11-2 추정2

점추정

확률표본의 정보를 이용하여 모수에 대한 특정 값을 지정하는 것

예 향후 엄청난 경제성장을 이룰 것으로 기대하는 서남아시아 지역 중 인도에 우리 기업들이 많이 진출하고 있다. 인도에 진출한 한국 기업 중 10개를 임의로 선정하여 투자액을 관찰하였더니 다음과 같았다고 가정하자.

표 투자액(단위:백만 달러)

30	5	44	9	48	4	30	80	18	264

- 모집단 평균과 분산에 대한 점추정값을 계산하라.
- 모집단의 집중화 경향을 추론하는 점추정량으로 표본평균 대신 표본중앙값, 표본최빈값을 이용하여 표본평균과 비교해 보라.

풀이

- 표본평균 $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{532}{10} = 53.2$

- 표본분산 $S^2 = \frac{1}{10-1}\sum_{i=1}^{10}(x_i - \bar{X})^2 = \frac{54279.6}{9} = 6031.0667$

따라서 모집단 평균의 점추정값은 53.2, 모집단 분산의 점추정값은 6031.0667이다.

- 표본중앙값과 표본최빈값을 이용한 모집단의 집중화 경향에 대한 점추정값은 모두 30으로 표본평균을 이용한 경우보다 그 추정값이 작다.

예 모비율 p의 추정량

여론조사에서 특정 후보에 대한 찬성비율과 같이 모집단의 비율 p를 추정하는 경우 이항분포를 이용해 모비율 p의 추정한다.

n명을 표본으로 추출하여 위와 같은 조사를 실시한다고 할 때, 표본은 $(X_1, X_2, \ldots, X_n)$과 같이 표현할 수 있으며, 특정 후보에 대한 찬성/반대 중 하나를 나타내는 확률변수이므로 이항확률변수의 정의에 의해 다음과 같이 표현할 수 있다.

$X_i=1$, i번째 사람이 찬성

$X_i=0$, i번째 사람이 반대

통계량 X를 $X = \sum_{i=1}^{n} X_i$와 같이 정의하면 X는 '표본으로 추출된 n명 중에서 찬성하는 사람의 수'를 의미하므로 전체 모집단에 있어서의 찬성률 p의 추정량은 다음과 같이 정의할 수 있다.

$$\hat{P} = \frac{\text{찬성하는 사람의 수}}{\text{표본의 수}} = \frac{X}{n}$$

따라서 표본비율 $\hat{p}$ 은 모비율 p의 점추정량이다.

구간추정

점추정량은 언제나 표본오차를 수반하기 때문에 전적으로 신뢰할 수 없다. 그러나 구간추정은 이런 점추정과 달리 모수가 빈번히 포함되는 범위를 제공하여 연구의 목적에 따라 원하는 만큼의 신뢰성을 가지고 모수를 추정할 수 있다.

신뢰구간

θ 는 알지 못하는 모수라고 하자. 표본정보에 근거하여 일정한 확률$(1-\alpha)$범위 내에 모수가 포함될 가능성이 있는 구간, 즉 다음을 만족하는 확률변수 A와 B를 구할 수 있다.

$$P(A<\theta<B)=1-\alpha$$

만약 확률변수 A와 B에 대한 측정값을 a와 b라고 하면,
구간 a < θ < b는 θ 에 대한 100$(1-\alpha)$% 신뢰구간이며
$1-\alpha$ 는 신뢰수준 그리고 α 는 유의수준이라고 한다.

모평균의 신뢰구간 추정 — 모집단 분산을 아는 경우

평균 μ 와 분산 σ^2 을 가지는 모집단이 있으며 μ는 모르고 σ^2 만 안다고 가정하자.

그리고 이 모집단으로부터 $x_1, x_2, \cdots, x_n$ 을 표본추출하였으며 σ^2 은 알지만 μ 는 모르기 때문에 신뢰구간을 통하여 μ 을 추정하고자 한다.

