제4강 추정

1. 용어 정의
2. 좋은 추정량이란?
3. 추정 및 신뢰구간
4. 모평균의 구간추정
5. 모분산의 구간추정

1. 용어 정의

(1) 추리(inference)

– 표본의 정보를 바탕으로 모집단에 관한 의사결정, 추정,예측을 하는 것으로 추정 및 가설검정으로 구분

(2) 추정(estimation)

– 표본을 이용하여 모집단의 미지의 수를 얻는 것

$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots x_n)$(즉, 추정량 $\hat{\theta}$는 표본관측의 함수이다)

(3) 추정량(estimator)

– 모수를 추정하는 공식을 나타내는 표본통계량

– (예) $\bar{X}, s^2$은 각각 모평균(μ)과 모분산(σ^2)의 추정량임

– (추정량의 종류)

·$E(\hat{\theta})$: 추정량의 평균

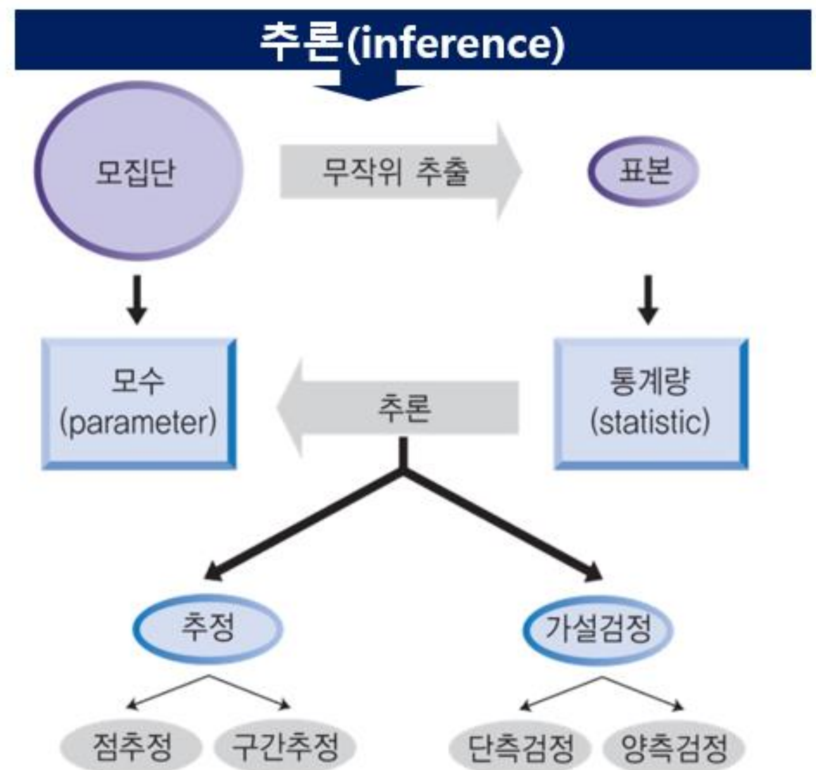
·$Var(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 = E(\hat{\theta}^2) - [E(\hat{\theta})]^2$: 추정량의 분산

·$\sqrt{Var(\hat{\theta})}$: 표준오차(standard error)(추정량의 표준오차)

·$\hat{\theta} - \theta$: 표본오차(sampling error)

·$E(\hat{\theta}) - \theta$: 편의(bias)

·$E(\hat{\theta} - \theta)^2$: 평균자승오차(mean square error)



(4) 추정치(estimate)

– 추정량에 실제 관찰값을 대입하여 계산한 표본통계량의 값

2. 좋은 추정량이란?

바람직한 추정량의 기준

- 불편성: $E(\hat{\theta}) = \theta$
- 효율성: $Var(\hat{\theta})$이 작을수록 효율적
- 일치성: $p\lim(\hat{\theta}) = \theta$

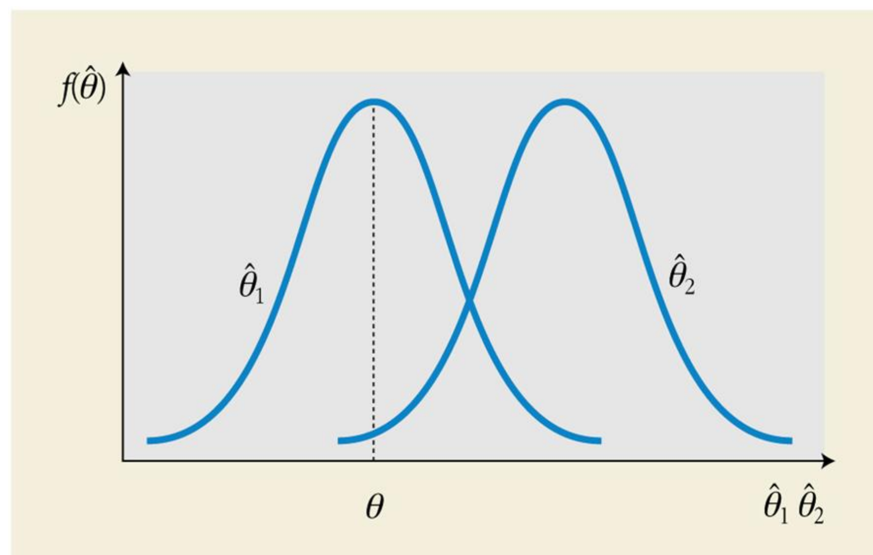
(1) 불편성(unbiasedness)

– (정의) 만약 $\boldsymbol{E(\hat{\theta}) = \theta}$ 이면 $\boldsymbol{\hat{\theta}}$ 는 $\boldsymbol{\theta}$ 의 불편추정량이라고 함. 즉, bias가 없는 추정량임

