韓國數學敎育學會誌 시리즈 E
<數學敎育 論文集>
제 21집 제 2호, 2007. 5. 199-225

J. Korea Soc. Math. Ed. Scr. E:
Communications of Mathematical Education
Vol. 21, No. 2, May. 2007. 199-225

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

장 대 홍 (부경대학교)

초·중·고등학교 확률 및 통계교육 시 우리는 통계패키지로서 R을 사용할 수 있다. R은 대화식 처리방식을 따르기 때문에 배우기가 쉽다. 또한 R에서의 그래픽스는 아주 강력하다. 가장 큰 장점은 R의 사용이 무료라는 것이다. 이러한 많은 장점을 갖고 있는 R을 초·중·고등학교 확률 및 통계교육 현장에서 사용하는 표준 통계패키지로서 고려할 필요가 있다.

I. 서 론

우리나라 초·중·고등학교 수학과 교육은 1997년 교육인적자원부 고시로 제 7차 수학과 교육과정(교육인적자원부 발간)이 개정되어 현재 초·중·고등학교 현장에서 시행되어 오고 있다. 확률 및 통계교육도 이러한 수학과 교육 과정의 한 부분으로서 시행되어진다. 통계학은 관심의 대상에 대한 자료를 수집하고 정리, 요약하며, 실험이나 관측에 의하여 얻은 자료나 정보를 토대로 불확실한 사실에 대하여 과학적인 판단을 내릴 수 있도록 그 방법을 제시하는 학문이다. 이러한 통계학의 본질상 확률 및 통계교육은 귀납법적인 사고를 요구하고 컴퓨터를 통한 자료분석이 필수적으로 이루어져야 한다. 그러나 대부분 초·중·고등학교 확률 및 통계교육은 지필교육 위주로 시행되고 있다. 초·중·고등학교 확률 및 통계교육이 컴퓨터를 통한 자료분석 위주로 운용되지 못하고 지필교육 위주로 시행되는 이유는 컴퓨터를 통한 자료분석을 위해서 컴퓨터실로 이동을 하여야 하고 컴퓨터실의 사용시간을 조정하여야 하는 등의 문제점 외에 통계패키지의 문제점도 있다고 하겠다.

자료분석을 위하여서는 통계패키지가 필수적으로 필요하다. 대표적인 통계패키지로는 SAS, SPSS, Minitab, S-Plus 등이 있다. Excel 내에서도 우리는 '도구(T) -> 데이터분석(D)' 메뉴를 통하여 기초통계학 수준의 자료분석을 행할 수 있다. 중·고등학교 확률 및 통계교육에서는 주로 이 Excel 내의 데이터분석 메뉴를 사용하거나 엑셀매크로(VBA)를 이용하여 프로그램을 작성한다. 그러나 이러한 통계패키지

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

들은 모두 상업용 통계패키지이다. 즉, 사용자가 돈을 지불하고 이 들 통계패키지를 구입하여야 한다는 것이다. 그러나 R은 공개용(GPL, General Public License) 통계적 분석도구이다. 즉 무료라는 것이다. R의 사용이 무료이기 때문에 R을 사용한 통계분석이나 교육과정 개발은 개별 학생이나 교사, 더 나아가 단위학교나 교육청에서 얼마든지 권장하거나 시행할 수 있다. '확률과 통계' 수학교과서(김원경외 5인, 2004)를 보면 15p에 '통계처리 및 분석이 가능한 어떤 통계 소프트웨어를 이용하여' 라는 문장이 나온다. 이 소프트웨어는 Excel을 지칭하는 데 상업용 소프트웨어이기 때문에 교과서에서 소프트웨어 이름을 직접 명시할 수가 없다. 반면 R은 공개소프트웨어이기 때문에 이러한 고민을 할 필요가 없게 된다. R은 다양한 통계 기법들과 우수한 그래픽 기능을 제공하며 행렬 계산을 위한 도구로서도 사용될 수 있다는 장점을 가지고 있는 반면 안정성이 부족하고 대용량의 자료를 분석하기에는 부적당하다는 단점도 지적되고 있으나 초·중·고등학교 확률 및 통계교육에서는 이러한 단점이 전혀 문제가 되지 않는다.

R은 통계언어인 S 언어를 기반으로 하여 1995년 Auckland 대학 Robert Gentleman과 Ross Ihaka에 의하여 만들어진 통계용 언어이다. 대부분의 통계패키지가 일괄처리방식(batch mode)인 데 비하여 R은 대화식 처리방식(interactive mode)을 따르기 때문에 R console이라는 창에서 프롬프트(>) 다음에 R 명령문을 입력하고 ENTER 키를 치면 명령문이 시행이 되어 그 결과가 그 다음 줄에 바로 나타나게 된다.

통상 기업에서의 자료분석에서는 SAS, SPSS, Minitab 등이 주종을 이루고 연구자들의 자료분석에서는 SAS, SPSS, S-Plus 등이 주로 쓰이나 R이 최근에 폭발적으로 전 세계적인 인기를 얻고 있다. 전 세계 통계학자들의 최근기법들이 R로 작성되어 발표되고 있는 실정이다. 연구용 통계패키지로서 각광을 받던 R이 통계학 교육의 수단으로서 서서히 쓰이기 시작하고 있는 데 최근에 R을 사용한 기초통계학 교재가 출간되기 시작하고 있다. 영어원서로서는 Dalgaard(2002), Venable외 다수(2004), Crawley(2005), Verzani(2005) 등이 있다. 기초통계학 교재가 아닌 좀 더 전문적인 R관련 통계학 책으로는 Fox(2002), Heiberger와 Holland(2004), Everitt(2005), Faraway(2005, 2006), Good(2005), Maindonald와 Braun(2005), Everitt와 Hothorn(2006), Murrell(2006), Pfaff(2006), Shumway와 Stoffer(2006), Wood(2006) 등이 있다. 국내에서는 R 입문서로서는 양경숙과 김미경(2007)이 있고 R을 사용한 기초통계학 교재로서 김연형(2006), 이정남과 김태수(2006)가 있고 기초통계학 교재가 아닌 좀 더 전문적인 R관련 통계학 책으로는 김달호(2005), 박태성외 5인(2005), 유충현외 2인(2005), 허문열외 4인(2005), 박동련(2006), 양경숙과 김미경(2007), 허명회(2007)가 있다.

허명회(2007)는 1장에서 R이라는 통계언어를 배워야 하는 이유를 다음과 같이 세 가지로 열거하고 있다.

첫째, R은 공짜(freeware)이기 때문이다.
둘째, 꽤 좋기 때문이다. R은 팩키지일 뿐만이 아니라 일종의 언어이다.
셋째, 전세계의 1급 연구자들이 개발한 각종 알고리즘들을 인터넷을 통하여 쉽게 획득하여 활용할 수 있다.

저자의 생각에는 첫째와 셋째 이유가 정말로 중요하다고 생각한다. 전 세계는 지금 양극화의 문제로 골머리를 앓고 있다. 양극화의 문제는 돈의 문제 뿐 만이 아니라 지식 소유의 문제에도 해당되는 문제이다. 부자나라나 부자 학교는 패키지의 구입이나 사용에 훨씬 자유롭다. 반면 저개발국이나 개발도상국에서는 인터넷 사용도 버거운 실정이다. 동남아 국가에서 과연 얼마나 많은 학교가 통계소프트웨어를 불법 사용하지 않고 정당한 가격을 주고 사용하고 있을까 하는 생각이 자주 들곤 한다. R은 이런 면에서 너무 매력적이다. 통계패키지의 문제로 제한하여 보면 R은 통계관련 지식소유의 양극화 문제를 해결할 수 있는 좋은 수단이 될 수 있다고 생각한다.

2절에서는 제 7차 수학과 교육과정 내의 확률 및 통계영역 목표와 내용을 중심으로 하고 제 7차 수학과 교육과정에 따라 집필된 '수학 I'(금성출판사외 11종, 2004)과 '실용수학'(교학사외 3종, 2004) 수학교과서들의 확률 및 통계단원 및 '확률과 통계' 수학교과서(김원경외 5인, 2004)를 참고로 하여 R 패키지를 구체적으로 수업에 어떻게 적용할 수 있는 지를 예제들을 통하여 살펴보았다. 3절에서 결론을 맺었다.

