응용통계연구 (2009)
22(4), 829–854

기초통계학 교육 시 확률에 관한 몇 가지 유용한 사례들

장대흥[1]

[1]부경대학교 수리과학부 통계학전공

(2009년 1월 접수, 2009년 3월 채택)

요 약

기초통계학 교육 시 확률 부분에 대한 강의는 추측통계학 영역의 시작 부분으로서 학생들이 통계학을 수강할 때 교수강의를 따라가기 어려워하기 시작하는 부분이다. 기초 통계학 확률부분 강의 시 유용하게 사용할 수 있는 사례들을 제시하고 R을 이용하여 구현하여 보았다.

주요용어: 통계교육, 확률, R 스크립트.

1. 서론

기초통계학 교육 내용은 교수들의 개성에 따라 배치가 달라질 수가 있고 내용에 대한 가감이 생길 수 있으나 대략 다음과 같은 내용으로 전개가 된다.

1. 통계학의 주제/자료의 종류
2. 자료의 요약(일변량, 이변량)
3. 확률/확률변수/확률분포
4. 표본추출분포
5. 추정/검정
6. 두 모집단 비교
7. 회귀분석

기초통계학 교육 시 확률 부분에 대한 강의는 추측통계학 영역의 시작 부분으로서 학생들이 통계학을 수강할 때 교수강의를 따라가기 어려워하기 시작하는 부분이다. 확률 부분을 어느 정도 깊이로 학생들에게 가르쳐야하는 지에 대하여는 교수들의 의견이 분분할 수가 있다. 이론적으로 접근할수록 학생들은 통계학을 점점 수학의 아류쯤으로 여기게 되는 경향이 있다. 그러므로 확률 부분에 대한 강의는 기초통계학 강의를 담당하는 교수로서는 매우 신경이 쓰이는 부분이다. 본 논문을 통하여 기초 통계학 확률부분 강의 시 유용하게 사용할 수 있는 사례들을 몇 가지 제시하고 R을 이용하여 구현하여 보고자 한다.

2. 사례들

다음 사례들은 김영일 등 (2008) 책에서 본 저자가 저술한 내용 중 확률에 관계된 사례들(일부 수정)에 R 스크립트를 첨부한 것이다.

[1](608-737) 부산광역시 남구 대연3동 599-1 부경대학교 수리과학부 통계학전공, 교수.
E-mail: dhjang@pknu.ac.kr

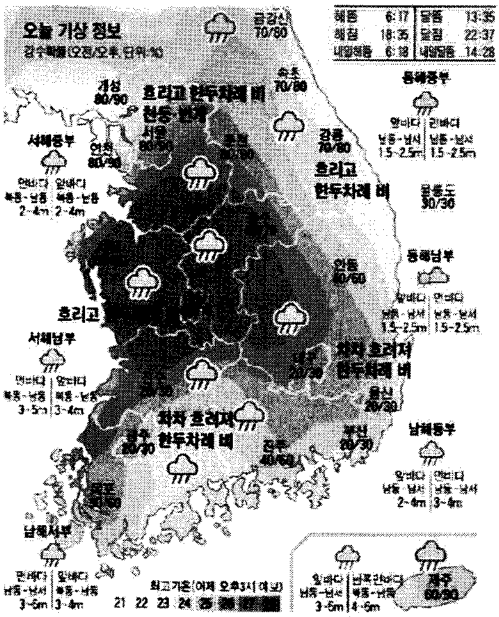

그림 2.1. 일기예보(제공처: 조선일보)

표 2.1. 의사결정과 예상손실

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

사례 1(비올 확률)

일간신문 지면을 보면 매일 '비 올 확률'이나 '눈 올 확률'이 나온다. 이 일간 신문에 매일 등장하는 '비 올 확률'이나 '눈 올 확률'을 우리는 어떻게 해석하며 이용하는가? 한 예로, 2007년 9월 태풍 '나리'로 인하여 제주도에 쏟아진 엄청난 양의 폭우로 제주도에 엄청난 재난을 가져다주었다. 기상청의 일기예보는 우리 생활에 점점 중요성을 더하고 있다. 다음 그림 2.1은 2007.09.19 조선일보에서 제공한 일기예보이다. 예로 '서울' 밑에 80/90이라 적혀 있다. 서울지역에서 비올 확률이 오전에는 80%, 오후에는 90%라는 말이다. 반면에 부산지역에서는 비올 확률이 오전에는 20%, 오후에는 30%이다. 이 숫자들(확률)의 의미는 무엇이며 우리는 이 숫자들을 어떻게 실생활에 사용하여야 하는 걸까?

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. 비올 확률이 p라 할 때 의사결정시의 예상손실을 정리하면 다음 표 2.1과 같다.

그러면 우리는 다음 표 2.2와 같이 항상 손실을 최소화하는 방향으로 의사결정을 한다.

이 두 가지 경우를 다음과 같이 그림 2.2로 나타내어 보자. 이 그림에서 빗금친 삼각형 부분이 '우산을

표 **2.2.** 의사결정

비교	의사결정
$m \leq pr \Leftrightarrow \frac{m}{r} \leq p$	우산을 가져간다.
$m > pr \Leftrightarrow \frac{m}{r} > p$	우산을 가져가지 않는다.

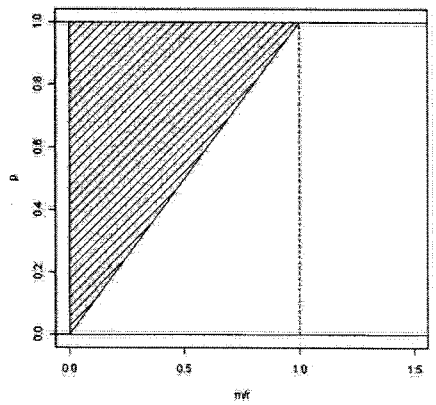

그림 **2.2.** 의사결정

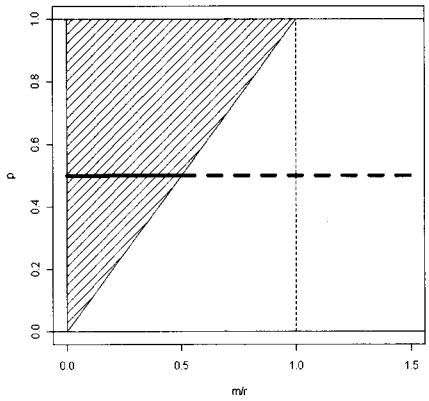

그림 **2.3.** 비올 확률 50%에 대한 해석

가져간다.'이고 흰색부분이 '우산을 가져가지 않는다.'이다. 극단적인 경우인 $m/r > 1$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. $0 < m/r < 1/2$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 그림 2.3에서 굵은 수평직선으로 표시하였다. $1/2 < m/r < 1$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 그림 2.3에서 굵은 수평점선으로 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.

그림 2.2와 그림 2.3을 그리기 위한 R 스크립트는 다음과 같다.

```
# 비올 확률과 의사결정
x=c(0,0,1)
y=c(0,1,1)
plot(0:1,0:1,type="n",xlab="m/r",ylab="p",xlim=c(0,1.5))
```

```
1 1 0 0 1 0 0 1 0 1
0 0 1 0 1 1 0 0 1 1
1 0 0 1 1 1 0 0 0 1
0 0 1 0 1 1 0 1 0 1
1 1 0 1 0 0 0 1 1 0
0 1 1 0 0 1 0 0 1 1
0 0 1 1 1 0 1 0 0 1
1 1 0 1 1 0 0 0 1 0
0 0 1 0 1 1 1 0 0 1
1 0 0 1 1 0 1 0 1 1
```

```
1 0 0 1 0 1 0 0 0 1
0 1 0 0 0 0 1 1 1 0
0 0 0 1 0 1 1 0 1 0
0 1 0 1 1 1 1 1 1 1
0 1 1 1 1 0 0 0 0 1
1 1 1 1 0 0 0 1 0 0
0 0 1 0 0 1 0 0 0 0
0 1 0 0 0 1 0 0 1 0
0 1 0 0 0 0 0 0 0 0
1 0 1 0 0 1 0 1 1 0
```

```
polygon(x,y,density=10)
abline(h=1)
abline(h=0)
lines(c(1,1),c(0,1),lty=2)

# 비올 확률: 0.5
x=c(0,0,1)
y=c(0,1,1)
plot(0:1,0:1,type="n",xlab="m/r",ylab="p",xlim=c(0,1.5))
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),lwd=5)
lines(c(0.5,1.5),c(0.5,0.5),lwd=5,lty=2)
```

사례 2(가짜를 찾아라.)

