UNISTUDY
유니스터디

박경은 교수님과 함께하는

확률과 통계

박경은 교수 제공

- 학습주제 -

PART 2. 통계적 추론: 추정과 검정
(Statistical Inference: Estimation and Testing)

7 통계적 추정
(Statistical Estimation)

7-2. 구간추정 (Interval Estimation)

7-2. 구간추정

7-2-1. 구간추정의 정의

(1) 구간추정(Interval Estimation)의 이해

① 점추정의 한계: 모수에 대한 점추정은 모수의 참값에 가까울 수 있지만, 모수의 참값과 완전히 일치하기는 매우 어려움. 이를 보완할 수 있는 방법으로 정해진 확신의 정도를 가지고 미지의 모수가 속할 것으로 기대되는 구간, 즉 신뢰구간을 추정

② 미지의 모수가 추정한 구간 안에 들어갈 확신의 정도인 신뢰수준에 따른 신뢰구간을 추정

(2) 신뢰수준과 신뢰구간

1) 신뢰수준(Confidence Level)

> 신뢰구간:
> 모수가 어느 구간 안에 있는지를 확률적으로 보여주는 방법

① 어떤 값이 알맞은 추정값이라고 믿을 수 있는 정도를 의미, 백분율(%)로 표현

② 신뢰수준(confidence level) $100(1-\alpha)\%$
: 모수가 추정한 구간 안에 들어갈 확신의 정도가 $100(1-\alpha)\%$ 라는 의미

③ 신뢰계수(Confidence Coefficient) $1-\alpha$: 신뢰구간이 모수를 포함할 확률

예 α가 0.1, 0.05, 0.01인 신뢰수준 90%, 95%, 99%

④ 신뢰수준을 90%, 95%, 98%, 99% 등으로 바꾸어 확신의 정도를 증가 또는 감소 가능

> 설문조사에서는 표본으로 계산한 추정치가 모수에서 벗어날 수 있는 최대 허용 오차를 오차범위라고 부르며, 통계학에서는 이를 오차한계(Margin of Error)라고 함. 일반적으로 신뢰구간 길이의 절반과 같음

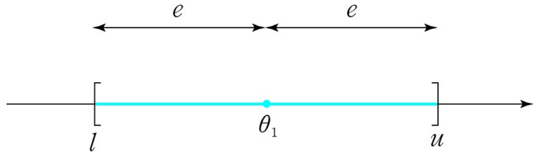
2) 신뢰구간(Confidence Interval)

① 정해진 확신의 정도(신뢰수준)를 가지고 미지의 모수 θ가 속할 것으로 기대되는 구간(신뢰구간)을 추정하는 방법
- 신뢰구간 = 모수를 포함한다고 확신하는 구간
- 확신한다 = '높은 확률을 가진다'는 의미

② 모수 θ에 대한 $100(1-\alpha)\%$ 신뢰구간(confidence interval) $[L,\ U]$
: 두 통계량 L(신뢰하한, lower bound)과 U(신뢰상한, upper bound)가 $P(L<\theta<U)=1-\alpha$를 만족하는 구간

예 모평균 μ 또는 모비율 p 등의 신뢰구간은 모수 θ에 대한 점추정량 $\hat{\theta}$의 관측값 θ_1을 중심으로 대칭인 구간 $[l=\theta_1-e,\ u=\theta_0+e]$을 사용

> 오차한계 (Margin of Error): 모수를 점추정 값으로부터 얼마나 벗어날 수 있는지의 최대 허용 오차 범위.
> 오차한계=(임계값)×(표준오차)



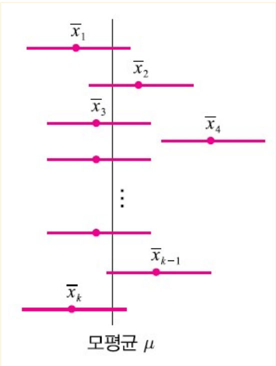
* 오차한계(Margin of Error) e

예 모집단이 정규분포일 때, 모분산 σ^2의 신뢰구간은 카이제곱분포가 비대칭 분포이므로 비대칭 구간 $\dfrac{(n-1)S^2}{\chi^2_{(1-\alpha)/2}} \le \sigma^2 \le \dfrac{(n-1)S^2}{\chi^2_{\alpha/2}}$ 임

③ 신뢰수준 $100(1-\alpha)\%$ 인 신뢰구간의 의미

- 동일한 모집단에서 동일한 추정방법으로 100회 반복하여 추정한 신뢰구간이 모수의 값을 $100(1-\alpha)$회 정도 포함하고 있을거라 기대된다는 의미
- 모수 θ에 대한 추정량 $\hat{\theta}$의 신뢰수준 $100(1-\alpha)\%$ 신뢰구간이 $[\hat{\theta}-e,\ \hat{\theta}+e]$일 때, $\hat{\theta}$의 오차 $|\hat{\theta}-\theta|$는 e를 초과하지 않는다고 $100(1-\alpha)\%$ 확신할 수 있음을 의미

$$P(|\hat{\theta}-\theta| \le e) = 1-\alpha$$



예 신뢰수준이 95%인 신뢰구간

- 동일한 모집단에서 동일한 크기의 표본을 추출하여 구한 신뢰구간이 모수의 값을 포함하고 있을 확률이 95%임
- 표본을 20번 추출하였을 때 20개의 신뢰구간 중에 95%, 즉 19개 정도가 모수 θ의 참값을 포함하고 5%에 해당하는 1개 정도는 모수 θ의 참값을 포함하지 않는다는 의미
- 95% 신뢰구간 [10, 11]이란 모수의 값을 포함하는 95개 중에 하나가 [10, 11]임

④ 다른 조건이 동일하다면 신뢰구간이 짧을수록 모수에 대한 추정의 정밀도가 높아지므로 신뢰구간은 짧을수록 바람직함

표본평균 $\overline{X}=100$에 대한 신뢰구간 (95, 105)에서

- 모평균이 100이라면, 95와 105 사이의 값이므로 확률 $P(95<\mu<105)=1$, 즉 100%
- 모평균이 110이라면, 95와 105 사이의 값이 아니므로 확률 $P(95<\mu<105)=0$, 즉 0%

⑤ 다른 조건이 동일하다면 표본의 크기가 클수록 신뢰구간의 폭은 좁아지며, 동일한 표본 크기에서는 신뢰수준이 높을수록 신뢰구간의 폭은 넓어짐

<표준정규분포 $N(0, 1)$에서 신뢰구간을 구할 때 자주 사용되는 Z값들>

신뢰계수$(1-\alpha)$	α	$\alpha/2$	$z_{\alpha/2}$
0.90	0.10	0.05	1.645
0.95	0.05	0.025	1.96
0.98	0.02	0.01	2.33
0.99	0.01	0.005	2.58

모분산이 알려진 경우에는 신뢰수준과 표본 크기가 동일하면 신뢰구간의 길이는 표본에 관계없이 일정. 반면, 모분산이 알려져 있지 않은 경우에는 표본표준편차에 의존하므로 신뢰구간의 길이가 표본마다 달라질 수 있음.

