통계학

– 제 7 장 통계적 추론 (1)

0011 0010 1010 1101 0001 0100 1011

통계적 추론

- 통계적 추론 (statistical inference)
  - 통계적 추론의 정의
    - 표본이 갖고 있는 정보를 분석하여
      모수에 관한 결론을 유도하고,
      모수에 대한 가설의 옳고 그름을 판단하는 것
  - 통계적 추론의 분류
    - 추정 (estimation)
      - 미지수인 모수에 대한 추측값을 그 수치화된 정확도와 함께 제시
      - 점추정 (point estimation) :
        미지인 모수의 참값으로 생각되는 단 하나의 값을 지정
      - 구간추정 (interval estimation) :
        미지인 모수의 참값이 포함되리라 예상되는 구간을 제시
    - 가설검정 (hypothesis testing)
      - 표본 자료의 정보를 통해 미지인 모수의 참값에 대한
        특정 가설이 적합한 지 적합하지 않은 지를 판별

추정 (estimation)

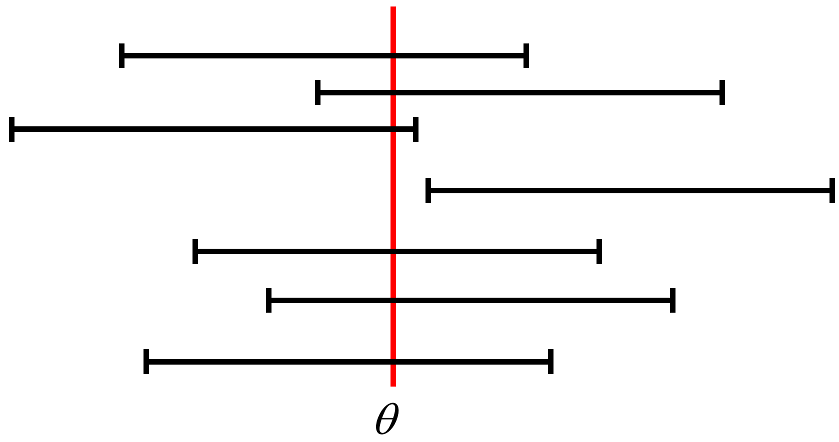
- 점추정 (point estimation)
  - 점추정의 의미
    - 추출된 표본으로부터
      미지인 모수의 참값에 가까우리라고 예상되는 하나의 값을 제시
  - 점추정과 관련된 용어들
    - 추정량 (estimator)
      - 모수를 추정하기 위해 만들어진 통계량
      - 일정한 확률분포를 갖는 확률변수
    - 추정치 (estimate)
      - 주어진 관측값으로부터 계산된 추정량의 실제 값
    - 표준오차 (standard error, $S.E.$)
      - 추정량의 표준편차
      - 추정량의 정확도를 측정하는 하나의 도구

- 구간추정 (interval estimation)
  - 구간추정의 의미
    - 추정량의 분포를 이용하여 표본으로부터
    - 미지인 모수의 참값을 포함하리라고 예상되는 구간을 제시
  - 신뢰구간 (confidence interval)
    - 미지인 모수의 참값이 포함되리라고 예상되는 구간
    - $100\times(1-\alpha)$% 신뢰구간 :
      미지의 모수 θ에 대해 다음이 성립하는 구간 (L, U)

$$\boxed{P[L \le \theta \le U] = 1-\alpha}$$

    - L : 신뢰하한
    - U : 신뢰상한
    - $1-\alpha$: 신뢰수준 (confidence level)
      - 일반적으로 90%, 95%, 99% 등을 사용

- 신뢰수준의 의미
  - 동일한 모집단으로부터 확률표본의 추출과정을 반복하여 신뢰구간을 무한히 많이 얻었을 때 그 중 $100(1-\alpha)\%$는 미지인 모수의 참값을 포함



- 오차한계 (limit of error)
  - 점추정량을 중심으로 한 구간추정의 범위
  - 오차범위(error margin)라고도 부름

- 구간추정의 절차
  - Step 1 :
    미지의 모수 θ에 대한 적절한 점추정량을 선택
  - Step 2 :
    선택된 점추정량과 추정하고자 하는 모수를 엮어
    잘 알려진 특정 확률분포를 따르는 확률변수의 함수를 만듦
  - Step 3 :
    만들어진 확률변수의 함수와 잘 알려진 확률분포를 이용하여
    확률이 $1-\alpha$가 되는 부등식을 구성
  - Step 4 :
    추정하고자 하는 모수 θ를 중심으로 부등식을 풀어
    신뢰구간 (L, U)을 찾음

모평균에 대한 추론

- 모평균에 대한 구간추정(Ⅰ)
  - 표본의 크기가 큰 경우
    - Step 1 :
      미지의 모수 μ에 대한 적절한 점추정량으로 $\bar{X}$ 을 선택
    - Step 2 :
      선택된 점추정량 $\bar{X}$ 과 추정하고자 하는 모수 μ를 엮어
      잘 알려진 특정 확률분포를 따르는 확률변수의 함수를 만듦
      - 확률변수의 함수 : $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
      - $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 의 확률분포 : $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ by C.L.T.

