3. 표본과 경험적 분포

임채영
서울대학교 통계학과

이번 강의에서 다룰 내용

- 표본의 개념과 분포
- 표본분포의 성질

표본의 개념과 분포

표본이란?

- 알고 싶은 전체집단의 모든 정보를 다 조사하기 힘든 경우 그 일부를 수집하여 전체집단의 정보를 알아보는것이 자연스럽다.
- 전체집단에서 일부를 수집하는 것을 표본조사(Sample Survey)라고 하며 수집된 정보는 표본(Sample)이라고 한다.
- 전체집단을 잘 설명할 수 있는 표본을 추출하는 과학적 추출방법과, 표본이 지닌 불확실성을 계량할 수 있는 통계적 개념이 필요하다.
- 관찰, 측정, 실험 등을 통해 얻어진 데이터를 과학적 추출방법으로 얻어진 표본으로 볼 수 있을 때, 분석결과의 불확실성을 계량화 할 수 있다.

표본조사의 예

- 선거에 출마한 특정 후보 A에 대한 유권자 성향분석을 위해 일부의 의견을 수집하는 경우
- 수집 방법 1: 선거구내의 지하철 역에 가서 1000명에게 지지 여부 조사
- 수집 방법 2: 유권자 명부에서 1000명을 임의 추출하여 전화를 걸어 지지 여부 조사
- 편의 표본: 개체가 표본에 포함되는 가능성을 알 수 없는 표본
- 확률 표본: 개체가 표본에 포함되는 가능성을 알 수 있는 표본
- 단순임의추출: 모든 유권자가 표본에 포함될 가능성이 같다는 의미

<참고> 전체집단을 다 조사하는것을 전수조사(Census)라고 한다.

필요한 몇가지 용어

- 모집단 (Population): 정보를 얻고자 하는 대상이 되는 집단 전체
- 모수 (Parameter): 모집단의 특성을 나타내는 대표값
- 표본 (Sample): 모집단에서 추출한 부분집합
  - 유한 모집단에서의 랜덤표본(Random Sample): 단순 랜덤 비복원추출 (임의추출)로 뽑은 표본
  - 무한 모집단에서 랜덤표본: 동일한 분포(모집단의 분포)를 따르는 독립인 확률변수들의 집합
- 통계량 (Statistic): 표본자료의 특성을 나타내는 값.
  - 통계량은 표본으로 구하므로, 표본의 함수라고 볼 수 있다.
- 추정량 (Estimator): 모수의 추정을 위해 구해진 통계량

표본의 경험적 분포

- 추출된 표본에 속하는 데이터의 분포를 경험적 분포(Empirical Distribution)이라고 한다.
- 표본의 갯수가 증가하면 확률 표본에서 얻은 정보는 모집단에 대한 정보와 점점 가까워 진다.
- 확률표본으로 구한 경험적 분포가 표본이 갯수가 증가하면 모집단의 분포에 점점 가까워 진다.

표본분포란?

- 표본이 지닌 불확실성에 대한 수학적 표현
- 확률 표본으로부터 구한 통계량 (평균, 분산 등)의 분포

- 표본은 임의로 추출하였으므로 추출전에는 어떤 표본이 나올지 알 수 없고 추출할 때마다 다른 관측값이 나옴. 따라서 표본으로부터 계산한 통계량도 임의성을 가짐.
- 표본이 확률표본인 경우, 통계량은 확률변수로 볼 수 있음. 따라서, 특정한 확률분포를 따르게 되고 이 분포를 표본분포(Sampling Distribution)라 한다.

표본분포의 예

시리얼제품에 5개에 1+1 할인 쿠폰이 한장씩 있다고 하자. 당첨, 당첨, 당첨,탈락,탈락으로 이루어졌다고 할때 여기서 크기 3인 표본을 단순 랜덤 비복원추출로 뽑아 당첨비율 θ (이 예시에서는 0.6)을 추정하는 상황(당첨=1, 탈락=0)을 생각해보자.

