

Homework2

(소수점은 넷째 자리에서 반올림하여 셋째 자리까지 구하시오.)

(7번 문제에 파이썬 실습 해당 부분을 제외하고는 전부 풀이까지 작성해주세요.)

1. 연세 반도체 공장에서 Etch 공정 시 일정 시간 동안 막질을 얼마나 제거하는지에 대한 Etch Rate를 측정했다. 아래는 6개의 샘플 Wafer에 대한 Etch Rate를 측정한 데이터이다. (단, Etch Rate는 정규분포를 따른다.)

Wafer	A	B	C	D	E	F
Etch Rate(A/min)	201	192	207	197	203	205

- (1) Etch Rate 모분산의 99% 양측 신뢰구간(confidence interval)을 구하여라.

$$\bar{X} = 200.83, n=6, \text{자릿수} = 5, S^2 = \frac{0.029 + 11.969 + 38.069 + 14.669 + 4.709 + 11.389}{5} = \frac{80.834}{5} = 16.167, S = 4.021$$

신뢰수준 = 0.99, σ^2 unknown 이므로 t분포를 사용
 $\alpha = 0.01, \frac{\alpha}{2} = 0.005, \chi^2_{0.005}(자릿수:5) = 16.153, \chi^2_{0.995}(자릿수:5) = 0.554$

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad \frac{5 \times 16.167}{16.153} < \sigma^2 < \frac{5 \times 16.167}{0.554}$$

$$4.924 < \sigma^2 < 145.168$$

- (2) 새로운 샘플 Wafer의 Etch Rate에 대한 95% 예측구간(prediction interval)을 구하여라.

$$\alpha = 0.05, \frac{\alpha}{2} = 0.025, t_{0.025} = 2.571$$

$$\bar{X} - t_{0.025} S \sqrt{1 + \frac{1}{n}} < \mu < \bar{X} + t_{0.025} S \sqrt{1 + \frac{1}{n}}$$

$$200.83 - 2.571 \times 4.021 \sqrt{1 + \frac{1}{6}} < \mu < 200.83 + 2.571 \times 4.021 \sqrt{1 + \frac{1}{6}}$$

$$200.83 - 15.354 < \mu < 200.83 + 15.354$$

$$185.476 < \mu < 216.184$$

2. 연세 반도체는 TSV 에칭 공정을 통하여 Wafer 에 깊은 마이크로 홀을 만든다. 아래는 공정 후 6 개의 Wafer 를 조사하였을 때의 홀의 깊이를 나타낸 표이다. 단, 홀의 깊이는 정규분포를 따른다.

Wafer 샘플	A	B	C	D	E	F
홀의 깊이 (um)	26.1	25.4	24.9	24.4	23.9	24.8

- (1) 홀의 깊이 모평균의 95% 양측 신뢰구간(confidence interval)을 구하여라.

$$\bar{X} = 24.917, S = 0.768, n = 6, 자유도 = 5$$

신뢰구간 = 0.95, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, 정규분포는 따르므로 σ^2 는 unknown 이므로 t분포 활용

$$\bar{X} - t_{0.025} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0.025} \frac{S}{\sqrt{n}}, \quad t_{0.025} = 2.571$$

$$24.11 < \mu < 25.723$$

- (2) 기존의 6개 Wafer 이외에 추가적으로 34개의 Wafer를 조사하였을 때, 홀의 깊이 평균은 25.87이고, 표준편차는 0.69가 나왔다. (1)의 경우와 어떤점이 다른지 설명하고, 95% 양측 신뢰구간(confidence interval)을 구하여라.

(1)과 다른점 : n이 30 이상이라서 z분포를 활용 할 수 있다.

$$\text{신뢰구간} = 0.95, \alpha = 0.05, \frac{\alpha}{2} = 0.025, \quad Z_{0.025} = 1.96$$

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \bar{X} + Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \quad t_{0.025} = 2.035 \text{ (자유도: 33)}$$

$$25.638 < \mu < 26.102$$

$$\hookrightarrow 25.629 < \mu < 26.111$$

z분포로 구한 값과 비슷!

3. 연세 반도체에서 두 가지 다른 공정(A와 B)을 사용하여 반도체 칩의 불량률을 비교하고자 한다. A 공정에서는 1000개의 칩을 조사하여 20개가 불량이었고, B 공정에서는 1200개의 칩을 조사하여 35개가 불량이었다. 이 정보를 이용하여 두 공정 간 반도체 칩 불량률의 차이에 대한 99% 신뢰 구간을 구하여라.

$$\text{A공정} \quad \hat{p}_1 = 0.02, \quad n\hat{p}_1 = 20, \quad n\hat{q}_1 = 980, \quad n\hat{p}_1 \geq 10, \quad n\hat{q}_1 \geq 10 \text{ 정규분포를 따름}$$

$$\text{B공정} \quad \hat{p}_2 = 0.029, \quad n\hat{p}_2 = 35, \quad n\hat{q}_2 = 1165, \quad n\hat{p}_2 \geq 10, \quad n\hat{q}_2 \geq 10 \text{ 정규분포를 따름}$$

$$\text{신뢰구간} \quad 99\%, \alpha = 0.01, \frac{\alpha}{2} = 0.005, \quad Z_{0.005} = 2.58$$

$$(\hat{p}_1 - \hat{p}_2) - Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$-0.009 - 2.58 \sqrt{\frac{0.02}{1000} + \frac{0.028}{1200}} < p_1 - p_2 < -0.009 + 2.58 \sqrt{\frac{0.02}{1000} + \frac{0.028}{1200}}$$

$$-0.026 < p_1 - p_2 < 0.008$$

4. 연세반도체는 A 장비를 이용하여 CMP 공정을 진행한다. 아래는 11 개의 웨이퍼에 대해 공정 후 웨이퍼 기판 중심부의 material removal rate 를 나타낸 표이다. material removal rate 측정 결과는 정규분포를 따른다.