Ⅱ. 제 7차 수학과 교육과정에 따른 R의 활용

1. 확률과 통계영역의 목표와 내용

제 7차 수학과 교육과정의 편성·운영 지침에 보면 1학년에서 10학년까지의 10년 동안은 국민 공통기본 교육과정을 편성, 운영하는 것으로 되어 있고, 10단계의 각 단계별로 학기를 단위로 하는 2개의 하위 단계(가, 나)를 설정하여 단계형 수준별 교육 과정을 운영한다. 1-10단계의 국민 공통 기본 교육과정 후 11, 12단계에서는 선택과목으로서 '실용수학', '수학 I', '수학 II', '미분과 적분', '확률과 통계', '이산수학'을 대상으로 선택하도록 하였다. 이 중 확률과 통계영역이 들어가 있는 과목은 '실용수학', '수학 I'과 '확률과 통계'이다. 1-10단계의 국민 공통 기본 교육과정에서는 확률과 통계영역은 6개의 영역(수와 연산, 도형, 측정, 확률과 통계, 문자와 식, 규칙성과 함수) 중 하나이다. 실용수학에서는 4개의 영역(계산기와 컴퓨터, 경제생활, 생활 통계, 생활 문제 해결) 중 하나의 영역, 수학 I에서는 3개의 영역(대수, 해석, 확률과 통계) 중 하나의 영역이다. 제 7차 수학과 교육과정 내의 '수학 I'과 '실용수학'의 확률과 통계영역 및 '확률과 통계' 교과목의 목표 체계와 내용 체계를 보면 모두 '실생활 문제'를 강조하고 있고 복잡한 계산이나 문제 해결을 위하여 계산기와 컴퓨터를 적극적으로 활용하도록 하고 있다.

2. R의 활용

제 7차 수학과 교육과정 '수학 I'과 '실용수학'의 확률과 통계영역 및 '확률과 통계' 교과목의 목표와

내용을 중심으로 하고 제 7차 수학과 교육과정에 따라 집필된 '수학 I'과 '실용수학' 수학교과서들의 확률 및 통계단원 그리고 '확률과 통계' 수학교과서를 참고로 하여 R 패키지를 활용하는 방법을 예제들을 통하여 살펴보자.

<예제 1> (자료의 정리와 요약) 다음은 밀가루에 있는 동(copper) 성분 함량의 측정값(단위: ppm) 24개를 나타내는 자료이다.

2.20 2.20 2.40 2.40 2.50 2.70 2.80 2.90 3.03 3.03 3.10 3.37 3.40 3.40 3.40 3.50
3.60 3.70 3.70 3.70 3.70 3.77 5.28 28.95

이 자료에 대하여 입력을 행하고 다양한 기술통계량(descriptive statistic)들을 구하여 보자.

(풀이) R에서 취급하는 자료를 R에서는 자료객체(data object)라고 한다. 자료객체에는 벡터(vector), 행렬(matrix), 배열(array), 리스트(list), 데이터프레임(data frame), 범주형 자료(factor), 시계열 자료(time series) 총 7 종류가 있다. R에서 취급하는 자료객체의 기본은 벡터이다.

```
> # 데이터의 입력
> copper=c(2.20, 2.20, 2.40, 2.40, 2.50, 2.70, 2.80, 2.90, 3.03, 3.03, 3.10, 3.37,
+  3.40, 3.40, 3.40, 3.50, 3.60, 3.70, 3.70, 3.70, 3.70, 3.77, 5.28, 28.95)
> copper
 [1]  2.20  2.20  2.40  2.40  2.50  2.70  2.80  2.90  3.03  3.03  3.10  3.37
[13]  3.40  3.40  3.40  3.50  3.60  3.70  3.70  3.70  3.70  3.77  5.28 28.95
> # 데이터의 개수
> length(copper)
[1] 24
> # copper의 정렬(sort)
> sort(copper)
 [1]  2.20  2.20  2.40  2.40  2.50  2.70  2.80  2.90  3.03  3.03  3.10  3.37
[13]  3.40  3.40  3.40  3.50  3.60  3.70  3.70  3.70  3.70  3.77  5.28 28.95
> sort(copper, decreasing=T)
 [1] 28.95  5.28  3.77  3.70  3.70  3.70  3.70  3.60  3.50  3.40  3.40  3.40
[13]  3.37  3.10  3.03  3.03  2.90  2.80  2.70  2.50  2.40  2.40  2.20  2.20
> # table
> table(copper)
```

```
copper
 2.2  2.4  2.5  2.7  2.8  2.9  3.03  3.1  3.37  3.4  3.5  3.6  3.7  3.77  5.28  28.95
   2    2    1    1    1    1     2    1     1    3    1    1    4     1     1     1
> # (copper에 대한 수치적 측도)
> # 합
> sum(copper)
[1] 102.73
> # 누적합(cumulative sum)
> cumsum(copper)
 [1]   2.20   4.40   6.80   9.20  11.70  14.40  17.20  20.10  23.13  26.16
[11]  29.26  32.63  36.03  39.43  42.83  46.33  49.93  53.63  57.33  61.03
[21]  64.73  68.50  73.78 102.73

> # 산술평균
> mean(copper)
[1] 4.280417
> sum(copper)/length(copper)
[1] 4.280417
> # 중앙값
> median(copper)
[1] 3.385
> # 10% 절사평균
> mean(copper,trim=1/10)
[1] 3.205
> # 분산
> var(copper)
[1] 28.06240
> sum((copper-mean(copper))^2)/(length(copper)-1)
[1] 28.06240
> # 표준편차
> sd(copper)
[1] 5.297396
> sqrt(var(copper))
[1] 5.297396

> # 다섯숫자 요약
> # (최소값, 제 1사분위수, 중앙값, 제 3사분위수, 최대
값)
> fivenum(copper)
[1]  2.200  2.750  3.385  3.700 28.950
> # 다섯숫자 요약과 같음
> quantile(copper)
    0%    25%    50%    75%   100%
 2.200  2.775  3.385  3.700 28.950
> # 수치적측도 요약(다섯숫자 요약+산술평균)
> summary(copper)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.200   2.775   3.385   4.280   3.700  28.950
> # 사분위수간범위=제 3사분위수 - 제 1사분위수
> IQR(copper)
[1] 0.925
> # MAD(Median Absolute Deviation):
> # 각 데이터에서 중앙값을 뺀 후 절대값을
> # 취한 값들의 중앙값
> mad(copper)
[1] 0.526323
```

```
> median(abs(copper-median(copper)))*1.4826
[1] 0.526323
> # 최대값
> max(copper)
[1] 28.95
> # 최소값
> min(copper)
[1] 2.2
> # 범위
# R에서는 범위를 호출하면 최소값과 최대값을 출력함
> range(copper)
[1]  2.20 28.95
> RANGE=max(copper)-min(copper)
> RANGE
[1] 26.75
```

<예제 2> (자료의 정리와 요약) 한국고용정보원이 발표한 2006년 직업지도에 보면 보건, 의료관련 직업에서 여자비율을 보면 다음과 같다. (단위: %)

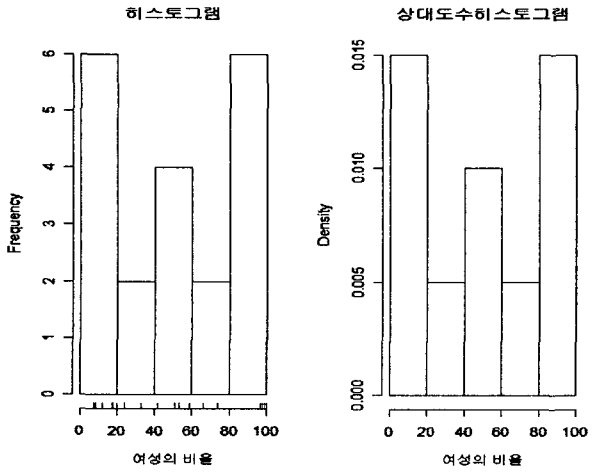
18, 7, 24, 8, 59, 99, 74, 51, 66, 17, 99, 12, 20, 42, 98, 100, 33, 97, 97, 53

(a) 도수분포표를 만들고 이를 이용하여 히스토그램을 그려라.
(b) 줄기와 잎 그림을 그려라.

