

학습계획서

| | | | |
|---|----------|-----|----------|
| 팀 | 호성이와 아이들 | 구성원 | 서호성, 최홍용 |
|---|----------|-----|----------|

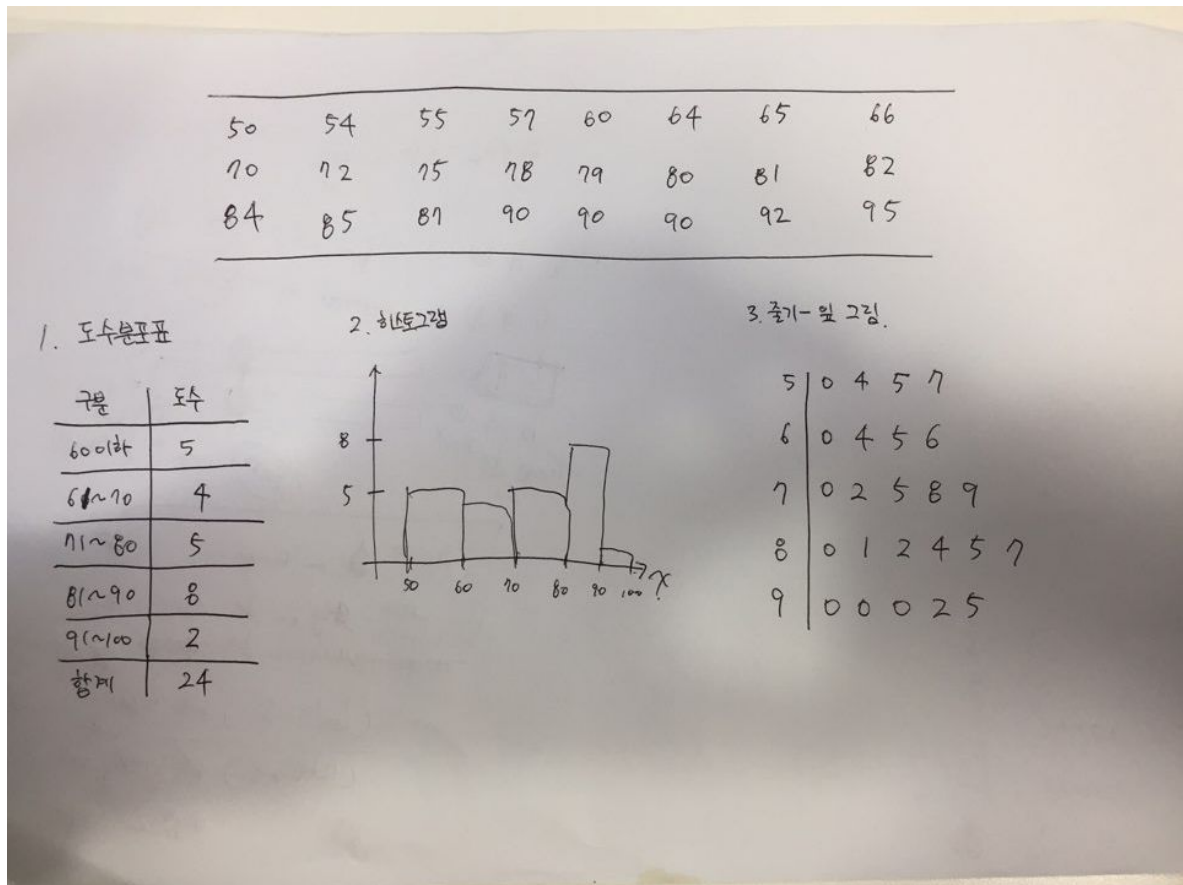
| 일정 | 발제자 | 주제 | 주요내용 |
|--------------------|-----|--|---|
| 1일차 (5 / 27) | 서호성 | Big Data Learning Packet : Basics of Statistics | 1. data form 2. random variable & distribution 1 3. random variable & distribution 2 |
| 2일차 (5 / 28) | 최홍용 | Big Data Learning Packet : Basics of Statistics | 4. normal distribution 5. sampling distribution & central limit theorem 6. statistical inference |
| 3일차 (5 / 29) | 서호성 | Big Data Learning Packet : Basics of Statistics | 7. statistical testing 8. population mean testing 9. correlation analysis |
| 4일차 (5 / 30) | 최홍용 | Big Data Learning Packet : Basics of Statistics | 10. simple linear regression 11. analysis of variance |
| 5일차 (5 / 31) | 서호성 | Computational Thinking and Data Science | Introduction and Optimization Problems |
| 6일차 (6 / 3) | 최홍용 | Computational Thinking and Data Science | Optimization Problems |
| 7일차 (6 / 4) | 서호성 | Computational Thinking and Data Science | Graph-theoretic Models |
| 8일차 (6 / 5) | 최홍용 | Computational Thinking and Data Science | Stochastic Thinking |
| 9일차 (6 / 7) | 서호성 | Computational Thinking and Data Science | Random Walks |
| 10일차 (6 / 10) | 최홍용 | Computational Thinking and Data Science | Monte Carlo Simulation |

