

확률과 통계

- 추정하기 및 신뢰구간 II -

참고서 : Head First Statistics

(11~12장 : 추정하기 및 신뢰구간)

작성자 : 이계식

환경대학교, 컴퓨터공학과



추정하기

1. 주요 내용

① 점추정 : 표본을 이용해서 모집단에 대한 정보를 추정하기

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

② 비율의 표본분포 \leftarrow 이번에 다룬 내용

③ 표본평균의 분포

④ 중심극한 정리 (Central Limit Theorem)

2. 주요 예제

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

표본 = 원형되지 않은 표본이 구성되어 있다면 가능

연습문제 = "달콤한 풍선껌" 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.494

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

⇒ 절추정을 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 ~~평균 및 분산~~ 절추정

i) 표본이 갖는 성공률을 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.

③ 확률과 비율의 관계

임의로 선택된 사람에게 현대통령에 대한 지지여부를 물어서
긍정적인 답을 얻으면 성공이라고 하였을 때, 모집단에서의
성공의 비율은 현대통령의 지지율과 동일한 것이다.

이와 같이, 성공할 확률을 계산하는 것과 성공의
비율을 계산하는 과정은 완전히 동일하다.

9.49) 5

예제: "달콤한 풍선껌" 회사에서 개발한껌에 대한
신호음을 조사하였다.

40명을 인의로 선택하여 물었더니 그중 10명이
불통식 풍선껌을 신호하였다.

질문: ① 조장간에서 불통식 풍선껌을 신호하는
사람의 비율은 얼마인가?
즉, 사람을 인의로 물었을 때 그 사람이
불통식 풍선껌을 신호할 확률은
얼마인가?

② 조장간에서 불통식 풍선껌을 좋아하지 않는
사람을 신호할 확률은 얼마인가?

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

② $P(\text{불통식을 신호하지 않는 사람})$
 $= 1 - \hat{P} = 1 - 0.25 = 0.75$

가 제

1) 과제 관리

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

직접 모의실험을 해 본 경우 강의 내용을 브라 췄기 이해 가능.

2) 모의실험 내용

(^o) 라나의 주사위가 주어졌다고 가정하자.

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

3) 질문

(^o) 주어진 주사위가 정상적인 주사위인가?

(^{oo}) 질문 (^o)에 대한 물의의 대처에 대한 글자를 설명하라.

(주제내용)

4) 친트

(i) 가정 : 주어진 주사위는 정상이다. 즉, "6면에 1면의 비율로 숫자 1이 나온다"고 추정한다.

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

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

주사위 30번
던지기

주사위는 한 번
던졌을 때 숫자 1이
나올 확률

(ii) ✓ 주사위를 30번 던졌을 때 숫자 1이 나온 횟수의 비율을 나타내는 확률은 $\frac{X}{30}$ 이다. 즉, P_S 는 아래의 확률을 따른다.

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

(iii) 이제 P_S 의 기대치, 즉 평균승률과 표준偏差를 구할 수 있다.

$$\begin{aligned} E(P_S) &= E\left(\frac{X}{30}\right) = \frac{1}{30} E(X) \\ &= \frac{1}{30} E(X) = \frac{1}{30} \cdot 30 \cdot \frac{1}{6} \\ &= \frac{1}{6} = 0.1667 \end{aligned}$$

$$\begin{aligned} \text{Var}(P_S) &= \text{Var}\left(\frac{X}{30}\right) \\ &= \frac{1}{30^2} \cdot \text{Var}(X) = \frac{1}{30^2} \cdot 30 \cdot \frac{1}{6} \cdot \frac{5}{6} \\ &= \frac{1}{6^3} = 0.0046 = (0.0667)^2 \end{aligned}$$

(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 초자 1이 나오는
비율이 $\frac{1}{3}$ 이상일 확률을 계산할 수 있다.

$$\begin{aligned} P(P_s \geq 0.3333) &= P(P_s > 0.333 - \frac{1}{2 \times 30}) \\ &= P(P_s > 0.3167) \\ &= P(z > \frac{0.3167 - 0.1667}{0.068}) \\ &= P(z > 2.21) \\ &= 1 - P(z \leq 2.21) \\ &= 1 - 0.9864 \\ &= 0.014 \quad (\text{즉}, 1.4\%) \end{aligned}$$

연속성보정
(뒤에 설명됨)

5) 결론

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

95% 확률로

즉, "주어진 주사위가 정상적이다라는 가정을 받아들이지
않는 게 일반적이다."

