2.5] (2.5,3] (3,3.5] (3.5,4] (4,4.5] (4.5,5]
   24      5       3       8      25      27      15
> # 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.35로 하는
> # 도수분포표(히스토그램)를 이용하여 구하는 평균과 분산
> m1=c(1.6,2.1,2.6,3.1,3.6,4.1,4.6,5.1)
```

```
> f1=c(16,12,2,5,20,30,21,1)
> mean1=sum(m1*f1)/sum(f1)
> mean1
[1] 3.441121
> var1=sum((m1-mean1)^2*f1)/sum(f1)
> var1
[1] 1.142982
> # 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.50으로 하는
> # 도수분포표(히스토그램)를 이용하여 구하는 평균과 분산
> m2=c(1.75,2.25,2.75,3.25,3.75,4.25,4.75)
> f2=c(24,5,3,8,25,27,15)
> mean2=sum(m2*f2)/sum(f2)
> mean2
[1] 3.432243
> var2=sum((m2-mean2)^2*f2)/sum(f2)
> var2
[1] 1.151367
```

원자료에 대한 평균은 3.460, 분산은 1.082, 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.35로 하는 도수분포표(히스토그램)를 이용하여 구한 평균은 3.441, 분산은 1.143이었고 계급의 폭을 0.5로 하고 제 1계급의 하한치를 1.50으로 하는 도수분포표(히스토그램)를 이용하여 구한 평균은 3.432, 분산은 1.151이었다. 두 개의 히스토그램과 원자료가 모두 거의 같은 평균과 분산을 나타내었다.

(d) 제 1계급의 하한치를 1.50으로 하고 계급의 폭이 0.1, 0.2, 0.3, 0.4, 0.5인 히스토그램을 이용하여 평균과 분산을 각각 구하면 다음과 같다.

```
> # 도수분포표(계급의 폭: 0.1,0.2,0.3,0.4,0.5)
> w=c(0.1,0.2,0.3,0.4,0.5)
> for(i in 1:5)
+ {
+ class1=seq(1.5,5.1,by=w[i])
+ cat.class1=cut(eruption.length,breaks=class1)
+ table(cat.class1)
+ cat("계급의 폭=", w[i], "\n")
+ print(table(cat.class1))
+ }
계급의 폭= 0.1
cat.class1
(1.5,1.6] (1.6,1.7] (1.7,1.8] (1.8,1.9]   (1.9,2]   (2,2.1] (2.1,2.2] (2.2,2.3]
       0        3        7       10        4        1        0        2
```

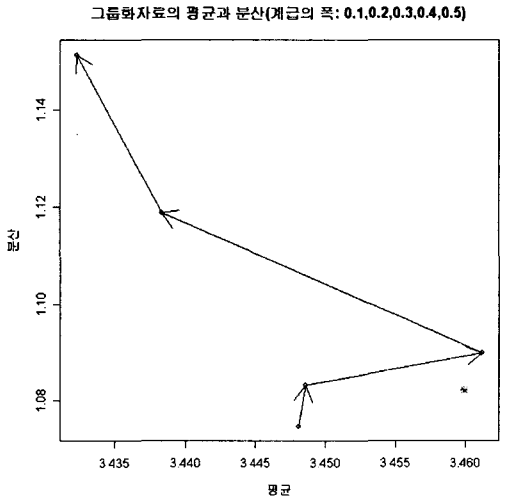
```
(2.3,2.4] (2.4,2.5] (2.5,2.6] (2.6,2.7] (2.7,2.8] (2.8,2.9]   (2.9,3]   (3,3.1]
        1         1         0         0         1         0         2         1
(3.1,3.2] (3.2,3.3] (3.3,3.4] (3.4,3.5] (3.5,3.6] (3.6,3.7] (3.7,3.8] (3.8,3.9]
        1         0         1         5         3         4         7         2
  (3.9,4]   (4,4.1] (4.1,4.2] (4.2,4.3] (4.3,4.4] (4.4,4.5] (4.5,4.6] (4.6,4.7]
        9         8         5         4         5         5         6         6
(4.7,4.8] (4.8,4.9]   (4.9,5]   (5,5.1]
        1         1         1         0
계급의 폭= 0.2
cat.class1
(1.5,1.7] (1.7,1.9] (1.9,2.1] (2.1,2.3] (2.3,2.5] (2.5,2.7] (2.7,2.9] (2.9,3.1]
        3        17         5         2         2         0         1         3
(3.1,3.3] (3.3,3.5] (3.5,3.7] (3.7,3.9] (3.9,4.1] (4.1,4.3] (4.3,4.5] (4.5,4.7]
        1         6         7         9        17         9        10        12
(4.7,4.9] (4.9,5.1]
        2         1
계급의 폭= 0.3
cat.class1
(1.5,1.8] (1.8,2.1] (2.1,2.4] (2.4,2.7]   (2.7,3]   (3,3.3] (3.3,3.6] (3.6,3.9]
       10        15         3         1         3         2         9        13
(3.9,4.2] (4.2,4.5] (4.5,4.8] (4.8,5.1]
       21        15        13         2
계급의 폭= 0.4
cat.class1
(1.5,1.9] (1.9,2.3] (2.3,2.7] (2.7,3.1] (3.1,3.5] (3.5,3.9] (3.9,4.3] (4.3,4.7]
       20         7         2         4         7        16        26        22
(4.7,5.1]
        3
계급의 폭= 0.5
cat.class1
(1.5,2] (2,2.5] (2.5,3] (3,3.5] (3.5,4] (4,4.5] (4.5,5]
     24       5       3       8      25      27      15
> # 그룹화자료의 평균과 분산(계급의 폭: 0.1,0.2,0.3,0.4,0.5)
> mean1=c(rep(0,5))
> var1=c(rep(0,5))
> m=list(c(1.55,1.65,1.75,1.85,1.95,2.05,2.15,2.25,2.35,2.45,2.55,2.65,2.75,2.85,2.95,3.05,3.15
+ ,3.25,3.35,3.45,3.55,3.65,3.75,3.85,3.95,4.05,4.15,4.25,4.35,4.45,4.55,4.65,4.75,4.85,4.95,5.05)
+ ,c(1.6,1.8,2.0,2.2,2.4,2.6,2.8,3.0,3.2,3.4,3.6,3.8,4.0,4.2,4.4,4.6,4.8,5.0)
+ ,c(1.65,1.95,2.25,2.55,2.85,3.15,3.45,3.75,4.05,4.35,4.65,4.95)
+ ,c(1.7,2.1,2.5,2.9,3.3,3.7,4.1,4.5,4.9)
```

```
+ ,c(1.75,2.25,2.75,3.25,3.75,4.25,4.75))
> f=list(c(0,3,7,10,4,1,0,2,1,1,0,0,1,0,2,1,1,0,1,5,3,4,7,2,9,8,5,4,5,5,6,6,1,1,1,0)
+ ,c(3,17,5,2,2,0,1,3,1,6,7,9,17,9,10,12,2,1)
+ ,c(10,15,3,1,3,2,9,13,21,15,13,2)
+ ,c(20,7,2,4,7,16,26,22,3)
+ ,c(24,5,3,8,25,27,15))
> for(i in 1:5)
+ {
+ cat("계급의 폭=", w[i], "\n")
+ mean1[i]=sum(m[[i]]*f[[i]])/sum(f[[i]])
+ cat("mean=", mean1[i], "\n")
+ var1[i]=sum((m[[i]]-mean1[i])^2*f[[i]])/sum(f[[i]])
+ cat("variance=", var1[i], "\n")
+ }
계급의 폭= 0.1
mean= 3.448131
variance= 1.074576
계급의 폭= 0.2
mean= 3.448598
variance= 1.083246
계급의 폭= 0.3
mean= 3.461215
variance= 1.089968
계급의 폭= 0.4
mean= 3.438318
variance= 1.118812
계급의 폭= 0.5
mean= 3.432243
variance= 1.151367
```

그룹화자료의 평균과 분산(계급의 폭: 0.1,0.2,0.3,0.4,0.5)의 변화를 그림으로 나타내면 다음과 같다.

