Activity_Explore hypothesis testing

October 25, 2023

1 Activity: 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. Can you rule out Michigan from being 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

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. `Pandas` was used to import the file `c4_epa_air_quality.csv` as a dataframe named `aqi`. As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the .csv file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

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

Load Dataset

[2]:
```
# RUN THIS CELL 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.

[23]:
```
# Explore your dataframe `aqi` here:

aqi.head(10)
```

[23]:

	Unnamed: 0	date_local	state_name	county_name	city_name	local_site_name	parameter_name
0	0	2018-01-01	Arizona	Maricopa	Buckeye	BUCKEYE	Carbon monoxide
1	1	2018-01-01	Ohio	Belmont	Shadyside	Shadyside	Carbon monoxide
2	2	2018-01-01	Wyoming	Teton	Not in a city	Yellowstone National Park - Old Faithful Snow …	Carbon monoxide
3	3	2018-01-01	Pennsylvania	Philadelphia	Philadelphia	North East Waste (NEW)	Carbon monoxide
4	4	2018-01-01	Iowa	Polk	Des Moines	CARPENTER	Carbon monoxide
5	5	2018-01-01	Hawaii	Honolulu	Not in a city		
6	6	2018-01-01	Hawaii	Honolulu	Not in a city		
7	7	2018-01-01	Pennsylvania	Erie	Erie		
8	8	2018-01-01	Hawaii	Honolulu	Honolulu		
9	9	2018-01-01	Colorado	Larimer	Fort Collins		

```
5                                          Kapolei  Carbon monoxide
6                                          Kapolei  Carbon monoxide
7                                              NaN  Carbon monoxide
8                                         Honolulu  Carbon monoxide
9              Fort Collins - CSU - S. Mason  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
5  Parts per million         0.994737   14
6  Parts per million         0.200000    2
7  Parts per million         0.200000    2
8  Parts per million         0.400000    5
9  Parts per million         0.300000    6
```

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 the preceding data exploration, what do you recognize? From the information in the dataset it looks like we can continue with answering the questions asked.

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.

[24]:
```
# Create dataframes for each sample being compared in your test

la_aqi = aqi[aqi['county_name']=='Los Angeles']
ca_aqi = 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:

[25]:
```
# For this analysis, the significance level is 5%

sig_level = 0.05
sig_level
```

[25]:
```
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

[26]:
```
# Compute your p-value here

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

[26]:
```
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 our "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 the P-value being 0.049 we are less than 5% so we reject the null and move forward with a metro strategy.

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.

[28]:
```
# Create dataframes for each sample being compared in your test

ny_aqi = aqi[aqi['state_name']=='New York']
print(ny_aqi.head(10))
oh_aqi = aqi[aqi['state_name']=='Ohio']
print(oh_aqi.head(10))
```

```
     Unnamed: 0  date_local state_name county_name      city_name  \
90           90  2018-01-01   New York        Erie    Cheektowaga
113         113  2018-01-01   New York       Bronx       New York
124         124  2018-01-01   New York      Monroe      Rochester
167         167  2018-01-01   New York    New York       New York
173         173  2018-01-01   New York      Queens       New York
182         182  2018-01-01   New York      Queens       New York
184         184  2018-01-01   New York     Steuben  Not in a city
195         195  2018-01-01   New York        Erie        Buffalo
196         196  2018-01-01   New York      Monroe      Rochester
234         234  2018-01-01   New York      Albany         Albany

              local_site_name   parameter_name   units_of_measure  \
90          Buffalo Near-Road  Carbon monoxide  Parts per million
113           PFIZER LAB SITE  Carbon monoxide  Parts per million
124               ROCHESTER 2  Carbon monoxide  Parts per million
167                      CCNY  Carbon monoxide  Parts per million
173  Queens College Near Road  Carbon monoxide  Parts per million
182          QUEENS COLLEGE 2  Carbon monoxide  Parts per million
184       PINNACLE STATE PARK  Carbon monoxide  Parts per million
195                   BUFFALO  Carbon monoxide  Parts per million
196       Rochester Near-Road  Carbon monoxide  Parts per million
```