크기 n인 표본의 평균 $\overline{X}$ 는 다음과 같이 표준정규분포하는 확률변수 Z로 표준화될 수 있음을 이미 살펴보았다.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

신뢰수준이 $1-\alpha$ 가 되는 표준정규분포의 신뢰구간은

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

여기서 확률변수 $Z=(\bar{X}-\mu)/(\sigma/\sqrt{n})$ 이므로 위의 확률등식을 다음과 같이 다시 쓸 수 있다.

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

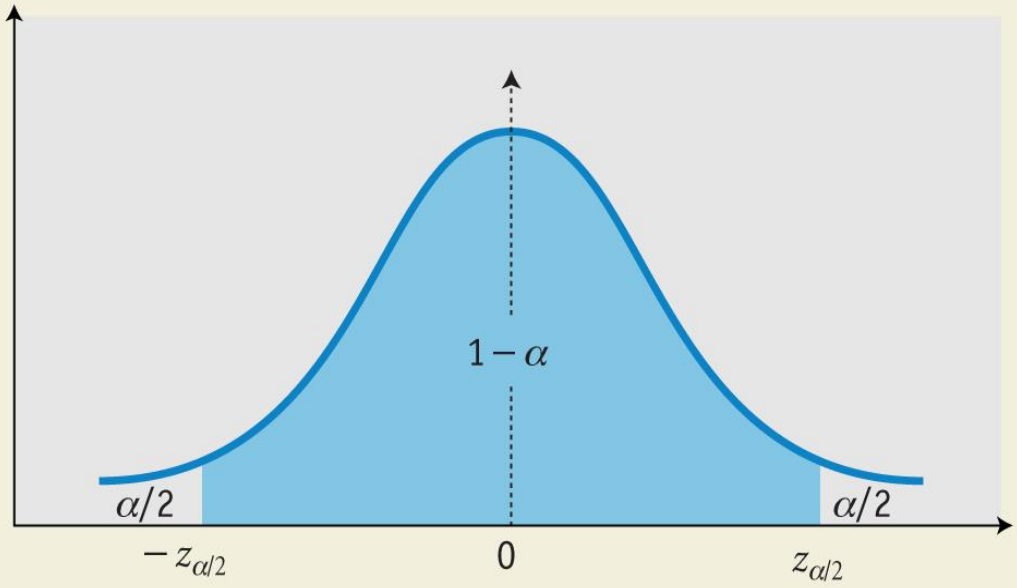
$$= P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

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

따라서 $a=\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$부터 $b=\left(\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ 까지의 임의구간이 모집단 평균 μ를 포함할 확률은 $1-\alpha$가 되므로 μ 에 대한 100($1-\alpha$)%의 신뢰구간은 다음과 같다.

$$\left(\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

그림 Z의 분포



예

어떤 축산업자가 목장에서 태어난 송아지들의 무게를 조사해 보니 분산이 36kg인 정규분포한다는 사실을 알았다. 갓 태어난 송아지 16마리를 임의로 선정하여 평균 중량을 계산해 보니 25kg이었다. 모집단 평균에 대한 90%신뢰구간을 구하라.

풀이

주어진 정보에 의하면 $\bar{X}=25,\ \sigma^2=36,\ n=16$ 이다.
90%신뢰구간의 신뢰수준은 $1-\alpha$ =0.9이므로 $\alpha/2=0.05$ 이다.
표준정규분포표로부터 $z_{0.05}=1.645$ 를 구할 수 있다.
따라서 μ 의 90%신뢰구간은

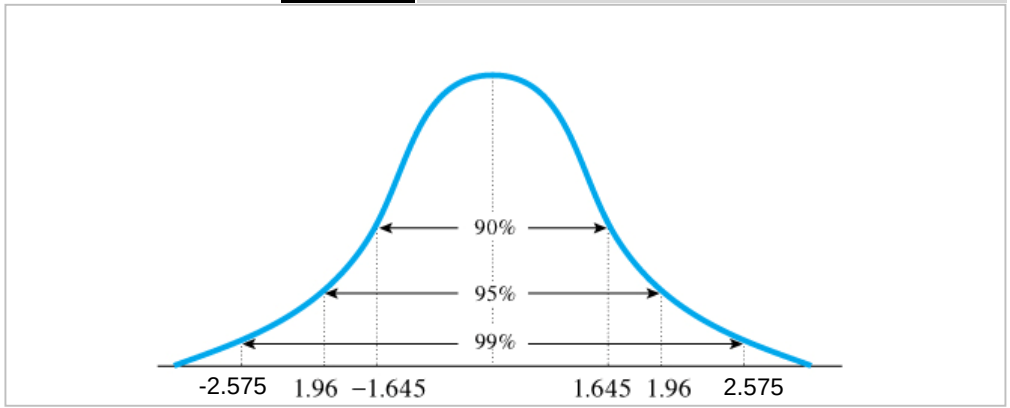
$$\bar{X}-1.645\frac{\sigma}{\sqrt{n}}<\mu<\bar{X}+1.645\frac{\sigma}{\sqrt{n}}$$

