11장 추정하기 (1부)

주의사항 : • 11장과 12장 내용을 1, 2부로 나누어 다룸.
• 1부 주요 내용 : 평균의 표본분포와 신뢰구간
• 2부 주요 내용 : 비율의 표본분포와 신뢰구간

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 (2부에서 다룸)

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

2부에서 다룸 → ② 타 회사 제품보다 "달콤한 회사" 제품을 선호하는 사람들의 비율

3. 모집단의 평균 및 분산 추정하기

p.483 ① 점추정 : 표본을 통해 구한 정보(평균, 분산 등)를 이용하여 모집단에 대한 정보를 구하는 방법

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

*주의 : 편향되지 않은 표본이 구성되도록 최선을 다해야 한다.

② 모집단 평균값의 점추정

p.484~485 i) 표본의 평균값 $\bar{x}$를 구한다 :

$$\bar{x} = \frac{\sum x}{n}$$

← 표본에 속한 모든 데이터 더하기

← 표본의 크기

ii) 모집단 평균값의 추정치를 설정한다 :

$$\hat{\mu} = \bar{x}$$

즉, 표본의 평균은 모집단 평균의 추정치로 사용.

* 주의 : 모집단의 진짜 평균은 μ로 표시한다. 하지만 μ는 일반적으로 거의 알 수가 없으며 평균값의 추정치인 $\hat{\mu}$를 μ 대신에 사용한다.

⟹ 보다 정확한 $\hat{\mu}$를 얻기 위해 편향되지 않으면서 "적당한" 크기의 표본을 구성하는 일이 매우 중요하다.

③ 모집단 분산의 점추정

p.488 ~490

평균값의 점추정과는 달리 모집단 분산의 점추정은 표본의 분산과 다르게 계산한다.

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

표본의 평균값

모집단 분산의 점추정 값으로 사용하는 값

표본의 크기

n 대신에 n−1 사용!

이유 : 모집단의 분산은 표본의 분산에 비해 좀 더 큰 값을 가짐.

⇒ 모집단 분산의 점추정 : $\boxed{\hat{\sigma}^2 = s^2}$

예제 : "달콤한 풍선껌" 회사의 제품의 표본에 속한 껌들의 맛의 지속시간이 다음과 같다. (단위 분)

p.485, p.491

61.9	62.6	63.3	64.8	65.1
66.4	67.1	67.2	68.7	69.9

$$\Rightarrow \hat{\mu} = \bar{x} = \frac{61.9 + 62.6 + \cdots + 69.9}{10} = 65.7$$

$$\hat{\sigma}^2 = s^2 = \frac{(61.9 - 65.7)^2 + \cdots + (69.9 - 65.7)^2}{9} = 6.92$$

(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) "한 개의 주사위를 6번 던지기"를 30회 반복하기

(ii) 주사위를 6번 던질 때마다 숫자 1이 나온 횟수를 아래 표에 작성하기

1회	2회	3회	...	29회	30회	평균

각각의 칸에는 주사위를 6번 던졌을 때 숫자 1이 나온 횟수 입력

숫자 1이 나온 횟수의 평균 입력

3) 질문

(i) 본인이 구한 평균값이 적절하다고 생각되는가?

(ii) 질문 (i)에 대한 본인의 답에 대한 근거를 설명하라.

(추가내용)

4) 힌트

(i) 먼저 주사를 6번 던져서 숫자 1이 나오는 횟수 X는 이항분포를 따른다.

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

주사위 6번 던지기

주사위를 한 번 던졌을 때 숫자 1이 나올 확률

(ii) "주사위 6번 던지기"를 30회 반복하여 숫자 1이 나온 횟수의 평균 $\bar{X}$는 아래 공식을 만족 시킨다.

$$\bar{X} = \frac{X_1 + \cdots + X_n}{30}$$

단, X_i 는 X의 독립관측이다.

"주사위 6번 던지기" 반복하는 행위는 서로 독립이며 동일한 확률분포를 따른다.

(iii) 이제 $\bar{X}$의 기대치, 즉 평균값과 분산을 구할 수 있다.