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)

비율

5. 표본평균의 분포 예제

p.500 예제

- 풍선껌을 100개씩 상자에 담아 판매
- 조사단에서 빨간색 풍선껌의 비율은 $P = 0.25$ 로 추정함.

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

$$P(P_s > 0.4) = ?$$

예제의 질문에 대답하기 위해서는 아래 그림을 확인해야 한다:

통산 100개들이 한 상자에 들어 있는 떨간색 풍선껌의
비율에 대한 표본률을 확인해 한다.

예제 답안:

X 를 100개 들이 한 상자에 들어 있는 떨간색 풍선껌의
개수라 하자. 그러면 X 는 이항분포를 따른다.

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

따라서 100개 들이 한 상자에 들어 있는 떨간색 풍선껌의
비율의 확률을 P_s 라 하면 다음과 성립한다.

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

P_s 의 평균값과 표준을 다음과 같이 구한다.

$$\begin{aligned} E(P_s) &= E\left(\frac{X}{100}\right) \\ &= \frac{1}{100} E(X) \\ &= \frac{1}{100} \cdot 100 \cdot \frac{1}{4} \\ &= 0.25 \end{aligned}$$

$$\begin{aligned} \text{Var}(P_s) &= \text{Var}\left(\frac{X}{100}\right) \\ &= \frac{1}{100^2} \cdot 100 \cdot \frac{1}{4} \cdot \frac{3}{4} \\ &= \frac{3}{1600} \\ &= 0.001875 \end{aligned}$$

이제 $P(P_s \geq 0.4)$ 를 구하자 한다.
그런데 P_s 는 어떤 확률을 따른가?
 \Rightarrow 정규분포를 따른다.

즉, $P_s \sim N(0.25, 0.001875)$ 가 성립한다.

이제 정규분포정리

$$\begin{aligned} \text{따라서 } P(P_s \geq 0.4) &= P\left(P_s > 0.4 - \frac{1}{2 \cdot 100}\right) \\ &= P(P_s > 0.395) \\ &\rightarrow P(z > \frac{0.395 - 0.25}{\sqrt{0.001875}}) \\ &= P(z > 3.35) \\ &= 1 - P(z \leq 3.35) \\ &= 1 - 0.9996 \\ &= 0.0004 \end{aligned}$$

* 연속성 보정: 표본의 크기가 1000 이하일 경우
연속성 보정 필요.

비율

6. 표본 ~~중간~~의 빈도 (일반화)

p.500~503 조정단에서 특정 사건이 발생할 확률이 P 라 하자.

그러면 n 개의 표본에서 해당 사건이 발생하는 횟수 X 는 이항분포를 따른다. 즉,

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

따라서, n 개의 표본에서 해당 사건이 발생할 확률 P_s 의 빈도는 $P_s = \frac{X}{n}$ 이다.

P_s 의 평균값과 흔적은 다음과 같다.

$$\begin{aligned} E(P_s) &= E\left(\frac{X}{n}\right) \\ &= \frac{1}{n} \cdot E(X) = \frac{n \cdot P}{n} = P \end{aligned}$$

$$\begin{aligned} \text{Var}(P_s) &= \text{Var}\left(\frac{X}{n}\right) \\ &= \frac{1}{n^2} \cdot \text{Var}(X) = \frac{1}{n^2} \cdot n \cdot P \cdot q \\ &= \frac{P \cdot q}{n} \quad (q = 1 - P) \end{aligned}$$

p.504~505

또한 $n \geq 30$ 일 때 P_s 는 중심극한정리에 의해 정규분포를 따른다. 즉,

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

7. P_s 의 연속성 보정

$$\left\{ \begin{array}{l} P_s = \frac{X}{n}, \\ X = \text{표본이 갖는 성공의 수} \end{array} \right.$$
$$\Rightarrow \text{연속성 보정 값} = \pm \frac{1}{2n}$$

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

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

하지만 $n > 1,000$ 인 경우 연속성 보정 출 필요!

