CHAPTER 07

추정(Estimation)

1

제7장 추정(Estimation)

1. 통계적 추론
2. 모평균에 대한 추론
3. 모비율에 대한 추론
4. 모분산에 대한 추론
5. 신뢰구간의 이해

2/22

2

1. 통계적 추론

[정의 10-1] 점추정(point estimation)
: 추정량의 관측된 값을 통해 모수의 참값을 추정하는 절차

- 불편성(unbiasedness) $E(\hat{\theta}) = \theta$ $\quad E(\bar{X}) = \mu,\ E(S^2) = \sigma^2$

[예 10-1] 평균이 μ인 모집단으로부터 n개의 표본을 추출하여 표본평균을 구했을 때, 표본평균은 μ에 대한 불편추정량임을 보이시오.

$$E(\bar{X}) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}n\mu = \mu$$

[예 10-2] 모분산이 σ^2인 모집단으로부터 n개의 표본을 추출하였을 때, 표본분산 S^2이 모분산 σ^2의 불편추정량임을 보이시오.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \Rightarrow E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1 \Rightarrow E(S^2) = \sigma^2$$

3/22

3

1.1 점추정

[예 10-3] 평균이 μ, 분산이 σ^2인 모집단으로부터 개의 표본을 추출하여 표본평균 X-bar와 또 다른 추정량 $Y=(X_1+X_2)/2$를 구했을 때, 두 추정량의 불편성을 검토하고, 분산을 비교하시오.

$$E(\bar{X}) = \mu,\ Var(\bar{X}) = \sigma^2 / n$$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

$$E(Y) = E\left(\frac{X_1 + X_2}{2}\right) = \frac{2\mu}{2} = \mu$$

$$Var(Y) = Var\left(\frac{X_1 + X_2}{2}\right) = \frac{2\sigma^2}{4} = \frac{\sigma^2}{2} > \frac{\sigma^2}{n} \quad (n \ge 3)$$

4/22

4

1.2 구간추정

[정의 10-2] 구간추정(interval estimation)
: 모수의 참값을 포함할 확률이 신뢰수준 1-α가 되는 신뢰구간을 결정하는 절차

- 신뢰구간(confidence interval) $\left[\hat{\theta}_L, \hat{\theta}_U\right]$
  모수의 참값을 포함할 확률이 신뢰수준 1-α가 되는 구간
- 신뢰수준(confidence level)
  신뢰구간이 모수의 참값을 포함할 확률 (1-α)
  $P\left(\hat{\theta}_L < \theta < \hat{\theta}_U\right) = 1 - \alpha$
- 유의수준(significance level)
  신뢰구간이 모수의 참값을 포함하지 못할 확률 (α)
  $1 - P\left(\hat{\theta}_L < \theta < \hat{\theta}_U\right) = \alpha$

5/22

5

2. 모평균에 대한 추론

[정리 10-1] 모평균의 신뢰구간 (모분산을 아는 경우)
: 정규 모집단의 모평균 μ에 대한 신뢰수준 (1-α) 신뢰구간(confidence interval), 100(1-α)% 신뢰구간

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[\bar{X} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

[증명]

$$1-\alpha = P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = P\left(-z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right)$$

$$= P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

6/22

6

2.1 모분산을 아는 경우의 추정

[예 10-4] 초콜릿 한 개의 무게는 모표준편차 5(g)인 정규분포를 따른다고 한다. 랜덤하게 샘플링한 초콜릿 50개 무게의 표본평균이 199.5(g)였을 때, 95% 신뢰구간을 구하시오.

$$\bar{X} = 199.5, \quad z_{0.975} \doteq 1.96$$

$$\left[199.5 \pm 1.96 \times \frac{5}{\sqrt{50}}\right] \doteq [199.5 \pm 1.386] = [198.114, 200.886]$$

- 신뢰구간의 오차(error) $z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
- 오차를 일정수준 이하로 유지하는데 필요한 표본의 개수

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \delta \Rightarrow n \ge \left(z_{1-\alpha/2}\frac{\sigma}{\delta}\right)^2$$

