Homework1

September 20, 2021

1 ECE 2556 Homework 1 - Avery Peiffer

[1]:
```
import pandas as pd
import random
import statsmodels.api as sm
from statistics import mean
from statsmodels.formula.api import glm
from scipy import stats
```

1.1 Question 1: For data in `scores.txt` (120x2), design a permutation algorithm (permutate 1000 times), and identify whether column 1 > column 2 (or column 1 < column 2) and report the p-value.

[2]:
```
# Set up the hyperparameters for the experiment, read in the data, and posit␣
 ↪the null and alternative hypotheses.
n = 1000
alpha = 0.05
top_x = n * alpha
perms = []

data = pd.read_csv('scores.txt', sep='\t', header=None, names=['col1', 'col2'])
samples = len(data)

orig_diff = data['col1'].mean() - data['col2'].mean()
print(f'Original difference = {orig_diff}.')

if orig_diff < 0:
    print(f'Since the original difference < 0: \
          H_0 = mean(col2) NOT > mean(col1) \
          H_a = mean(col2) > mean(col1).')
else:
    print(f'Since the original difference > 0: \
          H_0 = mean(col1) NOT > mean(col2) \
          H_a = mean(col1) > mean(col2).')
```

```
Original difference = -0.041666666666671404.
Since the original difference < 0:          H_0 = mean(col2) NOT > mean(col1)
```

H_a = mean(col2) > mean(col1).

[3]:
```
# Perform the permutations for the experiment.
for i in range(0, n):
    perm = data['col1'].tolist() + data['col2'].tolist()
    random.shuffle(perm)

    size = int(len(perm) / 2)
    new_col_1 = perm[:size]
    new_col_2 = perm[size:]

    mean1 = mean(new_col_1)
    mean2 = mean(new_col_2)

    if orig_diff < 0:
        perms.append(mean2 - mean1)
    else:
        perms.append(mean1 - mean2)
```

[4]:
```
# Tabulate and interpret the results of the experiment.
count = 0

for i in range(0, len(perms)):
    if perms[i] > abs(orig_diff):
        count += 1

p_value = count / len(perms)

print(f'The p-value is {p_value}.')

if p_value > alpha:
    print(f'The p-value is greater than alpha; the experiment fails to reject␣
 ↪the null hypothesis.')

    if orig_diff < 0:
        print(f'Result: mean(col2) is NOT > mean(col1).')
    else:
        print(f'Result: mean(col1) is NOT > mean(col2).')

else:
    print(f'The p-value is less than alpha; the experiment rejects the null␣
 ↪hypothesis.')

    if orig_diff < 0:
        print(f'Result: mean(col2) > mean(col1) at this significance.')
    else:
        print(f'Result: mean(col1) > mean(col2) at this significance.')
```

```
The p-value is 0.484.
The p-value is greater than alpha; the experiment fails to reject the null
hypothesis.
Result: mean(col2) is NOT > mean(col1).
```

1.2 Question 2: For data in hospital.txt:

- Divide subjects based on smoker (0: non-smoker, 1: smoker). Then, check whether weight has a
- Design a regression model to explore the relationship between blood pressure and smoke. Repo

[5]:
```
# Use student's t-test to determine the difference between the means of the␣
 ↪groups.
# Null hypothesis: Weight has no significant difference between the groups

df = pd.read_csv('hospital.txt', sep='\t')

df_smoker = df['Weight'][df['Smoker'] == 1]
df_non_smoker = df['Weight'][df['Smoker'] == 0]

ttest = stats.ttest_ind(df_smoker, df_non_smoker)

# The t value with alpha = 0.05 and degrees of freedom = 98 is about 1.984
alpha = 0.05
degrees_freedom = len(df_smoker) + len(df_non_smoker) - 2
t_const = stats.t.ppf(1 - (alpha / 2), degrees_freedom)

print(f'T-test statistic = {ttest.statistic} and p-value = {ttest.pvalue}.')
print(f'Constant from t-value lookup table = {t_const}.')

if ttest.statistic > t_const:
    print('Able to reject null hypothesis: there is a significant difference␣
 ↪between the means of the two groups.')
else:
    print('Fail to reject null hyopthesis: there is no significant difference␣
 ↪between the means of the two groups.')
```

```
T-test statistic = 2.185583781463617 and p-value = 0.03122827941228747.
Constant from t-value lookup table = 1.984467454426692.
Able to reject null hypothesis: there is a significant difference between the
means of the two groups.
```

