CHAPTER 8

대표본에서의 추론

통계적 추론(statistical inference)

- 표본 데이터를 분석하여 모집단 모수에 관한 결론을 이끌어내는 것

- 모수의 추정(estimation of parameter)
  - 모수와 가까울 것으로 생각되는 값을 추측
    - 점추정(point estimation)
    - 구간추정(interval estimation)
- 통계적 가설검정(testing of statistical hypothesis)
  - 모수의 참값에 관한 주장의 옳고 그름을 판정

예제 1

- 지역사회의 봉사에 참여하는 정도는 얼마나 될까? 학생들의 봉사 활동에 참여정도를 조사하기 위해 대학생 40명을 대상으로 지난 한 달 동안 지역사회의 봉사활동에 할애한 시간을 조사하였다.

표 1 한 달 동안 지역사회 봉사활동 시간

0	0	0	0	0	0	0	1	1	1
2	2	2	2	2	3	3	3	3	3
4	4	4	5	5	5	5	5	5	5
5	5	6	6	6	8	10	15	20	25

표본평균 $\bar{x}=4.525$,

표본표준편차 $s=5.17$,

표본중앙값= 3.50

제 1사분위수=1.5,

제 3사분위수=5.0

1. 미지의 μ를 한 값으로 추정한다(점추정 , point estimation).
2. μ에 대한 가능한 값들의 구간을 결정한다(구간추정 , interval estimation).
3. 월별 평균 봉사시간 μ가 앞선 연구에서의 평균인 2.6시간과 같은지를 결정한다(통계적 가설검정, testing statistical hypothesis).

모평균의 점추정

- 표본 데이터로부터 모수의 모르는 값에 가까울 것으로 생각되는 하나의 수를 계산하는 것
- 모집단에서 추출한 크기 n인 확률표본 $X_1, X_2, \cdots, X_n$을 이용

- 추정량(estimator)
  - 모수를 추정하기 위해 사용되는 통계량
- 추정량의 표준오차(standard error) S.E.
  - 추정량의 표준편차

모평균 μ 의 점추정

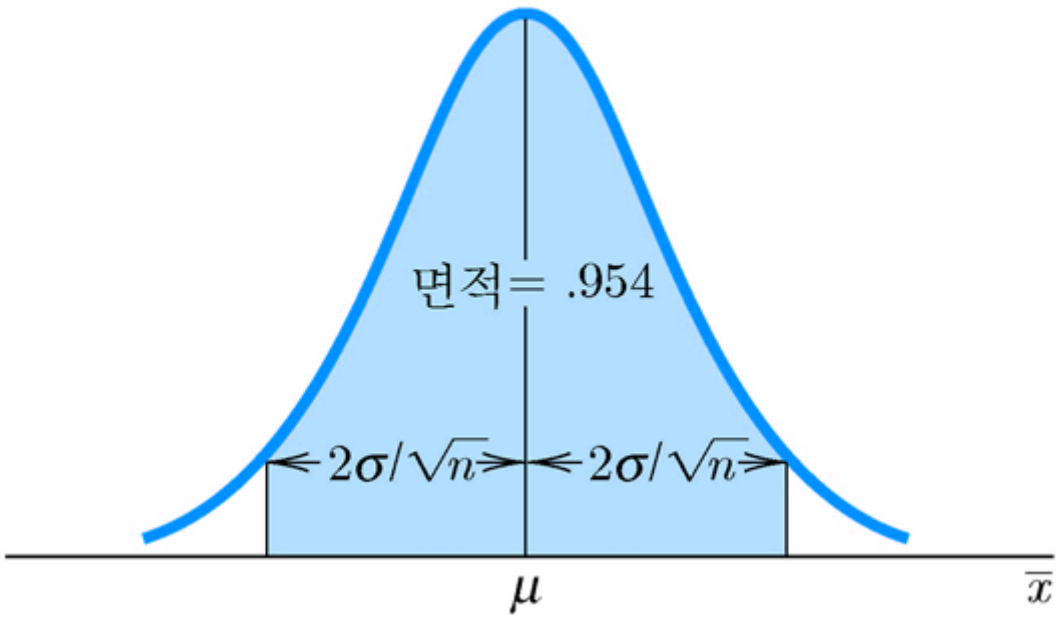
- 추정량
  - 표본평균 $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

- 표본평균 $\overline{X}$의 성질

1. $E(\overline{X}) = \mu$
2. $sd(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$, 그러므로 $\mathrm{S.E.}(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$.
3. n이 클 때, $\overline{X}$는 평균이 μ이고 표준편차가 $\sigma/\sqrt{n}$ 인 정규분포를 근사적으로 따른다.