학습 정리

| | | | |
|---|---------|-----|----------|
| 팀 | 호성이와 아이 | 구성원 | 서호성, 최홍용 |
|---|---------|-----|----------|

| 일정 | 발제자 | 주제 |
|------------|-----|---|
| (5 / 27) | 서호성 | Big Data Learning Packet : Basics of Statistics |

주요 내용 요약



1. 어느 전기부품이 고장 날 때까지 걸리는 시간을 조사하기 위하여 24개 부품을 실험한 결과 다음의 자료를 얻었다.

| | | | | | | | | | |
|-----|----|----|----|----|----|----|----|----|----|
| 44 | 48 | 64 | 51 | 32 | 29 | 48 | 39 | 51 | 55 |
| 101 | 49 | 74 | 59 | 56 | 62 | 60 | 37 | 61 | 73 |
| 122 | 45 | 69 | 52 | | | | | | |

- (1) 이 표본에서 고장 날 때까지 걸린 시간의 평균을 구하라.

$$m = \frac{\sum x_i}{n} = \frac{(44+48+\dots+52)}{24} = 57.19$$

- (2) 고장 날 때까지 걸린 시간의 표준편차를 구하라.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(44-57.19)^2 + \dots + (52-57.19)^2}{23}} = 20.51$$

1. 세 명의 학생이 각각 백화점에서 구두나 운동화 중 하나를 산다.
 서로의 구매에 영향을 받지 않고, 모두 반반의 가능성을 가지고 결정한다.
 여기서 확률 변수 X 를 세명 중 구두를 구매한 학생의 수라고 할 때,
 평균, 분산을 구하여라.

A: 구두 구매 B: 운동화 구매

| value of X | 0 | 1 | 2 | 3 |
|--------------|-----|-------------------|-------------------|-----|
| 사건 | BBB | ABB BAB BBA | AAB ABA BAA | AAA |

- 확률 분포표

| X | 0 | 1 | 2 | 3 | |
|-------|---------------|---------------|---------------|---------------|---|
| P_r | $\frac{1}{8}$ | $\frac{3}{8}$ | $\frac{3}{8}$ | $\frac{1}{8}$ | 1 |

(1) 평균

$$E(X) = \sum x \cdot P(X=x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}$$

(2) 분산

$$V(X) = \sum (X-M)^2 \cdot P(X=x)$$

$$= E(X^2) - \{E(X)\}^2$$

$$= 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} - \left(\frac{3}{2}\right)^2$$

$$= \frac{3}{4}$$

2. 확률 밀도 함수가

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{그 외} \end{cases}$$

(1) C의 값은 얼마인가?