7-2-2. 모평균에 대한 구간추정

(1) 단일 모평균 μ에 대한 $100(1-\alpha)$% 신뢰구간

1) 모분산 σ^2이 알려진 정규모집단

① 모분산 σ^2이 알려진 정규모집단에서 모평균 μ의 $100(1-\alpha)\%$ 신뢰구간을 구하는 방법

모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간: 모분산 σ^2이 알려진 정규모집단

$$\left[\overline{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

($\overline{x}$: 크기 n인 확률표본의 표본평균 값)

설명 정규분포 $N(\mu,\sigma^2)$을 따르는 모집단으로부터 추출한 크기 n인 확률표본 $X_1, X_2, \cdots, X_n$의 표본평균 $\overline{X}=\frac{1}{n}(X_1+X_2+\ \cdots\ +X_n)$에 대하여 표준화된 통계량 $Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$는 표준정규분포를 따름. 즉,

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

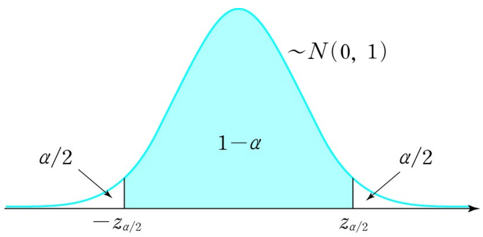
따라서 임의의 양수 $\alpha\,(0<\alpha<1)$에 대하여

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

$$\Leftrightarrow\ P\left(|\overline{X}-\mu|\le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

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

모평균 μ의 신뢰구간 표현
$[a-b,\ a+b]$,
$a-b\le\mu\le a+b$,
$\mu=a\pm b=[a-b,\ a+b]$
등



따라서 모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간의 오차한계는

$$e=z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

하한 L과 상한 U는 각각

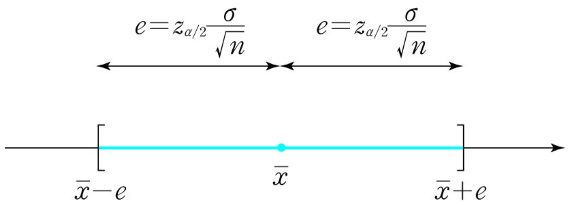
$$L=\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\quad U=\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

모집단에서 추출한 크기 n인 확률표본으로부터 관측값 $x_1, x_2, \cdots, x_n$을 얻었다고 할 때, 모평균 μ에 대한 점추정값은 표본평균

$$\overline{x}=\frac{1}{n}(x_1+x_2+\ \cdots\ +x_n)$$

이고 $100(1-\alpha)\%$ 신뢰구간은 $\bar{x}$를 중심으로 좌우대칭인 구간

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$ ▣

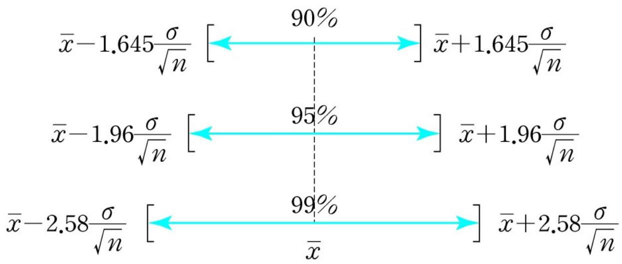


② μ에 대한 90%, 95%, 99% 신뢰구간

신뢰수준	$\overline{X}$의 표준오차	오차한계	신뢰구간
90%	$\sigma/\sqrt{n}$	$1.645\frac{\sigma}{\sqrt{n}}$	$\mu=\bar{x}\pm 1.645\frac{\sigma}{\sqrt{n}}$
95%	$\sigma/\sqrt{n}$	$1.96\frac{\sigma}{\sqrt{n}}$	$\mu=\bar{x}\pm 1.96\frac{\sigma}{\sqrt{n}}$
99%	$\sigma/\sqrt{n}$	$2.58\frac{\sigma}{\sqrt{n}}$	$\mu=\bar{x}\pm 2.58\frac{\sigma}{\sqrt{n}}$

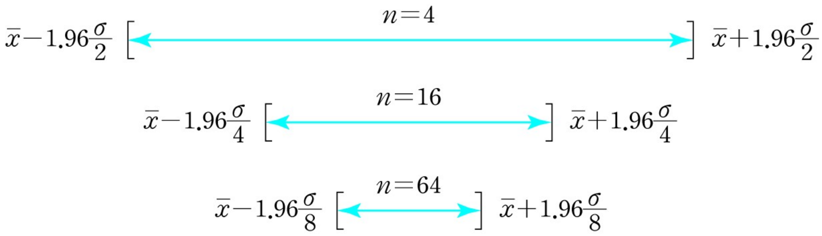
$\alpha=0.1,\ 0.05,\ 0.01$
에 대하여 $z_{\alpha/2}$값은
$z_{0.05}=1.645$,
$z_{0.025}=1.96$,
$z_{0.005}=2.58$

③ 표본의 크기가 같을 때 신뢰수준이 커질수록 신뢰구간은 넓어짐



④ 신뢰수준이 고정되면 표본의 크기가 커질수록 신뢰구간은 좁아짐.

예 신뢰수준 95%인 경우, 표본의 크기가 16일 때의 신뢰구간의 크기는 4일 때의 1/2이고 64일 때의 2배



예 $\sigma^2 = 16$인 정규모집단에서 모평균을 추정하기 위하여 크기가 25인 확률표본을 추출하였다. 표본평균이 19일 때 모평균에 대한 95% 신뢰구간과 99% 신뢰구간을 각각 구하시오.

풀이 $\bar{x} = 19$, $\sigma^2 = 16$, $n = 25$이므로 표본평균 $\bar{X}$의 분산은 $\dfrac{\sigma^2}{n} = \dfrac{16}{25}$이고

$z_{0.025} = 1.96$이므로 95% 오차한계는

$$e = z_{0.025}\frac{\sigma}{\sqrt{25}} = 1.96\frac{4}{5} = 1.568.$$

따라서 95% 신뢰구간은

$$19 \pm 1.568 = [17.432, 20.568].$$

같은 방법으로 99% 오차한계와 신뢰구간은

$$e = z_{0.005}\frac{\sigma}{\sqrt{25}} = 2.58\frac{4}{5} = 2.064, \quad 19 \pm 2.064 = [16.936, 21.064]$$

▣

2) 모분산 σ^2이 알려진 비정규모집단

모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간 : 모분산 σ^2이 알려진 모집단

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

($\bar{x}$: 크기 n인 확률표본의 표본평균 값)

설명 모평균과 모분산이 각각 μ, σ^2인 모집단에 대하여 표본의 크기 n이 클 때($n \geq 30$) 모집단의 분포에 상관없이 중심극한정리에 의해 $Z = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$는 근사적으로 표준정규분포를 따름

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

따라서 임의의 모집단에 대하여 모분산 σ^2이 알려져 있을 때 모평균 μ의 (근사)신뢰구간은 모집단이 정규분포를 따를 때와 동일하게 정의됨

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

▣

예 모분산 σ^2이 8로 알려진 모집단에서 크기가 32인 표본을 추출한 결과가 다음과 같다.

5.7	5.1	6.2	6.4	5.6	5.2	5.1	5.5
5.0	5.5	7.3	5.2	4.9	5.2	5.5	5.7
5.4	5.2	7.2	6.8	5.8	5.2	5.3	7.0
6.0	5.8	7.8	6.2	5.9	4.9	6.5	5.5

(a) 모평균 μ에 대한 점추정값을 구하시오.
(b) 모평균 μ에 대한 90% 신뢰구간을 근사적으로 구하시오.

풀이

(a) 표본평균 $\bar{x} = 5.8$을 점추정값으로 사용한다.

