확률통계 및 공통 알고리즘

3. 표본분포

임채영

서울대학교

이번강의에서 다룰 내용

- 표본분포의 개념
- 표본분포의 성질
- 표본평균의 분포
- 여러가지 표본분포

표본분포 소개를 위한 몇가지 개념

- 모집단 (population): 정보를 얻고자 하는 대상이 되는 집단 전체
- 모수 (parameter): 모집단의 특성을 나타내는 대표값
- 표본 (sample): 모집단에서 추출한 부분집합
  - 유한모집단에서 랜덤표본(Random Sample): 단순 랜덤 비복원추출로 뽑은 표본
  - 무한모집단에서 랜덤표본: 동일한 분포(모집단의 분포)를 따르는 독립인 확률변수들
- 통계량 (Statistic): 표본자료의 특성을 나타내는 값.
  - 통계량은 표본으로 구하므로, 표본의 함수라고 볼 수 있다.
- 추정량 (Estimator): 모수의 추정을 위해 구해진 통계량

표본분포란?

- 통계량의 확률 분포
  - 랜덤표본의 값에 따라 통계량의 값 역시 정해진다. 이 때, 통계량 역시 확률변수로서 특정한 확률분포를 따르게 되고, 그 분포는 모집단의 분포와 관계가 있다.

모수와 추정량 예시

5개의 시리얼 박스에 쿠폰이 한장씩 있다고 하자. 당첨, 당첨, 당첨,탈락,탈락으로 이루어졌다고 할때 이 모집단에서 크기 3인 표본을 단순 랜덤 비복원추출로 뽑아 모비율(당첨비율) p (이 예시에서는 0.6)을 추정하는 상황(당첨=1, 탈락=0)을 생각해보자.

Table: 가능한 표본과 그 확률, 표본비율, 오차

가능한 표본	표본비율 ($\hat{p}$)	모비율과의 오차
당첨 3, 탈락 0	1	0.4
당첨 2, 탈락 1	2/3	0.067
당첨 1, 탈락 2	1/3	-0.267

- ’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 방식의 영향을 받는다.

유한모집단의 표본분포

- 유한모집단의 랜덤표본의 경우:
  - 소수 알려진 분포가 존재하지만 비복원추출이기때문에 일반적으로 표본분포를 유도하기 복잡하다.
- 유한 모집단의 크기가 큰 경우:
  - 일반적으로 무한모집단에서의 랜덤표본으로 간주하고 표본분포를 구해서 실제 표본분포의 근사 (approximated) 분포로 사용한다.

표본평균의 분포

- 표본평균 (sample mean), $\bar{X}$
  - 표본의 중심경향성을 나타내는 통계량.
  - 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 랜덤표본일때,