$$25-1.645\times\frac{6}{4}<\mu<25+1.645\times\frac{6}{4}$$

$$22.5325<\mu<27.4675$$

갓 태어난 송아지의 평균 중량은 90%신뢰수준에서 22.535kg과 27.4675kg 사이에 속한다.

그림 표준정규분포의 확률구간



신뢰수준 90%, 95%, 99%에 대한 표준정규확률변수 Z의 구간이 다음과 같음을 알 수 있다.

$$\Pr(-1.645 \le Z \le 1.645) = 0.9$$
$$\Pr(-1.96 \le Z \le 1.96) = 0.95$$
$$\Pr(-2.575 \le Z \le 2.575) = 0.99$$

예

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.

표 분당 산소 소비량(단위 : 리터)

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	

이 연구팀은 일반인들의 모집단은 정규분포하며 분산이 0.36이라는 사실을 알고 있다고 가정하자. 모집단 평균의 95%신뢰구간을 계산해 보자.

풀이

주어진 자료로부터 표본평균을 계산하면

$$\overline{X}=\sum_{i=1}^{35} x_i/35=0.71643$$

이다.

표본크기 n=35,모집단 표준편차 $\sigma=\sqrt{0.36}=0.6$ 이며, 95% 신뢰구간의 신뢰수준은$1-\alpha=0.95$이므로 $\alpha/2=0.025$가 되어 표준정규분포표로부터 $z_{0.025}=1.96$을 구할 수 있다. 따라서 μ의 95%신뢰구간은

$$\overline{X}-1.960\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+1.960\frac{\sigma}{\sqrt{n}}$$

$$0.71643-1.960\times\frac{0.6}{\sqrt{35}}<\mu<0.71643+1.960\times\frac{0.6}{\sqrt{35}}$$

$$0.5174<\mu<0.9155$$

이다. 그러므로μ는 95%신뢰수준에서 (0.5174,0.9155)에 속하게 된다.

모평균의 신뢰구간 추정 | 모집단 분산을 모르는 경우

①대표본의 경우

σ^2을 모르며 n이 충분히 클 때 μ에 대한 $100(1-\alpha)\%$ 신뢰구간

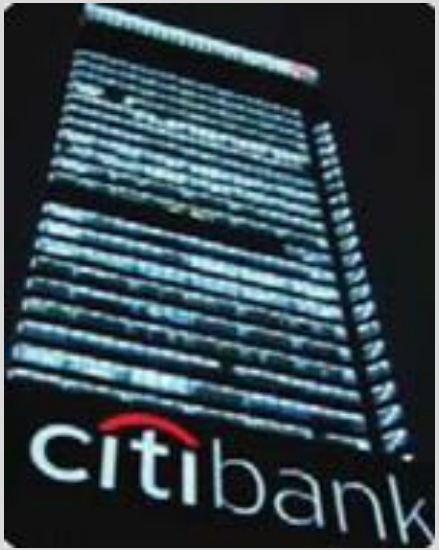
$$\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}} = \left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$$

예

금융기관에서 일하는 직장인 172명을 표본추출하여 현 직업을 선택한 중요한 요인들(직업의 안전성,보수, 사회 평판도 등)을 각각 1부터 5까지 평가하게 하였다.

숫자가 커질수록 중요도는 증가된다. 평가결과에 따르면 직업의 안정성에 대한 표본평균은 4.38, 표본 표준편차는 0.70이었으며, 직업의 안정성이 직업 선정 시 가장 중요한 요인으로 나타났다.