Table: 가능한 표본과 그 확률 및 대응되는 표본비율

가능한 표본	확률	표본비율 ($\hat{\theta}$)
당첨 3, 탈락 0	$\frac{{}_3C_3 \times {}_2C_0}{{}_5C_3}$	1
당첨 2, 탈락 1	$\frac{{}_3C_2 \times {}_2C_1}{{}_5C_3}$	2/3
당첨 1, 탈락 2	$\frac{{}_3C_1 \times {}_2C_2}{{}_5C_3}$	1/3

- '표본비율'은 모수인 모비율을 표본으로 추정한 추정량이다.
- 표본비율은 표본에 따라 다른 값을 가진다.

- 이 예시에서는 해당하는 표본이 나오는 확률이 표본비율의 값에 대응되는 확률로, 표본비율의 표본분포가 된다.
- 만약 모비율이나 표본의 크기가 달라진다면 표본비율의 분포 역시 달라진다. 즉, 표본분포는 모집단의 분포와 표본 추출 방식에 영향을 받는다.

표본의 크기가 클 때

- 만약 어느 회사에서 시리얼 제품 10000개를 생산하였고, 이 중 일부(30%)에 1+1 할인 쿠폰을 넣었다고 하자. 여기에서 크기가 100인 표본을 임의추출하여 할인 쿠폰의 비율을 알고 싶다고 하면, 표본을 추출했을때 나올 수 있는 경우의 수를 다 계산하여 표본비율의 분포를 구하는 것은 어렵다.
- 대신 가능한 표본을 모두 고려하지 않고 모집단에서 많은 수의 표본을 독립적으로 임의추출하는 모의실험(Simulation)을 통해 표본비율의 표본분포를 근사적 (경험적 분포)으로 구해볼 수 있다.

표본비율의 경험적 분포

- 시리얼 제품이 10000개 있고 이 중 100명을 표본으로 추출 할 때 표본 비율의 경험적 분포는 다음 절차로 구할 수 있다.
  - 모집단에서 100개의 표본을 임의 추출.
  - 추출된 표본으로부터 표본비율($\hat{\theta}$)을 계산.
  - 위 과정을 $B \geq 500$번 반복.
- $\hat{\theta}_i$를 i번째 표본에서 얻는 표본비율이라고 하면 위의 과정에서 B개의 표본 비율을 얻을 수 있음.

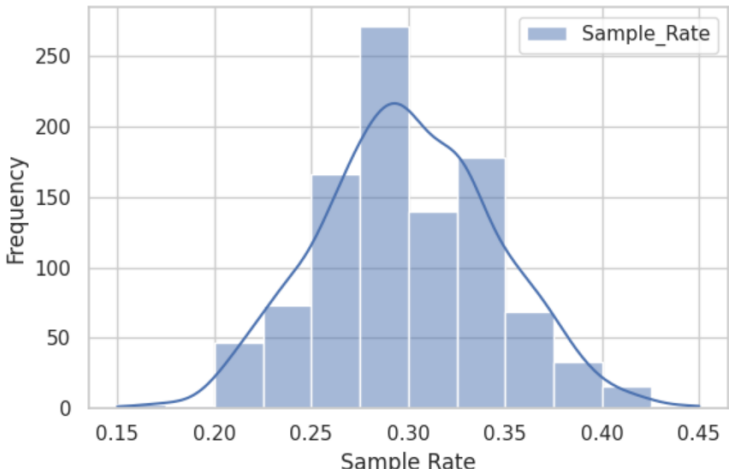
$$(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_B)$$

- 이렇게 모의실험을 통해 얻은 표본 비율의 경험적 분포는 표본비율의 실제 분포와 매우 유사함.

모의실험을 통한 통계량의 경험적 분포 $\approx$ 통계량의 분포

- 모의실험의 반복 횟수 B는 클수록 좋으며 일반적으로 500이상의 수를 고려한다.