– (의미) 추정량의 공식 $\boldsymbol{\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)}$을 이용하여 동일한 모집단으로부터 반복하여 모수 $\boldsymbol{\theta}$ 의 추정치를 계산할 때 그 추정치의 평균이 결국 모수와 같아진다는 것

– $\boldsymbol{\hat{\theta}_1}$은 $\boldsymbol{\theta}$ 의 불편추정량

– $\boldsymbol{\hat{\theta}_2}$은 $\boldsymbol{\theta}$ 의 불편추정량이 아님



- (예1) $\bar{X}$는 μ의 불편추정량임

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

- (예2) s^2은 σ^2의 불편추정량임

$$E(s^2) = \frac{1}{n-1}E[\sum_{i=1}^{n} (X_i - \overline{X})^2]$$

$$= \frac{1}{n-1}E[\sum_{i=1}^{n} ((X_i - \mu) - (\overline{X} - \mu))^2]$$

$$= \frac{1}{n-1}E[\sum_{i=1}^{n} (X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n} (X_i - \mu) + n(\overline{X} - \mu)^2]$$

$$= \frac{1}{n-1}E[\sum_{i=1}^{n} (X_i - \mu)^2 - 2n(\overline{X} - \mu)^2 + n(\overline{X} - \mu)^2]$$

$$= \frac{1}{n-1}E[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2]$$

$$= \frac{1}{n-1}[\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\overline{X} - \mu)^2]$$

$$= \frac{1}{n-1}[n\sigma^2 - n\frac{\sigma^2}{n}]$$

$$= \frac{1}{n-1}(n-1)\sigma^2$$

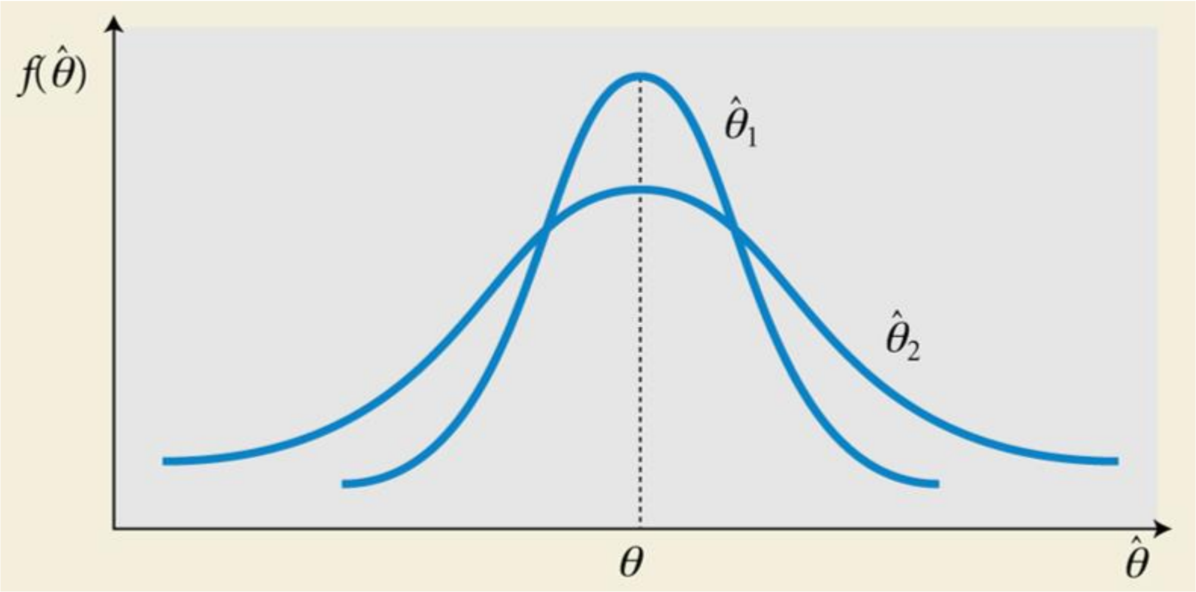
$$= \sigma^2$$

(2) 효율성(effectiveness)

– (정의) $\hat{\boldsymbol{\theta}}_1$과 $\hat{\boldsymbol{\theta}}_2$가 모두 불편추정량이라고 할 때, $Var(\hat{\boldsymbol{\theta}}_1) < Var(\hat{\boldsymbol{\theta}}_2)$ 이면 $\hat{\boldsymbol{\theta}}_1$이 $\hat{\boldsymbol{\theta}}_2$보다 더 효율적이라고 함

– $\hat{\boldsymbol{\theta}}_1$이 다른 모든 불편추정량보다 작은 분산을 가지면 $\hat{\boldsymbol{\theta}}_1$은 가장 효율적인(most efficient) 혹은 최소분산 (minimum variance) 불편추정량이라고 함

– 상대적 효율성(=$\frac{\hat{\theta}_2}{\hat{\theta}_1}$)이 1보다 크면 $\hat{\boldsymbol{\theta}}_1$이 $\hat{\boldsymbol{\theta}}_2$보다 효율적인 추정량임



