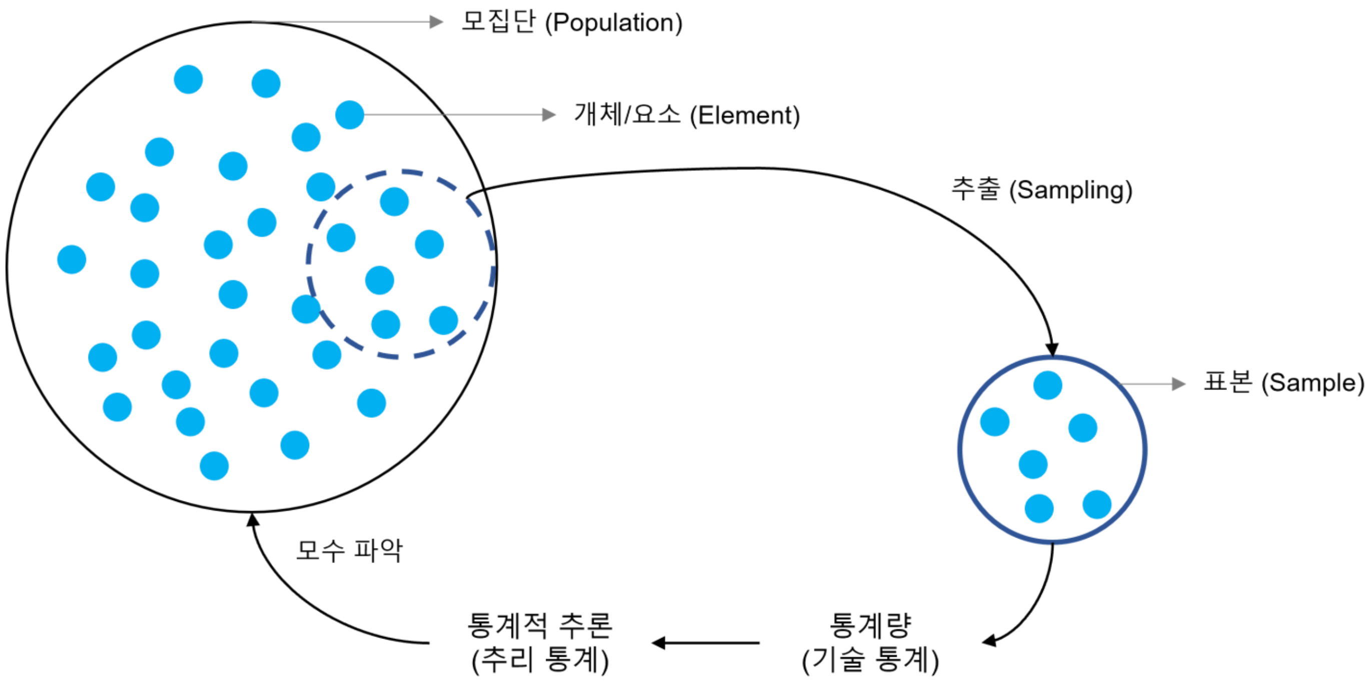
추론 통계학

유튜브 주소 : https://youtu.be/77C3YmpNU-0

IT융합공학부 이준희

‘추론 통계학’ 이란?

- **‘추론 통계학’**
  - ➢ 정의 : 모집단에 대한 미지의 양상을 알기 위해, 표본으로부터 얻은 통계량을 기초로 하여 **모집단의 모수(특성)을 추론하는 것**



