표본분포

숙명여자대학교 경영학부 오중산

모집단과 표본

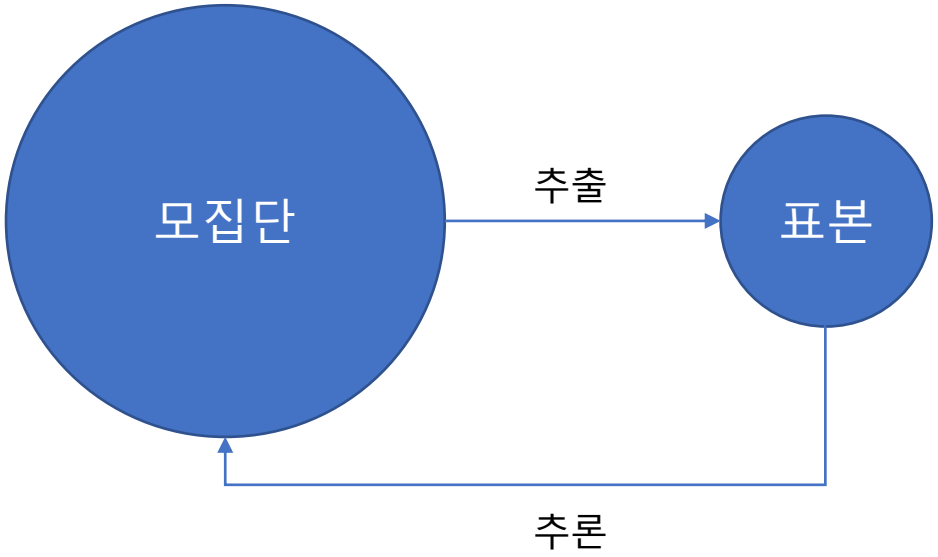
- 모집단(population)이란?
  - 데이터를 획득할 사례(case) 전체를 모집단이라고 함
    - 예1: 국내 200개 4년제 대학 재학생 150만 명
    - 예2: 18세 이상 유권자 4,000만 명
    - 모집단의 크기를 N이라고 표기함
  - 모집단을 대상으로 변수의 측정값, 즉 데이터 획득은 거의 불가능함
    - 시간과 비용 측면에서 효율성이 떨어지고, 누락되는 사례가 존재함

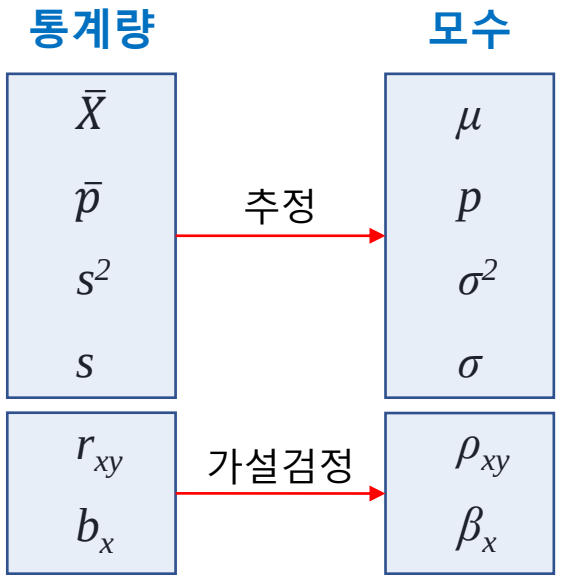
모집단과 표본

- 표본(sample)이란?
  - 모집단의 일부이며, 모집단 대신 실제 데이터 획득 대상
  - 모집단으로부터 표본을 뽑는 과정을 추출(sampling)이라고 함
    - 모집단으로 가정하는 표본프레임(sample frame)으로부터 추출
      - 예: 국내 200개 4년제 대학별 학적부가 표본프레임이 될 수 있음
    - 표본에 모집단의 특성이 잘 반영될 수 있도록 추출해야 표본의 선택편의(selection bias)가 적고, 모집단에 대한 대표성이 높아짐
      - 예: 국내 200개 4년제 대학별로 재학생의 5%를 학년/성별/전공 등을 감안하여 임의로 추출
    - 표본의 개수는 n으로 표시함

모집단과 표본

- 추론통계(inferential statistics)란?
  - 통계량(statistics)은 표본에 적용한 함수이고, 모수(parameter)는 모집단 특성을 반영한 수치
  - 추론통계는 통계량을 이용하여 모수를 추정(estimation)하거나 가설검정하는 것