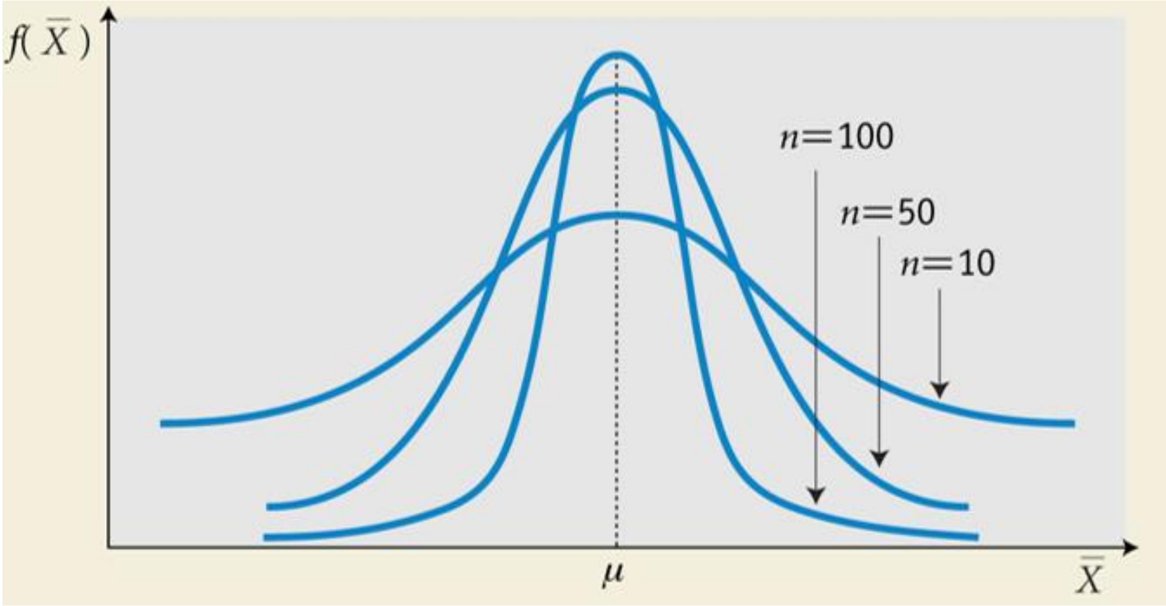
(3) 일치성(consistency)

– (정의) 표본크기가 무한히 증가할 때 추정량($\hat{\theta}$)이 모수(θ)에 근접하려는 특성을 일치성이라고 함

– $\hat{\theta}$의 분포가 θ 주위에서 높은 확률로 실현된다면 즉, $\hat{\theta}_n \xrightarrow{p} \theta$ 이면, $\hat{\theta}_n$은 θ의 일치추정량이 됨

– $n \to \infty$일 때, 임의의 $\varepsilon > 0$에 대해 $P[|\hat{\theta} - \theta| \le \varepsilon] \to 1$이 성립되면, $\hat{\theta}$은 일치추정량인데, 이를 간단히 다음과 같이 표시할 수 있으며, 이때 $\hat{\theta}$을 θ의 확률극한(probability limit)이라고 함

$$plim\,\hat{\theta} = \theta$$



(1) 추정(estimation)

- 추정량은 표본정보를 이용하여 알지 못하는 모수의 참값을 추정하는 방법이며, 알지 못하는 모수가 $\boldsymbol{\theta}$이라면 추정량은 $\hat{\boldsymbol{\theta}}$으로 표기함
- 추정량은 여러 가지가 있으며, 추정치는 수치로 계산된 추정량의 값임

① 점추정(point estimation) : 단일값으로 모수를 추정하는 것 또는 확률표본의 정보를 이용하여 모수에 대한 특정 값을 지정하는 것

② 구간추정(interval estimation) : 일정한 구간(신뢰구간)에 의해 모수를 추정하는 것. 점추정량은 언제나 표본오차를 수반하기 때문에 전적으로 신뢰할 수 없으나 구간추정은 이런 점추정과 달리 모수가 빈번히 포함되는 범위를 제공하여 연구의 목적에 따라 원하는 만큼의 신뢰성을 가지고 모수를 추정할 수 있음

(2) 신뢰구간(confidence interval)

① 신뢰구간(confidence interval) : 추정하고자 하는 모수 $\boldsymbol{\theta}$가 포함되어 있을 구간

- 모수 $\boldsymbol{\theta}$의 신뢰구간: $\boldsymbol{P(L \le \theta \le U) = 1} - \propto$

  단, $\boldsymbol{\alpha}$는 유의수준(significance level)

  $\boldsymbol{1-\alpha}$는 신뢰수준(confidence level)