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + \cdots + X_{30}}{30}\right) \\ &= \frac{1}{30} \cdot E(X_1 + \cdots + X_{30}) \\ &= \frac{1}{30} \cdot 30 \cdot E(X) \\ &= E(X) \end{aligned}$$

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{X_1 + \cdots + X_{30}}{30}\right) \\ &= \frac{1}{30^2} Var(X_1 + \cdots + X_{30}) \\ &= \frac{1}{30^2} \cdot 30 \cdot Var(X) \\ &= \frac{Var(X)}{30} \end{aligned}$$

(추가내용)

(iv) 또한 $\bar{X}$ 는 정규분포를 따른다. 즉,

$\bar{X} \sim N(1, \frac{1}{6^2})$ 중심극한정리에 의해!

이유: $X \sim B(6, \frac{1}{6})$

$$\Rightarrow \begin{cases} E(X) = 6 \cdot \frac{1}{6} = 1 \\ Var(X) = 6 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{6} \end{cases}$$

(v) 결론

각자의 실험결과는 아래 구간에 속할 것이다.

① $[1-\frac{1}{6}, 1+\frac{1}{6}]$ 구간에 속할 확률 68.2 %

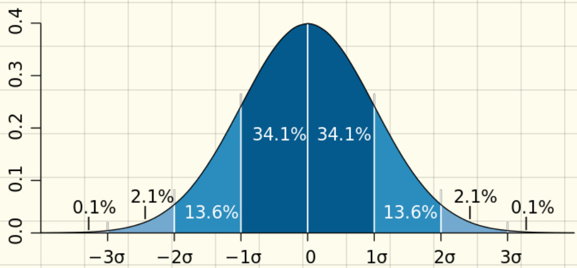
② $[1-\frac{1}{3}, 1+\frac{1}{3}]$ 구간에 속할 확률 95.4 %

③ $[1-\frac{1}{2}, 1+\frac{1}{2}]$ 구간에 속할 확률 99.7 %

5) 추가 질문

(v)의 결론을 어떻게 구했을까?

힌트: 아래 그림 참조



4. 표본분포

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

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

예) $\boxed{\text{• 평균의 표본분포}}$
• 비율의 표본분포 (2부에서 다룸)
• 분산의 표본분포 (여기서는 다루지 않음)

5. 평균의 표본분포 예제

p. 510

예제
- 풍선껌을 봉지에 담아 판매
- 한 봉지에 담긴 풍선껌 개수의 평균은 10, 분산은 1임.
- 한 상자에 30봉지씩 넣어 판매함.

질문 : ① 한 상자에 들어 있는 풍선껌 한 봉지에 담긴
풍선껌 개수의 평균값은 얼마인가?
② 한 상자에 들어 있는 풍선껌 한 봉지에 담긴
풍선껌 개수의 평균이 9.5개 이하일 확률은?

예제의 질문에 답하기 위해서는 아래 두 과정을 진행해야 한다.

① 30 개의 봉지를 무작위로 선택하여 하나의 봉지에 들어있는 풍선껌 개수의 평균을 구해보아야 한다.

② 과정 ① 을 여러 번 반복하여 풍선껌 개수의 평균의 분포를 알아 낸다.

일반적으로 30 회 이상.

주의사항 : 30개의 봉지의 선택은 매번 독립적으로 이루어져야 한다. 즉, 한 봉지의 선택은 무작위적으로 이루어지며 한 번 선택된 봉지를 다시 원래대로 돌려 넣는 **복원추출** 방식을 따른다.

예제 답안 :

X를 한 봉지에 들어 있는 풍선껌 개수의 분포를 나타낸다고 하자. 그러면 다음 사실이 성립한다.

① $E(X) = 10$

② $Var(X) = 1$

⇐ 하나의 봉지에 들어 있는 풍선껌의 개수에 대한 기대치 (평균값)와 분산

③ 봉지 30개 각각의 선택은 **독립관측**이다.

따라서 30 봉지에 들어있는 풍선껌 개수의 평균값의 분포를 $\bar{X}$ 라 하면 아래 공식이 성립한다.

① $$\bar{X} = \frac{X_1 + \cdots + X_{30}}{30}$$

② X_i 는 X 의 독립관측

X_i가 X의 독립관측이므로 $\bar{X}$의 평균값과 분산을 다음과 같이 구할 수 있다.

