Lesson5.1

Sung Won Kang

2016년 12월 1일

추정

- 추정: 표본의 특성에서 모집단의 특성을 추출
  - 추정의 목적: 모수값의 근사치를 파악??
  - 실제로 관심있는 값은? 모수일까, 기대값일까?

$$E(X|\theta) = \int x f(x|\theta) dx$$

- θ를 알면 $E(X|\theta)$는 당연히 알 수 있다
- 만약 θ를 모른다면? 아니, 만약 $f(x|\theta)$를 모른다면 $E(X|\theta)$ 는 포기해야 하나?
  - Predicitve Analysis : $E(y|X)$ 를 구하자. $f(y|\theta)$ 는 관심이 없다.

추정량(estimator)

- 추정량: 모수를 찍기 위해 만든 통계량 $\hat{\theta}$
- 추정치: 추정량에 표본을 대입하여 계산한 값.

추정량이 갖추었으면 하는 특징

1. 편향이 없을 것 (unbiasedness): 불편성(unbiasedness)

- 불편추정량(unbiased estimator)

$$E(\hat{\theta}) = \theta$$

$$E(\bar{X}) = \mu \quad E(X_i) = \mu, \quad i.i.d$$

$$E(S^2) = \sigma^2 \quad V(X_i) = \sigma^2 \quad i.i.d$$

(교과서 pp. 189-190)

2. 분산이 작을것 (efficiency): 유효성/효율성

- 최소분산불편추정량 (Minimum Variance unbiased estimator: MVUE)

$$V(\hat{\theta}_1) < V(\hat{\theta}_2)$$

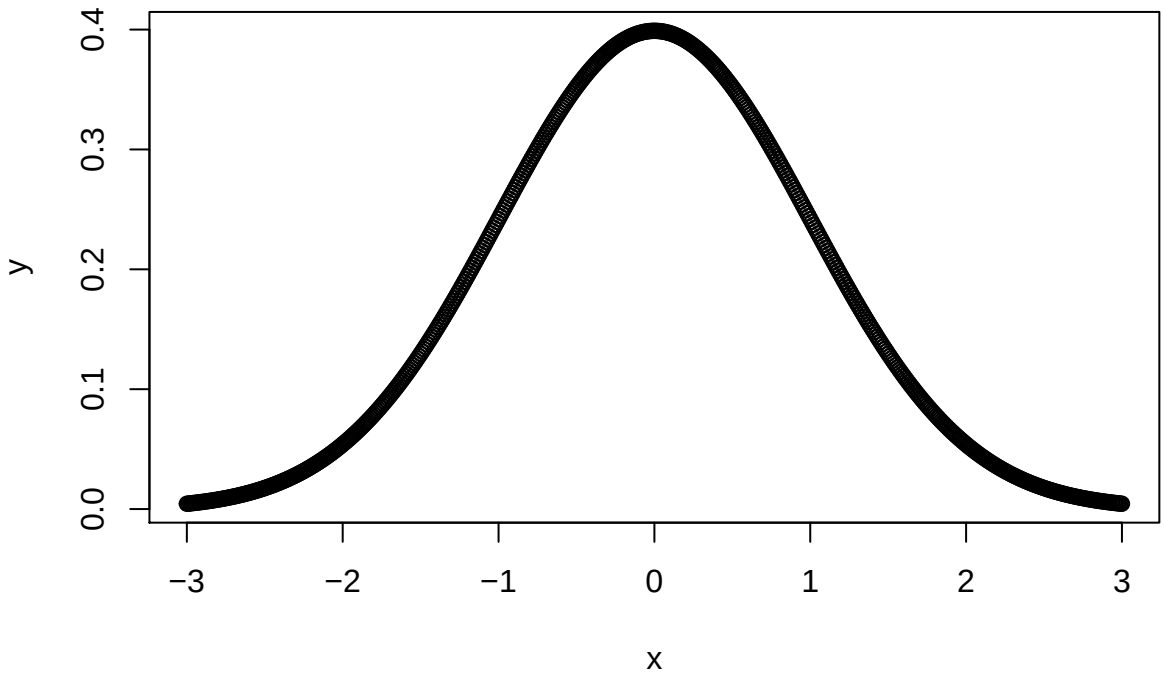
```
x=seq(-3,3,by=0.01)
y=dnorm(x)
y.1=dnorm(x,sd=sqrt(1/3))
y.2=dnorm(x,sd=sqrt(7/18))
pnorm(0.1,sd=sqrt(1/3))-pnorm(-0.1,sd=sqrt(1/3))
```

```
## [1] 0.1375098
```