(b) 90% 오차한계는 근사적으로 다음과 같다.

$$e = z_{0.05}\frac{\sigma}{\sqrt{32}} = 1.645\frac{2\sqrt{2}}{4\sqrt{2}} = 0.8225$$

따라서 90% 신뢰구간은 근사적으로

$$5.8 \pm 0.8225 = [4.9775, 6.6225]$$

▣

모분산 σ^2이 알려져 있는 모평균 μ의 신뢰구간
i) 정규모집단인 경우, 표본 크기와 무관하게 Z−신뢰구간 사용
ii) 비정규모집단이며 대표본이면 중심극한정리에 의해 Z−신뢰구간 사용
iii) 비정규모집단이며 소표본이면 모평균에 대한 모수적 추론이 어려움

** iii) 경우, 실무 / 응용 통계에서는 가능한 대안
◇ 비모수적 방법: 모평균 자체보다는 중앙값에 대한 신뢰구간
◇ 부트스트랩 신뢰구간: 표본에서 재표본추출로 경험적 분포 구성
◇ 분포를 가정: 예: 로그정규, 지수분포 등

【참고】 모분산의 알려진 경우, 모집단의 정규성 여부와 표본의 크기에 따른 신뢰구간

모집단 분포	표본 크기	신뢰구간
정규	작음	Z-신뢰구간 가능
정규	큼	Z-신뢰구간 가능
비정규	큼	Z-신뢰구간 가능 (중심극한정리에 의해)
비정규	작음	계산 불가능

3) 모평균 μ의 신뢰구간: 모분산 σ^2이 알려지지 않은 경우 표본분산 S^2로 대체

① 정규모집단인 경우(표본크기와 상관없이): t−분포

모평균의 $100(1-\alpha)\%$ 신뢰구간: 모분산을 모르는 정규모집단

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

확률표본의 크기 : n

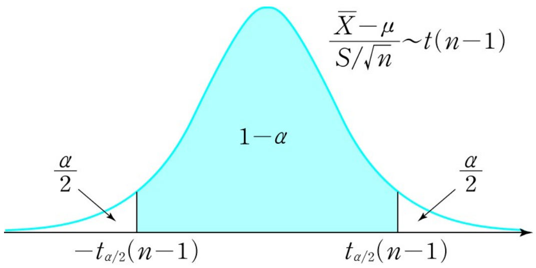
확률표본의 표본평균과 표본표준편차 : $\bar{x}$, S

모분산을 모르는 경우 모표준편차 σ 대신 표본표준편차의 값 S를 사용하고 표준정규분포 대신 자유도가 $(n-1)$인 t-분포 사용

설명 t-분포는 y축에 대하여 대칭이므로

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

$t_{\alpha/2}(n-1)$: 자유도가 $n-1$인 t-분포의 오른쪽 꼬리 확률이 $\frac{\alpha}{2}$인 t-값



$$\left(P(T > t_{\alpha/2}(n-1)) = \frac{\alpha}{2},\quad P(T < -t_{\alpha/2}(n-1)) = \frac{\alpha}{2}\right)$$

$T = \dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ 이므로 $\overline{X}$를 이용하여 다시 정리하면

$$P\left(-t_{\alpha/2}\frac{S}{\sqrt{n}} \le \overline{X}-\mu \le t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

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

모집단에서 추출한 크기 n인 확률표본의 값 $x_1, x_2, \cdots, x_n$을 얻었다고 할 때

모평균 μ에 대한 점추정값은 표본평균 $\overline{x} = \frac{1}{n}(x_1+x_2+\cdots+x_n)$,

표본편차 S에 대한 점추정값은 표본표준편차 $s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x})^2}$.

모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간의 오차한계는

$$e = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

이고 $100(1-\alpha)\%$ 신뢰구간은 $\overline{x}$를 중심으로 대칭인 구간

$$[\overline{x}-e,\ \overline{x}+e]$$ ▣

② 정규모집단이 아니지만 $n \ge 30$인 대표본: 중심극한정리 적용 근사적 $Z-$분포

$$\left[\overline{X}-z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X}+z_{\alpha/2}\frac{S}{\sqrt{n}}\right] \ (\text{단, } S\text{: 표본표준편차})$$

표본크기가 크면 $t-$분포가 Z분포와 유사해짐을 이용

설명 n이 크면($n \ge 30$) $t_{\alpha/2}(n-1) \dot{\sim} z_{\alpha/2}$이므로 오차한계 $e = z_{\alpha/2}\frac{s}{\sqrt{n}}$를 사용하여 $100(1-\alpha)\%$ 신뢰구간을 근사적으로 계산 ▣

③ 정규모집단이 아니면서 $n < 30$인 소표본: 모수적 신뢰구간 계산 불가능

예 휴대전화 배터리의 지속시간을 추정하기 위하여 100개의 배터리를 표본으로 하여 지속시간을 조사하였다. 표본평균이 51시간, 표본분산이 20(시간^2)일 때 배터리의 평균지속시간 μ에 대한 95% 신뢰구간을 구하시오.

풀이 모분산은 확인되지 않지만, 표본의 크기가 크므로($n=100$) 표준정규분포를 이용한다. 95% 오차한계는

$$z_{0.025}\frac{S}{\sqrt{n}} = 1.96\frac{\sqrt{20}}{10} = 0.88$$

이므로 95% 신뢰구간은

$$[51-0.88,\ 51+0.88] = [50.12,\ 51.88]$$ ▣

예 모분산이 알려지지 않은 정규모집단에서 크기가 16인 표본을 추출한 결과가 다음과 같다.

5.4	5.2	7.2	6.8	5.8	5.0	5.3	7.0
6.0	5.8	7.8	6.2	5.9	4.6	6.5	5.5

"등분산 가정"은 소표본 t-검정에서 필요한 전제: 정규모집단이면서 소표본인 경우에, 두 집단의 모분산이 같다고 믿을 만한 근거가 있을 때만 등분산 가정을 함. 근거가 없는 상태에서 무조건 가정하면 잘못된 자유도로 신뢰구간이 계산되고, 오류 가능성이 커짐

(a) 모평균 μ에 대한 점추정값을 구하시오.
(b) 표본분산을 구하시오.
(c) 모평균 μ에 대한 90%, 95% 오차한계와 신뢰구간을 구하시오.

풀이

(a) 표본평균 $\overline{x} = \dfrac{1}{16}(5.4+5.2+7.2+\cdots+5.5)=6$을 점추정값으로 사용

(b) 표본분산을 정의를 이용하여 구하면

$$\begin{aligned} S^2 &= \frac{1}{15}\sum_{k=1}^{16}(x_k-\overline{x})^2 \\ &= \frac{1}{15}\left((5.4-6)^2+(5.2-6)^2+\cdots+(5.5-6)^2\right)=0.76 \end{aligned}$$

(c) $n=16$이므로 자유도가 15인 t-분포를 사용. 자유도가 15인 t-분포에 대하여

$$t_{0.05}=1.75,\qquad t_{0.025}=2.13$$

이므로 90%, 95% 오차한계는 각각

$$t_{0.05}\frac{s}{\sqrt{n}}=1.75\frac{\sqrt{0.76}}{\sqrt{16}}\fallingdotseq 0.38,\quad t_{0.025}\frac{s}{\sqrt{n}}=2.13\frac{\sqrt{0.76}}{\sqrt{16}}\fallingdotseq 0.46.$$

따라서 90%, 95% 신뢰구간은 각각

$$6\pm0.38=[5.62,\ 6.38],\quad 6\pm0.46=[5.54,\ 6.46]$$

■

(2) 독립표본: 두 모평균 차 $\mu_1-\mu_2$에 대한 $100(1-\alpha)$% 신뢰구간

1) 두 모분산 σ_1^2, σ_2^2이 알려진 경우

① 두 정규모집단인 경우 (표본의 크기와 관계없음): Z-분포

$$\left[(\overline{X}_1-\overline{X}_2)-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ (\overline{X}_1-\overline{X}_2)+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

(단, σ_1^2, σ_2^2: 두 모분산)

설명 두 모집단이 독립이고 각각 평균이 μ_1, μ_2, 분산이 σ_1^2, σ_2^2인 정규분포를 따른다고 가정. 각 모집단에서 크기가 n_1, n_2인 확률표본을 추출하면 이들의 표본평균은 각각

$$\overline{X}_1=\frac{1}{n_1}(X_{11}+\cdots+X_{1n_1}),\qquad \overline{X}_2=\frac{1}{n_2}(X_{21}+\cdots+X_{2n_2}).$$

두 모평균의 차이 $\mu_1-\mu_2$에 대한 점추정량은 표본평균의 차이 $\overline{X}_1-\overline{X}_2$을 사용.

