11장 추정하기 (2부)

1. 주요 내용

① 점추정 : 표본을 이용해서 모집단에 대한 정보를 추정하기

표본 평균, 표본 분산 } 계산 $\xrightarrow{\text{점추정}}$ 추정 { 모집단 평균, 모집단 분산

② 비율의 표본분포 ← 이번에 다룰 내용

③ 평균의 표본분포

④ 중심극한 정리 (Central Limit Theorem)

2. 주요예제

모집단 : "달콤한 풍선껌" 회사에서 생산하는 "완전-오래가는-향" 풍선껌 전체

표 본 : 편향되지 않은 표본이 구성되어 있다고 가정

연습문제 : "달콤한 풍선껌" 모집단 전체에 대한 정보 구하기

① 향이 지속되는 시간의 평균과 분산

② 타 회사 제품보다 "달콤한 회사" 제품을 선호하는 사람들의 비율

3. 모집단의 ~~평균 및 분산~~ 비율 추정하기

p.494

① 점추정 : 표본을 통해 구한 정보(평균, 분산, 비율 등)를 이용하여 모집단에 대한 정보를 구하는 방법

⇒ 점추정을 이용하여 구한 모집단에 대한 정보(평균, 분산, 비율 등)가 정확하다고 장담할 수는 없다. 하지만 그렇게 하는 것이 최선이다.

*주의: 편향되지 않은 표본이 구성되도록 최선을 다해야 한다.

② 모집단 ~~평균값의~~ 비율의 점추정

i) 표본이 갖는 성공률 구한다:

$$P_s = \frac{\text{성공 횟수}}{\text{표본 크기}}$$

ii) 모집단 성공률의 추정치를 설정한다:

$$\hat{p} = P_s$$

즉, 표본의 성공률을 모집단 성공률의 추정치로 사용.

p.495

③ 확률과 비율의 관계

임의로 선택된 사람에게 대통령에 대한 지지여부를 물어서 긍정적인 답을 얻으면 성공이라고 하였을 때, 모집단에서의 성공의 비율은 대통령의 지지율과 동일한 값이다.
이와 같이, 성공할 확률을 계산하는 것과 성공의 비율을 계산하는 과정은 완전히 동일하다.

P.495

예제: "달콤한 풍선껌" 회사에서 개발한 껌에 대한 선호도를 조사하였다.

40명을 임의로 선택하여 물었더니 그중 10명이 분홍색 풍선껌을 선호하였다.

질문: ① 모집단에서 분홍색 풍선껌을 선호하는 사람의 비율은 얼마인가?

즉, 사람을 임의로 골랐을 때 그 사람이 분홍색 풍선껌을 선호할 확률은 얼마인가?

② 모집단에서 분홍색 풍선껌을 좋아하지 않는 사람을 선택할 확률은 얼마인가?

답안: ① $\hat{p} = p_s = \frac{10}{40} = 0.25$

② P(분홍색을 선호하지 않는 사람)

$= 1 - \hat{p} = 1 - 0.25 = 0.75$

(추가내용)

과 제

1) 과제 안내

11장~12장 내용을 이해하기 위해 다음의 모의실험을 각자 집에서 실행할 것.

직접 모의실험을 해 본 경우 강의내용을 보다 쉽게 이해 가능.

2) 모의 실험 내용

(i) 하나의 주사위가 주어졌다고 가정하자.

(ii) 주어진 주사위를 30번 던졌더니 숫자 1이 10번 나왔다.
즉, 숫자 1이 나온 비율이 $\frac{1}{3}$ 이었다.

3) 질문

(i) 주어진 주사위가 정상적인 주사위인가?

(ii) 질문 (i)에 대한 본인의 답에 대한 근거를 설명하라.

(추가내용)

4) 힌트

(i) 가정 : 주어진 주사위는 정상이다. 즉, "6번에 1번의 비율로 숫자 1이 나온다"고 가정한다.

그러면 주사를 30번 던져서 숫자 1이 나오는 횟수 X는 이항분포를 따른다.

$$X \sim B(30, \tfrac{1}{6})$$

주사위 30번 던지기

주사위를 한 번 던졌을 때 숫자 1이 나올 확률

따라서

(ii) 주사위를 30번 던졌을 때 숫자 1이 나온 횟수의 비율을 나타내는 분포는 $\frac{X}{30}$ 이다. 즉, P_s는 아래의 분포를 따른다.

$$P_s = \frac{X}{30}$$

(iii) 이제 P_s의 기대치, 즉 평균값과 분산을 구할 수 있다.

$$E(P_s) = E\left(\frac{X}{30}\right) = \frac{1}{30}E(X)$$
$$= \frac{1}{30}E(X) = \frac{1}{30}\cdot 30 \cdot \frac{1}{6}$$
$$= \frac{1}{6} = 0.1667$$

$$Var(P_s) = Var\left(\frac{X}{30}\right)$$
$$= \frac{1}{30^2}\cdot Var(X) = \frac{1}{30^2}\cdot 30 \cdot \frac{1}{6}\cdot\frac{5}{6}$$
$$= \frac{1}{6^3} = 0.0046 = (0.068)^2$$

(추가내용)

(iv) 또한 p_s 는 정규분포를 따른다. 즉,

$$p_s \sim N(0.1667, (0.068)^2)$$ ← 중심극한정리에 의해!

(v) 이제 정상적인 주사위를 던져서 숫자 1이 나오는 비율이 $\frac{1}{3}$ 이상일 확률을 계산할 수 있다.