직업의 안정성 평가에 대한 모평균의 99%신뢰구간을 구하라.



풀이

모집단 분포에 대한 가정이 없지만 표본의 크기가 172로 충분히 크기 때문에 중심극한정리에 의해 표본평균의 표본분포가 정규분포에 근접하게 되며, σ 를 S로 대체할 수 있을 것이다. 문제에서 주어진 정보에 의하면

$$\overline{X}=4.38,\quad S=0.70,\quad n=172$$

이다.

또한 99%신뢰구간을 구하기 위해서는 $\alpha/2=0.005$, 그리고 표준정규분포표에 의하면 $z_{\alpha/2}=2.575$ 이다. 따라서 99%신뢰구간은

$$4.38-\frac{(2.575\times 0.7)}{\sqrt{172}}<\mu<4.38+\frac{(2.575\times 0.7)}{\sqrt{172}}$$

$$4.24<\mu<4.52$$

이다. 이는 μ가 4.24와 4.52사이에 존재할 신뢰도가 0.99임을 의미한다.

②소표본의 경우

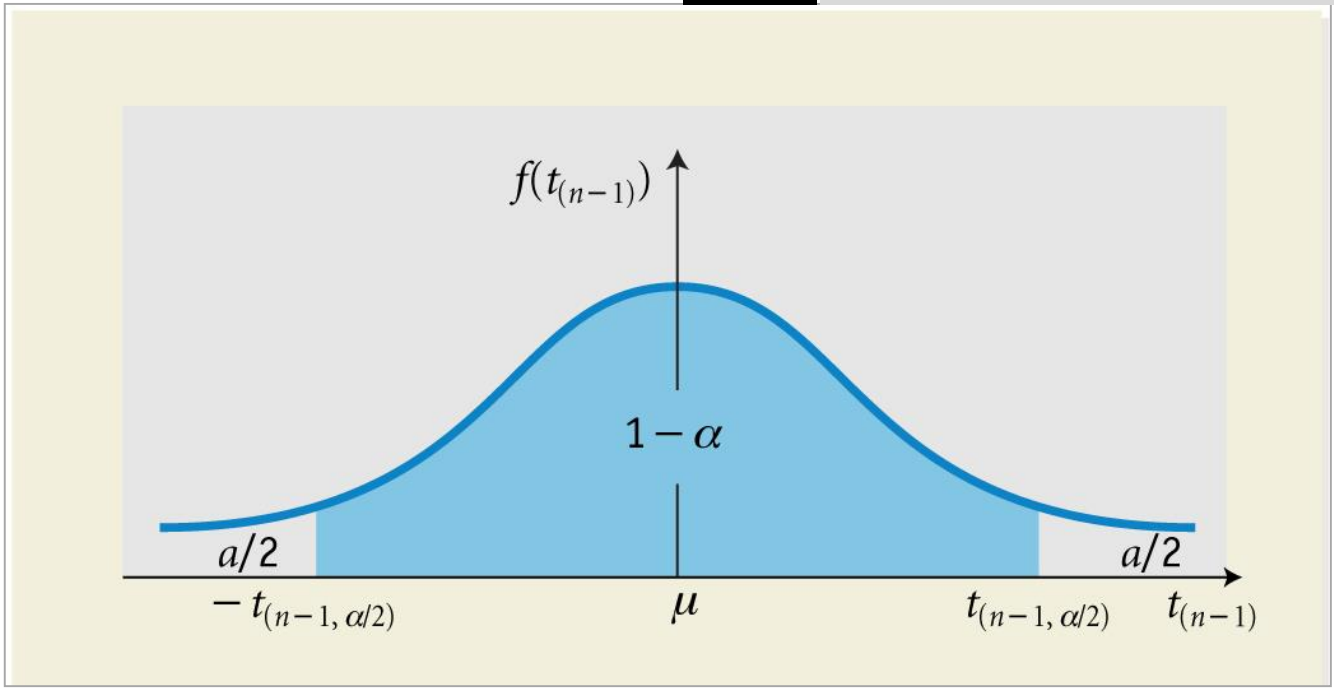
t-분포

모분산을 모르고 표본의 크기가 충분히 크지 않을 때, 확률변수

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

는 자유도 n-1의 t-분포(Student's t-distribution)를 이룬다.