(풀이) (a) 도수분포표와 그에 대응하는 히스토그램, 그리고 상대도수히스토그램을 그려보면 다음과 같다.

```
> # (도수분포표)
> # 자료의 입력
> prop.woman=c(18, 7, 24, 8, 59, 99, 74, 51, 66, 17, 99, 12, 20, 42, 98, 100, 33, 97, 97, 53)
> prop.woman
 [1]  18   7  24   8  59  99  74  51  66  17  99  12  20  42  98 100  33  97  97
[20]  53
> # 자료의 갯수
> n=length(prop.woman)
> n
[1] 20
> # 5개의 계급으로 나누기
> cat.job=cut(prop.woman, breaks=c(0, 20, 40, 60, 80, 100))
> cat.job
 [1] (0,20]   (0,20]   (20,40]  (0,20]   (40,60]  (80,100] (60,80]  (40,60]
 [9] (60,80]  (0,20]   (80,100] (0,20]   (0,20]   (40,60]  (80,100] (80,100]
```

```
[17] (20,40]  (80,100] (80,100] (40,60]
Levels: (0,20] (20,40] (40,60] (60,80] (80,100]
> # 도수분포표
> table(cat.job)
cat.job
  (0,20]  (20,40]  (40,60]  (60,80] (80,100]
     6       2        4        2        6
> levels(cat.job)=c("0-20%", "20-40%", "40-60%", "60-80%", "80-100%")
> table(cat.job)
cat.job
  0-20%  20-40%  40-60%  60-80% 80-100%
      6       2       4       2       6
> # x축에 자료가 표시되는 히스토그램
> hist(prop.woman, breaks=c(0, 20, 40, 60, 80, 100), main="여성의 비율에 대한 히스토그램", xlab="여성의 비율")
> rug(jitter(prop.woman))
> # 상대도수히스토그램
> hist(prop.woman, breaks=c(0, 20, 40, 60, 80, 100), probability=T,
+ main="여성의 비율에 대한 상대도수히스토그램", xlab="여성의 비율")
```



(b) 줄기와 잎 그림을 그리면 다음과 같다.

```
> # 줄기와 잎 그림
> stem(prop.woman)

  The decimal point is 1 digit(s) to the right of the |
   0 | 78278
   2 | 043
   4 | 2139
   6 | 64
   8 | 77899
  10 | 0
```

<예제 3> (이산분포) R에서는 특정 확률분포에 해당하는 이름이 있고 그 이름에 따른 인자(argument)가 있다. 각 확률분포의 이름 앞에 특정한 첫 글자가 나타나면 특정 함수를 실행하게 된다(d: density (확률질량함수), p: probability(누적확률), q: quantile(분위수), r: random(확률난수)).

1. 교통 사망사고 중 60%가 음주운전 때문이라고 하자. 3건의 교통 사망사고를 조사하였을 때 확률변수 X를 음주운전으로 인한 사망 사건의 수라하면 X는 이항분포 B(3,0.6)를 이룬다.

- 우선 팩토리얼과 조합을 구하여 보자. 3!, 100!, ${}_3C_1$을 각각 구하여 보면 다음과 같다.

```
> # 팩토리알
> factorial(3)
[1] 6
> factorial(100)
[1] 9.332622e+157
```

```
> # 조합
> choose(3,1)
[1] 3
> factorial(3)/(factorial(2)*factorial(1))
[1] 3
```

- 이항분포 B(3,0.6)에서 확률분포표, 누적확률, 확률난수(확률난수 값은 시행할 때마다 달라진다.) 등을 다음과 같이 구하여 보자.

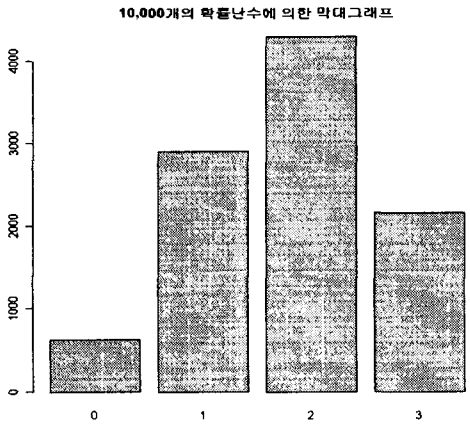
```
> # 이항분포 B(3,0.6)에서의 확률질량함수
> x=c(0,1,2,3)
> p.x=dbinom(0:3,size=3,prob=0.6)
```

```
> x
[1] 0 1 2 3
> p.x
```

```
[1] 0.064 0.288 0.432 0.216
> dbinom(0:3,3,0.6)
[1] 0.064 0.288 0.432 0.216
> # 이항분포표 B(3,0.6)
> names(p.x)=c("0","1","2","3")
> p.x
    0     1     2     3
0.064 0.288 0.432 0.216
> # 이항분포 B(3,0.6)에서의 이항계수
> choose(3,0:3)
[1] 1 3 3 1
> # 이항분포 B(3,0.6)에서의 확률질량함수
> choose(3,0:3)*(0.6)^(0:3)*(1-0.6)^(3:0)
[1] 0.064 0.288 0.432 0.216
> dbinom(0:3,3,0.6)
[1] 0.064 0.288 0.432 0.216
> # 이항분포 B(3,0.6)에서의 Pr[X<=2]
> pbinom(2,3,0.6)
[1] 0.784
> sum(dbinom(0:2,3,0.6))
[1] 0.784
> # 이항분포 B(3,0.6)에서의 Pr[X=2]
> dbinom(2,3,0.6)
[1] 0.432
> pbinom(2,3,0.6)-pbinom(1,3,0.6)
[1] 0.432
> # 이항분포 B(3,0.6)에서의 누적분포함수
> # (cumulative distribution function) Pr[X<=x]
> cdf.x=pbinom(x,3,0.6)
> names(cdf.x)=c("0","1","2","3")
> cdf.x
    0     1     2     3
0.064 0.352 0.784 1.000
> # 이항분포 B(3,0.6)에서의 난수 5개
> sample1=rbinom(5,3,0.6)
> sample1
[1] 2 1 2 1 2
```

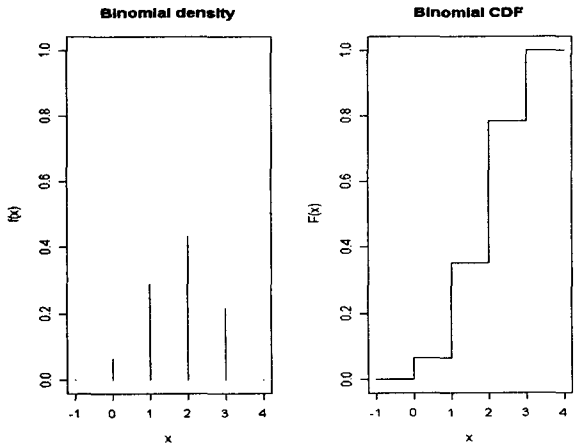
- 이항분포 B(3,0.6)에서의 평균과 분산을 구하여 보자. 또한 이항분포 B(3,0.6)에서 확률난수 10,000개를 뽑아(확률난수 값은 시행할 때마다 달라진다.) 이 10,000개로 이루는 분포가 이항분포 B(3,0.6)를 따르는 지 확인하여 보자. 이 10,000개로 이루는 분포의 평균과 분산이 이항분포 B(3,0.6)에서의 평균과 분산과 아주 유사함을 알 수 있다. 이 10,000개로 이루는 분포를 막대그래프로 그려 볼 수 있다. 이항분포 B(3,0.6)에서의 확률질량함수와 비교하면 유사하게 따라감을 알 수 있다.

```
> # 이항분포 B(3,0.6)에서의 평균 3*0.6=1.8,
> # 분산 3*0.6*0.4=0.72
> e.x=sum(x*px)
> e.x
[1] 1.8
> var.x=sum((x-e.x)^2*px)
> var.x
[1] 0.72
> sample2=rbinom(10000,3,0.6)
> mean(sample2)
[1] 1.7996
> var(sample2)
[1] 0.7187117
> table(sample2)
sample2
   0    1    2    3
 628 2912 4296 2164
> barplot(table(sample2), main="10,000개의
확률난수에 의한 막대그래프")
```



