비모수 통계학 (Nonparametric Statistics)

기본 개념 (Basic Concepts)

Kipoong Kim

Department of Statistics
Changwon National University

Fall, 2024

Syllabus

1. 일표본 위치 문제
1. 일표본 위치 문제
2. 이표본 위치 문제
2. 이표본 위치 문제
3. 이표본 척도 문제
3. 이표본 척도 문제
4. 분포함수 적합도 문제

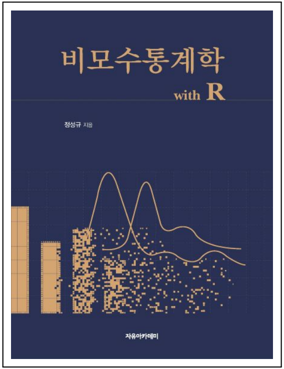
◦ 중간고사

5. 일원배치법
5. 일원배치법
6. 이원배치법
6. 이원배치법
7. 독립성과 순위상관
8. 회귀분석
8. 회귀분석

◦ 기말고사

Syllabus

- 연구실 : 32동 411호
- 상담시간 : 수 16:30 - 18:00
- 이메일 : statpng@changwon.ac.kr
- 참고자료 : **비모수통계학 with R**, 정성규



Syllabus

- 평가방법
  (1) 과제 (10%)
  (2) 중간고사 (40%)
  (3) 기말고사 (50%)

- 성적
  (1) A+ : $\approx$ 30%
  (2) B+ : $\approx$ 40%
  (3) C, D, F : tentative

통계학

- 통계학 (statistics)
  - 기술 통계학 (descriptive statistics)
    - 데이터 마이닝 (data mining)
    - 시각화 (visualization)
  - 추측 통계학 (inferential statistics)
    - 추정 (estimation)
      - **점추정** (point estimation)
      - **구간추정** (interval estimation)
    - **가설검정** (test of hypotheses)

$\Rightarrow$ 비모수통계는 모집단에 대한 가정을 적게 한 상태에서의 통계

Examples

- 어떤 소나무의 성장을 연구하기 위하여 1년생 소나무 묘목 16그루의 키를 조사하였다. 소나무 묘목의 평균키를 μ라 할 때

12	12	18	17	18	15	20	22
15	16	22	21	31	17	17	12

(1) 미지의 μ의 값을 추정하는 것. (점추정)

(2) μ가 속할 것으로 기대되는 구간을 결정하는 것. (구간추정)

(3) 다른 종류의 소나무 묘목의 평균키는 19cm라고 알려져있다고 할 때, 이 소나무 묘목의 평균키 μ도 역시 19cm인가를 결정하는 것. (가설검정)

점추정

- 모수 θ를 추정하기 위한 추정량으로 여러개의 추정량을 고려할 수 있다.
- 추정량은 변수이므로 추출된 표본에 따라 추정값이 달라질 수 있다.

구간추정

- 점추정에서는 관심이 있는 모수 θ의 참값 θ_0에 대한 추정으로 통계량 $\hat{\theta}$를 사용했지만 실제로 $\hat{\theta}$이 θ_0일 확률은 아주 낮다. $\hat{\theta}$이 연속형 분포를 따른다면 $P(\hat{\theta} = \theta_0) = 0$이다.

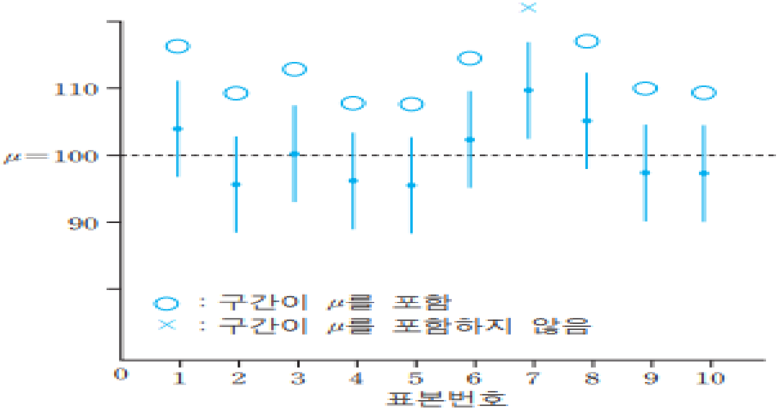
- 따라서 모수 θ를 추정하기 위해 참값이 포함될 것이라 예상되는 구간을 제시하는 것을 **구간추정**(interval estimation)이라고 한다.

- 참값이 포함될 확률에 따라 추정하는 참값의 하한값과 상한값으로 이루어진 구간을 **신뢰구간**(confidence interval)이라고 한다.

신뢰구간

- 주어진 확률 $1-\alpha$에 대하여 $P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1-\alpha$를 만족하는 구간 $(\hat{\theta}_L, \hat{\theta}_U)$를 모수 θ에 대한 $100(1-\alpha)\%$ 신뢰구간이라고 한다.
- 즉, 신뢰구간에서는 θ를 포함할 확률이 $1-\alpha$이고 이 확률을 **신뢰수준**(confidence level) 또는 **신뢰도**(confidence coefficient)라고 한다.
- α는 주로 $0.05, 0.01$를 사용한다.

신뢰구간의 해석



가설검정

- 가설검정의 절차
  (1) **귀무가설**(H_0, null hypothesis)과 **대립가설**(H_1, alternative hypothesis)을 세운다.
  (2) 검정통계량을 통해 기각역을 설정하고 검정결과를 도출한다.