A 장비											
Material removal rate	2623	2616	2614	2610	2617	2621	2616	2619	2613	2621	2618

- (1) A 장비의 평균 material removal rate 가 2620 인지 아닌지 유의수준 0.05 내에서 검정하시오.

$$\bar{X} = 2617.091, S = 3.859, n = 11, \text{자유도} = 10, \alpha = 0.05, \frac{\alpha}{2} = 0.025, \sigma^2 \text{ unknown, } t \text{ 분포 활용}$$

$$P\left(-t_{\frac{\alpha}{2}, n-1} < \frac{(\bar{X} - \mu)}{S/\sqrt{n}} < t_{\frac{\alpha}{2}, n-1}\right), t_{0.025, 10} = 2.228$$

$$P\left(-2.228 < -2.15 < 2.228\right) \text{ 해당 범위에 포함하므로, 귀무가설을 기각합니다. 2620 이 아닐 수 있다.}$$

$$\hookrightarrow p\text{-value} = 0.0414 \quad 0.05 \text{ 보다 크다} \quad \uparrow$$

(표미분으로 계산)

- (2) A 장비의 모분산은 7 이라고 가정할 때, 평균 material removal rate 가 2615 인지 아닌지 유의수준 0.1 내에서 검정하시오.

$$P\left(-Z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha. \text{ 모분산을 알고있으므로, } Z \text{ 분포 활용.}$$

$$\mu = 2615, \bar{X} = 2617.091, \sigma = 7, n = 11, \alpha = 0.1, \frac{\alpha}{2} = 0.05, Z_{0.05} \approx 1.65$$

$$P\left(-1.65 < 0.991 < 1.65\right) \text{ 해당 범위에 포함하므로, 귀무가설을 기각 할 수 없습니다, 2615 인 가능성이 있다.}$$

$$\hookrightarrow 0.322 \quad 0.1 \text{ 보다 크다} \quad \uparrow$$

(표미분 계산)

5. 연세반도체는 에칭(etching) 공정에서 두 가지 머신 러닝 모델 A, B를 통하여 CD에 대한 가상 계측(virtual metrology: VM)시스템을 도입하고자 한다. A, B의 모델은 아래 표와 같은 테스트를 거쳤다.

	A 모델	B 모델
Test 횟수	15	13
표본 평균	36	33
표본 표준편차	4	3

Data Science 팀장은 표 결과를 보고 모델 A, 모델 B의 오차 모평균이 비슷한 수준이라면 B 모델 대신 다른 모델과 비교 테스트를 진행할 생각이다. 유의 수준 0.05로 가설검정을 실시하여 다른 모델을 탐색해야 할 지 서술하시오. (단, A 모델과 B 모델의 테스트 결과는 정규분포를 따른다.)

귀무가설: 차이가 없다, 대립가설: 차이가 있다
 A, B 모델 정규분포를 따르고, σ unknown 이므로 t-분포 활용

$$P(-t_{\frac{\alpha}{2}, n} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < t_{\frac{\alpha}{2}, n}) \quad , \quad V = \frac{(S_1^2/n_1 + S_2^2/n_2)}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}} = 25.526$$

$\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $t_{0.025, n=25.526} \approx 2.058$

$P(-2.058 < 2.262 < 2.058)$, **범위안에 들어오지 못했으므로 두 모델 간 차이가 있다**
다른 모델 Test 까지 생각해 볼다

6. 다음 주어진 데이터 'Line_result.csv'는 A line과 B line에서 생산된 동일한 제품에 대한 품질 계측값이다. A line과 B line의 품질 계측값 평균이 차이가 있는지를 유의수준 0.05로 가설검정을 실시하여라. Python 코드를 통해 가설 검정을 실시한 결과를 사진으로 첨부하고, 직접 풀어서 결과를 비교하시오.

Line A $\bar{X}_1 = 42.745$, $S_1 = 9.337$, $n_1 = 50$
 Line B $\bar{X}_2 = 35.089$, $S_2 = 4.376$, $n_2 = 50$
 $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $t_{0.025}(\text{기분: } 49) \approx 2.009$

$P(-2.009 < \frac{7.656}{1.458} < 2.009)$

$P(-2.009 < 5.251 < 2.009)$

범위를 벗어났으므로 귀무가설 기각한다
(두 line은 차이가 있다)

< Python 코드 >

```

import pandas as pd
from scipy import stats

In [4]: data = pd.read_csv('exercises1_data/Line_result.csv')

In [15]: data.describe()

   lineA  lineB
count  50.000000  50.000000
mean   42.745361  35.088904
std     9.336688   4.37635
min    25.403299  21.90274
25%    36.390209  32.59443
50%    42.658548  35.230401
75%    48.362756  37.935370
max    63.522762  42.823218

In [19]: 1. t_statistic, p_value = stats.ttest_ind(Line_A, Line_B)
2.
3. # 결과 출력
4. print(f"t-검정 통계량: {t_statistic:.4f}")
5. print(f"p-value: {p_value:.4f}")
6.
7. # 결과 해석
8. alpha = 0.05
9. if p_value < alpha:
10.     print("두 공장의 생산량에는 차이가 있다. 귀무가설을 기각합니다.")
11. else:
12.     print("두 공장의 생산량에는 차이가 없다. 귀무가설을 채택합니다.")

t-검정 통계량: 5.2514
p-value: 8.88e-06
두 공장의 생산량에는 차이가 있다. 귀무가설을 기각합니다.

```