- Step 3 :
  만들어진 확률변수의 함수와 그것의 확률분포를 이용하여
  확률이 $1-\alpha$ 가 되는 부등식을 구성

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

- Step 4 :
  추정하고자 하는 모수 μ 를 중심으로 부등식을 풀어냄

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

  - 모분산(σ^2)을 모르는 경우 표본분산(S^2)으로 대체 가능

- 예 – [수확한 사과의 평균무게]
  - 어느 사과농장에서 올해 수확한 사과의 평균무게를 추정
  - 수확한 사과 중에서 임의로 36개 추출
  - 36개 사과의 평균무게 : $\bar{x} = 342$ (g), 표준편차 : $s = 20$ (g)
  - 올해 수확한 사과의 평균무게에 대한 95% 신뢰구간은?

    - Step 1 :

      모수 : 올해 수확한 사과의 평균 무게 (μ)

      점추정량 : 36개 사과의 평균 무게 ($\bar{X}$)

    - Step 2 :

      $$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \text{ by C.L.T.}$$

- 예 – [수확한 사과의 평균무게]
  - Step 3 :

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

  - Step 4 :

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

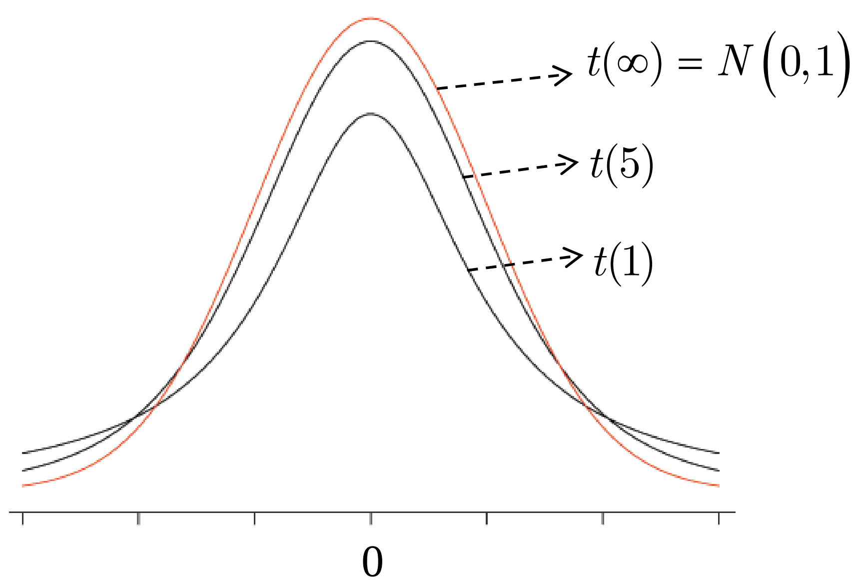
$$\Rightarrow \quad P\left[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right] = 1-\alpha$$

$$\Rightarrow \quad P\left[342 - 1.96\frac{20}{\sqrt{36}} < \mu < 342 + 1.96\frac{20}{\sqrt{36}}\right] = 0.95$$

  - 95% 신뢰구간 : $(335.47,\ \ 348.53)$ 혹은 342 ± 6.53

- 정규모집단에서의 추론 시 주의점
  - 표본의 크기가 큰 경우
    - 모분산(σ^2)을 모르는 경우 표본분산(S^2)으로 대체 가능
    - $$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \Rightarrow \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim N(0,1)$$
  - 표본의 크기가 작은 경우
    - 모분산(σ^2)을 모르는 경우 표본분산(S^2)으로 대체하면 확률분포가 정규분포를 따르지 않음
    - 표준정규분포와 비슷한 모양의 t-분포를 따름

- t-분포 (Student's t distribution)
  - 표준정규분포에 비해 두터운 꼬리를 가짐
  - 자유도(df)가 커져서 30이상이 되면 표준정규분포와 거의 유사
  - 자유도 (degree of freedom, df) = $n-1$
  - $$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(df)$$
  - 교재 p.329 부록 <표3> [t-분포표]참조