신뢰 수준

1. 절주정의 문제와 해결책

P.529

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

P.530

- ① 예를 들어, ~~평균값~~에 대한 추정값을 명확하게 말하는 대신에 ~~평균~~의 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

3. 절측정 신뢰구간 구하기 문제 정리

p. 543
~544

(1) 표본단 통계 선택

- 예) ~~평균~~ 한 자속시간의 평균에 대한
신뢰구간 설정 필요,
즉, ~~P~~에 대한 신뢰구간 설정 필요

(2) 표본분포 찾기

- 예) 표본분포의 기대치와 표준 계산
• P_s 의 정규분포 활용

(3) 신뢰도 정하기

- 예) 일반적으로 95%의 신뢰도로 사용

(4) 신뢰구간 찾기

- 예) 신뢰도와 표본분포를 사용하여 신뢰구간 구하기.

3

신뢰구간 설정 공식 모음

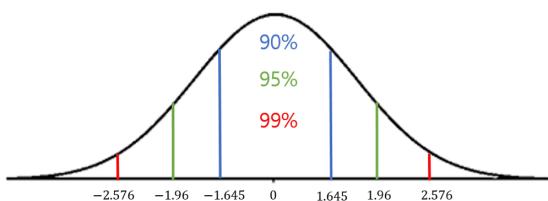
보정방법	보정방법	조건	신뢰구간
μ	X 정규분포	σ^2 알고 있음 n 표본크기 \bar{X} 표본평균	$(\bar{X} - C \frac{\sigma}{\sqrt{n}}, \bar{X} + C \frac{\sigma}{\sqrt{n}})$
μ	X 정규분포 아니면 \bar{X}	σ^2 알고 있음 $n \geq 30$ \bar{X} 표본평균	$(\bar{X} - C \frac{\sigma}{\sqrt{n}}, \bar{X} + C \frac{\sigma}{\sqrt{n}})$
μ	X 일정의 분포	σ^2 모름 $n \geq 30$ \bar{X} 표본평균	$(\bar{X} - C \frac{s}{\sqrt{n}}, \bar{X} + C \frac{s}{\sqrt{n}})$
P	이항	$n \geq 30$ p_s 표본비율 $q_s = 1 - p_s$	$(p_s - C \sqrt{\frac{q_s}{n}}, p_s + C \sqrt{\frac{q_s}{n}})$

오차범위는
기본적으로 아래의
형태이다 :

$C * (\text{표본 표준편차})$

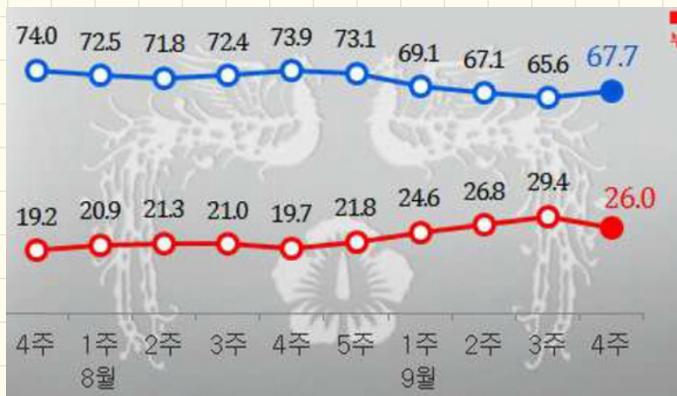
* C 의 값 : 신뢰수준에 의해 결정됨

신뢰수준	C 의 값
90%	1.64
95%	1.96
99%	2.58



4. 연습문제 (책에 있는)

지자율 선출조사가 대표적인 표본비율의 흐름을 사용한다. 아래 그림에서 표정오차 크기를 아래한 후 풀어야 한다.



총 응답자: 2,523명
선택률: 95%
오차범위: ±2.0%

CBS • 조사기관: 리얼미터 • 총응답자: 전국 성인 2,523명 • 응답률: 5.4%
무선 전화면접(10%), 무선(70%) • 유선(20%) 자동응답 훈용
95% 신뢰수준 ±2.0%p • 조사기간: 2017년 9월 25일(월) ~ 9월 29일(금)

* 신뢰오차 $\pm 2.0\%$ 이 사용된 "2.0%"에 대한 근거

$$\textcircled{1} \quad \text{신뢰수준} = 95\% \Rightarrow C = 1.96$$

$$\textcircled{2} \quad P_s = 0.697, q_s = 1 - P_s = 0.323$$

$$\textcircled{3} \quad M = 2,523$$

$$\Rightarrow C \cdot \sqrt{\frac{P_s \cdot q_s}{n}} = 1.96 \cdot \sqrt{\frac{0.697 \cdot 0.323}{2523}} \\ = 0.018$$

(약 2%)