# 데이터 분석 전문가 (ADP) 실기시험
# 기출유사 통계문제 모음.zip

## 들어가며

본 문서는 지난 2년 동안 **슬기로운 통계생활 데이터분석 전문가 (ADP) 실기반**을 운영하면서 학생들의 후기들을 참고하여 제작되었습니다. ADP 실기 시험에서 물어보고자 하는 것을 최대한 반영하여 **비슷하게 만들어낸 가상의 문제집**이니 이점 유의하시기 바랍니다.

- 문제에 사용된 데이터 역시 캐글이나 R, Python에서 제공하는 비슷한 구조 데이터를 차용하거나 시뮬레이션을 통하여 만들었습니다. **실제 기출이 아님**에 유의하세요! 문제 풀이가 매끄럽지 않을 수 있습니다.
- 각 회차별 문제에는 일부러 토픽은 적지 않았습니다. 문제를 보고 어떤 토픽에 관련한 문제인지 판단하는 것도 문제 푸는데 중요한 능력이죠?
- 시계열 기출의 경우 머신러닝 문제.Zip에서 다룹니다.

자료가 여러분의 공부에 도움이 되시길 바랍니다. 자료는 계속해서 업데이트 될 예정입니다. 문제 오류나 오타 발견 시 `statisticsplaybook@gmail.com` 으로 메일주세요! 데이터 분석 공부는 슬기로운 통계생활! :)

## 제 20회

### 문제 3

`boston.csv` 파일에는 미국 보스턴의 주택 가격과 관련된 다양한 환경 정보가 기록되어 있습니다.

1) Boston 데이터를 8:2의 비율로 분할한 후, 선형 회귀 모델을 학습시키고 결정계수와 RMSE 값을 계산하시오.

2) 같은 데이터 분할 비율을 사용하여 릿지 회귀 모델을 적용하고, 그 성능을 결정계수와 RMSE로 평가하시오.

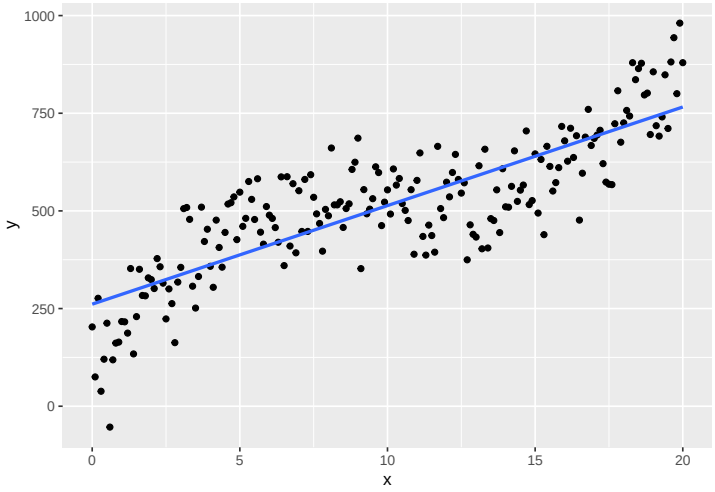
3) 데이터를 동일하게 8:2로 나눈 후, 라쏘 회귀 모델로 학습을 진행하고 결과의 결정계수와 RMSE 값을 구하시오.

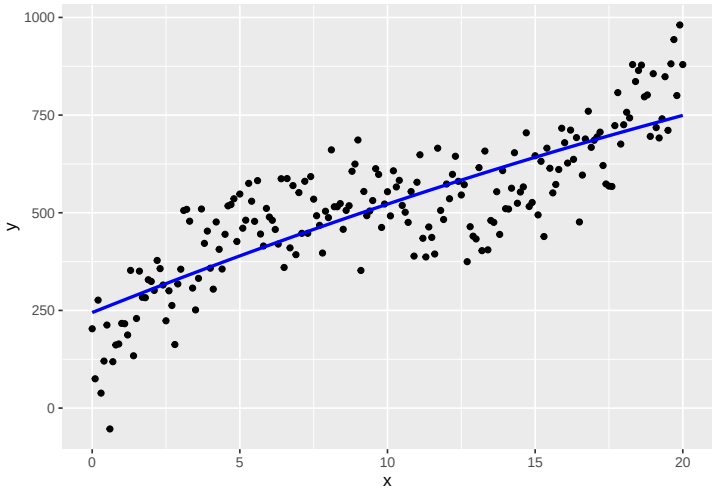
## 제 21회

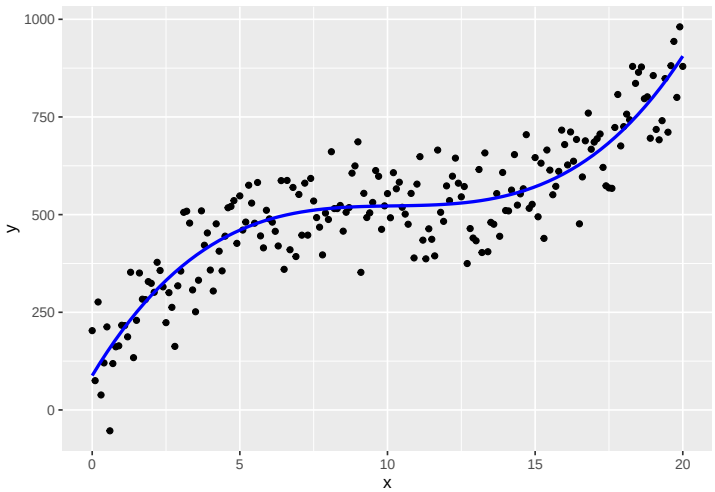
### 문제 3

주어진 데이터를 이용해서 아래에 해당하는 그림을 그리시오. 단, 파란색 직선은 데이터를 가장 잘 표현하는 1차, 2차, 3차 다항 회귀식을 나타낸다.

- 데이터: `data-visualization.csv`



**문제풀이**

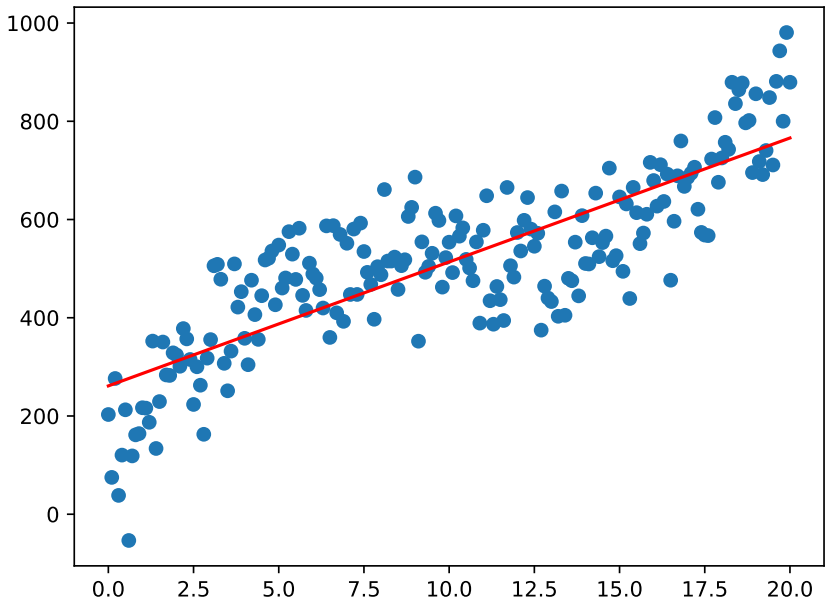
- 데이터 불러오기

```
import pandas as pd
import numpy as np

dat = pd.read_csv('./data/data-visualization.csv')
dat.head()
```

```
##      x           y
## 0  0.0  203.014823
## 1  0.1   75.006443
## 2  0.2  276.360400
## 3  0.3   38.323303
## 4  0.4  120.380259
```