```
> # 그룹화자료의 평균과 분산(계급의 폭: 0.1,0.2,0.3,0.4,0.5) 변화
> plot(mean1,var1,type="p",xlab="평균",ylab="분산",main="그룹화자료의 평균과 분산(계급의 폭:
0.1,0.2,0.3,0.4,0.5)")
> points(mean(eruption.length),var(eruption.length),pch="8",col="red")
> for(i in 1:4)
+ {
+ arrows(mean1[i],var1[i],mean1[i+1],var1[i+1])
+ }
```

```
> # 그룹화자료의 평균과 분산(계급의 폭: 0.1,0.2,0.3,0.4,0.5) 변화
>     plot(mean1,var1,type="p",xlab="평균",ylab="분산",main="그룹화자료의 평균과 분산(계급의 폭:
0.1,0.2,0.3,0.4,0.5)")
> points(mean(eruption.length),var(eruption.length),pch=8,col="red")
> for(i in 1:4)
+ {
+ arrows(mean1[i],var1[i],mean1[i+1],var1[i+1])
+ }
>
```



계급의 폭이 0.1에서 0.5로 변함에 따라 평균과 분산이 변하는 것을 알 수 있다. 물론 그 변화의 양은 크지는 않다.

<예제 7> (5-나, 9-나, 10-가 단계) 다음 자료는 환경부 홈페이지(www.me.go.kr)에 있는 2004년 전국시도별 대기오염물질 배출량(단위: 톤) 자료이다. 국가 대기오염배출량에서는 다양한 대기오염물질 중 환경기준 대기오염물질인 일산화탄소(CO), 질소산화물(NOx), 황산화물(SOx), 먼지농도(TSP, PM10) 및 VOC(Volatile Organic Compound : 휘발성 유기화합물) 등에 대한 배출량 자료를 제공한다. 대기 중 먼지 농도를 나타내는 통상적인 표현방법으로는 TSP, PM10, PM2.5 등이 있다. TSP(total suspended particulate)는 대기 중 부유상태에 있는 총먼지의 양이고, PM10은 직경 10㎛ 이하인 먼지의 양이며, PM2.5는 직경이 2.5㎛ 이하인 먼지의 양이다. 현재 우리나라 환경기준법에서는 TSP와 PM10에 대하여 농도 기준이 제시되어 있다.

시도명	CO	NOx	SOx	TSP	PM10	VOC
서울특별시	161,154	103,549	6,462	4,585	4,424	77,694
부산광역시	50,187	73,486	22,554	3,469	2,994	35,013
대구광역시	41,013	41,446	5,711	2,405	2,154	42,400
인천광역시	48,694	70,380	10,367	2,983	2,635	58,600
광주광역시	18,363	17,054	1,265	861	828	14,248
대전광역시	22,299	22,497	1,423	1,085	1,052	14,195
울산광역시	31,049	64,512	59,230	13,266	8,737	84,708
경기도	147,336	201,078	31,387	10,287	9,346	161,266
강원도	31,842	73,605	21,564	5,865	3,749	19,236
충청북도	35,832	52,147	13,265	5,442	3,333	26,759
충청남도	44,087	234,958	62,936	5,784	4,568	50,754
전라북도	32,547	49,099	12,633	2,906	2,513	28,797
전라남도	37,842	98,485	58,024	7,858	5,656	62,523
경상북도	55,696	80,348	33,791	7,436	5,582	46,387
경상남도	49,801	182,292	103,959	5,301	4,433	69,708
제주도	9,213	12,589	2,233	549	486	4,949

(a) 16개의 휘발성 유기화합물(VOC) 자료에 대하여 중심위치(대표값)와 퍼진 정도(산포도)를 나타내는 수치적 측도들을 구하고 줄기와 잎 그림, 히스토그램, 상자그림을 그리고 자료의 분포 특징에 대하여 서술하라.

(b) 16개의 휘발성 유기화합물(VOC) 자료 중 특이값인 경기도 자료를 제외한 나머지 15개 자료에 대하여 (a)와 같은 작업을 행하여라.

(c) PM10과 VOC 사이의 상관도(산점도)를 그려보아라. 무엇을 알 수 있나?

(d) CO와 NOx 사이의 상관도(산점도)를 그려보아라. 무엇을 알 수 있나?

(풀이) (a) 대기오염물질 배출량자료가 'c:/work/오염물질배출량.txt'에 위의 자료 형태로 저장되어 있다고 하자. 그러면 우리는 다음과 같이 read.table 함수를 이용하여 작업공간에 자료를 입력할 수 있다.

```
> # 데이터프레임 읽기
> pollutant=read.table(file="c:\\work\\오염물질배출량.txt",header=T)
> pollutant
      시도명    CO   NOx    SOx  TSP PM10  VOC
1  서울특별시 161154 103549   6462 4585 4424 77694
2  부산광역시  50187  73486  22554 3469 2994 35013
3  대구광역시  41013  41446   5711 2405 2154 42400
4  인천광역시  48694  70380  10367 2983 2635 58600
5  광주광역시  18363  17054   1265  861  828 14248
6  대전광역시  22299  22497   1423 1085 1052 14195
```

```
7  울산광역시  31049  64512  59230 13266  8737  84708
8       경기도 147336 201078  31387 10287  9346 161266
9       강원도  31842  73605  21564  5865  3749  19236
10   충청북도  35832  52147  13265  5442  3333  26759
11   충청남도  44087 234958  62936  5784  4568  50754
12   전라북도  32547  49099  12633  2906  2513  28797
13   전라남도  37842  98485  58024  7858  5656  62523
14   경상북도  55696  80348  33791  7436  5582  46387
15   경상남도  49801 182292 103959  5301  4433  69708
16      제주도   9213  12589   2233   549   486   4949
> # 대기오염물질배출량 자료
> CO=pollutant[,2]
> NOx=pollutant[,3]
> SOx=pollutant[,4]
> TSP=pollutant[,5]
> PM10=pollutant[,6]
> VOC=pollutant[,7]
```

16개의 VOC 자료에 대하여 중심위치와 퍼진 정도를 나타내는 수치적 측도들을 구하면 다음과 같다. 특이값 때문에 산술평균과 중앙값이 차이가 나고 분산이 매우 크다.