[예 10-5] 앞의 [예 10-4]에서 95% 신뢰수준에서 신뢰구간의 오차를 1.2 이하로 유지하기 위한 표본의 크기를 구하시오.

$$n \ge \left(1.96 \times \frac{5}{1.2}\right)^2 \doteq 66.7 \Rightarrow n \ge 67$$

7/22

7

2.3 모분산을 모르는 경우의 추정

[정리 10-3] 모평균에 대한 신뢰구간 (모분산을 모르는 경우)

: 모분산을 모르는 경우, 정규 모집단의 모평균에 대한 100(1-α)% 신뢰구간(confidence interval)

$$\left[\bar{X} - t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right] = \left[\bar{X} \pm t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right]$$

[증명]

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1) \Rightarrow P\left(-t_{1-\alpha/2;(n-1)} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{1-\alpha/2;(n-1)}\right) = 1-\alpha$$

$$P\left(\bar{X} - t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

$$\left[\bar{X} - t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right] = \left[\bar{X} \pm t_{1-\alpha/2;(n-1)}\frac{S}{\sqrt{n}}\right]$$

8/22

8

2.3 모분산을 모르는 경우의 추정

[예 10-8] 초콜릿 한 개의 무게는 정규분포를 따른다고 한다. 랜덤하게 샘플링한 초콜릿 16개 무게의 표본평균이 199.5(g), 표본분산이 25.0이었을 때, 모평균에 대한 95% 신뢰구간을 구하시오.

$$\bar{X} = 199.5,\ S = 5.0,\ t_{0.975;15} \doteq 2.131$$

$$\left[199.5 \pm 2.131 \frac{5}{\sqrt{16}}\right] \doteq [199.5 \pm 2.664] = [196.836, 202.164]$$

```
# 함수 정의 (표본개수, 표본평균, 표본표준편차, 유의수준)
tci1 <- function(n, xb, s, alp) {
        err <- qt(1-alp/2, n-1)*s/sqrt(n)
        cat("[", xb-err, ",", xb+err,"]₩n")}
# 함수 실행
tci1(16, 199.5, 5, 0.05)
[ 196.8357 , 202.1643 ]
```

9/22

9

2.3 모분산을 모르는 경우의 추정

[예 10-9] 초콜릿 한 개의 무게는 정규분포를 따른다고 한다. 랜덤하게 샘플링한 초콜릿 50개 무게의 측정 데이터가 다음과 같을 때, 모평균에 대한 95% 신뢰구간을 구하시오.

194	200	207	199	206	191	195	197	210	202
206	198	196	213	196	196	200	198	188	202
203	203	204	198	204	201	204	199	197	197
209	194	197	203	202	198	197	201	194	193
195	201	194	198	195	201	196	199	204	198

$$\sum_{i=1}^{50} x_i = 9{,}973 \qquad \sum_{i=1}^{50} x_i^2 = 1{,}990{,}407$$

$$\bar{X} = \frac{9{,}973}{50} = 199.46 \qquad S = \sqrt{\frac{1{,}990{,}407 - 9{,}973^2/50}{50-1}} \doteq 4.993$$

$$t_{0.975;49} \doteq 2.010$$

$$\left[199.46 \pm 2.010 \times \frac{4.993}{\sqrt{50}}\right] \doteq [199.46 \pm 1.402] = [198.058, 200.862]$$

10/22

10

3. 모비율에 대한 추론

[정리 10-5] 모비율의 근사적 신뢰구간
: 표본크기 n이 클 때, 모비율 p에 대한 100(1-α)% 근사 신뢰구간

$$\left[\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \qquad \hat{p} = \frac{X}{n}$$

$$\frac{X-np}{\sqrt{np(1-p)}} = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \overset{a}{\sim} N(0,1)$$

$$1-\alpha = P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) \simeq P\left(-z_{1-\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{1-\alpha/2}\right)$$

$$\simeq P\left(-z_{1-\alpha/2} < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z_{1-\alpha/2}\right)$$

$$\Rightarrow P\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \simeq 1-\alpha$$