다음 그림 중 하나는 100개의 동전을 실제로 던진 결과이고 하나는 100개의 동전을 실제로 던지지 않고 가짜로 머릿속에 떠오르는 대로 가짜로 작성한 결과이다. 여기서 0은 숫자면, 1은 그림면이다. 어느 것이 가짜이고 어느 것이 진짜일까?

답은 왼쪽이 가짜이고 오른쪽이 진짜이다. 이 결과는 어느 학생이 실험한 결과이다. 여기서 특이한 것은 가짜에서는 그림면이 나오는 확률이 51/100 = 0.51, 숫자면이 나오는 확률이 49/100 = 0.49로 비슷하나 진짜에서는 그림면이 나오는 확률이 41/100 = 0.41, 숫자면이 나오는 확률이 59/100 = 0.59로 차이가 많이 난다. 동전을 100번 정도만 던져가지고는 그림면이 나오는 확률이 0.5가 되기가 쉽지 않음을 알 수 있다. 컴퓨터 시뮬레이션으로 동전을 던져보면 어떤 결과가 나올까? 다음 그림 2.4는 컴퓨터 시뮬레이션으로 동전을 10,000번 던질 때까지의 그림면이 나올 확률의 변화를 나타내는 그림이다. 던진 횟수가 적을 때는 확률이 매우 불안정하게 요동치는 것을 알 수 있다. 최소 1,000번은 던져야 확률이 0.5에 겨우 수렴함을 알 수 있다.

우리는 다음과 같은 실습을 통하여 가짜를 알아내는 방법을 배울 수 있다(물론 100% 알아낼 수는 없

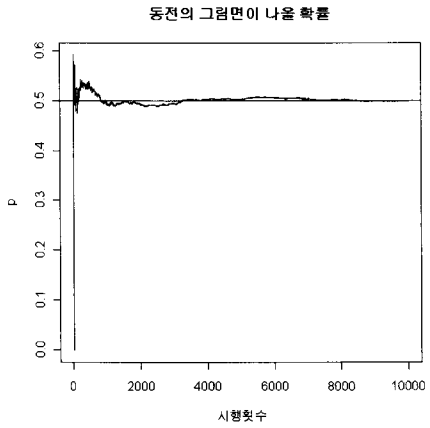

그림 **2.4.** 컴퓨터 시뮬레이션 결과

다.). 이 사례는 Gelman과 Nolan (2002) 책에 언급되어 있는 사례이다.

학생 각자 다음과 같은 실습을 행한다.

1. 100개의 동전을 실제로 던지지 않고 가짜로 머릿속에 떠오르는 데로 가짜로 작성한다. (힌트) 이론적으로 동전의 그림면(1)이 나올 확률은 숫자면(0)이 나올 확률과 같다.
2. 100개의 동전을 실제로 던져 나온 결과를 기록한다.
3. 1과 2 각각에 대하여 그림면과 숫자면이 바뀌는 횟수(number of switches)와 그림면이나 숫자면이 연속적으로 나오는 연(run) 중 가장 긴 길이(length of largest run)를 세어 기록한다.
4. 학생 전체의 결과를 수집하여 그림면과 숫자면이 바뀌는 횟수를 x축에, 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이를 y축으로 하여 학생 각자의 결과를 좌표로 하여 진짜점과 가짜점을 그린 후 이 들 점들을 가짜점에서 시작하여 진짜점에서 끝나는 화살표시가 있는 점선으로 연결한다.
5. 어떤 패턴이 있는 가에 대하여 토론한다.

이 실습의 목적은 통계의 조작이 발생할 때 이 조작을 발견할 수 있는 수단을 필히 강구하여야한다는 사실을 강조하기 위하여 실습을 행한다. 컴퓨터 시뮬레이션으로 100개의 동전을 던지는 모의실험을 10,000번을 행하여 각각에 대하여 그림면과 숫자면이 바뀌는 횟수(number of switches)를 x축에, 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이(length of largest run)을 y축으로 하여 점을 찍은 결과가 다음 그림 2.5와 2.6이다. 그림 2.5는 해바라기그림이다. 크게 보면 점들이 하나의 집락을 형성하고 있다. 그림 2.6은 jittering을 행한 후 점들을 회색으로 찍은 후 그 위에 등고선도를 그렸다. 전체적으로 뾰쪽한 산 모양 형태를 이룸을 알 수 있다. 자세히 보면 높이가 서로 다른 뾰쪽한 봉우리가 세 개가 있고(가운데 봉우리가 제일 높음) 같이 한 군 데 몰려 있음을 알 수 있다. 컴퓨터 시뮬레이션을 재차 시행하여 보면 또 다른 형태의 등고선도가 나타난다.

다음 그림 2.7은 위의 시뮬레이션 결과를 이용하여 결합밀도추정을 행하여 그린 3차원 그림이다. 우리는 앞의 등고선도와 비교하여 볼 수 있다.

25명의 학생에게 앞에서와 같은 실습을 시킨 결과가 다음 그림 2.8이다. 학생 전체의 결과를 수집하여 그림면과 숫자면이 바뀌는 횟수(number of switches)를 x축에, 그림면이나 숫자면이 연속적으로 나오

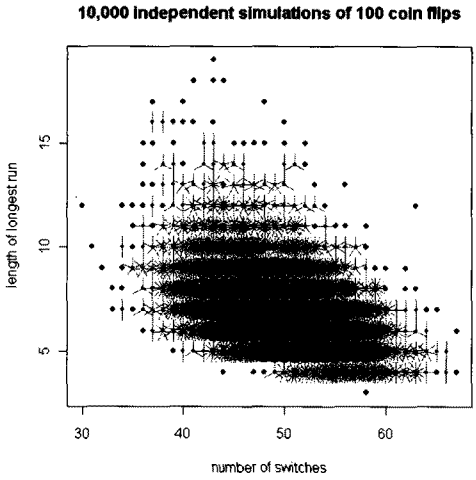

그림 **2.5.** 해바라기 그림(시뮬레이션 결과)

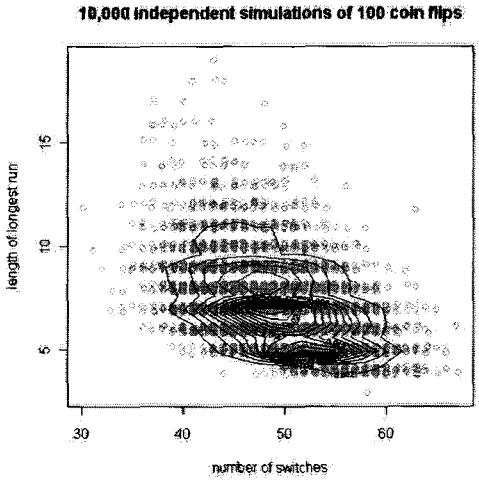

그림 **2.6.** 등고선도(시뮬레이션 결과)