```
> # VOC 자료
> VOC
 [1]  77694  35013  42400  58600  14248  14195  84708 161266  19236  26759
[11]  50754  28797  62523  46387  69708   4949
> # 정렬(오름차순)
> sort(VOC)
 [1]   4949  14195  14248  19236  26759  28797  35013  42400  46387  50754
[11]  58600  62523  69708  77694  84708 161266
> # 중심위치와 퍼진 정도를 나타내는 수치적 측도들을 구하기
> # 산술평균
> mean(VOC)
[1] 49827.31
> # 중앙값
> median(VOC)
[1] 44393.5
> # 분산
> var(VOC)
[1] 1452824192
> # 표준편차
```

```
> sd(VOC)
[1] 38115.93
> # 범위(R에서는 범위를 호출하면 최소값과 최대값을 출력함.)
> range(VOC)
[1]  4949 161266
> # 수치적 측도 요약(다섯숫자 요약(최소값, 제 1사분위수, 중앙값, 제 3사분위수, 최대값)+산술평균)
> summary(VOC)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4949   24880   44390   49830   64320  161300
> # 사분위수간범위=제 3사분위수-제 1사분위수
> IQR(VOC)
[1] 39441
```

16개의 VOC 자료에 대하여 줄기와 잎 그림을 그리면 다음과 같다. 경기도 자료값이 특이값임을 확연히 알 수 있다.

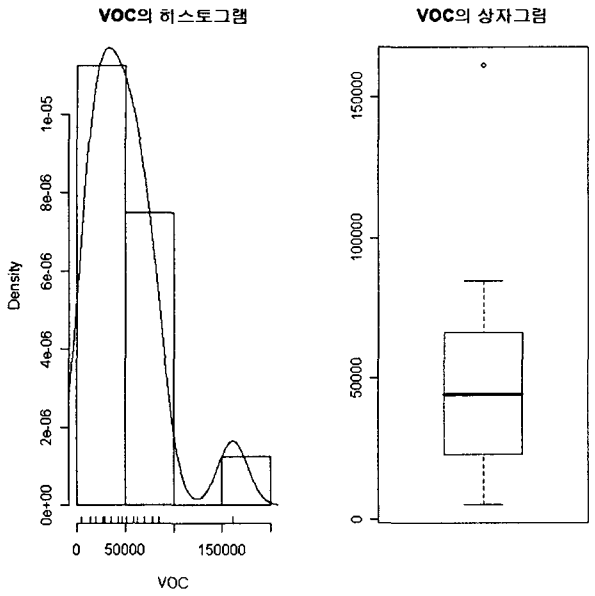
```
> # 줄기와 잎 그림
> stem(VOC)

  The decimal point is 5 digit(s) to the right of the |

  0 | 01123344
  0 | 5566788
  1 |
  1 | 6
```

16개의 VOC 자료에 대하여 상대도수밀도히스토그램(각각의 자료값과 밀도함수추정량이 같이 그려진)과 상자그림을 그리면 다음과 같다. 경기도 자료값이 특이값임을 확연히 알 수 있다. 종합하면 경기도 자료 때문에 오른쪽으로 매우 치우친 분포(비대칭분포)가 되었다.

```
> # 히스토그램
> par(mfrow=c(1,2))
> hist(VOC, prob=T, main="VOC의 히스토그램")
> lines(density(VOC))
> rug(VOC)
> # 상자그림
> boxplot(VOC, main="VOC의 히스토그램")
```



(b) 경기도 자료를 제외한 나머지 15개 자료에 대하여 중심위치와 퍼진 정도를 나타내는 수치적 측도들을 구하면 다음과 같다. 16개 자료에 비하여 평균, 분산, 표준편차가 많이 작아졌음을 알 수 있다.

```
> # 경기도자료를 제외한 자료에 대하여 중심위치와 퍼진 정도를 나타내는 수치적 측도들을 구하기
> VOC.sort=sort(VOC)
> VOC.sort
 [1]   4949  14195  14248  19236  26759  28797  35013  42400  46387  50754
[11]  58600  62523  69708  77694  84708 161266
> VOC1=VOC.sort[-16]
> VOC1
 [1]  4949 14195 14248 19236 26759 28797 35013 42400 46387 50754 58600 62523
[13] 69708 77694 84708
> # 산술평균
> mean(VOC1)
[1] 42398.07
> # 중앙값
> median(VOC1)
[1] 42400
> # 분산
> var(VOC1)
[1] 610419743
> # 표준편차
```

```
> sd(VOC1)
[1] 24706.67
> # 범위(R에서는 범위를 호출하면 최소값과 최대값을 출력함.)
> range(VOC1)
[1]  4949 84708
> # 수치적 측도 요약(다섯숫자 요약(최소값, 제 1사분위수, 중앙값, 제 3사분위수, 최대값)+산술평균)
> summary(VOC1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4949   23000   42400   42400   60560   84710
> # 사분위수간범위=제 3사분위수-제 1사분위수
> IQR(VOC1)
[1] 37564
```

경기도 자료를 제외한 나머지 15개 자료에 대하여 그린 줄기와 잎 그림은 다음과 같다. 특이값이 포함되어 있을 때의 줄기와 잎 그림에 비하여 비대칭 정도가 상당히 완화되었음을 알 수 있다.

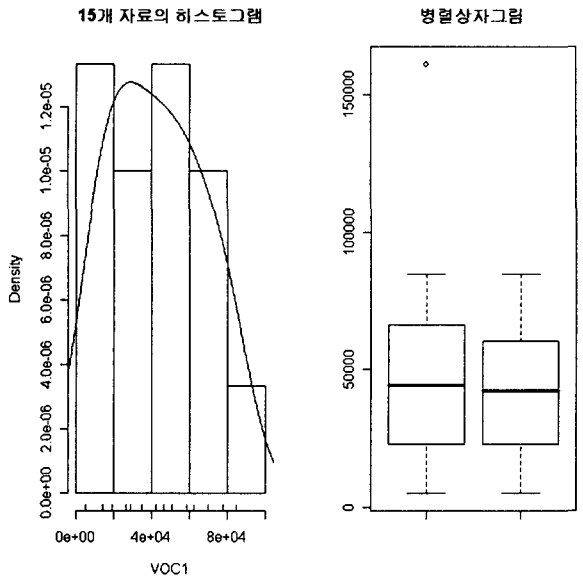
```
> # 경기도자료를 제외한 자료에 대한 줄기와 잎 그림
> stem(VOC1)

  The decimal point is 4 digit(s) to the right of the |

  0 | 5449
  2 | 795
  4 | 2619
  6 | 308
  8 | 5
```

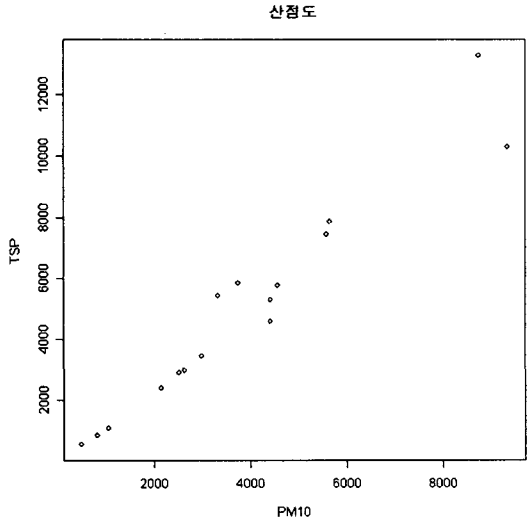
경기도 자료를 제외한 나머지 15개 자료에 대하여 그린 상대도수밀도히스토그램(각각의 자료값과 밀도함수추정량이 같이 그려진)과 병렬상자그림을 그리면 다음과 같다. 16개 자료에 비하여 비대칭성이 많이 완화되었음을 알 수 있다. 이처럼 특이값은 분포의 모양을 결정하는 데 큰 영향을 준다. 병렬상자그림에서 왼쪽이 16개 자료에 대한 상자그림이고 오른 쪽이 15개 자료에 대한 상자그림이다. 특이값(경기도자료) 하나를 제외하고는 두 개의 상자가 거의 차이가 없음을 알 수 있다.