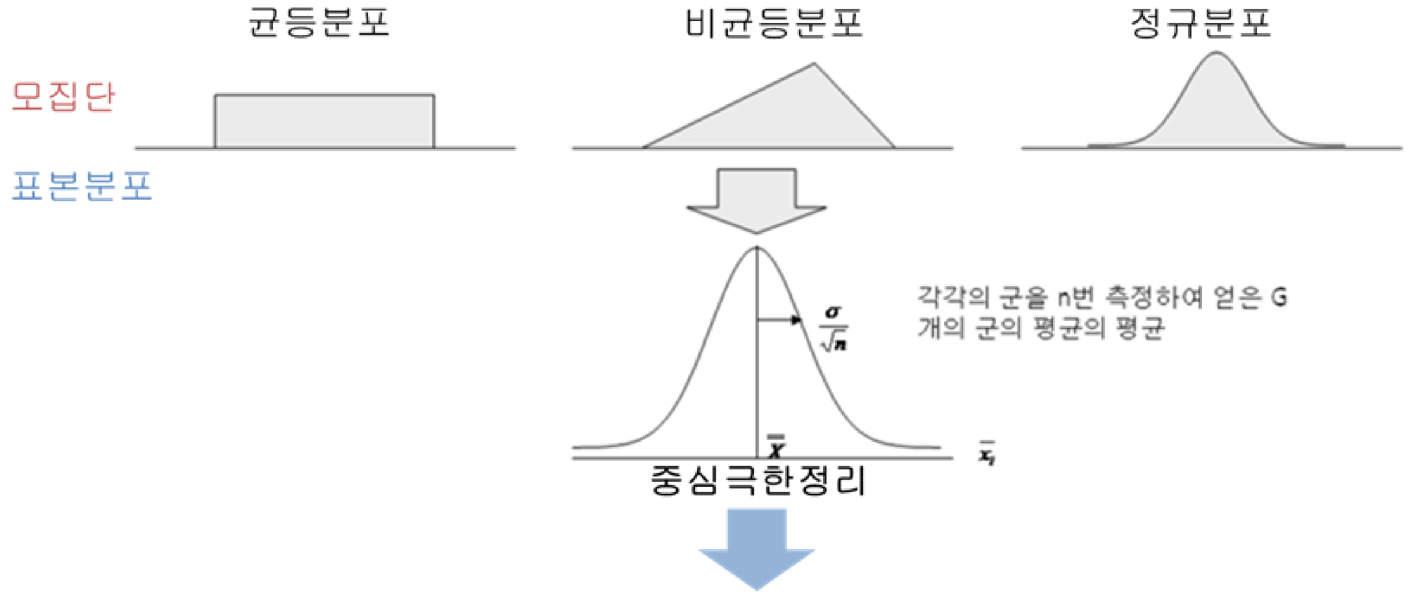
'모집단'과 '표본'

- **모집단의 모수**
  - 모집단의 다양한 특성을 나타내는 통계량, 특성값을 의미한다.
  - ex) 모집단의 평균 μ, 모집단의 표준편차 σ
- **표본의 통계량**
  - 표본 집단의 특성을 나타내는 통계량, 특성값을 의미한다.
  - 대부분의 표본 통계량은 모집단의 모수에 근사하지만**, 표본통계량과 모집단 모수의 개념은 분리**해야 한다.
  - ex) 표본집단의 평균 $\bar{x}$, 표본집단의 표준 편차 S

  - **표본평균 $\bar{x}$의 기댓값(평균)**은 **모집단의 평균 μ**와 일치한다.
  - 표본의 크기가 n일 때, **표본평균 $\bar{x}$들의 분산**은 **모집단의 분산 σ^2을 표본의 크기 n으로 나눈 것과 같다.**

‘중심극한정리’

- **중심극한정리(Central Limit Theorem)**
  - ➢ 동일한 확률분포를 가진 확률변수 n개의 평균 분포는 **n이 적당히 크다면 정규분포에 가까워진다는 정리**이다. (일반적으로 n은 30)
  - ➢ **모집단의 분포와 상관없이** 표본의 수가 큰 표본에서 **표본평균 $\bar{x}$의 분포는 정규분포에 가까운 분포**를 가진다.
  - ➢ 표본평균 $\bar{x}$는 정규분포의 확률변수로서 **평균이 μ**, **표준오차(표본평균의 표준편차)는 $\frac{\sigma}{\sqrt{n}}$**이다.
  - ➢ 이를 표준화하게 되면 다음과 같다. $Z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$



“모집단 분포에 상관없이” 큰 표본들의 “표본평균의 분포”가 정규분포로 수렴한다는 점을 이용하여, Z값을 구해 확률값을 구할 수 있게 된다. 즉, 수학적 확률 판단(추정)을 할 수 있다!

‘**점 추정**’ 이란?