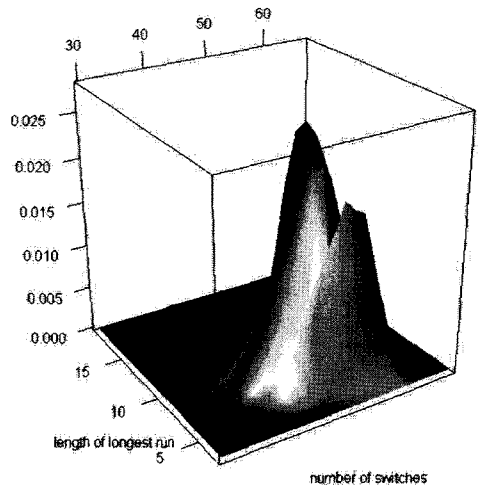

그림 **2.7.** 결합밀도추정량그림(시뮬레이션 결과)

는 연 중 가장 긴 길이(length of largest run)을 y축으로 하여 학생 각자의 결과를 좌표로 하여 진짜점과 가짜점을 그린 후 이들 점들을 가짜점에서 시작하여 진짜점에서 끝나는 화살표시가 있는 점선으로 연결하였다. 대체적으로 그래프의 남동쪽에 있던 점들(가짜 결과들)이 그래프의 북서쪽에 있는 점들(진짜 결과들)로 이동하고 있음을 알 수 있다. 왜 이런 패턴이 발생할까? 일반적으로 학생들은 그림면과 숫자면이 나오는 현상이 무작위하다(random)고 생각하기 때문에 그림면과 숫자면을 자주 바꾸기 때문에 그림면과 숫자면이 바뀌는 횟수가 많아지는 반면 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이는 짧게 된다. 그래서 가짜 결과들인 점들이 그래프의 남동쪽에 주로 위치하게 된다. 그런데 실제 던져보면 우리가 언뜻 생각하는 것만큼 그림면과 숫자면이 상대적으로 자주 바뀌지 않기 때문에 그림면

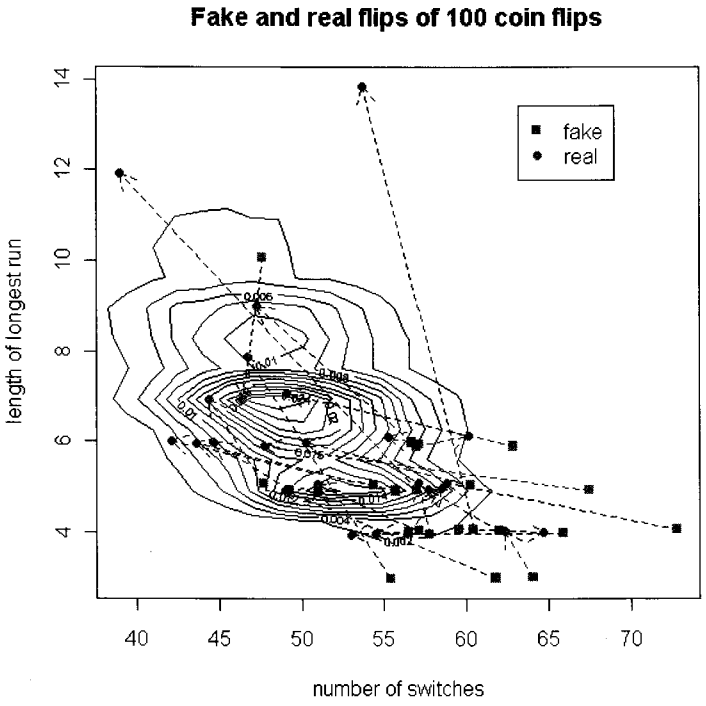

그림 **2.8.** 실습의 결과

과 숫자면이 바뀌는 횟수가 상대적으로 적어지는 반면 그림면이나 숫자면이 연속적으로 나오는 연 중 가장 긴 길이는 상대적으로 길게 된다. 그래서 진짜 결과들인 점들이 그래프의 북서쪽에 주로 위치하게 된다. 물론 이러한 패턴을 따르지 않는 학생들(특이값)도 있다. 그러나 이러한 학생들의 숫자는 그리 많지 않다.

그림 2.4–2.8을 그리기 위한 R 스크립트는 부록 A과 같다. 그림 2.8을 그릴 때 각 학생들의 자료파일이 필요하다. 예로 'student-1-fake.txt'는 student-1 학생의 가짜 자료이고 'student-1-real.txt'는 student-1 학생의 진짜 자료이다. 부록 B에서 'student-1-fake.txt'과 'student-1-real.txt' 구성 예를 보였다.

사례 3(생일이 같을 확률은?)

한 방에 모여 있는 사람의 수가 N이라 하자. N명 중 최소한 두 명이 같은 생일을 가질 확률을 구하여라(1년은 365일로 가정함.).

(풀이) 우선 $P[N$명이 모두 다른 생일을 가진다$]$를 구하여보자. 첫 번째 사람이 365일 중 하루를 선택할 확률은 365/365, 첫 번째 사람이 365일 중 하루를 선택하면 두 번째 사람은 나머지 364일 중 하나를 선택하여야 하므로 두 번째 사람이 하루를 선택할 확률은 364/365, 이런 식으로 생각하면 $P[N$명이 모두 다른 생일을 가진다$] = \{365\times364\times363\times\cdots\times(365-N+1)\}/365^N$이 된다. 그러므로 우리가 구하여야 할 확률은 여사건의 확률공식을 이용하여 $P[N$명 중 최소한 두 명이 같은 생일을 가진다$]$

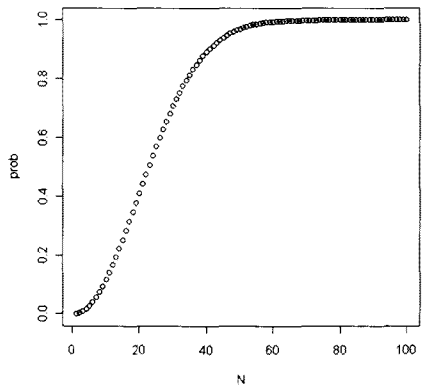

그림 **2.9.** 생일이 같은 사람이 적어도 한 쌍 이상 나올 확률

$= 1 - P$[명이 모두 다른 생일을 가진다] $= 1 - \{365\times364\times363\times\cdots\times(365-N+1)\}/365^N$이 된다. $N(N = 1, 2, \ldots, 100)$에 따른 확률 $P[N$명 중 최소한 두 명이 같은 생일을 갖는다]의 변화를 그림으로 그리면 다음 그림 2.9와 같다.

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. 이 문제는 365개의 면을 가지고 있는 주사위 40개를 동시에 던졌을 때 같은 숫자가 나오는 쌍이 하나라도 나오는 확률을 구하는 문제인 셈이다. 우리가 언뜻 생각할 때 1년은 365일이므로 최소한 두 명이 같은 생일을 갖게 되려면 365명 이상이 있어야할 것 같지만 70명만 되어도 최소한 두 명이 같은 생일을 갖게 될 확률이 무려 0.999가 된다.