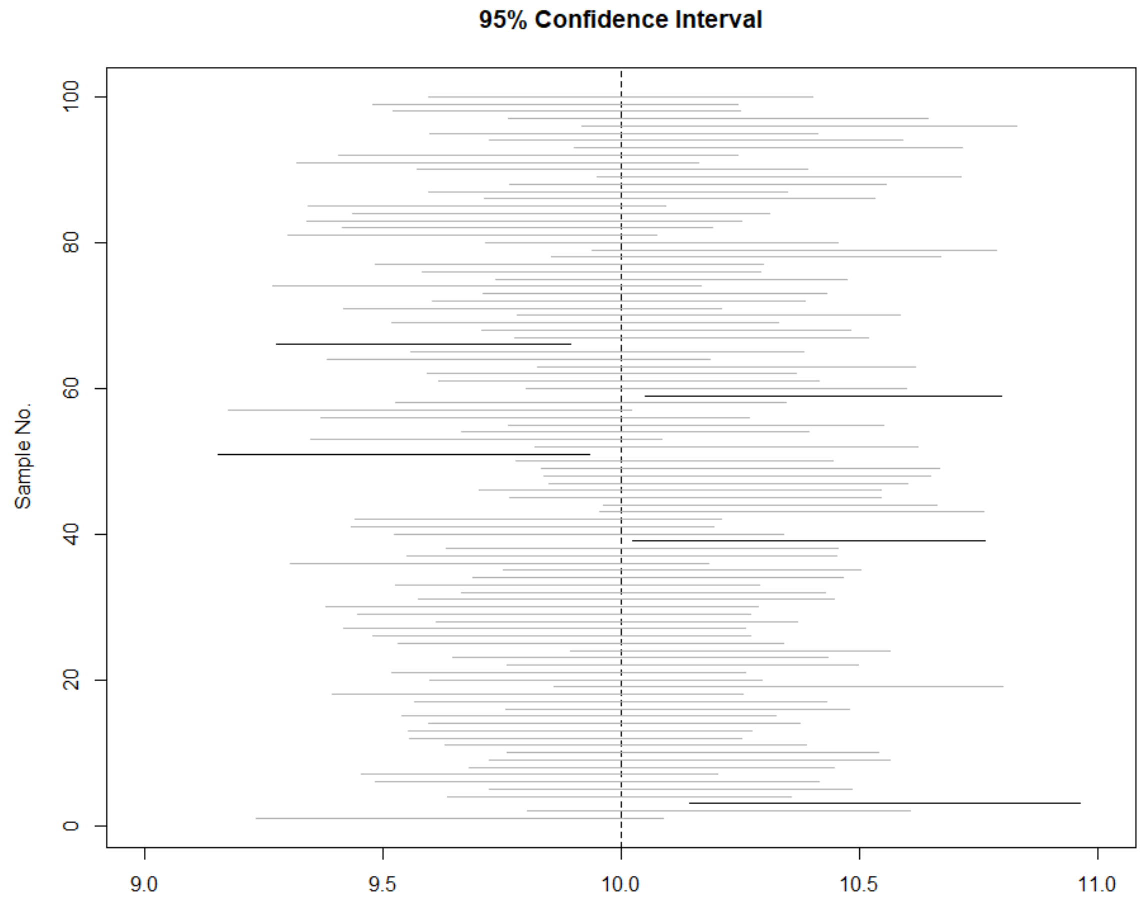
  $\boldsymbol{L}$은 신뢰구간의 하한(lower confidence limit)

  $\boldsymbol{U}$는 신뢰구간의 상한(upper confidence limit)

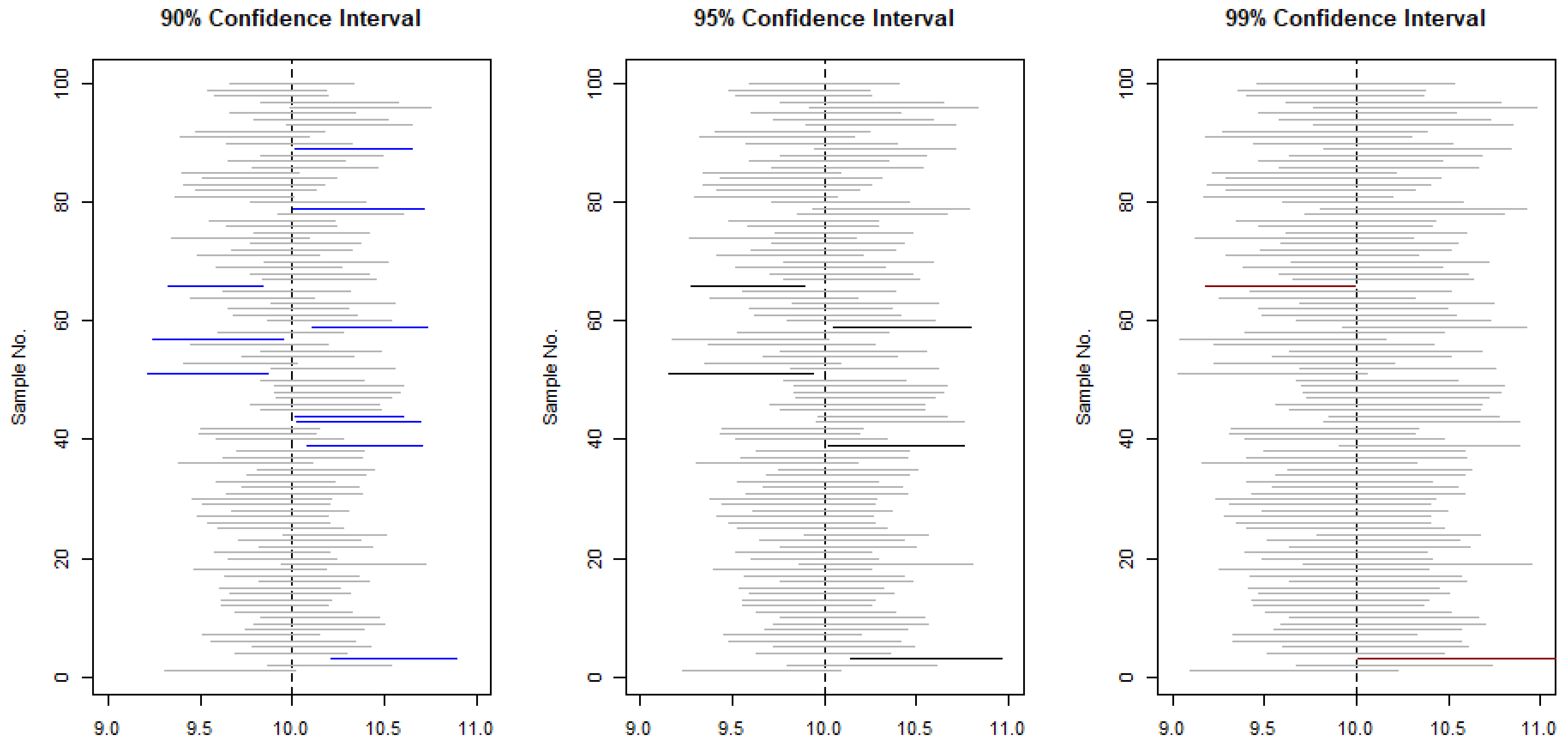
② 신뢰수준(confidence level) : 구간추정에 있어 모수 $\boldsymbol{\theta}$가 신뢰구간의 하한과 상한 사이에 있으리라는 것을 $(\boldsymbol{1-\alpha}) \times \boldsymbol{100}\%$ 확신을 가지고 신뢰할 수 있는 수준