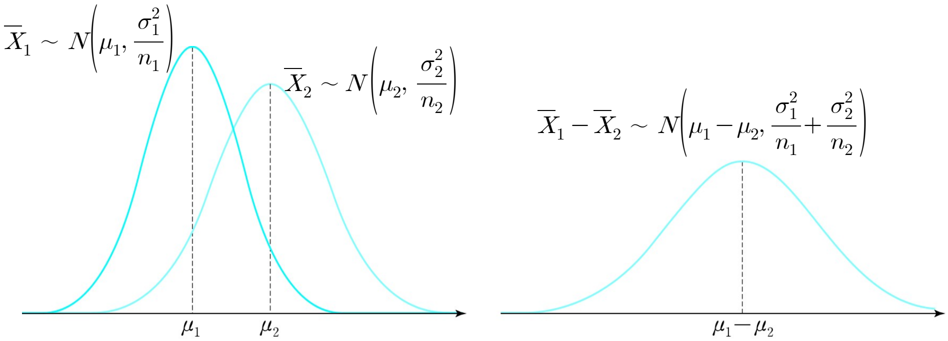
$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$이므로 $\overline{X}_1 - \overline{X}_2$은 $\mu_1 - \mu_2$의 불편추정량임.

표본평균 $\overline{X}_1$, $\overline{X}_2$에 대하여

$$\overline{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \qquad \overline{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

이고 두 표본이 독립이므로 $\overline{X}_1 - \overline{X}_2$의 분포는 다음과 같은 정규분포를 따름.

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$



따라서 표준화된 확률변수 Z는 표준정규분포를 따름

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1).$$

표준정규분포의 성질로부터 임의의 양수 $\alpha\,(0 < \alpha < 1)$에 대하여 다음 식이 성립

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} \le Z \le z_{\alpha/2}). \\ &= P\left(-z_{\alpha/2} \le \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \le z_{\alpha/2}\right) \\ &= P\left(\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right). \end{aligned}$$

이 식에서 모평균의 차 $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간의 오차한계는

$$e = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$100(1-\alpha)\%$ 신뢰구간은

$$\left[\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$ ▣

예 고속도로 휴게소와 서울의 휘발유 가격(단위 : 천원/L)에 차이가 있는지 알아보기 위하여 각각 서울에서 12개소, 고속도로에서 10개소를 무작위로 선정하여 조사한 결과가 다음과 같다. 이때 서울과 고속도로의 휘발유 가격은 독립이고 모표준편차가 각각 0.08, 0.06인 정규분포를 따른다고 한다.

서울	1.55	1.43	1.54	1.48	1.59	1.81	1.47	1.52	1.6	1.58	1.62	1.65
도속도로	1.46	1.47	1.52	1.56	1.49	1.53	1.62	1.43	1.54	1.58		

(a) 서울과 고속도로 휘발유 가격의 평균의 차에 대한 점추정값을 구하시오.
(b) 서울과 고속도로 휘발유 가격의 평균의 차에 대한 95% 신뢰구간을 구하시오.

풀이

(a) 서울의 휘발유 가격의 표분평균 값을 $\overline{x}$, 휴게소의 휘발유 가격의 표본평균 값을 $\overline{y}$라 하면

$$\overline{x} = \frac{1}{12}(1.55 + 1.43 + \cdots + 1.65) = 1.57$$

$$\overline{y} = \frac{1}{10}(1.46 + 1.47 + \cdots + 1.58) = 1.52$$

따라서 서울과 고속도로 휘발유 가격의 평균의 차에 대한 점추정값은

$$\overline{x} - \overline{y} = 1.57 - 1.52 = 0.05.$$

(b) 표본의 크기는 각각 $n = 12,\ m = 10$이므로 $\overline{X} - \overline{Y}$의 분산은

$$\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m} = \frac{0.08^2}{12} + \frac{0.06^2}{10} = 0.0009 = 0.03^2$$

$\alpha = 0.05$이면 $z_{0.025} = 1.96$이므로 95% 신뢰구간의 오차한계는

$$e = 1.96 \cdot 0.03 = 0.0588.$$

따라서 95% 신뢰구간은

$$\overline{x} - \overline{y} \pm e = [-0.0088,\ 0.1088]$$ ▣

② 두 비정규모집단, $n_1,\ n_2 \geq 30$인 대표본: (중심극한정리에 의해) 근사 Z-분포

$$\left[(\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

(단, σ_1^2, σ_2^2: 두 모분산)

③ 두 비정규모집단, $n_1,\ n_2 < 30$인 소표본: 비모수적 방법

* 부트스트랩(bootstrap)을 이용하면 비정규/소표본 환경에서 평균 또는 중앙값 차 신뢰구간 추정 가능

2) 두 모분산 σ_1^2, σ_2^2이 알려지지 않은 경우

① 두 정규모집단, n_1, $n_2 \geq 30$인 대표본: 근사적으로 Z-분포

$$\left[(\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\ (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right]$$

(단, S_1^2, S_2^2: 두 표본분산)

*등분산 가정 불필요

설명 독립인 두 모집단이 각각 정규분포를 따르고 모분산 σ_1^2, σ_2^2에 대하여 알려진 것이 없다고 가정. 크기가 각각 n_1, n_2(모두 ≥ 30)인 표본의 표본분산을 각각 S_1^2, S_2^2이라 할 때 표준화된 확률변수는 표준정규분포를 따름

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,\ 1).$$

Z에서 σ_1^2, σ_2^2 대신 S_1^2, S_2^2을 사용하면 근사적으로 표준정규분포를 따름

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \stackrel{\cdot}{\sim} N(0, 1).$$

따라서 표본분산의 값이 s_1^2, s_2^2일 때 두 모평균의 차이 $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$의 오차한계는

$$e = z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

또한 두 모평균의 차이 $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간은 근사적으로

$$\left[\overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

예 원래 생산되던 엔진과 새로 생산되는 엔진의 연비를 비교하는 실험이 이루어졌다. 사용된 휘발유와 기타 실험조건은 동일한 상태에서 기존 엔진과 새로운 엔진 모두 50회 실험한 결과 평균 연비는 각각 $\overline{x} = 11.1$, $\overline{y} = 11.9$ (단위 : km)이고 표본분산은 각각 $s_1^2 = 0.12$, $s_2^2 = 0.09$이었다. 새 엔진과 기존 엔진의 연비의 차이에 대한 90% 신뢰구간을 구하시오.

풀이 기존 엔진의 연비 평균을 μ_1, 새 엔진의 연비 평균을 μ_2라 하면 $\mu_2 - \mu_1$의 점추정값은

$$\overline{y} - \overline{x} = 11.9 - 11.1 = 0.8.$$

두 표본의 크기가 모두 50으로 크므로 모평균의 차 $\mu_2 - \mu_1$에 대한 90% 오차한계는 근사적으로

$$e = z_{0.05}\sqrt{\frac{0.12}{50} + \frac{0.09}{50}} = 1.645\sqrt{\frac{0.12}{50} + \frac{0.09}{50}} = 0.1066.$$

따라서 $\mu_2 - \mu_1$의 90% 신뢰구간은

$$0.8 \pm 0.1066 = [0.6934, 0.9066]$$

② 두 정규모집단, $n_1,\ n_2 < 30$인 소표본, $\sigma_1^{\ 2}=\sigma_2^{\ 2}$인 경우:
합동분산 활용, 자유도 n_1+n_2-2인 t-분포

$$\left[(\overline{X}_1-\overline{X}_2)-t_{\alpha/2}(n_1+n_2-2)\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\right.$$
$$\left.(\overline{X}_1-\overline{X}_2)+t_{\alpha/2}(n_1+n_2-2)\ S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ \right]$$

(단, $S_1^{\ 2}$, $S_2^{\ 2}$: 두 표본분산, $S_p^{\ 2}=\dfrac{(n_1-1)S_1^{\ 2}+(n_2-1)S_2^{\ 2}}{n_1+n_2-2}$: 합동분산)

합동분산 $S_p^{\ 2}$은 모분산의 추정량으로 $S_1^{\ 2}$과 $S_2^{\ 2}$의 표본의 크기에 비례한 가중평균.

