6. 통계적 추론 - 추정

통계적 추론 개요

- **통계적 추론**

  추정과 가설검정의 방식을 이용하여 표본으로부터의 정보를 이용하여 모집단에 대한 추측이나 결론을 이끌어 내는 과정

  - **추정** (estimation) : 모수의 값을 표본의 정보를 이용하여 예측
    - **점 추정** (point estimation) : 하나의 모수를 한 개의 값으로 추정
    - **구간 추정** (interval estimation) : 모수가 포함되리라 기대되는 구간으로 모수를 추정

  - **검정** (hypothesis testing) : 모수의 값에 대한 주장 또는 단순한 추측 등의 옳고 그름에 대한 결정을 하는 과정

통계적 추론 개요

- **추정량과 표준오차**
  - 추정량(Estimator)
    - 모집단의 모수 θ를 추정하기 위한 통계량 $\hat{\theta}$을 추정량이라고 함.
    - cf) 추정치 (Estimates) : 추정량 $\hat{\theta}$에 대한 관찰값
  - 표준오차 (Standard Error)
    - 추정량 $\hat{\theta}$의 표준편차

$$SE[\hat{\theta}] = \sqrt{V[\hat{\theta}]} = \sqrt{E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right]}$$

- 추정에서 발생하는 오차가 평균적으로 얼마인가를 나타냄.
- 예:

$$SE[\bar{X}] = \frac{\sigma}{\sqrt{n}}, \quad SE[\hat{p}] = \sqrt{\frac{p(1-p)}{n}}$$

$$\widehat{SE[\bar{X}]} = \frac{S}{\sqrt{n}}, \quad \widehat{SE[\hat{p}]} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

통계적 추론 개요

- 예제
  - 미국 주식시장에서 소형주 주식 중 49개의 표본을 무작위로 추출하여 주식가격을 조사한 결과 표본 평균은 56이고 표본 표준편차는 14로 계산되었다. 소형주 주식 전체의 평균 가격에 대한 추정치와 그 표준오차를 구하여라.

통계적 추론 개요

- 예제
  - 어느 포트폴리오의 월별 수익률이 (+)인지 아닌지 여부는 독립이라고 하자. 지난 4년간 이 포트폴리오의 월별 수익률을 조사한 결과 (+)인 경우는 30번에 해당했다. 이 포트폴리오를 운영한 전 기간 중 월별 수익률이 (+)인 경우의 비율에 대한 추정치와 그 표준오차를 구하여라.

바람직한 추정량이 갖추어야 할 성질

- **바람직한 추정량이 갖추어야 할 성질**
  - 불편성 (Unbiasedness)

$$E[\hat{\theta}] = \theta$$

을 만족하는 추정량 $\hat{\theta}$를 θ에 대한 불편추정량이라고 함.

- 불편 추정량의 예: $\overline{X}$, S^2, $\hat{p}$

바람직한 추정량이 갖추어야 할 성질

- 상대적 효율성 (Relative Efficiency)

$$E\left[\hat{\theta}_1\right] = E\left[\hat{\theta}_2\right] = \theta \quad \text{and} \quad SE[\hat{\theta}_1] < SE[\hat{\theta}_2]$$

인 경우, $\hat{\theta}_1$이 $\hat{\theta}_2$보다 상대적으로 효율적인 추정량이라고 함.

※ MVUB (최소 분산 불편 추정량, Minimum Variance Unbiased Estimator)
예 : 정규 모집단에서의 표본평균 $\overline{X}$는 모평균 μ에 대한 MVUE임.

바람직한 추정량이 갖추어야 할 성질

- 일치성 (Consistency)

임의의 양수 $\varepsilon > 0$에 대하여

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0 \ \ (확률적\ 수렴, convergence\ in\ probability : \hat{\theta} \xrightarrow{p} \theta\ )$$

를 만족하는 추정량 $\hat{\theta}$을 θ에 대한 일치추정량이라고 함.