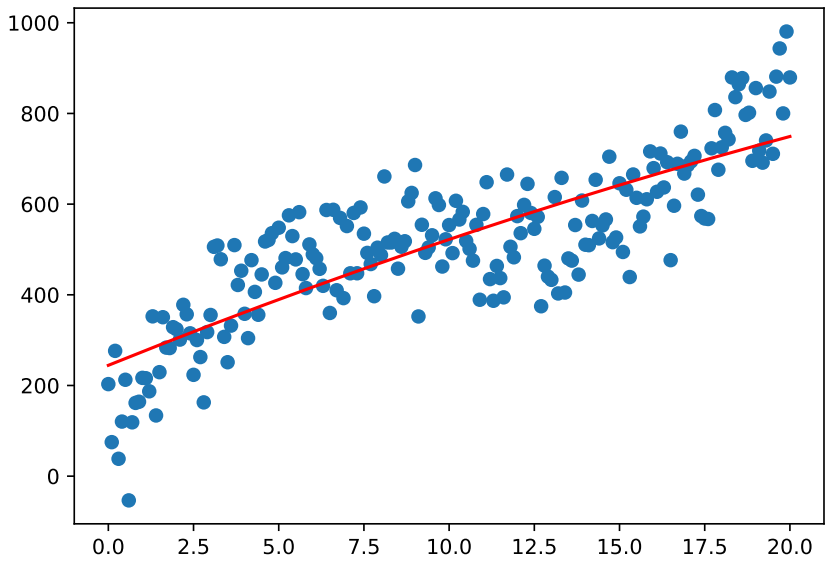
- 1차 선형회귀모형

```
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Linear fit
model = smf.ols(formula='y ~ x', data=dat).fit()
dat['y_est'] = model.predict(dat['x'])

# Plotting
plt.scatter(dat['x'], dat['y'])
plt.plot(dat['x'], dat['y_est'], color='red')
plt.show()
```



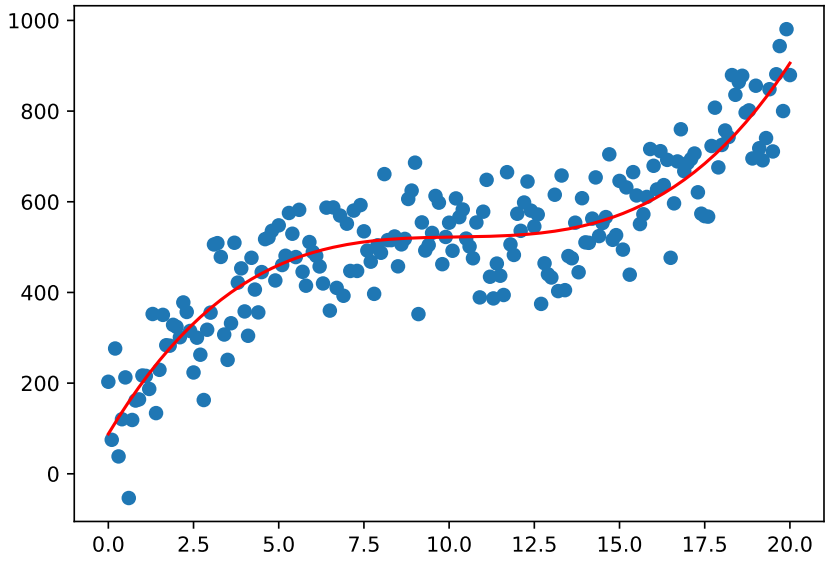

- 2차 다항회귀모형

```
# Quadratic fit
model2 = smf.ols(formula='y ~ np.power(x, 2) + x', data=dat).fit()
dat['y_est2'] = model2.predict(dat['x'])

# Plotting
plt.scatter(dat['x'], dat['y'])
plt.plot(dat['x'], dat['y_est2'], color='red')
plt.show()
```



- 3차 다항회귀모형

```
# Cubic fit
model3 = smf.ols(formula='y ~ np.power(x, 3) + np.power(x, 2) + x', data=dat).fit()
dat['y_est3'] = model3.predict(dat['x'])

# Plotting
plt.scatter(dat['x'], dat['y'])
plt.plot(dat['x'], dat['y_est3'], color='red')
plt.show()
```




### 문제 4

`toothgrowth.csv` 데이터를 사용하여 기니피그의 치아 길이의 성장이 비타민 투여량과 투약 방법에 따라 차이가 있는지 이원배치 분산분석을 수행하시오.

- 데이터 변수 설명

| Variable | Description |
|---|---|
| len | 기니피그의 치아길이 |
| dose | 비타민 C 투여량(mg/day) |
| sup | Orange Juice(OJ), Ascorbic Acid(VC) |

- 데이터 HEAD

표 1.2: toothgrowth.csv 데이터 HEAD

| len | supp | dose |
|---|---|---|
| 4.2 | VC | 0.5 |
| 11.5 | VC | 0.5 |
| 7.3 | VC | 0.5 |
| 5.8 | VC | 0.5 |
| 6.4 | VC | 0.5 |
| 10.0 | VC | 0.5 |

**문제풀이**

- 데이터 불러오기

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
from statsmodels.formula.api import ols
from statsmodels.graphics.api import interaction_plot, abline_plot
from statsmodels.stats.anova import anova_lm

dat = pd.read_csv('./data/toothgrowth.csv')
dat.head()
```

```
##     len supp  dose
## 0   4.2   VC   0.5
## 1  11.5   VC   0.5
## 2   7.3   VC   0.5
## 3   5.8   VC   0.5
## 4   6.4   VC   0.5
```

- 귀무가설 설정하기

  $H_0$ : 비타민 C 투여량에 따른 기니피그의 평균 치아 성장 차이는 존재하지 않는다.

  $H_0$ : 비타민 C 투약 방법에 따른 기니피그의 평균 치아 성장 차이는 존재하지 않는다.

  $H_0$ : 비타민 C 투약 방법과 비타민 C 투여량 사이에는 상호작용 효과가 존재하지 않는다.

- `dose`변수 범주형 변수로 변환

그룹 변수를 `factor`형으로 바꿔준다.

```
dat['dose'] = pd.Categorical(dat['dose'], categories=[0.5, 1, 2], ordered=True)
dat.head()
```

```
##     len supp dose
## 0   4.2   VC  0.5
## 1  11.5   VC  0.5
## 2   7.3   VC  0.5
## 3   5.8   VC  0.5
## 4   6.4   VC  0.5
```

- 모델 적합하기

```
model = ols('len ~ C(supp) * C(dose)', data=dat).fit()
results = sm.stats.anova_lm(model, typ=2)
results
```

```
##                       sum_sq    df          F        PR(>F)
## C(supp)           205.350000   1.0  15.571979  2.311828e-04
## C(dose)          2426.434333   2.0  91.999965  4.046291e-18
## C(supp):C(dose)   108.319000   2.0   4.106991  2.186027e-02
## Residual          712.106000  54.0        NaN           NaN
```

- 모델 해석

- supp의 p-value가 $0.000231$로 매우 작기 때문에 유의수준 $\alpha = 0.05$하에서 귀무가설을 기각한다. 따라서 비타민 C 투약방법에 따른 기니피그의 평균 치아 성장 차이는 유의미하다.

- dose의 p-value가 $< 2e-16$로 매우 작기 때문에 유의수준 $\alpha = 0.05$하에서 귀무가설을 기각한다. 따라서 비타민 C 투여량에 따른 기니피그의 치아 성장 차이는 유의미하다.

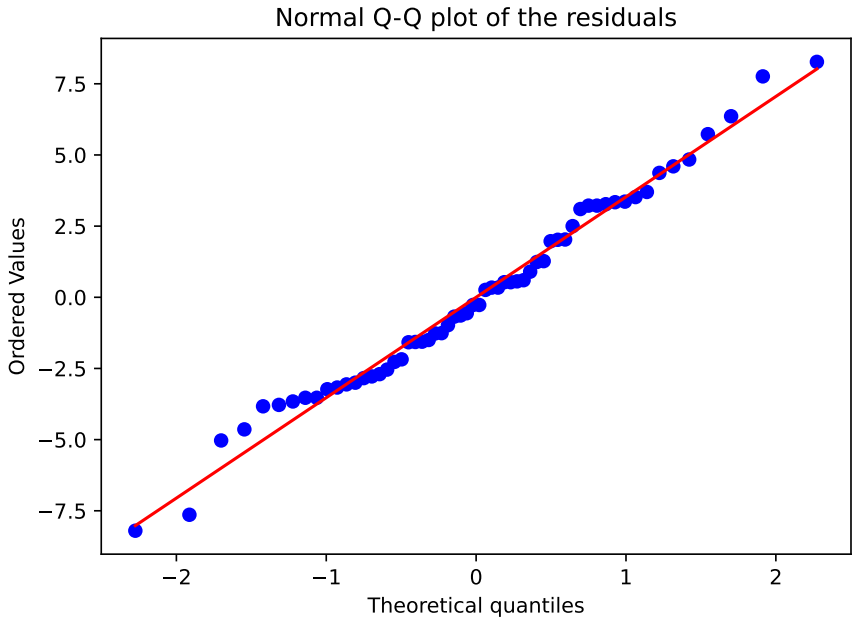
- supp:dose의 p-value가 $0.021860$으로 매우 작기 때문에 유의수준 $\alpha = 0.05$하에서 귀무가설을 기각한다. 따라서 비타민 C 투여량과 비타민 C 투약 방법 사이에는 상호작용 효과가 유의미하다.

모델 해석 결과를 신뢰할 수 있는지 모델 가정 체크를 시작합니다.

- 정규성 가정

첫번째 잔차 그래프에서 잔차 분포의 중심이 $0$이라는 것을 확인할 수 있으며, Normal Q-Q plot를 통하여 시각적으로 정규성을 확인하자.