– (실험1) 평균이 10이고 표준편차가 2인 모집단에서 표본크기가 100인 10000개의 표본을 추출하여 모평균 μ에 대한 신뢰구간을 추정해 보면 모평균에 대한 95% 신뢰구간 10000개 중 100개만 살펴보면 95개의 신뢰구간이 모평균 10을 포함하고 있는 것을 확인할 수 있음

– 모평균 μ가 확률변수가 아니고 고정된 상수이므로 μ에 대한 95% 신뢰구간의 의미는 포본크기가 동일한 10000개의 서로 다른 표본에 의해 동일한 공식으로 10000개의 신뢰구간을 구했을 때 그 중에서 95%의 구간이 모평균 μ를 포함한다고 볼 수 있다는 것임



– 모평균 μ에 대한 90%, 95% 및 99% 신뢰구간 100개 중 각각 90개, 95개, 98개의 신뢰구간이 모평균 10을 포함하고 있는 것을 확인할 수 있음



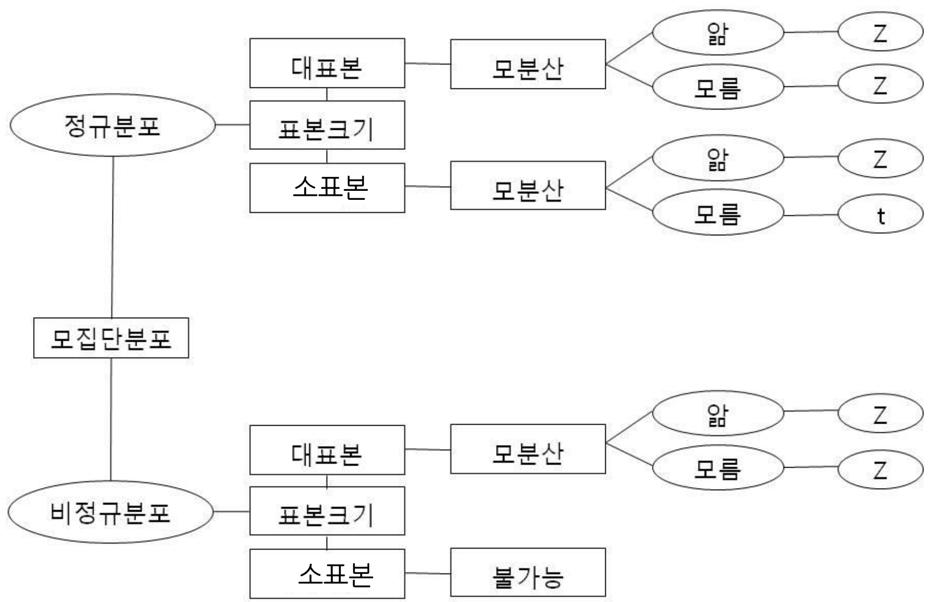
4. 모평균의 구간추정

- 모평균 구간추정을 위한 의사결정 트리는 다음과 같음

·모집단의 정규분포 여부에 관계없이, 모분산을 알든 모르든 관계없이, 표본의 크기가 30 이상의 대표본이면 Z-통계량을 이용

·표본의 크기가 30 미만의 소표본이면, 모집단이 정규분포에 따를 경우에만 모평균 구간추정이 가능하고, 이 경우에 모분산이 알려져 있으면 Z-통계량을 이용하고 모분산을 모를 경우 t-통계량을 이용

모평균 구간추정을 위한 의사결정 트리



(1) Case 1

– 모집단이 정규분포를 하며 모분산이 알려져 있을 경우

– 모집단이 정규분포를 하지 않더라도 모분산이 알려져 있고 표본의 크기가 30이상으로 충분히 클 경우

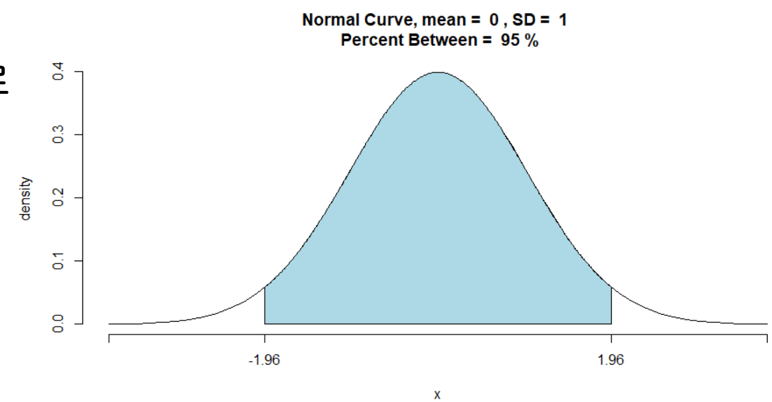
⇒ 표준정규분포를 이용

$\cdot X \sim N(\mu, \sigma^2)$

$\cdot \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$\cdot z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

```
> library(tigerstats)
> pnorm(1.96,0,1)
[1] 0.9750021
> qnorm(0.975,0,1)
[1] 1.959964
> qnormGC(0.95,region="between",m=0,s=1,graph=T)
[1] -1.959964  1.959964
```



– 표준정규분포표로부터 $p(-1.96 \le z \le 1.96) = 0.95$이므로

$$p\left(-1.96 \le \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right) = 0.95$$

$$p\left(-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$p\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

