응용통계연구 (2009)
22(5), 893–910

기초통계학 추측통계영역 교육시 R의 활용에 대한 연구

장대흥[1]

[1]부경대학교 수리과학부 통계학전공

(2009년 4월 접수, 2009년 5월 채택)

요 약

대학교 기초통계학 교육 시 우리는 통계패키지로서 R을 사용할 수 있다. R은 대화식 처리방식을 따르기 때문에 실행결과를 즉시 볼 수 있다. 또한 R에서의 그래픽스는 아주 강력하다. 그리고 가장 큰 장점은 R의 사용이 무료라는 것이다. 이러한 많은 장점을 갖고 있는 R을 대학교 기초통계교육 현장에서 사용하는 표준 통계패키지로서 고려할 필요가 있다. 본 논문은 기초통계학 내용 중 추측통계 영역을 대상으로 R의 활용에 대하여 기술하고자 한다.

주요용어: 기초통계학 교육, R 패키지.

1. 서론

기초통계학 강의 중 추측통계 영역은 기술통계 영역에 비하여 상대적으로 많은 비중을 차지한다. 예로 김진경 등 (2008)이 저술한 기초통계학 교재를 보면 기술통계 영역이 전체 분량의 25%, 추측통계 영역이 75%를 차지한다. 이처럼 기초통계학 내용 중 추측통계 영역의 비중은 매우 크고 통계모형을 다루는 영역이기 때문에 학생들이 기초통계학을 배우기 어려워하는 시발점이 된다. 이러한 추측통계 영역에서 R을 어떻게 활용하면 학생들이 낙오되지 않고 강사의 강의를 잘 이해하고 통계학의 여러 개념들을 습득할 수 있는 지를 살피고자 한다.

기초통계학 내용 중 기술통계 영역뿐만이 아니라 추측통계 영역에서도 R은 유용한 통계패키지가 될 수 있다. R은 많은 다양한 분포에 대하여 확률밀도함수, 누적분포함수, 분위수, 확률변수를 제공한다. 또한 우리는 기초통계학에서 언급되는 통계기법들을 모두 R로 구현하여 볼 수 있다.

본 논문은 기초통계학 내용 중 추측통계 영역을 대상으로 R의 활용에 대하여 기술하고자 한다. 2절에서는 저자 본인의 강의 경험을 바탕으로 추측통계 영역에서 R을 어떻게 활용할 수 있는 지에 대하여 생각하여 보고 3절에서는 표본추출분포부터 검정까지 예제들과 이 예제들을 풀기 위한 R 프로그램을 제시하였다. 본 논문에서 제시하지 않은 기술통계영역(표와 자료를 통한 자료의 요약, 수치를 통한 연속형자료의 요약, 두 변수 자료의 요약), 추측통계영역인 확률, 이항분포와 그에 관련된 분포들, 정규분포, 두 모집단의 비교, 회귀분석 부분은 저자의 홈페이지(http://myweb.pknu.ac.kr/daeheung) 내 강의자료실에 pdf 파일로 탑재되어 있다. 4절에서 결론을 맺었다.

2. R의 활용

본 저자는 2007년 1학기와 2학기에 각 1반씩 기초통계학 강의에서 R을 활용하여 보았다. 매 학기 강의는 두 모집단의 비교까지 진도가 나갔다. 교재로는 김진경 등 (2008)이 저술한 기초통계학 교재를 사용

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

하였다. 이 교재는 통계소프트웨어로서 엑셀을 사용한다. 엑셀의 사용은 학생 개개인들이 사용해보도록 권장하고 엑셀과 R을 비교하여 보도록 하였다.

R을 사용하기 위해서는 명령어들의 문법을 일일이 알아야하는 번거로움이 있어 학생들이 R을 사용하는 데 걸림돌이 된다. 이를 완화시키기 위하여 R Commander (Fox, 2005 참조)를 학생들에게 소개하고 R과 병행하여 사용하도록 권장하였다. R Commander는 2009년 4월 6일 현재 버전 1.4-7이 개발되어 있다. 기초통계학 강의에서 각 장의 내용을 가르치며 예제들을 R로 구현하여 학생들에게 제시하였다. R을 기본으로 하고 보조수단으로서 R Commander에 교수데모(teaching demos('RcmdrPlugin.TeachingDemos'))가 들어가 있는 R Commander를 사용하였다. 교수데모가 들어가 있는 R Commander는 다음과 같은 특징이 있다.

1. R Commander상의 9개의 주 메뉴(File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, Help) 외에 'Demos'라는 주메뉴가 있고 이 'Demos' 메뉴 하에 7개의 데모용 부메뉴(Central limit theorem(중심극한정리 데모), Confidence interval for the mean(모평균에 대한 신뢰구간 데모), Power of the test(검정력 데모), Flip a coin(동전던지기 데모), Roll a die(주사위굴리기 데모), Simple linear regression(단순선형회귀분석 데모), Simple correlation(상관계수 데모))가 있다. 중심극한정리 데모를 이용하면 학생들이 중심극한 정리의 의미를 파악할 수 있고, 모평균에 대한 신뢰구간 데모를 통해서는 학생들이 모평균에 대한 신뢰구간의 의미를 알 수 있고, 검정력 데모를 이용하면 학생들이 제1종 오류와 제2종 오류와의 관계를 쉽게 파악할 수 있고, 동전던지기 데모와 주사위굴리기 데모를 이용하면 동전던지기와 주사위굴리기를 시뮬레이션할 수 있다. 단순선형회귀분석 데모를 통하여 학생들이 회귀분석의 의미를 알 수 있고, 상관계수 데모를 통하여 학생들이 상관계수의 의미를 파악할 수 있다. 이 모든 데모들이 동적그래픽스로 구현된다.
2. 주 메뉴 'Distributions' 하에 연속형과 이산형 분포들 각각에 대하여 분위수, 누적분포함수, 분포(확률질량함수나 확률밀도함수)의 그림, 분포에 따른 난수를 구하는 부메뉴들 외에 'Visualize distributions'라는 부메뉴가 있다. 4가지 분포(이항분포, 정규분포, t-분포, 감마분포)에 대하여 모수들을 조정해가며 분포들의 변하는 모습을 실시간으로 동적그래픽스로 구현하여 볼 수 있다.