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x) dx \\ &= \int_{-\infty}^0 f(x) dx + \int_0^2 f(x) dx + \int_2^{\infty} f(x) dx \\ &= \int_0^2 C(4x - 2x^2) dx = C \left[2x^2 - \frac{2}{3}x^3 \right]_0^2 = \frac{8}{3} C. \\ \therefore C &= \frac{3}{8} \end{aligned}$$

(2) $P(X > 1)$ 의 값을 구하라.

$$\begin{aligned} P(X > 1) &= \int_1^{\infty} f(x) dx = \int_1^2 f(x) dx + \int_2^{\infty} f(x) dx \\ &= \int_1^2 f(x) dx = \frac{3}{8} \left[2x^2 - \frac{2}{3}x^3 \right]_1^2 \\ &= \frac{3}{8} \left\{ \left(8 - \frac{16}{3} \right) - \left(2 - \frac{2}{3} \right) \right\} = \frac{1}{2} \end{aligned}$$

학습 정리

| | | | |
|---|----------|-----|----------|
| 팀 | 호성이와 아이들 | 구성원 | 서호성, 최홍용 |
|---|----------|-----|----------|

| 일정 | 발제자 | 주제 |
|------------|-----|---|
| (5 / 28) | 최홍용 | Big Data Learning Packet : Basics of Statistics |

주요 내용 요약

정규분포 - 응용문제

문제 어느 회사에 입사를 희망한 자원자의 영어점수는 평균이 700이고 표준편차가 100인 정규분포를 따른다고 한다.

1) 합격자 중 영어점수가 최하인 사람의 점수가 870점인 때, 몇 퍼센트의 자원자가 합격하였을까?

X : 입사를 희망한 자원자의 영어점수

$$X \sim N(700, 100^2)$$

$$\begin{aligned} P(X \geq 870) &= P\left(Z \geq \frac{870 - 700}{100}\right) \\ &= P(Z \geq 1.7) \\ &= P(Z \leq -1.7) = 0.0446 \end{aligned}$$

2) 상위 15%를 선택하기 위한 기준과 같은 점수인가?

$$\begin{aligned} P(X \geq x) \\ &= P\left(Z \geq \frac{x - 700}{100}\right) \\ &= 0.15 \end{aligned}$$

$$P(Z \geq 1.036) = 0.15$$

$$\frac{x - 700}{100} = 1.036 \Rightarrow x = 803.6 \text{ 점}$$

$$Z = 1.036$$

표본분포와 중심극한정리 - 연습문제

문제 어느 도시 원자의 수명은 평균 250만 원이고, 표준편차는 50만 원이라고 한다.

(1) 100명을 표본으로 선택했을 때, 표본 평균의 분포는 무엇인가?

단위 : 만원

μ : 도시 원자의 수명

$$\bar{x} \sim N(250, 5^2)$$

$$E(\bar{x}) = 250$$

$$S(\bar{x}) = 50/\sqrt{100} = 5$$

(2) $P(\bar{x} > 260 \text{ 만원})$ 은 얼마인가?

$$P(\bar{x} \geq 260) = P(Z > \frac{260 - 250}{5})$$

$$= P(Z > 2) = 0.0228$$

문제 모량화 평균 550이고, 표준 편차가 70일 때, 다음의 각 경우에 표본평균 \bar{x} 의 분포를 구하라.

(1) 표본의 크기는 16으로 한다.

$$\bar{x} \sim N(M, (\frac{\sigma}{\sqrt{n}})^2)$$

$$E(\bar{x}) = 550$$

$$S(\bar{x}) = \frac{70}{\sqrt{16}} = 17.5$$

$$\bar{x} \sim N(550, 17.5^2)$$

(2) 표본의 크기는 160으로 한다.

$$\bar{x} \sim N(M, (\frac{\sigma}{\sqrt{n}})^2)$$

$$E(\bar{x}) = 550$$

$$S(\bar{x}) = \frac{70}{\sqrt{160}} = 5.53$$

$$\bar{x} \sim N(550, 5.53^2)$$

통계학 중간-연습문제

문제) 유산자에 거주하는 성인 1인당 신용카드 보유 개수는 추정하기 위해 25명을 무작위로 추출하여 조사한 결과 1인당 평균 7.314, 표준편차는 2.67였다. 유산자일 1인당 평균 신용카드 보유 개수에 대한 95% 신뢰구간을 구하시오.

