학습계획서

팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

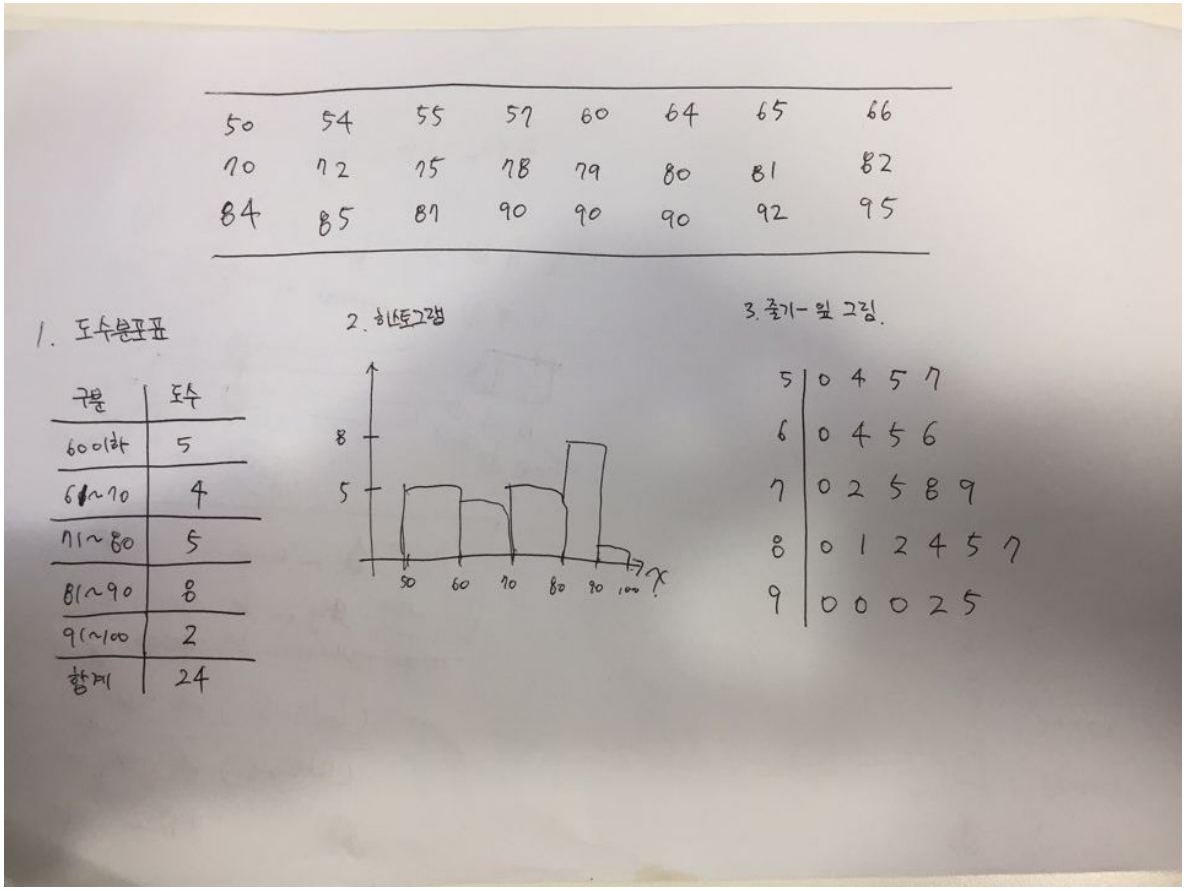
일정	발제자	주제	주요내용
1일차 (5 / 27)	서호성	Big Data Learning Packet : Basics of Statistics	1. data form 2. random variable & distribution 1 3. random variable & distribution 2
2일차 (5 / 28)	최홍용	Big Data Learning Packet : Basics of Statistics	4. normal distribution 5. sampling distribution & central limit theorem 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)	이현경	Big Data Learning Packet : Basics of Statistics	10. simple linear regression 11. analysis of variance
5일차 (5 / 31)	최홍용	Computational Thinking and Data Science	Introduction and Optimization Problems
6일차 (6 / 3)	서호성	머신러닝을 위한 Python 워밍업	Pythonic Code
7일차 (6 / 4)	최홍용	머신러닝을 위한 Python 워밍업	Data Structure - Collections && Linear Algebra - Vector
8일차 (6 / 5)	최홍용	머신러닝을 위한 Python 워밍업	Linear Combinations and Span && Linear Independence
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 아이들	구성원	서호성, 최홍용, 이현경