- 가설검정의 결과
  (1) 귀무가설 H_0를 기각한다.
  (2) 귀무가설 H_0를 기각하지 못한다.

단측검정과 양측검정

(1) 우측검정은 모수 θ에 대해 귀무가설과 대립가설이 다음 식으로 주어진다.

$$H_0 : \theta = \theta_0, \quad H_1 : \theta > \theta_0$$

(2) 좌측검정은 모수 θ에 대해 귀무가설과 대립가설이 다음 식으로 주어진다.

$$H_0 : \theta = \theta_0, \quad H_1 : \theta < \theta_0$$

(3) 양측검정은 모수 θ에 대해 귀무가설과 대립가설이 다음 식으로 주어진다.

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0$$

검정오류

미지의 실제 상황 \표본에 의한 검정	H_0 채택	H_0 기각
H_0가 참	옳은 판단	제 1종 오류
H_0가 거짓	제 2종 오류	옳은 판단

- 귀무가설 H_0가 참일 때 잘못하여 H_0를 기각하는 오류를 **제1종 오류**(type I error)라 하고 제1종 오류의 확률을 **유의확률**(significant probability, p-value, p-값), 유의확률의 최댓값을 **유의수준**(significant level, α)이라고 한다.

- 귀무가설 H_1이 참일 때 잘못하여 H_0를 채택하는 오류를 **제2종 오류**(type II error)라 하고 제2종 오류의 확률을 간단히 β로 표현한다. 또한 $1 - \beta$를 **검정력**(statistical power)이라고 한다.

부호와 순위

- **부호**(sign) : 관측값이 어떤 기준값 c보다 크면 1, 아니면 0

$$S(x,c) = \begin{cases} 1, & x > c \\ 0, & x \le c \end{cases}$$

- **순위**(rank) : 관측값을 작은 값에서 큰 값 순으로 나열하였을 때의 순서

변수	X_1	X_2	X_3	X_4	X_5
관측값	7	2	11	5	5
부호($c = mean(X)$)	1	0	1	0	0
순위	4	1	5	2.5	2.5

Example 0.1

임의로 선정된 10명의 A초등학교 어린이들이 IQ 점수를 높이기 위해 특별한 훈련을 1주 동안 받았다. 훈련을 마친 뒤, 초등학교 고학년에게 맞게 설계된 IQ 검사를 받아 다음과 같은 점수를 받았다.

$$98, 121, 110, 89, 109, 108, 102, 92, 131, 114 \Rightarrow \sum = 1074$$

IQ 검사의 점수는 평균이 100, 표준편차가 10이 되도록 표준화된 점수이다. 이 초등학생들에게 적용된 특별한 훈련이 효과가 있는가?

$$\text{i.e.,} \quad H_0 : \theta = 100, \quad v.s. \quad H_1 : \theta > 100$$

Solution 0.1 (a)

$$\text{검정통계량 } Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

$$Z = \frac{107.4 - 100}{10/\sqrt{10}} \approx 2.34$$

$$p.value = P(Z > 2.34) = 0.0096$$

$$\Rightarrow \text{유의수준 } \alpha = 0.05\text{에서 귀무가설 } H_0\text{를 기각한다.}$$

Solution 0.1 (b)

- 검정통계량: $B = 100$보다 큰 관측값의 개수 $\sim B(n, 0.5)$.

  $\Rightarrow B = \sum_{i=1}^{n} S(x_i, 100) = 7$

- p-value $= P(B \geq 7) = \sum_{k=7}^{10} \binom{10}{k} 0.5^k (1-0.5)^{10-k} = 0.17$

  $\Rightarrow$ 유의수준 $\alpha = 0.05$ 하에서 귀무가설 H_0를 기각하지 못한다.

Example 0.2

- 임의로 선정된 B초등학교 3명의 어린이들과 훈련을 받은 A초등학교 2명의 어린이들의 IQ 검사의 점수를 비교하고자 한다.

$$B = \{99, 88, 93\}, \quad A = \{101, 114\}$$

  이 A초등학교 학생들에게 적용된 특별한 훈련이 효과가 있는가?

$$\text{i.e.,} \quad H_0 : \mu_B = \mu_A, \quad v.s. \quad H_1 : \mu_B < \mu_A$$

$$\iff \text{Let } \mu_A = \mu_B + \Delta, \quad H_0 : \Delta = 0, \quad v.s. \quad H_1 : \Delta > 0$$

Solution 0.2

- 검정통계량 $W = A$의 순위의 합

$$W = \sum_{j=1}^{2} R_j, \quad R_j = \{B_1, B_2, B_3, A_1, A_2\}\text{에서 } A_j\text{의 순위}$$

번호	배열	W	확률
1	$AABBB$	3	0.1
2	$ABABB$	4	0.1
3	$ABBAB$	5	0.1
4	$ABBBA$	6	0.1
5	$BAABB$	5	0.1
6	$BABAB$	6	0.1
7	$BABBA$	7	0.1
8	$BBAAB$	7	0.1
9	$BBABA$	8	0.1
10	$BBBAA$	9	0.1

Solution 0.2

W	3	4	5	6	7	8	9	계
$P(W = w)$	0.1	0.1	0.2	0.2	0.2	0.1	0.1	1

$$W = \sum_{j=1}^{2} R_j = 9$$

$$p.value = P(W \geq 9) = 0.1$$

$$\Rightarrow \text{유의수준 } \alpha = 0.1\text{에서 귀무가설 } H_0\text{를 기각한다.}$$