$$\begin{aligned} n &= 25 & CI & \bar{x} \pm t^* \frac{s}{\sqrt{n}} \\ \bar{x} &= 7.3 & & 7.3 \pm 2.0797 \cdot \frac{2.6}{\sqrt{25}} \\ s &= 2.6 & & = (5.846, 8.754) \\ & & & 5.846 \leq \mu \leq 8.754 \end{aligned}$$

문제) 기존 컴퓨터의 시뮬레이션 실행을 조사한 결과 편집명령에 대한 반응시간이 표준편차가 25밀리 초인 정규분포를 따르는 사실을 알았다. 새로운 운영체제가 설치되었다. 이 시스템에 대한 평균 반응시간을 추정하고 싶다. 이 시스템의 반응시간도 표준편차가 25밀리의 구인 정규분포를 따르는 가정할 때, 모집단에 대한 95% 신뢰구간의 폭이 최대 10이 되게끔 추정하려면 표본 크기를 얼마로 정해야 할까?

$$\begin{aligned} \sigma &= 25 \\ l &= 2 \times \frac{\sigma}{\sqrt{n}} \leq 10 \\ 2 \times 25 \times \frac{1}{\sqrt{n}} &\leq 10 \\ \left(\frac{2 \times 25 \times 1}{10} \right)^2 &\leq n \rightarrow 96.04 \leq n \\ \therefore n &= 97 \end{aligned}$$

학습 정리

| | | | |
|---|----------|-----|----------|
| 팀 | 호성이와 아이들 | 구성원 | 서호성, 최홍용 |
|---|----------|-----|----------|

| 일정 | 발제자 | 주제 |
|------------|-----|---|
| (5 / 29) | 서호성 | Big Data Learning Packet : Basics of Statistics |

주요 내용 요약

문제

1. 지금까지 결핵 치료에 잘 듣는 항생제 A의 평균 치료율이 70%라고 하자. 새로운 항생제 B가 개발되었다. 새로운 항생제 B를 개발한 연구원들은 새 항생제 B가 기존의 항생제 A보다 치료율이 높다고 주장한다. 이것을 검증하기 위하여 결핵환자 100명을 랜덤으로 뽑아서 항생제 B를 일정 기간 투여한다. 그리고 100명 중 치료된 사람의 수를 X 라 하자. 항생제 B의 정확한 치료율 p 의 값은 전혀 모르지만, 표본에서 치료율은 $X/100$ 으로 추정된다. 이러한 경우에 적합한 가설은 무엇인가?

귀무가설 (H_0) : 항생제 B의 치료율이 항생제 A보다 낮지 않다.

대립가설 (H_1) : 항생제 B의 치료율이 항생제 A보다 낮다.

$$\Rightarrow H_0 : p \leq 0.7$$

$$H_1 : p > 0.7.$$

문제

2. 어느 정당에 대한 지지율은 지난 몇 달간 50%를 유지하고 있었다. 새로운 정책의 발표로 지지율이 변한 것 같아 표본조사를 하여 확인 하려 한다. 모두 10명의 사람을 랜덤 추출하여 지지 여부를 묻고, 지지하는 사람의 수를 확률변수 X 로 놓는다.

(1) 모집단의 지지율을 P 라 할 때, 지지율이 달라졌는지에 대한 가설을 세운다.

$$H_0 : P = 0.5 \quad / \quad H_1 : P \neq 0.5$$

(2) 만일 $X \leq 2$ or $X \geq 8$ 이면, H_0 을 기각한다.

이때, 제 1종 오류를 범할 확률 α 를 구하라.

$$X \sim B(10, \frac{1}{2})$$

$$\alpha = P(H_0 \text{ 기각} \mid H_0 \text{ 참})$$

$$= P(X \leq 2 \text{ or } X \geq 8)$$

$$= 0.11$$

문제

1. A대학 신입생의 영어 성적은 평균 75점에 표준편차가 15점이라고 한다. 90명의 신입생을 표본으로 뽑아 영어 모의고사를 치렀더니 평균 71점이었다. A대학 신입생의 영어성적이 75점이라고 할 수 있는가를 $\alpha=0.01$ 의 수준에서 검정하시오.

$$\mu = 75, \sigma = 15, n = 90, \bar{x} = 71$$

(1) 가설 설정: $H_0: \mu = 75$ vs $H_1: \mu \neq 75$

(2) 유의수준 결정: $\alpha = 0.01$, 양측검정.