기초통계학 내용 중 강사들이 가르치는 추측통계학 영역은 다음과 같다.

1. 확률
2. 이항분포와 그에 관련된 분포들
3. 정규분포
4. 표본추출분포
5. 구간추정
6. 검정
7. 두 모집단의 비교
8. 회귀분석

1학기 분량의 기초통계학 강의는 통상 두 모집단의 비교까지 진도가 나가며 조금 진도가 빠른 경우라면 회귀분석도 취급할 수 있다.

3. R의 활용 예

예제들은 김진경 등 (2008)이 저술한 기초통계학 교재에 나오는 예들을 중심으로 사용하였다. 기초통계학 내용 중 추측통계 영역에 대해서 기술하고자 한다.

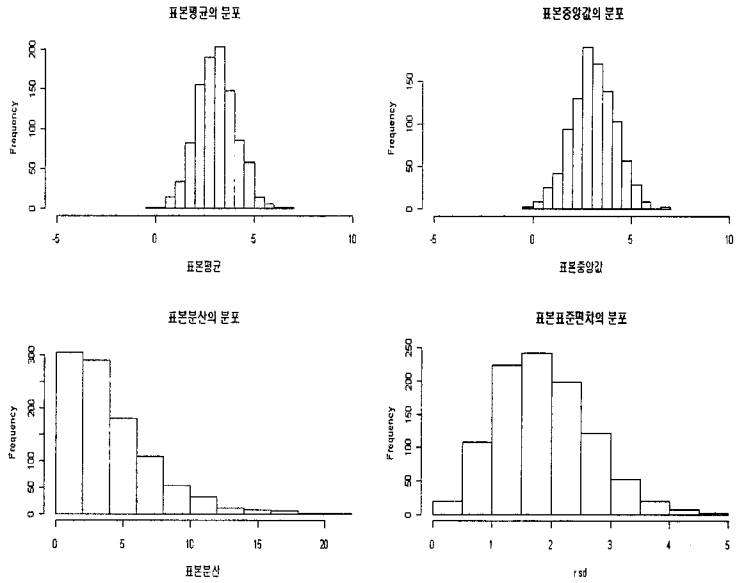

그림 3.1. 정규분포 $N(3, 2^2)$에서 추출한 표본추출분포

3.1. 표본추출분포

표본추출분포는 표본평균의 분포를 중심으로 전개되나 표본비율과 표본분산의 분포에 대해서도 고려해 볼 필요가 있다. 중심극한정리는 모집단의 분포가 이산적이든 연속적이든, 대칭적이든 비대칭적이든 상관없이 평균과 분산이 존재하면 표본의 크기가 30개 이상일 때 표본평균의 분포가 근사적으로 정규분포를 이룬다는 정리임으로 이 사실을 R을 이용한 컴퓨터시뮬레이션을 실행하여 각인시킨다.

예제 3.1: (a) 대칭분포인 평균이 3이고 표준편차가 2인 정규분포에서 표본을 4개 뽑아 표본평균, 표본중앙값, 표본분산, 표본표준편차를 구하는 과정을 1,000번 시행한 후 표본평균, 표본중앙값, 표본분산, 표본표준편차의 분포를 히스토그램으로 나타내어 보고 $E(\overline{X}) = \mu$, $\mathrm{Var}(\overline{X}) = \sigma^2/n$이고 $E(S^2) = \sigma^2$임을 계산하여 보아라. (b) 자유도가 5인 카이제곱분포는 평균이 5이고 분산이 10이 된다. 또한, 이 카이제곱분포는 전형적인 오른쪽으로 치우친 비대칭분포이다. 이 카이제곱분포에서 표본을 4개 뽑아 표본평균, 표본중앙값, 표본분산, 표본표준편차를 구하는 과정을 1,000번 시행한 후 표본평균, 표본중앙값, 표본분산, 표본표준편차의 분포를 히스토그램으로 나타내어 보고 $E(\overline{X}) = \mu$, $\mathrm{Var}(\overline{X}) = \sigma^2/n$이고 $E(S^2) = \sigma^2$임을 계산하여 보아라.

풀이: (a) 다음 그림 3.1은 평균이 3이고 표준편차가 2인 정규분포에서 표본을 4개 뽑아 표본평균, 표본중앙값, 표본분산, 표본표준편차를 구하는 과정을 1,000번 시행한 후 표본평균, 표본중앙값, 표본분산, 표본표준편차의 분포를 히스토그램으로 나타낸 그림이다. 표본의 크기에 관계없이 표본평균의 분포는 대칭분포가 된다. 표본표준편차의 분포는 오른쪽으로 치우친 분포, 표본분산은 오른쪽으로 매우 치우친 분포가 된다.