```
residuals = model.resid
stats.probplot(residuals, plot=plt);
plt.title("Normal Q-Q plot of the residuals")
plt.show()
```




Normal Q-Q plot에서 점들이 대부분 직선 상에 존재하기 때문에 정규성을 만족한다고 볼 수 있다. Shapiro 검정을 통하여 정규정을 검정하자. Shapiro 검정의 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \text{표본들이 정규분포를 따른다.}$$
$$H_1 : \text{표본들이 정규분포를 따르지 않는다.}$$

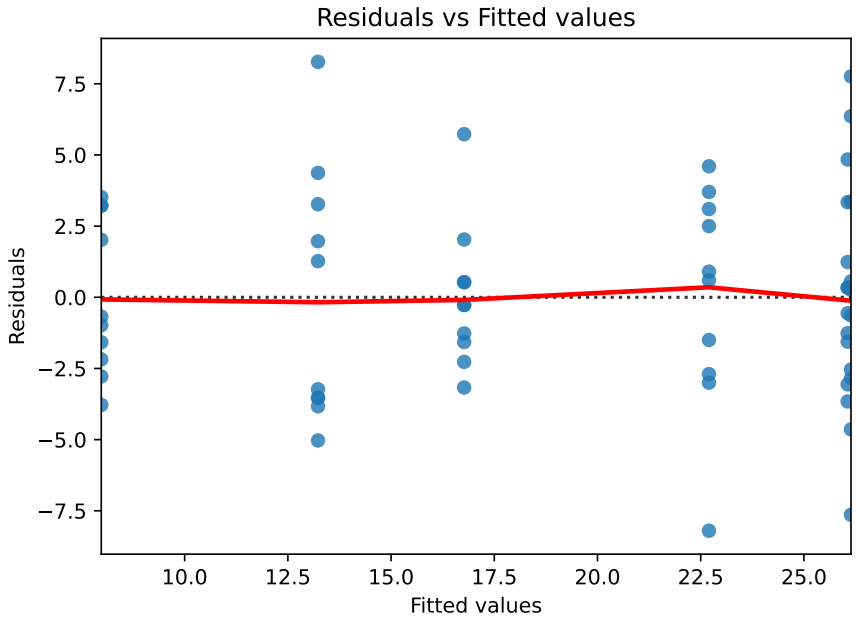
```
stats.shapiro(residuals)
```

```
## ShapiroResult(statistic=0.9849883913993835, pvalue=0.6694232821464539)
```

유의수준 $\alpha = 0.05$하에서 p-value가 $0.6694$로 매우 크기 때문에 $H_0$를 기각할 수 없다. 따라서 정규성 가정을 만족한다.

- 등분산성 가정

```
# Compute residuals
dat['residuals'] = model.resid
dat['fitted_values'] = model.fittedvalues

# Plot residuals
sns.residplot(x=model.fittedvalues, y=model.resid, lowess=True, line_kws={'color': 'red'})
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residuals vs Fitted values')
plt.show()
```




잔차 그래프로 보아 잔차의 값이 $\pm 5$ 사이에서 고르게 분포한다. 즉, 분산이 증가하거나 감소하는 경향이 없으며, 따라서 등분산성 가정을 만족한다고 볼 수 있다. 등분산성을 검정하기 위해서 levene 검정을 실시한다. levene 검정의 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \text{집단 간 분산이 동일하다.}$$
$$H_1 : \text{집단 간 분산이 동일하지 않다.}$$

```
from scipy import stats

# get residuals
dat['residuals'] = model.resid

# split residuals by groups
groups = dat.groupby(['dose', 'supp'])['residuals'].agg(list).tolist()

# conduct Levene's test
stats.levene(*groups, center= "mean")
```

```
## LeveneResult(statistic=1.9401303086614297, pvalue=0.10272977489885697)
```

유의수준 $\alpha = 0.05$ 하에서 p-value가 0.1027로 크기 때문에 귀무가설을 $H_0$를 기각하지 못한다. 즉, 등분산성 가정 만족한다고 판단한다. 따라서, ANOVA 분석 결과를 신뢰 할 수 있다.

## 제 22회

### 문제 1

제품에 금속 재질 함유량의 분산이 1.3을 넘으면 불량이라고 보고 있는데, 제조사별로 차이가 난다고 제보를 받았다. 주어진 회사 제품의 분산에 대해 검정을 수행하시오. (유의수준 5%)

10.67, 9.92, 9.62, 9.53, 9.14, 9.74, 8.45, 12.65, 11.47, 8.62

**문제 2**

슬통 전자는 매일 무작위로 완제품을 선택하여 불량품 유무를 조사한 자료이다. 다음은 지난 20일 동안의 불량품 갯수를 나타낸 데이터이다.

표 1.3: 날짜별 불량품 갯수

| 날짜 | 검사갯수 | 불량갯수 |
|---|---|---|
| 1 | 61 | 4 |
| 2 | 85 | 3 |
| 3 | 75 | 2 |
| 4 | 86 | 4 |
| 5 | 64 | 2 |
| 6 | 96 | 4 |
| 7 | 87 | 5 |
| 8 | 93 | 3 |
| 9 | 67 | 6 |
| 10 | 97 | 7 |
| 11 | 77 | 6 |
| 12 | 88 | 5 |
| 13 | 90 | 8 |
| 14 | 84 | 5 |
| 15 | 65 | 5 |
| 16 | 71 | 3 |
| 17 | 69 | 3 |
| 18 | 66 | 4 |
| 19 | 98 | 5 |
| 20 | 72 | 8 |

위의 데이터를 사용하여 p(불량률) 관리도에 따라 관리중심선, 관리 상한선 및 하한선을 구하시오.

**문제풀이** p 관리도는 불량품이 관찰되는 갯수를 이항분포를 따른다고 생각하고, 전체 불량률과 분산을 추정한 후 3 표준편차 범위를 그린 그래프라고 생각할 수 있다.

- 관리 중심선

전체 제품의 불량율을 구한다.

$$CL := \hat{p} = \frac{\sum_{i=1}^{d} x_i}{\sum n_i}$$

여기서 $x_i$는 불량품 검출 갯수, $n_i$은 검사 제품 갯수, $d$는 검사 날짜수를 의미한다.