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	31	61	73
122	45	69	52						

(1) 이 표본에서 고장 날 때까지 걸린 시간의 평균을 구하라.

$$m = \frac{\sum x_i}{n} = \frac{(44+48+\cdots+52)}{24} = 57.79$$

(2) 고장 날 때까지 걸린 시간의 표준편차를 구하라.

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(44-57.79)^2+\cdots+(52-57.79)^2}{23}} = 20.51$$

1. 세 명의 학생이 각각 백화점에서 구두나 운동화 중 하나를 산다.
서로의 구매에 영향을 받지 않고, 오두 반반의 가능성을 가지고 결정한다.
여기서 확률변수 X를 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

- 확률분포표

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

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

(1) C의 값은 얼마인가?

$$1 = \int_{-\infty}^{\infty} f(x)\,dx$$

$$= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{2} f(x)\,dx + \int_{2}^{\infty} f(x)\,dx$$

$$= \int_{0}^{2} C(4x-2x^2)\,dx = C\left[2x^2 - \frac{2}{3}x^3\right]_0^2 = \frac{8}{3}C.$$

$$\therefore \quad C = \frac{3}{8}$$

(2) $P(X>1)$을 구하라.

$$P(X>1) = \int_1^{\infty} f(x)\,dx = \int_1^2 f(x)\,dx + \int_2^{\infty} f(x)\,dx$$

$$= \int_1^2 f(x)\,dx = \frac{3}{8}\left[2x^2 - \frac{2}{3}x^3\right]_1^2$$

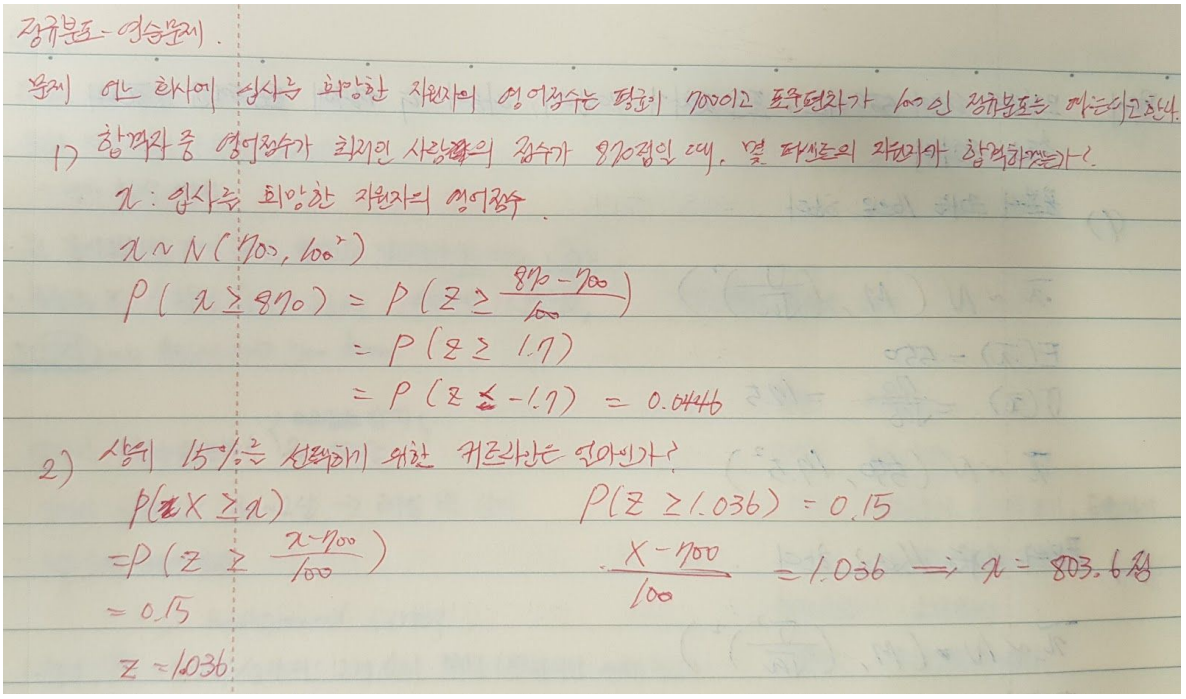
$$= \frac{3}{8}\left\{\left(8 - \frac{16}{3}\right) - \left(2 - \frac{2}{3}\right)\right\} = \frac{1}{2}$$