예

$S_1^{\ 2}=300$, $S_2^{\ 2}=370$, $n_1=10$, $n_2=20$일 때, 합동분산 $S_p^{\ 2}$은 $\dfrac{9\times300+19\times370}{28}=347.5$.

합동표준편차 S_p는 합동분산의 음이 아닌 제곱근임

설명 독립인 두 모집단이 정규분포를 따르고 모분산 σ_1^2, σ_2^2을 모르지만 $\sigma_1^2=\sigma_2^2=\sigma^2$이라 가정. 표본평균을 각각 $\overline{X}_1$, $\overline{X}_2$라 하면

$Var(\overline{X}_1-\overline{X}_2)=\dfrac{\sigma^2}{n_1}+\dfrac{\sigma^2}{n_2}$ 이므로 다음 확률변수는 표준정규분포를 따름

$$Z=\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{\sigma\sqrt{1/n_1+1/n_2}}\sim N(0,\ 1).$$

표본의 크기가 각각 n_1, n_2이고 표본분산을 각각 S_X^2, S_Y^2라 할 때 알려지지 않은 공통분산 σ^2의 추정량은 다음과 같은 합동표본분산(pooled sample variance)을 사용

$$S_p^2=\frac{\sum_{k=1}^{n_1}(X_{1k}-\overline{X}_1)^2+\sum_{k=1}^{n_2}(X_{2k}-\overline{X}_2)^2}{n_1+n_2-2}=\frac{(n_1-1)S_{X_1}^2+(n_2-1)S_{X_2}^2}{n_1+n_2-2}.$$

Z-통계량에서 σ 대신 S_p를 대입한 확률변수

$$T=\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}}$$

는 자유도가 $(n+m-2)$인 t-분포를 따름. 자유도가 (n_1+n_2-2)인 t-분포의 오른쪽 꼬리 확률이 $\dfrac{\alpha}{2}$인 값을 $t_{\alpha/2}$라 하면

$$1-\alpha=P(-t_{\alpha/2}\le T\le t_{\alpha/2}).$$
$$=P\left(-t_{\alpha/2}\le\frac{\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}}\le t_{\alpha/2}\right)$$
$$=P\left(\overline{X}_1-\overline{X}_2-t_{\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le\mu_1-\mu_2\le\overline{X}_1-\overline{X}_2+t_{\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

두 모평균의 차이 $\mu_1-\mu_2$에 대한 점추정치와 $100(1-\alpha)\%$의 오차한계는 각각

$$\overline{x}_1-\overline{x}_2,\quad e=t_{\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

($\overline{x}_1$, $\overline{x}_2$: 추출된 확률표본에서 얻어지는 표본평균,
s_p^2: 추출된 확률표본에서 얻어지는 합동표본분산의 값)

합동표본분산은 σ^2의 불편추정량($E(S_p^2)=\sigma^2$)

또한 두 모평균의 차이 $\mu_1-\mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$[\overline{x}_1-\overline{x}_2-e,\ \overline{x}_1-\overline{x}_2+e].$$

예 기존에 생산되던 엔진과 새로 생산되는 엔진의 연비를 비교하는 실험이 이루어졌다. 사용된 휘발유와 기타 실험조건은 동일한 상태에서 기존 엔진은 10회, 새로운 엔진은 12회 실험한 결과 평균연비는 각각 $\overline{x}=11.1$, $\overline{y}=11.9$(단위 : km/L)이고 표본분산은 각각 $s_1^2=0.12$, $s_2^2=0.09$이었다. 새 엔진과 기존 엔진의 연비의 차이에 대한 90% 신뢰구간을 구하시오. 단, 두 엔진의 연비는 서로 독립이고 모분산이 같은 정규분포를 따른다고 가정한다.

풀이

기존 엔진의 연비 평균을 μ_1, 새 엔진의 연비 평균을 μ_2라 하면 $\mu_2-\mu_1$의 점추정 값은 $\overline{y}-\overline{x}=11.9-11.1=0.8$. $n=10$, $m=12$에 대하여 합동표본분산 S_p^2의 값은

$$s_p^2=\frac{(n-1)s_1^2+(m-1)s_2^2}{n+m-2}=\frac{9\cdot 0.12+11\cdot 0.09}{20}=0.1035$$

$s_p=\sqrt{0.1035}=0.3217$이고 t-분포표에서 자유도 20일 때 $t_{0.05}$의 값은

$$t_{0.05}(20)=1.72$$

이므로 90% 오차한계는

$$e=t_{\alpha/2}s_p\sqrt{\frac{1}{10}+\frac{1}{12}}=1.72\cdot 0.3217\cdot\sqrt{\frac{11}{60}}=0.2369.$$

따라서 90% 신뢰구간은

$$[0.8-0.2369,\ 0.8+0.2369]=[0.5631,\ 1.0369]$$

③ 두 정규모집단, n_1, $n_2<30$인 소표본, $\sigma_1{}^2\neq\sigma_2{}^2$인 경우: Welch t-분포

$$\left[(\overline{X}_1-\overline{X}_2)-t_{\alpha/2}(welch\,df)\sqrt{\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}},\ (\overline{X}_1-\overline{X}_2)+t_{\alpha/2}(welch\,df)\sqrt{\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}}\right]$$

(단, $S_1{}^2$, $S_2{}^2$: 두 표본분산)

* 자유도 df는 Welch-Satterthwaite 방정식으로 근사 계산

$$\text{자유도 } df\fallingdotseq\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$$

SPSS에서는 Levene 검정을 통해 등분산 여부를 확인하고, 그 결과에 따라 풀링 t(등분산 가정) 또는 Welch t(등분산 불가)를 선택하여 자유도를 자동으로 계산

④ 두 비정규모집단, n_1, $n_2\geq 30$인 대표본: 중심극한정리에 의해 근사적으로 Z-분포

$$\left[(\overline{X}_1-\overline{X}_2)-z_{\alpha/2}\sqrt{\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}},\ (\overline{X}_1-\overline{X}_2)+z_{\alpha/2}\sqrt{\frac{S_1{}^2}{n_1}+\frac{S_2{}^2}{n_2}}\right]$$

(단, $S_1{}^2$, $S_2{}^2$: 두 표본분산)

대응표본(paired sample): 같은 대상에서 두 번 측정되었거나, 서로 짝지을 수 있는 두 관측값으로 이루어진 표본
- 약물 복용 전 · 후 혈압 비교
- 다이어트 프로그램 전 · 후 체중
- 일란성 쌍둥이 실험
- 동일 조건의 두 제품 비교 등

설명 모집단이 정규분포를 따르지 않더라도 표본의 크기가 크면 중심극한정리에 의하여 $\overline{X}_1 - \overline{X}_2$는 근사적으로 정규분포를 따름. 따라서 모평균의 차이 $\mu_1 - \mu_2$의 구간추정은 위의 결과들을 그대로 이용하여 근사적으로 유도