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

표본평균의 분포에 대한 성질

- 무한모집단에서 추출된 랜덤표본일 경우,

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

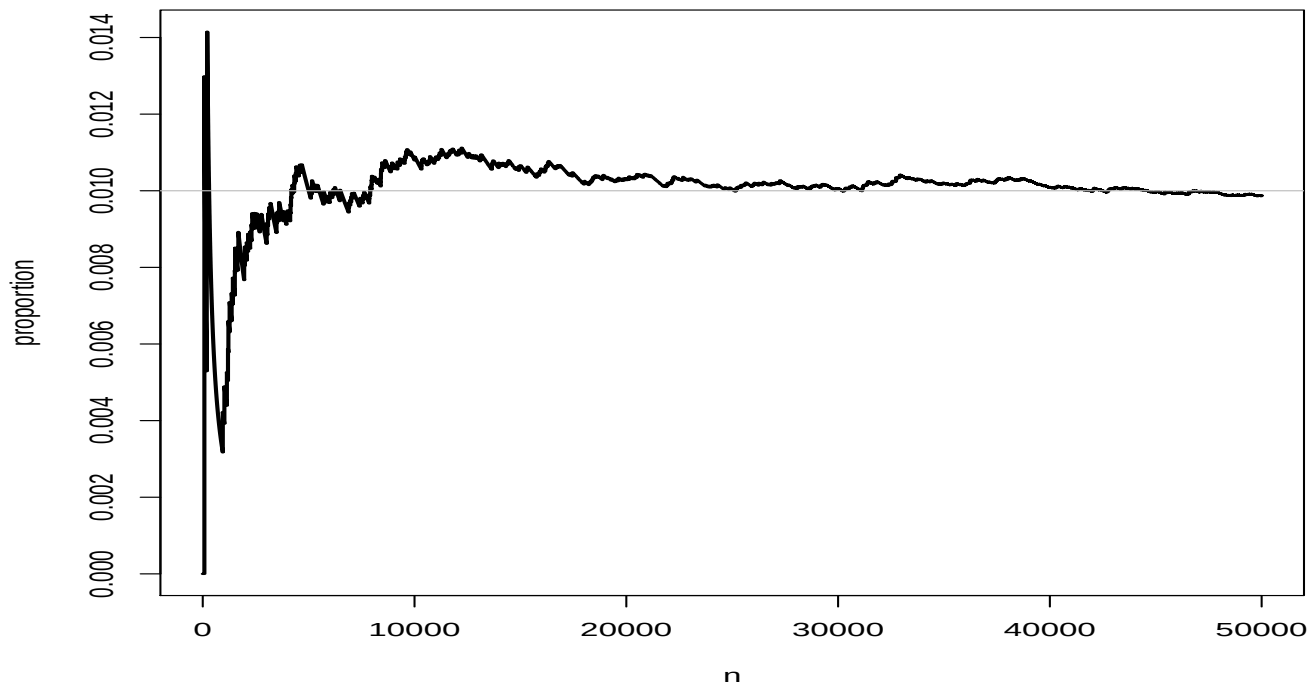
- 크기가 N인 유한모집단에서 추출된 랜덤표본일 경우,

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

큰수의 법칙 (Law of large numbers)

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

- 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 임의추출하였을때의불량품의 비율과 같고, $\bar{X}_n \to \mu = E(X_1) = p = 0.01$이 된다.

