5. 표집분포와 중심극한정리

<개요>

관심 대상의 전체집단에 대하여 그 특성이나 어떤 정보가 궁금

↓

But! 전체집단 모두를 조사하는 것은 불가능 OR 가능하다 하더라도 현실적으로 너무 많은 시간과 경비가 소요

↓

전체집단에서 일부를 추출하여 그 일부에서 얻는 정보로 전체집단의 특성을 알아보는 것이 현명. 단, 전체집단을 가능한 정확하게 대표할 수 있는 일부를 얻는 것이 매우 중요

모집단분포	표본분포	표집분포
모집단(population)을 구성하는 모든 요소들의 분포	모집단을 대표하기 위하여 추출된 표본(sample)의 분포	표집(sampling)으로부터 추리통계의 가설검정을 위한 이론적 분포
실재적으로 얻을 수 있는 분포		가상적 분포
모든 자료 수집	일부 표본 수집	·
⇓		·
모수치 - 모평균 μ - 모표준편차 σ	추정치 - 표본평균 $\overline{X}$ - 표본표준편차 s	중심극한 정리 - $\overline{X}_{\overline{X}}$ - 표준오차(SE)

5-1. 표집분포

5-1-1. 모집단분포

① 모집단(population) : 연구대상이 되는 사람 혹은 사물의 전체 집합

〚예〛 A지역 20세 성인들 전체 몸무게 측정

: 평균 μ, 표준편차 σ, 그래프 그리기

② 모집단의 속성은 평균이 μ이고 표준편차가 σ인 모수치(parameter)로 대표되며 이를 모집단분포(population distribution)라 한다.

③ 그러나 많은 경우에 모집단의 방대함과 역동성 때문에 모수치를 알기란 쉽지 않다.

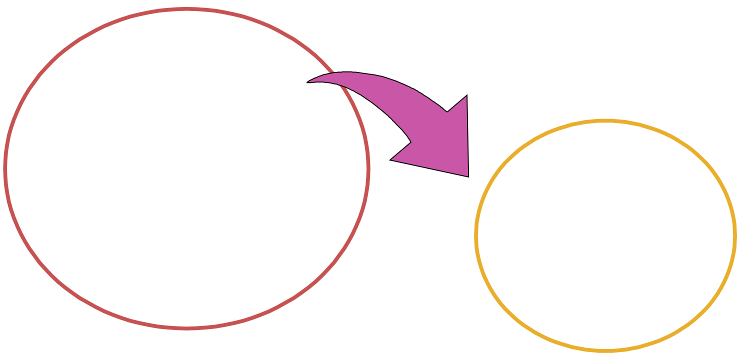
5-1-2. 표본분포

① 표본(sample) : 모집단의 속성을 알기 위하여 모집단으로부터 추출된 요소(표본의 속성으로 모집단의 속성을 추리)

〖예〗 A지역 20대 성인들 중 50명을 무작위 추출하여 몸무게 측정

② 모집단을 대표할 수 있게 추출된 표본의 평균 $\overline{X}$와 표준편차 s를 통계치(statistics) 혹은 추정치(estimates)라 하며, 이를 표본분포(sample distribution)라 한다.

③ 모집단분포는 일반적으로 정규분포라 하였으나 표본분포는 항상 정규분포는 아니다. 표본의 크기가 작으면 표집에 따라 정규분포 혹은 편포가 될 수 있다.



〖예〗 대학생의 영화 관람 횟수를 조사하기 위하여 표본으로 대학생 10명을 조사

1, 3, 1, 9, 0, 1, 2, 1, 5, 2

표본평균과 표본분산은?

[풀이]

표본평균 = $\dfrac{1+3+1+9+1+2+1+5+2}{10}=2.5$

표본분산 = $\dfrac{1+9+1+81+1+4+1+25+4}{10-1}-(2.5)^2=7.661$

【참고】 기술통계와 추측통계의 분산 계산, 표준편차의 편의추정량(biased estimator)과 불편추정량(unbiased estimator)

① 기술통계에서 모집단이나 표본의 표준편차를 계산할 시 분모를 n으로 한다.

$$V(X)=\frac{\sum(X_i-\overline{X})^2}{n}$$

② 그러나 추측통계를 위하여 사용되는 표본의 분산추정량은 분모를 $n-1$로 한다.

$$s^2(X)=\frac{\sum(X_i-\overline{X})^2}{n-1}$$

㉠ 왜? 분산 혹은 표준편차의 불편추정량을 얻기 위해서이다.

(참고) 불편추정량(unbiased estimator)이란 편의가 없는 추정량을 의미하며, 그 추정치의 기댓값이 모수의 참값과 일치하는 것이다.