```
n = np.array([61, 85, 75, 86, 64, 96, 87, 93, 67, 97, 77, 88, 90, 84, 65, 71, 69, 66, 98, 72])
x_i = np.array([4, 3, 2, 4, 2, 4, 5, 3, 6, 7, 6, 5, 8, 5, 5, 3, 3, 4, 5, 8])
p_hat = np.sum(x_i) / np.sum(n)
p_hat
```

```
## 0.05782526712759271
```

- 관리 상한선과 하한선

구한 불량율을 사용하여 표준편차를 추정한 후, 관리중심선에서 3 표준편차 떨어진 선

$$UCL = CL + 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n_i}}$$

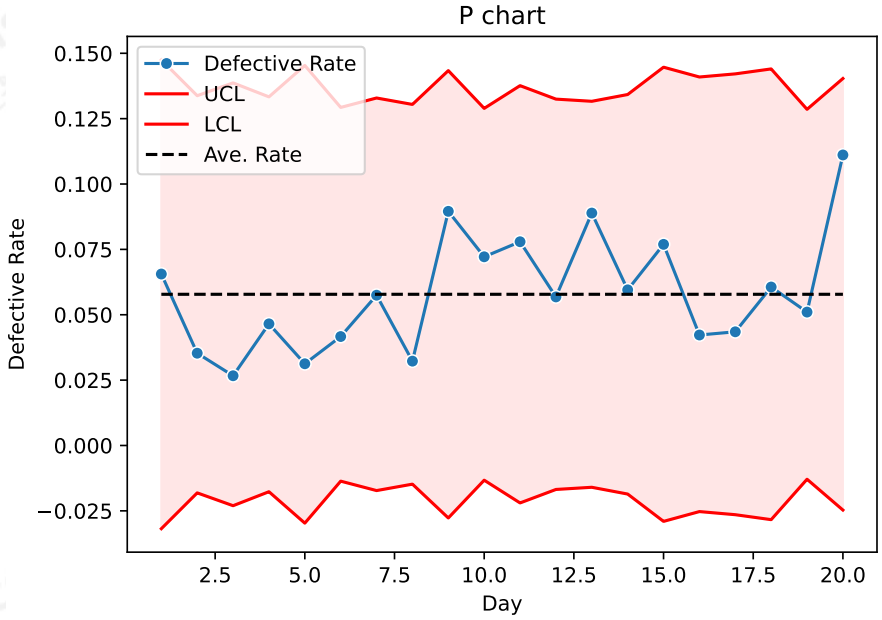
$$LCL = CL - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n_i}}$$

```
ucl = p_hat + 3 * np.sqrt(p_hat * (1 - p_hat) / n)
lcl = p_hat - 3 * np.sqrt(p_hat * (1 - p_hat) / n)
```

- P 관리도

추정한 관리중심선과 관리 상한/하한선을 표시하고, 일별 불량율을 날짜별로 그려준다.

```
day_i = np.arange(1, 21)
p_i = x_i / n

df = pd.DataFrame({
    "Day": day_i,
    "Defective Rate": p_i,
    "UCL": ucl,
    "LCL": lcl,
    "Ave. Rate": [p_hat] * 20
})
sns.lineplot(x="Day", y="Defective Rate", data=df, marker="o", label="Defective Rate")
sns.lineplot(x="Day", y="UCL", data=df, color='red', label="UCL")
sns.lineplot(x="Day", y="LCL", data=df, color='red', label="LCL")
sns.lineplot(x="Day", y="Ave. Rate", data=df, color='black', linestyle='--', label="Ave. Rate")

plt.fill_between(df["Day"], df["LCL"], df["UCL"], color='red', alpha=0.1)
plt.title('P chart')
plt.ylabel('Defective Rate')
plt.show()
```




### 문제 3

슬통이는 두 가지 종류의 빵을 판매하는데, 초코빵을 만들기 위해서는 밀가루 100g과 초콜릿 10g이 필요하고 밀빵을 만들기 위해서는 밀가루 50g이 필요하다. 재료비를 제하고 초코빵을 팔면 100원이 남고 밀빵를 팔면 40원이 남는다. 오늘 슬통이는 밀가루 3000g과 초콜릿 100g을 재료로 갖고 있다. 만든 빵을 전부 팔 수 있고 더 이상 재료 공급을 받지 않는다고 가정한다면, 슬통이는 이익을 극대화 하기 위해서 어떤 종류의 빵을 얼마나 만들어야 하는가?[1]

**문제풀이** $x_1$을 초코빵을 만드는 개수, $x_2$를 밀빵을 만드는 개수로 설정하자. 그렇다면, $x_1, x_2$는 정수값을 가져야하며, 다음과 같은 조건하에서 이익을 최대로 만드는 문제의 해가 된다.

$$Objective: \max_{z} z = 100x_1 + 40x_2$$

$$Constraints: 100x_1 + 50x_2 \leq 3000,$$

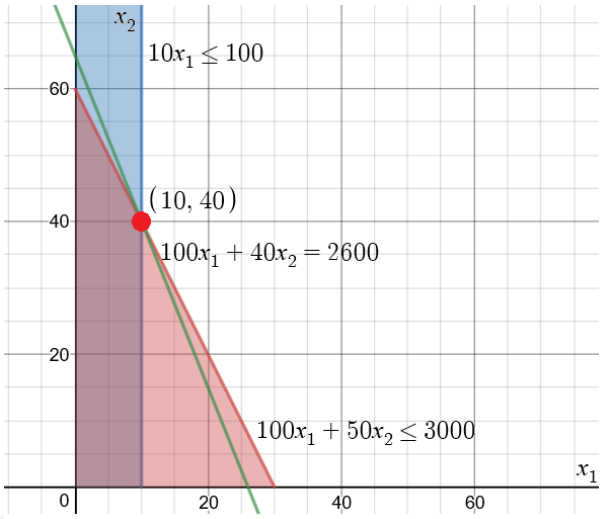
$$10x_1 \leq 100,$$

$$x_1, x_2 \geq 0.$$

위의 문제를 다음과 같은 Python코드를 통하여 풀 수 있다.

```
from scipy.optimize import linprog

#       (          )
c = [-100, -40]

#
A = [[100, 50],
     [10, 0]]
```

[1] 해당 문제는 위키피디아의 선형계획법 페이지에서 가져옴.




```
#
b = [3000, 100]

#
x0_bounds = (0, None)
x1_bounds = (0, None)

#
res = linprog(c, A_ub=A, b_ub=b, bounds=[x0_bounds, x1_bounds], method='highs')

#
print('Optimal value:', -res.fun, '\nX:', res.x)
```

```
## Optimal value: 2600.0
## X: [10. 40.]
```

위의 문제를 그림으로 풀면 다음과 같다.

### 문제 4

구매하는 패턴으로 봐서 두 상품이 연관이 있는지 가설을 세우고 검정하시오.

A, A, A, B, B, A, A, A, A, B, A, B, B, B, A, A, A, A, B, B, A, A, A, B, B,

**문제풀이** 주어진 수열이 무작위로 발생된 데이터인지 그렇지 않은 데이터인지를 검정하는 run test를 수행한다.

- 가설 설정

$H_0$ : 관측치는 randomness를 만족한다.$H_1$ : 관측치는 randomness를 만족하지 않는다.

- 검정 수행하기

Runs test의 검정통계량은 귀무가설하에서 표준정규분포를 따른다. Runs test를 수행하는 방법은 `tseries` 패키지의 `runs.test()`를 사용하면 된다.

```
import numpy as np
from statsmodels.sandbox.stats.runs import runstest_1samp

run_sample = np.array(['A', 'A', 'A', 'B', 'B', 'A', 'A', 'A', 'A',
                       'B', 'A', 'B', 'B', 'B', 'A', 'A', 'A', 'A',
                       'B', 'B', 'A', 'A', 'A', 'B', 'B'])

# Running the runs test
run_sample_encoded = np.where(run_sample == 'A', 1, 0)
test_stat, p_value = runstest_1samp(run_sample_encoded, correction=False)

print("Test statistic:", test_stat)
```

```
## Test statistic: -1.2792042981336627
```

```
print("p-value:", p_value)
```

```
## p-value: 0.2008251226951454
```

- 검정결과 해석

유의수준 0.05 하에서 검정통계량 값 -1.2792에 대응하는 p-value는 0.2008로 크기 때문에 귀무가설을 기각할 수 없다. 관측치는 무작위하게 발생된 표본으로 볼 수 있으며, 따라서 두 제품 구매 패턴에 연관성이 없다고 볼 수 있다.

**참고: Runs test 설명** Runs test 검정통계량은 다음과 같습니다.

$$Z = \frac{R - \bar{R}}{s_R}$$

$R$: 연속된 run 수 $n_1$: 데이터에서의 A 갯수 $n_2$: 데이터에서의 B 갯수

$$\bar{R} = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$s_R^2 = \frac{2n_1n_2\left(2n_1n_2 - n_1 - n_2\right)}{\left(n_1 + n_2\right)^2\left(n_1 + n_2 - 1\right)}$$

주어진 데이터에서 연속된 run 수는 10이며, $\bar{R}$와 $s_R^2$은 다음과 같이 계산 할 수 있습니다.