```
> # 경기도자료를 제외한 자료에 대한 히스토그램
> par(mfrow=c(1,2))
> hist(VOC1, prob=T, main="15개 자료의 히스토그램")
> lines(density(VOC1))
> rug(VOC1)
> # 상자그림
> boxplot(VOC, VOC1, main="병렬상자그림")
>
```



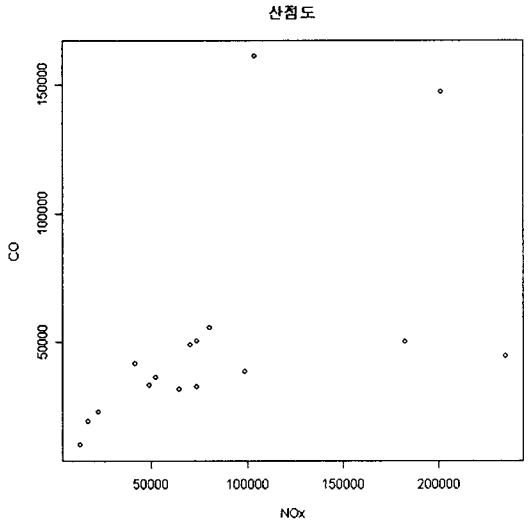
(c) PM10과 TSP 사이의 상관도(산점도)를 그려보면 다음과 같다. PM10과 TSP 사이에는 강한 양의 상관관계가 있음을 알 수 있다. 즉 PM10과 TSP 사이에는 기울기가 양수인 직선관계가 강하게 있음을 알 수 있다.

```
> # 산점도 1
> par(mfrow=c(1,1))
> plot(PM10, TSP, main="산점도")
```



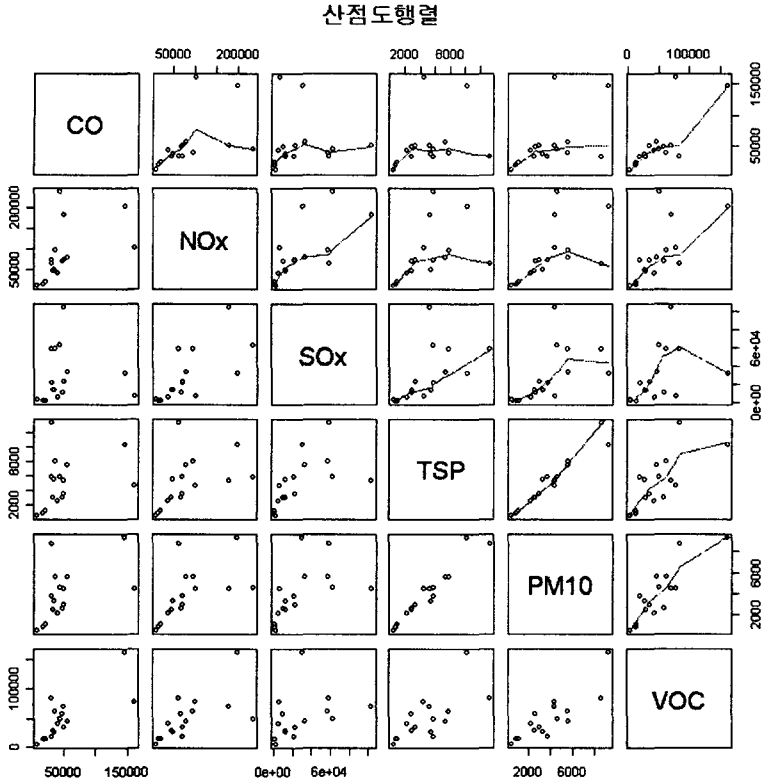
(d) NOx와 CO 사이의 상관도(산점도)를 그려보면 다음과 같다. CO 값이 큰 두 개의 특이값을 제거하고 생각하면 NOx가 100,000톤까지 커지면서 CO 값도 직선적으로 커지나 100,000톤을 넘으면 CO 값이 거의 변화가 없는 패턴을 이룬다. 그러나 CO 값이 큰 두 개의 특이값을 그대로 두고 보면 상관관계가 있는 지 알기가 쉽지 않다. 이처럼 특이값은 두 변수 사이의 상관관계에 큰 영향을 준다.

```
> # 산점도 2
> plot(NOx, CO, main="산점도")
```



참고로 6개의 변수를 상대로 각 두 개씩의 산점도를 행렬의 행과 열에 배당하는 산점도행렬(scatterplot matrix)를 그려보면 다음과 같다. 대각선 위 쪽 패널들에 있는 곡선은 평활모수 α=0.75인 LOWESS이다. 이 회귀선을 통하여 두 변수 사이의 관계를 한 눈에 알 수 있다.

```
> # 산점도행렬
> air.pollutant=cbind(CO,NOx,SOx,TSP,PM10,VOC)
> pairs(air.pollutant, upper.panel=panel.smooth, main="산점도행렬")
```



각 변수의 분산과 공분산을 나타내는 분산-공분산행렬(variance-covariance matrix)은 다음과 같다. 대각선 값은 각 변수의 분산이고 비대각선 값은 해당되는 두 변수 사이의 공분산이다.

```
> # 분산-공분산행렬
> cov(air.pollutant)
```

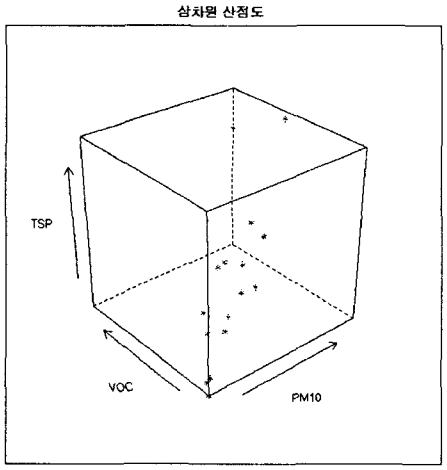
	CO	NOx	SOx	TSP	PM10	VOC
CO	1782355351	1411686171	16031109	49824506	56548999	1223005196
NOx	1411686171	4325553562	1340404842	111319330	101308604	1674652816
SOx	16031109	1340404842	852863622	59093578	41806581	451761208
TSP	49824506	111319330	59093578	12133540	8560741	94580415
PM10	56548999	101308604	41806581	8560741	6460923	82939539
VOC	1223005196	1674652816	451761208	94580415	82939539	1452824192

변수들 사이의 상관계수를 나타내는 상관계수행렬(correlation coefficient matrix)을 구하면 다음과 같다. TSP, PM10, VOC 사이에는 서로 양의 상관관계가 강함을 알 수 있다.

```
> # 상관계수행렬
> cor(air.pollutant)
           CO       NOx       SOx       TSP      PM10       VOC
CO   1.00000000 0.5084171 0.01300249 0.3388068 0.5269634 0.7600193
NOx  0.50841712 1.0000000 0.69787131 0.4859102 0.6060078 0.6680322
SOx  0.01300249 0.6978713 1.00000000 0.5809072 0.5631934 0.4058473
TSP  0.33880678 0.4859102 0.58090717 1.0000000 0.9668750 0.7123623
PM10 0.52696344 0.6060078 0.56319341 0.9668750 1.0000000 0.8560671
VOC  0.76001930 0.6680322 0.40584725 0.7123623 0.8560671 1.0000000
```

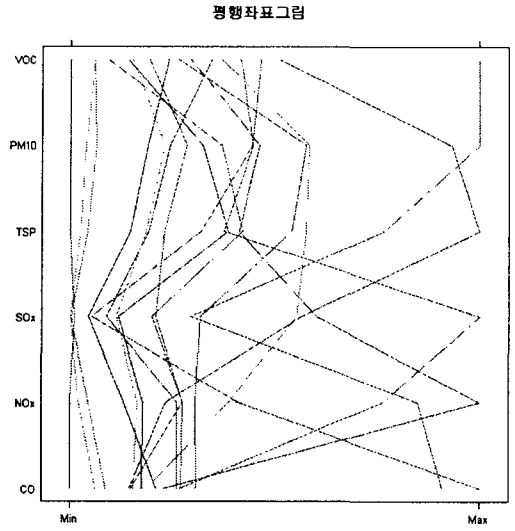
TSP, PM10, VOC 세 변수에 대한 삼차원 산점도를 그리면 다음과 같다. 대략 삼차원 직선 띠 패턴을 이루고 있음을 알 수 있다.