- **‘점 추정’**
  - **‘모집단의 특성을 나타내는 모수’**를 특정한 값으로 추정하는 것
  - 모집단에서 표본을 추출하고, 표본 데이터를 바탕으로 특정한 값을 계산하여 **모집단의 특성(예: 평균, 분산, 비율 등)을 추정**하는 것

ex) 어떤 학교의 학생들 전체(모집단)의 평균 시험 점수를 알고 싶다.
하지만 모든 학생의 점수를 조사하는 것은 현실적으로 어려울 수 있다.
그래서 일부 학생(표본)을 무작위로 선택하여 이들의 시험 점수를 조사한다.

표본으로 30명의 학생을 선택했는데, 이들의 평균 점수가 85점이라면, "학교 전체 학생들의 평균 점수는 약 85점일 것“
이것이 **점추정**이다.

- **점추정량의 조건**

1. 불편성 : 표본에서 얻은 추정값과 모수는 차이가 없다.
2. 효율성 : 최소의 분산을 가진 추정량이 효율적이다.
3. 일치성 : 표본의 크기가 증가할수록 추정량이 정확하다.
4. 충분성 : 모수에 대한 정보를 충분히 제공한다.

‘점추정량’

- 점추정량
  - ➢ 모집단의 특정 특성(모수)을 하나의 값으로 추정하기 위해 **표본 데이터를 바탕으로 계산되는 통계량**

- ‘점추정량’의 종류

① 모평균 μ의 추정량

- 모집단의 평균을 추정하기 위한 추정량으로**, 표본들의 평균**이다.
- 표본을 반복해서 추출할 때마다 다른 값을 가질 수 있기 때문에 **표본평균**은 확률변수이다.
- **표본평균도 확률변수**이기 때문에 특정한 확률분포를 가진다.
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$

② 모분산 σ²의 추정량

- 모집단의 분산을 추정하기 위한 추정량으로, **표본들의 분산**이다.
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - \bar{x})^2$

‘최대우도추정’

- **최대우도추정(Maximum Likelihood Estimation)**
  - ➢ **주어진 표본에서 우도를 최대화하는 모수 θ를 찾는 기법**

- **우도(Likelihood)**
  - ➢ 이미 주어진 표본에 비추어 봤을 때, 모집단의 **모수 θ에 대한 추정이 그럴듯한 정도**
  - ➢ **우도 L(θ | x)** 는 θ 가 전제되었을 때 주어진 표본이 등장할 확률인 **p(x | θ) 에 비례**

✓ 정확하게 이해하기 위해 예를 들어 설명하겠음

- ex) **동전 던지기 100번을 수행**하는 예시에서 반복적인 동전 던지기는 **앞면이 나올 확률이 p** 인 베르누이 시행을
  n 번 반복 시행할 때 성공 횟수의 분포인 이항분포(binomial distribution)를 따른다.
  여기서 **미지의 모수 θ** 는 동전을 한 번 던졌을 때 **앞면이 나올 확률 p** 가 된다.
  이를 위해 앞면이 나올 확률이 p 인 이항분포에서 뽑은 표본 x 를 활용한다.

- 이항분포의 확률분포함수는 다음과 같다.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

'최대우도추정'

- 이 동전이 **앞면 뒷면 나올 확률이 (θ =0.5) 라고 가정**하고 **우도를 계산**하면 다음과 같다.

$$p(X=56|\theta=0.5)=\binom{100}{56}0.5^{56}0.5^{44}\approx 0.0389$$

- 여기서 표본(주어진 데이터)는 "동전 던지기를 100번 했을 때, 앞면이 56번 나온 상황" 이다.
- 모수 θ는 우리가 알고자 하는 미지의 값이다. 그러하여 θ를 가정한 것이다.
- 여기서 중요한게 **θ를 0.5라고 가정하고 계산한 확률값** "0.0389" 이 "**우도(Likelihood)**"이다.
- **θ 를 변형해보면서 우도를 여러 개 구해보자.**

θ	likelihood
0.48	0.0222
0.50	0.0389
0.52	0.0587
0.54	0.0739
0.56	0.0801
0.58	0.0738
0.60	0.0576
0.62	0.0378

‘최대우도추정’

- θ 를 변형해가면서 구한 우도 중에서 **제일 높은 우도에 해당하는 θ 를 구하는 것**이 ‘최대우도추정’ 이다.