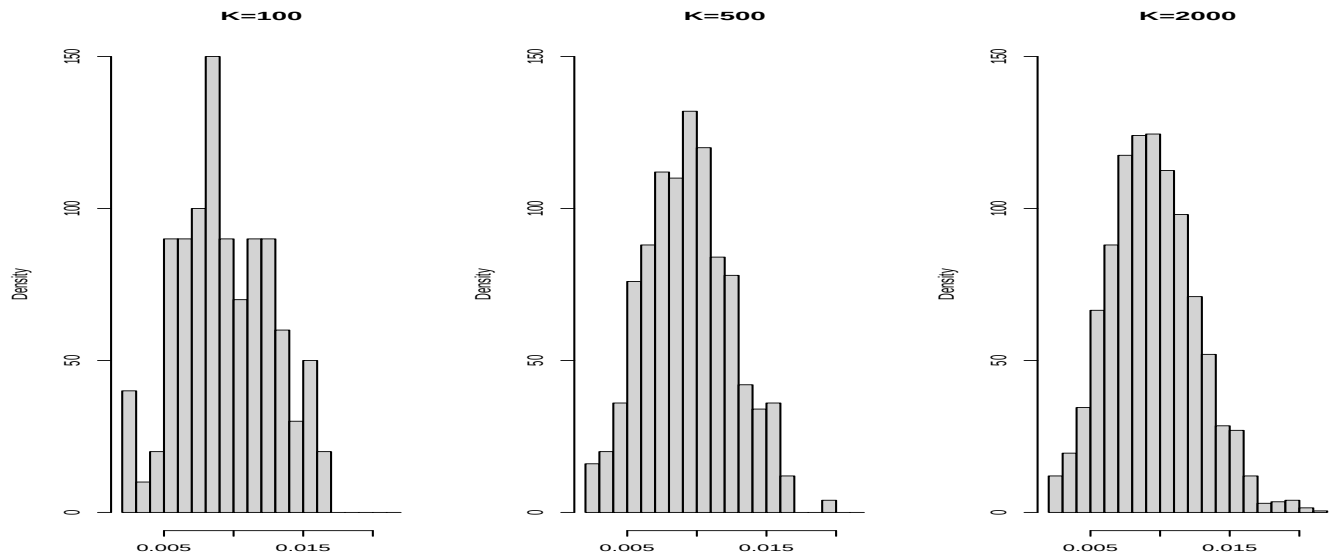


중심극한정리 (Central limit theorem)

- 임의의 모집단에 대해 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$의 분포는 표준정규분포 $N(0,1)$에 근사한다.
- 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 $\approx N\left(\mu, \frac{\sigma^2}{n}\right)$.

- 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 임의추출하였을때의불량품의 비율과 같다.
- $\bar{X}_n$의 분포를 알기 위해서는 여러개의 $\bar{X}_n$이 필요하다.
- $n = 1000$이라하고, K개의 $\bar{X}_n^{(1)}, \cdots, \bar{X}_n^{(K)}$를 구해서 분포를 확인해 보자.



이항분포의 정규분포 근사

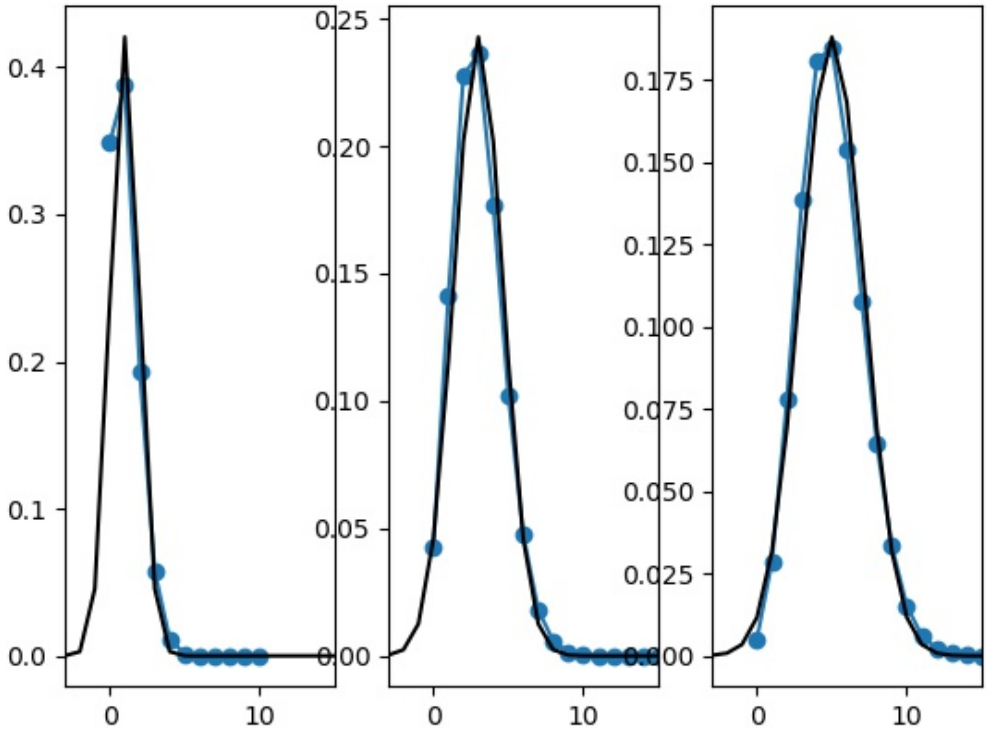
- $X_1, X_2, \cdots, X_n$이 성공률이 p 인 베르누이분포를 따르는 무한모집단의 랜덤표본이라고 하자
- 이 경우, $S = \sum_{i=1}^{n} X_i$ 은 이항분포 $B(n, p)$을 따른다.
- 중심극한정리를 적용하면, n이 충분히 클 때

$$\frac{S - np}{\sqrt{(np(1-p))}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

의 분포는 표준정규분포 $N(0, 1)$에 근사한다.
($\hat{p}$= 베르누이분포의 표본비율 $\frac{S}{n}$)

- 즉, n이 충분히 크고, np가 적당한 값이면, $B(n, p)$를 이용하는 확률계산을 $N(np, np(1-p))$를 이용하여 근사할수 있다.