㉡ 표본의 분산 계산에서 총 사례수 n으로 나누어 계산하면 모집단의 분산 결과보다 항상 작게 나온다. 즉, 표본의 분산이 모집단의 분산을 정확하게 추정하지 못하며 n으로 나누면 항상 과소 추정된다. 따라서 분모를 $n-1$로 해야 정확한 모집단의 분산을 추정할 수 있다.

㉢ 모집단을 추정하기 위하여 계산된 분산의 공식 s_X^2의 기댓값은 모집단의 분산 σ_X^2와 같은 것이 아니라 항상 작은 값을 추정하게 된다. 그러므로 모집단의 분산을 추정하기 위해서는 $E(s_X^2)$으로 $E\left(\frac{\sum(X_i-\overline{X})^2}{n}\right)$을 사용하여서는 안 되고 $E\left(\frac{\sum(X_i-\overline{X})^2}{n-1}\right)$을 사용하여야 정확한 모집단의 분산을 추정할 수 있다.

$$E(s^2)=E\left[\frac{\sum(X_i-\overline{X})^2}{n-1}\right]$$

$$=\frac{1}{n-1}E\left[\sum(X_i-\overline{X})^2\right]=\frac{1}{n-1}E\left[\sum(X_i-\mu+\mu-\overline{X})^2\right]$$

$$=\frac{1}{n-1}E\left[\sum\left((X_i-\mu)-(\overline{X}-\mu)\right)^2\right]$$

$$=\frac{1}{n-1}E\left[\sum\left((X_i-\mu)^2-2(X_i-\mu)(\overline{X}-\mu)+(\overline{X}-\mu)^2\right)\right]$$

$$=\frac{1}{n-1}E\left[\sum(X_i-\mu)^2-2(\overline{X}-\mu)\sum(X_i-\mu)+\sum\left(\overline{X}-\mu\right)^2\right)\Big]$$

$$\left(\sum(X_i-\mu)=\sum X_i-\sum\mu=n\overline{X}-n\mu=n(\overline{X}-\mu)\text{이므로}\right)$$

$$=\frac{1}{n-1}E\left(\sum(X_i-\mu)^2-2n(\overline{X}-\mu)^2+n(\overline{X}-\mu)^2\right)$$

$$=\frac{1}{n-1}E\left[\sum(X_i-\mu)^2-n(\overline{X}-\mu)^2\right]$$

$$=\frac{1}{n-1}\left[\sum E(X_i-\mu)^2-nE(\overline{X}-\mu)^2\right]$$

$$\left(E(X_i-\mu)^2=\sigma^2,\ E(\overline{X}-\mu)^2=\sigma_{\overline{X}}^2=\frac{\sigma^2}{n}\text{이므로}\right)$$

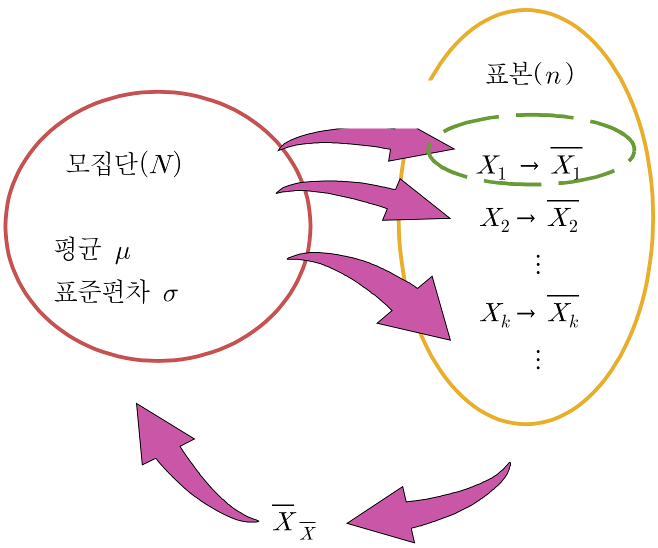
$$=\frac{1}{n-1}\left(\sum\sigma^2-n\frac{\sigma^2}{n}\right)=\frac{1}{n-1}(n\sigma^2-\sigma^2)=\frac{1}{n-1}(n-1)\sigma^2=\sigma^2$$

5-1-3. 표집분포

① 표집(sampling) : 표본의 크기가 n인 표본을 무한번 반복 추출하는 것

〖예〗 A지역 20대 성인들의 평균 몸무게를 알고 싶다.

② 표본의 크기가 n인 표본을 무한번 반복추출한 후, 무한개의 표본들의 평균(추정치)들을 가지고 그린 분포를 표집분포(sampling distribution)라 한다.