학습 정리

팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

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

주요 내용 요약



표본분포와 중심극한정리 - 연습문제

문제 어느 도시 근로자의 수입은 평균 250만원이고, 표준편차는 50만원이라고 한다.

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

단위 : 만원

x : 도시 근로자의 수입.

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

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

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

(2) $P(\bar{x} > 260$만원$)$은 얼마인가?

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

$= P(Z > 2) = 0.0228$

문제 모집단의 평균 550이고, 표준편차가 70일 때, 다음의 각 경우에 표본평균 $\bar{x}$의 분포를 구하라.

(1) 표본의 크기 16으로 하라.

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

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

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

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

(2) 표본의 크기 100으로 한다.

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

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

$\sigma(\bar{x}) = \frac{70}{\sqrt{100}} = 5.53$

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

통계적 추론 - 연습문제.

문제 울산시에 거주하는 성인 1인당 신용카드 보유 개수를 추정하기 위해 25명을 무작위로 추출하여 조사한 결과 1인당 평균 7.3개. 표준편차는 2.6개였다. 울산시민 1인당 평균 신용카드 보유 개수에 대한 95% 신뢰구간을 구하시오.

$n = 25$
$\bar{x} = 7.3$
$S = 2.6$

CI $\bar{x} \pm t^* \frac{S}{\sqrt{n}}$

$7.3 \pm 2.797 \cdot \frac{2.6}{\sqrt{25}}$

$= (5.846, 8.754)$

$5.846 \leq \mu \leq 8.754$.

문제 개론 컴퓨터의 시분할시스템을 조사한 결과 편집명령에 대한 반응시간이 표준편차가 25일의 초인 정규분포를 따르는 사실을 알았다. 새로운 운영체제가 설치되었다. 이 시스템에 대한 평균 반응시간을 추정하고 싶다. 이 시스템의 반응시간도 표준편차가 25일의 초인 정규분포를 따른다고 가정할 때, 모평균에 대한 95% 신뢰구간의 폭이 최대 10이 되게끔 추정하려면 표본크기는 얼마쯤으로 정해야 할까?

$\sigma = 25$

$l = 2 \times \frac{\sigma}{\sqrt{n}} \leq 10$

$2 \times 1.96 \times \frac{25}{\sqrt{n}} \leq 10$.

$\left(\frac{2 \times 1.96 \times 25}{10}\right)^2 \leq n$. $\longrightarrow$ $96.04 \leq n$.