- 일치 추정량의 예: $\overline{X},\ S^2,\ S,\ \hat{p}$

점 추정

- **주요 점추정량**
  - 모평균 μ의 추정량
    - $\hat{\mu} = \bar{X}$
  - 모분산 σ^2의 추정량
    - $\widehat{\sigma^2} = S^2$
  - 모비율 p의 추정량
    - $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$

구간 추정

- **신뢰구간 개념 및 도출 방법**
  - **신뢰구간**

모수 θ에 대한 통계량 $\hat{\theta}_L$ 과 $\hat{\theta}_U$ 이 있어 다음을 만족하는 경우,

$$P\left[\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\right] = 1 - \alpha, \qquad 0 \leq \alpha \leq 1,$$

구간 $(\hat{\theta}_L, \hat{\theta}_U)$을 θ의 $100(1-\alpha)\%$ 신뢰구간(confidence interval)이라고 함.

- $\hat{\theta}_L$: 신뢰구간의 하한
- $\hat{\theta}_U$: 신뢰구간의 상한
- $1 - \alpha$: 신뢰수준(confidence level)

구간 추정

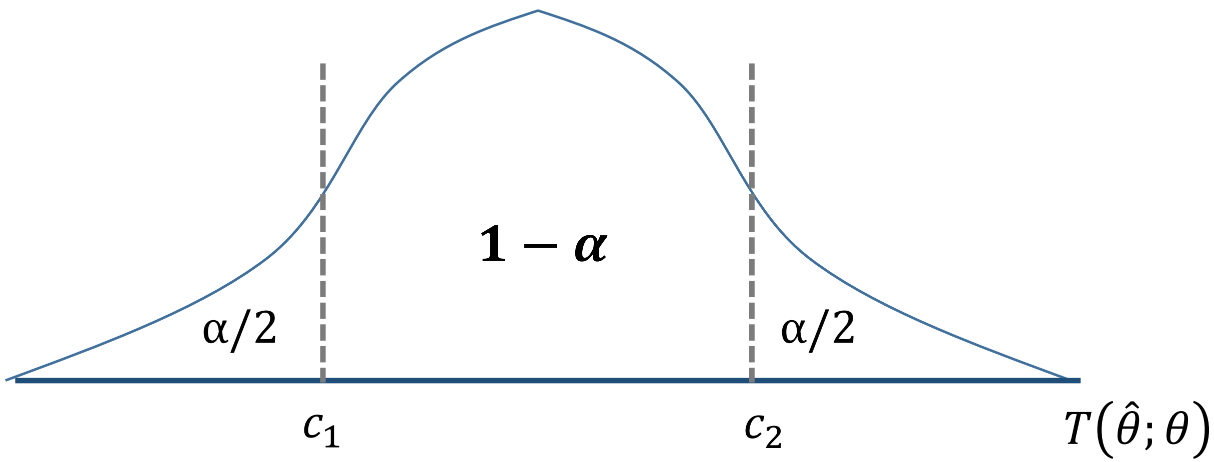
- **신뢰구간 추정량의 도출**

  신뢰구간은 일반적으로 모수 θ의 점추정량 $\hat{\theta}$을 이용한 다음의 과정을 통해 도출함.

  - 그 분포가 모수 θ에 의존하지 않는, 점추정량 $\hat{\theta}$과 모수 θ의 함수인 $T(\hat{\theta};\theta)$을 정의한 뒤,
  - 고정된 $0 < \alpha < 1$에 대하여 $P\left(c_1 < T(\hat{\theta};\theta) < c_2\right) = 1 - \alpha$를 만족하는 c_1, c_2를 구한다.
  - 부등식 $c_1 < T(\hat{\theta};\theta) < c_2$를 $\hat{\theta}_L \le \theta \le \hat{\theta}_U$의 꼴로 바꿔쓸 수 있다면,

$$P\left(\hat{\theta}_L \le \theta \le \hat{\theta}_U\right) = 1 - \alpha$$

가 되므로 $[\hat{\theta}_L, \hat{\theta}_U]$는 θ에 대한 $100(1-\alpha)\%$ 신뢰구간이 된다.