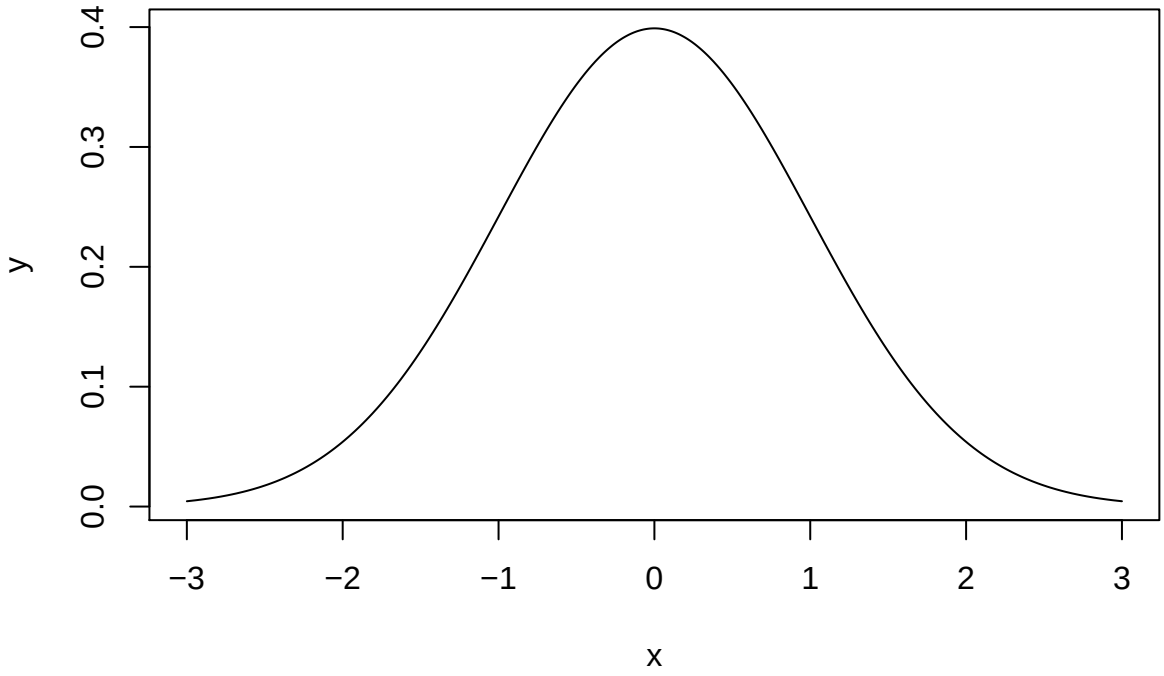
```
pnorm(0.1,sd=sqrt(7/18))-pnorm(-0.1,sd=sqrt(7/18))
```

```
## [1] 0.1273999
```