```
r = 10
n1 = np.sum(np.array(run_sample) == "A")
n2 = np.sum(np.array(run_sample) == "B")
r_bar = (2 * (n1 * n2) / (n1 + n2)) + 1
var_r = (2 * (n1 * n2) * (2 * (n1 * n2) - n1 - n2)) / ((n1 + n2)**2 * (n1 + n2 - 1))
test_stat = (r - r_bar) / np.sqrt(var_r)
test_stat
```

```
## -1.2792042981336627
```

## 제 23회

### 문제 1

슬통 회사는 자신들이 만든 진공관 제품의 수명이 1만 시간이라는 광고를 하고 있다. 이에 슬통 회사의 품질 관리팀에서 12개 샘플을 뽑아서 수명을 측정 한 데이터이다. 유의수준 5% 하에서 부호 검정하시오.

1) 연구가설과 귀무가설 작성하시오.

2) 유효한 샘플의 수를 계산하시오.

3) 검정 통계량을 계산하고, 연구가설 채택 여부를 작성하시오.

### 문제 3

학과별 학생들의 평점 분포에 관한 연구 - 사회과학, 자연과학, 공학의 각 학과별 학생들의 입학 성적 분포를 조사하였습니다. 아래 제시된 표는 각 학과의 학생들이 어떤 평점 구간에 속하는지를 나타냅니다. 평점 구간은 50-66, 67-83, 84-100로 나누어져 있으며, 각 셀에는 해당 구간에 속하는 학생의 수가 기록되어 있습니다. 학과와 평점 분포 간의 관계를 검정하여 주세요.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| Score.50.66 | 90 | 35 | 45 | 40 | 50 | 60 |
| Score.67.83 | 50 | 55 | 55 | 50 | 45 | 40 |
| Score.84.100 | 20 | 70 | 60 | 70 | 65 | 60 |

1) 연구의 귀무가설과 대립가설을 명확하게 작성하시오.

2) 학과와 평점이 서로 독립적일 때, 각 셀의 기대 빈도를 계산하시오.

3) 검정 통계량을 계산하고, 연구의 귀무 가설을 기각 할 지 여부를 결정하시오.

## 제 24회

### 문제 1

아래는 슬통 회사의 2023년 10개월 간 광고비, 연구 개발비 및 해당 기간의 판매액 정보에 대한 데이터입니다. 주어진 데이터를 기반으로 다음의 작업을 수행하시오.

표 1.5: 광고비, 연구개발비 및 판매액에 대한 데이터

| 광고비 | 연구개발비 | 판매액 |
|---|---|---|
| 낮음 | 52 | 1322.53 |
| 낮음 | 63 | 824.10 |
| 낮음 | 74 | 1492.06 |
| 낮음 | 81 | 1566.05 |
| 높음 | 96 | 1422.84 |
| 높음 | 112 | 1887.65 |
| 높음 | 127 | 1221.44 |
| 높음 | 135 | 877.59 |
| 높음 | 143 | 1570.82 |
| 높음 | 153 | 1402.99 |

1) 광고비를 가변수화하여 다중 선형 회귀방정식을 구성하시오.
2) 회귀 모형을 검정하시오.

**문제 2**

A 생산라인에서 샘플링된 100개의 제품의 평균 크기는 5.7mm이며, 해당 제품의 표준편차는 0.03입니다. 반면, B 생산라인에서 샘플링된 120개의 제품의 평균 크기는 5.6mm이며, 해당 제품의 표준편차는 0.04입니다.

두 생산라인에서 샘플링된 제품의 평균 크기에 차이가 있는지를 검정하시오.

**문제 3**

Covid-19의 발병률은 1%라고 한다. 다음은 이번 코로나 사태로 인하여 코로나 의심 환자들 1,085명을 대상으로 슬통 회사의 "다잡아" 키트를 사용하여 양성 반응을 체크한 결과이다.

| 키트 \ 실제 | 양성 | 음성 |
|---|---|---|
| 양성 | 370 | 10 |
| 음성 | 15 | 690 |

1) 다잡아 키트가 코로나 바이러스에 걸린 사람을 양성으로 잡아낼 확률을 계산하세요.
2) 슬통 회사에서 다잡아 키트를 사용해 양성으로 나온 사람이 실제로는 코로나 바이러스에 걸려있을 확률을 97%라며, 키트의 우수성을 주장했다. 이 주장이 옳지 않은 이유를 서술하세요.
3) Covid-19 발병률을 사용하여, 키트의 결과값이 양성으로 나온 사람이 실제로 코로나 바이러스에 걸려 있을 확률을 구하세요.

**문제 4**

다음은 슬통 고등학교의 3학년 학생들 중 16명을 무작위로 선별하여 몸무게를 측정한 데이터이다. 이 데이터를 이용하여 해당 고등학교 3학년 전체 남학생들의 몸무게 평균을 예측하고자 한다.

71.2, 62.2, 53.2, 70.1, 65.7, 82.9, 62.9, 82, 68, 67.3, 75.3, 67.9, 77.6, 78.6, 66, 79

단, 슬통 고등학교 3학년 남학생들의 몸무게 분포는 정규분포를 따른다고 가정한다. 또한, 신뢰구간 계산시 아래 값들을 사용하여 계산하시오.

- $Z_{0.05} = 1.65$, $Z_{0.025} = 1.96$, $t_{0.05,16} = 1.745$, $t_{0.025,16} = 2.119$, $t_{0.05,15} = 1.753$, $t_{0.025,16} = 2.131$

1) 모평균에 대한 95% 신뢰구간을 구하시오.

2) 작년 남학생 3학년 전체 분포의 표준편차는 6kg 이었다고 합니다. 이 정보를 이번 년도 남학생 분포의 표준편차로 대체하여 모평균에 대한 95% 신뢰구간을 구하시오.

## 제 25회

### 문제 3

1) 어떤 사람이 갈 때는 시속 4km로 이동하고, 돌아올 때는 시속 5km로 이동하였다. 이 사람의 왕복 여정에서의 평균 속도를 계산하시오.

2) 슬통 회사의 연매출이 연속적으로 2천만원, 3천5백만원, 7천만원으로 증가하였다. 이 기간 동안 매년 연매출은 평균적으로 몇 배로 증가하였는지 계산하시오.

3) 슬통 화장품에서 개발한 신제품 향수 12개의 지속시간을 측정 한 후 표본 표준편차를 계산했더니 9.74분이 나왔다. 신제품의 지속시간 분포의 모분산 추정을 위한 95% 신뢰 구간을 구하시오.

### 문제 4

슬통 제약회사에서 혈압약을 개발하여 20명을 대상으로 약의 효능을 검사한 결과 혈압이 평균 25mmHG 내려갔으며, 표준편차 9.1mmHG으로 계산되었다. 유의수준 5%하에서 가설 검정하여라.

1) 연구가설과 귀무가설을 설정하시오

2) 약이 혈압을 실제로 낮추는 것인지 검정통계량과 가설 채택여부에 대하여 작성하시오.

### 문제 5

슬통 대학교 남학생과 여학생 각각의 축구와 배구에 대한 스포츠 선호도를 조사하였습니다. 아래 제시된 교차표를 참고하여, 여학생 중에서 배구를 선호하는 확률을 구하시오.

표 1.7: 대학생들의 스포츠 선호도

| 성별 | 축구 | 배구 |
|---|---|---|
| 남성 | 45 | 55 |
| 여성 | 35 | 65 |

### 문제 6

슬통 제조회사에서 다음 X, Y, Z공장에서 생산되는 제품 무게의 중앙값이 동일한지 95% 신뢰수준에서 가설 검정을 시행하려고 한다.

표 1.8: 슬통 제조 공장 A, B, C 제품의 무게 측정값

| A | B | C |
|---|---|---|
| 52.48 | 47.68 | 72.33 |
| 49.31 | 47.67 | 63.87 |
| 53.24 | 51.21 | 65.34 |
| 57.62 | 40.43 | 57.88 |
| 48.83 | 41.38 | 62.28 |
| 48.83 | 47.19 | 65.55 |
| 57.90 | 44.94 | 59.25 |
| 53.84 | 51.57 | 66.88 |
| 57.62 | 57.62 | 62.00 |
| 57.90 | 57.90 | |

1) 주어진 측정결과의 혼합표본 순위를 계산하시오. (단, 동점이 있는 경우 평균순위를 사용하시오.)

2) 연구가설과 귀무가설을 설정하시오.

3) 설정한 가설을 크러스칼-윌리스 검정을 사용하여 검정하고, 검정 통계량값, 가설채택 여부 설명하시오.

**문제 7**

슬통이는 순수 현재가치 (Net Present Value; NPV)를 최대화하는 투자 계획을 세우려고 한다. 정해진 예산은 1년차 50억, 2년차 60억, 3년차 80억을 넘지 않는 선에서 포트폴리오를 운영하려고 할 때, 다음의 조건을 만족하면서 현재 가능한 최대 NPV를 달성할 수 있는 최적의 투자안을 구하시오.

- 단, 각 자산은 1개까지만 투자 할 수 있으며, 공매도는 허용하지 않는다.
- 한 번 결정한 투자 포트폴리오는 3년 동안 변하지 않는다.

| | 1년차 | 2년차 | 3년차 | NPV |
|---|---|---|---|---|
| 자산 1 | 23 | 23 | 15 | 30 |
| 자산 2 | 15 | 15 | 12 | 20 |
| 자산 3 | 17 | 25 | 12 | 31 |
| 자산 4 | 16 | 12 | 13 | 42 |
| 자산 5 | 24 | 23 | 17 | 44 |

**문제풀이** 위의 문제는 해가 0과 1로 이루어진 이진수 조건이 있으므로 이를 직접적으로 푸는 파이썬 패키지는 `ortools`에 있습니다. 하지만, 이 패키지는 ADP의 기본적으로 제공되는 패키지에 속하지는 않습니다.

위와 같은 문제는 다음의 코드처럼 주어진 모든 경우의 수를 구한 후, 제약에 걸리는 경우는 제외 시키는 방법으로 구하는 게 훨씬 빠르다.