(3) 검정통계량 계산:

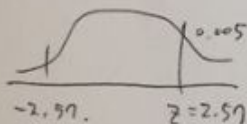
$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{71 - 75}{15 / \sqrt{90}} = -2.53.$$

Sol 1) p-value 이용

$$P\text{-value} = 2 \cdot P(Z \geq |-2.53|) = 0.011 > 0.01 \rightarrow H_0 \text{ 채택}$$

Sol 2) 구간 비교

$$\alpha = 0.01$$



$$\text{채택영역: } -2.57 < Z < 2.57$$

$$Z = -2.53 \text{ 이므로 } H_0 \text{ 채택}$$

Sol 3) 신뢰구간 이용

$$99\% \text{ CI: } 71 \pm Z_{0.005} \cdot \frac{15}{\sqrt{90}} = (66.936, 75.063)$$

$$\mu = 75 \text{ 가 범위안에 포함되지 때문에 } H_0 \text{ 채택}$$

문제

2. 어떤 종류의 토양은 자연 상태에서 평균 8.75의 pH값을 갖는다고 한다. 대체 토양을 합성하였다. 5개의 시료에서 pH의 평균이 8.00, 표준편차가 0.05로 측정되었다. 자연상태의 토양과 차이가 나는가? 유의 수준 $\alpha=0.01$ 에서 검정하라.

$$n = 5, \quad \bar{x} = 8, \quad S = 0.05$$

(1) 가설 설정: $H_0: \mu = 8.75$ vs $H_1: \mu \neq 8.75$

(2) 유의수준 설정: $\alpha = 0.01$, 양측검정

(3) 검정통계량 계산:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{8 - 8.75}{0.05/\sqrt{5}} = -33.54$$

(4) p-value 계산:

$$p = 2 \cdot P(t \geq | -33.54 |) \approx 0$$

(5) p-value와 유의수준 비교.

$$0 < 0.01 \Rightarrow H_0 \text{ 기각}$$

학습 정리

| | | | |
|---|----------|-----|----------|
| 팀 | 호성이와 아이들 | 구성원 | 서호성, 최홍용 |
|---|----------|-----|----------|

| 일정 | 발제자 | 주제 |
|------------|-----|---|
| (5 / 30) | 최홍용 | Big Data Learning Packet : Basics of Statistics |

주요 내용 요약

선형 회귀의 - 연습문제

문제

다음과 같이 자료가 주어졌다.

| | | | | | | |
|---|---|---|---|---|---|---|
| x | 1 | 2 | 3 | 5 | 6 | 7 |
| y | 4 | 6 | 3 | 1 | 3 | 1 |

(1) 최소제곱법으로 β_0, β_1 의 추정치를 구하고, 회귀선을 구하여라.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\bar{x} = \sum x_i / n = \frac{1}{6} (1+2+3+5+6+7) = 4$$

$$\bar{y} = \sum y_i / n = \frac{1}{6} (4+6+3+1+3+1) = 3$$

$$\hat{\beta}_1 = \frac{(1-4)(4-3) + (2-4)(6-3) + \dots + (7-4)(1-3)}{(1-4)^2 + (2-4)^2 + (3-4)^2 + \dots + (7-4)^2} = -0.607$$

$$\hat{\beta}_0 = 3 - (-0.607) \cdot 4 = 5.428$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x = 5.428 - 0.607x$$

(2) $x=6$ 일 때, y 값을 추정하라.

$$\hat{y} = 5.428 - 0.607x = 5.428 - 0.607 \cdot 6 = 1.786$$

문제

다음과 같이 계산된 결과가 다음과 같다.

$$n=14, \quad \bar{x}=1.2, \quad \bar{y}=5.1, \quad \sum (x_i - \bar{x})^2 = 14.1$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 2.31, \quad \sum (y_i - \bar{y})^2 = 2.01$$

y 의 총 변동량 중 회귀성에 의해 설명되는 변동량의 비율을 구하라.