● 이항분포 B(3,0.6)와 누적분포함수 Pr[X<=x]를 그려보면 다음과 같다. plot 함수에서 type="h"란 수직으로 그리라는 것이고 type="s"란 값을 만날 때마다 수평으로 간 후 그 다음에 수직으로 가라는 것이다.

```
> # 이항분포 B(3,0.6)와 누적분포함수 Pr[X<=x]
> par(mfrow=c(1,2))
> x=-1:4
> y=dbinom(x,3,0.6)
> plot(x,y,type="h",xlab="x",ylab="f(x)",ylim=c(0,1.0))
> title("Binomial density")
> z=pbinom(x,3,0.6)
> plot(x,z,type="s",xlab="x",ylab="F(x)",ylim=c(0,1.0))
> title("Binomial CDF")
```

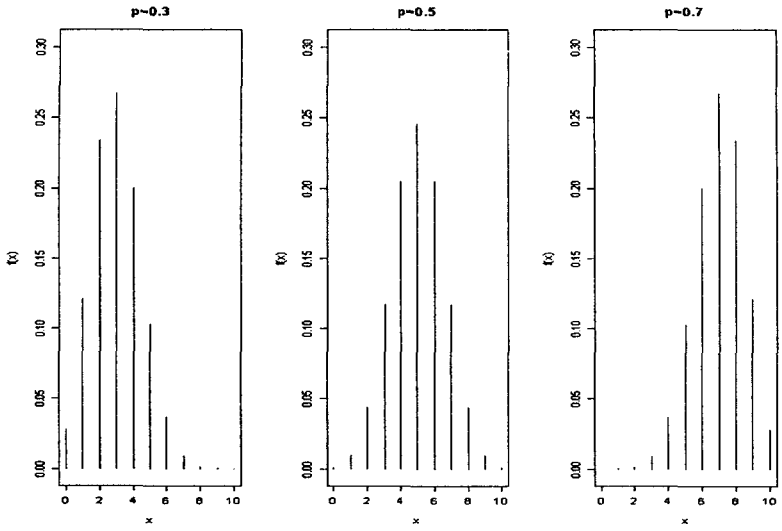


2. 동전을 100번 던지는 시행을 이항분포를 이용하여 시행하여 보자. '1'이면 앞면, '0'이면 뒷면이다. 동전의 앞면이 나올 확률을 구하여 보면 시행순서가 커지면서 0.5에 다가감을 알 수 있다.

```
> # 동전던지기
> rbinom(100,1,0.5)
  [1] 0 1 0 0 0 0 1 1 1 1 1 0 0 1 0 0 1 1 0 1 0 0 1 1 0 1 0 1 0 1 0 1 1 1 0 1
 [38] 1 1 1 0 1 1 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0
 [75] 0 1 0 0 1 0 1 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 1 0 0
> # 동전의 앞면이 나올 확률
> n=length(rbinom(100,1,0.5))
> cumsum(rbinom(100,1,0.5)/n)
  [1] 0.00 0.01 0.01 0.01 0.02 0.02 0.02 0.02 0.03 0.04 0.05 0.06 0.06 0.07 0.07
 [16] 0.07 0.08 0.08 0.08 0.08 0.09 0.10 0.11 0.12 0.13 0.14 0.15 0.15 0.15 0.16
 [31] 0.16 0.17 0.18 0.18 0.19 0.19 0.19 0.19 0.20 0.20 0.20 0.21 0.21 0.22 0.22
 [46] 0.22 0.23 0.23 0.23 0.24 0.24 0.25 0.25 0.25 0.25 0.26 0.27 0.28 0.29 0.30
 [61] 0.30 0.30 0.30 0.31 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.39 0.39
 [76] 0.40 0.41 0.41 0.41 0.42 0.42 0.43 0.44 0.44 0.45 0.46 0.46 0.46 0.46 0.47
 [91] 0.48 0.48 0.49 0.49 0.50 0.50 0.50 0.51 0.51 0.52
```

3. 시행횟수가 10이고 성공률 $p = 0.3,\ 0.5,\ 0.7$일 때의 이항분포의 변화에 대하여 알아보자.

```
> # 성공률의 변화에 따른 이항분포의 변화
> par(mfrow=c(1,3))
> x=c(0:10)
> y1=dbinom(x,10,0.3)
> plot(x,y1,type="h",xlab="x",ylab="f(x)",ylim=c(0,0.3))
> title("p=0.3")
> y2=dbinom(x,10,0.5)
> plot(x,y2,type="h",xlab="x",ylab="f(x)",ylim=c(0,0.3))
> title("p=0.5")
> y3=dbinom(x,10,0.7)
> plot(x,y3,type="h",xlab="x",ylab="f(x)",ylim=c(0,0.3))
> title("p=0.7")
```



4. 어떤 제조업체에서는 한 상자에 25개의 제품을 집어넣는다. 상자별로 검사를 시행하는 데 상자에서 4개를 랜덤하게 뽑아 제품을 검사한 후 불량품의 개수가 없을 때 이 상자를 합격시킨다고 한다.
(a) 원래 불량품의 개수가 25개 중 5개일 때 상자를 합격할 확률을 구하라.
(b) 원래 불량품의 개수가 25개 중 $r(0 \leq r \leq 25)$개일 때 합격할 확률을 그림으로 그려 보아라.

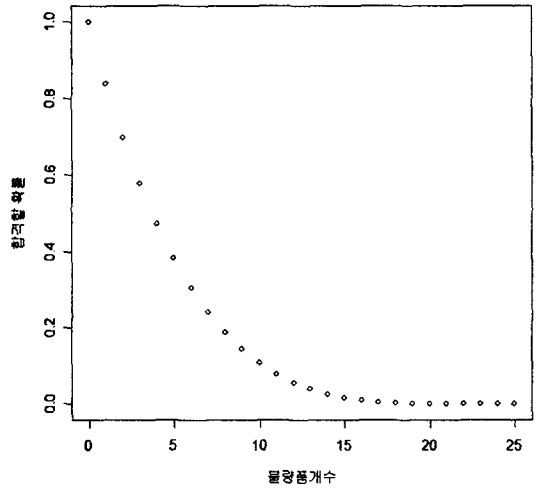
(풀이) (a)

```
> # 초기하분포(불량품개수=5,양품개수=20,검사개수=4)
> accept.box=dhyper(0,5,20,4)
> accept.box
[1] 0.3830040
```

(b)

```
> # 초기하분포(불량품개수=r,양품개수=25-r,검사개수=4)
> par(mfrow=c(1,1))
> r=c(0:25)
> accept.box.r=dhyper(0,r,25-r,4)
> accept=cbind(r,accept.box.r)
> accept
      r accept.box.r
 [1,] 0 1.000000e+00
 [2,] 1 8.400000e-01
```

```
[3,]  2 7.000000e-01
[4,]  3 5.782609e-01
[5,]  4 4.731225e-01
[6,]  5 3.830040e-01
[7,]  6 3.064032e-01
[8,]  7 2.418972e-01
[9,]  8 1.881423e-01
[10,] 9 1.438735e-01
[11,] 10 1.079051e-01
[12,] 11 7.913043e-02
[13,] 12 5.652174e-02
[14,] 13 3.913043e-02
[15,] 14 2.608696e-02
[16,] 15 1.660079e-02
[17,] 16 9.960474e-03
[18,] 17 5.533597e-03
[19,] 18 2.766798e-03
[20,] 19 1.185771e-03
[21,] 20 3.952569e-04
[22,] 21 7.905138e-05
[23,] 22 0.000000e+00
[24,] 23 0.000000e+00
[25,] 24 0.000000e+00
[26,] 25 0.000000e+00
> plot(r,accept.box.r,xlab="불량품개수",ylab="합격할 확률")
```