그림 2.9를 그리기 위한 R 스크립트는 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")
```

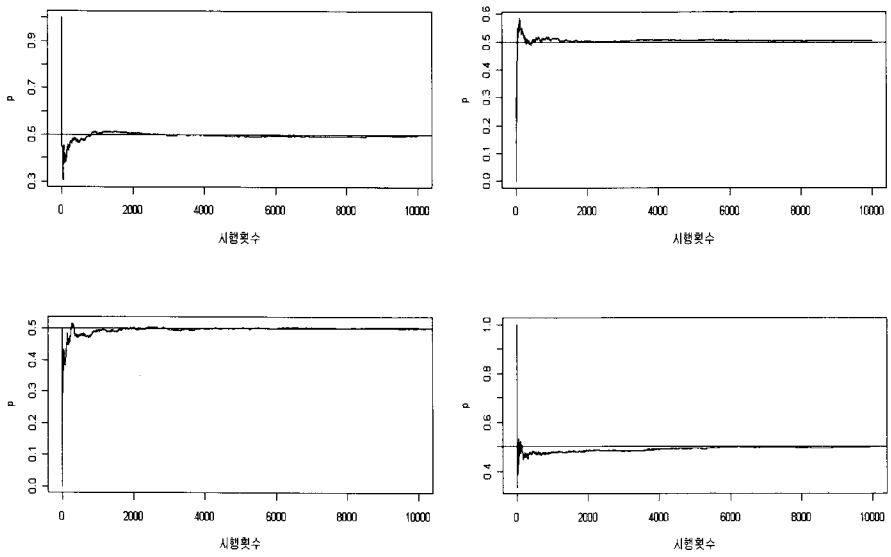

그림 2.10. 동전의 그림면이 나올 확률(컴퓨터 시뮬레이션 결과, 각 10,000번 시행)

사례 4(빈도적 확률에 대하여 알아보자.)

중고등학교 수학교과서에서 논리적확률(고등학교 수학교과서에서는 '수학적 확률'이라 칭함)로 구할 수 없는 빈도적확률(고등학교 수학교과서에서는 '통계적 확률'이라 칭함.)의 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) 확률 등이 있다.

동전을 컴퓨터시뮬레이션을 이용하여 10,000번 던져 동전의 그림면이 나오는 확률을 네 차례 구하여보니 다음 그림 2.10과 같은 네 가지 그림이 나왔다. 우리는 네 가지 그림에서 각각 0.5에 수렴하는 것을 볼 수 있다. 그러나 수렴하는 패턴은 서로 다름을 알 수 있다.

반면에 동전을 컴퓨터시뮬레이션을 이용하여 100번 던져 동전의 그림면이 나오는 확률을 네 차례 구하여보니 다음 그림 2.11과 같은 네 가지 그림이 나왔다. 첫 번째와 두 번째 그림을 보면 0.5에 수렴한다고 보기가 어렵다. 즉 시행횟수가 적으면 빈도적 확률을 구하기가 어렵게 된다는 것을 알 수 있다. 그러므로 빈도적 확률을 우리가 보기 위해서는 컴퓨터를 이용한 시행횟수를 크게 하여 통계적 시뮬레이션을 행하여 볼 필요가 있다.

빈도적 확률을 구할 때 주의할 또 한 가지 사항은 시행횟수가 커짐에 따라 동전의 그림면이 나오는 상대도수의 극한값이 0.5가 된다고 해서 동전의 그림면이 나오는 횟수와 동전의 숫자면이 나오는 횟수의 차이가 0에 가까워진다는 것이 아니다. 시행횟수가 커짐에 따라 동전의 그림면이 나오는 횟수와 동전의 숫자면이 나오는 횟수의 차이는 커졌다 작아졌다 한다. 즉 이 차이는 랜덤하게 된다. 이를 다음 그림 2.12와 같은 통계적 시뮬레이션으로 확인하여 보자.

다음 그림 2.13은 시행횟수를 백만 번으로 했을 때 시행횟수가 커짐에 따라 동전의 그림면이 나오는 횟수와 동전의 숫자면이 나오는 횟수의 차이를 나타내는 그림이다. 시행횟수가 커짐에 따라 동전의 그림면이 나오는 횟수와 동전의 숫자면이 나오는 횟수의 차이는 커졌다 작아졌다 하나 그 변동 폭이 시행횟

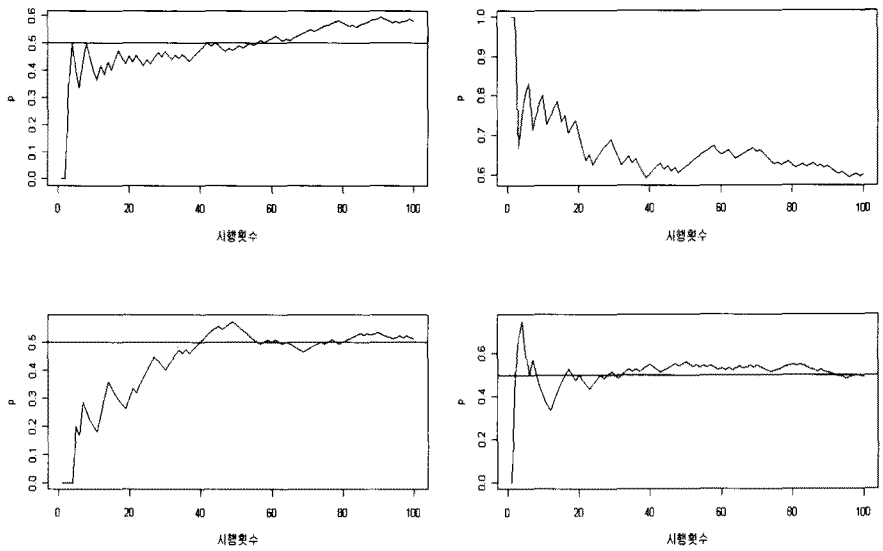

그림 **2.11.** 동전의 그림면이 나올 확률(컴퓨터 시뮬레이션 결과, 각 100번 시행)

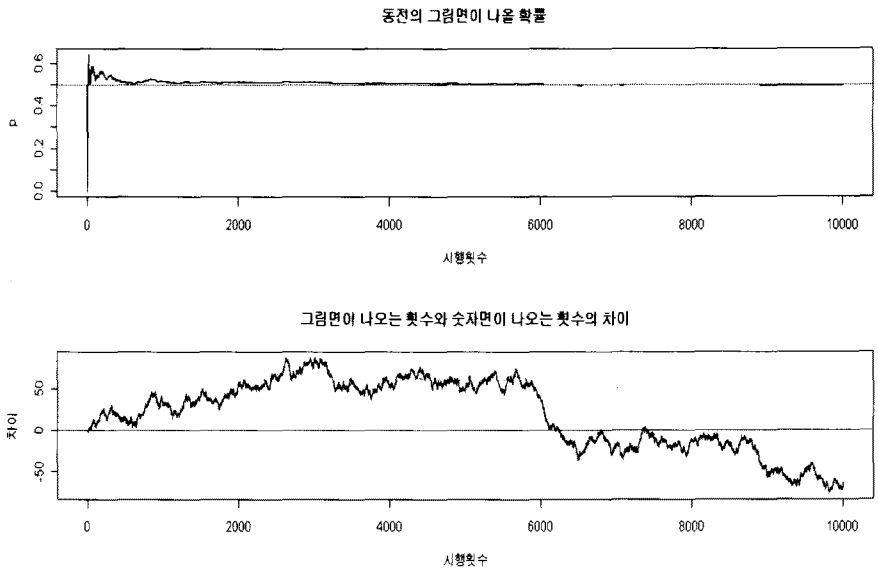

그림 **2.12.** 동전의 그림면이 나올 확률 및 동전의 그림면이 나오는 횟수와 동전의 숫자면이 나오는 횟수의 차이(컴퓨터 시뮬레이션 결과 – 만 번 시행)

수가 만 번일 때보다 큼을 알 수 있다.

사례 4에 나오는 그림들을 그리기 위한 R 스크립트는 다음과 같다.

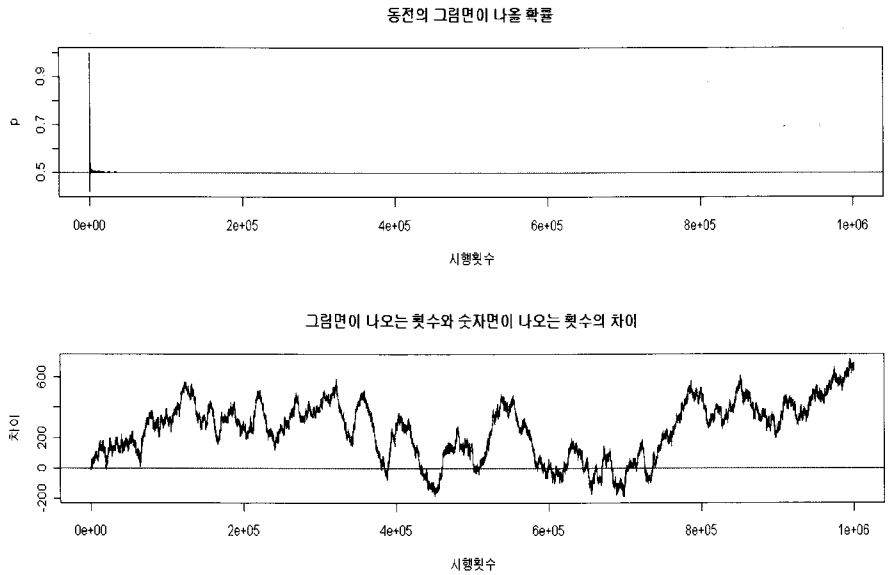

그림 2.13. 동전의 그림면이 나올 확률 및 동전의 그림면이 나오는 횟수와 동전의 숫자면이 나오는 횟수의 차이(컴퓨터 시뮬레이션 결과 – 백만번 시행)

