

확률과 통계

- 추정하기 및 신뢰구간 I -

참고서 : Head First Statistics

(11~12장 : 추정하기 및 신뢰구간)

작성자 : 이계식

환경대학교, 컴퓨터공학과



추정하기

1. 주요 내용

① 절측정 : 표본을 이용해서 모집단에 대한 정보를 추정하기

$$\left. \begin{array}{l} \text{표본 평균} \\ \text{표본 통산} \end{array} \right\} \text{계산} \xrightarrow{\text{절측정}} \text{추정} \left\{ \begin{array}{l} \text{모집단 평균} \\ \text{모집단 통산} \end{array} \right.$$

② 비율의 표본분포

③ 표본평균의 분포

④ 중심극한 정리 (Central Limit Theorem)

2. 주요 예제

모집단 = "달콤한 통선껌" 회사에서 생산하는 "완전一樣" 풍선껌 전체

표본 = 원형되지 않은 표본이 구성되어 있다면 가능

연습문제 = "달콤한 통선껌" 모집단 전체에 대한 정보 구하기

① 양이 지속되는 시간의 평균과 통산

② 타 회사 제품과 "달콤한 회사" 제품을 선호하는 사람들의 비율

3. 모집단의 평균 및 분산 추정하기

p. 483

① 절추점 = 표본을 통해 구한 정보(평균, 분산 등)를 이용하여 모집단에 대한 정보를 구하는 방법

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

* 주의: 편향되지 않은 표본이 구성되도록 치선을 다해야 한다.

② 모집단 평균값의 절추점

p. 484~485

i) 표본의 평균값 $\bar{x} \approx$ 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 평균을 모집단 평균의 추정치로 사용.

* 주의: 모집단의 진짜 평균은 서로 표시된다.

하지만 서는 일반적으로 거의 알 수가 없으며 평균값의 추정치인 $\hat{\mu}$ 를 서 대신에 사용한다.

⇒ 보다 정확한 $\hat{\mu}$ 를 얻기 위해 편향되지 않은면서 "적합한" 크기의 표본을 구성하는 일이 매우 중요하다.

③ 모집단 분산의 절추정

P. 488
~ 490

평균값의 절추정과는 달리 모집단 분산의 절추정은

표본의 분산과 다르게 계산한다.

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

모집단 분산의
절추정 값으로
사용하는 경우
표본의 크기
n 대신에
n-1 사용!

표본의
평균값

n 대신에
n-1 사용!

↑ 이유: 모집단의 분산은 표본의 분산에
비해 좀 더 큰 값을 가짐.

\Rightarrow 모집단 분산의 절추정 : $\hat{\sigma}^2 = S^2$

예제: “달콤한 풍선껌” 회사의 제품의 표본에
속한 개체의 끊임없는 차수시차가 다음과 같다.
(현위 2)

61.9	62.6	63.3	64.8	65.1
66.4	67.1	67.2	68.7	69.9

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

$$= 65.7$$

$$\hat{\sigma}^2 = S^2 = \frac{(61.9 - 65.7)^2 + \dots + (69.9 - 65.7)^2}{9}$$

$$= 6.92$$

가 제

1) 과제 소개

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

직접 모의실험을 해 본 경우 강의 내용을 브라 첨기 이해 가능.

2) 모의 실험 내용

- (i) "한 개의 주사위를 6번 던지기"를 30회 반복하기
- (ii) 주사위를 6번 휘젓을 때마다 숫자 1이 나온 횟수를
아래 표에 작성하기

1회	2회	3회	...	29회	30회	평균

각각의 번에는 주사위를
6번 던졌을 때마다
나온 횟수 입력

숫자 1이 나온
횟수의 평균 입력

3) 질문

- (i) 본인이 주한 평균값이 적절하다고 생각되는가?
- (ii) 질문 (i)에 대한 본인의 답에 대한 근거를 설명하라.

(주제내용)

4) 힌트

(i) 주제 주사를 6번 했을 때 숫자 1이 나오는 횟수 X 는 이항분포를 따른다.

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

주사의 6번
인자기

주사의 6번
인자기
인자를 때 숫자 1이
나올 확률

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

$$\bar{X} = \frac{x_1 + \dots + x_{30}}{30}$$

단, x_i 는 X 의 독립간측이다.