이항분포의 정규근사 그래프



파랑: p=0.1인 이항분포, 검정: 평균이 np, 분산이 np(1-p)인 정규분포. 왼쪽부터 차례대로 n=10,30,50

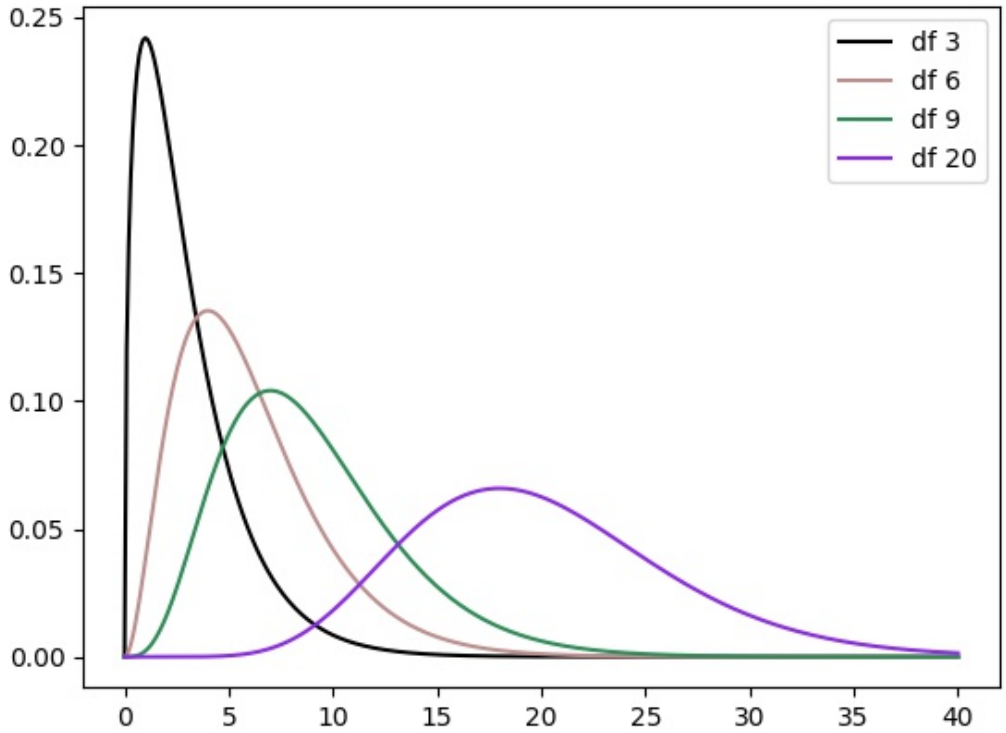
여러 가지 표본분포 : 카이제곱분포

- 표본분산의 분포와 관련이 있다.
- 확률변수 $Z_1, Z_2, \cdots, Z_k$이 정규분포 $N(0,1)$로부터 추출한 k개의 랜덤표본일 때,

$$V = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

의 분포를 자유도(degrees of freedom, df)가 k인 카이제곱분포(Chi-square distribution)라 한다.

- $V \sim \chi^2(k)$



자유도를 3,6,9,20으로 변화시킨 카이제곱분포의 확률밀도함수.
자유도가 작을 수록 0 근방에 밀집되어있다.

