확률 및 통계학

- 8장 확률표본과 표본분포 -

명 세인(sein@pel.smuc.ac.kr)

상명대학교 프로토콜공학연구실

목 차

- 확률표본
- 대표적 통계량
- 표본분포와 표본평균의 분포

확률표본

- 모집단(Population)
  - 관심이 있는 대상과 관련된 모든 관측 가능한 값의 집합

- 표본(Sample)
  - 모집단의 부분 집합
  - 전체 모집단을 측정할 수 없으므로 표본을 추출
    - 모집단의 성질을 가장 잘 반영할 수 있도록 추출
    - 편향(Bias)되지 않게 추출

확률표본

- 정의
  - 서로 독립인 n개의 확률변수 $X_1, X_2, \ldots X_n$이 동일한 확률분포 $f(x)$를 따를 때, $X_1, X_2, \ldots X_n$을 모집단$f(x)$로부터, 크기 n인 **확률표본**(Random Sample)
    - 결합확률분포는 $f(x_1, x_2, \ldots x_n) = f(x_1)f(x_2) \ldots f(x_n)$으로 표현

대표적 통계량

- 통계량(Statistic)
  - 미지의 모집단에 대한 정보
    - 모집단의 모든 원소에 대해 측정 불가능
    - 커피 선호도를 조사하기 위해 대표본(Large Sample)을 추출, 특정 커피 선호도 p에 대한 정보 $\hat{p}$ (Hat)를 조사하고, $\hat{p}$는 참값 P를 추론하기 위함

- 정의
  - 확률표본을 구성하는 확률변수들의 함수를 **통계량**으로 함

대표적 통계량

- 중심 통계량
  - $X_1, X_2, \ldots X_n$을 크기가 n인 확률표본이라 할 때
    - 표본평균(Sample Mean)

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

    - 표본중앙값(Sample Median)

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, n\text{이 홀수일 때} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), n\text{이 짝수일 때} \end{cases}$$

대표적 통계량

- 중심 통계량
  - $X_1, X_2, \ldots X_n$을 크기가 n인 확률표본이라 할 때
    - 표본최빈값(Sample Mode)
      - 가장 많이 발생한 표본값
    - 예제 8.1
      - 다음과 같은 관측값이 주어진 경우 0.43이 가장 많이 관측됨으로 표본최빈값
        - 0.32, 0.53, 0.28, 0.37, 0.47, 0.43, 0.36, 0.42, 0.38, 0.43

대표적 통계량

- 산포 통계량
  - 산포(Variability)란 관측값이 평균을 중심으로 어떻게 분포되어있는지를 알 수 있는 통계량
  - $X_1, X_2, \ldots X_n$을 크기가 n인 확률표본이라 할 때
    - 표본분산(Sample Variance)

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

    - S^2이 크기 n인 확률표본의 분산 이면, 다음과 같이 표현 가능

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2\right]$$

대표적 통계량

- 산포 통계량
  - $X_1, X_2, \ldots X_n$을 크기가 n인 확률표본이라 할 때
    - 표본표준편차(Sample Standard Deviation)

$$S = \sqrt{S^2}$$

    - 표본범위(Sample Range)
      - X_i중 가장 큰 값을 X_{max}, 가장 작은 값을 X_{min}이라고 할 때

$$R = X_{max} - X_{min}$$

대표적 통계량

- 예제 8.2
  - 어느 지역의 상점 중 임의로 4곳을 선택하여 200g 짜리 병 커피의 가격을 비교하면, 지난 달보다 12, 15, 17, 20원 씩 인상되었을 때, 가격인상에 대한 확률표본의 분산은?
    - 표본평균을 계산하면

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = 16$$

    - 분산을 계산하면

$$s^2 = \frac{1}{3}\sum_{i=1}^{4}(x_i - 16)^2 = \frac{34}{3} = 11.\dot{3}$$

대표적 통계량

- 예제 8.3
  - 어느 호수에서 임의로 선정된 6명의 낚시꾼들에 의해 잡힌 송어의 수는 3,4,5,6,6,7 마리일 때, 표본 분산은?

$$\sum_{i=1}^{6} x_i^2 = 171, \sum_{i=1}^{6} x_i = 31, n = 6 \text{ 이므로}$$

$$s^2 = \frac{1}{(6)(5)}[(6)(171) - (31)^2] = \frac{13}{6} = 2.1\dot{6}$$

  - 따라서
    - 표본표준편차는 $s = \sqrt{\frac{13}{6}} \approx 1.47$
    - 표본범위는 $7 - 3 = 4$

표본분포와 표본평균의 분포

- 표본분포
  - 통계적 추론은 모집단의 모수에 대한 예측
    - 길거리에 통행하는 사람들에게 설문조사 하여, 그것을 근거로 지지율을 측정
    - 알려진 모집단 통계량(평균, 표준편차, …)과 측정된 표본 통계량을 이용하여 표본과 모집단이 같거나 다르다고 말할 수 있도록 계산
  - 정의
    - 표본분포(Sampling Distribution)는 통계량의 확률분포