↑ "주사의 6번 인자기" 반복하는

행위는 서로 독립이며

동일한 확률분포를 따른다.

(iii) 이제 \bar{X} 의 기대치, 즉 평균과 표준을 구할 수 있다.

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{x_1 + \dots + x_{30}}{30}\right) \\ &= \frac{1}{30} \cdot E(x_1 + \dots + x_{30}) \\ &= \frac{1}{30} \cdot 30 \cdot E(X) \\ &= E(X) \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{x_1 + \dots + x_{30}}{30}\right) \\ &= \frac{1}{30^2} \text{Var}(x_1 + \dots + x_{30}) \\ &= \frac{1}{30^2} \cdot 30 \cdot \text{Var}(X) \\ &= \frac{\text{Var}(X)}{30} \end{aligned}$$

(추가내용)

(iv) 또한 \bar{X} 는 정규분포를 따른다. 즉,

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

중심극한정리에 의해!

이유: $X \sim B(6, \frac{1}{6})$

$$\Rightarrow E(X) = 6 \cdot \frac{1}{6} = 1$$

$$Var(X) = 6 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{6}$$

(v) 결론

각자의 실험 결과는 아래 구간에 속할 것이다.

① $[10 - \frac{1}{6}, 10 + \frac{1}{6}]$ 구간에 속할 확률 81.1%

② $[10 - \frac{1}{3}, 10 + \frac{1}{3}]$ 구간에 속할 확률 97.7%

③ $[10 - \frac{1}{2}, 10 + \frac{1}{2}]$ 구간에 속할 확률 99.9%

5) 추가 질문

(v)의 결론을 어떻게 구했을까?

4. 표본분포

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

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

- 예) • 표본평균의 분포
• 표본비율의 분포
• 표본총산의 분포 (여기서는 다루지 않음)

5. 표본평균의 분포 예제

p.510

예제

- 풍선껌을 봉지에 담아 판매
- 한 봉지에 담긴 풍선껌 개수의 평균은 10, 표준은 1인.
- 한 상자에 30봉지씩 넣어 판매했.

질문: ① 한 상자에 들어 있는 풍선껌 한 봉지에 담긴

풍선껌 개수의 평균은 얼마인가?

② 한 상자에 들어 있는 풍선껌 한 봉지에 담긴

풍선껌 개수의 평균이 8.5개 이하일 확률은?

예제의 질문에 대하기 위해서는 아래의 과정을 진행해야 한다.

- ① 30개의 물자를 무작위로 선택하여 하나의 물자에 들어 있는 풍선개의 개수의 평균을 구해보아야 한다.
- ② 과정①을 여러 번 반복하여 풍선개 개수의 평균의 확률을 알아낸다.

일반적으로 30회 이상.

주의사항은 30개의 물자의 선택은 매번 독립적으로 이루어져야 한다. 즉, 한 물자의 선택은 무작위적으로 이루어지며 한 번 선택된 물자를 다시 원래대로 들여 넣는 **복원 추출** 방식을 따른다.

예제 답안:

X 를 한 물자에 들어 있는 풍선개의 개수의 확률로 나타낸다고 하자. 그러면 다음 사실이 성립한다.

$$\begin{aligned} \textcircled{1} \quad E(X) &= 10 \\ \textcircled{2} \quad \text{Var}(X) &= 1 \end{aligned}$$

\Leftrightarrow \left\{ \begin{array}{l} \text{하나의 물자에 들어 있는 풍선개의 개수가} \\ \text{대한 기대치 (평균값)와 분산} \end{array} \right.

$\textcircled{3}$ 물자 30개 각각의 선택은 **독립관측**이다.

따라서 30물자에 들어 있는 풍선개 개수의 평균값의 분포를 \bar{X} 라 하면 아래 공식이 성립한다.

$$\textcircled{1} \quad \bar{X} = \frac{X_1 + \dots + X_{30}}{30}$$

$\textcircled{2} \quad X_i$ 는 X 의 독립관측

$X_1 + X_2$ 의 평균과 표준偏差으로 \bar{X} 의 평균과 표준偏差를 다음과 같이 구할 수 있다.