– 모평균 μ에 대한 95% 신뢰구간은 $[\bar{X}-1.96\frac{\sigma}{\sqrt{n}}, \bar{X}+1.96\frac{\sigma}{\sqrt{n}}]$

– $(1-\alpha)\times 100\%$ 신뢰구간의 일반적인 표시 : $\bar{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

90% 신뢰구간	$z_{\frac{\alpha}{2}}$=1.645
95% 신뢰구간	$z_{\frac{\alpha}{2}}$=1.96
99% 신뢰구간	$z_{\frac{\alpha}{2}}$=2.58

– (연습 1) 한 기업에서 현대인들이 대기오염에 시달린다는 사실에 착안하여 언제나 신선한 산소를 마실 수 있는 휴대용 산소제품을 개발하기로 하였다. 따라서 이 기업의 연구팀은 일반인들의 산소 소비량을 측정하기 위해 임의로 35명을 선정, 분당 산소 소비량을 조사하여 다음 자료를 얻었다. 이 연구팀은 일반인들의 모집단은 정규분포하며 분산이 0.36이라는 사실을 알고 있다고 가정하자. 모집단 평균의 95% 신뢰구간을 구하라.

⇒ 표준정규분포를 이용

0.360	1.189	0.614	0.788	0.273	2.464	0.517	1.827	0.537	0.374	0.449	0.262
0.448	0.971	0.372	0.898	0.411	0.348	1.925	0.550	0.622	0.610	0.319	0.406
0.413	0.767	0.385	0.674	0.521	0.603	0.533	0.662	1.177	0.307	1.499	

– $\bar{X} = 0.7164286, \sigma = 0.6, n = 35$

– 표준정규분포표로부터 $p(-1.96 \le z \le 1.96) = 0.95$이므로

$$p\left(-1.96 \le \frac{0.7164286-\mu}{\frac{0.6}{\sqrt{35}}} \le 1.96\right) = 0.95$$

$$p\left(-1.96\frac{0.6}{\sqrt{35}} \le 0.7164286 - \mu \le 1.96\frac{0.6}{\sqrt{35}}\right) = 0.95$$

$$p\left(0.7164286 - 1.96\frac{0.6}{\sqrt{35}} \le \mu \le 0.7165286 + 1.96\frac{0.6}{\sqrt{35}}\right) = 0.95$$

– 모평균 μ에 대한 95% 신뢰구간은 $[0.7164286 - 1.96\frac{0.6}{\sqrt{35}}, 0.7164286 + 1.96\frac{0.6}{\sqrt{35}}]$

– 모평균 μ에 대한 95% 신뢰구간은 $[0.517648, 0.915209]$

(2) Case 2

– 모집단이 정규분포 여부에 관계 없이 모분산이 알려져 있지 않으나 표본의 크기가 30이상으로 충분히 클 경우

⇒ s^2으로 σ^2을 대체한 후 표준정규분포를 이용

·$X \sim N(\mu, ?)$ 또는 $X \sim (\mu, ?), n \geq 30$

·$\bar{X} \sim N\left(\mu, \frac{s^2}{n}\right)$

·$z = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim N(0, 1)$

– 표준정규분포표로부터 $p(-1.96 \leq z \leq 1.96) = 0.95$이므로

$$p\left(-1.96 \leq \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \leq 1.96\right) = 0.95$$

$$p\left(-1.96\frac{s}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96\frac{s}{\sqrt{n}}\right) = 0.95$$

$$p\left(\bar{X} - 1.96\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{s}{\sqrt{n}}\right) = 0.95$$

– 모평균 μ에 대한 95% 신뢰구간은 $[\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}}]$

– $(1-\alpha) \times 100\%$ 신뢰구간의 일반적인 표시 : $\bar{X} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$

(3) Case 3

– 모집단이 정규분포를 하지 않고 모분산이 알려져 있지 않으며 표본의 크기가 30이하로 충분히 크지 않을 경우 ⇒ t–분포를 이용

$\cdot t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$

– t–분포표로부터 $p\left(-t_{0.025,n-1} \leq t \leq t_{0.025,n-1}\right) = 0.95$이므로

$$p\left(-t_{0.025,n-1} \leq \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \leq -t_{0.025,n-1}\right) = 0.95$$

$$p\left(-t_{0.025,n-1}\frac{s}{\sqrt{n}} \leq \bar{X}-\mu \leq t_{0.025,n-1}\frac{s}{\sqrt{n}}\right) = 0.95$$

$$p\left(\bar{X}-t_{0.025,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X}+t_{0.025,n-1}\frac{s}{\sqrt{n}}\right) = 0.95$$

– 모평균 μ에 대한 95% 신뢰구간은 $[\bar{X}-t_{0.025,n-1}\frac{s}{\sqrt{n}}, \bar{X}+t_{0.025,n-1}\frac{s}{\sqrt{n}}]$

– $(1-\alpha)\times 100\%$ 신뢰구간의 일반적인 표시 : $\bar{X} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$

– (연습 2) 한 피자 체인점의 지배인은 피자 배달시간이 오래 걸린다는 소비자들의 불평을 확인해 보기 위해 피자 배달주문 중 임의로 20개를 선정하여 배달시간(단위 : 분)을 측정하였더니 다음과 같았다. 모집단이 정규분포를 한다고 가정하고 모평균 배달시간에 대한 95% 신뢰구간을 구하라.