<예제 4> (정규분포)

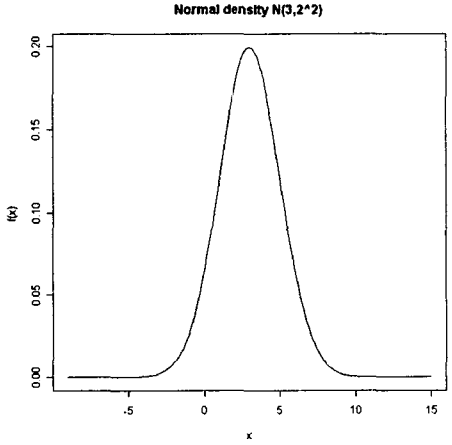
1. (a) $Z \sim N(0, 1)$일 때 $Pr[-1 \le Z \le 1]$, $Pr[-2 \le Z \le 2]$, $Pr[-3 \le Z \le 3]$을 각각 구하라.
   (b) 평균이 3이고 표준편차가 2인 정규분포를 그려라.

(풀이) (a) $Pr[-1 \le Z \le 1] = 0.683$, $Pr[-2 \le Z \le 2] = 0.954$, $Pr[-3 \le Z \le 3] = 0.997$

```
> # 표준정규분포에서의 확률 구하기
> pnorm(1)-pnorm(-1)
[1] 0.6826895
> pnorm(2)-pnorm(-2)
[1] 0.9544997
> pnorm(3)-pnorm(-3)
[1] 0.9973002
```

(b) 평균이 3이고 표준편차가 2인 정규분포를 그리면 다음과 같다.

```
> # 평균이 3이고 표준편차가 2인 정규분포
> n.mean=3;n.sd=2
> x=seq(n.mean-6*n.sd,n.mean+6*n.sd,length=200)
> y=dnorm(x=x,mean=3,sd=2)
```




```
> plot(x,y,type="l",ylab="f(x)",main="Normal density N(3,2^2)")
```

2. (a) $Z \sim N(0,1)$일 때 $Pr[Z \leq z] = 0.995, 0.975, 0.99, 0.95, 0.9$인 z를 각각 구하라.
(b) 평균이 3이고 표준편차가 2인 정규분포 $X \sim N(3,2^2)$에서 $Pr[X \leq x] = 0.975$인 x를 구하라.

(풀이) (a) z은 각각 2.58, 1.96, 2.33, 1.645, 1.282가 된다.

(b) $Pr[X \le x] = Pr[\frac{X-3}{2} \le \frac{x-3}{2}] = Pr[Z \le \frac{x-3}{2}] = 0.975,\ \therefore \frac{x-3}{2} = 1.96 \therefore x = 6.92$

```
> # 표준정규분포에서의 분위수(quantile)
> qnorm(0.995)
[1] 2.575829
> qnorm(0.975)
[1] 1.959964
> qnorm(0.99)
[1] 2.326348
> qnorm(0.95)
[1] 1.644854
> qnorm(0.9)
[1] 1.281552
> # 정규분포에서의 분위수
> qnorm(p=0.975,mean=3,sd=2)
[1] 6.919928
> qnorm(0.975,3,2)
[1] 6.919928
```

3. (이항분포의 정규근사) $X \sim B(15, 0.4)$에서
(a) $Pr[X=6]$을 구하라. 이 확률을 이항분포의 정규근사를 이용하여 구하라.
(b) $Pr[8 \le X \le 10]$을 구하라. 이 확률을 이항분포의 정규근사를 이용하여 구하라.

(풀이) (a) $X \sim B(15, 0.4)$에서 $\mu = np = 15 \times 0.6 = 6,\ \sigma^2 = npq = 15 \times 0.4 \times 0.6 = 3.6,$
$\sigma = \sqrt{3.6} = 1.897,\ Pr[X=6] = Pr[X \le 6] - Pr[X \le 5] = 0.610 - 0.403 = 0.207$
이 확률을 이항분포의 정규근사를 이용하여 구하면

$$Pr[X=6] = Pr[5.5 \le X \le 6.5] = Pr\left[\frac{5.5-6}{1.897} \le \frac{X-6}{1.897} \le \frac{6.5-6}{1.897}\right]$$
$$= Pr[-0.26 \le Z \le 0.26] = Pr[Z \le 0.26] - Pr[Z \le -0.26] = -0.6026 - 0.3974 = 0.2052$$

(b) $Pr[8 \le X \le 10] = Pr[X \le 10] = Pr[X \le 7] = 0.991 - 0.787 = 0.204$
이 확률을 이항분포의 정규근사를 이용하여 구하면

$$Pr[8 \le X \le 10] = Pr[X \le 10] = Pr[7.5 \le X \le 10.5] = Pr\left[\frac{7.5-6}{1.897} \le \frac{X-6}{1.897} \le \frac{10.5-6}{1.897}\right]$$
$$= Pr[0.79 \le Z \le 2.37] = Pr[Z \le 2.37] - Pr[Z \le 0.79] = 0.991 - 0.7852 = 0.2059$$

```
> # 이항분포의 정규근사
> # Pr[X=6]
> pbinom(6,15,0.4)-pbinom(5,15,0.4)
[1] 0.2065976
> pnorm(0.26)-pnorm(-0.26)
[1] 0.2051362
> # Pr[8<=X<=10]
> pbinom(10,15,0.4)-pbinom(7,15,0.4)
[1] 0.2037555
> pnorm(2.37)-pnorm(0.79)
[1] 0.2058698
```

4. (a) 표준정규분포 $Z \sim N(0,1)$에서 확률난수를 5개 뽑으시오.

(b) 평균이 3이고 표준편차가 2인 정규분포 $X \sim N(3,2^2)$에서 확률난수를 5개 뽑으시오.

(풀이) 의사난수발생기(pseudo-random number generator)를 이용할 때마다 확률난수는 달라진다.

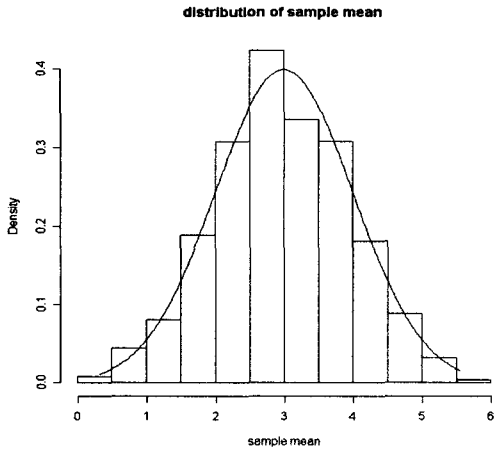
```
> # 표준정규분포로부터 5개의 난수 추출
> rnorm(5)
[1] 2.5178708 0.7331323 0.8809315 0.3972354 0.5009881
> # 정규분포로부터 5개의 난수 추출
> rnorm(n=5,mean=3,sd=2)
[1] 2.734831 2.874476 4.425200 3.261234 4.933814
> rnorm(5,3,2)
[1] 2.210338 2.716910 2.367497 1.895204 3.279857
```

<예제 5> (표본추출분포) 1. (a) 평균이 3이고 표준편차가 2인 정규분포 $X \sim N(3,2^2)$에서 확률난수를 4개를 뽑아 표본평균을 구하라. 이런 과정을 500번 시행한 후 표본평균의 분포를 히스토그램으로 그려라. 무엇을 알 수 있나?
(b) 평균이 3이고 표준편차가 2인 정규분포 $X \sim N(3,2^2)$에서 확률난수를 4개를 뽑아 표본분산을 구하라. 이런 과정을 500번 시행한 후 표본평분산의 분포를 히스토그램으로 그려라. 무엇을 알 수 있나?