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

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

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

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

여기서 정규분포 정리

수학에서 설명!

$$\begin{aligned} \text{따라서 } P(\bar{X} < 8.5) &= P(Z < \frac{8.5 - 10}{\sqrt{0.0333}}) \\ &= P(Z < -1.822) \\ &= 0 \end{aligned}$$

학점을 해이를에서 확인한 결과,
사실상 0인을 의미함.

(주의: 연속성분정 폴도때문)

6. 표본평균의 분포 (일반화)

p.512 ~ 519 $E(X) = \mu$, $\text{Var}(X) = \sigma^2$ 을 따른 X 의 분포가 주어짐.

n 개의 X 를 독립간측으로 인의로 선택했을 때
 X 의 평균값을 \bar{X} 라고 하자. 즉,

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

(X_i 는 X 의 독립간측)

그러면 \bar{X} 의 평균값과 표준은 다음과 같다.

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + \dots + X_n}{n}\right) \\ &= \frac{1}{n} \cdot E(X_1 + \dots + X_n) = \frac{1}{n} \cdot n \cdot E(X) \\ &= E(X) = \mu \end{aligned}$$

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) \\ &= \frac{1}{n^2} \cdot \text{Var}(X_1 + \dots + X_n) = \frac{1}{n^2} \cdot n \cdot \text{Var}(X) \\ &= \frac{1}{n} \cdot \text{Var}(X) = \frac{\sigma^2}{n} \end{aligned}$$

~한번抽取로 n 이 커질수록 표준오차는 줄어든다.

p.520 ~ 521

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

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

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

주의 : X 가 반드시 정규분포를 따를 필요는 없다.

7. 중심극한정리

- ① 조건은 X 에서 통계를 추출할 때 표본의 크기 n 이 충분히 크면
 (예를 들어 $n \geq 30$), \bar{X} 의 분포가 개략적으로 정규분포를 따른다.
 또한, $E(X) = \mu$ 이고 $\text{Var}(X) = \sigma^2$ 이면 다음이
 성립한다.

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

(단, n 의 표본의 크기이며 30 이상이다.)

② 확률 찾기

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

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

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

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

(주의: 연속성보정 필요 없음.)

d. 중성주한 정리 활용

주의: 책 설명과 약간 다른
(책 설명을 일반화 시킨 내용임)

p. 522

1) 이항분포

전제: $X \sim B(n, p)$
 $\Rightarrow \mu = n \cdot p, \sigma^2 = n \cdot p \cdot q \quad (q = 1 - p)$

이항분포를 따르는 X 를 득점 만족으로 보면 반복할 때
중성주한 정리에 의해 X 의 평균값 \bar{X} 는 정규분포를
따른다. 즉,

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

예제) 동전을 10번 던졌을 때 앞면이 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
앞면이 나오는 횟수의 평균은 다음의 봄프를 따른다.

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

$\frac{10 \cdot \frac{1}{2}}{30}$

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

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

$$\begin{aligned} P(\bar{X} > 6) &= P(Z > \frac{6-5}{\sqrt{\frac{1}{12}}}) \\ &= P(Z > \sqrt{12}) = P(Z > 3.46) \\ &= 1 - P(Z \leq 3.46) \\ &= 1 - 0.8987 = 0.0003 \end{aligned}$$

↑ 매우 낮음.

2) 평균 분포

P.522

$$\text{전제: } X \sim P_6(\lambda)$$

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

평균 분포를 따르는 X 를 높임 간격으로 n 번 반복할 때
중심 주된 정리에 의해 X 의 평균값 \bar{X} 는 정규분포를
따른다. 즉,

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

예제) 영화관의 편리 기계의 주당 평균 고장 횟수는
평균 분포를 따르며 $X \sim P_6(3.4)$ 이다.

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 고장 횟수의 평균이
2회 이하일 확률은?

정답: $\bar{X} \sim N(3.4, 0.11)$ 이므로

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

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

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

$$= 0$$

↑ 거의 0임.

주의:
 1) X 자체는
 정규분포를 따르지
 않는다.

