Exemplar_Explore hypothesis testing

August 20, 2024

1 Exemplar: Explore hypothesis testing

1.1 Introduction

You work for an environmental think tank called Repair Our Air (ROA). ROA is formulating policy recommendations to improve the air quality in America, using the Environmental Protection Agency's Air Quality Index (AQI) to guide their decision making. An AQI value close to 0 signals "little to no" public health concern, while higher values are associated with increased risk to public health.

They've tasked you with leveraging AQI data to help them prioritize their strategy for improving air quality in America.

ROA is considering the following decisions. For each, construct a hypothesis test and an accompanying visualization, using your results of that test to make a recommendation:

1. ROA is considering a metropolitan-focused approach. Within California, they want to know if the mean AQI in Los Angeles County is statistically different from the rest of California.
2. With limited resources, ROA has to choose between New York and Ohio for their next regional office. Does New York have a lower AQI than Ohio?
3. A new policy will affect those states with a mean AQI of 10 or greater. Would Michigan be affected by this new policy?

Notes: 1. For your analysis, you'll default to a 5% level of significance. 2. Throughout the lab, for two-sample t-tests, use Welch's t-test (i.e., setting the `equal_var` parameter to `False` in `scipy.stats.ttest_ind()`). This will account for the possibly unequal variances between the two groups in the comparison.

1.2 Step 1: Imports

To proceed with your analysis, import `pandas` and `numpy`. To conduct your hypothesis testing, import `stats` from `scipy`.

Import Packages

[1]:
```
# Import relevant packages

### YOUR CODE HERE ###
```

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

You are also provided with a dataset with national Air Quality Index (AQI) measurements by state over time for this analysis. Use `pandas` to import the file `c4_epa_air_quality.csv` as a dataframe named `aqi`.

Note: For purposes of your analysis, you can assume this data is randomly sampled from a larger population.

Load Dataset

[2]:
```
# Use read_csv() to import your data

### YOUR CODE HERE ###

aqi = pd.read_csv('c4_epa_air_quality.csv')
```

1.3 Step 2: Data Exploration

1.3.1 Before proceeding to your deliverables, explore your datasets.

Use the following space to surface descriptive statistics about your data. In particular, explore whether you believe the research questions you were given are readily answerable with this data.

[3]:
```
# Explore your dataframe `aqi` here:

### YOUR CODE HERE ###

print("Use head() to show a sample of data")
print(aqi.head())

print("Use describe() to summarize AQI")
print(aqi.describe(include='all'))

print("For a more thorough examination of observations by state use␣
 ↪values_counts()")
print(aqi['state_name'].value_counts())

print('for a more')
```

```
Use head() to show a sample of data
   Unnamed: 0  date_local    state_name   county_name      city_name  \
0           0  2018-01-01       Arizona      Maricopa        Buckeye
1           1  2018-01-01          Ohio       Belmont      Shadyside
2           2  2018-01-01       Wyoming         Teton  Not in a city
```

```
3           3  2018-01-01  Pennsylvania  Philadelphia   Philadelphia
4           4  2018-01-01          Iowa          Polk     Des Moines

                                    local_site_name   parameter_name  \
0                                             BUCKEYE  Carbon monoxide
1                                           Shadyside  Carbon monoxide
2  Yellowstone National Park - Old Faithful Snow …  Carbon monoxide
3                            North East Waste (NEW)  Carbon monoxide
4                                         CARPENTER  Carbon monoxide

    units_of_measure  arithmetic_mean  aqi
0  Parts per million         0.473684    7
1  Parts per million         0.263158    5
2  Parts per million         0.111111    2
3  Parts per million         0.300000    3
4  Parts per million         0.215789    3
Use describe() to summarize AQI
        Unnamed: 0  date_local  state_name  county_name      city_name  \
count   260.000000         260         260          260            260
unique         NaN           1          52          149            190
top            NaN  2018-01-01  California  Los Angeles  Not in a city
freq           NaN         260          66           14             21
mean    129.500000         NaN         NaN          NaN            NaN
std      75.199734         NaN         NaN          NaN            NaN
min       0.000000         NaN         NaN          NaN            NaN
25%      64.750000         NaN         NaN          NaN            NaN
50%     129.500000         NaN         NaN          NaN            NaN
75%     194.250000         NaN         NaN          NaN            NaN
max     259.000000         NaN         NaN          NaN            NaN

       local_site_name   parameter_name   units_of_measure  arithmetic_mean  \
count              257              260                260       260.000000
unique             253                1                  1              NaN
top            Kapolei  Carbon monoxide  Parts per million              NaN
freq                 2              260                260              NaN
mean               NaN              NaN                NaN         0.403169
std                NaN              NaN                NaN         0.317902
min                NaN              NaN                NaN         0.000000
25%                NaN              NaN                NaN         0.200000
50%                NaN              NaN                NaN         0.276315
75%                NaN              NaN                NaN         0.516009
max                NaN              NaN                NaN         1.921053

               aqi
count   260.000000
unique         NaN
top            NaN
freq           NaN
```

```
mean      6.757692
std       7.061707
min       0.000000
25%       2.000000
50%       5.000000
75%       9.000000
max      50.000000
For a more thorough examination of observations by state use values_counts()
California              66
Arizona                 14
Ohio                    12
Florida                 12
Texas                   10
New York                10
Pennsylvania            10
Michigan                 9
Colorado                 9
Minnesota                7
New Jersey               6
Indiana                  5
North Carolina           4
Massachusetts            4
Maryland                 4
Oklahoma                 4
Virginia                 4
Nevada                   4
Connecticut              4
Kentucky                 3
Missouri                 3
Wyoming                  3
Iowa                     3
Hawaii                   3
Utah                     3
Vermont                  3
Illinois                 3
New Hampshire            2
District Of Columbia     2
New Mexico               2
Montana                  2
Oregon                   2
Alaska                   2
Georgia                  2
Washington               2
Idaho                    2
Nebraska                 2
Rhode Island             2
Tennessee                2
Maine                    2
```

```
South Carolina           1
Puerto Rico              1
Arkansas                 1
Kansas                   1
Mississippi              1
Alabama                  1
Louisiana                1
Delaware                 1
South Dakota             1
West Virginia            1
North Dakota             1
Wisconsin                1
Name: state_name, dtype: int64
for a more
```