$B = 1000$일때의 모의실험을 통한 표본비율의 히스토그램



통계적 추정

- 분포는 관측할 수 있는 값의 가능성을 나타내므로 표본 통계량의 정확한 분포를 알고 있으면, 표본통계량을 이용해 모집단의 모수를 추정하는 방법이 얼마나 합리적인지 판단 할 수 있는 근거를 제시할 수 있음
- 그러나, 모집단의 크기가 크고, 표본의 크기도 크면 표본 통계량의 분포를 정확하게 구하기 어렵다.
- 따라서, 무한모집단에서의 랜덤표본으로 간주하고 구한 이론적 표본분포를 실제 표본분포의 근사분포(approximated distribution)로 사용하거나 모의실험을 통해 구한 경험적 분포를 사용한다.
- 확률표본으로 자료를 수집하고 이를 통해 알고자 하는 모집단의 모수를 추정하는것을 통계적 추정(Estimation)이라고 한다.

표본분포의 성질

표본평균

- 표본에 속한 데이터들의 평균을 표본평균(Sample Mean)이라고 하고 $\bar{X}$로 표시한다.
- 표본의 중심 경향성을 나타내는 통계량이다.
- 모집단의 평균 (모평균)을 μ라고 하면, 표본평균은 μ의 추정량 (estimator)이다.
- 표본 $\{X_1, X_2, \cdots, X_n\}$가 모평균 μ, 모분산 σ^2인 모집단에서 추출된 랜덤표본(i.i.d.)일때,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

표본평균의 평균과 분산

- 무한모집단에서 추출된 랜덤표본일 경우,

$$E(\bar{X}) = \mu,\ Var(\bar{X}) = \frac{\sigma^2}{n},\ sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- 크기가 N인 유한모집단에서 추출된 랜덤표본일 경우,

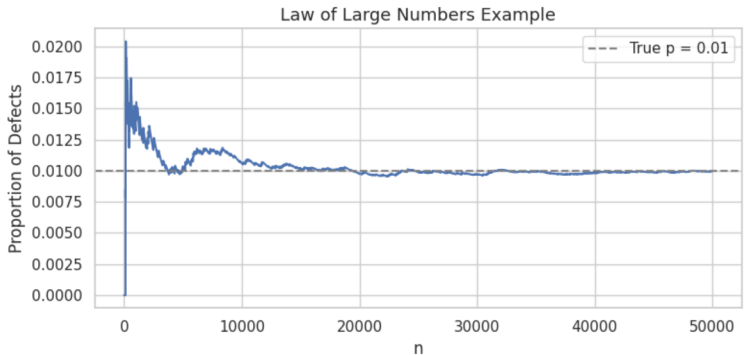
$$E(\bar{X}) = \mu,\ Var(\bar{X}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}.$$

큰수의 법칙

- 표본의 크기 n 이 커질수록 표본평균의 분산은 0에 가까워진다.
- 표본평균의 기댓값은 모평균과 같고, 분산이 작아지므로, $\bar{X}$는 모평균 μ의 근처에 밀집되어 분포함을 알 수 있다.
- 이러한 결과를 큰수의 법칙(Law of Large Numbers, LLN, 대수의 법칙)이라고 한다.

큰수의 법칙 예

- 공장 A에서 생산되는 배터리는 불량품일 확률이 1%라고 하자.
- X_i는 i번째 임의 추출한 배터리가 불량이면 1, 정상이면 0을 갖는 확률변수라고 하자.
- $\bar{X}_n$는 공장 A에서 생산되는 배터리를 n개 임의추출하였을때의 불량품의 비율과 같고, $\bar{X}_n \to \mu = E(X_1) = p = 0.01$이 된다.