그림 3.1을 그리기 위한 R 스크립트는 다음과 같다,

```
# (표본평균, 표본중앙값, 표본분산, 표본표준편차의 분포)
```

```
# 평균이 3이고 표준편차가 2인 정규분포에서
# 표본의 크기가 4인 확률표본을 1000번 구하기
r.mean=rep(0,1000);r.median=rep(0,1000);r.var=rep(0,1000);r.sd=rep(0,1000)
for(i in seq(1000))
r=rnorm(n=4,mean=3,sd=2)
r.mean[i]=mean(r);r.median[i]=median(r);r.var[i]=var(r);r.sd[i]=sd(r)
par(mfrow=c(2,2))
hist(r.mean,xlim=c(-5,10),xlab="표본평균",main="표본평균의 분포")
hist(r.median,xlim=c(-5,10),xlab="표본중앙값",main="표본중앙값의 분포")
hist(r.var,xlab="표본분산",main="표본분산의 분포")
hist(r.sd,main="표본표준편차의 분포")
```

이렇게 구한 통계량들의 평균, 분산, 표준편차를 구하면 다음과 같다. 난수에 따른 계산상의 작은 차이를 무시하면 $E(\overline{X}) = \mu$, $\mathrm{Var}(\overline{X}) = \sigma^2/n$이고 $E(S^2) = \sigma^2$임을 알 수 있다. 또한, 표본평균의 분산이 표본중앙값의 분산보다 작음을 알 수 있고, $E(S) < \sigma$임을 알 수 있다.

```
> # 표본평균, 표본중앙값, 표본분산, 표본표준편차의 기대값, 분산, 표준편차
> mean(r.mean);mean(r.median);mean(r.var);mean(r.sd)
[1] 3.055991
[1] 3.043938
[1] 4.097572
[1] 1.865406
> var(r.mean);var(r.median);var(r.var);var(r.sd)
[1] 0.9890213
[1] 1.214045
[1] 11.08310
[1] 0.6184521
> sd(r.mean);sd(r.median);sd(r.var);sd(r.sd)
[1] 0.9944955
[1] 1.101837
[1] 3.32913
[1] 0.7864173
```

(b) 다음 그림 3.2는 자유도가 5인 카이제곱분포에서 표본을 4개 뽑아 표본평균, 표본중앙값, 표본분산, 표본표준편차를 구하는 과정을 1,000번 시행한 후 표본평균, 표본중앙값, 표본분산, 표본표준편차의 분포를 히스토그램으로 나타낸 그림이다. 표본이 4개임에도 불구하고 표본평균의 분포는 대칭분포가 되려고 노력하고 있음을 알 수 있고 표본표준편차의 분포는 오른쪽으로 치우친 분포, 표본분산은 오른쪽으로 매우 치우친 분포가 된다.

그림 3.2를 그리기 위한 R 스크립트는 다음과 같다.

```
# (표본평균, 표본중앙값, 표본분산, 표본표준편차의 분포)
# 자유도가 5인 카이제곱분포에서
# 표본의 크기가 4인 확률표본을 1000번 구하기
```

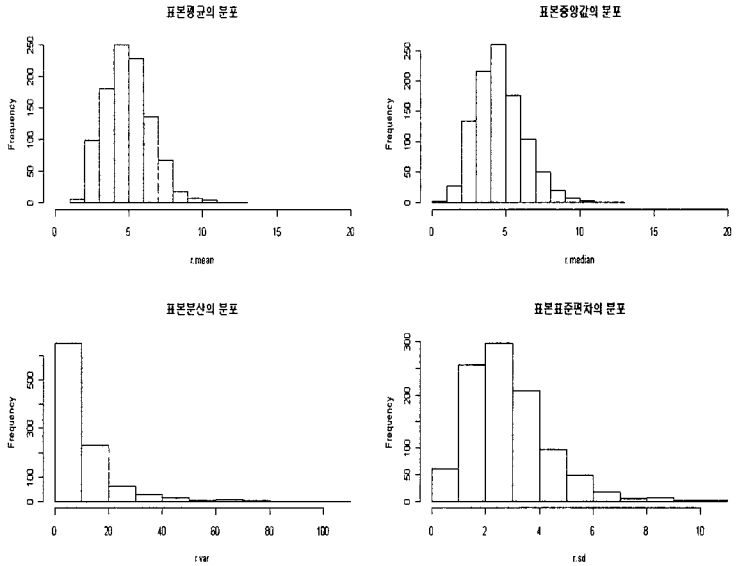

그림 **3.2.** 자유도가 5인 카이제곱분포에서 추출한 표본추출분포

```
r.mean=rep(0,1000);r.median=rep(0,1000);r.var=rep(0,1000);r.sd=rep(0,1000)
for(i in seq(1000))
r=rchisq(n=4,df=5)
r.mean[i]=mean(r);r.median[i]=median(r);r.var[i]=var(r);r.sd[i]=sd(r)
par(mfrow=c(2,2))
hist(r.mean,xlim=c(0,20),main="표본평균의 분포")
hist(r.median,xlim=c(0,20),main="표본중앙값의 분포")
hist(r.var,main="표본분산의 분포")
hist(r.sd,main="표본표준편차의 분포")
```

이렇게 구한 통계량들의 평균, 분산, 표준편차를 구하면 다음과 같다. 난수에 따른 계산상의 작은 차이를 무시하면 $E(\overline{X}) = \mu$, $\mathrm{Var}(\overline{X}) = \sigma^2/n$이고 $E(S^2) = \sigma^2$임을 알 수 있다. 또한, 표본평균의 분산이 표본중앙값의 분산보다 작음을 알 수 있고, $E(S) < \sigma$임을 알 수 있다.

```
> # 표본평균, 표본중앙값, 표본분산, 표본표준편차의 기대값, 분산, 표준편차
> mean(r.mean);mean(r.median);mean(r.var);mean(r.sd)
[1] 4.958888
[1] 4.599006
[1] 10.09648
[1] 2.819139
> var(r.mean);var(r.median);var(r.var);var(r.sd)
[1] 2.495527
[1] 2.687176
[1] 125.3678
```