```
> # 삼차원 산점도
> library(lattice)
> cloud(air.pollutant[,4]~air.pollutant[,5]*air.pollutant[,6]
+ , xlab="PM10", ylab="VOC", zlab="TSP", main="삼차원 산점도")
```



대기오염물질 배출량자료에 대한 평행좌표그림을 그리면 다음과 같다. 대략 5개의 특이값이 있음을 알 수 있다.

```
> # 평행좌표그림
> parallel(~air.pollutant, main="평행좌표그림")
```



<예제 8> (6-나, 8-나 단계) R에서의 난수를 이용하여 동전의 앞면이 나오는 확률을 구하여 보아라.

(풀이) 6-나 단계 수학교과서에 통계적 확률과 수학적 확률이 모두 나타나고 있다. 여기에 두 가지 문제점이 있다. 첫 번째 문제점은 수학적 확률의 정의를 '모든 경우의 수에 대한 어떤 사건이 일어날 경우의 수의 비율'이라고 정의하였다. 이 정의에서 문제가 되는 것은 각각의 경우가 일어날 가능성이 일정하다는 조건을 달지 않았다는 것이다. 두 번째 문제점은 통계적 확률의 예로 100원짜리 동전을 30번 던지는 시행에서 그림면이 나온 경우의 수의 비율, 100원짜리 동전을 40번 던지는 시행에서 숫자면이 나오는 경우의 수의 비율, 주사위를 20번 던지는 시행에서 홀수의 눈이 나오는 경우의 수의 비율을 구하는 문제 총 3 가지 예가 나오는 데 이 세 가지 경우 모두 시행횟수가 너무 적다는 것이다. 시행횟수가 이렇게 적으면 우리가 예상하는 확률 0.5가 거의 나오지 않는다. 8-나 단계 수학교과서 16종에서 언급한 통계적 확률의 예와 시행횟수는 다음 <표 3>과 같다. 수학적 확률로 구할 수 없는 예로서 병뚜껑을 던지는 시행에서 겉면이 나올 확률과 압정을 던지는 시행에서 침이 아래로 향하는 확률, 윷가락 한 개를 던지는 시행에서 안면이 나오는(평평한 면이 위로 향할) 확률이 있다.

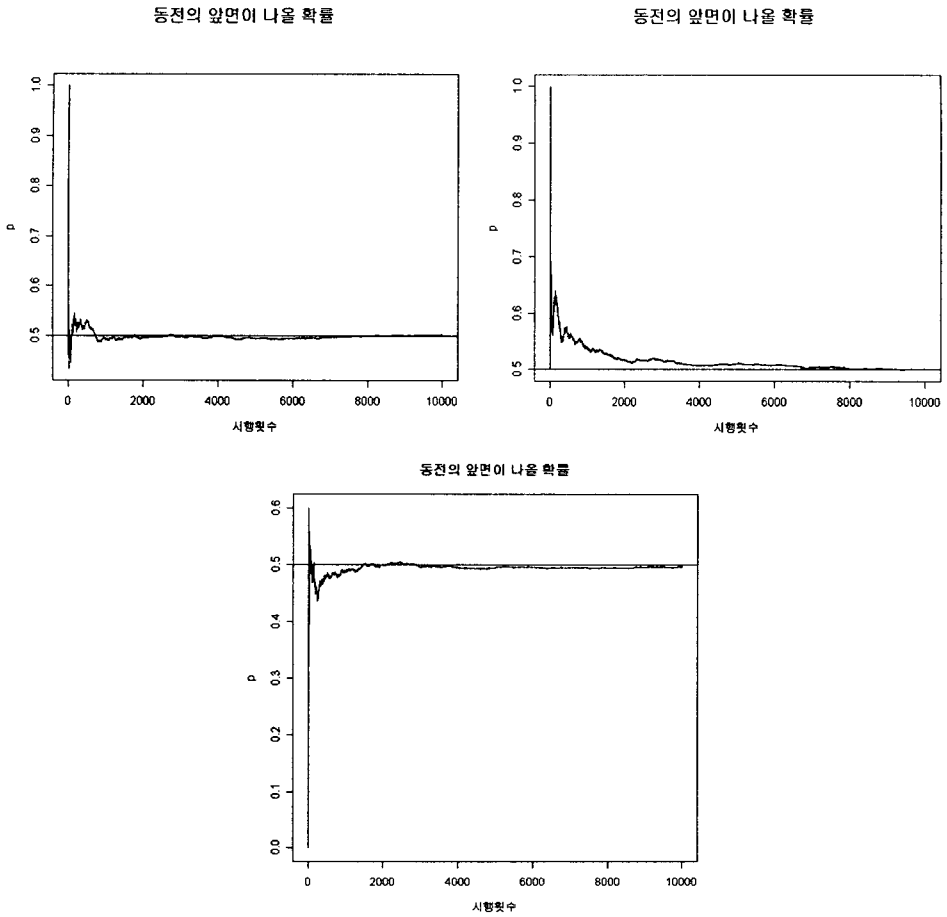
<표 3> 8-나 단계 수학교과서 16종에서 언급한 통계적 확률의 예

단계	출판사	통계적 확률의 예와 시행횟수
8-나	A	동전(600), 태어난 아이의 남녀비(3,000)
	B	동전(1500)
	C	동전(100->횟수를 늘임), 주사위(2,000), 병뚜껑(500), 컴퓨터 모의실험
	D	동전(800)
	E	주사위(1명 30번->반 전체), 동전(1,000), 옻짝 한 개(20,000)
	F	동전(400), 주사위(500)
	G	주사위(1,000)
	H	동전(2,000)
	I	동전(1,000)
	J	구슬주머니(흰색 4, 빨강 3, 녹색 2, 파랑 1)(50), 옻가락 한 개(900)
	K	동전(500->1,000)
	L	동전(1,000), 주사위(1,000)
	M	듀퐁의 바늘실험(100), 압정(1,000), 병마개(1,000), 동전(1,000), 주사위(1,000)
	N	동전(100), 컴퓨터 모의실험
	O	주사위(1,200)
	P	동전(1,000), 주사위(2,000), 병뚜껑(1,000)

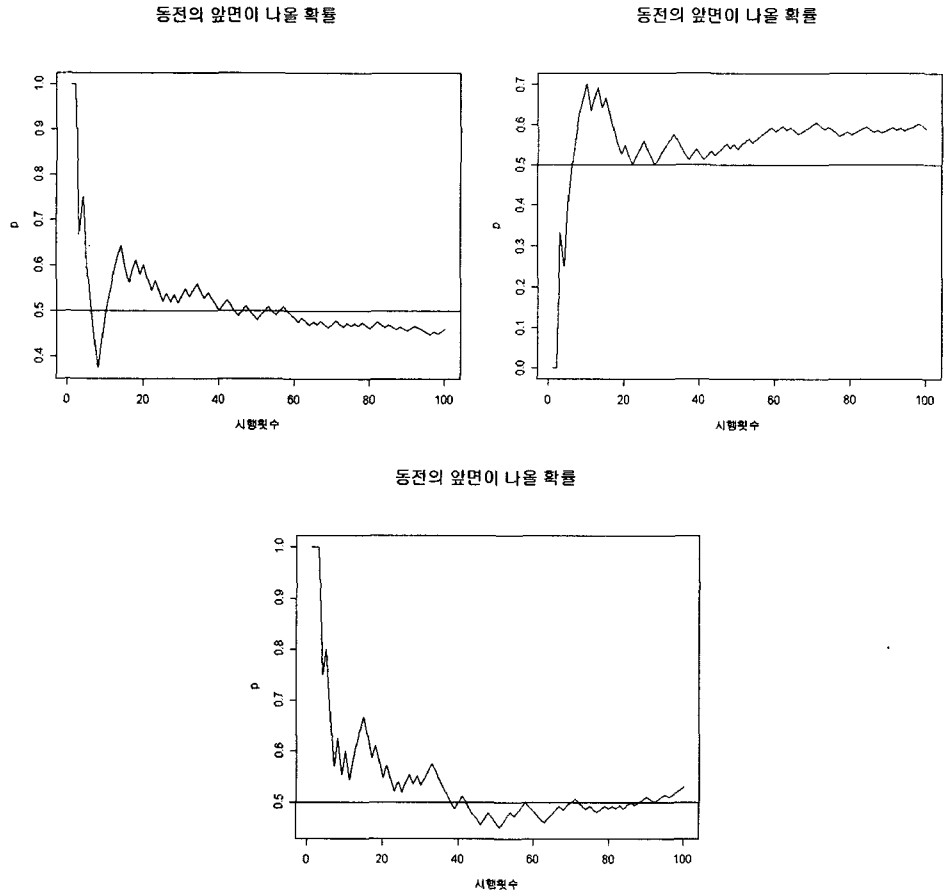
동전의 앞면이 나오는 확률을 구하기 위한 R 프로그램은 다음과 같다.(시행횟수를 10,000번으로 하였으나 조정이 가능함.)