```
law.large.number= function(n)
{
par(mfrow=c(2,2))
for(i in 1:4)
{
# 베르누이분포에서의 난수
x1=rbinom(n,1,0.5)
# 상대도수
xbar1=cumsum(x1)/1:n
plot(xbar1,ylab="p",xlab="시행횟수",type="l")
abline(0.5, 0)
}
par(mfrow=c(1,1))
}
law.large.number(10000)
law.large.number(100)

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 그림면이 나올 확률",type="l")
abline(0.5, 0, col="red")
# 동전의 그림면이 나오는 횟수와 동전의 숫자면이 나오는 횟수의 차이
H=cumsum(x1)
T=1:n-H
difference=H-T
plot(difference,ylab="차이",xlab="시행횟수", main="그림면이 나오는 횟수와 숫자면이 나오는 횟수의 차이",type="l")
abline(0, 0, col="red")
par(mfrow=c(1,1))
}
law.large.number2(10000)
law.large.number2(1000000)
```

사례 5(승률과 표준오차)

다음 그림 2.14는 프로야구 2007년 시즌이 끝난 후 8개 구단의 승률을 나타내고 있는 꺾은선그래프이다. 그래프의 x축은 날짜, y축은 승률을 나타낸다. 각 구단의 승률이 시즌 초기에는 불안정하게 변동이 심하다가 시즌 후반기에 가면 안정적으로 특정값에 수렴하는 것을 볼 수 있다. 왜 이런 현상이 벌어질까?

각 팀에서 치룬 경기 수 중 이긴 경기수와 진 경기수를 합친 경기수를 n이라 하고 이긴 경기수를 w라 하면 승률은 $\hat{p} = w/n$이 된다. 예로 SK의 경우 이긴 경기수가 73, 진 경기수가 48이므로 승률은 $73/(73+48) \approx 0.603$이 된다. 이러한 승률의 표준오차(승률의 변동의 크기를 나타내는 값)는 $\sqrt{p(1-p)/n}$가 된다. 여기서 p는 모승률이다. $0 \leq p \leq 1$이므로 $p(1-p)$의 최대값은 다음 그림 2.15에서 보는 것과 같이 $p = 1/2$에서 $1/4$이 되므로 승률의 표준오차의 최대값은 $1/(2\sqrt{n})$이 되어 n이 점점 커지면 승률의 표준오차의 최대값은 점점 작아지게 된다. 그래서 각 구단의 승률이 시즌 초기에는 불안정하게 변동이 심하다가 시즌 후반기에 가면 안정적으로 특정값에 수렴하게 되는 것이다. p를 우리는 모르므로 승률의 표준오차로서 $\sqrt{\hat{p}(1-\hat{p})/n}$을 주로 사용한다.

각 팀의 승/패/무와 승률 및 승률의 표준오차는 다음 표 2.3과 같다.

각 팀 승률의 표준오차 값이 0.044에서 0.046까지 큰 차이가 없이 비슷하고 값이 작다. 그러므로 각 팀이 126 게임을 치룬 후의 승률을 우리는 참 승률에 대한 근사값으로 여기고 각 팀의 2007년 실력으로 순위를 정하는 것이다.

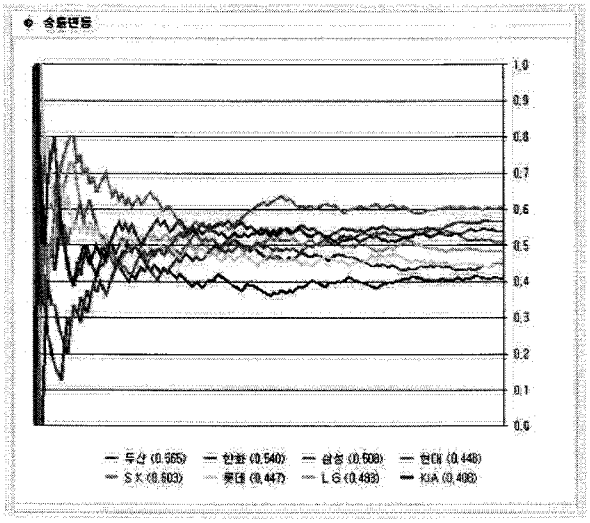

그림 2.14. 프로야구 8개 구단의 승률(2007년 시즌, 제공처: 스포츠조선)

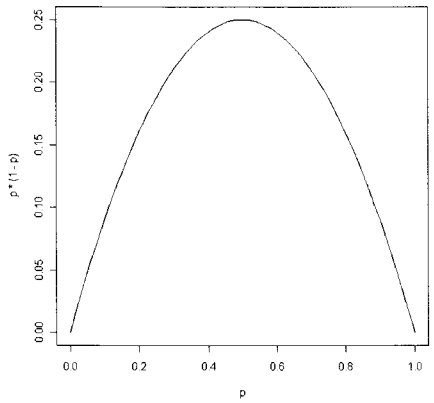

그림 2.15. $p(1-p)$의 그래프

표 2.3. 프로야구 8개구단의 승률표(2007년 시즌)

순위	팀	승	패	무	승률	승률의 표준편차
1	SK	73	48	5	0.603	0.045
2	두산	70	54	2	0.565	0.045
3	한화	67	57	2	0.540	0.045
4	삼성	62	60	4	0.508	0.045
5	LG	58	62	6	0.483	0.046
6	현대	56	69	1	0.448	0.045
7	롯데	55	68	3	0.447	0.045
8	KIA	51	74	1	0.408	0.044

표 2.4. 프로야구 8개구단의 승률 및 피타고라스 승률표(2007년 시즌)

순위	팀	승	패	무	총득점	총실점	승률	피타고라스 승률
1	SK	73	48	5	603	465	0.603	0.627
2	두산	70	54	2	578	480	0.565	0.592
3	한화	67	57	2	534	481	0.540	0.552
4	삼성	62	60	4	497	509	0.508	0.488
5	LG	58	62	6	532	600	0.483	0.440
6	현대	56	69	1	530	615	0.448	0.426
7	롯데	55	68	3	533	554	0.447	0.481
8	KIA	51	74	1	499	602	0.408	0.407

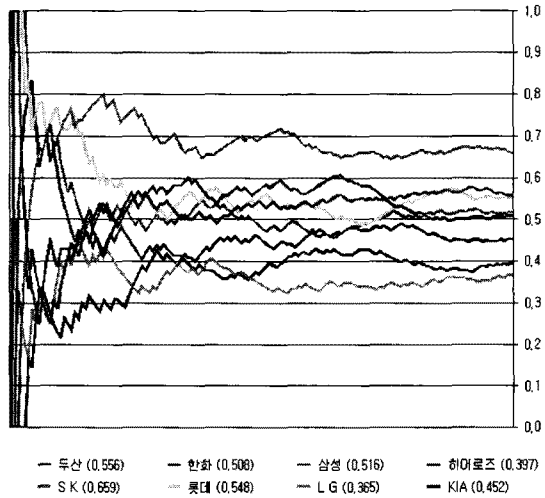

그림 **2.16.** 프로야구 8개 구단의 승률(2008년 시즌, 제공처: 스포츠조선)

야구에서는 그 팀의 기록과 전력을 바탕으로 기대되는 다른 형태의 승률이 존재한다. 이러한 승률중 하나인 '피타고라스 승률'은 야구통계학자 빌 제임스가 1980년대 초에 고안한 승률로서 '총득점의 제곱/(총득점의 제곱 + 총실점의 제곱)'의 값으로 그 팀의 기록과 전력을 바탕으로 기대되는 승률을 파악할 수 있다. 실제로 피타고라스 승률이 시즌 종료 시점에는 실제 팀 승률과 거의 근접하는지 조사하여 보자.

2007년 시즌이 끝난 후 각 팀의 승/패/무, 승률, 총득점, 총실점 및 피타고라스 승률은 다음과 표 2.4와 같다. 승률에서의 순위와 피타고라스 승률에서의 순위가 다른 팀이 한 팀으로 롯데임을 알 수 있다. 롯데는 총득점과 총실점의 입장에서는 5위인 LG나 6위인 현대보다도 성적이 좋고 피타고라스 승률로서는 4위인 삼성에 필적한다. 그럼에도 불구하고 승률에서의 순위는 7위밖에 안 된다. 경제적인 야구를 못했다는 결론을 얻을 수 있다.

그럼 2008년의 결과는 어떨까? 다음 그림 2.16은 2008년 시즌이 끝난 후 8개 구단의 승률을 나타내고 있는 꺾은선그래프이다. 그래프의 x축은 날짜, y축은 승률을 나타낸다. 2007년과 마찬가지로 각 구단의 승률이 시즌 초기에는 불안정하게 변동이 심하다가 시즌 후반기에 가면 안정적으로 특정값에 수렴하는 것을 볼 수 있다.

표 **2.5.** 프로야구 8개구단의 승률 및 피타고라스 승률표(2008년 시즌)

순위	팀	승	패	무	총득점	총실점	승률	승률의 표준오차	피타고라스 승률
1	SK	83	43	0	632	461	0.659	0.042	0.653
2	두산	70	56	0	647	542	0.556	0.044	0.588
3	롯데	69	57	0	624	518	0.548	0.044	0.592
4	삼성	65	61	0	557	596	0.516	0.045	0.466
5	한화	64	62	0	592	595	0.508	0.045	0.497
6	KIA	57	69	0	503	555	0.452	0.044	0.451
7	히어로즈	50	76	0	499	609	0.397	0.044	0.402
8	LG	46	80	0	468	646	0.365	0.043	0.344

2007년과 마찬가지로 승률과 피타고라스 승률을 구하면 다음 표 2.5와 같다. 롯데는 승률에서는 3위이나 피타고라스 승률로는 승률이 2위인 두산보다 높고 한화는 승률에서는 5위이나 피타고라스 승률로는 승률이 4위인 삼성보다 높다.

사례 5를 위한 R 스크립트는 다음과 같다. 부록 C에서 '프로야구2007.txt'과 '프로야구2008.txt'의 내용을 보였다.

