7.5절

10주1강 확률분포와 추정2

복습

중심극한정리

- 모집단의 평균이 μ이고 분산이 σ^2인 확률분포로부터 크기가 n인 확률표본 $(X_1, X_2, \ldots, X_n)$을 추출할 때, 표본평균 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$는 n이 클수록 평균이 μ이고, 분산이 $\frac{\sigma^2}{n}$인 정규분포와 근사한 분포를 갖는다.
- 즉, $\overline{X}$의 분포는 다음과 같이 표현한다.

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

- 만약에 확률표본 $(X_1, X_2, \ldots, X_n)$ 이 평균 μ와 분산 σ^2을 갖는 정규분포에서 추출되었다면, 표본평균 $\overline{X}$의 분포는 n의 크기에 관계없이 평균 μ와 분산 $\frac{\sigma^2}{n}$을 갖는 정규분포를 따른다.
- 즉, $\overline{X}$의 분포는 다음과 같이 나타낼 수 있다.

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

점추정

- 점추정이란 확률표본의 정보를 이용하여 모수에 대한 특정값을 지정하는 것을 말한다.
- 추정량은 여러 가지 형식으로 정의될 수 있는데, 예를 들면 $(X_1, X_2, \dots, X_n)$ 을 평균이 μ인 어떤 확률분포로부터 구한 확률표본이라고 할 때 모집단의 평균(모평균) μ의 추정량으로 생각할 수 있는 통계량은 다음과 같이 여러 가지가 있다.

- **모평균 μ 의 추정량**
- 표본평균, 표본중위수, 최소값, 최대값, 최소값과 최대값의 평균

- **추정량에 대해 고려해야할 사항**

① 추정량은 확률표본 $(X_1, X_2, \dots, X_n)$에 있는 확률변수 $X_1, \dots, X_n$의 함수 이므로 추정량도 또한 확률변수이다. **따라서 추정량도 특정한 확률분포를 갖는다.**

② 특정 모수에 대한 여러 가지 추정량 중에서 **가장 바람직한 추정량을 선택**하여야 한다.
가장 바람직한 추정량이란 추정량의 분포에서 분포의 중심이 추정하고자 하는 모수이고, 분포의 흩어진 정도가 작은 추정량을 말한다.

여러 가지 추정량

① 불편추정량 (unbiased estimator)

- 분포의 중심이 모수인 추정량

② 최소분산추정량 (minimum variance estimator)

- 분산이 가장 작은 추정량

③ 최소분산불편추정량 (minimum variance unbiased estimator)

- 위의 두 조건을 모두 만족하는 추정량.
- 추정량을 구할 때는 항상 최소분산불편추정량을 구하는 것이 바람직하다.

모분산의 추정량

- $(X_1, X_2 \ldots, X_n)$을 평균 μ와 분산 σ^2을 갖는 모집단으로부터의 확률표본이라고 할 때, 모분산 σ^2의 추정량은 다음과 같이 정의할 수 있다.

(i) 평균 μ가 알려져 있는 경우

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

(ii) 평균 μ가 알려져 있지 않은 경우

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

모비율 p의 추정량

- 자료분석에 있어서 특정법안에 대한 찬성비율과 같이 모집단의 비율 P를 추정하는 경우가 있다. 이러한 경우, 이항분포를 이용해 모비율 P의 추정한다.
- n명을 표본으로 추출하여 위와 같은 조사를 실시한다고 할 때, 표본은 $(X_1, X_2, \dots, X_n)$과 같이 표현할 수 있으며, 특정안건에 대한 찬성/반대 중 하나를 나타내는 확률변수이므로 이항확률변수의 정의에 의해 다음과 같이 표현할 수 있다.

$$X_i = \begin{cases} 1, & i\text{번째 사람이 찬성} \\ 0, & i\text{번째 사람이 반대} \end{cases}$$

- 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.

$$X \sim B(n, p), \quad \hat{P} = \frac{\text{찬성하는 사람의 수}}{\text{표본의 수}} = \frac{X}{n}$$

$$\Longrightarrow \quad E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}np = p, \; Var(\hat{P}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}npq = \frac{pq}{n}$$