⑤ 두 비정규모집단, $n_1,\ n_2 < 30$인 소표본: 모수적 방법 불가

(3) 대응표본(Paired samples): 두 모평균 차 $\mu_1 - \mu_2$에 대한 $100(1-\alpha)\%$ 신뢰구간

- 각 관측값이 서로 짝을 이루는 경우, 두 값의 차이 $D_i = X_i - Y_i$를 분석

예 동일한 사람의 치료 전 · 후 결과, 쌍둥이나 부부와 같이 짝을 맞춘 실험

① 대응표본의 수 $n \geq 30$인 대표본: 중심극한정리에 의해 근사 Z-분포

$$\left[\overline{D} - z_{\alpha/2}\frac{S_D}{\sqrt{n}},\ \overline{D} + z_{\alpha/2}\frac{S_D}{\sqrt{n}}\right]$$

(단, D: 각 표본요소의 값들의 차이, $\overline{D}$: D의 평균, S_D: D의 표준편차)

② 대응표본의 수 $n < 30$인 소표본: (정규성 가정 가능) 자유도 $n-1$인 t-분포

대응표본에 의한 $(\mu_1 - \mu_2)$의 신뢰구간

모평균이 각각 $\mu_1,\ \mu_2$이고 정규분포를 따르는 두 모집단

$d_k = x_k - y_k,\ \ k = 1, 2, \cdots, n$: n개의 대응표본 값의 차이

$$\overline{d} = \frac{1}{n}(d_1 + d_2 + \cdots + d_n)$$

$$\left[\overline{D} - t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}},\ \overline{D} + t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}\right]$$

(단, D: 각 표본요소의 값들의 차이, $\overline{D}$: D의 평균, S_D: D의 표준편차)

설명 표본의 크기가 n이고 처리 1을 한 결과인 X_k는 정규분포 $N(\mu_1, \sigma_1^2)$을 따르고 처리 2를 한 결과인 Y_k는 정규분포 $N(\mu_2, \sigma_2^2)$를 따른다고 하자. 순서쌍

$$(X_k,\ Y_k),\ k = 1, 2, \cdots, n$$

는 서로 독립이지만 일반적으로 X_k와 Y_k는 상관관계를 가짐. 다시 말해서

$D_1, D_2, \cdots, D_n$은 서로 독립이고 평균과 분산이 각각

$$\mu_D = \mu_1 - \mu_2,\quad \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\,Cov(X_1,\ Y_1)$$

인 정규분포를 따름. 따라서 $D_1, D_2, \cdots, D_n$의 표본평균

$$\overline{D} = \frac{1}{n}(D_1 + D_2 + \cdots + D_n)$$

은 정규분포 $N\left(\mu_1 - \mu_2, \dfrac{\sigma_D^2}{n}\right)$을 따르고 이를 표준화한 Z-통계량은 표준정규분포를 따름

$$Z = \frac{\overline{D} - (\mu_1 - \mu_2)}{\sigma_D / \sqrt{n}} \sim N(0,1).$$

따라서 σ_D^2이 알려져 있으면 표준정규분포의 성질로부터 임의의 양수 $\alpha\,(0<\alpha<1)$에 대하여

$$\begin{aligned} 1-\alpha &= P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) \\ &= P\left(|\overline{D} - (\mu_1 - \mu_2)| \le z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}\right) \\ &= P\left(\overline{D} - z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}} \le \mu_1 - \mu_2 \le \overline{D} + z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}\right) \end{aligned}$$

이므로 d_k가 D_k의 측정값이고 $\overline{d} = \frac{1}{n}(d_1 + d_2 + \cdots + d_n)$이면 $\mu_1 - \mu_2$의

$100(1-\alpha)\%$ 신뢰구간은

$$\left[\overline{d} - z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}},\ \overline{d} + z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}\right].$$

σ_D^2이 알려지지 않은 경우, σ_D^2 대신 표본분산

$$S_D^2 = \frac{1}{n-1}\sum_{k=1}^{n}(D_k - \overline{D})^2$$

을 사용한 통계량 T는 자유도가 $(n-1)$인 t-분포를 따름

$$T = \frac{\overline{D} - (\mu_1 - \mu_2)}{S_D / \sqrt{n}} \sim t(n-1).$$

따라서 $\overline{D}$의 측정값이 $\overline{d}$, 표본분산 S_D^2의 측정값이 s_D^2이면 $(\mu_1 - \mu_2)$의

$100(1-\alpha)\%$ 신뢰구간은

$$\left[\overline{d} - t_{\alpha/2}\frac{s_D}{\sqrt{n}},\ \overline{d} + t_{\alpha/2}\frac{s_D}{\sqrt{n}}\right].$$

이때 $t_{\alpha/2}$는 자유도가 $(n-1)$인 t-분포의 오른쪽 꼬리 확률이 $\frac{\alpha}{2}$인 t-값이고 s_D^2은

표본분산의 값

$$s_D^2 = \frac{1}{n-1}\sum_{k=1}^{n}(d_k - \overline{d})^2$$

② 대응표본의 수 $n<30$인 소표본: (정규성 가정 불가능으로) 비모수적 방법 이용

(또는) 신뢰구간이 필요하면 부트스트랩(bootstrap)으로 평균/중앙값 차 신뢰구간 계산 또는 순위 기반 신뢰구간 방법을 따로 적용 가능

예 고단백저탄수화물 다이어트의 체중감소 효과를 확인하기 위하여 참가자 중 20명을 임의로 추출하여 다이어트를 시작할 때의 체중 x_k와 다이어트 3개월 후의 체중 y_k를 측정한 결과가 다음과 같다.

k	x_k	y_k	k	x_k	y_k
1	55.2	56.0	11	82.5	79.4
2	67.5	66.1	12	77.4	78.2
3	73.5	65.5	13	63.5	59.9
4	88.7	79.2	14	79.9	81.3
5	76.5	72.3	15	55.8	55.5
6	92.4	85.4	16	77.5	76.3
7	61.3	63.2	17	73.5	68.5
8	58.4	55.2	18	82.4	78.3
9	103.6	91.2	19	73.2	73.5
10	58.4	59.6	20	91.2	88.4

μ_1, μ_2를 각각 고단백저탄수화물 다이어트를 하는 사람들의 시작할 때와 다이어트 3개월 후의 체중의 모평균이라 할 때 $\mu_D = \mu_1 - \mu_2$의 95% 신뢰구간을 구하시오. 단, 다이어트 시작 전과 3개월 후의 체중은 모두 정규분포를 따른다고 가정한다.

풀이 체중의 변화량을 $d_k = x_k - y_k$로 놓으면 다음 표를 얻는다.

x_k	y_k	d_k	x_k	y_k	d_k
55.2	56.0	- 0.8	82.5	79.4	3.1
67.5	66.1	1.4	77.4	78.2	- 0.8
73.5	65.5	8.0	63.5	59.9	3.6
88.7	79.2	9.5	79.9	81.3	- 1.4
76.5	72.3	4.2	55.8	55.5	0.3
92.4	85.4	7.0	77.5	76.3	1.2
61.3	63.2	- 1.9	73.5	68.5	5.0
58.4	55.2	3.2	82.4	78.3	4.1
103.6	91.2	12.4	73.2	73.5	- 0.3
58.4	59.6	- 1.2	91.2	88.4	2.8

$\mu_D = \mu_1 - \mu_2$의 점추정값은 $d_k,\ k = 1, 2, \cdots, n$의 표본평균

$$\overline{d} = \frac{1}{20}(-0.8 + 1.4 + \cdots + 2.8) = 2.97.$$

또한 표본표준편차는

$$s_D = \sqrt{\frac{1}{19}\sum_{k=1}^{20}(d_k - \overline{d}\,)^2} = 3.92.$$

자유도가 19일 때 $t_{0.025}(19) = 2.09$이므로 95% 오차한계는

$$e = t_{\alpha/2}\frac{s_D}{\sqrt{n}} = 2.09 \cdot \frac{3.92}{\sqrt{20}} = 1.83.$$