㉠ 표집분포의 평균 $\overline{X}_{\overline{X}}$ = 표본분포 평균들의 평균

㉡ 표집분포의 표준편차 $\sigma_{\overline{X}}$ =표본분포 평균들의 표준편차

㉢ 추리통계의 의사결정을 위한 이론적 분포

: 표집분포는 추리통계의 가설검정을 위한 판단의 기준을 제시하는 기각역과 채택역을 나타내어 준다.

㉣ 어떤 가정을 전제로 하여 이론적으로 그리는 이론적 분포(theoretical distribution)

: 표집분포는 모집단의 분포가 정규분포가 아니더라도 정규분포의 형태를 나타낸다.

5-2. 중심극한정리

5-2-1. 중심극한정리의 정의

① 중심 극한 정리(中心 極限 定理, central limit theorem, 약자 CLT)

② 수학자 피에르시몽 라플라스가 1774년~1786년 사이의 일련의 논문에서 발견 및 증명을 시도하였다.

③ 모집단의 분포가 어떤 형태이든지 간에 표집을 거의 무한에 가깝게 반복하면(표본의 크기가 커질수록) 표본 평균의 분포가 정규분포에 가까워진다는 내용의 정리이다.

④ 표본평균의 분포에서 평균은 모집단의 평균과 같고, 표준편차는 모집단의 표준편차를 표본 크기의 제곱근으로 나눈 것과 같다.

$$\overline{X}_{\overline{X}} = \mu_X, \quad {\sigma_{\overline{X}}}^2 = \frac{{\sigma_X}^2}{n}, \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

$$\overline{X} \sim N\left(\mu_X, \frac{{\sigma_X}^2}{n}\right)$$

㉠ 표집분포의 평균 = 모집단의 평균

㉡ 표집분포의 분산 = 모집단의 분산을 표본의 크기로 나눈 것

㉢ 표본의 크기가 충분히 클 때($n > 30$) 모집단의 분포와 상관없이 정규분포가 됨

⑤ 표집분포의 평균이 모집단의 평균과 같음을 보이면

$$\begin{aligned}
\overline{X}_{\overline{X}} &= E(\overline{X}) = E\left(\frac{\sum X_i}{n}\right) \\
&= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= E\left(\frac{1}{n}X_1\right) + E\left(\frac{1}{n}X_2\right) + \cdots + E\left(\frac{1}{n}X_n\right) \\
&= \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n) \\
&= \frac{1}{n}\mu_X + \frac{1}{n}\mu_X + \cdots + \frac{1}{n}\mu_X = \mu_X
\end{aligned}$$

⑥ 표집분포의 분산이 모집단의 분산을 표본의 크기로 나눈 것과 같음을 보이면

$$\begin{aligned}
{\sigma_{\overline{X}}}^2 &= Var(\overline{X}) = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\
&= Var\left(\frac{1}{n}X_1\right) + Var\left(\frac{1}{n}X_2\right) + \cdots + Var\left(\frac{1}{n}X_n\right) \\
&= \frac{1}{n^2}Var(X_1) + \frac{1}{n^2}Var(X_2) + \cdots + \frac{1}{n^2}Var(X_n) \\
&= \frac{1}{n^2}{\sigma_X}^2 + \frac{1}{n^2}{\sigma_X}^2 + \cdots + \frac{1}{n^2}{\sigma_X}^2 = \frac{1}{n}{\sigma_X}^2
\end{aligned}$$

⑦ 표집분포의 표준오차는 분산의 제곱근이므로

$$SE = \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

〚예〛 치매가 발병했을 때부터 사망까지의 지속시간은 3년에서 20년에 이르며, 평균은 8년이고 표준편차는 4년이라 한다. A 병원의 관계자는 병원의 자료로부터 30명의 치매환자를 임의로 선택하고 평균 생존시간을 측정하였다.

① 평균지속시간이 7년보다 작을 확률은?

② 평균지속시간이 7년보다 클 확률은?

③ 평균지속시간이 모평균 8년에서 1년 이내에 있을 확률은?

[풀이]

모집단 분포의 모양에 관계없이 표집분포는 평균 $\mu=8$, 표준편차 $\sigma/\sqrt{n}=4/\sqrt{30}=0.73$을 가지며, 표본의 크기가 30이므로 중심극한정리에 의해 근사적으로 정규분포를 가진다.

