금융인을 위한

통계와 데이터 분석 입문

평균을 중심으로 추정 이해하기

한국금융연수원
KOREA BANKING INSTITUTE

학습 내용

1. 점추정과 구간추정 비교
2. 모평균의 신뢰구간
3. 파이썬을 활용한 모평균의 추정

점추정

추정

표본을 통해 모집단의 특성을 추측하는 과정

- 추정에는 점추정(point estimation)과 구간추정(interval estimation)이 있음

점추정

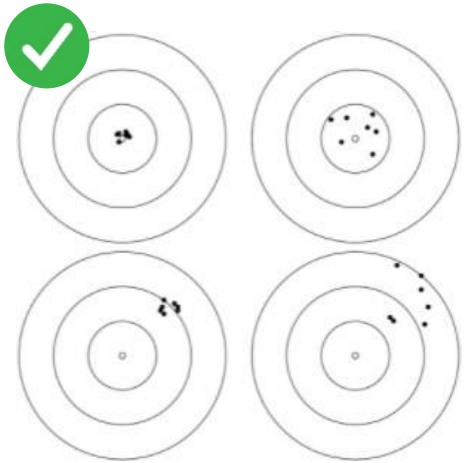
모집단의 모수(parameter)를 추정할 때
표본으로부터 얻을 수 있는 통계량(statistic) 중에서
적절한 것을 선택하여 값을 구하는 것

점추정

- 추정량(estimator)
  : 추정의 목적으로 이용되는 통계량, 수식의 형태
- 추정치(estimate)
  : 표본에서 구한 추정량의 구체적인 값
- 바람직한 추정량을 선택하는 대표적인 기준
  - 불편추정량(unbiased estimator)
    : 추정량의 기대값이 모수

    예 $$E(\bar{X}) = E(\sum_{1}^{n} X_i/n) = \mu$$

  - 효율성(efficiency): 최소분산
  - 불편성(unbiasedness)를 만족시키는 추정량 중에서 효율성을 만족시키는 추정량을 선택함



- 최소분산불편추정량(minimum variance unbiased estimator) : 불편성, 효율성을 모두 만족시키는 추정량

예

표본평균 $X = \sum_1^n X_i/n$

예

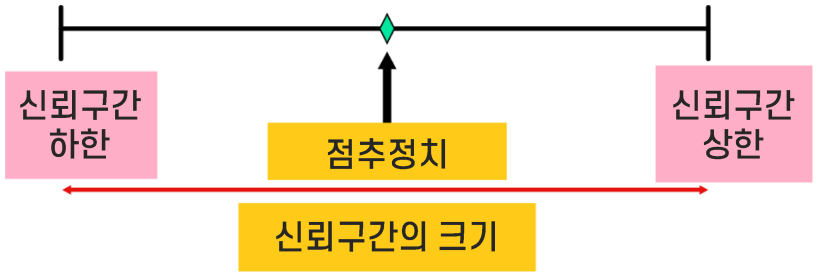
표본분산 $S^2 = \sum_1^n (X_i - \bar{X})^2/(n-1)$

- 일관성(consistency) 등 다른 기준들이 있음

구간추정

구간추정

모수를 추정할 때 하나의 '값'으로 추정하지 않고,
모수의 참값이 포함되어 있으리라 생각되는
'구간'으로 추정하는 것



- 알고 싶은 모수 θ 에 대한 $(1-\alpha)100\%$ 신뢰구간은 (L, U)
  - $(1-\alpha)100\%$은 **신뢰수준**이라 함
  - 신뢰구간의 하한 L, 상한 U는 $P(L<\theta<U)=1-\alpha$를 만족시키는 값

