Statistics Final Personal Assignment

110550085 房天越

1. a. By assumption, the population variance of age in each level of class are the same. We can apply the ANOVA test.

   Let H0 be the null hypothesis that the means are the same, then H1 is the statement that at least one mean is different.

   Use this following code to obtain the useful columns and values.

```python
from scipy import stats
import numpy as np
import pandas as pd

csv_file_path = "Titanic_R.csv"

useful_cols = ['name', 'pclass', 'survived', 'age']

df = pd.read_csv(csv_file_path, usecols=useful_cols)

print("After dropping rows with missing values:")
df['age'] = df['age'].replace(" ", np.nan)
df = df.dropna(subset=['pclass', 'age'])

print(f"Count of age : {df['age'].count()}")
print(f"Count of pclass : {df['pclass'].count()}")

df['age'] = pd.to_numeric(df['age'])

print(df.describe())
```

   After running this code, we can get this result:

```
After dropping rows with missing values:
Count of age : 1046
Count of pclass : 1046
            pclass     survived          age
count  1046.000000  1046.000000  1046.000000
mean      2.207457     0.408222    29.881135
std       0.841497     0.491740    14.413500
min       1.000000     0.000000     0.166700
25%       1.000000     0.000000    21.000000
50%       2.000000     0.000000    28.000000
75%       3.000000     1.000000    39.000000
max       3.000000     1.000000    80.000000
```

   This result shows that there are 1046 useful rows, and some descriptive data of the given csv file.

   Then, we move on to apply the ANOVA test with this code:

```
print("--------------------")
print("Apply the ANOVA test")
print("--------------------")
print("H0: The means are equal")
print("Ha: At least one of the means is different")

class1 = df[df['pclass'] == 1]['age']
class2 = df[df['pclass'] == 2]['age']
class3 = df[df['pclass'] == 3]['age']

f_statistic, p_value = stats.f_oneway(class1, class2, class3)

print(f"F-statistic: {f_statistic}")
print(f"P-value: {p_value}")

alpha = 0.05

if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

After running this part, we get this result:

```
H0: The means are equal
Ha: At least one of the means is different
F-statistic: 108.32597769816022
P-value: 1.7972028849109706e-43
Reject the null hypothesis
```

We can see that the F-statistic is 108.326, and P-value is about 0.

Assume, alpha = 0.05, we reject the null hypothesis.

There is a strong evidence to indicate that there is at least one of the means of the passengers' ages is different from the others.

b.

For this problem, we perform a chi-square test of independence using this code:

```python
print("1b")
print("H0: There is no significant association between pclass and survived")
print("Ha: There is a significant association between pclass and survived")

from scipy.stats import chi2_contingency

contingency_table = pd.crosstab(df['pclass'], df['survived'])
chi2, p, dof, expected = chi2_contingency(contingency_table)

print(f"Chi2: {chi2}")
print(f"P-value: {p}")
print(f"Degrees of freedom: {dof}")
print("Expected:")
print(expected)

alpha = 0.05

if p < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

Notice that here we use the original df to construct the contingency table instead of the df that dropped useless values in 1a, we can obtain this result:

```
1b
H0: There is no significant association between pclass and survived
Ha: There is a significant association between pclass and survived
Chi2: 127.85915643930326
P-value: 1.7208259588256175e-28
Degrees of freedom: 2
Expected:
[[199.62337662 123.37662338]
 [171.19404125 105.80595875]
 [438.18258212 270.81741788]]
Reject the null hypothesis
```

So we have the chi-square = 127.859, and P-value is about 0.

Assume alpha = 0.05, we reject the null hypothesis.

So there is a strong evidence to indicate that there is a significant association between pclass and survived.

2. Let's apply a single tail 2-sample t-test.

   Let H0 be the null hypothesis that the ratio in 2022 is greater than or equal to that of 2021.

