인공지능 수학: 통계학
6.추정

윤상민

E-mail : smyoon@kookmin.ac.kr
Office : 02-910-4645

모평균의 추정

- 표본평균의 특성
  - 모집단이 정규분포인경우
    - 표본평균 (sample mean) 사용
    - $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} = \frac{\sum_{i=1}^{n} X_1}{n}$
    - $\bar{X}$는 모평균 μ의 추정에 사용되는 통계량
  - 대표본인경우
    - 중심극한 정리에 의해 표본평균이 정규분포를 따른다고 가정함

모평균의 추정

- 점추정
  - 표본평균이 점 추정값 (추정량)이 됨
  - import numpy as np
  - samples = [9, 4, 0, 8, 1, 3, 7, 8, 4, 2]
  - print(np.mean(samples))

모평균의 추정

- 구간추정
  - 모평균 μ 의 $100(1-\alpha)$% 신뢰구간 (confidence interval)
    - (μ 의 추정량) $\pm z_{\alpha/2}$ (추정량의 표준편차)
  - 정규분포에서 σ 를 알 때,

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

    - 실용적이지 못함: 정규분포가 아니거나 표준편차가 알려져 있지 않음

모평균의 추정

- 구간추정
  - 표본의 크기가 클때 중심극한 정리 사용
    - (μ 의 추정량) $\pm z_{\alpha/2}$ (추정량의 표준편차)

$$\left(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \qquad \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

  - s: 표본표준편차

모평균의 추정

- 구간추정
  - 어떤 학교의 고1 남학생의 평균키를 추정하기 위해 36명을 표본으로 추출하여 그 표본평균과 표본표준편차를 계산하여 그 결과가 아래와 같다.
  - $\bar{x} = 173.6, s = 3.6$
  - 평균키에 대한 95% 신뢰 구간을 구하시오.

모평균의 추정

- 구간추정
  - $\alpha = 0.05$
  - $z_{\alpha/2} = z_{0.025} = 1.96$
  - $z_{\alpha/2}\frac{s}{\sqrt{n}} = 1.96 \times \frac{3.6}{\sqrt{36}} = 1.96 \times \frac{3.6}{6} = 1.176$
  - 95% 신뢰 구간
    - $(173.6 - 1.176\,, 173.6 + 1.176) = (172.4, 174.8)$

모평균의 추정

- 구간추정
  - 어떤 농장에서 생산된 계란 30개의 표본을 뽑았더니 그 무게가 아래와 같다.
    - w=[10.7, 11.7, 9.8, 11.4, 10.8, 9.9, 10.1, 8.8, 12.2, 11.0, 11.3, 11.1, 10.3, 10.0, 9.9, 11.1, 11.7, 11.5, 9.1, 10.3, 8.6, 12.1, 10.0, 13.0, 9.2, 9.8, 9.3, 9.4, 9.6, 9.2]
  - 계란의 평균 무게에 대한 95% 신뢰 구간을 구하시오.

모평균의 추정

- 구간추정
  - import numpy as np
  - w=[10.7, 11.7, 9.8, 11.4, 10.8, 9.9, 10.1, 8.8, 12.2, 11.0, 11.3, 11.1, 10.3, 10.0, 9.9, 11.1, 11.7, 11.5, 9.1, 10.3, 8.6, 12.1, 10.0, 13.0, 9.2, 9.8, 9.3, 9.4, 9.6, 9.2]
  - xbar=np.mean(w)
  - sd=np.std(w, ddof=1)
  - print("평균 %.2f, 표준편차: %.2f" %(xbar, sd))
  - import scipy.stats
  - alpha=0.05
  - zalpha = scipy.stats.norm.ppf(1-alpha/2)
  - print("zalpha: ", zalpha)

모평균의 추정

- 구간추정
  - $\alpha = 0.05$
  - $z_{\alpha/2} = z_{0.025} = 1.96$
  - $\bar{X} = 10.43$
  - $s = 1.11$
  - $z_{\alpha/2}\frac{s}{\sqrt{n}} = 1.96 \times \frac{1.11}{\sqrt{30}} = 0.397$
  - 95% 신뢰 구간
    - $(10.43 - 0.397\,, 10.43 + 0.397) = (10.033, 10.827)$

모비율의 추정

- 점 추정
  - 확률변수 X:
    - n개의 표본에서 특정 속성을 갖는 표본의 개수
  - 모비율 p 의 점추정량
    - $\hat{p} = \frac{X}{n}$

모비율의 추정

- 점 추정
  - 대학교 1학년생의 흡연율을 조사하기 위해 150명을 랜덤하게 선택하여 흡연여부를 조사하였다. 이 중 48명이 흡연을 하고 있었다. 이 대학교 1학년생의 흡연율의 평균을 점추정하시오.
  - $n = 150, X = 48$
  - $\hat{p} = \frac{X}{n} = \frac{48}{150} = 0.32$
  - 평균흡연율을 32%로 추정됨.

모비율의 추정

- 구간추정
  - n 이 충분히 클 때,
    - $n\hat{p} > 5, n(1 - \hat{p}) > 5$일 때를 의미
    - $X \sim N(np, np(1 - p))$
  - 확률변수 X 의 표준화
    - $Z = \frac{X - np}{\sqrt{n\hat{p}(1 - \hat{p})}} = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}}$
    - 근사적으로 표준정규분포 $N(0,1)$를 따름

모비율의 추정

- 구간추정
  - $P(|Z| \le z_{\alpha/2}) = 1 - \alpha$
    - $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = P\left(-z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le z_{\alpha/2}\right)$
    - $= P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$
  - 모비율 p 의 $100(1-\alpha)\%$ 신뢰구간 (confidence interval)
    - $\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

모비율의 추정

- 구간추정
  - 대학교 1학년생의 흡연율을 조사하기 위해 150명을 랜덤하게 선택하여 흡연여부를 조사하였다. 이 중 48명이 흡연을 하고 있었다.
  - 흡연율 p 의 95% 신뢰구간 (confidence interval) 을 구하시오.

모비율의 추정

- 구간추정
  - $\alpha = 0.05, z_{\frac{\alpha}{2}} = z_{0.025} = 1.96, \hat{p} = 0.32$
  - $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.32)(0.68)}{150}} = 0.038$
  - $(0.32 - 1.96 \times 0.038, 0.32 + 1.96 \times 0.038) = (0.245, 0.395)$

모비율의 추정

- 구간추정
  - import numpy as np
  - x=48
  - n=150
  - phat = x / n
  - alpha=0.05
  - zalpha = scipy.stats.norm.ppf(1-alpha/2)
  - sd=np.sqrt(phat*(1-phat)/n)
  - print("phat %.3f, zalpha: %.3f, sd: %.3f"%(phat, zalpha, sd))
  - ci = [phat -zalpha * sd, phat + zalpha * sd]
  - print(ci)