HINT 1

Consider referring to the material on descriptive statisics.

HINT 2

Consider using `pandas` or `numpy` to explore the `aqi` dataframe.

HINT 3

Any of the following functions may be useful: - `pandas`: `describe()`,`value_counts()`,`shape()`, `head()` - `numpy`: `unique()`,`mean()`

Question 1: From preceding data exploration, what do you recognize?

- You have county-level data for the first hypothesis.
- Ohio and New York both have a higher number of observations to work with in this dataset.

1.4 Step 3. Statistical Tests

Before you proceed, recall the following steps for conducting hypothesis testing:

1. Formulate the null hypothesis and the alternative hypothesis.
2. Set the significance level.
3. Determine the appropriate test procedure.
4. Compute the p-value.
5. Draw your conclusion.

1.4.1 Hypothesis 1: ROA is considering a metropolitan-focused approach. Within California, they want to know if the mean AQI in Los Angeles County is statistically different from the rest of California.

Before proceeding with your analysis, it will be helpful to subset the data for your comparison.

[4]:
```
# Create dataframes for each sample being compared in your test

### YOUR CODE HERE ###

ca_la = aqi[aqi['county_name']=='Los Angeles']
ca_other = aqi[(aqi['state_name']=='California') & (aqi['county_name']!='Los␣
↪Angeles')]
```

HINT 1

Consider referencing the material on subsetting dataframes.

HINT 2

Consider creating two dataframes, one for Los Angeles, and one for all other California observations.

HINT 3

For your first dataframe, filter to `county_name` of `Los Angeles`. For your second dataframe, filter to `state_name` of `Calfornia` and `county_name` not equal to `Los Angeles`.

Formulate your hypothesis: Formulate your null and alternative hypotheses:

- H_0: There is no difference in the mean AQI between Los Angeles County and the rest of California.
- H_A: There is a difference in the mean AQI between Los Angeles County and the rest of California.

Set the significance level:

[5]:
```
# For this analysis, the significance level is 5%

significance_level = 0.05
significance_level
```

[5]:
```
0.05
```

Determine the appropriate test procedure: Here, you are comparing the sample means between two independent samples. Therefore, you will utilize a **two-sample -test**.

Compute the p-value

[6]:
```
# Compute your p-value here

### YOUR CODE HERE ###

stats.ttest_ind(a=ca_la['aqi'], b=ca_other['aqi'], equal_var=False)
```

[6]:
```
Ttest_indResult(statistic=2.1107010796372014, pvalue=0.049839056842410995)
```

HINT 1

Consider referencing the material on how to perform a two-sample t-test.

HINT 2

In `ttest_ind()`, a is the aqi column from the "Los Angeles" dataframe, and b is the aqi column from the "Other California" dataframe.

HINT 3

Be sure to set `equal_var` = False.

Question 2. What is your p-value for hypothesis 1, and what does this indicate for your null hypothesis? With a p-value (0.049) being less than 0.05 (as your significance level is 5%), reject the null hypothesis in favor of the alternative hypothesis.

Therefore, a metropolitan strategy may make sense in this case.

1.4.2 Hypothesis 2: With limited resources, ROA has to choose between New York and Ohio for their next regional office. Does New York have a lower AQI than Ohio?

Before proceeding with your analysis, it will be helpful to subset the data for your comparison.

[7]:
```
# Create dataframes for each sample being compared in your test

### YOUR CODE HERE ###

ny = aqi[aqi['state_name']=='New York']
ohio = aqi[aqi['state_name']=='Ohio']
```

HINT 1

Consider referencing the materials on subsetting dataframes.

HINT 2

Consider creating two dataframes, one for New York, and one for Ohio observations.

HINT 3

For your first dataframe, filter to `state_name` of `New York`. For your second dataframe, filter to `state_name` of 'Ohio'.

Formulate your hypothesis: Formulate your null and alternative hypotheses:

- H_0: The mean AQI of New York is greater than or equal to that of Ohio.
- H_A: The mean AQI of New York is **below** that of Ohio.