$$
\begin{aligned}
E(\bar{X}) &= E\left(\frac{X_1+\cdots+X_{30}}{30}\right) \\
&= E\left(\frac{1}{30}X_1+\cdots+\frac{1}{30}X_{30}\right) \\
&= \frac{1}{30}E(X_1)+\cdots+\frac{1}{30}E(X_{30}) \\
&= \frac{1}{30}\left(E(X_1)+\cdots+E(X_{30})\right) \\
&= \frac{1}{30}\cdot 30\cdot E(X) \\
&= E(X) = 10
\end{aligned}
$$

$$
\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{X_1+\cdots+X_{30}}{30}\right) \\
&= Var\left(\frac{1}{30}X_1+\cdots+\frac{1}{30}X_{30}\right) \\
&= \frac{1}{30^2}Var(X_1)+\cdots+\frac{1}{30^2}Var(X_{30}) \\
&= \frac{1}{30^2}\cdot 30\cdot Var(X) \\
&= \frac{Var(X)}{30} = \frac{1}{30}
\end{aligned}
$$

상식적으로 이해 가능.
분산의 경우 : 한 뭉치에 포함된 풍선껌의 개수가 늘어날 수록 분포가 비례하여 상쇄됨

이제 $P(\bar{X} < 8.5)$를 구하고자 한다.
그런데 $\bar{X}$는 어떤 분포를 따르는가?

⇒ 정규분포를 따른다.

즉, $\bar{X} \sim N(10, \frac{1}{30})$ 이 성립한다.

이유 : 중심극한정리 ← 뒤에서 설명!

따라서 $P(\bar{X} < 8.5) = P\left(Z < \frac{8.5-10}{\sqrt{0.0333}}\right)$

$= P(Z < -8.22)$

$= 0$

확률 테이블에서 확인 불가.
사실상 0임을 의미함.

(주의 : 연속성보정 필요 없음)
↑ 평균값들은 어떠한 값도 임의로 가질 수 있기 때문.

6. 평균의 표본분포 (일반화)

p.512~519

$E(X) = \mu$, $Var(X) = \sigma^2$ 을 따르는 X의 분포가 주어짐.
X를 n번 독립관측했을 때, 관측된 X들의 평균값을 가리키는 확률변수를 $\bar{X}$로 표기한다. 그러면 아래 식이 성립한다.

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

(X_i는 X의 독립관측)

또한 $\bar{X}$의 평균값과 분산은 다음과 같다.

$$E(\bar{X}) = E\left(\frac{X_1 + \cdots + X_n}{n}\right)$$
$$= \frac{1}{n} \cdot E(X_1 + \cdots + X_n) = \frac{1}{n} \cdot n \cdot E(X)$$
$$= E(X) = \mu$$

$$Var(\bar{X}) = Var\left(\frac{X_1 + \cdots + X_n}{n}\right)$$
$$= \frac{1}{n^2} \cdot Var(X_1 + \cdots + X_n) = \frac{1}{n^2} \cdot n \cdot Var(X)$$
$$= \frac{1}{n} \cdot Var(X) = \frac{\sigma^2}{n}$$

~ 은 일반적으로 n이 커질 수록 표준오차는 줄어든다.

p.520~521

특히 $n \geq 30$ 일때 $\bar{X}$는 중심극한정리에 의해 정규분포를 따른다. 즉,

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

(단, $E(X) = \mu$, $Var(X) = \sigma^2$)

주의 : X가 반드시 정규분포를 따를 필요는 없다.

7. 중심극한정리

p.521

① 서로 독립이며 동일한 분포를 따르는 n개의 확률변수 $X_1, X_2, \cdots, X_n$에 대해 n이 충분히 크면, 예를 들어 $n \geq 30$, 아래와 같이 정의된

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

$\bar{X}$의 분포가 개략적으로 정규분포를 따른다.

또한, $E(X) = \mu$ 이고 $Var(X) = \sigma^2$ 이고 X_i가 X와 동일한 확률분포를 가지면 다음이 성립한다.

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (\text{단}, n \geq 30)$$

② 확률 찾기

$\bar{X}$가 정규분포를 따르므로 확률테이블을 이용하여 확률값을 계산할 수 있다.

(i) $P(\bar{X} > a) = P\left(z > \frac{a-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$

(ii) $P(a < \bar{X} < b) = P\left(\frac{a-\mu}{\frac{\sigma}{\sqrt{n}}} < z < \frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right)$

(iii) $P(a < \bar{X}) = P\left(\frac{a-\mu}{\frac{\sigma}{\sqrt{n}}} < z\right)$

(주의: 연속성보정 필요 없음)

↑ 평균값들은 어떠한 값도 임의로 가질 수 있기 때문.