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{2.01}$$

$$= \frac{1}{2.01} \sum ((\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i - \bar{y})^2 = \frac{1}{2.01} \sum (\hat{\beta}_1 (x_i - \bar{x}))^2$$

$$= \frac{\hat{\beta}_1^2}{2.01} \sum (x_i - \bar{x})^2 = \left\{ \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right\}^2 \times \frac{14.10}{2.01} = \left(\frac{2.31}{14.10} \right)^2 \times \frac{14.10}{2.01}$$

$$R^2 = 0.1882$$

부산물 - 영수증

문제

A 한림대학교는 세 개의 다른 진열 방식에 대해 몇 차례 판매 실험을 해 본 결과 다음과 같은 데이터를 얻었다. 다음 분석표를 작성하고 진열 방식에 따라 판매량이 다른지 $\alpha = 0.01$ 수준에서 검정하여라.

| 1안 | 2안 | 3안 |
|-----|-----|-----|
| 120 | 122 | 144 |
| 122 | 135 | 144 |
| 119 | 138 | 134 |
| 140 | 138 | 142 |
| 118 | 140 | 152 |

관측치 $n = 15$

처리수 = 3

| 요인 | 제곱합 | 자유도 | 평균제곱 | F값 |
|-----------------|---------|------|-----------|--------|
| 집단 간 (SS_c) | 994.93 | (2) | (492.465) | (7.95) |
| 집단 내 (SS_E) | 712.8 | (12) | (59.4) | |
| 합계 (SS_T) | 1657.73 | (14) | | |

$$df_T = n - 1 = 15 - 1 = 14$$

$$df_E = 3 - 1 = 2$$

$$df_c + df_E = df_T \Rightarrow df_E = 12$$

$$MS_c = SS_c / df_c = 994.93 / 2 = 492.465$$

$$MS_E = SS_E / df_E = 712.8 / 12 = 59.4$$

$$F = MS_c / MS_E = 492.465 / 59.4 = 7.95$$

분석표 (ANOVA)

* 가설 설정

1. 가설 설정 : $H_0 : \mu_1 = \mu_2 = \mu_3$, $H_1 : \text{not } H_0$

2. 유의수준 설정 : $\alpha = 0.01$

3. 검정통계량 계산 : $F_{0.01, 2, 12} = 6.93 \Rightarrow F = 7.95 \sim F_{2, 12}$

4. 임계치 계산 : $F_{0.01, 2, 12} = 6.93$

$$6.93 < 7.95$$

H_0 기각

문제

다음 부산부산물에서 $\alpha = 0.10$ 일 때 평균의 동질성에 대한 F-검정을 설명하라.

| 요인 | 제곱합 | 자유도 |
|----|-----|-----|
| 처리 | 104 | 5 |
| 에러 | 109 | 20 |

| 요인 | 제곱합 | 자유도 | 평균제곱 | F값 |
|------------|-------|------|--------|---------|
| 집단 간 (SSC) | 104 | 5 | (20.8) | (3.816) |
| 집단 내 (SSE) | 109 | 20 | (5.45) | |
| 총계 (SST) | (213) | (25) | | |

$$SS_T = SS_C + SS_E = 104 + 109 = 213$$

$$df_T = df_C + df_E = 5 + 20 = 25$$

$$MS_C = SS_C / df_C = 104 / 5 = 20.8$$

$$MS_E = SS_E / df_E = 109 / 20 = 5.45$$

$$F = MS_C / MS_E = 20.8 / 5.45 = 3.816$$

$$\sim F_{5, 20}$$

*가설 검정

1. 가설 설정 : H_0 : 평균이 동일하다 . H_1 : not H_0

2. 유의수준 설정 : $\alpha = 0.10$

3. 검정통계량 계산 : $F = 3.816$

4. 임계치 계산 : $F_{5, 20, 0.10} = 2.158$

$$2.158 < 3.816$$

$\therefore H_0$ 기각