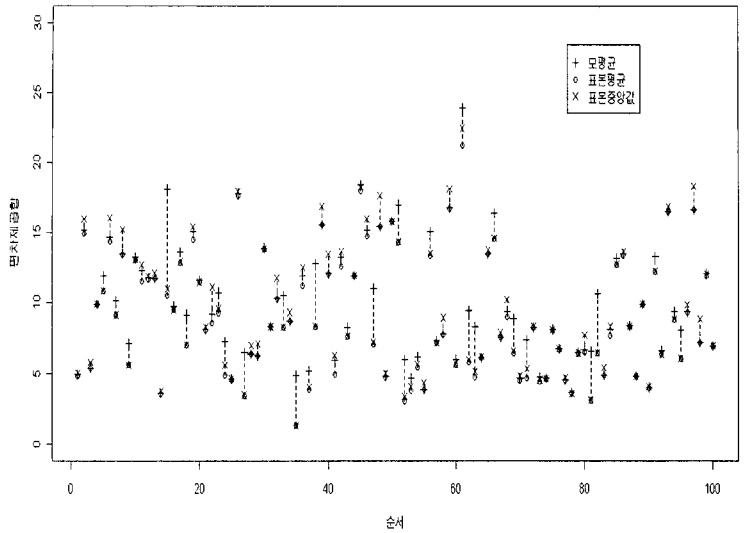

그림 **3.3.** 편차제곱합

[1] 2.151090
> sd(r.mean);sd(r.median);sd(r.var);sd(r.sd)
[1] 1.579724
[1] 1.639261
[1] 11.19678
[1] 1.466659

예제 3.2: 확률표본을 $X_1, X_2, \ldots, X_n$이라 할 때 (a) 편차제곱합 $\sum_{i=1}^{n}(X_i - M)^2$을 최소화하는 값 M은 모평균, 표본평균, 표본중앙값 중 어느 것인가? (b) 편차절대값 $\sum_{i=1}^{n}|X_i - M|$을 최소화하는 값 M은 모평균, 표본평균, 표본중앙값 중 어느 것인가?

풀이: (a) 이를 확인하기 위하여 표준정규분포로부터 10개의 표본을 뽑아 모평균, 표본평균, 표본중앙값을 M으로 하여 편차제곱합을 구한다. 이러한 과정을 100번 시행한 결과가 다음 그림 3.3이다. 100번 시행한 결과 모두 표본평균을 M으로 하여 구한 편차제곱합이 제일 작음을 알 수 있다.

그림 3.3을 그리기 위한 R 스크립트는 다음과 같다.

```
# 편차제곱합
A=matrix(rep(0,300),ncol=3)
plot(0:10,type="n",xlim=c(1,100),ylim=c(0,30),xlab="순서",ylab="편차제곱합")
for(i in 1:100)
x=rnorm(10)
mean.x=mean(x)
median.x=median(x)
A[i,1]=sum((x-0)^2)
A[i,2]=sum((x-mean.x)^2)
```

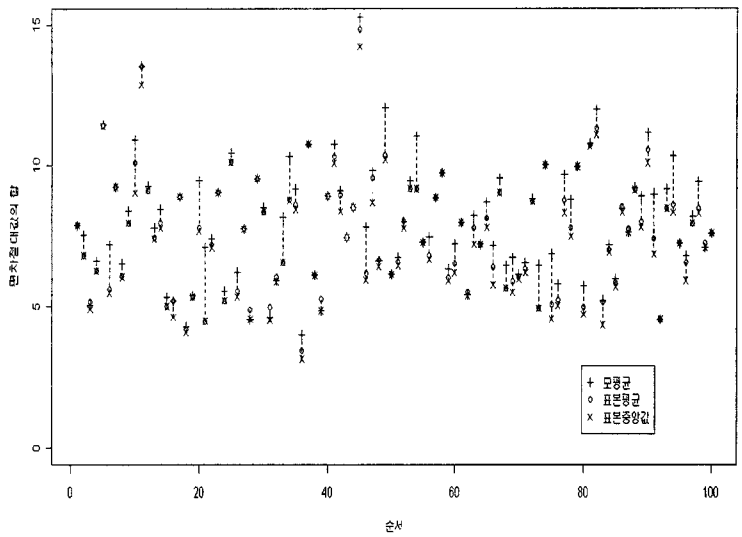

그림 **3.4.** 편차절대값의 합

```
A[i,3]=sum((x-median.x)^2)
points(i,A[i,1],pch=3)
points(i,A[i,2])
points(i,A[i,3],pch=4)
va=c(A[i,1],A[i,2],A[i,3])
lines(c(i,i),c(min(va),max(va)),lty=2)
legend(locator(1),c("모평균","표본평균","표본중앙값"),pch=c(3,1,4))
```

(b) 이를 확인하기 위하여 표준정규분포로부터 10개의 표본을 뽑아 모평균, 표본평균, 표본중앙값을 M으로 하여 편차절대값의 합을 구한다. 이러한 과정을 100번 시행한 결과가 다음 그림 3.4이다. 100번 시행한 결과 모두 표본중앙값을 M으로 하여 구한 편차절대값의 합이 제일 작음을 알 수 있다.

