EPA Air Quality Hypothesis Testing

August 4, 2023

1 EPA Air Quality Hypothesis Testing

1.1 Step 1: Imports

Import Packages

[1]:
```
# Import relevant packages

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

Load Dataset

[2]:
```
aqi = pd.read_csv('c4_epa_air_quality.csv')
```

1.2 Step 2: Data Exploration

1.2.1 Before proceeding to your deliverables, explore your datasets.

[3]:
```
# Explore your dataframe `aqi` here:
aqi.describe()
```

[3]:

	Unnamed: 0	arithmetic_mean	aqi
count	260.000000	260.000000	260.000000
mean	129.500000	0.403169	6.757692
std	75.199734	0.317902	7.061707
min	0.000000	0.000000	0.000000
25%	64.750000	0.200000	2.000000
50%	129.500000	0.276315	5.000000
75%	194.250000	0.516009	9.000000
max	259.000000	1.921053	50.000000

[4]:
```
aqi.shape
```

[4]:
```
(260, 10)
```

[5]:
```
aqi.head()
```

[5]:

	Unnamed: 0	date_local	state_name	county_name	city_name	local_site_name	parameter_name	units_of_measure	arithmetic_mean	aqi
0	0	2018-01-01	Arizona	Maricopa	Buckeye	BUCKEYE	Carbon monoxide	Parts per million	0.473684	7
1	1	2018-01-01	Ohio	Belmont	Shadyside	Shadyside	Carbon monoxide	Parts per million	0.263158	5
2	2	2018-01-01	Wyoming	Teton	Not in a city	Yellowstone National Park - Old Faithful Snow …	Carbon monoxide	Parts per million	0.111111	2
3	3	2018-01-01	Pennsylvania	Philadelphia	Philadelphia	North East Waste (NEW)	Carbon monoxide	Parts per million	0.300000	3
4	4	2018-01-01	Iowa	Polk	Des Moines	CARPENTER	Carbon monoxide	Parts per million	0.215789	3

[6]:
```
aqi["state_name"].value_counts()
```

[6]:
```
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
```

[7]:
```
aqi["aqi"].mean()
```

[7]:
```
6.757692307692308
```

1.3 Step 3. Statistical Tests

1.3.1 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.

[8]:
```
# Create dataframes for each sample being compared in your test

la_aqi=aqi[aqi["county_name"]=="Los Angeles"]
```

```
cal_aqi=aqi[(aqi["state_name"]=="California") & (aqi["county_name"]!="Los␣
↪Angeles")]
```

Formulate your hypothesis:

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

Set the significance level: For this analysis, the significance level is 5%

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

Compute the P-value

[25]:
```
# Compute your p-value here
stats.ttest_ind(a=la_aqi["aqi"], b=cal_aqi["aqi"], equal_var=False)
```

[25]:
```
Ttest_indResult(statistic=2.1107010796372014, pvalue=0.049839056842410995)
```

1.3.2 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?

[15]:
```
# Create dataframes for each sample being compared in your test

aqi_ny=aqi[aqi["state_name"]=="New York"]
aqi_oh=aqi[aqi["state_name"]=="Ohio"]
```

Formulate your hypothesis:

- 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, we are comparing the sample means between two independent samples. Therefore, we will apply a **two-sample -test**.

Compute the P-value

[20]:
```
# Computer your p-value here
stats.ttest_ind(a=aqi_ny["aqi"], b=aqi_oh["aqi"], alternative="less",␣
↪equal_var=False)
```

[20]:
```
Ttest_indResult(statistic=-2.025951038880333, pvalue=0.030446502691934697)
```

1.3.3 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?

[21]:
```
# Create dataframes for each sample being compared in your test

aqi_mic=aqi[aqi["state_name"]=="Michigan"]
```

Formulate your hypothesis:

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

Significance Level (remains at 5%)

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

Compute the P-value

[24]:
```
# Computer your p-value here

stats.ttest_1samp(aqi_mic["aqi"], 10, alternative="less")
```

[24]:
```
Ttest_1sampResult(statistic=-1.7395913343286131, pvalue=0.06005948068598906)
```

1.4 Step 4. Conclusion

- Clearly articulating null and alternative hypotheses for each test is important to choose the relevant hypothesis test and related conclusions to be drawn.
- Even with small sample sizes, the data variation allows for statistically significant conclusions.
- At the 5% significance level we can conclude that:
  - The mean AQI in Los Angeles is statistically different from the rest of California.
  - New York has a lower mean AQI than Ohio.
  - Michigan's mean AQI was less than 10.