```
234              LOUDONVILLE  Carbon monoxide  Parts per million

     arithmetic_mean  aqi
90          0.252632    3
113         0.289474    3
124         0.200000    2
167         0.200000    2
173         0.273684    3
182         0.200000    2
184         0.200000    2
195         0.300000    3
196         0.200000    2
234         0.221053    3
     Unnamed: 0  date_local state_name county_name             city_name  \
1             1  2018-01-01       Ohio     Belmont             Shadyside
12           12  2018-01-01       Ohio    Hamilton            Cincinnati
22           22  2018-01-01       Ohio       Stark                Canton
51           51  2018-01-01       Ohio      Summit                 Akron
59           59  2018-01-01       Ohio    Cuyahoga             Cleveland
120         120  2018-01-01       Ohio    Cuyahoga             Cleveland
149         149  2018-01-01       Ohio    Franklin              Columbus
191         191  2018-01-01       Ohio    Franklin              Columbus
215         215  2018-01-01       Ohio    Cuyahoga  Warrensville Heights
231         231  2018-01-01       Ohio  Montgomery                Dayton

         local_site_name   parameter_name   units_of_measure  arithmetic_mean  \
1              Shadyside  Carbon monoxide  Parts per million         0.263158
12           Taft NCore  Carbon monoxide  Parts per million         0.252632
22               Canton  Carbon monoxide  Parts per million         0.394737
51         NIHF STEM MS  Carbon monoxide  Parts per million         0.083333
59        GT Craig NCore  Carbon monoxide  Parts per million         0.250000
120             Galleria  Carbon monoxide  Parts per million         0.273684
149            Morse Rd  Carbon monoxide  Parts per million         0.184211
191  Smoky Row Near Road  Carbon monoxide  Parts per million         0.115789
215  Cleveland Near Road  Carbon monoxide  Parts per million         0.321053
231              Reibold  Carbon monoxide  Parts per million         0.163158

     aqi
1      5
12     3
22     6
51     3
59     3
120    3
149    3
191    2
215    5
231    2
```

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

[31]:
```
# Compute your p-value here

tstat, pvalue = stats.ttest_ind(a=ny_aqi['aqi'],b=oh_aqi['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, an 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? P-value equalling 0.030 and less than the 5% new york has a lower aqi than ohio. The t-stat is -2.025

1.4.3 Hypothesis 3: A new policy will affect those states with a mean AQI of 10 or greater. Can you rule out Michigan from being affected by this new policy?

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

[32]:
```
# Create dataframes for each sample being compared in your test

mi_aqi=aqi[aqi['state_name']=='Michigan']
print(mi_aqi.head(10))
```

```
     Unnamed: 0  date_local state_name county_name      city_name  \
65           65  2018-01-01   Michigan       Wayne        Livonia
122         122  2018-01-01   Michigan       Wayne        Detroit
123         123  2018-01-01   Michigan       Wayne        Detroit
129         129  2018-01-01   Michigan       Wayne        Detroit
192         192  2018-01-01   Michigan       Wayne     Allen Park
207         207  2018-01-01   Michigan       Wayne  Not in a city
226         226  2018-01-01   Michigan        Kent   Grand Rapids
242         242  2018-01-01   Michigan       Wayne        Detroit
248         248  2018-01-01   Michigan       Wayne        Detroit

              local_site_name   parameter_name   units_of_measure  \
65                 LIVONIA-NR  Carbon monoxide  Parts per million
122               West corner  Carbon monoxide  Parts per million
123  MARK TWAIN MIDDLE SCHOOL  Carbon monoxide  Parts per million
129                   ELIZA-NR  Carbon monoxide  Parts per million
192                Allen Park  Carbon monoxide  Parts per million
207            Eliza Downwind  Carbon monoxide  Parts per million
226                 GR-MONROE  Carbon monoxide  Parts per million
242        (Northeast corner)  Carbon monoxide  Parts per million
248                 NORTHWEST  Carbon monoxide  Parts per million

     arithmetic_mean  aqi
65          0.338889    5
122         0.394737    8
123         0.515789    9
129         0.616667   11
192         0.811111   13
207         0.516667   10
226         0.200000    2
242         0.378947    7
248         0.415789    8
```

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

[33]:
```
# Compute your p-value here

tstat, pvalue = stats.ttest_1samp(mi_aqi['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? The p-value for Michigan is 0.939 which is lower than 10. So Michigan will not be affected but should be looked at as it will soon be greater than 10.

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, after performing the testing it does show Los Angeles was above the rest of the state.

Question 6. Did New York or Ohio have a lower AQI? New York has a lower AQI than Ohio.

Question 7: Will Michigan be affected by the new policy impacting states with a mean AQI of 10 or greater? No I fail to reject the null because it is lower than 10. It's close to 10 but still less than.

2 Conclusion

What are key takeaways from this lab? Los Angeles is a problem area for AQI for the state of California. New York is less in AQI than Ohio, and that Michigan does not fall under being higher than 10 AQI.

What would you consider presenting to your manager as part of your findings? I would be presenting that LA does need a metropolitan strategy in place. That Michigan falls out of the over 10 areas but it's growing close to being over the mark, we should look more into that to prevent future issues.

What would you convey to external stakeholders? I would convey that the metro project should be done based off testing and that the state with the lower AQI is New York so if they wish to build an office New York would be their best option.

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.