   H1 be the alternative hypothesis that is less

   Use this code to perform the test:

```python
from scipy import stats
import numpy as np
import pandas as pd

data1 = pd.read_csv("07-07-2022.csv")
data2 = pd.read_csv("07-07-2021.csv")

print("H0: The Case Fatality Ratio in 2022 is greater than or equal to the Case Fatality Ratio in 2021")
print("Ha: The Case Fatality Ratio in 2022 is less than the Case Fatality Ratio in 2021")

# perform test
t_statistic, pval= stats.ttest_ind(
    data1['Case_Fatality_Ratio'].dropna(),
    data2['Case_Fatality_Ratio'].dropna(),
    equal_var=False,
    alternative='less'
)

print(f"t-statistic: {t_statistic}")
print(f"P-value: {pval}")

alpha = 0.05
if pval < alpha:
    print("Reject the null hypothesis")

else:
    print("Fail to reject the null hypothesis")
```

Then we can get the result:

```
H0: The Case Fatality Ratio in 2022 is greater than or equal to the Case Fatality Ratio in 2021
Ha: The Case Fatality Ratio in 2022 is less than the Case Fatality Ratio in 2021
t-statistic: -0.2546296007662016
P-value: 0.3995078941229881
Fail to reject the null hypothesis
```

So, t-statistic = -0.255, P-value = 0.4.

With alpha = 0.05, P-value > alpha, we fail to reject H0.

There is not enough evidence to show that the mean in 2022 is less than the mean in 2021.

3. a. Apply the chi-square test,

   Let H0: The customer rating is not affected by reached on-time status.

   Ha: The customer rating is affected by reached on-time status.

   Use this code:

```python
from scipy import stats
import numpy as np
import pandas as pd

csv_file_path = "Train.csv"

print("H0: There is no significant association between Customer_rating and Reached.on.Time_Y.N")
print("Ha: There is a significant association between Customer_rating and Reached.on.Time_Y.N")

useful_cols = ['Customer_rating', 'Reached.on.Time_Y.N']
df = pd.read_csv(csv_file_path, usecols=useful_cols)

contingency_table = pd.crosstab(df['Customer_rating'], df['Reached.on.Time_Y.N'])
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)

print(f"Chi2: {chi2}")
print(f"P-value: {p}")

alpha = 0.05

if p < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

After running this code, we get this result:

```
H0: There is no significant association between Customer_rating and Reached.on.Time_Y.N
Ha: There is a significant association between Customer_rating and Reached.on.Time_Y.N
Chi2: 3.200045474831146
P-value: 0.5249236018493662
Fail to reject the null hypothesis
```

Since chi2 = 3.200, P-value = 0.525, we fail to reject H0.

There is not enough evidence to indicate that there is a significant association between Customer_rating and Reached.on.Time_Y.N.

b. Apply the ANOVA test.

Let H0 be there is no different among the modes of shipment.

Ha be there is at least one mode of shipment different from others.

Use this code:

```python
from scipy import stats
import numpy as np
import pandas as pd

csv_file_path = "Train.csv"

useful_cols = ['Mode_of_Shipment', 'Cost_of_the_Product']

df = pd.read_csv(csv_file_path, usecols=useful_cols)

print("H0: The mean cost of the product is the same for all modes of shipment")
print("Ha: The mean cost of the product is not the same for all modes of shipment")

ship = df[df['Mode_of_Shipment'] == 'Ship']['Cost_of_the_Product']
flight = df[df['Mode_of_Shipment'] == 'Flight']['Cost_of_the_Product']
road = df[df['Mode_of_Shipment'] == 'Road']['Cost_of_the_Product']

f_stat, p_val = stats.f_oneway(ship, flight, road)

print(f"F-statistic: {f_stat}")
print(f"P-value: {p_val}")

alpha = 0.05

if p_val < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

We can obtain the results:

```
H0: The mean cost of the product is the same for all modes of shipment
Ha: The mean cost of the product is not the same for all modes of shipment
F-statistic: 0.368844184436174
P-value: 0.6915417092271511
Fail to reject the null hypothesis
```

So, with F-statistic = 0.369, P-value = 0.692, we fail to reject H0.

There is not enough evidence to show that there is difference between all modes of shipment.