⇒ t–분포를 이용

14	10	9	10	11	16	15	8	6	18
17	4	12	15	14	15	9	8	7	16

– $\overline{X} = 11.7, s = 4.040584, n = 20$

– t–분포표로부터 $t_{0.025,19} = 2.093$이므로,

$$p\left(-2.093 \le \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \le 2.093\right) = 0.95$$

$$p\left(-2.093 \le \frac{11.7-\mu}{\frac{4.040584}{\sqrt{20}}} \le 2.093\right) = 0.95$$

$$p\left(-2.093\frac{4.040584}{\sqrt{20}} \le 11.7-\mu \le 2.093\frac{4.040584}{\sqrt{20}}\right) = 0.95$$

$$p\left(11.7-2.093\frac{4.040584}{\sqrt{20}} \le \mu \le 11.7+2.093\frac{4.040584}{\sqrt{20}}\right) = 0.95$$

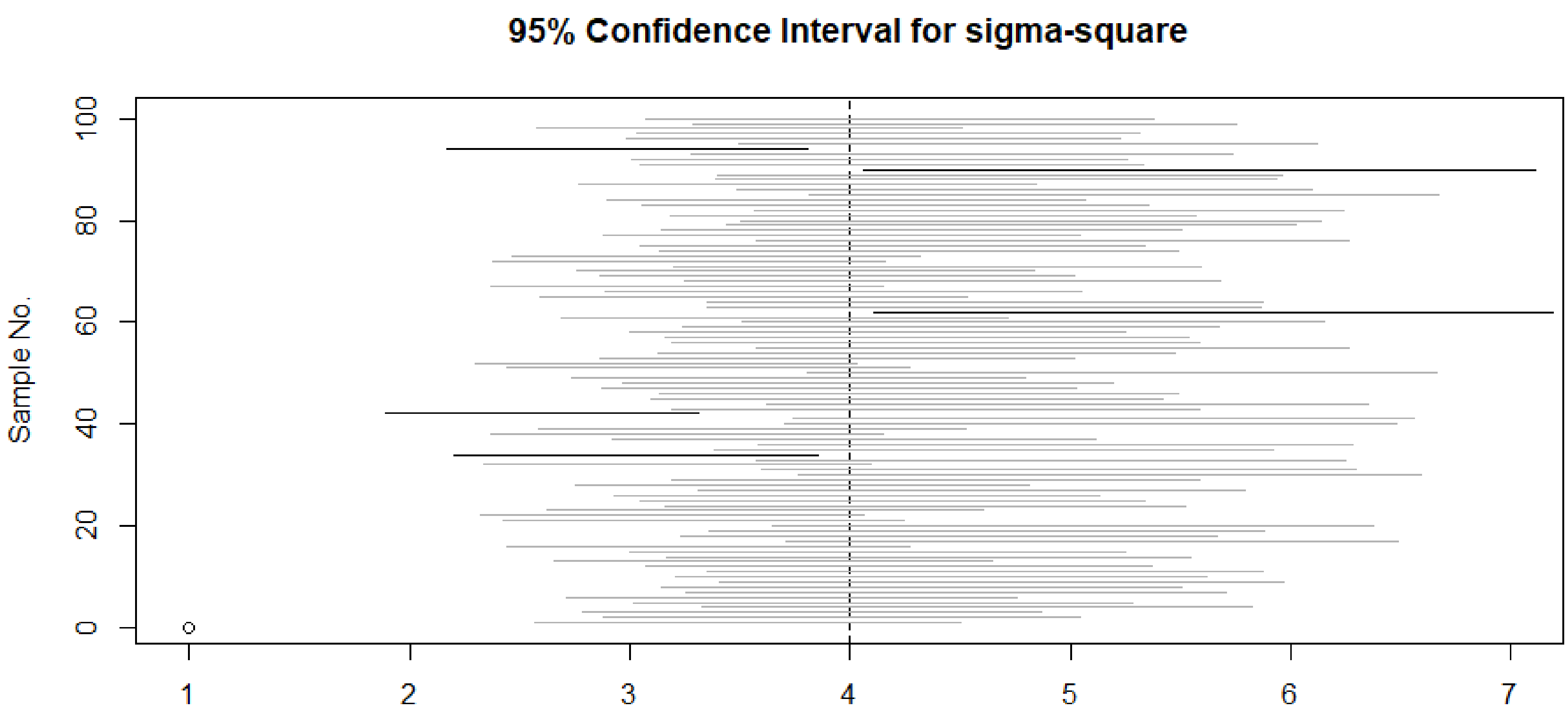
– 모평균 μ에 대한 95% 신뢰구간은 $[11.7-2.093\frac{4.040584}{\sqrt{20}}, 11.7+2.093\frac{4.040584}{\sqrt{20}}]$