11/22

11

3.1 모비율의 추정

[예 10-13] 공정에서 200개의 제품을 랜덤 샘플링하여 검사한 결과 15개의 불량품이 발견. 공정 불량률에 대한 95% 신뢰구간?

$$\hat{p} = \frac{X}{n} = \frac{15}{200} = 0.075 \qquad z_{0.975} \doteq 1.96$$

$$\left[0.075 \pm 1.96 \times \sqrt{\frac{0.075(1-0.075)}{200}}\right] \doteq [0.075 \pm 0.037] = [0.038, 0.112]$$

```
# 정규근사 신뢰구간 함수 정의 (표본개수, 성공회수, 유의수준)
pci1 <- function(n, x, alp) {
            p <- x/n
            err <- qnorm(1-alp/2)*sqrt(p*(1-p)/n)
            cat("[", p, "±", err, "] = [", p-err, ",", p+err,"]₩n")}
# 함수 실행
pci1(200, 15, 0.05)
[ 0.075 ± 0.03650351 ] = [ 0.03849649 , 0.1115035 ]
# 이항분포를 이용한 정확한 신뢰구간 추정 (다소 차이 발생)
binom.test(15,200)$conf
[1] 0.0425828 0.1206842
attr(,"conf.level")
[1] 0.95
```

12/22

12

3.1 모비율의 추정

- 신뢰구간 오차를 일정 수준 이하로 유지하기 위해 필요한 표본의 개수

$$z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \delta \;\Rightarrow\; n \ge \hat{p}(1-\hat{p})\left(\frac{z_{1-\alpha/2}}{\delta}\right)^2$$

$$p(1-p) = -(p-0.5)^2 + 0.25 \;\Rightarrow p^* = 0.5 \quad \Rightarrow n \ge 0.5^2 \times \left(\frac{z_{1-\alpha/2}}{\delta}\right)^2$$

[예 10-14] 어떤 프로세스의 불량률은 과거의 경험을 통하여 7.5% 95% 신뢰수준에서 모비율에 대한 신뢰구간의 오차가 0.02 이하가 되기 위해 필요한 표본의 크기를 구하시오.

$$p = 0.075 \quad \Rightarrow n \ge 0.075 \times (1-0.075)\left(\frac{1.96}{0.02}\right)^2 \doteq 666.3 \quad \Rightarrow n \ge 667$$

- 모비율 p에 대한 정보가 전혀 없는 경우

$$\Rightarrow n \ge 0.5^2 \times \left(\frac{1.96}{0.02}\right)^2 \doteq 2400.9 \;\Rightarrow n \ge 2401$$

13/22

13

3.1 모비율의 추정

[예 10-14] (계속)

```
# 표본 개수를 구하는 함수 정의 (오차한계, 유의수준, 비율추정치)
nsample  <- function(err, alp=0.05, ph=0.5) {
         n  <- qnorm(1-alp/2)^2 * ph*(1-ph) /err^2
         cat("n ≥", n, "⇒ n ≥", ceiling(n), "₩n")}
# 함수 실행
nsample(0.02,  ph=15/200)
n ≥ 666.253 ⇒ n ≥ 667
nsample(0.02)
n ≥ 2400.912 ⇒ n ≥ 2401
```

14/22

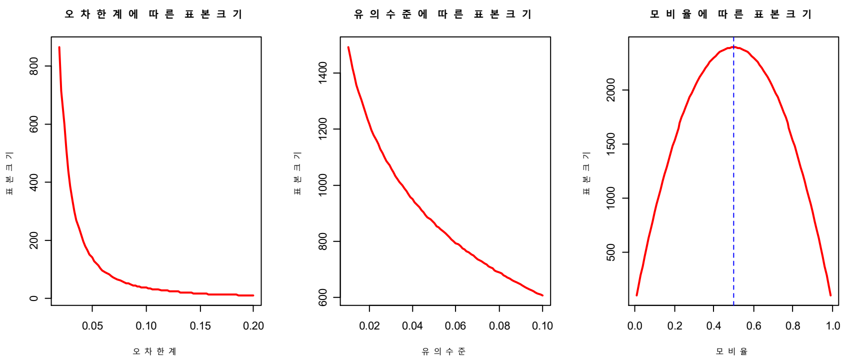
14

[예 10-15] 아래의 경우에 대해 모비율에 대한 신뢰구간의 오차한계를 유지하기 위해 필요한 표본크기의 그래프를 작성하시오.