$\therefore n = 97$

학습 정리

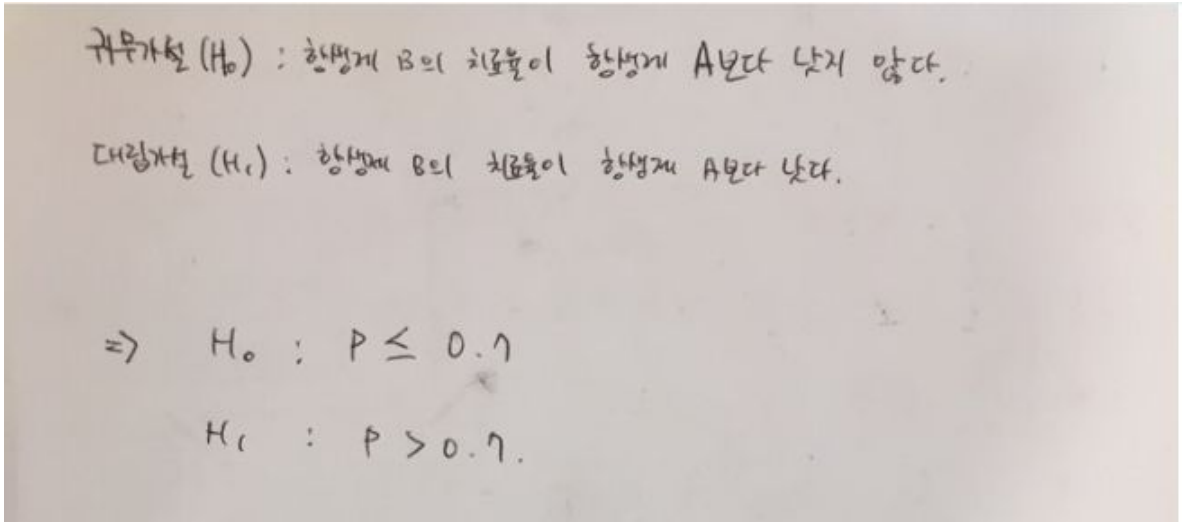
팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

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 X라 하자. 항생제 B의 정확한 치료율 p의 값은 전혀 모르지만, 표본에서 치료율은 X/100으로 추정된다. 이러한 경우에 적합한 가설은 무엇인가?



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 X로 놓는다.

(1) 모집단의 지지율을 P라 할때, 지지율이 달라졌는지에 대한 가설을 세워라.

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

(2) 만일 $X \leq 2$ or $X \geq 8$ 이면, H_0을 기각한다.
이때, 제 1종 오류를 범할 확률 α를 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 α=0.01의 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:

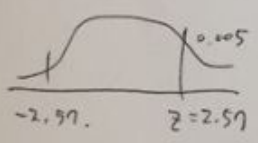
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{71-75}{15/\sqrt{90}} = -2.53.$$

Sol 1) p-value 이용

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

Sol 2) z값 비교

$\alpha = 0.01$



채택영역 : $-2.57 < z < 2.57$.

z값 -2.53 이므로 H_0 채택.

Sol 3) 신뢰구간 이용.

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

$\mu = 75$가 범위안에 포함되기 때문에 H_0 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 α=0.01에서 검정하라.

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

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

(2) 유의수준 결정 : $\alpha = 0.01$, 양측검정

(3) 검정통계량 계산 :

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

(4) p-value 계산 :