표본분포와 표본평균의 분포

- 표본평균의 분포
  - 평균이 μ, 분산이 σ^2인 정규모집단에서 크기가 n인 확률표본을 추출했을 때, 확률표본 X_i, $i = 1,2,\dots,n$은 모두 모집단과 동일한 정규분포를 따름
    - 정리를 이용하여

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

    - 이 분포는 평균과 분산이 각각 아래와 같은 정규분포를 따름

$$\mu_{\bar{X}} = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu$$

$$\sigma_{\bar{X}}^2 = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}$$

표본분포와 표본평균의 분포

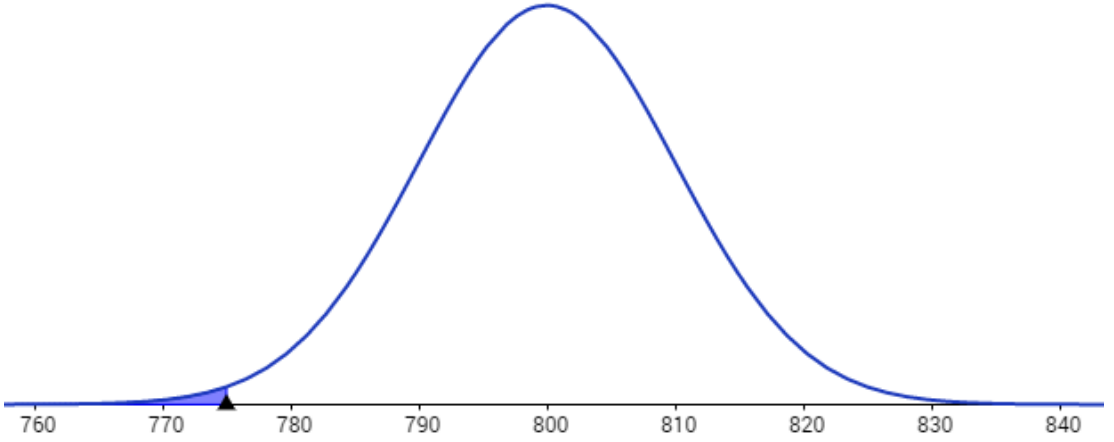
- 중심극한정리
  - 모집단의 분포가 알려져 있지 않은 경우에도 중심극한정리에 의해, 표본의 크기가 상당히 크면 표본평균의 분포는 근사적으로 평균과 분산이 각각 $\mu, \frac{\sigma^2}{n}$인 정규분포를 따름
    - 평균과 분산이 각각 μ, σ^2인 모집단으로부터 크기 n인 확률표본을 추출했을 때, 표본의 평균 $\bar{X}$에 대해 다음 식은 n이 상당히 클 때 표준정규분포 $n(z; 0,1)$에 근접

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

    - 일반적으로 n이 30이상인 경우 적합

표본분포와 표본평균의 분포

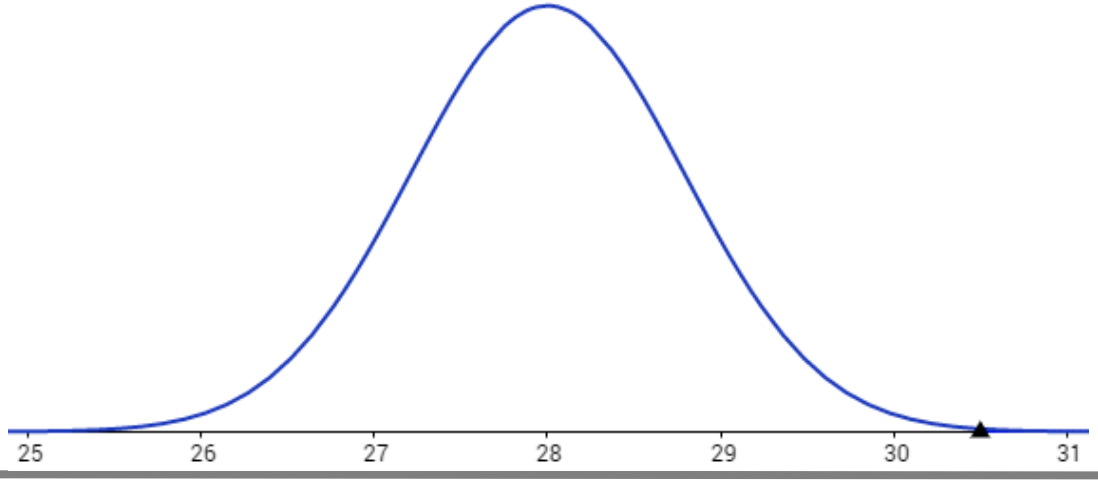
- 예제 8.4
  - 전구생산공장에서 생산되는 전구의 수명은 평균이 800시간이고, 표준편차가 40시간인 정규분포를 따를 때, 임의로 추출된 16개의 전구의 평균 수명이 775시간 미만일 확률
    - $\bar{X}$의 분포는 $\mu_{\bar{X}} = 800, \sigma_{\bar{X}} = \frac{40}{\sqrt{16}} = 10$인 정규분포, 구하고자 하는 확률 값은 아래 그림에 해당하는 영역