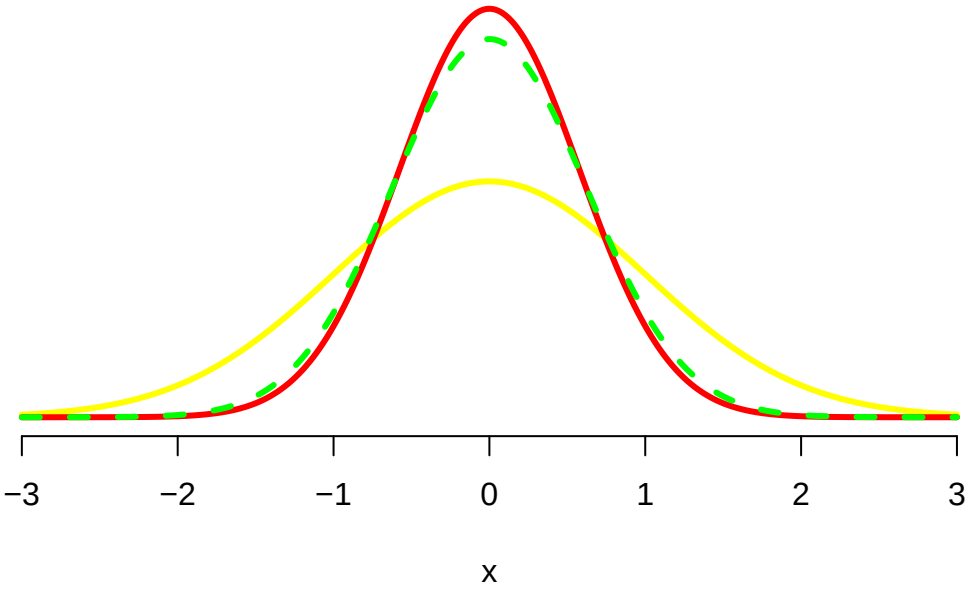
```
plot(x,y)
```




```
plot(x,y,type="l")
```




```
plot(x,y,type="l",ylim=c(0,0.8),axes=F,ylab="",lwd=3,col="yellow")
lines(x,y.1,col="red",lwd=3)
lines(x,y.2,col="green",lty=2,lwd=3)
axis(1)
```



- 효율성 비교

$$\bar{Y}_1 = (Y_1 + Y_2 + Y_3)/3, \qquad \bar{Y}_2 = (Y_1 + 2Y_2 + 3Y_3)/6$$

$$E(\bar{Y}_1) = E(Y)$$

$$E(\bar{Y}_2) = \frac{E(Y_1) + 2E(Y_2) + 3E(Y_3)}{6} = (6/6)E(Y) = E(Y)$$

```
options(digits=3)
set.seed(1)
mean.seq=function(x){
  n=length(x)
  sum=0
  n2=0
  for (i in 1:n){
    newx=i*x[i]
    sum=sum+newx
    n2=n2+i
  }
  return(sum/n2)
```