그림 3.4를 그리기 위한 R 스크립트는 다음과 같다.

```
# 편차절대값의 합
B=matrix(rep(0,300),ncol=3)
plot(0:15,type="n",xlim=c(1,100),ylim=c(0,15),xlab="순서",ylab="편차절대값의 합")
for(i in 1:100)
x=rnorm(10)
mean.x=mean(x)
median.x=median(x)
B[i,1]=sum(abs(x-0))
B[i,2]=sum(abs(x-mean.x))
B[i,3]=sum(abs(x-median.x))
points(i,B[i,1],pch=3)
points(i,B[i,2],pch=4)
```

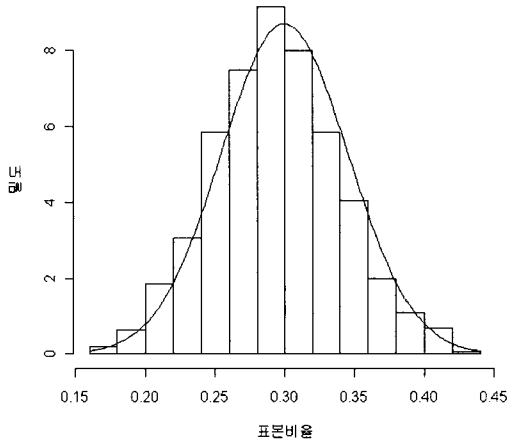

그림 **3.5.** 표본비율의 분포

```
points(i,B[i,3])
vb=c(B[i,1],B[i,2],B[i,3])
lines(c(i,i),c(min(vb),max(vb)),lty=2)
legend(locator(1),c("모평균","표본평균","표본중앙값"),pch=c(3,4,1))
```

예제 3.3: 모비율이 $p = 0.3$인 임의의 모집단에서 표본을 100개 뽑아 원하는 속성을 갖고 있는 것의 개수를 세고 표본비율을 구하라. 이런 과정을 1,000번 시행한 후 표본비율의 분포를 히스토그램으로 그려라. 무엇을 알 수 있나?

풀이: 표본의 크기가 대표본이므로 표본비율의 분포는 그림 3.5에서 보는 것처럼 근사적으로 평균이 $p = 0.3$이고 표준편차가 $\sqrt{(pq)/n} = \sqrt{(0.3 \times 0.7)/100} \approx 0.0458$인 정규분포를 따른다.

그림 3.5를 그리기 위한 R 스크립트는 다음과 같다.

```
# (표본비율의 분포)
# 성공률 p=0.3인 모집단에서
# 표본의 크기가 100인 확률표본을 1000번 구하기
r.mean=rep(0,1000);r.var=rep(0,1000)
for(i in seq(1000))
r=rbinom(n=100,1,0.3)
r.mean[i]=mean(r)
# 표본비율의 기대값과 분산
mean(r.mean)
var(r.mean)
sd(r.mean)
# 표본비율의 분포
h=hist(r.mean,plot=F)
ylim=range(0,h$density,dnorm(0.3,mean=0.3,sd=0.0458))
```

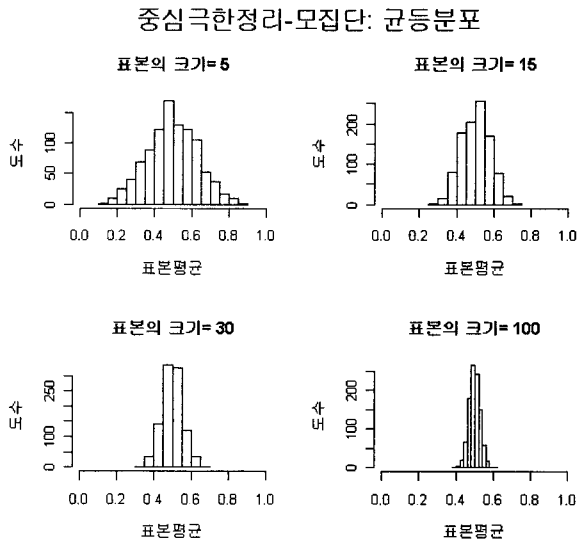

그림 **3.6.** 모집단이 균등분포인 경우의 표본평균의 분포

```
hist(r.mean,prob=TRUE,ylim=ylim,xlab="표본비율",ylab="밀도",main="")
x1=seq(min(r.mean),max(r.mean),length=200)
y1=dnorm(x=x1,mean=0.3,sd=0.0458)
lines(x1,y1)
```

예제 3.4: 모집단이 균등분포일 때 중심극한정리가 성립함을 모의실험을 통하여 보여라.

풀이: 모집단이 균등분포 $f(x) = 1(0 \leq X \leq 1)$일 때 다음 그림 3.6과 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.

그림 3.6을 그리기 위한 R 스크립트는 다음과 같다.

```
# 중심극한정리-모집단: 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(0,1000)
for(i in 1:4)
for(j in 1:1000)
xbar[j] = sum(runif(nt[i],a,b))/nt[i]
hist(xbar,main=paste("표본의 크기=",nt[i]),xlim=c(0,1),xlab="표본평균",ylab="도수")
mtext("중심극한정리-모집단: 균등분포",side=3,outer=T,cex=1.5)
central.Uniform(0,1)
```

3.2. 구간추정

신뢰구간의 의미를 파악하는 일은 매우 중요하다. R을 통하여 이 신뢰구간의 의미를 파악해 볼 수 있다.