```
options(digits=3)
# 2007년 프로야구 승률
프로야구2007=read.table("c:/프로야구2007.txt",header=T)
attach(프로야구2007)
승률=승/(승+패)
승률표준오차=sqrt(승률*(1-승률)/(승+패))
프로야구2007승률표=cbind(프로야구2007,승률,승률표준오차)
프로야구2007승률표
# p(1-p)의 그래프
p=seq(0,1,by=0.01)
plot(p,p*(1-p),type="l",main="p(1-p)의 그래프")
# 2007년 프로야구 승률과 피타고라스 승률
피타고라스승률=총득점^2/(총득점^2+총실점^2)
프로야구2007승률표2=cbind(프로야구2007승률표,피타고라스승률)
프로야구2007승률표2

# 2008년 프로야구 승률
프로야구2008=read.table("c:/프로야구2008.txt",header=T)
attach(프로야구2008)
승률=승/(승+패)
승률표준오차=sqrt(승률*(1-승률)/(승+패))
```

표 **2.6.** 지원자수와 합격자수를 남녀별로 정리한 분할표(괄호 안은 지원자수)

학부	합격한 남학생	합격한 여학생
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)

표 **2.7.** 합격률표

학부	남자합격률	여자합격률	전체합격률
A	0.619	0.824	0.643
B	0.629	0.680	0.631
C	0.337	0.352	0.344
D	0.059	0.070	0.064
전체	0.472	0.309	0.426