    - $\bar{x} = 775$에 대응하는 표준정규계수$z = \frac{775-800}{10} = -2.5$
    - 따라서 $P(\bar{X} < 775) = P(Z < -2.5) = 0.0062$

표본분포와 표본평균의 분포

- 예제 8.5
  - 어느 대학의 두 캠퍼스 사이를 운행하는 셔틀버스의 운행 시간은 평균 28분, 표준편차 5분의 분포를 따르게 운행하고 있으며, 어느 한 주에 평균 40번의 운행이 있을 때, 평균 운행시간이 30분보다 길 확률(평균시간은 분단위 반올림)
    - $\mu = 28, \sigma = 5, n = 40$ 일 때 $P(\bar{X} > 30)$을 구하면 되며, 분단위 반올림으로 측정하므로 $\bar{x} \geq 30.5$ 의 영역
      - $P(\bar{X} > 30) = P\left(\frac{\bar{X}-28}{5/\sqrt{40}} \geq \frac{30.5-28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008$



표본분포와 표본평균의 분포

- 두 표본평균 차이의 분포
  - 두 모집단이 서로 독립이고 각각의 평균이 μ_1, μ_2 분산이 σ_1^2, σ_2^2 일 때, 각 모집단으로부터 추출된 크기 n_1, n_2인 두 표본평균의 차이 $\bar{X}_1 - \bar{X}_2$의 분포는 아래와 같은 **정규분포에 근사**
    - $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$
    - $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  - $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$는 근사적으로 표준정규분포를 따름

표본분포와 표본평균의 분포

- 예제 8.6
  - A,B 두 회사에서 텔레비전 브라운관을 생산하고 있고, 아래 표를 따를 때, A회사 제품의 표본평균이 B회사 제품의 표본평균보다 적어도 1년 이상 길 확률

	평균	표준편차	표본 추출
A 회사	6.5년	0.9년	36
B 회사	6년	0.8년	49

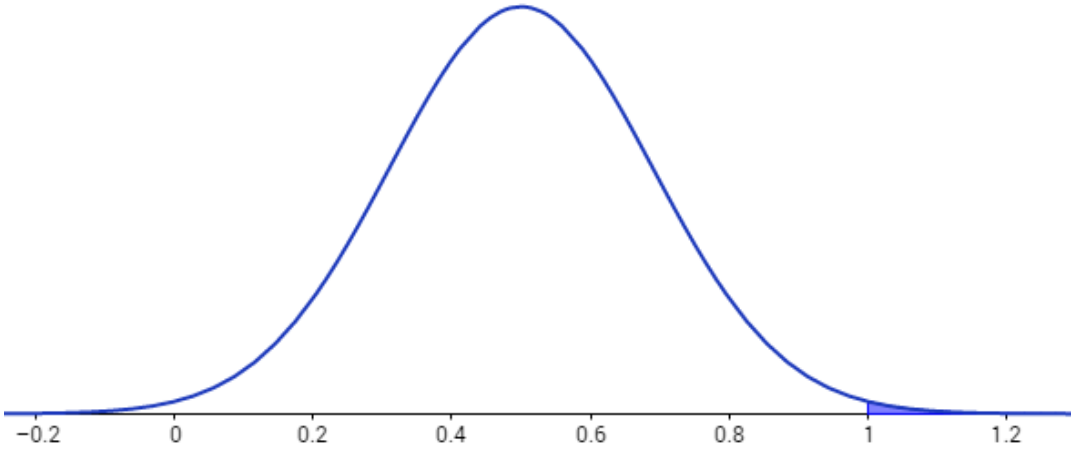
    - 다음과 같은 정규분포를 따름
      - $\mu_{\bar{X}_A - \bar{X}_B} = 6.5 - 6.0 = 0.5, \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189$
      - $\bar{x}_A - \bar{x}_B = 1.0$에 대응하는 $z = \frac{1.0 - 0.5}{0.189} = 2.65$
    - 따라서, 구하고자 하는 확률은
      - $P(\bar{X}_A - \bar{X}_A \geq 1.0) = P(Z > 2.65) = 1 - P(Z < 2.65) = 0.0040$

표본분포와 표본평균의 분포

- 예제 8.6
  - A,B 두 회사에서 텔레비전 브라운관을 생산하고 있고, 아래 표를 따를 때, A회사 제품의 표본평균이 B회사 제품의 표본평균보다 적어도 1년 이상 길 확률

	평균	표준편차	표본 추출
A 회사	6.5년	0.9년	36
B 회사	6년	0.8년	49

  - 표준정규분포의 영역



감사합니다!