- $\hat{P}$의 평균과 분산은 각각 $E(\hat{P}) = p, Var(\hat{P}) = \frac{pq}{n}$이며, n이 클 때 중심극한정리에 의하여 다음과 같이 정규분포를 따른다고 할 수 있다.

$$\hat{P} \sim N\left(p, \frac{pq}{n}\right)$$

t분포

- $(X_1, X_2, \cdots, X_n)$을 $N(\mu, \sigma^2)$으로부터의 확률표본이라고 할 때, 확률변수 T를 다음과 같이 정의한다.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

- 여기에서 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$으로 정의하면, T는 자유도(d.f.)가 $n-1$인 t분포를 따르며 $T \sim t(n-1)$로 표현한다.

t-분포의 확률계산

- t분포는 자유도 $n-1$에 따라 분포의 형태가 다른데 부록 III의 [표 4]에 각각의 자유도에 따른 t분포의 확률값이 주어져 있다.
- 각각의 자유도에서 확률 α에 대한 $\mathrm{P_r}(T \geq t_\alpha) = \alpha$가 되는 t_α값이 주어져 있는데 자유도가 5일 때 $t_{0.05} = 2.015$이고, 자유도가 9일 때 $t_{0.025} = 2.262$이다. 이를 식으로 표현하면 다음과 같다.

$$\mathrm{P_r}(T \geq 2.015) = 0.05\,,\ \mathrm{P_r}(T \geq 2.262) = 0.025$$

- 자유도가 30 이상인 경우는 inf.로 표현되어 있는데, 이 경우는 표준정규분포와 동일하다.

$$\mathrm{P_r}(T \geq 1.645) = 0.05\,,\ \mathrm{P_r}(T \geq 1.96) = 0.025$$

χ^2분포

- 확률변수 $Z_1, Z_2, \cdots, Z_n$이 서로 독립적으로 표준정규분포 $N(0,1)$을 따를 때, $Z_1, Z_2, \cdots, Z_n$의 제곱합 $\sum_{i=1}^{n} Z_i^2$은 자유도가 n인 χ^2-분포를 따른다.
- 자유도가 n인 χ^2-분포의 평균과 분산은 다음과 같다.
- 즉, $X^2 = \chi^2_{(n)}$일 때 $E(X) = n,\ Var(X) = 2n$이다.

χ^2–분포의 확률계산

- 부록 V의 [표 5]에는 각각의 자유도에 따르는 χ^2-분포의 확률이 주어져 있는데, t-분포와 같이 ${\chi^2}_{\alpha}$는 χ^2-분포에서 그 값 이상의 확률이 α인 값으로 다 음과 같이 표현할 수 있다.

$$\mathrm{P_r}(X^2 \geq {\chi^2}_{\alpha}) = \alpha$$

- 즉, ${\chi^2}_{0.995}$란 χ^2-분포에서 그 값 이상의 확률이 99.5%인 점의 위치를 의미한다. 자유도가 5인 경우와 자유도가 19인 경우의 확률값은 다음과 같이 구할 수 있다.

$$\mathrm{P_r}(X^2 \geq 1.145476) = 0.95\ ,\mathrm{P_r}(X^2 \geq 11.0705) = 0.05$$
$$\mathrm{P_r}(X^2 \geq 6.84398) = 0.995\ ,\mathrm{P_r}(X^2 \geq 27.2036) = 0.100$$

7.5절 구간추정

구간추정(interval estimation)

- 점추정은 "모수가 특정한 값일 것이다" 라고 선언하는 것으로 사실상 추정이 얼마나 정확한가를 판단하기가 불가능하다. 이러한 점추정의 정확성을 보완하는 방법이 구간추정이다.
- 구간추정이란 확률로 표현된 믿음의 정도 하에서 모수가 특정한 구간에 있을 것이라고 선언하는 것이다. 구간추정을 하려면 아래의 두 가지가 주어져야 한다.

① 추정량의 분포에 대한 전제가 주어져야 한다.

② 구하여진 구간 안에 모수가 있을 가능성의 크기가 주어져야 한다.

*** 신뢰수준(confidence level)**

- 구하여진 구간 안에 모수가 있을 가능성의 크기로, 일반적으로 90%, 95%, 99%의 확률을 이용하는 경우가 많다.