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



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

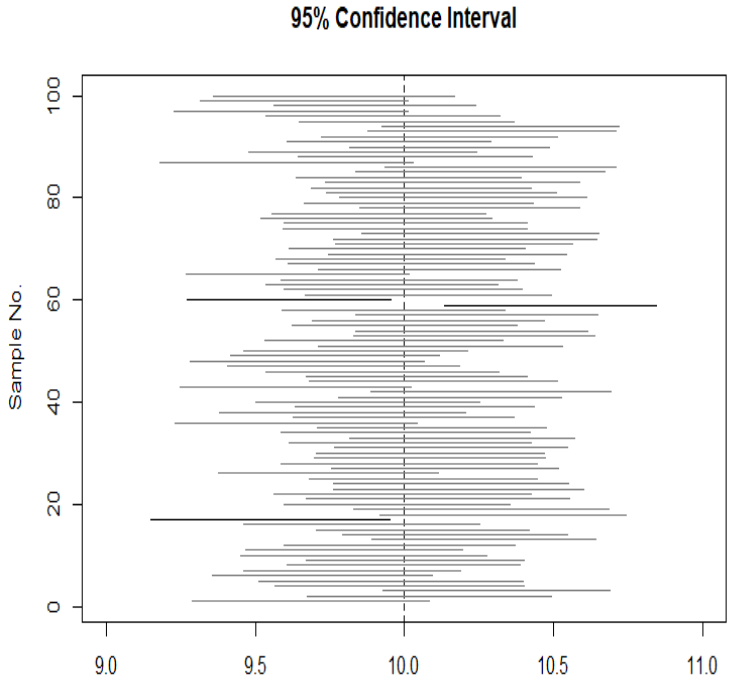
$$=P\left(-t_{(n-1,\,\alpha/2)}<\frac{\overline{X}-\mu}{S/\sqrt{n}}<t_{(n-1,\,\alpha/2)}\right)$$

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

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

σ^2을 모르며 n이 충분히 크지 않을 때 μ에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

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


95% Confidence Interval
Sample No.
0
20
40
60
80
100
9.0
9.5
10.0
10.5
11.0

예

한 피자 체인점의 지배인은 피자 배달시간이 오래 걸린다는 소비자들의 불평을 확인해 보기 위해 피자 배달주문 중 임의로 20개를 선정하여 배달시간을 측정하였더니 다음과 같았다.

표 피자 배달시간(단위 : 분)

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

모집단이 정규분포를 한다고 가정하고 모평균 배달시간에 대한 95%신뢰구간을 설정하라.

풀이

표의 자료를 이용하여 다음 값을 구할 수 있다.

$$\bar{X} = 11.7, \quad S = \sqrt{16.32632} = 4.041, \quad t_{(19,\,0.025)} = 2.0930$$