8. 중심극한 정리 활용

주의 : 책 설명과 약간 다름.
(책 설명을 일반화 시킨 내용임)

p.522

1) 이항분포

전제 : $X \sim B(n, p)$

$\Rightarrow \mu = n \cdot p$, $\sigma^2 = n \cdot p \cdot q$ $(q = 1 - p)$

이항분포를 따르는 X를 독립관측으로 n번 $(n \geq 30)$ 반복할 때 중심극한 정리에 의해 X의 평균값 $\bar{X}$는 정규분포를 따른다. 즉,

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

예제) 동전을 10번 던졌을 때 숫자가 나오는 횟수는 이항분포를 따른다. 즉, $X \sim B(10, \frac{1}{2})$.

주의: X 자체는 정규분포를 따르지 않는다..

이제, "동전 10번 던지기"를 30번 반복했을 때 숫자가 나오는 횟수의 평균은 다음의 분포를 따른다.

$$\bar{X} \sim N\left(\underset{10 \cdot \frac{1}{2}}{\underset{\uparrow}{5}}, \underset{\frac{10 \cdot \frac{1}{2} \cdot \frac{1}{2}}{30}}{\underset{\uparrow}{\frac{1}{12}}}\right)$$

예제) "동전 10번 던지기"를 30번 반복했을 때, 숫자가 평균 6번 이상 나올 확률은?

정답 : $\bar{X} \sim N(5, \frac{1}{12})$ 이므로

$$P(\bar{X} > 6) = P\left(Z > \frac{6-5}{\frac{1}{\sqrt{12}}}\right)$$

$$= P(Z > \sqrt{12}) = P(Z > \underline{3.46})$$

$$= 1 - P(Z \leq 3.46)$$

$$= 1 - 0.999730$$

$$\fallingdotseq 0.000270$$

← 매우 낮음.

* 주의사항 (책에 없는 내용임)

"동전 10번 던지기"를 30회 반복해서 얻은 통계량과 동전을 300번 던져서 얻는 통계량은 어떻게 다를까?

즉, 아래 두 질문의 결과를 비교해 보자.

① "동전 10번 던지기"를 30번 반복했을 때, 숫자가 평균 6번 이상 나올 확률은?

② 동전을 300번 던져서 숫자가 180번 이상 나올 확률은?

⇒ ① 의 답 : 0.00027 (앞서 구했음)

② 의 답은 아래와 같이 구할 수 있다.

X가 동전을 300번 던져서 숫자가 나오는 횟수를 가리킨다면 $X \sim B(300, \frac{1}{2})$ 이다.

그리고 $300 * \frac{1}{2} = 150 > 5$ 이므로 X는 정규 분포를 따른다. 즉,

$$X \sim N(150, 75).$$

따라서

$$P(X > 180) = P\left(Z > \frac{180.5 - 150}{\sqrt{75}}\right)$$
$$= P(Z > 3.52)$$
$$= 1 - P(Z < 3.52)$$
$$= 1 - 0.999784$$
$$\simeq 0.000216$$

0.00027 과 비교하여 오차는 0.000054, 즉, 0.0054 % 임. 확률로는 어떤 의미도 없는 오차임.

2) 푸아송 분포 (다루지 않음)

p.522

전제: $X \sim Po(\lambda)$

$\Rightarrow \mu = \lambda \; , \; \sigma^2 = \lambda$

푸아송분포를 따르는 X를 독립관측으로 n번 반복할 때 ($n \geq 30$)
중심극한 정리에 의해 X의 평균값 $\bar{X}$는 정규분포를 따른다. 즉,

$$\bar{X} \sim N\left(\lambda, \frac{\lambda}{n}\right)$$

예제) 7장에서 살펴본 영화관의 팝콘 기계의 주당 평균 고장 횟수는 푸아송 분포를 따르며 $X \sim Po(3.4)$ 이다.

주의: X 자체는 정규분포를 따르지 않는다..

이제, 30주 동안 주당 고장 횟수의 평균은 다음의 정규분포를 따른다:

$$\bar{X} \sim N(3.4, \underbrace{0.11}_{\frac{3.4}{30}})$$

예제) 임의로 지정된 30주 동안 주당 고장횟수의 평균이 2회 이하일 확률은?

정답: $\bar{X} \sim N(3.4, 0.11)$ 이므로

$$P(\bar{X} < 2) = P\left(Z < \frac{2-3.4}{\sqrt{0.113}}\right)$$

$$= P\left(Z < \frac{-1.4}{0.337}\right)$$

$$= P(Z < -4.15)$$

$$= 0$$

거의 0임.