```
import numpy as np
from itertools import product
```

```python
# Set cost, NPV, budget
costs = np.array([[23, 23, 15],
                  [15, 15, 12],
                  [17, 25, 12],
                  [16, 12, 13],
                  [24, 23, 17]])
npv = np.array([30, 20, 31, 42, 44])
budgets = np.array([50, 60, 80])

# Possible outcome and check the condition
comb = list(product([0, 1], repeat=5))
valid_comb = []
valid_npv = []
for c in comb:
    total_costs = np.dot(c, costs)
    if np.all(total_costs <= budgets):
        valid_comb.append(c)
        valid_npv.append(np.dot(c, npv))

# Get result
max_npv_idx = np.argmax(valid_npv)
optimal_comb = valid_comb[max_npv_idx]
optimal_npv = valid_npv[max_npv_idx]

print('Optimal Plan: ', optimal_comb); print('Maximum NPV: ', optimal_npv)
```

```
## Optimal Plan:  (0, 1, 1, 1, 0)
## Maximum NPV:  93
```

따라서, 최적의 포트폴리오는 2, 3, 4 안에 투자하는 것이고, 이때의 NPV는 93으로 최대가 됩니다.

## 제 26회

### 문제 4

슬통 전구회사는 자사의 제품 생산 라인에서 최근 불량률이 급증했다는 내부 보고서를 받았습니다. 초기 보고서에 따르면 불량률이 약 90%에 달한다고 합니다. 회사의 경영진은 이를 확인하기 위해 독립적인 품질 검사팀에 조사를 의뢰하였습니다. 검사팀은 신뢰도 95%로 불량률을 확인하기 위해 필요한 표본 크기를 계산하려고 합니다. 오차한계가 3% 내외로 허용된다면, 검사팀은 최소 몇 개의 제품을 표본으로 선택해야 하는지 계산하시오.

**문제 5**

어느 도시에 있는 3개의 선거구에서 특정후보 A를 지지하는 유권자의 비율을 비교하기 위해 각 선거구에서 300명을 무작위를 추출하여 조사한 데이터이다. 주어진 데이터를 대상으로 후보A를 지지하는 비율이 3개 선거구 간에 차이가 있는지를 5% 유의수준에서 검정하라.

표 1.10: 지역별 대선 후보의 지지율

| 구분 | 선거구 1 | 선거구 2 | 선거구 3 |
|---|---|---|---|
| 지지함 | 176 | 193 | 159 |
| 지지하지 않음 | 124 | 107 | 141 |

1) 연구가설과 귀무가설을 설정하시오.

2) 가설 검증에 대한 검정통계량을 계산하시오. (단, 반올림하여 소수점 셋째 자리까지 표시하시오.) 연구가설의 채택여부를 결정하시오.

**문제 6**

슬통 초등학교에서는 학생들의 건강 관리를 위해 일부 학생들의 혈압을 측정하였습니다. 총 25명의 학생 중 남학생은 16명, 여학생은 9명이었습니다. 학교 건강 관리팀은 남학생과 여학생 사이에 평균 혈압에 차이가 있는지 궁금해하였고, 이를 확인하기 위해 5%의 유의수준에서 검정하려 합니다. (단, 남녀 학생의 혈압이 정규분포를 따르며, 두 집단의 분산이 동일하다고 가정합니다.)

표 1.11: 남학생 여학생 혈압 측정 결과

| No. | 남학생 | 여학생 |
|---|---|---|
| 1 | 124.97 | 114.87 |
| 2 | 118.62 | 128.14 |
| 3 | 126.48 | 115.92 |
| 4 | 135.23 | 110.88 |
| 5 | 117.66 | 139.66 |
| 6 | 117.66 | 122.74 |
| 7 | 135.79 | 125.68 |
| 8 | 127.67 | 110.75 |
| 9 | 115.31 | 119.56 |
| 10 | 125.43 | |
| 11 | 115.37 | |
| 12 | 115.34 | |
| 13 | 122.42 | |
| 14 | 100.87 | |
| 15 | 102.75 | |
| 16 | 114.38 | |

1) 연구가설과 귀무가설 작성하시오.

2) 가설검증에 대한 검정통계량를 계산하고, 연구가설의 채택여부를 설명하시오.

3) 가설검정에 대한 신뢰구간을 계산하고, 계산된 신뢰구간이 어떻게 2의 결과를 지지하는지 설명하시오. (단, 신뢰구간 계산시 다음의 값을 사용하시오.)

- $t_{23,0.025} = 2.069$

### 문제 7

다음은 슬통시의 20대 남성 411명을 대상으로 키(height)와 몸무게(weight), 그리고 허리둘레(waist)를 측정한 데이터이다. 20대 남성의 키와 허리둘레가 체중과 어떠한 관계에 있는지 베이지안 회귀분석을 사용하여 분석하시오.

- 데이터: `height_weight_waist.csv`

표 1.12: 20대 남성 신체정보 측정 데이터

| 순번 | 컬럼명 | 의미 | 타입 |
|---|---|---|---|
| 1 | height | 키 (cm) | number |
| 2 | weight | 몸무게 (kg) | number |
| 3 | waist | 허리둘레 (cm) | number |

1) 아래 조건들을 참고하여 회귀계수를 구하라.

- 베이지안 회귀를 이용한다.
- 시드넘버는 1234로 지정
- 1000번의 burn-in 후 10,000번의 MCMC 수행
- 사전분포 정보
  - 회귀계수: 부적절한 균일분포(improper uniform prior distribution)
  - 오차항의 분산의 사전분포: 역감마분포 형상(shape) 모수와 척도(scale) 모수는 각각 0.0005로 지정

2) 도출된 결과에 근거하여 키 180 센티미터, 허리둘레 80 센티미터인 20대 남성 체중의 추정값을 구하시오.

## 제 27회

### 문제 5

슬통 전자는 과거 수년간 전통적인 제조 방식을 사용하여 전자제품을 생산해 왔습니다. 2년 전, 회사의 연간 생산량은 10만 개였습니다. 그러나 기술의 발전과 시장의 변화를 감지하고, 회사는 생산 효율을 향상 방법을 모색했습니다. 그 결과, 1년 전에는 생산량을 15만 개로 늘렸습니다. 최근 회사는 스마트 팩토리 도입의 결정으로 연간 생산량은 무려 25만 개로 증가하였습니다. 이러한 변화를 통해 연평균 몇 배의 증가가 이루어졌는지 계산하시오. (단, 반올림하여 소수점 셋째 자리까지 표시하시오.)

### 문제 6

엘리베이터에 설치된 미디어 보드에서 8개의 광고 영상의 평균 광고시간 (단위: 초)을 조사하였다.

$$19.26, 17.09, 16.71, 19.76, 17.25, 18.88, 20.12, 16.46$$

이 데이터가 정규분포를 따른다고 할 때, 광고시간의 90% 신뢰구간을 소수점 둘째 자리까지 구하시오.

### 문제 7

`streams.csv` 파일에는 16개 강의 상류와 하류에서의 생물 다양성 데이터가 포함되어 있습니다. 강물은 상류에서 하류로 흐르며, 같은 강의 상류와 하류 생물 다양성은 서로 종속적인 관계에 있습니다. 상류와 하류의 생물 다양성 점수 평균에 차이가 있는지 유의수준 0.05로 검증하시오.

1) 귀무가설과 연구가설을 제시하시오.

2) 검정 통계량 및 유의확률을 산출하고, 연구가설 채택 여부를 판단하시오.

### 문제 8

부산시에서 교통 관리 담당자 슬통이는 교통량이 기상 조건에 어떻게 영향을 받는지 알아보고자 한다. 특히, 비나 눈, 바람과 같은 기상 조건이 도로의 교통량에 어떤 영향을 미치는지 궁금해한다. 지난 1년 동안의 교통량과 기상정보 데이터 `traffic.csv` 파일을 사용하여 다음에 답하시오.

1) 분위수 회귀(quantile regression)를 이용하여 회귀 계수를 구하십시오. (단, 여기서 분위수는 50백분위수를 사용합니다.)

2) 기온이 15.5℃, 강수량이 16.5$mm$, 풍속이 1.6$m/s$일 때의 교통량은 어떻게 되는지 예측하십시오.

### 문제 9

`subway.csv` 파일에는 최근 서울시에서 서울역의 지하철 승차 인원수를 조사하며, 1월과 2월의 출근시간(7시~9시) 동안의 승차 인원을 관찰한 데이터가 들어있습니다. 서울시의 데이터 분석 팀은 다양한 호선과 월별 승차 인원의 차이를 알고 싶어합니다. 호선과 월의 상호작용에 따라 승차 인원수에 차이가 있는지 유의수준 0.05에서 검정하시오. (단, 제곱합 계산에는 제3종 Type III을 이용하시오.)

1) 귀무가설과 대립가설을 제시하라.

2) 검정통계량 및 유의확률을 계산하고, 연구가설 채택 여부에 대하여 서술하시오.

## 제 28회

### 문제 4

다음의 질문에 답하시오.

1) Geartool 데이터 셋을 이용하여 시간별, 제조사별 불량률 데이터로 생존분석을 시행한 후 25, 30, 35개월 후의 불량률을 계산하시오.

2) 로그 순위법으로 제조사별 불량률이 차이가 있는지 검정하시오.

**문제풀이**

1) Geartool 데이터 셋을 이용하여 시간별, 제조사별 불량률 데이터로 생존분석을 시행한 후 25, 30, 35 개월 후의 불량률을 계산하시오.

- 데이터 불러오기

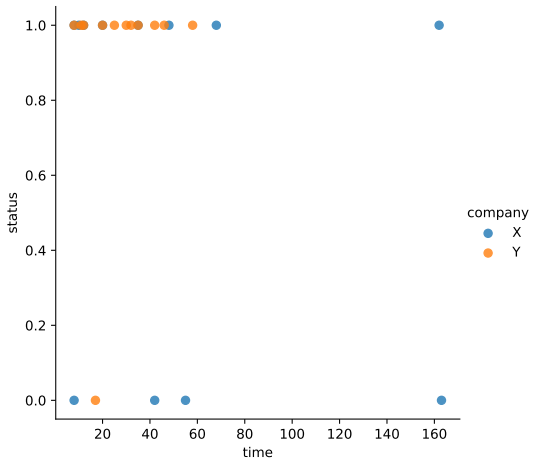
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from lifelines import KaplanMeierFitter
from lifelines.statistics import logrank_test
from scipy.stats import ks_2samp

# Read and manipulate the data
geartool = pd.read_csv("data/geartool.csv")
geartool['company'] = geartool['company'].astype('category')
```