(풀이) (a) 평균이 3이고 표준편차가 2인 정규분포 $X \sim N(3,2^2)$에서 확률난수를 4개를 뽑아 표본평균을 구하는 과정을 500번 시행한 후 표본평균의 분포를 히스토그램으로 그리면 다음과 같다. 이론적으로는 표본평균의 기대값이 $E(\overline{X}) = \mu = 3$이고 표본평균의 표준편차가 $Var(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{4}} = 1$이 되어야 하나 확률난수가 완전한 난수가 아니어서 2.970과 0.985가 나왔다. 히스토그램과 같이 그린 그림은 평균이 3이고 표준편차가 1인 정규분포이다. 표본평균의 분포가 평균이 3이고 표준편차가 1인 정규분포가 됨을 알 수 있다.

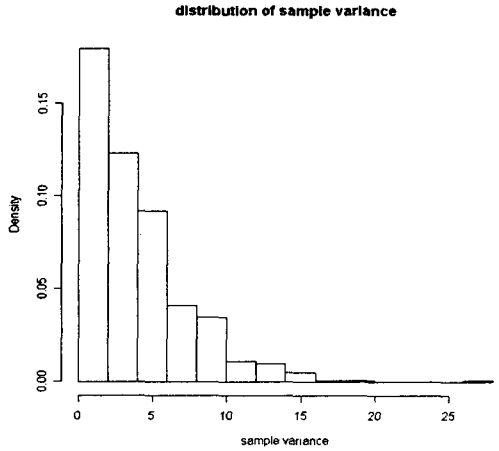
```
> # (표본평균의 분포)
> # 평균이 3이고 표준편차가 2인 정규분포에서
> # 표본의 크기가 4인 확률표본을 500번 구하기
> r.mean=rep(0,500);r.var=rep(0,500)
> for(i in seq(500)) {
+ r=rnorm(n=4,mean=3,sd=2)
+ r.mean[i]=mean(r);r.var[i]=var(r)
+ }
> # 표본평균의 기대값과 분산
> mean(r.mean)
[1] 2.97032
> var(r.mean)
[1] 0.97072
> sd(r.mean)
[1] 0.9852513
> # 표본평균의 분포
> hist(r.mean,prob=TRUE,main="distribution of
sample mean",xlab="sample mean")
```

```
> x1=seq(min(r.mean),max(r.mean),length=200)
> y1=dnorm(x=x1,mean=3,sd=1)
> lines(x1,y1)
```



(b) 평균이 3이고 표준편차가 2인 정규분포 $X \sim N(3,2^2)$에서 확률난수를 4개를 뽑아 표본분산 $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$을 구하는 과정을 500번 시행한 후 표본분산의 분포를 히스토그램으로 그리면 다음과 같다. 이론적으로는 표본평균의 기대값이 $E(S^2) = \sigma^2 = 4$가 되어야 하나 확률난수가 완전한 난수가 아니어서 3.947이 나왔다. 표본분산의 분포가 매우 오른쪽으로 치우친(skewed to the right) 분포가 됨을 알 수 있다.

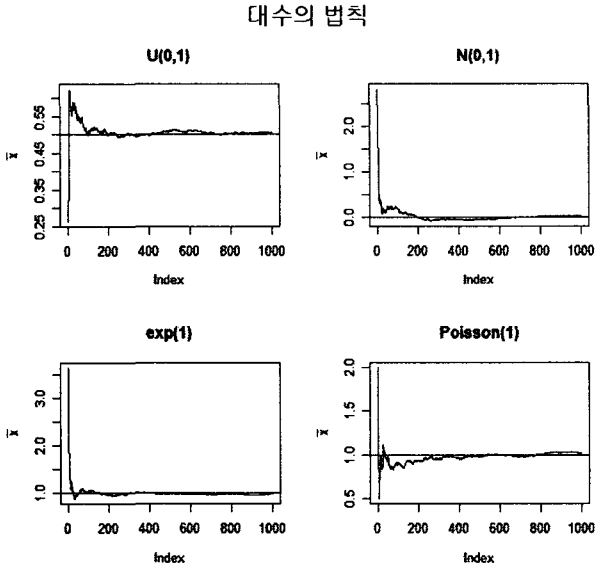
```
> # 표본분산의 기대값과 분산
> mean(r.var)
[1] 3.946905
> var(r.var)
[1] 11.81551
> # 표본분산의 기대값과 분산
> mean(r.var)
[1] 3.946905
> var(r.var)
[1] 11.81551
> # 표본분산의 분포
> hist(r.var,prob=TRUE,main="distribution of
sample variance",xlab="sample variance")
```



2. (대수의 법칙) 모집단이 각각 균등분포, 표준정규분포, 지수분포, 포아송분포일 때 대수의 법칙이 성립함을 모의실험을 통하여 보여라.

(풀이)

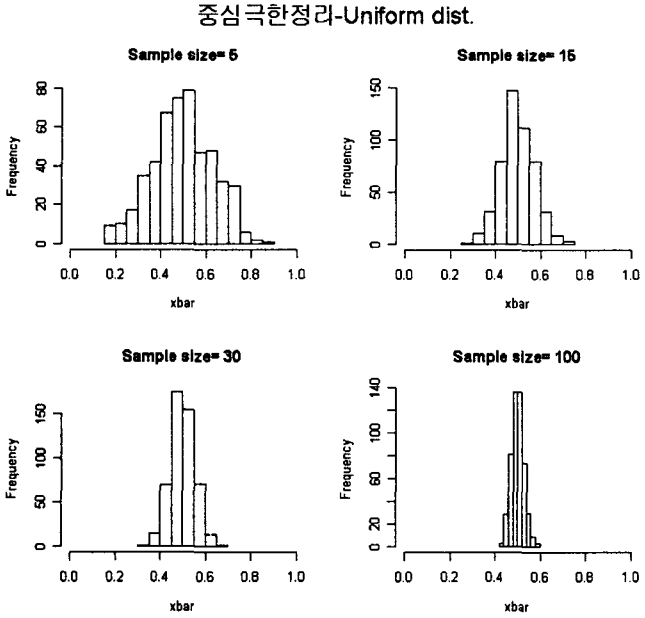
```
> # 대수의 법칙
> law.large.number=
+ function(n)
+ {
+ win.graph()
+ par(oma=c(0,0,5,0))
+ par(mfrow=c(2,2))
+ x1=runif(n)
+ xbar1=cumsum(x1)/1:n
+ plot(xbar1,ylab=expression(bar(x))
,type="l",main="U(0,1)")
+ abline(0.5, 0)
+ x2=rnorm(n)
+ xbar2=cumsum(x2)/1:n
+ plot(xbar2,ylab=expression(bar(x))
,type="l",main="N(0,1)")
+ abline(0, 0)
+ x3=rexp(n,1)
+ xbar3=cumsum(x3)/1:n
+ plot(xbar3,ylab=expression(bar(x))
,type="l",main="exp(1)")
+ abline(1, 0)
+ x4=rpois(n,1)
+ xbar4=cumsum(x4)/1:n
+ plot(xbar4,ylab=expression(bar(x))
,type="l",main="Poisson(1)")
+ abline(1, 0)
+ mtext("대수의 법칙",side=3,outer=T,cex=1.5)
+ par(mfrow=c(1,1))
+ }
> law.large.number(1000)
```



3. (중심극한정리) 모집단이 균등분포일 때 중심극한정리가 성립함을 모의실험을 통하여 보여라.