[6]:
```
# Use two GLMs to examine the effect that smoking has on blood pressure.
X = df['Smoker']
y = df['BloodPressure_high']
z = df['BloodPressure_Low']

model = sm.GLM(y, X, family=sm.families.Gaussian())
results = model.fit()
```

```
model2 = sm.GLM(z, X, family=sm.families.Gaussian())
results2 = model2.fit()

print(f'These models examine the effect that smoking has on blood pressure - both high and low.')
print(f'Note: in the summary, the beta values are listed under the coefficient column and the p-values are under the P>|z| column.\n')
print(results.summary())
print(f'AIC: {results.aic}.')
print(f'BIC: {results.bic}.')
print('\n\n')
print(results2.summary())
print(f'AIC: {results2.aic}.')
print(f'BIC: {results2.bic}.')
```

```
These models examine the effect that smoking has on blood pressure - both high and low.
Note: in the summary, the beta values are listed under the coefficient column and the p-values are under the P>|z| column.

                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:     BloodPressure_high   No. Observations:                  100
Model:                            GLM   Df Residuals:                       99
Model Family:                Gaussian   Df Model:                           0
Link Function:               identity   Scale:                         9525.9
Method:                          IRLS   Log-Likelihood:                -599.48
Date:                Mon, 20 Sep 2021   Deviance:                   9.4306e+05
Time:                        14:08:44   Pearson chi2:                 9.43e+05
No. Iterations:                     3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Smoker       129.3529     16.738      7.728      0.000      96.546     162.159
==============================================================================
AIC: 1200.9591817235264.
BIC: 942603.8528574695.



                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:      BloodPressure_Low   No. Observations:                  100
Model:                            GLM   Df Residuals:                       99
Model Family:                Gaussian   Df Model:                           0
```

```
Link Function:                identity   Scale:                          4223.6
Method:                           IRLS   Log-Likelihood:                -558.81
Date:                 Mon, 20 Sep 2021   Deviance:                   4.1813e+05
Time:                         14:08:44   Pearson chi2:                 4.18e+05
No. Iterations:                      3
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Smoker        89.9118     11.146      8.067      0.000      68.067     111.757
==============================================================================
AIC: 1119.6263481652995.
BIC: 417677.8234457048.
```

[7]:
```
# Use another GLM to examine the opposite: the likelihood that an individual is␣
 ↪a smoker based on their blood pressure.
X = df[['BloodPressure_high', 'BloodPressure_Low']]
y = df['Smoker']
model = sm.GLM(y, X, family=sm.families.Gaussian())
results = model.fit()

print(f'This model measures the likelihood that an individual is a smoker based␣
 ↪on their blood pressure.')
print(f'Note: in the summary, the beta values are listed under the coefficient␣
 ↪column and the p-values are under the P>|z| column.\n')
print(results.summary())
print(f'AIC: {results.aic}.')
print(f'BIC: {results.bic}.')
```

```
This model measures the likelihood that an individual is a smoker based on their
blood pressure.
Note: in the summary, the beta values are listed under the coefficient column
and the p-values are under the P>|z| column.

                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 Smoker   No. Observations:                  100
Model:                            GLM   Df Residuals:                       98
Model Family:                Gaussian   Df Model:                           1
Link Function:                identity   Scale:                        0.19486
Method:                           IRLS   Log-Likelihood:                -59.109
Date:                 Mon, 20 Sep 2021   Deviance:                       19.096
Time:                         14:08:44   Pearson chi2:                     19.1
No. Iterations:                      3
Covariance Type:             nonrobust
==============================================================================
======
```

```
                          coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
------
BloodPressure_high    -0.0134      0.005     -2.698      0.007      -0.023
-0.004
BloodPressure_Low      0.0242      0.007      3.291      0.001       0.010
0.039
==============================================================================
======
AIC: 122.21862732487348.
BIC: -432.21066860643856.
```