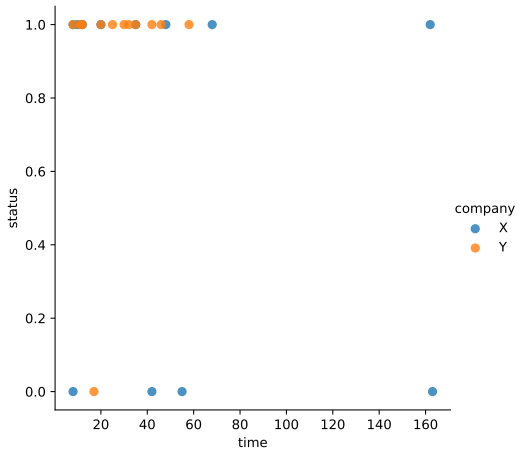
데이터 셋에 의하면 고장여부(0: 정상, 1:고장) 변수를 사용하여 그 시점까지 고장이 난 것을 확인 할 수 있는 것들 (1)이 있고, 그 시점까지 고장인 난 것을 확인 못한 관찰값들 (2)이 있는 것을 알 수 있다.

- 시각화

각 제조사별 불량품 데이터를 시각화하면 다음과 같다.

```
sns.lmplot(data=geartool, x='time', y='status', hue='company', fit_reg=False)
```

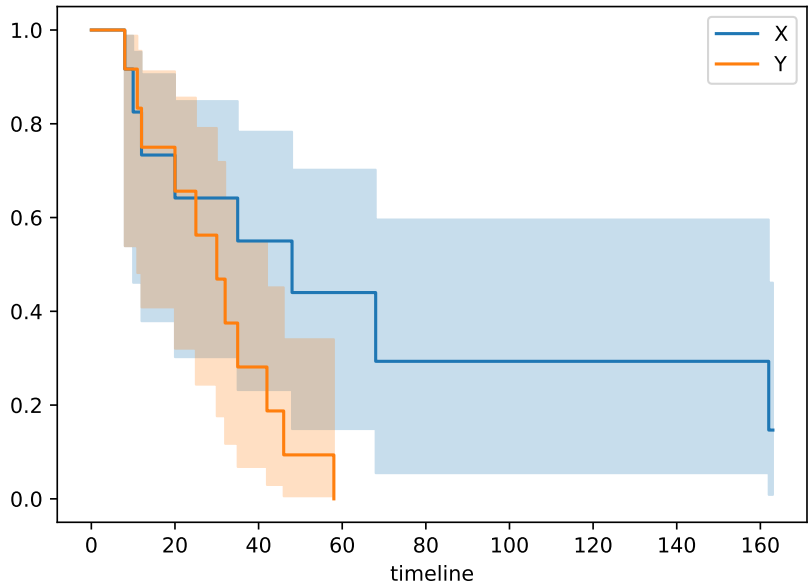





- `KaplanMeierFitter`를 사용한 불량률 적합

KaplanMeierFitter 모듈의 `fit()` 함수는 `Kaplan-Meier` 생존함수를 적합시켜주는 함수이며, 관찰 데이터와 레이블을 입력값으로 받습니다.

```
# Kaplan-Meier survival estimate
kmf = KaplanMeierFitter()
for company in geartool['company'].unique():
    data = geartool[geartool['company'] == company]
    kmf.fit(data['time'], event_observed=data['status'], label=str(company));
    kmf.plot_survival_function()
plt.show()
```




- 불량률 추정하기

```
def compute_survival(data):
    kmf.fit(data['time'], event_observed=data['status'])
    return kmf.survival_function_at_times([25, 30, 35])

geartool.groupby('company').apply(compute_survival)
```

```
## Y              25        30       35
## company
## X        0.641667  0.641667  0.55000
## Y        0.562500  0.468750  0.28125
```

생존함수는 주어진 $t$ 시점까지 살아있을 확률이므로, 제조사 X와 Y의 25, 30, 35개월에 대응하는 불량률은 각각 다음과 같이 계산할 수 있다.

```
# X
1 - np.array([0.642, 0.642, 0.550])
# Y
```

```
## array([0.358, 0.358, 0.45 ])
```

```
1 - np.array([0.562, 0.469, 0.281])
```

```
## array([0.438, 0.531, 0.719])
```

2) 로그 순위법으로 제조사별 불량률이 차이가 있는지 검정하시오.

로그 순위법은 두 함수가 차이가 있는지 검정하는 비모수적 방법이며, `logrank_test` 함수에 구현되어 있습니다. 유의 수준 0.05 하에서 검정을 실시하도록 하겠습니다.