카이제곱분포의 성질

- $V \sim \chi^2_k$, $E(V) = k$, $var(V) = 2k$.
- $V_1 \sim \chi^2(k_1)$, $V_2 \sim \chi^2(k_2)$, V_1과 V_2가 서로 독립,

$$V_1 + V_2 \sim \chi^2(k_1 + k_2)$$

표본분산의 분포

- 표본분산: 표본의 산포 (spread)를 나타내는 통계량
  - 모분산의 추정량

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- 모집단이 정규분포 $N(\mu, \sigma^2)$인 경우,

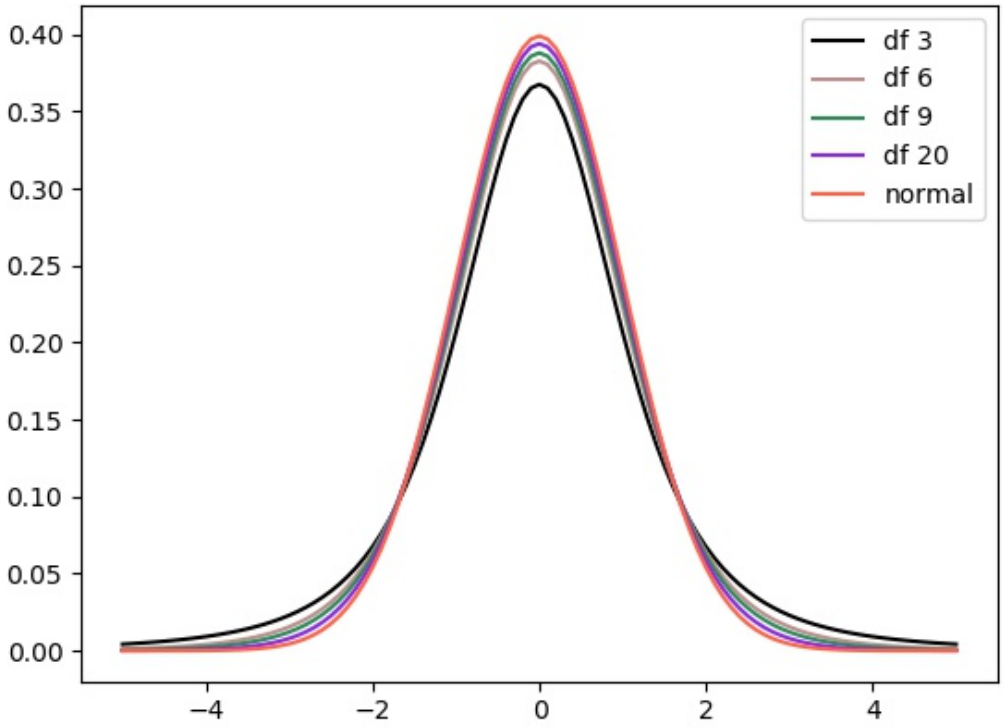
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

여러 가지 표본분포 : t분포

- 스튜던트화 된 표본평균의 분포와 관련이 있다.
- $Z \sim N(0,1)$, $V \sim \chi^2_k$, Z와 V는 서로 독립,

$$T = \frac{Z}{\sqrt{V/k}}$$

  의 분포를 자유도가 k인 t 분포라고 한다.
- $T \sim t(k)$



자유도를 3,6,9,20으로 변화시킨 t분포의 확률밀도함수. 자유도가 작을 수록 표준정규분포에 비해 두꺼운 꼬리를 가진다.

t분포의 성질

- $T \sim t(k)$, $E(T) = 0$
- 정규분포의 확률밀도함수보다 꼬리가 두껍다.
- $k \to \infty$, $t(k) \to N(0, 1)$.

스튜던트화된 표본평균의 분포

- 모집단의 분포가 정규분포 $N(\mu, \sigma^2)$인 경우,

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

- $\frac{\bar{X}-\mu}{S/\sqrt{n}}$: 스튜던트화 된 표본평균
  - 모분산 대신 표본분산을 사용
- 스튜던트화된 표본평균에는 모분산 대신 분산의 추정량이 들어가게 되므로 그 분포는 정규분포와는 약간 다르다.

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