모평균 μ 의 점추정

- 대표본 n 에 대해, 표본평균 $\overline{X}$로 모평균 μ 를 추정할 때의 95.4% 오차한계 $2\dfrac{\sigma}{\sqrt{n}}$



모평균 μ 의 점추정

- $\overline{X}$로 μ를 추정할 때의 $100(1-\alpha)\%$ 오차한계

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

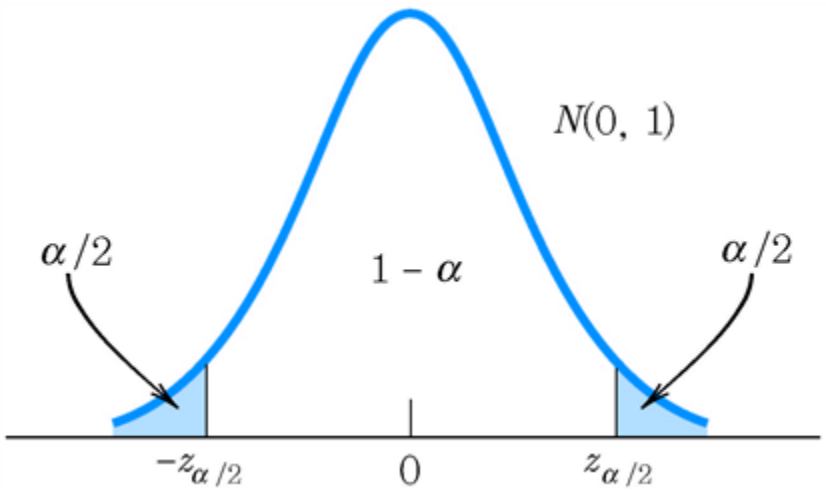


표 2 $z_{\alpha/2}$ 값

$1-\alpha$	0.80	0.85	0.90	0.95	0.99
$z_{\alpha/2}$	1.28	1.44	1.645	1.96	2.58

평균의 점추정

모수 : 모평균 μ

데이터 : $X_1, \cdots, X_n$ (크기 n인 확률표본)

추정량 : $\overline{X}$ (표본평균)

$$\text{S.E.}(\overline{X}) = \frac{\sigma}{\sqrt{n}}\text{, 추정된 S.E.}(\overline{X}) = \frac{S}{\sqrt{n}}$$

대표본 n에 대해, $100(1-\alpha)\%$ 오차한계는 $z_{\alpha/2}\sigma/\sqrt{n}$ 이다.

(만일 모표준편차 σ를 모르면, σ대신에 표본표준편차 S를 사용한다.)

예제 3

- 지난 달 지역봉사 활동시간을 나타낸 40개의 측정값으로 구성된 예제 1의 데이터로부터, 모집단 봉사활동 평균시간의 점추정값을 구하고 95% 오차한계를 정하라.

표본크기의 결정

- $\overline{X}$에 의한 μ의 추정에서 $100(1-\alpha)\%$ 오차한계

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

- 추정오차 $|\overline{X}-\mu|$ 가 지정된 값 d를 초과하지 않는다고 $100(1-\alpha)\%$ 확신하기 위해 필요한 표본의 크기

$$n=\left[\frac{z_{\alpha/2}\sigma}{d}\right]^2$$

  - 이 때, n 은 표본크기이므로 그 값이 정수가 아니면 계산된 값보다 큰 최소의 정수로 결정한다.

예제 4

- 어느 생물학자는 호수 물의 단위용적 당 평균 인산염 함유량을 추정하고자 한다. 전년도의 연구에서 표준편차는 대략 $\sigma = 4$임을 알았다. 추정오차가 0.8 이내임을 90% 확신하기 위해 얼마나 많은 표본을 분석해야 하는가?

모평균의 신뢰구간

- 구간추정(interval estimation)
  - 모수의 참값이 포함될 것으로 기대되는 구간을 추측
- 신뢰수준(level of confidence)
  - 추정된 구간들이 모수의 참값을 포함할 확률
  - 추정된 구간이 참값을 포함할 것으로 신뢰할 수 있는 수준
  - $1 - \alpha$로 표기
- $100(1 - \alpha)\%$ 신뢰구간

모평균의 신뢰구간

- 모집단이 정규분포이고 표준편차 σ를 알고 있을 때

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

이고, 다음과 같이 나타낼 수 있다.

$$P\left[\mu - 1.96\,\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1.96\,\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

$$P\left[\overline{X} - 1.96\,\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

모평균의 신뢰구간

- 모집단이 정규분포이고 표준편차 σ를 알고 있을 때,
  모평균 μ의 95% 신뢰구간(confidence interval)

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

모평균 μ의 $100(1-\alpha)\%$ 신뢰구간(confidence interval)