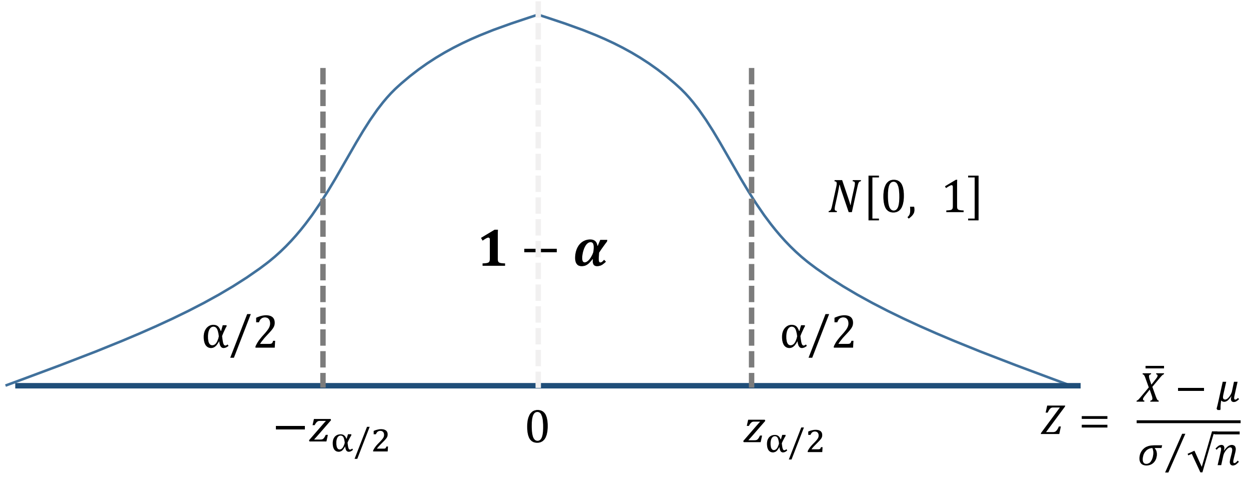


구간 추정

- **모평균에 관한 신뢰구간 도출**
  - $X_1,\ldots,X_n$이 모분산 σ^2이 알려진 정규 모집단 $N[\mu,\ \sigma^2]$ 으로부터의 확률표본이라고 할 때, 모평균 μ에 관한 신뢰수준 $100(1-\alpha)\%$의 신뢰구간을 도출해보자.
  - μ 의 점 추정량은 $\bar{X}$ 이고, 그 분포는 아래와 같다.

$$\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right) \quad \leftrightarrow \quad Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