```
> law.large.number=
+ function(n)
+ {
+ par(mfrow=c(1,1))
+ # 베르누이분포에서의 난수
+ x1=rbinom(n,1,0.5)
+ # 상대도수
+ xbar1=cumsum(x1)/1:n
+ plot(xbar1,ylab="p",xlab="시행횟수",main="동전의 앞면이 나올 확률",type="l")
+ abline(0.5, 0)
+ }
> law.large.number(10000)
>
```

동전을 10,000번 던져 동전의 앞면이 나오는 확률을 세 차례 구하여보니 다음과 같은 세 가지 그림이 나왔다. 우리는 세 가지 그림에서 각각 0.5에 수렴하는 것을 볼 수 있다. 그러나 수렴하는 패턴은 서로 다름을 알 수 있다.



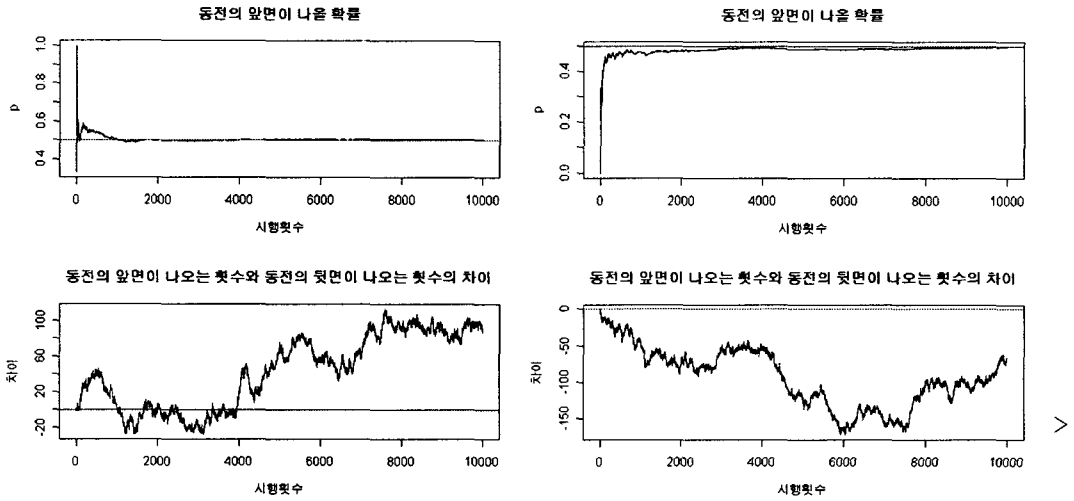
반면에 동전을 100번 던져 동전의 앞면이 나오는 확률을 세 차례 구하여보니 다음과 같은 세 가지 그림이 나왔다. 우리는 세 가지 그림을 종합하여 보면 0.5에 수렴한다고 보기가 어렵다. 즉 시행횟수가 적으면 통계적 확률을 구하기가 어렵게 된다는 것을 알 수 있다. 그러므로 통계적 확률을 학생들에게 보이기 위해서는 컴퓨터를 이용한 통계적 시뮬레이션을 시행횟수를 크게 하여 제시할 필요가 있다.



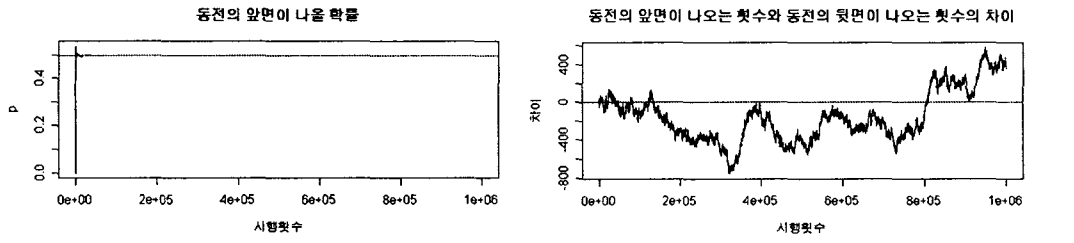
통계적 확률을 구할 때 주의할 또 한 가지 사항은 시행횟수가 커짐에 따라 동전의 앞면이 나오는 상대도수의 극한값이 0.5가 된다고 해서 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이가 0에 가까워진다는 것이 아니다. 시행횟수가 커짐에 따라 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이는 커졌다 작아졌다 한다. 즉 이 차이는 랜덤하게 된다. 이를 통계적 시뮬레이션으로 확인하여 보자.

```
> law.large.number2=
+ function(n)
+ {
+ par(mfrow=c(2,1))
+ # 베르누이분포에서의 난수
+ x1=rbinom(n,1,0.5)
+ # 상대도수
```