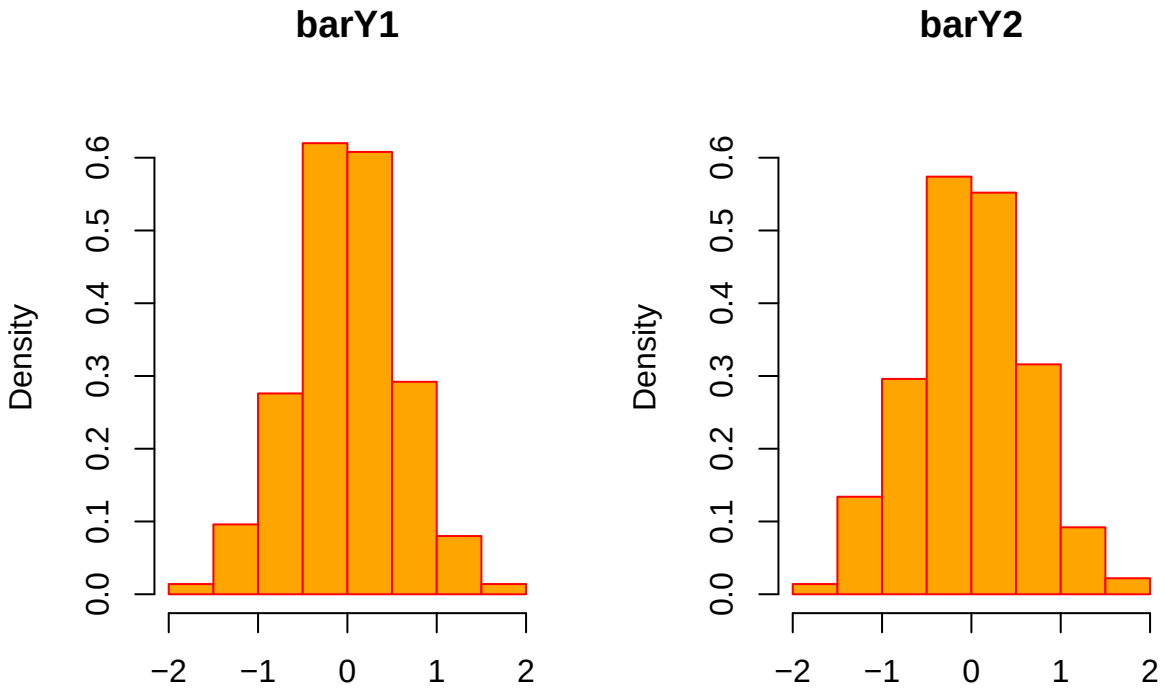
```
}
mean.seq2=function(x){
  n=length(x)
  sum(x*(1:n))/sum(1:n)
}
y1=rep(NA,1000)
y2=rep(NA,1000)

for (i in 1:1000){
  smp=rnorm(3)
  y1[i]=mean(smp)
  y2[i]=mean.seq(smp)
}
n1=length(y1[(y1>-0.1)&(y1<0.1)])
n2=length(y2[(y2>-0.1)&(y2<0.1)])

data.frame(mean=mean(y1),var=var(y1),n=n1)
```

```
##      mean  var   n
## 1 -0.0042 0.36 134
```

```
data.frame(mean=mean(y2),var=var(y2),n=n1)
```

```
##      mean   var   n
## 1 -0.0113 0.427 134
```

```
par(mfrow=c(1,2))
hist(y1,prob=T,xlim=c(-2,2),ylim=c(0,0.65),main="barY1",xlab="",col="orange",border="red")
hist(y2,prob=T,xlim=c(-2,2),ylim=c(0,0.65),main="barY2",xlab="",col="orange",border="red")
```



3. '수렴'할 것: 일치성(consistency)

$$\lim_{n\to \text{inf}} P(|\hat{\theta} - \theta| \geq \epsilon) = 0$$

표준오차(Standard error)

- 정의: 추정량의 표준편차의 추정량

추정량의 표준편차

$$\sqrt{V(\hat{\theta})} = \sqrt{E(\hat{\theta} - E(\hat{\theta}))^2}$$

$\hat{\theta}$ 가 불편추정량이면

$$= \sqrt{E(\hat{\theta} - \theta)^2}$$

그런데 이 값은 모집단의 모수에 의해 결정되는 경우가 많아서 관측이 불가능한 경우가 대부분이다. 예를 들면

$$\hat{\theta} = \bar{X}$$

$$\sqrt{E(\hat{\theta}-E(\hat{\theta}))^2}=\sqrt{E(\bar{X}-\mu)^2}=\sqrt{\sigma^2/N}=\sigma/\sqrt{N}$$

여기서 σ는 관찰이 불가능하다. 따라서 이 σ 대신 그 불편추정량인 표본표준편차 s를 사용하는 다음 추정량이 $\bar{X}$의 표준오차가 된다.

$$SE(\bar{X})=s/\sqrt{N} \qquad s=\sqrt{\frac{\sum_i X_i-\bar{X}}{N-1}}$$

$SE(\hat{\theta})$ 와 $S\hat{E}(\theta)$ 의 구분은 무의미하다. $SE(\hat{\theta})$가 실제 관측이 안 되는 값이기 때문에.