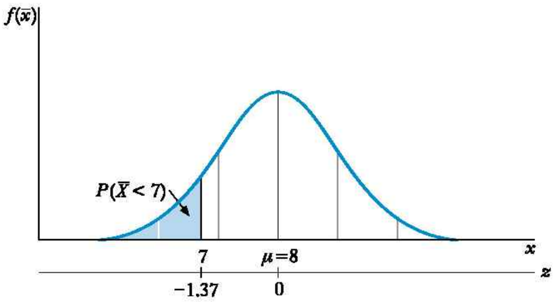
㉠ $\overline{X}=7$에 대응되는 Z값을 계산

$$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}=\frac{7-8}{0.73}=-1.37$$

그리고 $Z=-1.37$에 대응되는 누적면적을 구하면

$$P(\overline{X}<7)=P(Z<-1.37)=0.0853$$

즉, 8.53% 정도가 7년 정도 생존한다.



㉡ 7년 이상 생존할 확률은

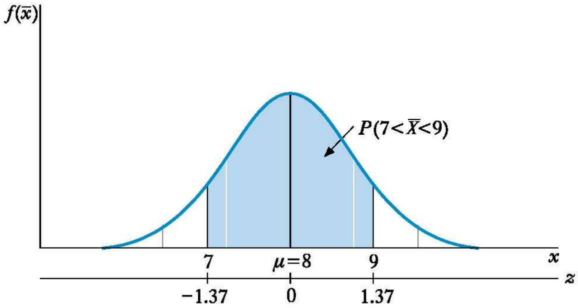
$$P(\overline{X}>7)=1-P(Z<-1.37)=1-0.0853=0.9147$$

즉, 91.47%이다.

㉢ 8년에서 1년 이내에 있을 확률은 7년부터 9년까지 생존할 확률이다[그림2]). 따라서

$$P(7<\overline{X}<9)=P(-1.37<Z<1.37)=0.9147-0.0853=0.8294$$

즉, 82.94%이다.



5-2-2. 중심극한정리의 역할

통계적 유의성 검정을 위한 이론적 토대가 된다.

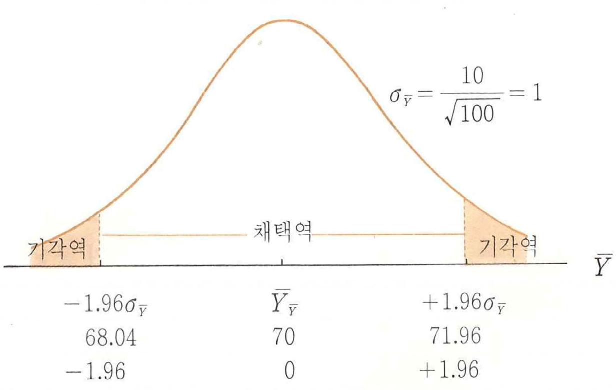
채집한 표본의 평균값이 어떤 특정한 값에 비해 통계적으로 유의한 정도로 더 큰지 혹은 더 작은지를 검토한다고 할 때, 표집분포가 대략 정규분포를 이룬다는 전제(=중심극한정리)가 있기 때문에 채집한 표본의 값이 이론적으로 전개된 표집분포에 비추어 봤을 때 나올 확률이 5% (통상적으로 상정되는 유의기준) 미만인지를 검토할 수 있다.

〚예〛 30세 성인 남자 100명을 무작위 추출하고 체중을 측정하여 30세 성인 남자의 체중을 추정하려 한다. 즉, 연구자는 성인 30세 남자의 체중이 70kg, 표준편차 10이라는 잠정적 진술을 유의수준 0.05에서 검정하려한다.

[설명]

㉠ 귀무가설 : 성인 30세 남자의 체중은 70kg이다.

㉡ 연구가설 : 성인 30세 남자의 체중은 70kg이 아니다.



㉢ 채택역 : $70 - 1.96 <$ 모집단의 체중 $< 70 + 1.96$

㉣ 기각역 : 채택역 외

㉤ 표본의 평균값이 70kg에 가까우면 성인의 체중이 70kg이라는 귀무가설을 수용 또는 70kg보다 멀리 떨어져 있다면 귀무가설을 기각

㉥ 수용하거나 기각하는 기준이 표집분포에 그려지고 유의수준 0.05에서 얻어진 표본의 평균이 $70 \pm 1.96\sigma_{\overline{Y}}$ 사이에 위치하면 귀무가설을 수용하고 그렇지 않으면 귀무가설을 기각해야 한다.

㉦ 체중의 평균이 65kg이 나왔다면, 성인남자의 체중이 70kg이라 가정하였을 때의 기각역에 65kg이 위치하므로 한국 30세 남성의 체중이 70kg이라는 가설을 수용할 수 없다고 판단할 수 있다.