*** 신뢰구간(confidence interval)**

- 각각의 신뢰수준 하에서 모수가 존재할 것이라고 구한 구간을 신뢰구간이라 한다.

모평균 μ의 신뢰구간

- 확률표본 $(X_1, X_2, \ldots, X_n)$을 $N(\mu, \sigma^2)$으로부터의 확률표본이라고 할 때, 모평균 μ의 점 추정량은 표본평균 $\overline{X}$임을 알고 있다. 모평균에 대한 신뢰구간은 $\overline{X}$의 분포를 이용하여 구하는데 모분산 σ^2이 알려져 있는 경우와 알려져 있지 않은 경우를 구분하여 다음과 같이 신뢰구간을 구한다.

① 모분산 σ^2이 알려져 있는 경우

- 모평균 μ의 $(1-\alpha)\times 100\%$ 신뢰구간은 다음과 같다.

$$\left(\overline{X} - Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

② 모분산 σ^2 이 알려져 있지 않은 경우

- 모평균 μ의 $(1-\alpha)\times 100\%$ 신뢰구간은 다음과 같다.

  (a) $n > 30$인 경우

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

- (b) $n \le 30$인 경우

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

- 여기에서 $Z_{\alpha/2}$는 $N(0,1)$에서 $P_r\left(Z \ge Z_{\alpha/2}\right) = \frac{\alpha}{2}$인 값이고,

  $t_{\alpha/2}$는 자유도가 $n-1$인 t분포에서 $P_r\left(T \ge t_{\alpha/2}\right) = \frac{\alpha}{2}$인 값을 말한다.

표준정규분포의 확률구간

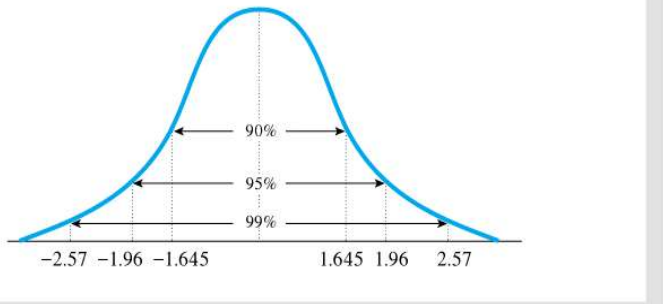
- 부록의 정규확률분포표를 이용해 신뢰수준 90%, 95%, 99%에 대한 표준정규확률변수 Z의 구간을 위의 모평균 μ의 신뢰구간을 구하는 공식에 적용하면 다음과 같이 나타낼 수 있다. 각 신뢰구간을 그림으로 표현하면 다음과 같다.

$$P_r(-1.645 \leq Z \leq 1.645) = 0.90$$

$$P_r(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P_r(-2.57 \leq Z \leq 2.57) = 0.99$$

그림 9-11 표준정규분포의 확률구간



- 여기에서 각각의 신뢰수준에 따른 신뢰구간은 표준오차인 $\frac{\sigma}{\sqrt{n}}$ 앞의 계수만 서로 다름을 알 수 있는데, 1.645, 1.96, 2.57 등을 각각 90%, 95%, 99% 신뢰수준 하에서의 **신뢰계수(confidence coefficient)**라고 한다.

예7-9. 100명의 사람들에 대하여 혈중 콜레스테롤 수준을 측정한 결과, 평균 245.69를 얻었다. 콜레스테롤 수준은 정규분포를 따르며 모표준편차가 $\sigma = 46.02$라고 할 때 모집단의 콜레스테롤 수준의 평균에 대한 95% 신뢰구간을 구하라.

(sol) 모평균 μ의 95% 신뢰구간을 구하는 공식은 다음과 같다.

$$\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$

평균, n, 모표준편차의 값을 대입하면 μ에 대한 95% 신뢰구간은 다음과 같이 계산된다.

$$\left(245.69 - 1.96 \cdot \frac{46.02}{\sqrt{100}},\ 245.69 + 1.96 \cdot \frac{46.02}{\sqrt{100}}\right)$$
$$= (245.69 - 9.02, 245.69 + 9.02)$$
$$= (236.14, 263{,}86)$$

예7-10. 한 대학교 신입생의 학력고사 성적을 추정하기 위하여 임의로 200명의 학생을 추출하여 조사한 결과, 표본평균 250점을 구하였다. 그 대학교 신입생의 학력고사 성적의 표준편차가 $\sigma = 100$점이라고 할 때 신입생 학력고사의 평균에 대한 95% 신뢰구간을 구하라.