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

(5) p-value 와 유의 수준을 비교

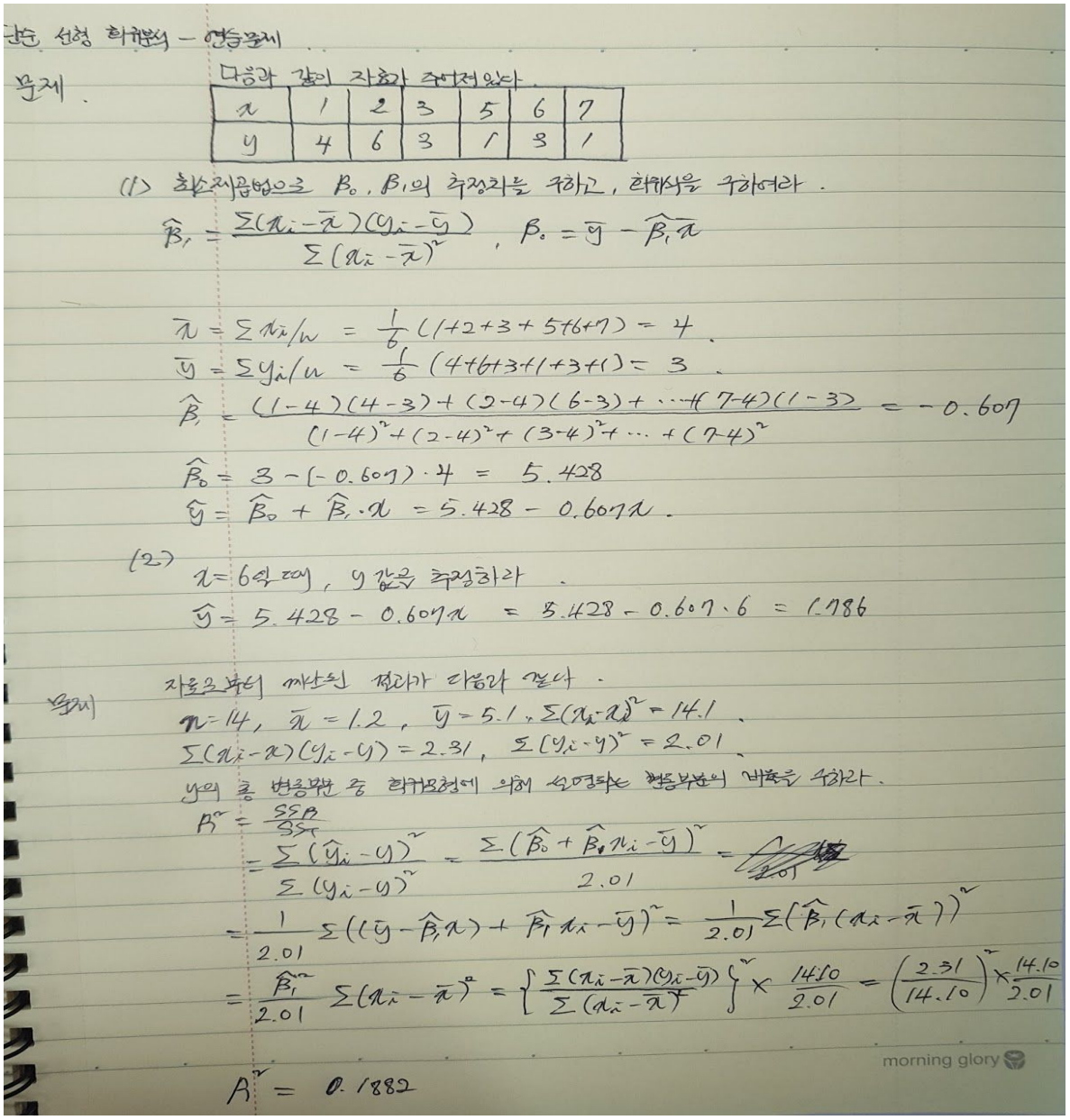
$0 < 0.01$ => H_0 기각

학습 정리

팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

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

주요 내용 요약



분산분석 - 영승율씨.

문제 A 할인점에서는 세 개의 다른 진열 방식에 대해 몇 차례 판매 실험을 해 본 결과 다음과 같은 데이터를 얻었다. 다음 분산분석표을 완성하고 진열방식에 따라 판매량이 같은지 $\alpha = 0.01$ 수준에서 검정하여라.

10안	20안	30안
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
집단간(SS_t)	994.93	(2)	(472.465)	(7.95)
집단내(SS_E)	712.8	(12)	(59.4)	
합계(SS_T)	1657.73	(14)		

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

$df_t = 3-1 = 2$

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

$MS_t = SS_t / df_t = 994.93/2 = 472.465$

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

$F = MS_t / MS_E = 472.465/59.4 = 7.95$

분산분석(ANOVA)

* 가설 검정

1. 가설 설정 : $H_0 : \mu_1 = \mu_2 = \mu_3$, H_1 : 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 분산분석표에서 $\alpha = 0.10$일 때 평균의 등질성에 대한 F-검정을 시행하라.