– 모평균 μ에 대한 95% 신뢰구간은 $[9.80897, 13.59013]$

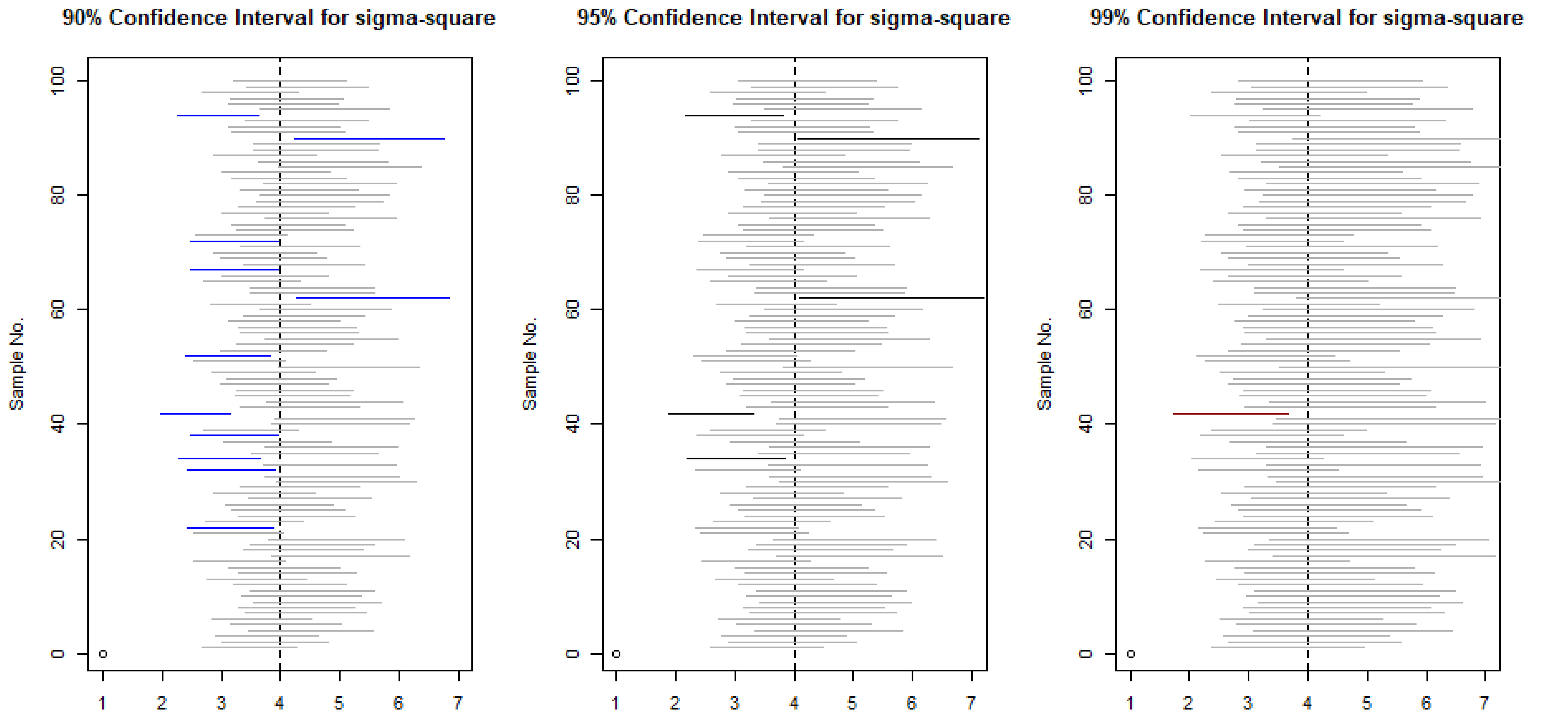
5. 모분산의 구간추정

– (실험3) 평균이 10이고 표준편차가 2인 모집단에서 표본크기가 100개인 1000개의 표본을 추출하여 모분산 σ^2에 대한 신뢰구간을 추정해 보면 모분산에 대한 95% 신뢰구간 10000개 중 100개만을 살펴보면 95개의 신뢰구간이 모평균 10을 포함하고 있는 것을 확인할 수 있음

– 모분산 σ^2이 확률변수가 아니고 고정된 상수이므로 σ^2에 대한 95% 신뢰구간의 의미는 표본크기가 동일한 10000개의 서로 다른 표본에 의해 동일한 공식으로 10000개의 신뢰구간을 구했을 때 그 중에서 95%의 구간이 모분산 σ^2를 포함한다고 볼 수 있다는 것임



– 모분산 σ^2에 대한 90%, 95% 및 99% 신뢰구간 100개 중 각각 89개, 95개, 99개의 신뢰구간이 모평균 10을 포함하고 있는 것을 확인할 수 있음



– 모집단이 정규분포에 따르더라도 n이 충분히 크지 않으면 표본분산 s^2은 정규분포에 따르지 않고 다음의 통계량을 가진 χ^2–분포에 따름.

$\cdot \chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

– χ^2–분포표로부터 $p\left(\chi^2_{0.975} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{0.025}\right) = 0.95$이므로

$$p\left(\frac{\chi^2_{0.975}}{(n-1)s^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{0.025}}{(n-1)s^2}\right) = 0.95$$

$$p\left(\frac{(n-1)s^2}{\chi^2_{0.025}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{0.975}}\right) = 0.95$$

– 모분산 σ^2에 대한 95% 신뢰구간은 $[\frac{(n-1)s^2}{\chi^2_{0.025}}, \frac{(n-1)s^2}{\chi^2_{0.975}}]$

– $(1-\alpha) \times 100\%$ 신뢰구간의 일반적인 표시 : $[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}]$

– (연습 3) K제약회사에서 두통약을 개발하였으며 이 약의 효과를 알아보기 위해 두통환자 10명을 임의로 선정하여 두통약을 복용하게 한 후 두통 억제 시간(단위 : 분)을 측정하였다. 두통 억제 시간이 정규분포할 때, 모분산에 대한 99%의 신뢰구간을 구하라.

⇒ χ^2–분포를 이용

66	37	18	31	85	63	73	83	65	80

– $S^2 = 547.4333,\ n = 10$

– χ^2–분포표로부터 $p\left(\chi^2_{0.995} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{0.005}\right) = 0.99$ 단, $\chi^2_{0.995} = 1.7349,\ \chi^2_{0.005} = 23.5894$

$$p\left(\frac{\chi^2_{0.995}}{(n-1)s^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{0.005}}{(n-1)s^2}\right) = 0.99$$

$$p\left(\frac{(10-1)\times 547.4333}{23.589} \le \sigma^2 \le \frac{(10-1)\times 547.4333}{1.734}\right) = 0.99$$

$$p(208.8607 \le \sigma^2 \le 2839.875) = 0.99$$

– 모분산 S^2 에 대한 99% 신뢰구간은 $[208.8607, 2839.875]$