(sol) 모평균 μ의 95% 신뢰구간은 다음과 같이 계산된다.

$$\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right)$$
$$= \left(250 - 1.96 \cdot \frac{100}{\sqrt{200}}, 250 + 1.96 \cdot \frac{100}{\sqrt{200}}\right)$$
$$= (250 - 13.86,\ 250 + 13.86)$$
$$= (236.14, 263.86)$$

예7-11. 10마리의 쥐가 미로를 통과하는 데 걸린 시간이 다음과 같다.
(32, 38, 42, 29, 30, 37, 33, 40, 37, 35) (단위 : 초)
쥐가 미로를 통과하는 데 걸리는 평균시간의 95% 신뢰구간을 구하라.

(sol) 주어진 관찰값에 의하여, 표본평균과 표본표준편차는 다음과 같이 계산된다.

$$\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 35.3 \qquad S = \sqrt{\frac{1}{9}\sum_{i=1}^{10}\left(X_i - \overline{X}\right)^2} = 4.27$$

부록의 t확률분포표에서 자유도가 9이고 신뢰구간이 95%일 때의 확률은 다음과 같다.

$$P_r(-2.26 \le T \le 2.26) = 0.95$$

μ의 95% 신뢰구간은 다음과 같이 계산된다.

$$\left(\overline{X} - 2.26 \cdot \frac{S}{\sqrt{n}}, \overline{X} + 2.26 \cdot \frac{S}{\sqrt{n}}\right) = \left(35.3 - 2.26 \cdot \frac{4.27}{\sqrt{10}}, 35.3 + 2.26 \cdot \frac{4.27}{\sqrt{10}}\right)$$
$$= (35.3 - 3.0517, 35.3 + 3.0517)$$
$$= (32.2483, 38.3517)$$

예7-12. 산성비는 공해에 기인한다고 알려져 있으며 오늘날에는 점차적으로 비의 산도가 높아지고 있다고 한다. 보통비는 pH 값이 5.7(pH 는 0이 산성이고 14가 알칼리성임)이라고 하는데, 40개의 장소에서 빗물을 수거하여 pH를 측정한 결과 표본평균은 3.7, 표본표준편차은 1.2를 얻었다. 빗물의 pH 평균에 대한 95%와 99% 신뢰구간을 구하라. 이 결과에 의할 때 우리는 무엇을 알 수 있는가?

(sol) 이 문제에서 $n = 40$으로 $n > 30$이므로 모표준편차 대신에 표본표준편차를 이용하여도 표준정규분포에 의하여 신뢰구간을 구할 수 있다.

-- μ의 95% 신뢰구간

$$\left(\overline{X} - 1.96 \cdot \frac{S}{\sqrt{n}}, \overline{X} + 1.96 \cdot \frac{S}{\sqrt{n}}\right) = \left(3.7 - 1.96 \cdot \frac{1.2}{\sqrt{40}}, 3.7 + 1.96 \cdot \frac{1.2}{\sqrt{40}}\right) = (3.33, 4.07)$$

-- μ의 99% 신뢰구간

$$\left(\overline{X} - 2.57 \cdot \frac{S}{\sqrt{n}}, \overline{X} + 2.57 \cdot \frac{S}{\sqrt{n}}\right) = \left(3.7 - 2.57 \cdot \frac{1.2}{\sqrt{40}}, 3.7 + 2.57 \cdot \frac{1.2}{\sqrt{40}}\right) = (3.21, 4.19)$$

모비율 P의 신뢰구간

- 확률변수 X가 이항분포를 따를 때 $X \sim B(n, P)$라 하면, 모비율 P의 추정량 $\hat{P}$과 $\hat{P}$의 평균과 분산은 다음과 같다.