```
+ xbar1=cumsum(x1)/1:n
+ plot(xbar1,ylab="p",xlab="시행횟수",main="동전의 앞면이 나올 확률",type="l")
+ abline(0.5, 0, col="red")
+ # 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이
+ H=cumsum(x1)
+ T=1:n-H
+ difference=H-T
+ plot(difference,ylab="차이",xlab="시행횟수",
+ main="동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이",type="l")
+ abline(0, 0, col="red")
+ par(mfrow=c(1,1))
+ }
> law.large.number2(10000)
```



다음 그림은 시행횟수를 백만 번으로 했을 때 시행횟수가 커짐에 따라 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이를 나타내는 그림이다. 시행횟수가 커짐에 따라 동전의 앞면이 나오는 횟수와 동전의 뒷면이 나오는 횟수의 차이는 커졌다 작아졌다 하나 그 변동 폭이 시행횟수가 만 번일 때보다 큼을 알 수 있다.

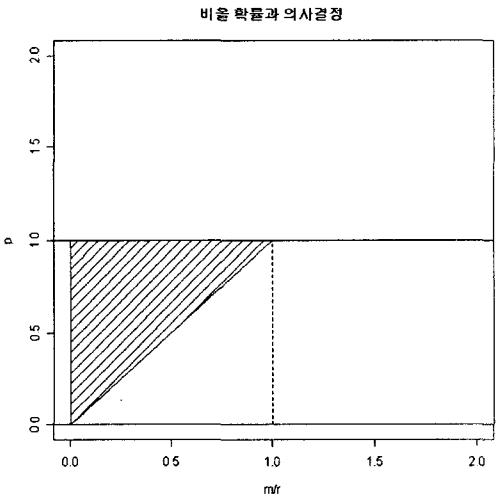


<예제 9> (6-나, 8-나 단계) 2-나 교과서 6단원 '표와 그래프' 중 '실생활에 적용하여 봅시다'(100p)에서나 8-나 교과서들을 보면 '비 올 확률'이 나온다. 이러한 교과서들은 비올 확률이 의미하는 바를 설명하고는 있으나 이 비올 확률이 통계학에서 이야기하고 있는 '불확실성(uncertainty)의 모형화'와 어떤 관계가 있는지에 대한 설명이 부족하다. Rao(2003)는 통계학을 정의하며 '불확실성(uncertainty)의 모형화', '불확실성 길들이기'라는 표현들을 사용하고 있다. 불확실성을 내포한 지식에 불확실성의 계량화를 시행하면 사용가능한 지식이 된다. Rao(2003)의 설명을 다시 음미하여 보자. 비올 확률이 50%이면 집을 나설 때 우산을 가지고 가야 하나 말아야 하나? 이를 해결하기 위해서 우산을 가지고 외출해서 겪게 되는 불편이 m원이고 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실을 r원이라 하자. 의사결정과 예상손실을 정리하면 다음 표와 같다.

의사결정	예상손실
우산을 가져간다.	m
우산을 가져가지 않는다.	$p \times r + (1-p) \times 0 = pr$

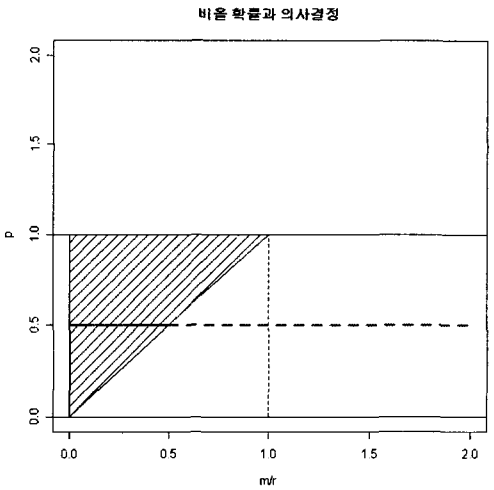
우리는 항상 손실을 최소화하는 방향으로 의사결정을 한다. 만일 $m \leq pr$이면 즉 $\frac{m}{r} \leq p$면 우산을 가져가고 $m > pr$이면 즉 $\frac{m}{r} > p$면 우산을 가져가지 않는다. 이 두 가지 경우를 R을 이용하여 그림으로 나타내어 보자. 이 그림에서 빗금친 삼각형 부분이 '우산을 가져간다'이고 흰색부분이 '우산을 가져가지 않는다'이다.

```
> # 비올 확률과 의사결정
> x=c(0,0,1)
> y=c(0,1,1)
> plot(0:2,0:2,type="n",main="비올 확률과 의사결정
",xlab="m/r",ylab="p")
> polygon(x,y,density=10)
> abline(h=1)
> abline(h=0)
> lines(c(1,1),c(0,1),lty=2)
>
```



극단적인 경우인 $\frac{m}{r} > 1$이면 즉 $m > r$이면 항상 우산을 가져가지 않는다. 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실보다 우산을 가지고 외출해서 겪게 되는 불편을 더 크게 생각하는 사람에게는 비올 확률이 크든 작든 항상 우산을 가져가지 않으므로 비올 확률이 유용한 정보가 되지 못한다. 그러나 보통의 사람들에게는 우산을 가지고 외출해서 겪게 되는 불편보다 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실이 더 크다. 앞의 질문인 '비올 확률이 50%이면 집을 나설 때 우산을 가지고 가야 하나 말아야 하나?'에 대하여 다시 생각하여 보자. $0 < \frac{m}{r} < \frac{1}{2}$라고 여기는 사람, 즉 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실이 우산을 가지고 외출해서 겪게 되는 불편의 2배 이상이 된다고 여기는 사람은 우산을 가지고 가게 된다. 이것을 다음의 그림에서 빨간 직선으로 표시하였다. $\frac{1}{2} < \frac{m}{r} < 1$라고 여기는 사람, 즉 우산을 두고 외출해서 옷이 비에 젖게 됨으로써 받는 손실이 우산을 가지고 외출해서 겪게 되는 불편의 2배 이하가 된다고 여기는 사람은 우산을 가져가지 않게 된다. 이것을 다음 그림에서 빨간 점선으로 표시하였다. 이처럼 비올 확률이라는 정보가 우리들이 합리적인 의사 결정을 하는 데 중요한 역할을 하게 된다. 불확실성을 내포한 지식에 불확실성의 계량화를 시행함으로써 사용가능한 지식으로 바꿀 수가 있게 되는 것이다.

```
> # 비올 확률: 0.5
> x=c(0,0,1)
> y=c(0,1,1)
> plot(0:2,0:2,type="n",main="비올 확률과 의사결정
",xlab="m/r",ylab="p")
> polygon(x,y,density=10)
> abline(h=1)
> abline(h=0)
> lines(c(1,1),c(0,1),lty=2)
> lines(c(0,0.5),c(0.5,0.5),col="red",lwd=3)
> lines(c(0.5,2),c(0.5,0.5),col="red",lwd=3,lty=2)
>
```



<예제 10> (6-나, 8-나 단계) 한 방에 모여 있는 사람의 수가 N이라 하자. 한 방의 사람들 중 서로 같은 생일을 갖게 될 확률을 구하여라(1년은 365일로 가정함.).

(풀이) 우선 P[N명이 같은 생일을 갖지 않는다]를 구하여보자. 첫 번째 사람이 365일 중 하루를 선택할 확률은 $\frac{365}{365}$, 첫 번째 사람이 365일 중 하루를 선택하면 두 번째 사람은 나머지 364일 중 하

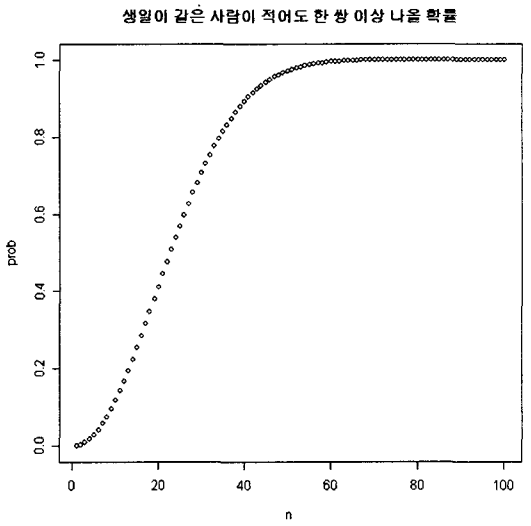
나를 선택하여야 하므로 두 번째 사람이 하루를 선택할 확률은 $\frac{364}{365}$, 이런 식으로 생각하면 P[N명이 같은 생일을 갖지 않는다]= $\frac{365\times364\times363\times\cdots\times(365-N+1)}{365^N}$이 된다. 그러므로 우리가 구하여야 할 확률은 여사건의 확률공식을 이용하여 P[N명 중 최소한 두 명이 같은 생일을 갖는다]=1-P[N명이 같은 생일을 갖지 않는다]= $1-\frac{365\times364\times363\times\cdots\times(365-N+1)}{365^N}$이 된다.

이 확률을 표로 작성하면 다음 표와 같다.

N	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75
P	0.027	0.117	0.253	0.411	0.569	0.706	0.814	0.891	0.941	0.970	0.986	0.994	0.998	0.999	0.9997

$N(N=1, 2, \cdots, 100)$에 따른 확률 P[N명 중 최소한 두 명이 같은 생일을 갖는다]의 변화를 그림으로 그리면 다음과 같다.