신뢰 수준

1. 절주정의 문제와 해결책

1) 문제

- ① 모집단의 평균과 표준을 위해 표본을 활용한 절주정 기법을 살펴보았다. 그런데 추정값을 얻기 위해 단 하나의 표본만을 이용하였다.
- ② 표본이 편향되지 않았다 하더라도 표본이 모집단을 100% 정확하게 반영하는지 여부는 절대로 알 수 없다.
- ③ 절주정은 단지 칙선을 예측하는 것 뿐이다.

2) 해결책

- ① 예를 들어, 평균값에 대한 추정값을 명확하게 말하는 대신에 평균의 추정값이 어느 구간에 속하는지를 말할 수 있다.
- ② 구간의 크기는 결과를 어느 정도 신뢰할 수 있게 만들 것인가에 따라 결정된다.
- ③ 신뢰할 수 있는 정도를 신뢰수준이라 부른다.

2. 절주정 신뢰구간

p. 528

예제: 달콤한 흥선기 회사의 표본조사 결과

$$\begin{cases} \hat{\mu} = \bar{x} = 62.7 (\text{kg}) \\ s^2 = 25 \\ n = 100 \end{cases} \leftarrow \text{표본의 크기}$$

질문: $\hat{\mu}$ 를 어느 정도로 신뢰할 수 있나?

해결책: 신뢰구간 활용.

전비사항: ① 표본평균의 분포 확인

② 신뢰수준 설정

p. 533 ① 예제의 표본분포

$$E(\bar{X}) = \mu, \quad \text{Var}(\bar{X}) = \frac{s^2}{n} = \frac{25}{100} = 0.25$$

\sigma^2을 모를 때
대신 사용
30 보다 큼

따라서, $\bar{X} \sim N(\mu, 0.25)$

주의: i) 모평균 (μ)가 속할 수 있는 구간을 알아내고자 함.

ii) 모분산 (σ^2)을 모를 때 s^2 을 대신 사용

p. 534~535

② 신뢰수준 설정

i) 신뢰수준: 설정한 신뢰구간이 모평균을 어느 정도의 확률로 포함할지에 대한 기준 제공

ii) 일반적으로 90%, 95%, 99% 사용하며,
이 중에 95% 신뢰구간이 가장 많이 사용됨.

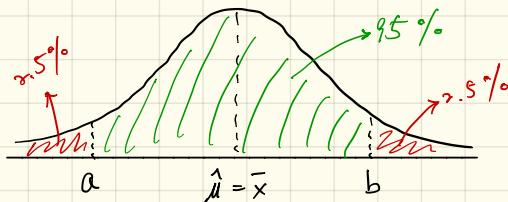
iii) 신뢰수준을 높일수록 신뢰구간이 넓어짐.

iv) 신뢰구간이 너무 크면 신뢰구간의 의미가 약해짐.

(5) 예제의 모평균의 신뢰구간

전제: 여기서 사용하는 신뢰수준: 95 %

\Rightarrow 모평균의 신뢰구간 $[a, b]$ 는 다음을 만족해야 한다.



즉, $P(a < \bar{X} < b) = 0.95$ 를 만족시키는 a 와 b 를 구해야 함.

방법: 표준화 과정 거꾸로하기.

$$i) \text{ 표준화: } z = \frac{\bar{x} - \mu}{\sqrt{0.25}}$$

ii) $P(z_a < z < z_b) = 0.95$ 가 되도록 z_a, z_b 구하기
즉, 다음이 성립해야 한다.

$$\begin{aligned} P(z < z_a) &= 0.025 \\ P(z > z_b) &= 0.025 \end{aligned} \quad \left. \begin{array}{l} z_a = -1.96 \\ z_b = 1.96 \end{array} \right\}$$

iii) z_a, z_b 를 이용하여 μ 의 신뢰구간 구하기.

힌트: 다음 흔히 활용 ↗

$$-1.96 < z = \frac{\bar{x} - \mu}{\sqrt{0.25}} < 1.96$$

$$\bar{x} - 0.98 < \mu < \bar{x} + 0.98$$

\Rightarrow 모평균의 신뢰구간: [61.72, 63.68]

3. 절측정 신뢰기간 일반화

4. 신뢰구간 설정 공식 모음