예제 3.5: 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 같은 R 스크립트를 사용하여 정규분포를 이용한 95% 근사신뢰구간을 구하면 (45.54, 61.46)가 된다.

```
# 모평균에 대한 구간추정(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(60,40,50,30,60,50,90,30,60,60,60,80,90,90,60,30,20,120,60,50
,20,60,30,120,50,30,90,20,30,40,50,40,30,40,20,30,60,50,60,80)
one.sample.z.confidence.interval(x,0.95)
```

이 경우 t-분포를 이용한 95% 신뢰구간을 다음과 같은 R 스크립트를 사용하여 구하면 (45.29, 61.71)이다. 정규분포를 이용한 신뢰구간보다 폭이 조금 큰 것을 확인할 수 있다.

```
t.test(x)
```

예제 3.6: 모평균에 대한 95% 신뢰구간을 구한다는 것은 이러한 신뢰구간을 100개 구하였을 때 95개의 신뢰구간이 모평균을 포함하고 5개의 신뢰구간은 모평균을 포함하지 않는다는 의미이다. 우리는 다음과 같은 시뮬레이션 (허문열 등 (2005)이 작성한 R 스크립트)을 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 그림 3.7과 같이 95%에 수렴함을 알 수 있다.

```
# 허문열 등 (2005)이 작성한 R 스크립트
confidence.normal =
function(n, nz, mu, std)
trial = rep(1,nz)
z.val = qnorm(0.975)
x = matrix(rnorm(n*nz, mu, std), nrow=nz)
xbar = apply(x,1,mean)
```

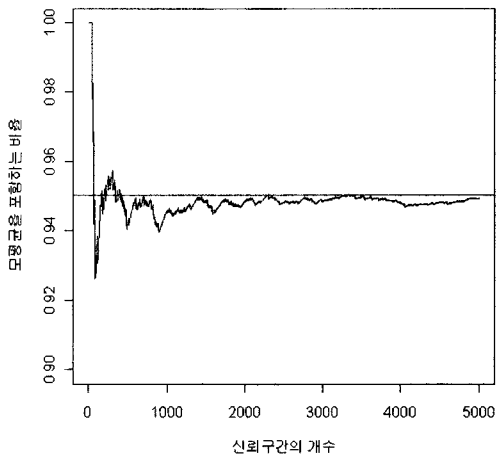

그림 **3.7.** 95% 신뢰구간의 의미(1)

```
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)
```

위의 R 스크립트를 변형한 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 300개까지 만들어보며 신뢰구간이 모평균 0을 포함하는 지의 여부와 그 비율을 계산하여 보면 다음 그림 3.8과 같이 95%에 대략 수렴함을 알 수 있다. 물론 그림 3.7처럼 시행횟수를 많이 하면 95%에 수렴함을 알 수 있다.

```
confidence.normal2 =
function(n, nz, mu, std)
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
```

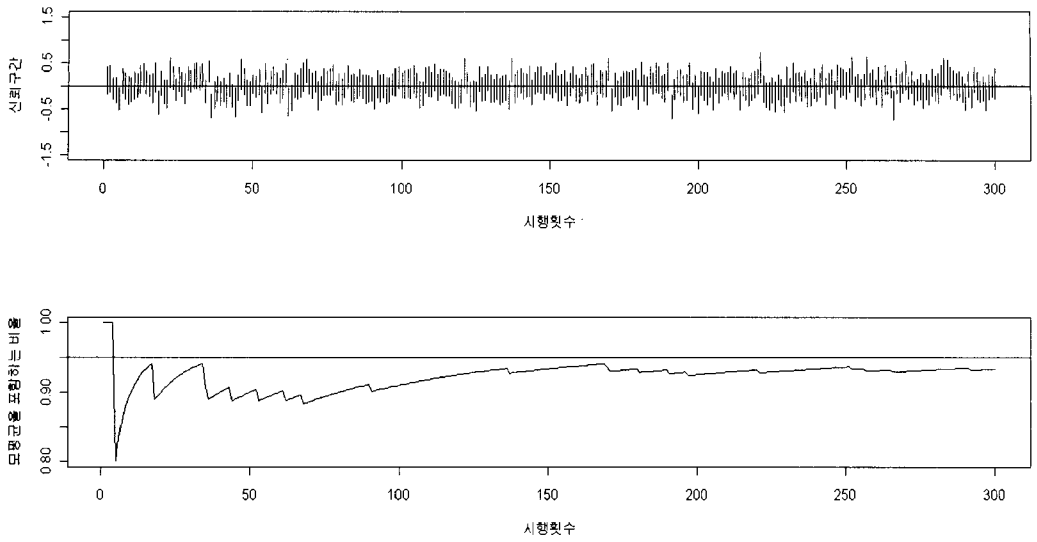

그림 **3.8.** 95% 신뢰구간의 의미(2)

```
trial[abs(xbar) > limit] = 0
trial = cumsum(trial)/ 1:nz
plot(c(0,nz),c(-1.5,1.5),type="n",xlab="시행횟수",ylab="신뢰구간")
for (k in 1:nz)
points(c(k,k),c(llimit[k],ulimit[k]),type="l")
abline(0, 0)
plot(trial,type="l",,xlab="시행횟수",ylab="모평균을 포함하는 비율")
par(mfrow=c(1,1))
confidence.normal2(50,300,0,1)
```

예제 3.7: 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
```

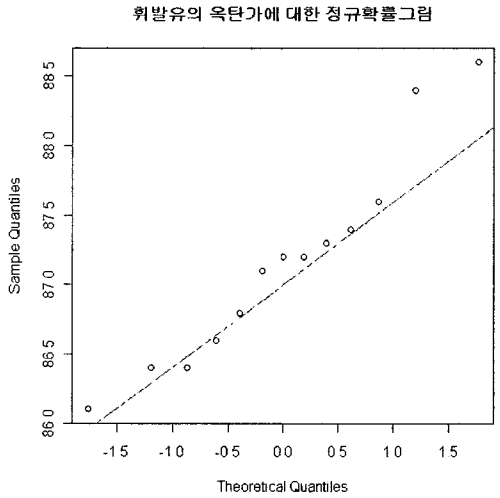

그림 3.9. 휘발유의 옥탄가에 대한 정규확률그림

alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
86.71302 87.61006
sample estimates: mean of x
87.16154

참고로 만일 휘발유의 옥탄가에 대하여 정규분포로 가정한다는 가정이 없다면 우리는 자료가 정규분포를 이루는 지를 검토하여야 한다. 다음 3.9와 같은 정규확률그림을 그리고 Shapiro-Wilk 검정을 행하여 보면 p-값이 0.44로서 정규분포라 할 만하다는 결론을 내릴 수 있다.