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



구간 추정

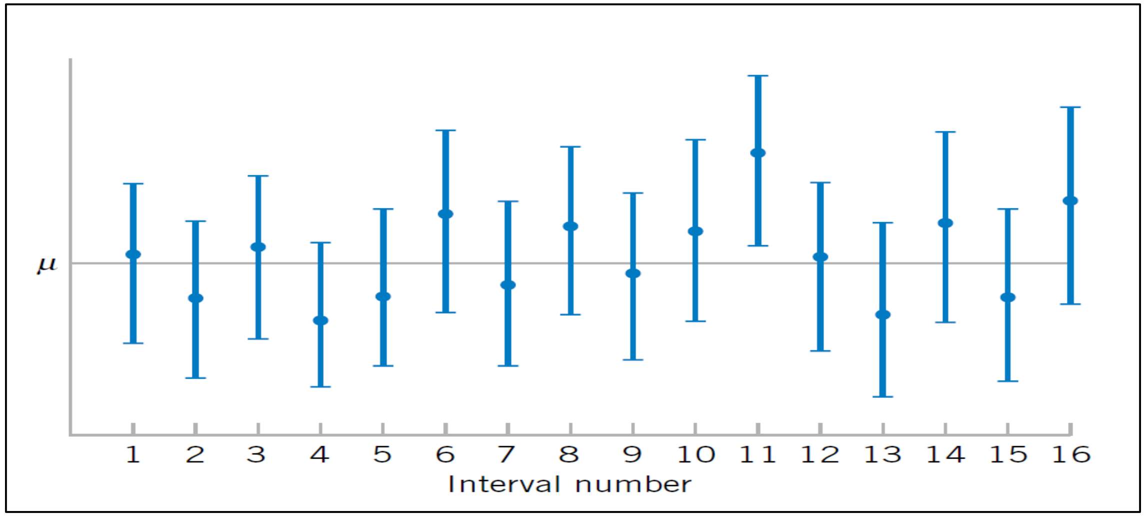
- **신뢰수준과 오차한계**
  - **오차한계** (margin of error) : 신뢰구간 길이의 절반에 해당함.

  - **좋은 신뢰구간의 조건**
    - 신뢰수준은 높을수록 좋음
    - 구간의 길이(또는 오차한계)는 짧을수록 좋음.

  - **신뢰수준과 구간의 길이는 서로 상충관계**
    - 신뢰수준을 높이면 구간의 길이는 길어지며, 구간의 길이를 줄이면 신뢰수준이 떨어짐.
    - 이를 감안하여, 주어진 신뢰수준에 대해 가장 짧은 구간의 길이를 가지도록 만드는 것을 가장 이상적인 방법으로 여김.

구간 추정

- **신뢰수준의 해석**
  - 신뢰수준은 동일한 구간 추정법을 반복적으로 사용할 때 얻어지는 신뢰구간들이 참값 θ를 품을 확률을 의미함.



  - μ에 관한 신뢰수준 90%의 신뢰구간이 $\left(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}},\ \ \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\right)$라고 할 때, 신뢰수준 90%의 의미는
    - 어떤 n개의 표본이 주어졌을 때, 이 표본을 위 식에 대입하여 계산된 신뢰구간은 참값(모평균)을 포함할 수도 포함하지 않을 수도 있다.
    - 만약 n개의 표본을 100번 반복적으로 뽑은 뒤, 같은 방식으로 100개의 신뢰구간을 구성한다면, 그 중 90개 정도는 참값(모평균)을 포함하고 있을 것으로 기대된다.

모평균에 관한 신뢰구간

- **모평균 μ 에 관한 신뢰수준 $100(1-\alpha)\%$의 신뢰구간**
  - 정규모집단인 경우
    - σ^2 알려진 경우

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

    - σ^2 알려지지 않은 경우

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

모평균에 관한 신뢰구간

- 정규가 아닌 모집단. 표본의 수 n이 $n \geq 30$ 으로 충분히 큰 경우.
  - σ^2 알려진 경우

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

  - σ^2 알려지지 않은 경우

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

모평균에 관한 신뢰구간

- 예제
  - 분산이 알려진 정규모집단에서 모평균에 관한 신뢰구간을 구하는 경우, 표본의 수가 증가하면 오차한계는
    - 증가함.
    - 감소함.
    - 변하지 않음.
    - 표본평균에 따라 증가 혹은 감소할 수 있음.

모평균에 관한 신뢰구간

- 예제
  - 모집단에서 추출한 n개의 표본을 이용하여 모평균에 관한 신뢰구간 추정치를 구하고자 한다. 신뢰수준만 95% 대신 99%로 바꾸고, 다른 모든 조건은 변경하지 않는다고 할 때, 다음 중 예상되는 변화는 무엇인가?
    - 신뢰구간의 길이가 증가
    - 신뢰구간의 길이가 감소
    - 신뢰구간의 길이가 변하지 않음

모평균에 관한 신뢰구간

- 예제
  - 모분산이 900인 정규 모집단에서 225개의 표본을 무작위로 선택하였는데, 표본 평균이 100이었다고 하자. 이 정보로 모평균에 대한 95% 신뢰수준의 신뢰구간 추정치를 구한다면, 오차한계는 얼마인가?