- 모평균에 대한 구간추정(II)
  - 표본의 크기가 작은 경우
    - Step 1 :

      미지의 모수 μ에 대한 적절한 점추정량으로 $\bar{X}$ 을 선택

    - Step 2 :

      선택된 점추정량 $\bar{X}$ 과 추정하고자 하는 모수 μ를 엮어 잘 알려진 특정 확률분포를 따르는 확률변수의 함수를 만듦

      - 확률변수의 함수 : $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \Rightarrow \dfrac{\bar{X}-\mu}{S/\sqrt{n}}$
      - $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ 의 확률분포 : $\dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(df)$

- Step 3 :
  만들어진 확률변수의 함수와 그것의 확률분포를 이용하여
  확률이 $1-\alpha$ 가 되는 부등식을 구성

$$P\left[-t_{\alpha/2}(df) < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2}(df)\right] = 1-\alpha$$

- Step 4 :
  추정하고자 하는 모수 μ 를 중심으로 부등식을 풀어냄

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

- 예 – [소비자 만족도 지수]
  - S회사가 개발한 신제품을 사용하는 25명의 소비자 대상
  - 소비자 만족도 조사 결과

75	63	49	86	53
80	70	72	81	80
69	76	85	95	66
77	77	63	58	74
68	90	82	59	60

  - S회사가 개발한 신제품의 소비자 만족도에 대한 95% 신뢰구간?

- 예 – [소비자 만족도 지수]
  - 정규분포 가정 하에서의 95% 신뢰구간
    - 표본평균 : $\bar{x} = 72.32$
    - 표본표준편차 : $s^2 = 11.60$
    - $z_{\alpha/2} = z_{0.025} = 1.96$
    - 95% 신뢰구간 :

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 72.32 \pm 1.96 \times \frac{11.60}{\sqrt{25}} = 72.32 \pm 4.5472$$

  - t–분포 하에서의 95% 신뢰구간
    - 자유도 : $df = n - 1 = 24$
    - $t_{\alpha/2}(24) = t_{0.025}(24) = 2.064$
    - 95% 신뢰구간 :

$$\bar{X} \pm t_{\alpha/2}(24)\frac{S}{\sqrt{n}} = 72.32 \pm 2.064 \times \frac{11.60}{\sqrt{25}} = 72.32 \pm 4.78848$$

모비율에 대한 추론

- 모비율에 대한 추정
  - 점추정
    - 모비율 p 에 대한 점추정 :
      임의의 n 명 중에서 성공의 개체 수를 X 라 하면

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

    - 모비율 p 의 점추정량에 대한 기대값과 표준오차
      - 확률변수 X 의 확률분포는 이항분포 $\mathrm{Bin}(n, p)$ 이므로

$$E[\hat{p}] = \frac{1}{n}E[X] = \frac{1}{n}np = p$$

$$S.E.[\hat{p}] = S.E.\left[\frac{X}{n}\right] = \sqrt{\frac{p(1-p)}{n}}$$

- 구간추정(표본의 크기가 큰 경우)
  - Step 1 :
    미지의 모수 p에 대한 적절한 점추정량으로 $\hat{p}$ 을 선택
  - Step 2 :
    선택된 점추정량 $\hat{p}$ 과 추정하고자 하는 모수 p를 엮어
    잘 알려진 특정 확률분포를 따르는 확률변수의 함수를 만듦
    - 확률변수의 함수 : $\dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} \Rightarrow \dfrac{\hat{p}-p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}$
    - $\dfrac{\hat{p}-p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}}$ 의 확률분포 : $\dfrac{\hat{p}-p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$ by C.L.T.

- Step 3 :
  만들어진 확률변수의 함수와 그것의 확률분포를 이용하여
  확률이 $1-\alpha$ 가 되는 부등식을 구성

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

- Step 4 :
  추정하고자 하는 모수 p 를 중심으로 부등식을 풀어냄

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

- 예 – [도시의 실업률]
  - 한 도시의 노동인구 중 500명을 임의추출 하였더니 41명이 실업자
  - 이 도시의 실업률에 대한 점추정치와 그에 따른 표준오차는?
    - 실업자 수 : X
    - 실업률에 대한 점추정치 :

$$\hat{p} = \frac{X}{n} = \frac{41}{500} = 0.082$$

    - 점추정치의 표준오차 :

$$S.E.[\hat{p}] = \sqrt{\frac{p(1-p)}{n}} \doteq \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.082 \times (1-0.082)}{500}} = 0.0123$$

    - 95% 신뢰구간 :

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.082 \pm 1.960 \times 0.0123 = 0.082 \pm 0.0240$$