Multivariate Statistical Analysis

Answer to Assignment 4

12112627 李乐平

Question 1

There is a certain relationship between the amount of sweating and the content of potassium in human body. The amount of perspiration (x_1) of 20 healthy adult women was measured. The content of sodium (x_2) and potassium content (x_3) were also measured. The data are listed in the following table. It is assumed that the data obeys a trivariate normal distribution.

(1) Construct a Hotelling T^2 statistic to test the hypothesis $H_0 : \mu = \mu_0 = [4\ 50\ 10]'$ against $H_1 : \mu \neq \mu_0 (\alpha = 0.05)$.

In [1]:
```python
import numpy as np
import pandas as pd

from scipy.stats import f, t
```

In [2]:
```
df = pd.read_csv("4_1.csv")
df
```

Out[2]:

	x1	x2	x3
0	3.7	48.5	9.3
1	5.7	65.1	8.0
2	3.8	47.2	10.9
3	3.2	53.2	12.0
4	3.1	55.5	9.7
5	4.6	36.1	7.9
6	2.4	24.8	14.0
7	7.2	33.1	7.6
8	6.7	47.4	8.5
9	5.4	54.1	11.3
10	3.9	36.9	12.7
11	4.5	58.8	12.3
12	3.5	27.8	9.8
13	4.5	40.2	8.4
14	1.5	13.5	10.1
15	8.5	56.4	7.1
16	4.5	71.6	8.2
17	6.5	52.8	10.9
18	4.1	44.1	11.2
19	5.5	40.9	9.4

In [4]:
```
n = len(df)
p = len(df.columns)
data = np.array(df)
means = np.array(df.mean()).reshape(-1, 1)
cov = np.array(df.cov())

u0 = np.array([4, 50, 10]).reshape(-1, 1)
alpha = 0.05

F = f.ppf(1 - alpha, p, n - p)
T2_alpha = p * (n - 1) * F / (n - p)

T2 = n * (means - u0).T @ np.linalg.inv(cov) @ (means - u0)

print(f"Hotelling's T2 = {T2[0][0] : .4f}")
print(f"T2(α/2, p, n - 1) = {T2_alpha: .4f}")
```

```
Hotelling's T2 =  9.7388
T2(α/2, p, n - 1) =  10.7186
```

Solution: Because Hotelling's $T^2 = 9.7388 < T_\alpha^2(3, 19) = 10.7186$, hence we cannot reject H_0.

(2) Find the 95% confidence region for μ.

In [6]:
```
np.linalg.inv(cov)
```

Out[6]:
```
array([[ 0.58615531, -0.02208572,  0.25796874],
       [-0.02208572,  0.00606723, -0.00158093],
       [ 0.25796874, -0.00158093,  0.40184677]])
```

Solution: The confidence region with significance level $1 - \alpha = 95\%$ is given by

$$\{\mu : 20 \times (\begin{bmatrix} 4.64 \\ 45.4 \\ 9.965 \end{bmatrix} - \mu)' \begin{bmatrix} 0.58615531 & -0.02208572 & 0.25796874 \\ -0.02208572 & 0.00606723 & -0.00158093 \\ 0.25796874 & -0.00158093 & 0.40184677 \end{bmatrix} (\begin{bmatrix} 4.64 \\ 45.4 \\ 9.965 \end{bmatrix} - \mu) \leq 10.7186\}$$

Question 2

Using the data in the table, test the following hypotheses of the female baby population at the significance level of $\alpha = 0.05$.

(1) $H_0 : \mu = [80\ 60\ 15]'$ against $H_1 : \mu \neq [80\ 60\ 15]'$.

In [5]:
```
df = pd.read_csv("4_3.csv")
df
```

Out[5]:

	x1	x2	x3
0	80	58.4	14.0
1	75	59.2	15.0
2	78	60.3	15.0
3	75	57.4	13.0
4	79	59.5	14.0
5	78	58.1	14.5
6	75	58.0	12.5
7	64	55.5	11.0
8	80	59.2	12.5

In [6]:
```
n = len(df)
p = len(df.columns)
data = np.array(df)
means = np.array(df.mean()).reshape(-1, 1)
cov = np.array(df.cov())

u0 = np.array([80, 60, 15]).reshape(-1, 1)
alpha = 0.05

F = f.ppf(1 - alpha, p, n - p)
T2_alpha = p * (n - 1) * F / (n - p)

T2 = n * (means - u0).T @ np.linalg.inv(cov) @ (means - u0)

print(f"Hotelling's T2 = {T2[0][0]: .4f}")
print(f"T2(α/2, p, n - 1) = {T2_alpha: .4f}")
```

```
Hotelling's T2 =  13.3700
T2(α/2, p, n - 1) =  19.0283
```

Since $T^2 = 13.3700 < T^2(\alpha/2, p, n-1) = 19.0283$, we cannot reject H_0.