(풀이) 모집단이 균등분포 $f(x)=1(0 \le x \le 1)$일 때 다음과 같이 중심극한정리가 성립함을 모의실험을 통하여 보일 수 있다. 표본의 크기가 커짐에 따라 점점 표본평균의 분포가 정규분포가 됨을 알 수 있고 평균 0.5를 중심으로 점점 집중됨을 알 수 있다.

```
> # 중심극한정리-균등분포
> par(oma=c(0,0,5,0))
> par(mfrow=c(2,2))
> central.Uniform<-
+ function(a,b)
+ {
+     nt <- c(5, 15, 30, 100)
+     xbar <- rep(0,500)
+     for(i in 1:4)
+     {
+         for(j in 1:500)
+         {
+         xbar[j] <- sum(runif(nt[i],a,b))/nt[i]
+         }
+                     hist(xbar,main=paste("Sample
                       size=",nt[i]),xlim=c(0,1))
+     }
+                     mtext("중심극한정리-Uniform
                      dist.",side=3,outer=T,cex=1.5)
+ }
> central.Uniform(0,1)
```

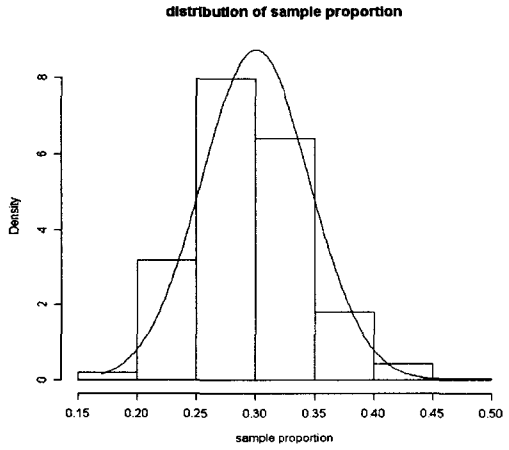


4. 모비율이 $p=0.3$인 임의의 모집단에서 표본의 크기를 100개를 뽑아 원하는 속성을 갖고 있는 것의 개수를 세고 표본비율을 구하라. 이런 과정을 500번 시행한 후 표본비율의 분포를 히스토그램으로 그려라. 무엇을 알 수 있나?

(풀이) 모비율이 $p=0.3$인 임의의 모집단에서 표본의 크기를 100개를 뽑아 원하는 속성을 갖고 있는 것의 개수를 세고 표본비율을 구하는 과정을 500번 시행한 후 표본비율의 분포를 히스토그램으로 그리면 다음과 같다. 이론적으로는 표본비율의 기대값이 $E(\hat{p})=p=0.3$이고 표본비율의 표준편차가 $sd(\hat{p})=\sqrt{\frac{pq}{n}}=\sqrt{\frac{0.3\times 0.7}{100}}\approx 0.0458$이 되어야 하나 확률난수가 완전한 난수가 아니어서 0.3002와 0.0474가 나왔다. 히스토그램과 같이 그린 그림은 평균이 0.3이고 표준편차가 0.0458인 정규분포이다. 표본비율의 분포가 평균이 0.3이고 표준편차가 0.0458인 정규분포가 됨을 알 수 있다.

```
> # (표본비율의 분포)
> # 성공률 p=0.3인 모집단에서
> # 표본의 크기가 100인 확률표본을 500번 구하기
> r.mean=rep(0,500);r.var=rep(0,500)
> for(i in seq(500))
+ {
```

```
+ r=rbinom(n=100,1,0.3)
+ r.mean[i]=mean(r)
+ }
> # 표본비율의 기대값과 분산
> mean(r.mean)
[1] 0.30022
> var(r.mean)
[1] 0.002249851
> sd(r.mean)
[1] 0.0474326
> # 표본비율의 분포
> h=hist(r.mean,plot=F)
>
ylim=range(0,h$density,dnorm(0.3,mean=0.3,sd=0.0458))
> hist(r.mean,prob=TRUE,ylim=ylim
,main="distribution of sample proportion",xlab="sample
proportion")
> x1=seq(min(r.mean),max(r.mean),length=200)
> y1=dnorm(x=x1,mean=0.3,sd=0.0458)
> lines(x1,y1)
```



<예제 6> (구간추정) 1. 다음 자료는 40명에 대하여 심장병을 줄이기 위한 한 주당 육체 훈련 양을 분 단위로 조사한 값이다.

60,40,50,30,60,50,90,30,60,60,60,80,90,90,60,30,20,120,60,50,20,60,30,120,50,30,90,20,30,40,50,40,30,40,20,30,60,50,60,80

한 주당 평균적인 육체 훈련 양에 대한 95% 신뢰구간을 구하라.

(풀이) 표본의 크기가 40개이어서 대표본이므로 다음과 같이 정규분포를 이용한 95% 신뢰구간을 구하면 (45.54, 61.46)이다.

```
> # 모평균에 대한 추정(대표본의 경우)
> one.sample.z.confidence.interval=function(x,
confidence.level)
+ {
+ n=length(x)
+ xbar=mean(x)
+ se=sd(x)/sqrt(n)
+ alpha.half=(1-confidence.level)/2
+ z.alpha.half=qnorm(1-alpha.half)
+ c(xbar-z.alpha.half*se,xbar+z.alpha.half*se)
+ }
>
x=c(60,40,50,30,60,50,90,30,60,60,60,80,90,90,60,30,20,120,6
0,50
+,20,60,30,120,50,30,90,20,30,40,50,40,30,40,20,30,60,50,60,8
0)
> one.sample.z.confidence.interval(x,0.95)
[1] 45.54323 61.45677
```

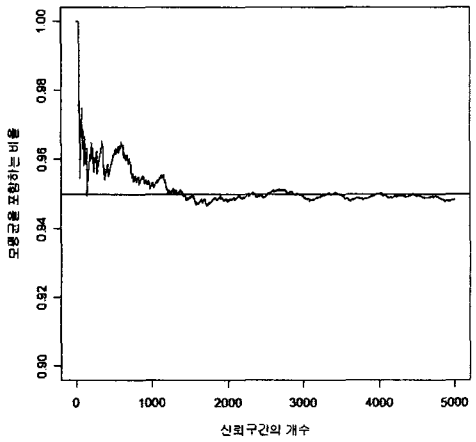
이 경우 t분포를 이용한 95% 신뢰구간을 다음과 같이 구하면 (45.29, 61.71)이다. 정규분포를 이용한 신뢰구간보다 폭이 조금 큰 것을 확인할 수 있다.

```
> t.test(x)

        One Sample t-test

data:  x
t = 13.1785, df = 39, p-value = 6.035e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 45.28858 61.71142
sample estimates:
mean of x
     53.5
```

모평균에 대한 95% 신뢰구간을 구한다는 것은 이러한 신뢰구간을 100개 구하였을 때 95개의 신뢰구간이 모평균을 포함하고 5개의 신뢰구간은 모평균을 포함하지 않는다는 의미이다. 우리는 다음과 같은 시뮬레이션을 통하여 정규분포를 이용한 모평균에 대한 95% 신뢰구간의 의미를 파악할 수 있다. 모집단이 평균이 0이고 표준편차가 1인 정규모집단이라 할 때 표본의 크기가 50개인 신뢰구간의 개수를 5,000개까지 만들어보며 신뢰구간이 모평균 0을 포함하는 비율을 계산하여 보면 다음 그림과 같이 95%에 수렴함을 알 수 있다.

```
> confidence.normal <-
+ function(n, nz, mu, std)
+ {
+    win.graph()
+    trial <- rep(1,nz)
+    z.val <- qnorm(0.975)
+    x <-matrix(rnorm(n*nz, mu, std), nrow=nz)
+    xbar <- apply(x,1,mean)
```

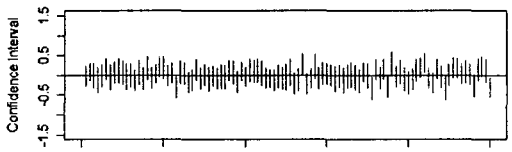
```
+   xvar <- apply(x,1,var)
+   limit <- z.val * sqrt(xvar/n)
+   xbar <- xbar - mu
+   trial[abs(xbar) > limit] <- 0
+   trial <- cumsum(trial)/ 1:nz
+   plot(trial,ylim=c(0.9,1),type="l"
,xlab="신뢰구간의 개수",ylab="모평균을 포함하는 비율
")
+   abline(0.95, 0)
+ }
> confidence.normal(50,5000,0,1)
```