모비율의 점추정

1. 추정량 : 출현빈도

$$\hat{p}=\frac{X}{n}$$
$$E(\hat{p})=E(X)/n=np/n=p$$
$$\sqrt{V(X/n)}=\sqrt{V(X)/n^2}=\sqrt{np(1-p)/n^2}=\sqrt{p(1-p)/n}$$
$$SE(\hat{p})=\sqrt{\hat{p}(1-\hat{p})/n}$$

```
library(prob)
n=3
smps.all=rolldie(n)
str(smps.all)
```

```
## 'data.frame':    216 obs. of  3 variables:
##  $ X1: int  1 2 3 4 5 6 1 2 3 4 ...
##  $ X2: int  1 1 1 1 1 1 2 2 2 2 ...
##  $ X3: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
head(smps.all,n=3)
```

```
##   X1 X2 X3
## 1  1  1  1
## 2  2  1  1
## 3  3  1  1
```

```
is.even=function(x) return(!x%%2)# 나머지 (2로 나누어 나머지가 0이면 TRUE)
var.p=function(x){
  return(sum((x-mean(x))^2)/(length(x)))
}
```

```r
p.even=function(x, s.size=3){
  return(sum(is.even(x))/s.size)
}

phat=apply(smps.all,1,p.even)
mean(phat)
```

```
## [1] 0.5
```

```r
p.p=0.5
var.p(phat)
```

```
## [1] 0.0833
```

```r
p.p*(1-p.p)/3
```

```
## [1] 0.0833
```

```r
sqrt(var.p(phat))# 표준오차
```

```
## [1] 0.289
```

구간추정

1. 정의 : 모수 θ에 대한 $100(1-\alpha)\%$ 신뢰구간

$$P(\theta_L < \theta < \theta_U) = 1 - \alpha$$

$\alpha = 0.01$: 99% 신뢰구간 $\alpha = 0.05$: 95% 신뢰구간

2. 의미: 100번 반복해서 표본을 추출하면 $100(1-\alpha)$ 번 정도는 θ 가 구간 안에 들어감

3. 모평균 구간 추정 3.1. 분산을 알 경우 (정규분포)

$$X_i \sim N(\mu, \sigma^2)$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim N(0,1)$$

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha = 0.95$$

좌우 대칭을 이용하면

$$2 \times P(Z < -z_{\alpha/2}) = 0.05$$

$$P(Z < -z_{\alpha/2}) = 0.025 \rightarrow z_{\alpha/2} = 1.96$$

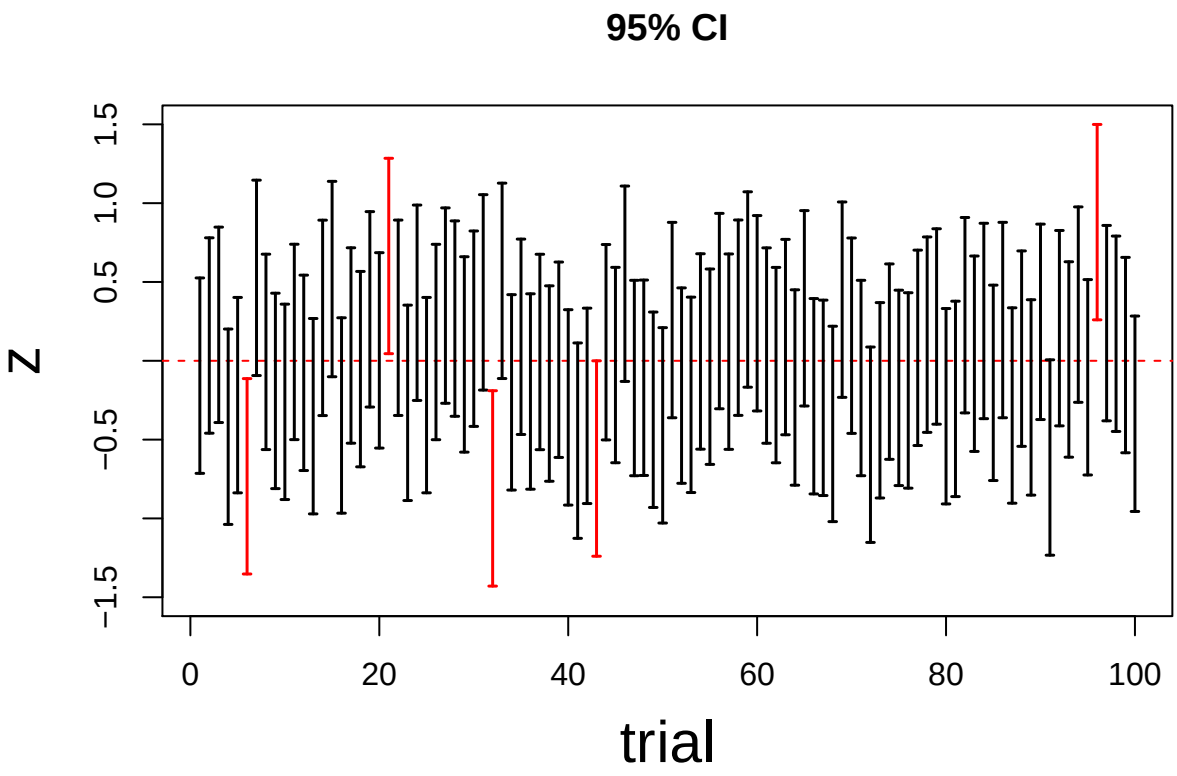
$$P\left[-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{N}} < 1.96\right] = 95\%$$