요인	제곱합	자유도
처리	104	5
오차	109	20

요인	제곱합	자유도	평균제곱	F값
집단 간 (SS_t)	104	5	(20.8)	(3.816)
집단 내 (SS_E)	109	20	(5.45)	
합계 (SS_T)	(213)	(25)		

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

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

$MS_t = SS_t / df_t = 104/5 = 20.8$

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

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

$\sim F_{5,20}$

*가설검정

1. 가설 설정 : H_0 : 평균이 동일하다. H_1 : not H_0
2. 유의수준 결정 : $\alpha = 0.10$
3. 검정통계량 계산 : $F = 3.816$
4. 임계치 계산 : $F_{5,20,0.10} = 2.158$

$2.158 < 3.816$

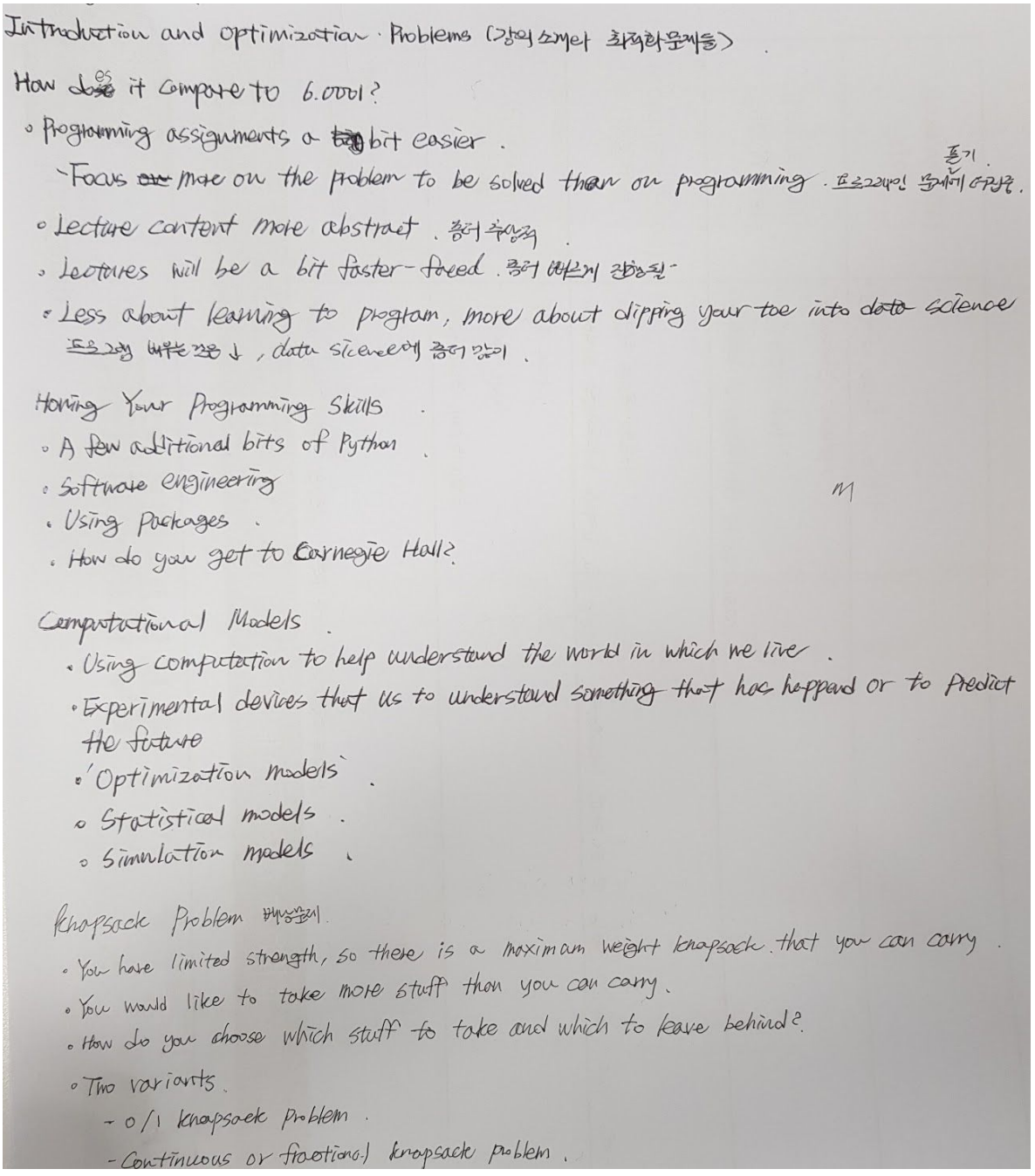
$\therefore H_0$ 기각.