$\mu_D = \mu_1 - \mu_2$의 95% 신뢰구간은

$$\left[\overline{d} - t_{\alpha/2}\frac{s_D}{\sqrt{n}},\ \overline{d} + t_{\alpha/2}\frac{s_D}{\sqrt{n}}\right] = [2.97 - 1.83,\ 2.97 + 1.83]$$

$$= [1.14,\ 4.80]$$

따라서 고단백저탄수화물 다이어트 결과 체중 감소가 1.14kg과 4.80kg 사이임을 95%로 확신 가능 ▣

[정리] 표본 조건에 따른 모평균 신뢰구간 방법 정리

:모평균 및 평균 차이에 대한 신뢰구간 방법의 선택

구분	모분산	표본크기	모집단 정규성	등분산	공식/방법
하나의 모평균	알려짐	$n \geq 30$ (대표본)	상관없음	-	$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$: Z−신뢰구간
	알려짐	$n < 30$ (소표본)	정규	-	$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$: Z−신뢰구간
	미지	$n \geq 30$ (대표본)	상관없음	-	$\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$: t−근사 신뢰구간
	미지	$n < 30$ (소표본)	정규	-	$\overline{X} \pm t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}$: t−신뢰구간
	미지	$n < 30$ (소표본)	비정규	-	비모수/부트스트랩
두 모평균 차 (독립)	알려짐	$n_1,\ n_2 \geq 30$ (대표본)	상관없음	불필요	$(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$: Z−신뢰구간
	미지	$n_1,\ n_2 \geq 30$ (대표본)	상관없음	불필요	$(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$: t−근사 신뢰구간
	미지	$n_1,\ n_2 < 30$ (소표본)	정규	가능	$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}(n_1+n_2-2)\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$: $pooled\ t$−신뢰구간
	미지	$n_1,\ n_2 < 30$ (소표본)	정규	불가능	Welch t, Welch-Satterthwaite 근사 자유도
	미지	$n_1,\ n_2 < 30$ (소표본)	비정규	-	비모수/부트스트랩 사용
두 모평균 차 (대응)	미지	$n \geq 30$ (대표본)	상관없음	-	$\overline{D} \pm z_{\alpha/2}\frac{S_D}{\sqrt{n}}$: t−근사 신뢰구간
	미지	$n < 30$ (소표본)	정규	-	$\overline{D} \pm t_{\alpha/2}(n-1)\frac{S_D}{\sqrt{n}}$: t−신뢰구간
	미지	$n < 30$ (소표본)	비정규	-	비모수/부트스트랩 사용

CLT가 필요한 경우는
- 표본크기: 대표본
- 모집단 정규성: 상관없음
- 표본평균(또는 평균의 차)을 사용

즉, 정규성 가정 없이 정규근사를 쓰는 경우

두 모평균 차(독립)에서 비모수 방법이나 부트스트랩을 사용하는 경우 '등분산 가정' 자체를 보통 두지 않음. 즉, 필요/불필요를 구분할 대상이 아님

7-2-3. 모비율에 대한 구간추정

(1) 단일 모비율 p에 대한 $100(1-\alpha)$% 근사 신뢰구간: Z-분포

① 단일 모비율 p에 대한 근사 신뢰구간

$$\left[\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \quad (\text{단, } \hat{p}\text{: 표본비율})$$

설명 무한 모집단에서 어떤 속성을 가지는 개체의 비율이 p라 하자. 이 모집단에서 크기가 n인 확률표본을 추출하였을 때 이 속성을 가진 표본의 개수를 X라 하면 확률변수 X는 이항분포 $Bin(n,p)$를 따름. 이때 표본비율(sample ratio)을 나타내는 통계량

$$\hat{p}=\frac{X}{n}$$

을 모비율 p의 추정량으로 사용. 이항확률변수 X의 평균과 분산은

$$E(X)=np,\quad Var(X)=np(1-p).$$

따라서 표본비율 $\hat{p}$의 평균과 분산은 각각

$$E(\hat{p})=p,\quad Var(\hat{p})=\frac{p(1-p)}{n}.$$

표본비율 $\hat{p}$은 모비율 p의 불편추정량이고, 표준오차는

$$SE(\hat{p})=\sqrt{\frac{p(1-p)}{n}}.$$

표본의 크기 n이 크면 중심극한정리(드무아브르 라플라스 정리)에 의하여 확률변수 X는 근사적으로 정규분포 $N(np,np(1-p))$를 따름

$$X \stackrel{\cdot}{\sim} N(np,np(1-p)).$$

모비율 p가 알려져 있지 않으므로 표준오차에서 모비율 p 대신 불편추정량 $\hat{p}$을 사용하면

$$Z=\frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}}=\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

는 근사적으로 표준정규분포를 따름. 따라서 임의의 양수 $\alpha\,(0<\alpha<1)$에 대하여

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

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

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

따라서 모비율 p에 대한 $100(1-\alpha)$% 오차한계는

$$e=z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

신뢰구간은

$$\left[\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$ ▣

예 시청률 조사기관에서 텔레비전을 소유한 1000가구를 대상으로 대통령 후보 TV 토론회 시청여부를 조사한 결과 $x = 177$가구가 시청하는 것으로 나타났다. 이 토론회의 실제 시청률 p에 대한 95% 신뢰구간을 구하시오.

풀이 시청률 p의 점추정값은

$$\hat{p} = \frac{177}{1000} = 0.177$$

$z_{0.025} = 1.96$이므로 p의 95% 오차한계는

$$e = z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.177 \cdot 0.823}{1000}} = 0.0237$$

이고 p의 95% 신뢰구간은 $0.177 \pm 0.0237 = [0.1533, 0.2007]$ ▣

② 점추정값 $\hat{p}$은 모비율 p 사이의 오차 $|p-\hat{p}|$가 $100(1-\alpha)\%$로 오차한계 $e = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ 보다 작을 것임을 확신

(2) 두 모비율 차 $p_1 - p_2$에 대한 $100(1-\alpha)\%$ 근사 신뢰구간: $Z-$분포

$$\left[(\hat{p}_1-\hat{p}_2)-z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ (\hat{p}_1-\hat{p}_2)+z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$ (단, $\hat{p}_1$, $\hat{p}_2$: 두 표본비율)

설명 독립인 두 모집단에 대하여 어떤 같은 속성을 가지는 모비율이 각각 p_1, p_2, 두 모집단에서 각각 n_1, n_2개의 확률표본을 독립적으로 추출하였을 때 주어진 속성을 가지는 개체 수를 각각 X, Y라 하면 표본비율은

$$\hat{p}_1 = \frac{X}{n_1}, \quad \hat{p}_2 = \frac{Y}{n_2} \quad .$$

두 모집단이 독립이므로 두 표본비율 $\hat{p}_1$, $\hat{p}_2$도 독립이고

$$X \sim B(n_1, p_1), \quad Y \sim B(n_2, p_2)$$

이므로 표본비율의 차 $\hat{p}_1 - \hat{p}_2$의 기댓값과 분산은

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

따라서 표본비율의 차 $(\hat{p}_1 - \hat{p}_2)$은 모비율의 차 $(p_1 - p_2)$의 불편추정량이고, 표준오차는

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

이다. 두 표본의 크기 n_1, n_2이 크면 $(\hat{p}_1 - \hat{p}_2)$을 표준화한 Z-통계량은 중심극한정리에 의하여 근사적으로 표준정규분포를 따름

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \stackrel{\cdot}{\sim} N(0,1)$$

알려져 있지 않은 모비율 p_1, p_2 대신 불편추정량 $\hat{p}_1$, $\hat{p}_2$을 사용하여 표준오차의 추정값

$$S = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

을 사용하면

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

는 근사적으로 표준정규분포를 따름. 따라서 모비율의 차 $p_1 - p_2$에 대한 $100(1-\alpha)\%$ 오차한계는

$$e = z_{\alpha/2} S = z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

이고 신뢰구간은

$$\left[(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ \ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

▣

예 같은 제품을 생산하는 두 공장 A, B에서 불량률은 각각 10%, 6%라 한다. 임의로 공장 A의 제품 90개와 공장 B의 제품 94개를 조사하였을 때 공장 A의 제품의 불량률이 더 높을 확률을 구하시오.