```
프로야구2008승률표=cbind(프로야구2008,승률,승률표준오차)
프로야구2008승률표
# 2008년 프로야구 승률과 피타고라스 승률
피타고라스승률=총득점^2/(총득점^2+총실점^2)
프로야구2008승률표2=cbind(프로야구2008승률표,피타고라스승률)
프로야구2008승률표2
```

사례 6(심슨의 역설(Simpson's Paradox))

분할표를 전체적으로 볼 때와 부분적으로 볼 때의 결과가 서로 다른 결과를 나타내는 경우가 종종 있는데 이를 심슨의 역설(Simpson's paradox)이라고 한다. 표 2.6은 어느 대학교의 4개의 학부(이 대학교는 4개의 학부로 구성되어 있음.)에 지원한 학생 수와 합격한 학생 수를 남녀별로 정리한 분할표이다. 다음 문제를 풀어 보자.

(a) 전체 합격률, 남자합격률과 여자합격률을 막대그래프로 그리시오. 무엇을 알 수 있습니까?

(b) 각 학부별 합격률, 남자합격률과 여자합격률을 막대그래프로 그리시오. 무엇을 알 수 있습니까?

(c) (a)와 (b)를 종합하여 결론을 내려 보시오.

(d) 합격한 남학생과 합격한 여학생의 학부 분포율을 원형그래프로 그려 비교하시오.

(풀이)

(a) 분할표를 이용하여 합격률표를 다음 표 2.7과 같이 작성한다.
전체 합격률, 남자합격률과 여자합격률을 막대그래프로 그려보면 다음 그림 2.17과 같다. 전체 남자의 합격률이 여자의 합격률보다 약 16% 높음을 알 수 있다.

(b) 각 학부별 합격률, 남자합격률과 여자합격률을 막대그래프로 그려보면 다음 그림 2.18과 같다. 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 큼을 알 수 있다.

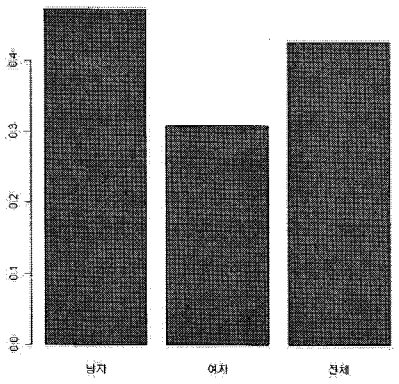

그림 2.17. 합격률 막대그래프

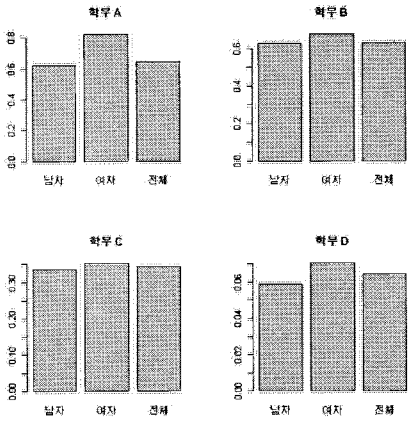

그림 2.18. 각 학부별 합격률 막대그래프

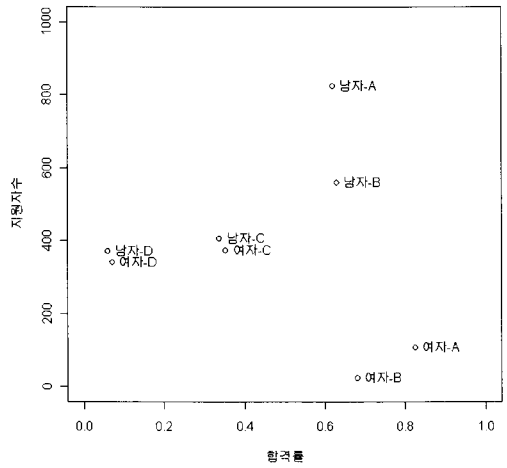

그림 2.19. 학부와 성별 합격률과 지원자수를 나타내는 산점도

(c) 전체적으로는 남자의 합격률이 여자의 합격률보다 크나 각 학부별로는 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 큼을 알 수 있다. 이러한 현상을 심슨의 역설이라고 한다. 이러한 현상이 발생한 이유는 무엇일까? 이 이유를 알기 위하여 다음 그림 2.19와 같은 산점도를 살펴보자.

여자의 경우 학부 C와 D에서 상대적으로 지원자수는 많으나 합격률이 낮은 반면 학부 A와 B에서는 합격률이 높으나 상대적으로 지원자수가 적다. 즉, 여자들은 합격률이 낮은 학부에 많이 지원한 반면 합격률이 높은 학부에는 적게 지원하였다. 반면 남자의 경우는 학부 C와 D에서 상대적으로 지원자수가 적고 합격률도 낮은 반면 학부 A와 B에서는 상대적으로 지원자수가 많고 합격률도 높다. 즉, 남자들은 여자들과 달리 합격률이 낮은 학부에 적게 지원한 반면 합격률이 높은 학부에는 많이 지원하였다. 이런 연유로 전체적으로는 남자의 합격률이 여자의 합격률보다 크나 각 학부별로

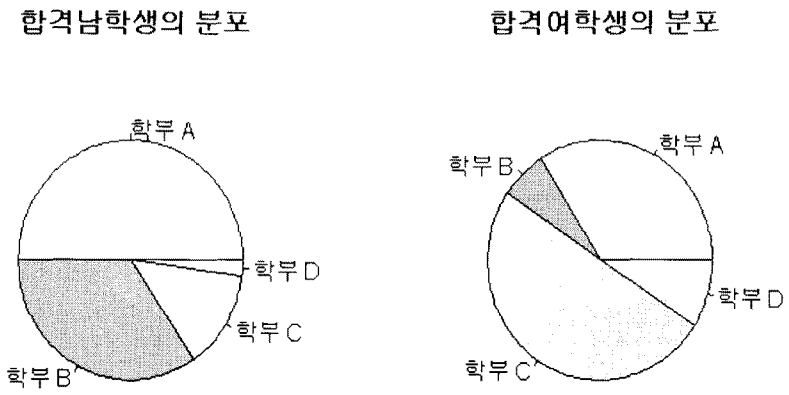

그림 **2.20.** 합격남학생과 합격여학생의 학부 분포율을 나타내는 원형그래프

는 4개의 학부 모두 여자의 합격률이 남자의 합격률보다 크게 된 것이다. 우리가 관심이 있는 남녀 간의 전체 합격률의 차이는 남녀 간의 차별 때문이 아니라 남녀 간의 학부별 지원 선호도 차이로 인하여 발생하였음을 알 수 있다. '남녀 간의 학부별 지원 선호도 차이'라는 중요한 변수를 고려하지 않고 '남녀'라는 성별의 차이로만 합격률을 보면 잘못된 결론을 내릴 수가 있다.

(d) 합격한 남학생과 합격한 여학생의 학부 분포율을 원형그래프로 그리면 다음 그림 2.20과 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.

사례 6에 나오는 그림들이나 표들을 그리기 위한 R 스크립트는 다음과 같다.

```
options(digits=3)
# 도수분포표
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)
colnames(entrance)=c("합격 남학생", "지원 남학생", "합격 여학생", "지원 여학생")
rownames(entrance)=c("학부 A", "학부 B", "학부 C", "학부 D", "전체")
entrance
남자합격률=entrance[,1]/entrance[,2]
여자합격률=entrance[,3]/entrance[,4]
전체합격률=(entrance[,1]+entrance[,3])/(entrance[,2]+entrance[,4])
prop.entrance=cbind(남자합격률, 여자합격률, 전체합격률)
prop.entrance

# 막대그래프
```

```
prop.total=prop.entrance[5,]
names(prop.total)=c("남자", "여자", "전체")
barplot(prop.total, col="green")

# 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=rownames(entrance)[i])
}

# 심슨의 역설
# 남자
합격률=c(0.619,0.629,0.337,0.059)
지원자수=c(825,560,407,373)
plot(합격률,지원자수,xlim=c(0,1),ylim=c(0,1000))
text(0.619,825,"남자-A",pos=4)
text(0.629,560,"남자-B",pos=4)
text(0.337,407,"남자-C",pos=4)
text(0.059,373,"남자-D",pos=4)
# 여자
합격률=c(0.824,0.680,0.352,0.070)
지원자수=c(108,25,375,341) par(new=T)
plot(합격률,지원자수,xlim=c(0,1),ylim=c(0,1000))
text(0.824,108,"여자-A",pos=4)
text(0.680,25,"여자-B",pos=4)
text(0.352,375,"여자-C",pos=4)
text(0.070,341,"여자-D",pos=4)

# 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="합격남학생의 분포")
pie(합격여학생, main="합격여학생의 분포")
```

3. 결론

기초 통계학 확률부분 강의 시 유용하게 사용할 수 있는 사례들을 6가지 제시하였다. 각 사례에 R스크립트를 첨부하였으므로 기초 통계학 확률부분 강의 시 사례들을 제시하며 R 패키지를 사용하여 구현하여 볼 수 있다. 사례 5는 '통계적 추정'과도 관계를 갖고 사례 6는 '분할표'와도 관계를 갖는다.

부록 A: 그림 2.4-그림 2.8을 그리기 위한 R 스크립트

```
# 베르누이분포에서의 난수
x=rbinom(100,1,0.5)
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 그림면이 나올 확률",type="l")
abline(0.5, 0)
}
law.large.number(10000)

# 동전 100개 던지기(10,000번 시뮬레이션 시행)
x.val=c(0,10000)
y.val=c(0,10000)
for(j in seq(1,10000))
{
x=rbinom(100,1,0.5)
longest=rep(1,100)
switch=0
for(i in seq(1,99))
{
```

```
if(x[i]!=x[i+1]) switch=switch+1
if(x[i]==x[i+1]) longest[i+1]=longest[i]+1
}
x.val[j]=switch
y.val[j]=max(longest)
}
# jittering
jitter.x=jitter(x.val)
jitter.y=jitter(y.val)
xmin=min(x.val);xmax=max(x.val)
ymin=min(y.val);ymax=max(y.val)
j.xmin=min(jitter.x);j.xmax=max(jitter.x)
j.ymin=min(jitter.y);j.ymax=max(jitter.y)

# scatterplot with jittering
plot(jitter.x,jitter.y,type="p", xlim=c(j.xmin,j.xmax),ylim=c(j.ymin,j.ymax),xlab="number of switches"
,ylab="length of longest run",main="10,000 independent simulations of 100 coin flips")

# sunflower plot without jittering
sunflowerplot(x.val,y.val, xlim=c(xmin,xmax),ylim=c(ymin,ymax), xlab="number of switches"
,ylab="length of longest run",main="10,000 independent simulations of 100 coin flips")