Significance Level (remains at 5%)

Determine the appropriate test procedure: Here, you are comparing the sample means between two independent samples in one direction. Therefore, you will utilize a **two-sample -test**.

Compute the p-value

[8]:
```
# Compute your p-value here

### YOUR CODE HERE ###

tstat, pvalue = stats.ttest_ind(a=ny['aqi'], b=ohio['aqi'], alternative='less',␣
↪equal_var=False)
print(tstat)
print(pvalue)
```

```
-2.025951038880333
0.030446502691934697
```

HINT 1

Consider referencing the material on how to perform a two-sample t-test.

HINT 2

In `ttest_ind()`, a is the aqi column from the "New York" dataframe, and b is the aqi column from the "Ohio" dataframe.

HINT 3

You can assign `tstat`, `pvalue` to the output of `ttest_ind`. Be sure to include `alternative = less` as part of your code.

Question 3. What is your p-value for hypothesis 2, and what does this indicate for your null hypothesis? With a p-value (0.030) of less than 0.05 (as your significance level is 5%) and a t-statistic < 0 (-2.036), **reject the null hypothesis in favor of the alternative hypothesis**.

Therefore, you can conclude at the 5% significance level that New York has a lower mean AQI than Ohio.

1.4.3 Hypothesis 3: A new policy will affect those states with a mean AQI of 10 or greater. Will Michigan be affected by this new policy?

Before proceeding with your analysis, it will be helpful to subset the data for your comparison.

[9]:
```
# Create dataframes for each sample being compared in your test

### YOUR CODE HERE ###

michigan = aqi[aqi['state_name']=='Michigan']
```

HINT 1

Consider referencing the material on subsetting dataframes.

HINT 2

Consider creating one dataframe which only includes Michigan.

Formulate your hypothesis: Formulate your null and alternative hypotheses here:

- H_0: The mean AQI of Michigan is less than or equal to 10.
- H_A: The mean AQI of Michigan is greater than 10.

Significance Level (remains at 5%)

Determine the appropriate test procedure: Here, you are comparing one sample mean relative to a particular value in one direction. Therefore, you will utilize a **one-sample -test**.

Compute the P-value

[10]:
```
# Compute your p-value here

### YOUR CODE HERE ###

tstat, pvalue = stats.ttest_1samp(michigan['aqi'], 10, alternative='greater')
print(tstat)
print(pvalue)
```

```
-1.7395913343286131
0.9399405193140109
```

HINT 1

Consider referencing the material on how to perform a one-sample t-test.

HINT 2

In `ttest_1samp`), you are comparing the AQI column from your Michigan data relative to 10, the new policy threshold.

HINT 3

You can assign `tstat`, `pvalue` to the output of `ttest_1samp`. Be sure to include `alternative = greater` as part of your code.

Question 4. What is your p-value for hypothesis 3, and what does this indicate for your null hypothesis? With a p-value (0.940) being greater than 0.05 (as your significance level is 5%) and a t-statistic < 0 (-1.74), **fail to reject the null hypothesis**.

Therefore, you cannot conclude at the 5% significance level that Michigan's mean AQI is greater than 10. This implies that Michigan would most likely not be affected by the new policy.

1.5 Step 4. Results and Evaluation

Now that you've completed your statistical tests, you can consider your hypotheses and the results you gathered.

Question 5. Did your results show that the AQI in Los Angeles County was statistically different from the rest of California? Yes, the results indicated that the AQI in Los Angeles County was in fact different from the rest of California.

Question 6. Did New York or Ohio have a lower AQI? Using a 5% significance level, you can conclude that New York has a lower AQI than Ohio based on the results.

Question 7: Will Michigan be affected by the new policy impacting states with a mean AQI of 10 or greater? Based on the tests, you would fail to reject the null hypothesis, meaning you can't conclude that the mean AQI is greater than 10. Thus, it is unlikely that Michigan would be affected by the new policy.

2 Conclusion

What are key takeaways from this lab?

Even with small sample sizes, the variation within the data is enough to allow you to make statistically significant conclusions. You identified at the 5% significance level that the Los Angeles mean AQI was stastitically different from the rest of California, and that New York does have a lower mean AQI than Ohio. However, you were unable to conclude at the 5% significance level that Michigan's mean AQI was greater than 10.

What would you consider presenting to your manager as part of your findings?

For each test, you would present the null and alternative hypothesis, then describe your conclusion and the resulting p-value that drove that conclusion. As the setup of t-test's have a few key configurations that dictate how you interpret the result, you would specify the type of test you chose, whether that tail was one-tail or two-tailed, and how you performed the t-test from `stats`.

What would you convey to external stakeholders?

In answer to the research questions posed, you would convey the level of significance (5%) and your conclusion. Additionally, providing the sample statistics being compared in each case will likely provide important context for stakeholders to quickly understand the difference between your results.

Congratulations! You've completed this lab. However, you may not notice a green check mark next to this item on Coursera's platform. Please continue your progress regardless of the check mark. Just click on the "save" icon at the top of this notebook to ensure your work has been logged.