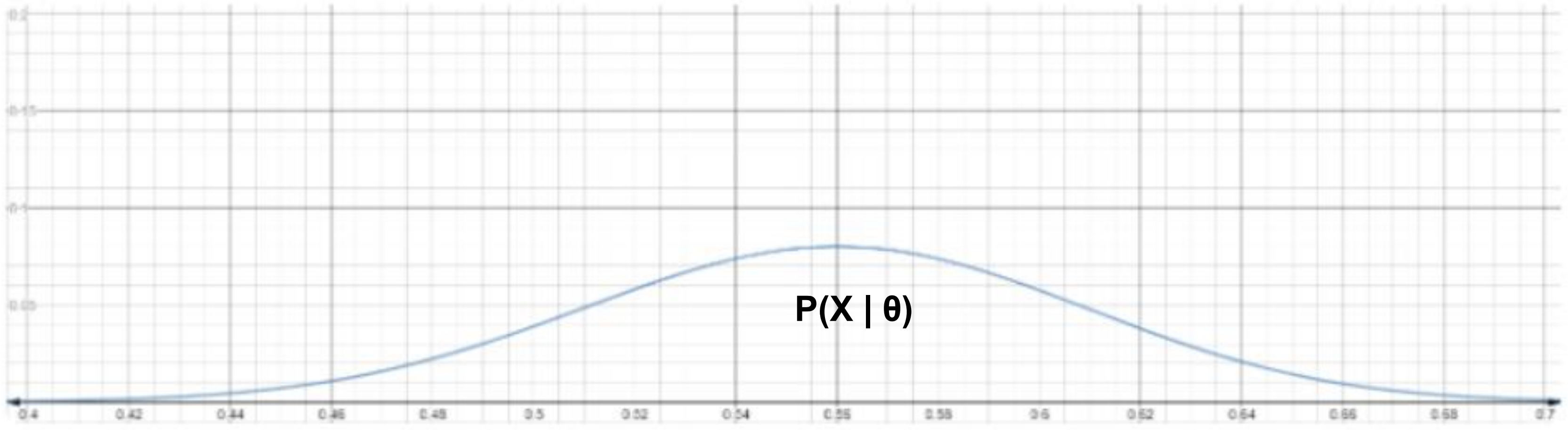
- **‘확률’**과 **‘우도’**는 비슷하지만 엄밀히 다른 개념이다.

- **‘확률** (모수 θ에 해당)’은 미래에 발생할 사건에 대해 예측할 때 사용된다. **사건(표본)이 발생하기 전에 정의**됨.
  - 모델 및 추정치 => 데이터
- **‘우도’**는 이미 발생한 사건(표본)을 설명하기 위해 **모수 θ** (예 : 확률)**를 변화시키며** 그 사건을 가장 잘 설명하는 **모수 θ** 를 찾기 위해 사용
  - 데이터 => 모델 및 추정치

θ	likelihood
0.48	0.0222
0.50	0.0389
0.52	0.0587
0.54	0.0739
0.56	0.0801
0.58	0.0738
0.60	0.0576
0.62	0.0378

‘최대우도추정’

- **P(X | θ) 확률분포함수**를 그래프로 그려보면 다음과 같다.



- 확률분포함수를 보면 **미분이 가능**하다.
- 따라서 모수 θ 에 대해 **편미분을 해 0이 되는 지점**을 구하면 **우도를 최대화 하는 θ** 를 구할 수 있다.
- 하지만 미분이 불가능할 경우에는 ‘**경사하강법(Gradient Descent)’ 과 같은** 반복적이고 점진적인 방식으로 θ 를 추정하게 된다.
- ‘**최대우도추정(MLE)**’은 다양한 통계 모델에서 **파라미터(매개변수, 모수)를 추정**하기 위한 중요한 기법이다.
  - 주어진 데이터를 기반으로 가장 그럴듯한 모델을 찾는 데 필수적인 도구 (데이터 => 모델 및 추정치)

Q & A