$$\hat{P} = \frac{X}{n}, \qquad E(\hat{P}) = P, \qquad Var(\hat{P}) = \frac{P(1-P)}{n}$$

- 평균과 분산을 가지고 중심극한정리를 이용해 다음과 같이 나타낼 수 있다.

$$\frac{\hat{P} - P}{\sqrt{\frac{P(1-P)}{n}}} \sim N(0,1)$$

- 모비율 P에 대한 신뢰구간은 위에 주어진 중심극한정리에 의해 구한 표준정규분포를 이용해 구한다.
- 모비율 P의 $(1-\alpha) \times 100\%$ 신뢰구간은 다음과 같다.

$$\left(\hat{P} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + Z_{\alpha/2} \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right)$$

* 여기서 $\hat{P}$의 표준오차인 $\sqrt{\frac{P(1-P)}{n}}$에 모수 P가 있으므로 이를 P의 추정량인 $\hat{P}$으로 대치해 구한다.)

예7-13. 2006년 5월에 실시된 지방선거에서 선거일 하루 전에 1,000명을 대상으로 한 후보에 대한 지지율을 조사한 결과 540명이 그 후보를 지지한다고 말하였다. 그 후보가 과반수의 표를 얻을 수 있는가를 설명하라.

(sol) 그 후보에 대한 지지율의 추정치가 0.54이므로 그 후보에 대한 지지율 P의 95% 신뢰구간은 다음과 같다.

$$\left(\hat{P} - 1.96 \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}, \hat{P} + 1.96 \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right)$$

$$= \left(0.54 - 1.96 \cdot \sqrt{\frac{0.54 \times 0.46}{1{,}000}}, 0.54 + 1.96 \cdot \sqrt{\frac{0.54 \times 0.46}{1{,}000}}\right) = (0.51, 0.57)$$

따라서 95%의 신뢰수준 하에서 그 후보에 대한 지지율의 범위가 51%에서 57% 사이에 있다고 믿을 수 있으므로 그 후보는 과반수의 표를 얻을 수 있을 것으로 생각된다.

모분산 σ^2의 신뢰구간

- $(X_1, X_2, \dots, Xn)$을 $N(\mu, \sigma^2)$으로부터의 확률표본이라 할 때, 표본분산 S^2의 분포는 자유도가 $n-1$인 χ^2-분포로 다음과 같이 나타낼 수 있다.

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

- 일반적으로 σ^2에 대한 $(1-\alpha)\times 100\%$ 신뢰구간은 주어진 자유도(d.f)하에서 $\chi^2_{1-\alpha/2}$와 $\chi^2_{\alpha/2}$을 구한 후 다음과 같은 수식에 의해 구해진다.

$$P_r\left(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}\right) = 1-\alpha$$

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

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

- 모분산 σ^2에 대한 $(1-\alpha)\times 100\%$ 신뢰구간은 다음과 같다.

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

예7-14. 한 학생이 야구공을 10번 던졌을 때의 결과가 다음과 같다.

68.2, 57.3, 62.6, 58.9, 67.4, 55.2, 63.5, 61.2, 57.2, 64.5 (단위 : m)

이 학생이 야구공을 던지는 거리의 분산에 대한 95% 신뢰구간을 구하라.

(sol) 주어진 자료에서 표본평균과 표본분산은 각각 다음과 같이 계산된다.

$\overline{X} = 61.8$ $S^2 = 18.05$

$n = 10$이므로 $d.f. = n - 1 = 9$이고 카이제곱분포표를 이용해 분위수를 구할 수 있다.

$\chi^2_{0.975} = 2.70039$ $\chi^2_{0.025} = 19.0228$

따라서 σ^2에 대한 95% 신뢰구간은 다음과 같이 계산된다.

$$\left(\frac{(n-1)S^2}{\chi^2_{0.025}}, \frac{(n-1)S^2}{\chi^2_{0.975}}\right) = \left(\frac{9\times18.05}{19.0228}, \frac{9\times18.05}{2.70039}\right) = \left(\frac{162.45}{19.0228}, \frac{162.45}{2.70039}\right) = (8.54, 60.16)$$

끝~~ ♥♥