(2) $H_0 : \frac{1}{5}\mu_1 = \frac{1}{4}\mu_2 = \mu_3$ against H_1 : At least 2 terms in $\frac{1}{5}\mu_1,\ \frac{1}{4}\mu_2,\ \mu_3$ are unequal.

Solution: Let $a = \begin{bmatrix} 0.2 & 0.2 \\ -0.25 & 0 \\ 0 & -1 \end{bmatrix}$, then our test turns out to be $H_0 : a'\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ against $H_1 : a'\mu \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

In [7]:
```python
a = np.array([
    [0.2, -0.25, 0],
    [0.2, 0, -1]
]).T
a_means = a.T @ means
a_cov = a.T @ cov @ a
a_u = np.array([0, 0, 0]).reshape(-1, 1)

k = len(a.T)
F = f.ppf(1 - alpha, p, n - p)
T2_alpha = p * (n - 1) * F / (n - p)
T_alpha = np.sqrt(T2_alpha)
t_alpha = t.ppf(1 - alpha / 2 / k, n - 1)

print(f"""
T({alpha}, {n - 1}) = {T_alpha}
t({alpha / 2 / k}, {n - 1}) = {t_alpha}
""")
```

```
T(0.05, 8) = 4.362138311924285
t(0.0125, 8) = 2.751523593712948
```

Because $t_{\alpha/2k}(n-1) = 2.7515 < T_\alpha(p, n-1) = 4.3621$, so we should consider to use Bonferroni simultaneous confidence interval.

In [8]:
```
print(f"""
The Bonferroni simultaneous confidence interval is given by
    {"×".join([f"[{((a_means[i]) - t_alpha * np.sqrt(a_cov[i, i]) / np.sqrt(n))[0]}, {((a_means[i]) + t_alpha * np.sqrt(a
""")
```

```
The Bonferroni simultaneous confidence interval is given by
    [-0.07359273266943156, 1.2735927326694303]×[0.7523762685529175, 2.6476237314470845]
```

Notice that $\begin{bmatrix} 0 \\ 0 \end{bmatrix} \notin [-0.0736, 1.2736] \times [0.7524, 2.6476]$, hence we must reject H_0.

Question 3

A prison divides prisoners into 3 parts: ordinary prisoners (catagory 1), crazy prisoners(catagory 2) and other prisoners(catagory 3). 20 prisoners were selected from each of the 3 parts to measure the length of their ears. Under the hypothesis of multivariate normality, we tried to test whether there was significant difference in the length of ears of 3 parts ($\alpha = 0.05$).

In [9]:
```
df = pd.read_csv("4_10.csv")
df
```

Out[9]:

	x1	y1	x2	y2	x3	y3
0	59	59	70	69	63	63
1	60	65	69	68	56	57
2	58	62	65	65	62	62
3	59	59	62	60	59	58
4	50	48	59	56	62	58
5	59	65	55	58	50	57
6	62	62	60	58	63	63
7	63	62	58	64	61	62
8	68	72	65	67	55	59
9	63	66	67	62	63	63
10	66	63	60	57	65	70
11	56	56	53	55	64	64
12	62	64	66	65	65	65
13	66	68	60	53	67	67
14	65	66	59	58	55	55
15	61	60	58	54	56	56
16	60	64	60	56	65	67
17	60	57	54	59	62	65
18	58	60	62	66	55	61
19	58	59	59	61	58	58

In [10]:
```
means = np.array(df.mean()).reshape(-1, 1)
cov = np.array(df.cov())
```

In [11]:
```
n = len(df)
k = 3
p = 2
alpha = 0.05
mmeans = sum([means[2 * i: 2 * i + 2] for i in range(0, k)]) / k

E = (n - 1) * sum([cov[2 * i : 2 * i + 2, 2 * i : 2 * i + 2] for i in range(0, k)])
H = n * sum([(means[2 * i : 2 * i + 2] - mmeans) @ (means[2 * i : 2 * i + 2] - mmeans).T for i in range(0, k)])
L = np.linalg.det(E) / np.linalg.det(E + H)
F = (k * n - k - p + 1) * (1 - np.sqrt(L)) / p / np.sqrt(L)
F_alpha = f.ppf(1 - alpha, 2 * p, 2 * (k * n - k - p + 1))

print(f"""
F = {F : .4f}
F({alpha}, {2 * p}, {2 * (k * n - k - p + 1)}) = {F_alpha : .4f}
""")
```

```
F =  1.1069
F(0.05, 4, 112) =  2.4527
```

Solution: We want to test whether $H_0 : \mu_1 = \mu_2 = \mu_3$ or H_1 : At least 2 of μ_1, μ_2, μ_3 are unequal, with the significance level of $1 - \alpha = 0.95$.

The test statistic $F = \frac{(n-k-p+1)\sqrt{\Lambda}}{p\sqrt{\Lambda}} = 1.1069 < F(\alpha, 2p, 2(n-k-p+1)) = 2.4527$, hence we cannot reject H_0.