- 귀무가설: 두 생존함수는 동일하다. ( $S_X(t) = S_Y(t)$ for all $t$)
- 대립가설: 두 생존함수는 같지 않다.( $S_X(t) \neq S_Y(t)$ for some $t$)

```
results = logrank_test(geartool[geartool['company']=='X']['time'],
                       geartool[geartool['company']=='Y']['time'],
                       event_observed_A=geartool[geartool['company']=='X']['status'],
                       event_observed_B=geartool[geartool['company']=='Y']['status'])
results.print_summary()
```

```
## <lifelines.StatisticalResult: logrank_test>
##                t_0 = -1
##  null_distribution = chi squared
## degrees_of_freedom = 1
##          test_name = logrank_test
##
## ---
```

```
##  test_statistic    p  -log2(p)
##            3.64 0.06      4.15
```

로그 순위 검정의 통계량 값 3.6에 대응하는 p-value값이 0.06으로 이는 유의수준 0.05보다 크므로, 귀무가설을 기각할 수 없습니다. 따라서, 두 제조사의 불량률에는 유의미한 차이가 있다고 판단할 통계적 근거가 충분치 않다고 판단할 수 있습니다.

### 문제 5

다음 표는 슬통 Food의 신제품 홍보 설문 조사 결과이다. 시식 행사에 참여한 고객들의 시식 후 구매 의사의 변화가 있는지 없는지 검정하시오.

표 1.13: 시식 전후 비교

| 구분 | 시식전 | 있음 | 없음 |
|---|---|---|---|
| 시식전 | 있음 | 23 | 7 |
| | 없음 | 18 | 12 |

### 문제 6

school_exam.csv 파일에는 2개의 고등학교 시험 표준 점수가 들어있습니다. 두 학교 표준 점수의 분포 차이가 있는지 검정하시오. (단, 각 학생들의 성적은 독립이라고 가정)

**문제풀이** 데이터를 불러오고, 알맞은 형태로 변환합니다.

```
school_exam = pd.read_csv("data/school_exam.csv")
school_exam = pd.melt(school_exam, value_vars=['school_A', 'school_B'],
                      var_name='school', value_name='score')
```

- 각 학교별 표본 수 계산

```
print(school_exam['school'].value_counts())
```

```
## school_A    24
## school_B    24
## Name: school, dtype: int64
```

두 학교의 표준 점수 분포를 검정하기 위하여 two-sample Kolmogorov-Smirnov 검정을 유의수준 0.05 하에서 수행합니다.

- 귀무가설: 두 표본이 같은 분포에서 뽑혀져 나왔다.
- 대립가설: 두 표본이 다른 분포에서 뽑혀져 나왔다.

```
school_A_scores = school_exam[school_exam['school'] == 'school_A']['score']
school_B_scores = school_exam[school_exam['school'] == 'school_B']['score']
ks_statistic, p_value = ks_2samp(school_A_scores, school_B_scores)
print(f'KS-statistic: {ks_statistic}, p-value: {p_value}')
```

```
## KS-statistic: 0.5416666666666666, p-value: 0.0014013568629808897
```

검정 통계량 값 0.29에 대응하는 p-value값 0.4616이 유의수준 값 0.05보다 크므로 귀무가설을 기각 할 수 없습니다. 두 표본이 같은 분포에서 뽑혀져 나왔다는 가설을 기각 할 수 있는 통계적 근거가 충분히 않으므로 같은 분포에서 나왔다고 판단합니다.

#### 문제 7

몸무게를 제어했을 때, 나이와 콜레스테롤 상관계수 및 유의확률 구하라.

### 제 29회

#### 문제 5

제품 A의 불량률은 0.03이다. 25개의 제품을 뽑았을 때 3개가 불량일 확률을 구하시오. (소수점 다섯 째 자리에서 반올림)

#### 문제 6

C사 생산 제품 1000개 중 양품이 600개, D사 생산 제품 500개 중 양품이 200개 이다. 두 회사의 양품률에 차이가 있는지 검정하여라.

#### 문제 7

아래 데이터는 a,b,c,d 네 차종 각각 5회 실험 시 범퍼 파손 정도 이다. (단, 각 모집단은 정규분포를 따르며 모집단 간 등분산성을 가정한다.)

1) 각 차종 별 범퍼 파손의 정도에 차이가 유의한지 검정하라.

2) 귀무가설을 채택한다면 그 의미를 해석하고, 귀무가설을 기각하였다면 사후분석을 시행하라.

#### 문제 8

L1, L2, L3 세 개의 생산라인에서 각각 13%, 37%, 50%를 생산하며 각각 1.1% , 2.1%, 3.3% 불량률을 갖는다. 불량 제품이 나왔을 때 L1 라인에서 생산되었을 확률을 구하시오. (소수점 둘째자리에서 반올림)

### 제 30회

#### 문제 5

아래 데이터는 3개의 철강 제조공장(공장A,공장B,공장C)에서 생산된 제품을3개의 지역(지역1, 지역2, 지역3)으로 배송할 때 발생하는 운송비용과 공장별 총 생산량, 지역별 총 수요량이다.

이 데이터를 활용하여 총 운송비를 최소로 하는 운송계획을 수립하시오. (단, 각 공장에서는 3개 지역으로만 운송되고, 공장간 또는 지역 간 운송은 없다고 가정한다.)

표 1.14: 운송비용과 공장별 총 생산량, 지역별 총 수요량

| 구분 | 지역1 | 지역2 | 지역3 | 총생산량 |
|---|---|---|---|---|
| 공장 A | 12만원 | 5만원 | 34만원 | 70개 |
| 공장 B | 22만원 | 2만원 | 21만원 | 55개 |

| 구분 | 지역1 | 지역2 | 지역3 | 총생산량 |
|---|---|---|---|---|
| 공장 C | 3만원 | 23만원 | 15만원 | 25개 |
| 총 수요량 | 30개 | 50개 | 70개 | |

```
import numpy as np
from itertools import product
from scipy.optimize import linprog

# Data
cost_matrix = [[12, 5, 34], [22, 2, 21], [3, 23, 15]]
supply = [70, 55, 25]
demand = [30, 50, 70]

costs = [item for sublist in cost_matrix for item in sublist]  # Flattened cost matrix

# Constraints
# For A_eq and b_eq (equality constraints)
A_eq = [
    [1, 1, 1, 0, 0, 0, 0, 0, 0],  # Supply constraint for Factory A
    [0, 0, 0, 1, 1, 1, 0, 0, 0],  # Supply constraint for Factory B
    [0, 0, 0, 0, 0, 0, 1, 1, 1],  # Supply constraint for Factory C
    [1, 0, 0, 1, 0, 0, 1, 0, 0],  # Demand constraint for Region 1
    [0, 1, 0, 0, 1, 0, 0, 1, 0],  # Demand constraint for Region 2
    [0, 0, 1, 0, 0, 1, 0, 0, 1],  # Demand constraint for Region 3
]

b_eq = supply + demand  # Combining supply and demand into one list

# Solve using linprog
result = linprog(c=costs, A_eq=A_eq, b_eq=b_eq, method='simplex')

# Extract and reshape the solution
solution_scipy = [list(result.x[i:i+3]) for i in range(0, len(result.x), 3)]
np.array(solution_scipy)
```

**문제풀이**

```
## array([[20., 50.,  0.],
##        [ 0.,  0., 55.],
##        [10.,  0., 15.]])
```

**문제 6**

아래 데이터를 이용하여 헤드셋에 대한 연령대별 선호도 차이가 있는지를 유의수준5%로 검정하시오. (단, 반올림하여 소수점 셋째 자리까지 표시하시오.)

- 데이터: `headset.csv`
- 데이터는 ID, 헤드셋 종류, 연령대로 구성

1) 연구가설($H_1$)과 귀무가설($H_0$)을 설정하시오.

2) 유의확률을 계산하고 가설의 채택 여부를 결정하시오.

**문제 7**

각각 6명의 자녀를 가진 다섯 가족이 있다. 각각의 자녀가 아들 또는 딸일 확률은 0.5일 때 아래 질문에 답하시오. (단, 반올림하여 소수점 셋째 자리까지 표시하시오.)

1) 4명 이상의 딸을 가진 가족이 세 가족 이상일 확률을 0에서 1 사이 숫자로 구하시오.

2) 다섯 가족 중 몇 가족이 4명 이상의 딸을 가질 것으로 기대되는지 계산하시오.