풀이 X, Y를 표본에 포함된 불량품의 개수라 하면 공장 A의 표본의 불량률은 $\hat{p}_1 = \dfrac{X}{90}$, 공장 B의 표본의 불량률은 $\hat{p}_2 = \dfrac{Y}{94}$.

$$E(\hat{p}_1 - \hat{p}_2) = 0.1 - 0.06 = 0.04$$

$$\sqrt{Var(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{0.1 \cdot 0.9}{90} + \frac{0.06 \cdot 0.94}{94}} = \sqrt{0.0016} = 0.04 = \frac{1}{25}$$

이므로 $25(\hat{p}_1 - \hat{p}_2 - 0.04)$는 근사적으로 표준정규분포를 따름. 따라서

$$P(\hat{p}_1 > \hat{p}_2) = P(\hat{p}_1 - \hat{p}_2 > 0) = P(25(\hat{p}_1 - \hat{p}_2 - 0.04) > -1)$$
$$\stackrel{\cdot}{\sim} P(Z > -1) = P(Z < 1) = 0.8413$$

▣

예 한국갤럽이 2016년 12월 셋째 주(13~15일) 전국 성인 1,004명에게 요즘 어느 방송사의 뉴스를 가장 즐겨보는지 물은 결과 J사가 45%로 압도적으로 1위를 차지하였다. 남성은 응답자 498명 중 229명이, 여성 응답자 506명 중 218명이 J사 뉴스를 가장 즐겨본다고 답했다. J사 뉴스를 가장 즐겨보는 남성과 여성의 비율의 차이에 대한 95% 신뢰구간을 구하시오.

풀이 J사 뉴스를 가장 즐겨보는 남성과 여성의 비율을 각각 p_1, p_2라 하자. 표본비율은 각각

$$\hat{p}_1 = \frac{229}{498} = 0.4598, \quad \hat{p}_2 = \frac{218}{506} = 0.4308.$$

따라서 95% 오차한계는

$$e = z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$$

$$= 1.96\sqrt{\frac{0.4598(1-0.4598)}{498} + \frac{0.4308(1-0.4308)}{506}} = 0.0615$$

따라서 95% 신뢰구간은

$$[\hat{p}_1 - \hat{p}_2 - e,\ \hat{p}_1 - \hat{p}_2 + e] = [0.0290 - 0.0615,\ 0.0290 + 0.0615]$$

$$= [-0.0325,\ 0.0905]$$

▣

7-2-4. 모분산에 대한 구간추정

(1) 단일 모분산 σ^2에 대한 $100(1-\alpha)\%$ 신뢰구간: 자유도 $n-1$인 χ^2-분포

모분산 σ^2의 $100(1-\alpha)\%$ 신뢰구간

모집단이 정규분포 $N(\mu, \sigma^2)$를 따르고 크기 n인 확률표본의 값이 $x_1, x_2, \cdots, x_n$. 표본평균이 $\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$, 표본분산이 $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2$일 때 σ^2에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\ \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right] \quad \text{(단, } S^2\text{: 표본분산)}$$

설명 모평균이 μ이고 모분산이 σ^2인 모집단에서 추출한 크기 n인 확률표본 $X_1, X_2, \cdots, X_n$의 표본분산 S^2은

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$

표본분산은 모분산의 불편추정량이며 모분산에 대한 점추정량으로 사용

$$E(S^2) = \sigma^2.$$

확률표본이 정규분포로부터 추출된 것이라면 통계량

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

는 자유도가 $(n-1)$인 χ^2-분포를 따름. 따라서 σ^2의 신뢰구간은 χ^2-통계량 V를 이용하여 구함. 즉 임의의 양수 $\alpha\,(0<\alpha<1)$에 대하여

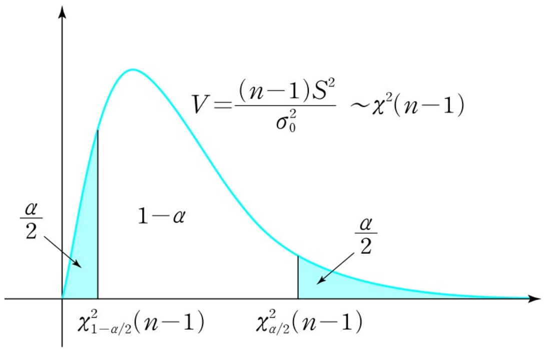
> $\chi^2_{1-\alpha/2}(n-1)$, $\chi^2_{\alpha/2}(n-1)$
> : 자유도가 $(n-1)$인 카이제곱분포에서 각각 왼쪽 넓이와 오른쪽 넓이가 $\frac{\alpha}{2}$인 값

$$P(\chi^2_{1-\alpha/2}(n-1) < V < \chi^2_{\alpha/2}(n-1)) = 1-\alpha.$$

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

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

> $P(V < \chi^2_{1-\alpha/2}(n-1)) = \frac{\alpha}{2}$
>
> $P(V > \chi^2_{\alpha/2}(n-1)) = \frac{\alpha}{2}$



따라서 추출된 표본의 표본분산 값이 $s^2\left(=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\overline{x})^2\right)$인 경우 σ^2에 대한 $100(1-\alpha)\%$ 신뢰구간은

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

표본표준편차 S는 모표준편차 $\sigma=\sqrt{\sigma^2}$에 대한 편의추정량이지만 표본의 크기가 커지면 편의가 0에 가까워짐. 따라서 모표준편차 σ에 대한 $100(1-\alpha)\%$ 신뢰구간은 모분산의 신뢰구간으로부터 다음과 같이 양끝 값에 제곱근을 취하여 사용 가능

예 모분산이 알려지지 않은 정규모집단에서 크기가 16인 표본을 추출한 결과가 다음과 같다. 모분산 σ^2과 모표준편차 σ에 대한 95% 신뢰구간을 구하시오.

5.4	5.2	7.2	6.8	5.8	5.0	5.3	7.0
6.0	5.8	7.8	6.2	5.9	4.6	6.5	5.5

풀이 표본평균과 표본분산은 각각

$$\overline{x}=\frac{1}{16}(5.4+5.2+7.2+\cdots+5.5)=6,\quad s^2=\frac{1}{15}\sum_{k=1}^{16}(x_k-\overline{x})^2=0.76$$

카이제곱분포표에서 자유도가 15일 때

$$\chi^2_{0.025}(15)=27.49,\quad \chi^2_{0.975}(15)=6.26$$

이므로 모분산에 대한 95% 신뢰구간은

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)},\ \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right]=\left[\frac{15\cdot 0.76}{27.49},\ \frac{15\cdot 0.76}{6.26}\right]$$
$$=[0.41,\ 1.82]$$

따라서 모표준편차에 대한 95% 신뢰구간은

$$[\sqrt{0.41},\ \sqrt{1.82}]=[0.64,\ 1.35]$$ ▣

(2) 독립인 두 정규모집단의 모분산 비 ${\sigma_1}^2/{\sigma_2}^2$에 대한 $100(1-\alpha)\%$ 신뢰구간: 자유도 n_1-1, n_2-1인 $F-$분포

$$\left[\frac{1}{F_{\alpha/2}(n_2-1,\ n_1-1)}\frac{S_1^{\ 2}}{S_2^{\ 2}},\ F_{\alpha/2}(n_1-1,\ n_2-1)\frac{S_1^{\ 2}}{S_2^{\ 2}}\right]$$

(단, $S_1^{\ 2}$, $S_2^{\ 2}$: 두 표본분산)

설명 (생략) ▣