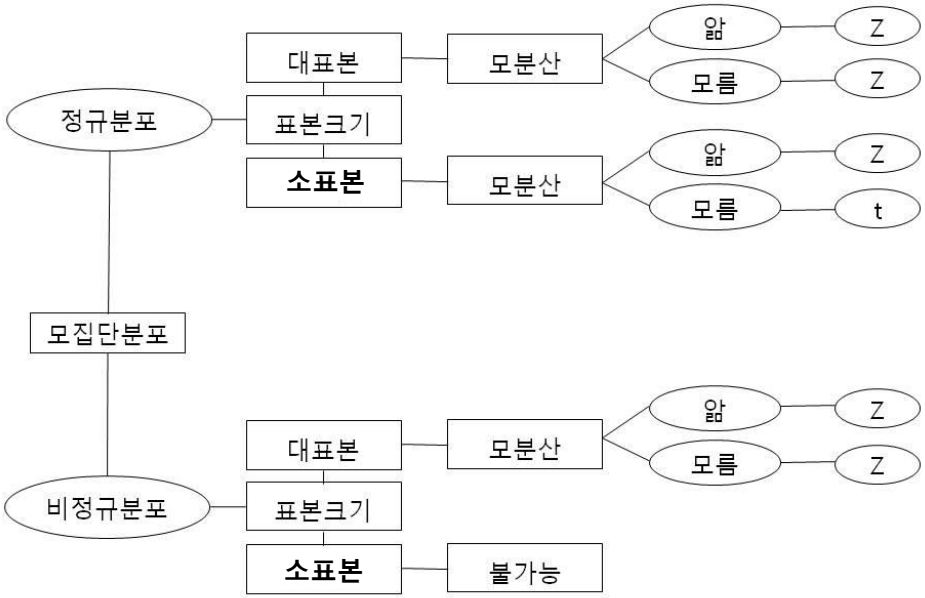
따라서 95%신뢰구간은

$$11.7 - \left(2.0930 \times \frac{4.041}{\sqrt{20}}\right) < \mu < 11.7 + \left(2.0930 \times \frac{4.041}{\sqrt{20}}\right)$$

$$9.8 < \mu < 13.6$$

이다. 95%신뢰수준을 가지고 지배인은 평균 피자 배달시간이 9.8~13.6분 사이가 될 것이라고 고객들에게 말할 수 있을 것이다.

모평균 구간추정을 위한 의사결정 트리



모분산의 신뢰구간 추정

확률변수 $\chi^2_{(n-1)}=\frac{(n-1)S^2}{\sigma^2}$ 은 자유도 n-1인 χ^2- 분포를 따른다.

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

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

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

정규모집단의 분산σ^2에 대한 100$(1-\alpha)$% 신뢰구간은 다음과 같다.

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

예

K제약회사에서 두통약을 개발하였으며
이 약의 효과를 알아보기 위해
두통환자 10명을 임의로 선정하여
두통약을 복용하게 한 후
두통 억제 시간을 측정하였다.

표 두통 억제 시간(단위 : 분)

66	37	18	31	85	63	73	83	65	80

두통 억제 시간이 정규분포할 때, 모분산에 대한 95%의 신뢰구간을 계산하라.

풀이

두통 억제 시간에 대한 10개의 관측값과 자유도 n-1=9를 이용하여 통계량을 계산하면

$$\bar{X}=60.1,\ S^2=4926.9/9$$

가 된다.

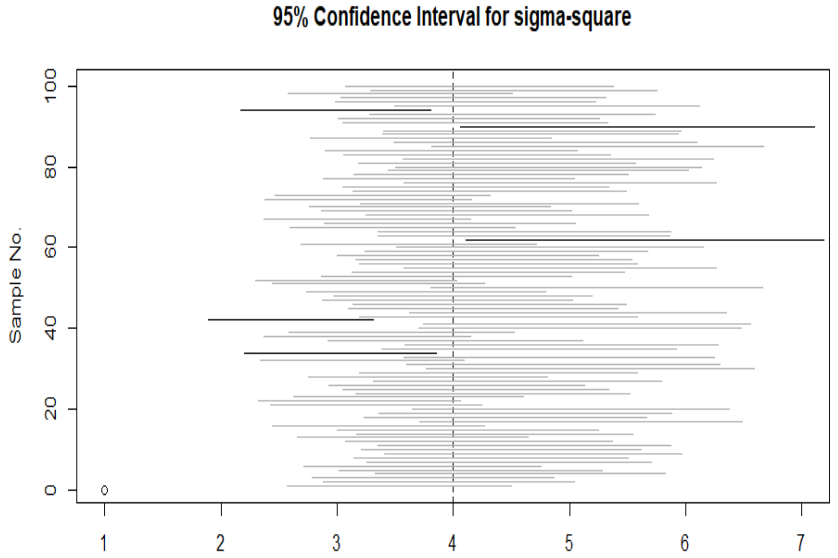
이 통계량의 값들을 정규모집단의 분산에 대한 신뢰구간 공식에 대입하면

$$\frac{4926.9}{19.022}<\sigma^2<\frac{4926.6}{2.7}$$
$$259.01<\sigma^2<1824.77$$

이 된다.

따라서 모분산에 대한 95%의 신뢰구간은 (259.01, 1824.77)이다.

- (실험2) 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 모분산 4를 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 모분산 4를 포함하고 있는 것을 확인할 수 있음