(1) p=0.1, α=0.05일 때, 오차한계 δ=0.02 ~ 0.2에서 변화

(2) p=0.1, δ=0.02일 때, 유의수준 α=0.01 ~ 0.1에서 변화

(3) α=0.05, δ=0.02 일 때, 모비율 p=0.01 ~ 0.99에서 변화



15/22

15

4. 모분산에 대한 추론

[정리 10-8] 모분산에 대한 신뢰구간

: 정규 모집단의 모분산 σ^2에 대한 100(1-α)% 신뢰구간

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}}\right]$$

- 카이제곱분포의 분위수

$$P(\chi^2_v \le \chi^2_{p;v}) = p \quad \Rightarrow P(\chi^2_{\alpha/2;v} \le \chi^2_v \le \chi^2_{1-\alpha/2;v}) = 1-\alpha$$

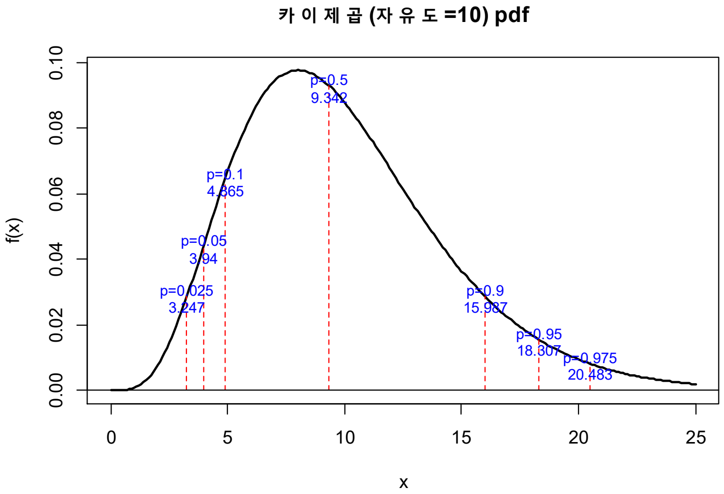
$$\Rightarrow P(\chi^2_{\alpha/2;n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{1-\alpha/2;n-1}) = 1-\alpha$$

$$= P(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2;n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\alpha/2;n-1}})$$

16/22

16

[예 10-18] 자유도 10인 카이제곱분포의 확률밀도함수를 그리고, 누적확률 p=0.025, 0.05, 0.1, 0.5, 0.9, 0.95, 0.975에 대응하는 분위수를 표시하시오.



17/22

17

4.1 모분산의 추정

[예 10-19] 품질특성치가 정규분포를 따르는 생산공정에서 품질특성에 대한 분산을 추정하기 위하여 10개 제품을 랜덤 샘플링한 결과 다음의 데이터를 구하였다. 모분산의 95% 신뢰구간을 구하시오.

20.0 21.5 20.9 19.8 22.5 20.3 23.6 18.0 23.3 17.8

$$S^2 = \frac{4350.13 - 207.7^2/10}{9} = \frac{36.201}{9} \doteq 4.022$$

$$\chi^2_{0.025;9} \doteq 2.700 \quad \chi^2_{0.975;9} \doteq 19.023$$

$$\Rightarrow \left[\frac{36.201}{19.023}, \frac{36.201}{2.700}\right] \doteq [1.903, 13.408]$$