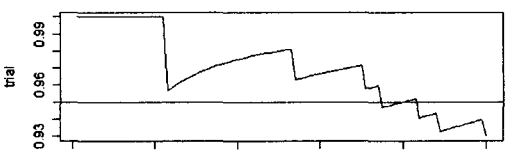


우리는 다음과 같은 시뮬레이션을 통하여 정규분포를 이용한 모평균에 대한 95% 신뢰구간의 의미를 다시 한 번 파악할 수 있다. 모집단이 평균이 0이고 표준편차가 1인 정규모집단이라 할 때 신뢰구간의 개수를 100개까지 만들어보며 신뢰구간이 모평균 0을 포함하는 지의 여부와 그 비율을 계산하여 보면 다음 그림과 같이 95%에 수렴함을 대략 알 수 있다.

```
> confidence.normal2 <-
+ function(n, nz, mu, std)
+ {
+   win.graph()
+   par(mfrow=c(2,1))
+   trial <- rep(1,nz)
+   z.val <- qnorm(0.975)
+   x <-matrix(rnorm(n*nz, mu, std), nrow=nz)
+   xbar <- apply(x,1,mean)
+   xvar <- apply(x,1,var)
+   limit <- z.val * sqrt(xvar/n)
+   ulimit=xbar+limit
+   llimit=xbar-limit
+   xbar <- xbar - mu
+   trial[abs(xbar) > limit] <- 0
+   trial <- cumsum(trial)/ 1:nz
+
plot(c(0,nz),c(-1.5,1.5),type="n",xlab="index",ylab="Confi
dence Interval")
+   for (k in 1:nz){
+   points(c(k,k),c(llimit[k],ulimit[k]),type="l")
+   }
```

```
+   abline(0, 0)
+   plot(trial,type="l")
+   abline(0.95, 0)
+   par(mfrow=c(1,1))
+ }
> confidence.normal2(50,100,0,1)
```





2. 휘발유의 옥탄가(정규분포로 가정함.)를 13일 연속 조사하니 다음과 같았다.

88.6 86.4 87.2 88.4 87.2 87.6 86.8 86.1 87.4 87.3 86.4 86.6 87.1

옥탄가 모평균에 대한 95% 신뢰구간을 구하라.

(풀이) 표본의 크기가 13개이어서 소표본이므로 t분포를 이용한 95% 신뢰구간을 다음과 같이 구하면 (86.71, 87.61)이다.

```
> # 모평균에 대한 구간추정(소표본의 경우)
> x=c(88.6,86.4,87.2,88.4,87.2,87.6,86.8,86.1,87.4,87.3,86.4,86.6,87.1)
> t.test(x)

        One Sample t-test

data:  x
t = 423.4101, df = 12, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 86.71302 87.61006
sample estimates:
mean of x
 87.16154
```

3. 어느 지역의 실업률을 조사하기 위하여 1,000명을 조사하니 15명이 실업자이었다.
모실업률에 대한 95% 신뢰구간을 구하라.

(풀이) 표본의 크기가 1,000개이어서 대표본이므로 다음과 같이 정규분포를 이용한 95% 신뢰구간을 구하면 (0.0087, 0.0252)이다.

```
> # 모비율에 대한 구간추정
> prop.test(15,1000)

        1-sample proportions test with continuity correction

data:  15 out of 1000, null probability 0.5
X-squared = 938.961, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.008733248 0.025217456
sample estimates:
    p
0.015
```

Ⅲ. 맺는 말

초·중·고등학교 확률 및 통계교육은 지필교육과 더불어 컴퓨터를 통한 자료분석이 필수적으로 이루어져야 한다. 우리는 통계패키지로서 R을 사용할 수 있다. R은 대화식 처리방식을 따르기 때문에 배우기가 쉽다(물론, 명령어의 숙지라는 조건이 붙는다.). 또한 R에서의 그래픽스는 아주 강력하고 통계적 모의실험도 쉽게 행할 수 있다. 가장 큰 장점은 R의 사용이 무료라는 것이다. 그러므로 R의 사용에 따른 사용료나 저작권료에 대하여 고민할 필요가 없다. 이러한 많은 장점을 갖고 있는 R을 초·중·고등학교 확률 및 통계교육 현장에서 표준 통계패키지로서 정착시킬 필요가 있다고 본다. 초·중·고등학교 확률 및 통계교육에서는 주로 Excel을 사용한다. 그러나 통상 기업에서의 자료분석에서는 SAS, SPSS, Minitab 등이 주종을 이루고 연구자들의 자료분석에서는 SAS, SPSS, S-Plus 등이 주로 쓰이고 있으나 R이 최근에 폭발적으로 전 세계적인 인기를 얻고 있다. 즉 자료분석의 입장에서는 Excel이 통계패키지 사용의 연속성을 가지지 못한다는 것이다. 반면 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 사료된다.

참 고 문 헌

교육인적자원부 (1997). 제 7차 수학과 교육과정.

교학사외 3종 (2004). 실용수학 수학교과서 4종.

금성출판사외 11종 (2004). 수학 Ⅰ수학교과서 12종.

김달호 (2005). R과 WinBUGS를 이용한 베이지안 통계학, 자유아카데미.

김연형 (2006). 통계학의 개념과 응용-R 프로그램 활용, 교우사.

김원경 · 박석윤 · 이성덕 · 황선영 · 정상일 · 이종학 (2004). 확률과 통계, 교육인적자원부.

박동련 (2006). R에 의한 통계 그래픽스, 자유아카데미.

박태성 · 이승연 · 김기웅 · 이성곤 · 최호식 · 윤단규 (2005). 마이크로어레이 자료의 통계적 분석, 자유아카데미.

양경숙 · 김미경 (2007). R입문 및 기초 프로그래밍, 자유아카데미.

양경숙 · 김미경 (2007). R을 활용한 회귀분석, 자유아카데미.

유충현 · 이상호 · 김정일 (2005). R 그래픽스, 자유아카데미.

이정남 · 김태수 (2006). 통계학(R활용), 자유아카데미.

허명희 (2007). R을 사용한 탐색적 자료분석, 자유아카데미.

허문열 · 이승천 · 차경준 · 박종선 · 유종영 (2005). R & 통계계산, 박영사.

Crawley, M. J. (2005). *Statistics: An Introduction using R*. Hoboken: John Wiley.

Dalgaard, P. (2002). *Introductory Statistics with R*, New York: Springer.

Everitt, B. S. (2005). *An R and S-Plus Companion to Multivariate Analysis*, London: Springer.

Everitt, B. S. & Hothorn, T. (2006). *A Handbook of Statistical Analyses using R*, Boca Raton: Chapman & Hall/CRC.

Faraway, J. J. (2005). *Linear Models with R*, Boca Raton: Chapman & Hall/CRC.

Faraway, J. J. (2006). *Extending the Linear Model with R*, Boca Raton: Chapman & Hall/CRC.

Fox, J. (2002). *An R and S-Plus Companion to Applied Regression*, Thousand Oaks: Sage Publications.

Good, P. I. (2005). *Introduction to Statistics through Resampling Methods and R/S-Plus*, Hoboken: John Wiley.

Heiberger, R. M. & Holland, B. (2004). *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS*, New York: Springer.

Maindonald, J. & Braun, J. (2005). *Data Analysis and Graphs using R-An Example-based Approach*, New York: Cambridge University Press.

Murrell, P. (2006). *R Graphics*, Boca Raton: Chapman & Hall/CRC.

Pfaff, B. (2006). *Analysis of Integrated and Cointegrated Time Series with R*, New York: Springer.

Shumway, R. H. & Stoffer, D. S. (2006). *Time Series Analysis and Its Applications with R Examples*, New York: Springer.

Venables, W. N.; Smith, D. M. & the R Development Core Team (2004). *An Introduction to R*, Bristol: Network Theory Ltd.

Verzani, J. (2005). *Using R for Introductory Statistics*, Boca Raton: Chapman & Hall/CRC.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*, Boca Raton: Chapman & Hall/CRC.

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

Jang, Dae-Heung

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

E-mail: dhjang@pknu.ac.kr

We can use R package as a statistical package on the education of probability and statistics in elementary, middle and high school mathematics. R is an interactive mode package and graphical presentation tools in R are powerful. The greatest advantage is that R is a general public license package. We need to consider R package as a standard statistical package on the education of probability and statistics 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