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

예제 5

- 어느 대규모 생산 공장에서 매일 일산화탄소(CO) 배출량에 대해 알아보기 위해 주중 25일을 무작위로 추출하여 조사하였다. 생산 과정이 항상 변하기 때문에 일일 일산화탄소 배출량의 현재 평균 값 μ는 알 수 없다. 여러 해 동안 조사된 자료에 의하면 일일 일산화탄소 배출량은 표준편차가 0.8톤인 정규분포를 따른다고 알려져 있다.
- 조사된 일산화탄소 배출량 자료의 표본평균이 $\bar{x} = 2.7$톤이라고 할 때, μ 의 95% 신뢰구간을 구하라.
- 모평균 μ는 신뢰구간에 포함되는가?

신뢰수준의 의미

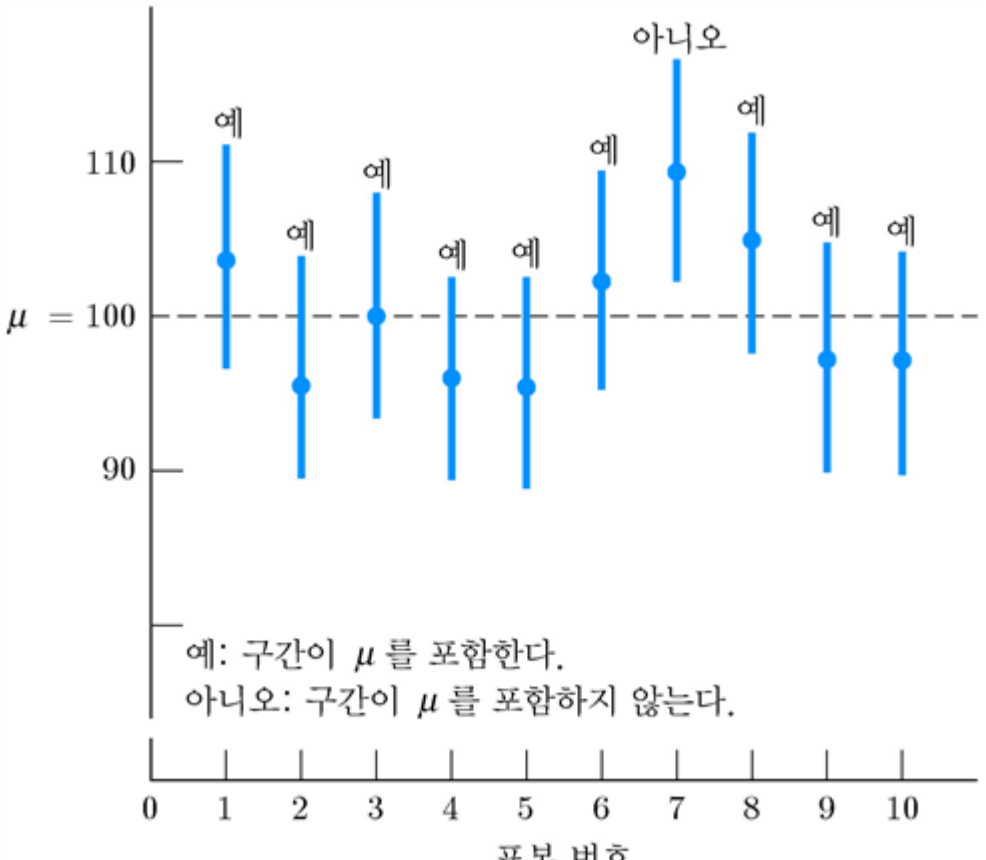
- $\mu = 100$, $\sigma = 10$인 정규모집단
- 표본크기 7인 확률표본 10개를 뽑아 μ의 95% 신뢰구간 계산

$$\overline{x} \pm 1.96\frac{10}{\sqrt{7}}$$

- 표본 1의 $\overline{x} = 104.3$이면,

95% 신뢰구간

$$104.3 \pm 7.4 = (96.9, 111.7)$$



대표본에서 μ의 신뢰구간

- 중심극한정리에 의해

$$P\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha$$

- n 이 충분히 클 때, σ 대신에 S를 사용하여 신뢰구간 계산

대표본에서 μ의 신뢰구간

n이 충분히 클 때, μ의 $100(1-\alpha)\%$ 신뢰구간은

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

이다. 여기서 S는 표본표준편차이다.

예제 6

- 지난 한 달 동안 지역사회 봉사활동에 참여한 시간 데이터

$$\mathrm{n} = 40, \qquad \bar{x} = 4.525, \qquad s = 5.17$$

매월 봉사활동 평균 시간의 (a) 90%, (b) 80% 신뢰구간을 구하라.