18/22

18

5. 신뢰구간의 이해

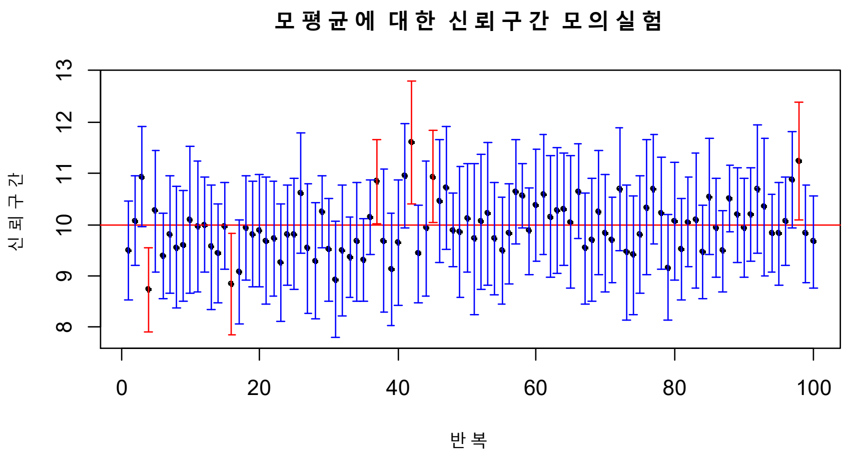
- **신뢰구간에 대한 오해**

모수 θ가 신뢰구간 [a, b]에 포함될 확률은 100(1-α)%이다.

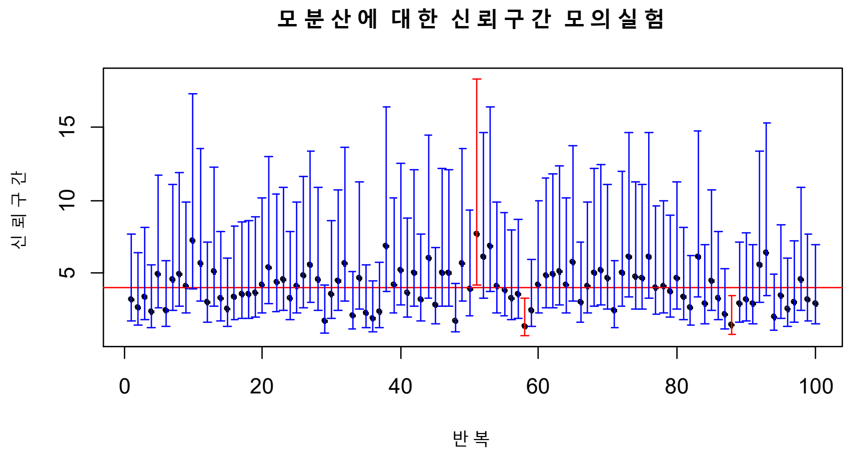
- **신뢰구간에 대한 바른 해석**

랜덤한 신뢰구간 [L, U]가 모수 θ의 참값을 포함할 확률은 100(1-α)%인데, 그 신뢰구간 중 하나의 관측치가 [a, b]이다.

[예 10-22] 평균 10, 분산 4인 정규모집단에서 크기 16의 확률표본을 추출하여 모평균에 대한 95% 신뢰구간을 구하여 그래프로 나타내고, 이 과정을 100회 반복하여 모평균 10을 포함하지 않는 신뢰구간이 몇 회 나타나는지 확인하시오.



[예 10-23] 평균 10, 분산 4인 정규모집단에서 크기 16의 확률표본을 추출하여 모분산에 대한 95% 신뢰구간을 구하여 그래프로 나타내고, 이 과정을 100회 반복하여 모분산 4를 포함하지 않는 신뢰구간이 몇 회 나타나는지 확인하시오.



[예 10-24] 모비율 0.6인 모집단에서 크기 16의 확률표본을 추출하여 모비율에 대한 95% 신뢰구간을 구하여 그래프로 나타내고, 이 과정을 100회 반복하여 모비율 0.6을 포함하지 않는 신뢰구간이 몇 회 나타나는지 확인하시오.