```
> qqnorm(x,main="휘발유의 옥탄가에 대한 정규확률그림")
> qqline(x,col="red")
> shapiro.test(x)

Shapiro-Wilk normality test

data: x
W = 0.9386, p-value = 0.4395
```

예제 3.8: 어느 지역의 실업률을 조사하기 위하여 1,000명을 조사하니 15명이 실업자이었다. 모실업률에 대한 95% 신뢰구간을 구하라.

풀이: 피어슨의 카이제곱통계량을 이용한 95% 신뢰구간을 구하면 (0.0087, 0.0252)이다.

```
> # 모비율에 대한 구간추정-피어슨의 카이제곱통계량 이용
> 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

표본의 크기가 1,000개이어서 대표본이므로 다음과 같은 R 스크립트를 사용하여 정규분포를 이용한 95% 근사신뢰구간을 구하면 (0.0075, 0.0225)이다.

```
# 모비율에 대한 구간추정-근사정규분포 이용
one.sample.p.confidence.interval=function(n,x,confidence.level)
p.hat=x/n
se=sqrt(p.hat*(1-p.hat)/n)
alpha.half=(1-confidence.level)/2
z.alpha.half=qnorm(1-alpha.half)
c(p.hat-z.alpha.half*se,p.hat+z.alpha.half*se)
one.sample.p.confidence.interval(1000,15,0.95)
```

3.3. 검정

검정에서 제 1종 오류와 제 2종 오류 사이의 관계, 유의수준의 의미, 양측검정과 단측검정의 차이, p-값의 의미에 대하여 예를 들어가며 학생들에게 자세히 가르친다.

예제 3.9: 검정에서 제 1종 오류와 제 2종 오류 사이의 관계를 알기위하여 교수데모가 들어가 있는 R Commander에서 교수데모 중 검정력 데모를 보면 다음 그림 3.10과 같다. 귀무가설 $H_0 : \mu \leq \mu_0$, 대립가설 $H_1 : \mu > \mu_0$일 때 모표준편차가 σ인 정규모집단에서 표본을 n개 뽑을 때 제 1종 오류의 값을 α로 놓으면 표준오차(standard error)는 $\text{se} = \sigma/\sqrt{n}$, 기각역(rejection region)은 $\overline{X} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$, 검정력(= 1 − 제 2종 오류)는 $r(\mu_1) = \Pr[Z \geq (\mu_0 - \mu_1)/(\sigma/\sqrt{n}) + z_\alpha]$이 된다. 그림 3.10에서는 $\mu_0 = 0$이다. $n = 10$, σ(Standard Deviation로 표시) = 2, $\mu_1 - \mu_0$(True Difference로 표시) = 2, $\alpha = 0.05$로 선택하면 표준오차는 $\text{se} = 0.63$, 기각역에서 $z^\star = \mu_0 + z_\alpha \sigma/\sqrt{n} = 1.04$, 검정력은 $r(2) = 0.935$가 된다. 제 1종 오류를 작게 하면 제 2종 오류가 커지고 제 1종 오류를 크게 하면 제 2종 오류가 작게 되는 반비례 관계가 성립하는 것을 동적 그래픽스로 확인할 수 있다.

예제 3.10: 35세에서 55세까지의 여성 100명에 대한 비만지수(BMI, body mass index)를 조사하니 표본평균이 25.12이고 표본표준편차가 5.3이었다. 비만지수가 25보다 크다고 할 수 있나? 유의수준 5%에서 검정을 행하라.

풀이: 가설은 다음과 같다.

$$H_0 : \mu \leq 25, \qquad H_1 : \mu > 25$$

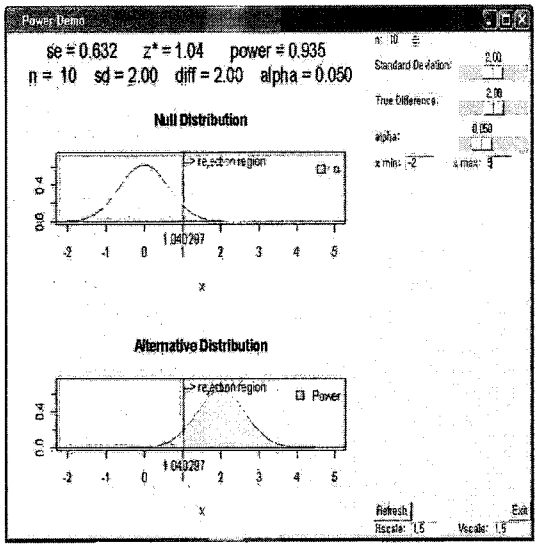

그림 **3.10.** 교수데모가 들어가 있는 R Commander에서 검정력 데모 화면

표본의 크기가 100개로 대표본이므로 정규검정을 다음과 같이 행할 수 있다. p-값이 0.410으로서 0.05보다 크므로 귀무가설을 기각할 수 없다. 즉 비만지수가 25보다 크다고 할 수 없다.

```
> # 모평균에 대한 검정(대표본의 경우)
> one.sample.z.test=function(mean.x,sd.x,n,alt,mu)
+ {
+ se=sd.x/sqrt(n)
+ z=(mean.x-mu)/se
+ if(alt=="큼") print(1-pnorm(z))
+ if(alt=="작음") print(pnorm(z))
+ if(alt=="같음") print(1-pnorm(abs(z))+pnorm(-abs(z)))
+ }
> one.sample.z.test(25.12,5.3,100,"큼",25)
[1] 0.4104393
```

예제 3.11: 다음 자료는 어느 체리농장에서 15개 지역의 수확량(단위: 에이커당 톤수)을 조사한 결과이다. 평균 수확량이 4.35t/acre보다 크다고 할 수 있나? 유의수준 5%에서 검정을 행하라(단, 수확량은 정규분포를 이룬다고 가정하자.).