구간추정

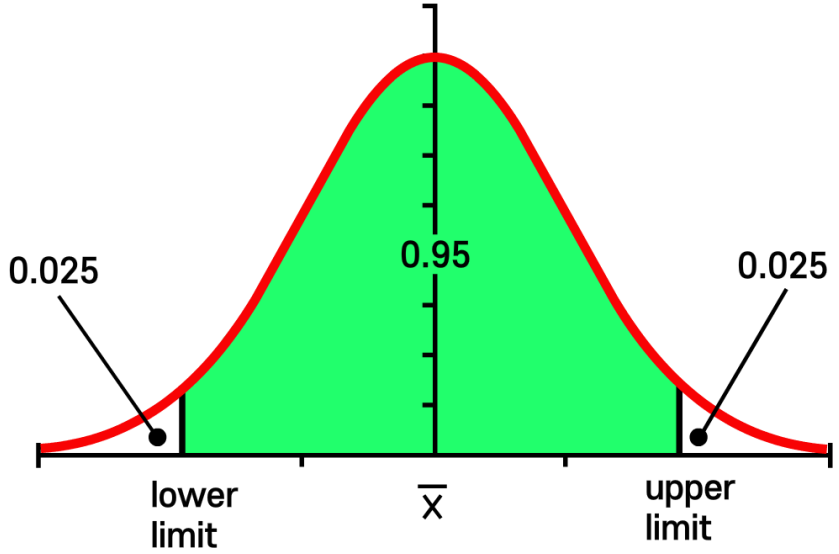
신뢰수준

참값을 구하기 위한 작업을 많이 반복했을 때,
참값이 구해진 구간에 포함되어 있을 비율
→ **정확도**의 개념

- 알지 못하는 모수에 대해 신뢰한다는 것은 불가능하므로 100% 신뢰수준은 의미 없음
- 신뢰수준이 높을수록 구간의 크기가 커짐
- 모수에 대한 제대로 된 정보를 얻기 위해서는 적당한 신뢰수준을 정할 필요가 있음

구간추정

- 일반적으로 95% 신뢰구간($\alpha = 0.05$)을 사용하고, 경우에 따라 90%, 99% 신뢰구간도 사용함



구간추정

신뢰구간

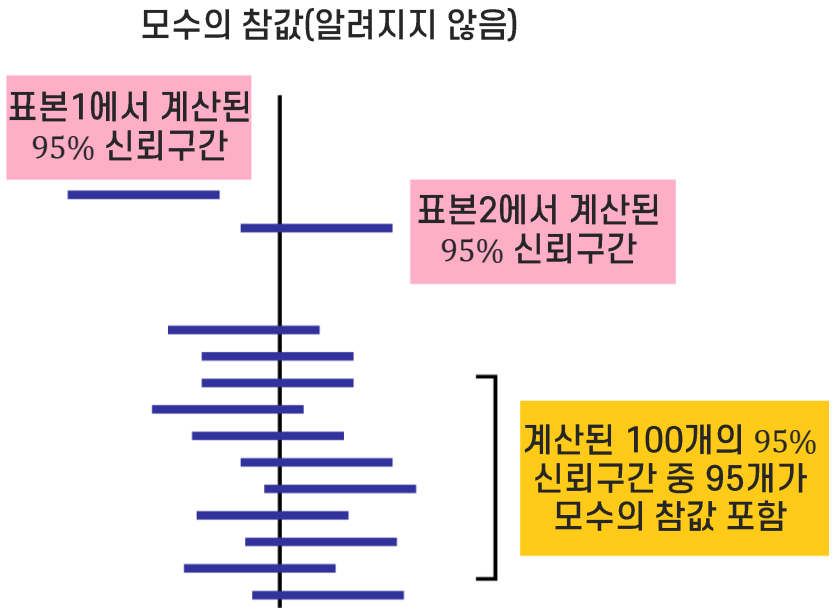
알고 싶은 모수의 참값이 포함되어 있다고 추정되는 구간

- 95% 신뢰구간이란, 신뢰구간을 구하는 일을 무한히 반복할 때 95%의 경우에는 신뢰구간 안에 모수의 참값이 있다는 의미
- 모수의 참값을 포함하고 있을 확률이 95%라는 의미는 아님
  - 표본에서 신뢰구간이 계산이 되면 모수의 참값은 포함/미포함으로 확률의 개념이 들어갈 수 없음