$$P\left[\bar{X} - 1.96\sigma/\sqrt{N} < \mu < \bar{X} + 1.96\sigma/\sqrt{N}\right] = 95\%$$

신뢰구간

$$(\bar{X} - 1.96\sigma/\sqrt{N}, \bar{X} + 1.96\sigma/\sqrt{N})$$

```r
set.seed(9)
n=10
x=1:100
y=seq(-3,3,by=0.01)

smps=matrix(rnorm(n*length(x)),ncol=n)
xbar=apply(smps,1,mean)
se=1/sqrt(10)
alpha=0.05
z=qnorm(1-alpha/2)
ll=xbar-z*se
ul=xbar+z*se

plot(y,type="n",xlab="trial",ylab="z",main="95% CI", xlim=c(1,100),ylim=c(-1.5, 1.5),cex.lab=1.8)
abline(h=0,col="red",lty=2)
l.c=rep(NA,length(x))
l.c=ifelse(ll*ul>0,"red","black")
arrows(1:length(x),ll,1:length(x),ul,code=3,angle=90,length=0.02,col=l.c,lwd=1.5)
```



3.2. 분산을 모를 경우 (t 분포)

$$T = \frac{\bar{X} - \mu}{s/\sqrt{N}} \sim t(N-1)$$

$$P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1 - \alpha = 0.95$$

좌우 대칭을 이용하면

$$2 \times P(T < -t_{\alpha/2,n-1}) = 0.05$$

$$P(T < -t_{\alpha/2,n-1}) = 0.025 \rightarrow t_{\alpha/2,n-1} = 2.78 \qquad (N = 5)$$

$$P\left[\bar{X} - 2.78 \times s/\sqrt{N} < \mu < \bar{X} + 2.78 \times s/\sqrt{N}\right] = 95\%$$

신뢰구간

$$(\bar{X} - 2.78s/\sqrt{N}, \bar{X} + 2.78s/\sqrt{N})$$

```r
ci.t=function(x,alpha=0.05){
  n=length(x)
  m=mean(x)
  s=sd(x)
  t=-qt((alpha/2),df=n-1)
  #t=-qt(1-(alpha/2),df=n-1)
  ll=m-t*(s/sqrt(n))
  ul=m+t*(s/sqrt(n))
  ci=c(1-alpha,ll,m,ul)
  names(ci)=c("Confidence level","Lower limit","Mean","Upper limit")
  return(ci)
}

smp=c(520,498,481,512,515,542,520,518,527,526)
ci.t(smp)
```

```
## Confidence level      Lower limit             Mean      Upper limit
##             0.95           503.98           515.90           527.82
```

```r
ci.t(smp,0.1)
```

```
## Confidence level      Lower limit             Mean      Upper limit
##              0.9            506.2            515.9            525.6
```

6장 준비