# joint density plot & contour plot(with jittering)
library(rgl)
library(MASS)
density2=kde2d(jitter.x,jitter.y,n=25)
open3d()
bg3d("white")
material3d(col="black")
persp3d(density2$x,density2$y,density2$z,col="lightblue", xlab="number of switches",ylab="length of longest run",zlab="density")

image(density2$x,density2$y,density2$z, xlim=c(j.xmin,j.xmax),ylim=c(j.ymin,j.ymax),xlab="number of switches",ylab="length of longest run", main="10,000 independent simulations of 100 coin flips")
contour(density2$x,density2$y,density2$z,xlim=c(j.xmin,j.xmax),ylim=c(j.ymin,j.ymax),add=T)
```

```
contour(density2$x,density2$y,density2$z, xlim=c(j.xmin,j.xmax),ylim=c(j.ymin,j.ymax),xlab="number of switches",ylab="length of longest run", main="10,000 independent simulations of 100 coin flips")

plot(jitter.x,jitter.y,type="p",col="gray", xlim=c(j.xmin,j.xmax),ylim=c(j.ymin,j.ymax),xlab="number of switches",ylab="length of longest run", main="10,000 independent simulations of 100 coin flips")
par(new=T)
contour(density2$x,density2$y,density2$z, xlim=c(j.xmin,j.xmax),ylim=c(j.ymin,j.ymax),xlab="number of switches",ylab="length of longest run", main="10,000 independent simulations of 100 coin flips")

fake=matrix(rep(0,5000),ncol=50,byrow=T)
real=matrix(rep(0,5000),ncol=50,byrow=T)
fake[,1]=scan("c:/사례-2/student-1-fake.txt")
fake[,2]=scan("c:/사례-2/student-2-fake.txt")
fake[,3]=scan("c:/사례-2/student-3-fake.txt")
(생략)
fake[,25]=scan("c:/사례-2/student-25-fake.txt")
real[,1]=scan("c:/사례-2/student-1-real.txt")
real[,2]=scan("c:/사례-2/student-2-real.txt")
real[,3]=scan("c:/사례-2/student-3-real.txt")
(생략)
real[,25]=scan("c:/사례-2/student-25-real.txt")
N=25
# jittering
x.fake=c(rep(0,50))
y.fake=c(rep(0,50))
x.real=c(rep(0,50))
y.real=c(rep(0,50))
x=matrix(rep(0,100),,ncol=50,byrow=T)
y=matrix(rep(0,100),,ncol=50,byrow=T)
for(j in 1:N)
{
longest=rep(1,100)
switch=0
for(i in seq(1,99))
{
if(fake[i,j]!=fake[i+1,j]) switch=switch+1
```

```
if(fake[i,j]==fake[i+1,j]) longest[i+1]=longest[i]+1
}
x.fake[j]=jitter(switch)
y.fake[j]=jitter(max(longest))
longest=rep(1,100)
switch=0
for(i in seq(1,99))
{
if(real[i,j]!=real[i+1,j]) switch=switch+1
if(real[i,j]==real[i+1,j]) longest[i+1]=longest[i]+1
}
x.real[j]=jitter(switch)
y.real[j]=jitter(max(longest))
x[,j]=c(x.fake[j],x.real[j])
y[,j]=c(y.fake[j],y.real[j])
}

xmin2=min(x.fake[1:N],x.real[1:N]);xmax2=max(x.fake[1:N],x.real[1:N])
ymin2=min(y.fake[1:N],y.real[1:N]);ymax2=max(y.fake[1:N],y.real[1:N])

# 등고선도(jittering) + 화살표시
contour(density2$x,density2$y,density2$z, xlim=c(xmin2,xmax2),ylim=c(ymin2,ymax2),xlab=
"number of switches",ylab="length of longest run",main="Fake and real flips of 100 coin flips")
for(j in 1:N)
{
par(new=T)
plot(x.fake[j],y.fake[j],type="p",pch=15, xlim=c(xmin2,xmax2),ylim=c(ymin2,ymax2),col="red",
xlab="number of switches",ylab="length of longest run",main="Fake and real flips of 100 coin
flips")
par(new=T)
plot(x.real[j],y.real[j],type="p",pch=16, xlim=c(xmin2,xmax2),ylim=c(ymin2,ymax2),col="blue",
xlab="number of switches",ylab="length of longest run",main="Fake and real flips of 100 coin
flips")
arrows(x.fake[j],y.fake[j],x.real[j],y.real[j],length=0.15,lty=2)
}
legend(locator(1),legend=c("fake","real"),pch=c(15,16),col=c("red","blue"))
```

부록 B: 'STUDENT-1-FAKE.TXT'과 'STUDENT-1-REAL.TXT' 구성 예

'student-1-fake.txt' 구성 예

1 0 0 1 0 1 1 0 0 1
0 0 1 1 0 1 1 1 0 0
0 0 1 0 1 1 0 1 1 0
0 1 1 1 1 0 0 0 1 0
1 0 0 0 1 1 0 1 0 0
1 0 0 1 1 1 0 0 1 1
0 0 1 1 0 0 1 0 1 1
0 0 0 1 1 0 1 1 0 0
0 1 0 1 1 0 1 0 1 0
1 0 0 1 1 0 0 0 1 1

'student-1-real.txt' 구성 예

0 1 0 1 0 0 1 0 0 0
1 0 1 0 0 0 0 0 1 1
0 1 0 1 0 1 0 1 1 1
0 0 0 1 1 1 1 1 0 0
0 1 1 0 1 1 0 0 1 0
0 0 1 0 1 1 0 0 1 0
1 0 1 0 1 0 0 1 1 1
0 0 1 1 1 0 0 0 1 0
1 0 0 1 0 0 1 1 1 1
0 1 0 0 0 1 0 1 1 0

부록 C: '프로야구2007.TXT'과 '프로야구2008.TXT'의 내용

'프로야구2007.txt'의 내용

순위 팀 승 패 무 총득점 총실점
1 SK 73 48 5 603 465
2 두산 70 54 2 578 480
3 한화 67 57 2 534 481
4 삼성 62 60 4 497 509

5 LG 58 62 6 532 600

6 현대 56 69 1 530 615

7 롯데 55 68 3 533 554

8 KIA 51 74 1 499 602

'프로야구2008.txt'의 내용

순위 팀 승 패 무 총득점 총실점

1 SK 83 43 0 632 461

2 두산 70 56 0 647 542

3 롯데 69 57 0 624 518

4 삼성 65 61 0 557 596

5 한화 64 62 0 592 595

6 KIA 57 69 0 503 555

7 히어로즈 50 76 0 499 609

8 LG 46 80 0 468 646

참고문헌

김영일, 장대흥, 이태림, 강명희 (2008). <통계마인드>, 통계교육원.

Gelman, A. and Nolan, D. (2002). *Teaching Statistics: A Bag of Tricks*, Oxford University Press, Oxford.

Rao, C. R. (2003). <혼돈과 질서의 만남>(원 제목: *Statistics and Truth: Putting Chance to Work*), 이재창 옮김, 나남출판사.

Useful Cases for the Probability Education in the Elementary Statistics Course

Dae-Heung Jang[1]

[1]Division of Mathematical Sciences, Pukyong National University

(Received January 2009; accepted March 2009)

Abstract

Because the probability education in the elementary statistics course is the starting part of the inferential statistics, it is hard to follow the lecture for undergraduate students. In this paper, we suggest the useful cases for the probability education in the elementary statistics course and give the corresponding R scripts.

Keywords: Statistical education, probability, R script.

[1]Professor, Division of Mathematical Sciences, Pukyong National University, 599-1 Daeyeon-dong, Nam-gu, Busan 608-737, Korea. E-mail: dhjang@pknu.ac.kr