구간추정

신뢰구간

알고 싶은 모수의 참값이 포함되어 있다고 추정되는 구간

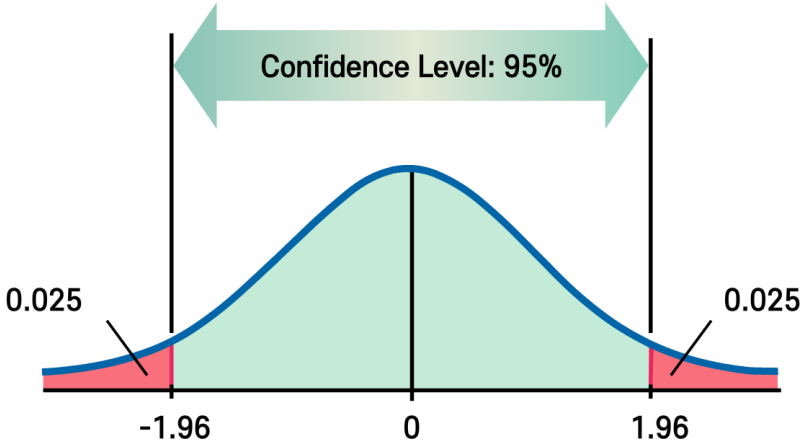


모평균에 대한 신뢰구간

- 모집단으로부터 추출한 크기가 n인 표본에서 계산된 표본평균을 $\bar{X}$이라고 한다면 모평균 μ에 대한 $(1-\alpha)100\%$ 신뢰구간은 **표본평균의 표본분포**를 통해 구해짐
- $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$은 **표준정규분포**를 따른다는 성질을 활용함
  - $P\left(-Z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1-\alpha$ 를 정렬하여 신뢰구간을 구함
  - 정렬한 결과:

$$P\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

모평균에 대한 신뢰구간



- 모분산 (σ^2)이 알려져 있는 경우, 모평균에 대한

  $(1-\alpha)100\%$ 신뢰구간: $\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

모평균에 대한 신뢰구간

- 일반적으로 모분산(σ^2)은 알려져 있지 않고 이 경우는 $\frac{\bar{X}-\mu}{S/\sqrt{n}}$은 **자유도가** $(n-1)$**인 t-분포를 따른다는** 성질을 활용함
  - $P\left(-t_{\alpha/2,n-1} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1-\alpha$를 정렬하여 신뢰구간을 구함
- 모분산을 모르는 경우(일반적), 모평균에 대한 $(1-\alpha)100\%$ 신뢰구간 :

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

모평균에 대한 신뢰구간

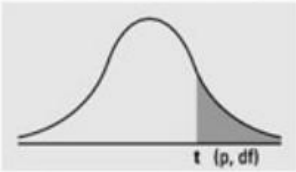
- 모평균에 대한 $(1-\alpha)100\%$ 신뢰구간 :

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

- 다른 조건이 동일하다고 가정했을 때 신뢰구간의 폭을 결정하는 요인
  - 표본의 크기가 작으면 표준오차가 크므로 구간의 폭이 넓어짐
  - 신뢰수준이 커질수록 구간의 폭이 넓어짐
  - 표본의 표준편차가 커지면 표준오차가 커지므로 구간의 폭이 넓어짐
- 너무 넓지 않은 신뢰구간을 얻기 위해서는 충분한 크기를 가진 표본을 구해야 함
  - 하지만 이는 비용 문제와 직결됨

모평균에 대한 신뢰구간

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

학습정리

- 추정은 표본을 통해 모집단 특성을 추측하는 과정
- 모수에 대한 추정은 점추정과 구간추정으로 나누어 볼 수 있음
- 점추정은 하나의 '값'으로 모수를 추정하는 것이라면 구간추정은 '구간'으로 모수를 추정하는 것임
- 표본평균의 표본분포를 활용하여 모평균의 신뢰구간을 도출할 수 있음
- 파이썬 함수들을 활용하여 모평균의 추정을 할 수 있음