모집단과 표본

- 추정량(estimator)이란?
  - 여러 개의 통계량 중에서 모수 추정에 사용되는 통계량을 추정량이라고 함
    - 예: 모평균(모수) 추정을 위한 통계량 후보군으로는 표본평균, 범위(range: max – min), trimmed(상·하위 10%를 제외한 표본평균) 등이 있으며, 이때 표본평균이 추정량이 됨
  - 추정량이 되기 위해서는 다음과 같은 조건을 만족하는, 즉 편의(bias)가 없어야 하며, 이 때 추정량을 불편추정량(unbiased estimator)이라고 함

$$E[\theta(\text{모수})] - E[\hat{\theta}(\text{추정량})] = E[\theta - \hat{\theta}] = 0$$

모집단과 표본

- 표본평균 vs. 표본분산
  - 표본평균은 모평균의 불편추정량 $\quad \mathrm{E}[\mu] - \mathrm{E}[\bar{X}] = \mu - \mu = 0$
  - 표본분산(표본표준편차)은 모분산(모표준편차)의 불편추정량이 아닌 편의추정량

$$\mathrm{E}[\sigma^2] - \mathrm{E}[s^2] = \mathrm{E}[\sigma^2 - s^2] = \mathrm{E}[\frac{1}{n}\sum_1^n (\mathrm{X} - \mu)^2 - \frac{1}{n}\sum_1^n (\mathrm{X} - \bar{X})^2]$$

$$= \mathrm{E}[\frac{1}{n}\sum_1^n \{(\mathrm{X}^2 - 2\mathrm{X}\mu + \mu^2) - \{\mathrm{X}^2 - 2\mathrm{X}\bar{X} + \bar{X}^2\}] = \mathrm{E}[\frac{1}{n}\sum_1^n \{\mu^2 - \bar{X}^2 + 2\mathrm{X}(\bar{X} - \mu)\}]$$

$$= \mathrm{E}[\mu^2 - \bar{X}^2 + \frac{1}{n}\sum_1^n 2\mathrm{X}(\bar{X} - \mu)] = \mathrm{E}[\mu^2 - \bar{X}^2 + 2\bar{X}(\bar{X} - \mu)]$$

$$= \mathrm{E}[\bar{X}^2 - 2\bar{X}\mu + \mu^2] = \mathrm{E}[(\bar{X} - \mu)^2] = \mathrm{E}[(\bar{X} - \mathrm{E}[\bar{X}])^2] = \mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$$

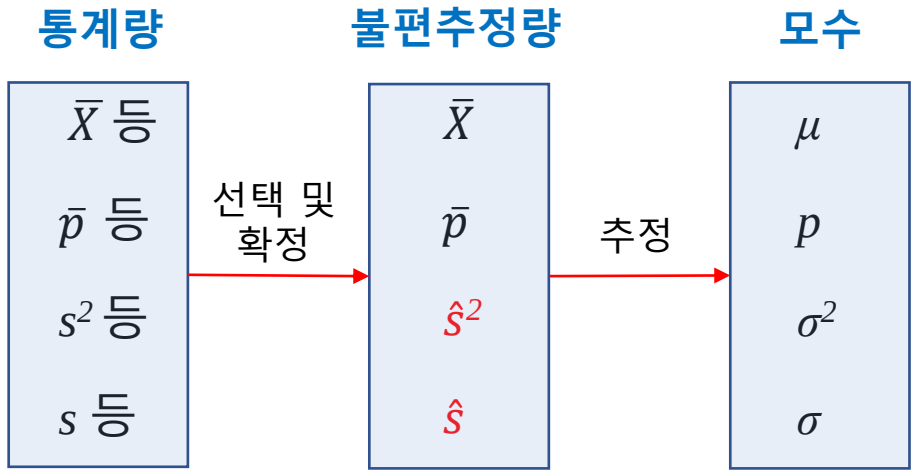
$$Var[\mathrm{X}] = \mathrm{E}[(\mathrm{X} - \mathrm{E}[\mathrm{X}])^2]$$

$$\bar{X} \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2)$$

$$\mathrm{E}[\sigma^2 - s^2] = \sigma^2 - s^2 = \frac{\sigma^2}{n} \rightarrow \sigma^2 = \frac{n}{n-1} \times s^2 = \frac{\sum(X - \bar{X})^2}{n-1}$$ (불편표본분산)