3.56, 5.00, 4.88, 4.93, 3.92, 4.25, 5.12, 5.13, 5.35, 4.81, 3.48, 4.45, 4.72, 4.79, 4.45

풀이: 가설은 다음과 같다.

$$H_0 : \mu \leq 4.35, \qquad H_1 : \mu > 4.35$$

표본의 크기가 15개로 소표본이므로 t검정을 다음과 같이 행할 수 있다. p-값이 0.063로서 0.05보다 크므로 귀무가설을 기각할 수 없다. 즉 평균 수확량이 4.35t/acre보다 크다고 할 수 없다.

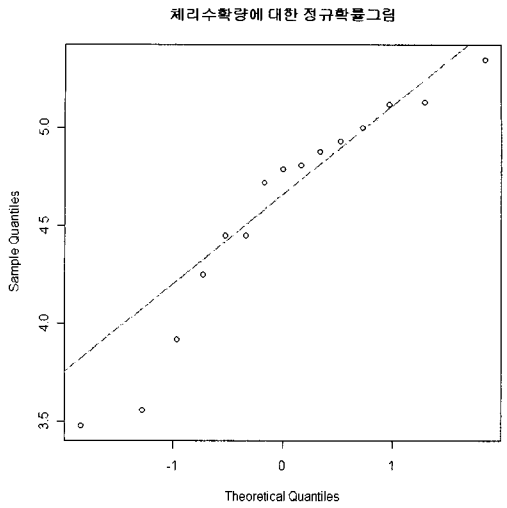

그림 **3.11.** 체리수확량에 대한 정규확률그림

```
> # 모평균에 대한 검정(소표본의 경우)
> x=c(3.56,5.00,4.88,4.93,3.92,4.25,5.12,5.13,5.35,4.81,3.48,4.45,4.72,4.79,4.45)
> t.test(x,alt="greater",mu=4.35)

One Sample t-test

data:  x
t = 1.6314, df = 14, p-value = 0.06255
alternative hypothesis: true mean is greater than 4.35
95 percent confidence interval:
4.330935 Inf
sample estimates:  mean of x
4.589333
```

참고로 만일 정규모집단이라는 가정이 없다면 검정을 행하기 전 자료를 이용하여 정규모집단이라고 할 수 있는 지를 검토하여야 한다. 다음 그림 3.11과 같은 정규확률그림 및 Shapiro-Wilk 검정을 통하여 정규모집단이라고 할 수 있다.

```
> qqnorm(x,main="체리수확량에 대한 정규확률그림")
> qqline(x,col="red")
> shapiro.test(x)

Shapiro-Wilk normality test

data:  x
W = 0.9109, p-value = 0.1396
```

예제 3.12: 인스턴트커피와 원두커피에 대한 선호도를 조사하기 위하여 100명을 대상으로 구별이 불가능한 두 개의 잔에 각각의 커피를 따른 후 좋아하는 커피를 선택하도록 하였더니 61명이 원두커피를 선호하였다. 원두커피에 대한 선호도가 65%보다 작다고 주장할 수 있는 지 유의수준 5%에서 검정을 행하라.

풀이: 가설은 다음과 같다.

$$H_0 : p \geq 0.65, \qquad H_1 : p < 0.65$$

표본의 크기가 100개로 대표본이므로 정규검정을 다음과 같이 행할 수 있다. p-값이 0.232로서 0.05보다 크므로 귀무가설을 기각할 수 없다. 즉 선호도가 65%보다 작다고 할 수 없다. R에서는 모비율의 검정에서 피어슨의 카이제곱통계량을 이용하여 검정을 행한다.

```
> # 모비율에 대한 검정
> prop.test(61,100,p=0.65,alt="less")

1-sample proportions test with continuity correction

data: 61 out of 100, null probability 0.65
X-squared = 0.5385, df = 1, p-value = 0.2315
alternative hypothesis: true p is less than 0.65
95 percent confidence interval:
0.0000000 0.6910052
sample estimates:
p
0.61
```

4. 결론

R을 대학교 기초통계학 추측통계 영역 강의 시 활용하는 문제에 대하여 앞 절에서 살펴보았다. 추후 과제로서 한 학기 기초통계학 추측통계 영역 강의 시 통계패키지로서 기초통계학용 R GUI인 R Commander을 활용하는 연구가 있겠다.

참고문헌

김진경, 박진호, 박헌진, 이재준, 전홍석, 황진수 (2008). <통계학-엑셀을 이용한 분석>, 자유아카데미.

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

Fox, J. (2005). The R commander: A basic-statistics graphical user interface to R, *Journal of Statistical Software*, **14**, 1–42.

Applications of R for Inferential Statistics in the Elementary Statistics Course

Dae-Heung Jang[1]

[1]Division of Mathematical Sciences, Pukyong National University

(Received April 2009; accepted May 2009)

Abstract

We can use R package as a statistical package on the statistical education for college students. R is an interactive mode package and graphical presentation tools are powerful in R. 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 statistical education for college students. We can consider the applications of R for inferential statistics in elementary statistics course.

Keywords: Statistical education, R package.

[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