```
> # 생일이 같은 사람이 적어도 한 쌍 이상 나올 확률
> pr=c(rep(0,101))
> prob=c(rep(0,100))
> n=1:100
> pr[1]=1
> for (i in 1:100)
+ {
+ pr[i+1]=pr[i]*(365-i)/365
+ prob[i]=1-pr[i]
+ }
> plot(n,prob,type="p",main="생일이 같은 사람이 적어도 한 쌍 이상 나올 확률")
>
```



한 방에 23명이 모여 있으면 최소한 두 명이 같은 생일을 갖게 될 확률이 0.507로 0.5를 넘게 되고 한 방에 40명이 모여 있으면 최소한 두 명이 같은 생일을 갖게 될 확률이 0.891로 매우 큼을 알 수 있다.

Ⅲ. 결 론

우리는 제 7차 수학과 교육과정 내의 확률 및 통계영역 내용을 중심으로 한 10 가지 예제들을 통하여 R 패키지를 구체적으로 수업에 어떻게 적용할 수 있는 지를 살펴보았다. 물론 다른 예제들을 통하여서도 R을 적용하여 볼 수 있다. 예제에서 살펴 본 것처럼 우리는 R을 이용하면 기술통계에 필요한 다양한 수치적 측도들을 쉽게 구할 수 있고 다양한 그림들을 어렵지 않게 그릴 수가 있다.

참 고 문 헌

강옥기 · 정순영 · 이환철 (2003). 중학교 7, 8, 9단계 수학, (주)두산.

강행고 · 양윤택 · 설명환 · 이종배 · 변임수 · 정호집 · 김경현 (2003). 고등학교 10단계 수학, 동화사.

강행고 · 이화영 · 박진석 · 이용완 · 한경연 · 이준홍 · 이혜련 · 송미현 · 박정숙 (2003). 중학교 7, 8, 9단계 수학, (주)중앙교육진흥연구소.

고성은 · 박복현 · 김준희 · 최수일 · 강운중 · 소순영 (2003). 중학교 7, 8, 9단계 수학, (주)블랙박스.

교육 인적 자원부 (1997). 제 7차 수학과 교육과정.

교육 인적 자원부 (2002). 1-6단계 수학교과서와 수학익힘책.

김수환 · 이강섭 · 임영훈 · 왕규채 · 송교식 · 이동수 · 강영길 (2003). 고등학교 10단계 수학, (주)지학사.

금종해 · 이만근 · 이미라 · 김영주 (2003). 중학교 7, 8, 9단계 수학, (주)고려출판.

박규홍 · 고성군 · 김성국 · 김유태 · 박재용 · 육상국 · 임창우 · 한옥동 (2003). 중학교 7, 8, 9단계 수학, 두레교육(주).

박규홍 · 임성근 · 양지청 · 김수영 (2004). 고등학교 10단계 수학, (주)교학사.

박두일 · 신동선 · 강영환 · 윤재성 · 김인종 (2003). 중학교 7, 8, 9단계 수학, (주)교학사.

박두일 · 신동선 · 김기현 · 박복현 · 안훈 · 소순영 · 송건수 · 김주석 · 이미선 (2003). 고등학교 10단계 수학, (주)교학사.

박배훈 · 김원경 · 조민식 · 김원석 · 이대현 (2003). 고등학교 10단계 수학, 법문사.

박세희 · 정광식 · 강병개 · 서정인 (2003). 고등학교 10단계 수학, 동아서적(주).

박윤범 · 박혜숙 · 권혁천 · 김홍섭 · 육인선 · 송상헌 (2004). 고등학교 10단계 수학, 대한교과서(주).

박윤범 · 박혜숙 · 권혁천 · 육인선 (2003). 중학교 7, 8, 9단계 수학, 대한교과서(주).

배종수 · 박종률 · 윤행원 · 유종광 · 김문환 · 민기열 · 박동익 · 우현철 (2003). 중학교 7, 8, 9단계 수학, 한성교육연구소.

신항균 (2003). 중학교 7, 8, 9단계 수학, 형설출판사.

신현성 · 최용준 (2004). 고등학교 10단계 수학, (주)천재교육.

양승갑 · 박영수 · 박원선 · 배종숙 · 심덕현 · 이성길 · 홍우철 (2003). 중학교 7, 8, 9단계 수학, (주)금성

출판사.

양승갑 · 배종숙 · 이성길 · 박원선 · 박영수 · 홍우철 · 신준국 · 성덕현 · 김대희 (2004). 고등학교 10단계 수학, (주)금성출판사.

우정호 · 류희찬 · 문광호 · 박경미 (2004). 고등학교 10단계 수학, 대한교과서(주).

이광복 · 김광환 · 김호영 · 이덕실 (2004). 고등학교 10단계 수학, 새한교과서(주).

이방수 · 기호삼 (2004). 고등학교 10단계 수학, (주)천재교육.

이영하 · 허민 · 박영훈 · 여태경 (2003). 중학교 7, 8, 9단계 수학, (주)교문사.

이준열 · 장훈 · 최부림 · 남호영 · 이상은 (2003). 중학교 7, 8, 9단계 수학, (주)도서출판 디딤돌.

임재훈 · 기우항 · 김진호 · 윤오영 · 반용호 · 조동석 · 남승진 · 오명성 (2004). 고등학교 10단계 수학, (주)두산.

장건수 · 안재문 · 김의석 · 정연석 · 박은주 (2003). 고등학교 10단계 수학, 지구문화사.

장대홍 (1995). 우리나라 언론매체에 나타나는 통계적 그래픽의 실태조사와 통계적, 제도적 해결방안에 대한 연구, 응용통계연구, **제 8권 제 2호**, pp.1-26.

장대홍 · 박용범 · 이혜영 (2000). A Study on Probability and Statistics Education in Middle School's Mathematics Textbooks in Korea, 한국통계학회논문집, **제 7권 제 1호**, pp.337-355.

장대홍 (2005). 1-10단계 수학교과서 확률 및 통계단원 내용상의 문제점 분석과 해결방안, 한국수학교육학회 제 35회 전국수학교육연구대회프로시딩, pp.105-120.

장대홍 (2007). 초·중·고등학교 확률 및 통계교육에서의 R 통계패키지의 활용(Ⅰ), 수학교육논문집, **제 21집 제 2호**, pp.199-225.

전평국 · 신동윤 · 방승진 · 황현모 · 정석규 (2003). 중학교 7, 8, 9단계 수학, 교학연구사.

조태근 · 임성모 · 정상권 · 이재학 · 이성재 (2003). 중학교 7, 8, 9단계 수학, (주)금성출판사.

최봉대 · 강옥기 · 황석근 · 이재돈 · 김영욱 · 전무근 · 홍진철 (2004). 고등학교 10단계 수학, (주)중앙교육진흥연구소.

최상기 · 이만근 · 이재실 · 백한미 (2003). 고등학교 10단계 수학, (주)고려출판.

최용준 (2003). 중학교 7, 8, 9단계 수학, (주)천재교육.

황석근 · 이재돈 (2003). 중학교 7, 8, 9단계 수학, 한서출판사.

Rao, C. R./이재창 옮김 (2003). 혼돈과 질서의 만남, 나남,

Applications of R statistical package on Probability and Statistics Education in Elementary, Middle and High School(II)

Jang, Dae-Heung

Division of Mathematical Sciences, Pukyong National University, Busan Korea, 608-737

E-mail: dhjang@pknu.ac.kr

We described the overall explanation about R statistical package in Jang(2007). With referring the contents of the 7th national mathematics curriculum, we suggest the plan for applications of R package on probability and statistics education in elementary, middle and high school mathematics.

* ZDM Classification : N80

* 2000 Mathematics Subject Classification : 97U70

* Key Words : the education of probability and statistics, R statistical package