모평균에 관한 신뢰구간

- 예제
  - 기업의 신뢰도는 모표준편차가 12인 정규분포를 따른다고 하자. 어느 금융기관의 여신담당자로부터 임의로 추출된 9개 기업의 신뢰도는 평균이 60이었다. 전체 기업에 대한 신뢰도 모평균에 관한 95%의 신뢰구간을 구하여라.

모평균에 관한 신뢰구간

- 예제
  - 코네티컷의 한 부동산 중개인은 이 주의 주택가격의 평균을 알고자 한다. 이를 위해 20개의 무작위 표본 주택을 추출하였더니, 표본의 평균 주택가격은 $175,622이고, 표본 표준편차가 $37,221이었다고 하자. 코네티컷 주택가격은 정규분포를 따른다고 할 때, 이 주택가격의 모평균에 대한 95% 신뢰구간을 구하여라.

모비율에 관한 신뢰구간

- **모비율 p에 관한 신뢰수준 $100(1-\alpha)\%$ 의 신뢰구간**

  - 표본의 수 n이 $np \geq 5, n(1-p) \geq 5$를 만족할 만큼 충분히 큰 경우

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \qquad \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

모비율에 관한 신뢰구간

- 예제
  - 어느 은행에서 판매하는 적금 상품에 관심있는 고객의 비율을 알고자 한다. 이 은행 고객 중 200명을 랜덤하게 추출한 뒤, 적금 상품에 관심이 있는지 여부를 확인한 결과 28명이 관심이 있다고 답을 하였다. 이 적금상품에 관심있는 고객의 모비율을 95%의 신뢰구간으로 추정하여라.

모비율에 관한 신뢰구간

- 예제
  - 어느 택배 회사는 주문 후 72 시간 이내에 제품이 배송될 것을 고객에게 약속하였다. 이 회사의 품질 관리 부서에서는 이 약속이 이행되었는지 수시로 확인하고 있는데, 최근 품질 관리 부서에서 50개의 주문을 무작위로 추출한 결과, 주문 후 72 시간 이내에 35 개가 배송된 것으로 나타났다.
    - 주문 후 72 시간 이내에 배송되는 주문의 비율에 대한 98% 신뢰구간을 도출하여라.
    - 위의 질문에서 얻은 신뢰구간이 다소 넓어서, 더 좁은 너비의 구간 추정치를 구하고자 한다면 어떻게 해야 하는가?

모분산에 관한 신뢰구간

- **모분산 σ^2에 관한 신뢰수준 $100(1-\alpha)\%$ 의 신뢰구간**
  - 정규 모집단인 경우

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

모분산에 관한 신뢰구간

- 예제
  - 어느 은행 고객 중 21명을 무작위로 추출하여 예금액을 조사한 결과 일일 잔액의 평균은 \$430이고 표준편차는 \$50이었다고 한다. 이 은행 고객 별 예금액의 분포는 정규분포를 따른다고 할 때, 예금액의 모분산에 관한 95% 신뢰구간을 구하여라.

모분산에 관한 신뢰구간

- 예제
  - 어느 금융기관에서 운영하고 있는 포트폴리오의 로그수익률은 정규분포를 따른다. 이 금융기관에서 운영하는 포트폴리오 중 10개를 무작위로 추출한 뒤 로그수익률을 구해보았더니 표준편차가 4%였다. 전체 포트폴리오의 로그수익률의 모표준편차게 관한 90%의 신뢰구간을 구하여라.