학습 정리

팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

일정	발제자	주제
(5 / 31)	최홍용	Computational Thinking and Data Science

주요 내용 요약



Introduction and optimization · Problems (강의소개와 최적화문제들)

How does it compare to 6.0001?

- Programming assignments a ~~bag~~ bit easier.
  - Focus ~~on~~ more on the problem to be solved than on programming. 프로그래밍 문제에 초점.
- Lecture content more abstract. 좀더 추상적.
- lectures will be a bit faster-paced. 좀더 빠르게 진행됨.
- Less about learning to program, more about dipping your toe into data science
  프로그램 배우는 것 ↓, data science에 좀더 깊이.

Honing Your Programming Skills

- A few additional bits of Python
- Software engineering
- Using Packages
- How do you get to Carnegie Hall?

Computational Models

- Using computation to help understand the world in which we live.
- Experimental devices that us to understand something that has happend or to predict the future
- Optimization models
- Statistical models
- Simulation models

Knapsack Problem 배낭문제

- You have limited strength, so there is a maximum weight knapsack that you can carry.
- You would like to take more stuff than you can carry.
- How do you choose which stuff to take and which to leave behind?
- Two variants
  - 0/1 knapsack problem.
  - Continuous or fractional knapsack problem.

0/1 knapsack problem, Formalized?

- Each item is represented by a pair, <value, weight>
- The knapsack can accommodate items with a total weight of no more than W.
- A vector, I, of length n, represents the set of available items. Each element of the vector is an item.
- A vector, V, of length n, is used to indicate whether or not items are taken. If V[i]=1, item I[i] is taken. If V[i] = 0, item I[i] is not taken.
- Find a V that maximizes

$$\sum_{i=0}^{n-1} V[i] * I[i].value$$

subject to the constraint that

$$\sum_{i=0}^{n-1} V[i] * I[i].weight \leq W$$

Brute Force Algorithm

- 1. Enumerate all possible combinations of items. That is to say, generate all subsets of the set of items. This is called the <power set>
- 2. Remove all of the combinations whose total wits exceeds the allowed weight.
- 3. From the remaining combinations choose any one whose value is the largest.

Often Not Practical

- How big is power set?
- Recall
  - A vector, V, of length n, is used to indicate whether or not items are taken. If V[i] = 1, item I[i] is taken. If V[i] = 0, item I[i] is not taken.
- How many possible different values can V have?
  - As many different binary numbers as can be represented in n bits.

Greedy Algorithm a practical Alternative

- while knapsack not full put "best" available item in knapsack.
- But what does best mean?
  - Most valuable
  - Least expensive
  - Highest value/units.

학습 정리

팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

일정	발제자	주제
(6 / 3)	서호성	Pythonic Code

주요 내용 요약

Pythonic Code

- Split & Join
- List Comprehension
- Enumerate & Zip
- Lambda
- Map
- Reduce
- Asterisk

학습 정리

팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

일정	발제자	주제
(6 / 4)	최홍용	Data Structure - Collections && Linear Algebra - Vector

주요 내용 요약

Data Structure - Collections

- Deque
- Counter
- OrderedDict
- Defaultdict
- Namedtuple

Linear Algebra - Vector

- Vector intro for Linear Algebra
- Real Coordinate Space
- Adding Vectors
- Multiplying a Vector by a Scalar
- Unit Vector Notation
- Parametric Representations of Lines

학습 정리

팀	호성이와 아이들	구성원	서호성, 최홍용, 이현경

일정	발제자	주제
(6 / 5)	최홍용	Linear Combinations and Span && Linear Independence
주요 내용 요약		

Linear Combinations and Span

- Linear Combination
- Span

Introduction to Linear Independence

- Position Vector
- Linear Independence Set
- Linear Independence iff(if and only if)

Introduction to Linear Independence

- Linear Independence iff(if and only if)