모집단과 표본

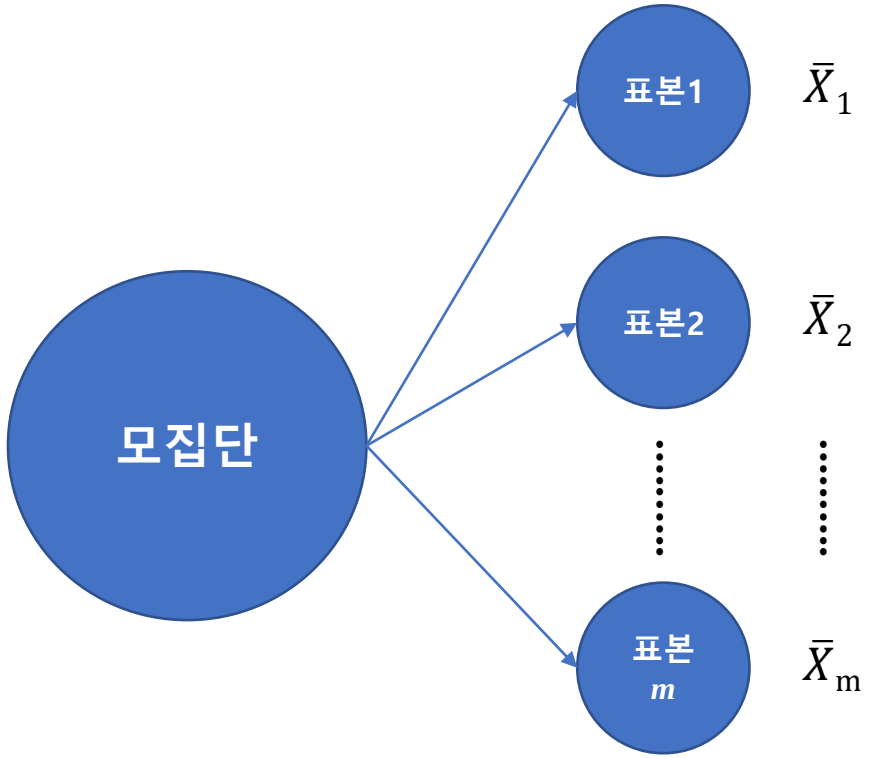
- 통계량과 불편추정량 및 모수 정리
  - STEP1: 복수의 통계량 중에서 통계량 선택
  - STEP2: 해당 통계량이 불편추정량인지 확인
  - STEP3: 불편추정량 확정
  - STEP4: 모수에 대해 추정



- 주의사항! 표본분산과 불편표본분산은 목적이 다름
  - 전자는 기술통계를, 후자는 추론통계에 활용됨
  - 표본크기가 증가하면 양자는 일치하게 됨

표본분포

- 표본분포(sampling distribution)의 정의
  - (표본)통계량이 띠게 되는 분포를 표본분포라고 함
    - 표본분포를 구하려면 표본크기(n)가 동일한 복수의 표본이 있어야 함
    - 만약 N/n이 충분히 크지 않아서 복수의 표본 구성이 어려우면 복원 추출
- 표본분포의 유형
  - 대표적으로 표본평균의 분포와 표본비율의 분포가 있음



표본평균의 표본분포

- N이 상당히 큰 수라고 가정할 때(무한모집단), 표본평균의 표본분포
  - 다음 두 가지 조건 중 하나를 만족할 경우, 표본평균의 표본분포는 정규분포임
    - 조건1: 확률변수(X)가 모집단에서 정규분포를 띰
    - 조건2: 표본크기(n)가 최소 30개 이상
      - 중심극한정리(Central Limit Theorem): 확률변수가 모집단에서 어떤 분포를 띠든, 표본크기가 커지면 표본평균은 정규분포를 띠게 됨
  - 표준오차(SE: standard error)
    - (표본)통계량의 표준편차를 SE라고 함

$$\bar{X} \sim N(\mu, (\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}})^2) \quad \text{z} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$$

표본평균의 표본분포

- 모표준편차를 모를 경우 무한모집단에서의 표본평균의 표본분포
  - 표본평균이 정규분포를 띨 조건을 만족한 상태에서 모표준편차를 모르면, 모표준편차를 불편표준편차로 대체함
  - 표준화변량 z가 아니라, 스튜던트화변량 t로 표준화됨

$$\sigma = \sqrt{\frac{n}{n-1}} \times s \quad \bar{X} \sim N(\mu, (\sigma_{\bar{X}} = \frac{s}{\sqrt{n-1}})^2) \quad t = \frac{(\bar{X} - \mu)}{s/\sqrt{n-1}}$$

표본비율의 표본분포

- 이항분포(Binomial distribution)
  - 확률변수 X(성공회수) ~ B(n, p)
    - E[X] = np, Var[X] = np(1-p)
    - n이 커지면 이항분포는 정규분포에 수렴하므로, X ~ N(np, np(1-p))
- 표본비율의 표본분포
  - X는 이항분포를 띠므로, 표본비율($\bar{p}$)도 이항분포를 띰
  - E[$\bar{p}$] = E[X/n] = E[X]/n = np/n = p
  - Var[$\bar{p}$] = Var[X/n] = Var[X]/n^2 = np(1-p)/n^2 = p(1-p)/n
  - n이 증가하여 np & n(1-p) ≥ 5이면 $\bar{p}$는 정규분포로 수렴: $\bar{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}}^2)$