연속성보정 (뒤에 설명됨)

$$p(p_s \geq 0.3333) = p\left(p_s > 0.333 - \frac{1}{2*30}\right)$$
$$= p(p_s > 0.3167)$$
$$= p\left(z > \frac{0.3167 - 0.1667}{0.068}\right)$$
$$= p(z > 2.21)$$
$$= 1 - p(z \leq 2.21)$$
$$= 1 - 0.9864$$
$$= 0.014 \quad (\text{즉, } 1.4\%)$$

5) 결론

주어진 주사위가 정상적이다라는 가정하에서, 주사위를 30번 던져서 숫자 1이 나오는 비율이 $\frac{1}{3}$ 이상일 확률은 1.4% 에 불과하다. 어떤 가정하에서 발생할 확률이 5% 미만인 사건이 실제로 발생하였다면 해당 가정이 잘못되었다고 결론 내린다.

95% 확률로

즉, "주어진 주사위가 정상이다"라는 가정을 받아들이지 않는 게 일반적이다.

4. 표본분포

① 통계량 : 표본으로부터 얻은 표본평균, 표본분산, 표본비율과 같은 통계적인 양

② 표본분포 : 모집단에서 일정한 크기의 표본을 반복적으로 선정하여 얻은 통계량의 확률분포

예) • 평균의 표본분포

$\boxed{\text{• 비율의 표본분포}}$

• 분산의 표본분포 (여기서는 다루지 않음)

5. ~~평균의~~ 비율의 표본분포 예제

p.500

예제

• 풍선껌을 100개씩 상자에 담아 판매

• 모집단에서 빨간색 풍선껌의 비율은 $p = 0.25$로 주어짐.

질문 : 풍선껌 100개 들이 한 상자에 빨간색 풍선껌의 비율이 40% 이상일 확률은?, 즉,

$P(P_s \geq 0.4) = ?$

예제의 질문에 답하기 위해서는 아래 과정을 확인해야 한다:

풍선껌 100개들이 한 상자를 무작위로 선택하였을 때 빨간색 풍선껌의 비율에 대한 표본분포를 확인한다.

예제 답안:

X를 100개 들이 한 상자에 들어 있는 빨간색 풍선껌의 개수라 하자. 그러면 X는 이항분포를 따른다.

$$X \sim B(100, 0.25)$$

따라서 100개 들이 한 상자에 들어 있는 빨간색 풍선껌의 비율의 분포를 P_s 라 하면 다음이 성립한다.

$$P_s \sim \frac{X}{100}$$

P_s 의 평균값과 분산을 다음과 같이 구한다.

$E(P_s) = E\left(\frac{X}{100}\right)$ $= \frac{1}{100}E(X)$ $= \frac{1}{\cancel{100}} \cdot \cancel{100} \cdot \frac{1}{4}$ $= 0.25$	$Var(P_s) = Var\left(\frac{X}{100}\right)$ $= \frac{1}{100^2} \cdot 100 \cdot \frac{1}{4} \cdot \frac{3}{4}$ $= \frac{3}{1600}$ $= 0.001875$

이제 $P(P_s \geq 0.4)$를 구하고자 한다.
그런데 P_s는 어떤 분포를 따르는가?

⇒ 정규분포를 따른다.

즉, $P_s \sim N(0.35, 0.001875)$가 성립한다.

이유: 중심극한정리

연속성 보정 필요!

따라서 $P(P_s \geq 0.4) = P\left(P_s > 0.4 - \frac{1}{2 \times 100}\right)$
$= P(P_s > 0.395)$
$= P\left(Z > \frac{0.395 - 0.35}{\sqrt{0.001875}}\right)$
$= P(Z > 3.35)$
$= 1 - P(Z \leq 3.35)$
$= 1 - 0.9996$
$= 0.0004$

0.04%. 사실상 가능성 없음!

* 연속성 보정: 표본의 크기 n이 1,000 이하일 경우
연속성 보정 필요. $\left(\pm \frac{1}{2n}\right)$

6. ~~평균의~~ 비율의 표본분포 (일반화)

p.500~503

모집단에서 특정 사건이 발생할 확률이 p라 하자.
그러면 n개의 표본에서 해당 사건이 발생하는 횟수 X는 이항분포를 따른다. 즉,

$$X \sim B(n, p)$$

따라서, n개의 표본에서 해당 사건이 발생할 확률 P_s의 분포는 $P_s = \frac{X}{n}$ 이다.

P_s의 평균값과 분산은 다음과 같다.

$$E(P_s) = E\left(\frac{X}{n}\right)$$
$$= \frac{1}{n} \cdot E(X) = \frac{\cancel{n} \cdot p}{\cancel{n}} = p$$

$$Var(P_s) = Var\left(\frac{X}{n}\right)$$
$$= \frac{1}{n^2} \cdot Var(X) = \frac{1}{n^2} \cdot n \cdot p \cdot q$$
$$= \frac{p \cdot q}{n} \qquad (q = 1 - p)$$

p.504~505

또한 $n \geq 30$ 일때 P_s는 중심극한정리에 의해 정규분포를 따른다. 즉,

$$P_s \sim N\left(p, \frac{p \cdot q}{n}\right)$$

7. P_s의 연속성 보정

$$\begin{cases} P_s = \dfrac{X}{n}, \\ X = \text{표본이 갖는 성공의 수} \end{cases}$$

$\Rightarrow$ 연속성보정 값 $= \pm \dfrac{1}{2n}$

즉, P_s를 위한 확률을 구하기 위해 정규분포를 이용한 근사치를 구하고자 할 때 $\pm\frac{1}{2n}$ 이라는 연속성보정을 적용해야 한다.

예) $P(P_s \geq 0.40) = P(P_s > 0.40 - \frac{1}{2*100})$

하지만 n가 1,000 인 경우 연속성보정 불필요!