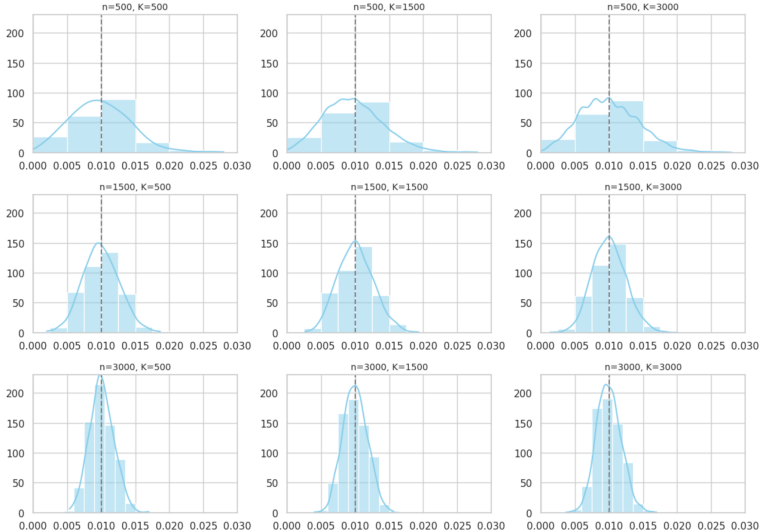
중심극한정리

- 임의의 모집단에 대해 $(\bar{X} - \mu)/(\sigma/\sqrt{n})$의 분포는 표준정규분포 $N(0, 1)$에 근사한다. 이를 중심극한정리(Cenltral Limit Theorem)라 한다.
- 유한모집단의 경우, 모집단의 크기 N과 표본의 크기 n이 충분히 크면(단 $N \gg n$) $\frac{N-n}{N-1}$의 값이 1에 근사하므로, 위의 성질이 성립한다.
- 중심극한정리를 통해, 모집단의 분포가 어떤 형태이든지 표본의 크기가 크면 표본평균의 분포를 정규분포로 근사할 수 있다.

  즉, $\bar{X}$의 분포 $\approx N\left(\mu, \frac{\sigma^2}{n}\right)$.

중심극한 정리의 예

- 공장 A에서 생산되는 배터리는 불량품일 확률이 1%라고 하자.
- X_i는 i번째 임의 추출한 배터리가 불량이면 1, 정상이면 0을 갖는 확률변수라고 하자.
- $\bar{X}_n$는 공장 A에서 생산되는 배터리를 n개 임의추출하였을때의 불량품의 비율과 같다.
- $\bar{X}_n$의 분포를 알기 위해서는 여러개의 $\bar{X}_n$이 필요하다.
- K개의 $\bar{X}_n^{(1)}, \cdots, \bar{X}_n^{(K)}$를 n을 바꿔가며 분포를 확인해 보자.


n=500, K=500
n=500, K=1500
n=500, K=3000
n=1500, K=500
n=1500, K=1500
n=1500, K=3000
n=3000, K=500
n=3000, K=1500
n=3000, K=3000
0 50 100 150 200
0.000 0.005 0.010 0.015 0.020 0.025 0.030

표본오차

- 표본으로 구한 추정량(통계량)과 모집단의 모수사이의 차이를 표본오차(Sampling Error)라고 한다.
- 중심극한정리에 의해 추정량의 분포가 정규분포에 가까우므로

  추정량($\bar{X}$)과 그 평균 (μ)의 차이가 1.96$\times$ 표준오차 (SE, $\sqrt{\sigma^2/n}$) 이내 일 확률이 근사적으로 95%임을 알 수 있다. 즉, $P(\mu - 2SE \leq \bar{X} \leq \mu + 2SE) \approx 0.95$
- 이를 정리하여 구한 $\bar{X} - 1.96SE \leq \mu \leq \bar{X} + 1.96SE$를 μ의 95% 신뢰구간이라고 한다.
- 이때 사용한 확률 95%를 신뢰수준이라고 부르며 $1.96 \times SE$가 표본오차이다.

- 표준오차(Standard Error)는 통계량(추정량)의 표준편차이다.

스튜던트화된 표본평균의 분포

- 모집단의 분포가 정규분포 $N(\mu, \sigma^2)$인 경우 표준화한 분포는 다음과 같다.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- 표준편차 σ를 알 수 없으므로 데이터로 구한 추정량인 표본표준편차 S를 대입한 경우를 스튜던트화 된 표본평균 (Studentized sample mean)이라고 한다.
- 스튜던트화된 표본평균의 분포는 정규분포와는 약간 다르다.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/\frac{\sigma}{\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t(n-1)$$

- $t(n-1)$은 자유도가 $n-1$인 $t-$분포라고 한다. 정규분포보다 꼬리가 두껍지만 자유도가 커질수록 정규분포에 가까워 진다.

The End