Hypothesis testing - Air Quality Index (AQI)

July 11, 2024

By Nitish Adhikari

Email id :nitishbuzzpro@gmail.com , +91-9650740295

Linkedin : https://www.linkedin.com/in/nitish-adhikari-6b2350248

1 Hypothesis testing - Air Quality Index (AQI)

1.1 Introduction

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. Will Michigan be affected by this new policy?

Notes: 1. 5% level of significance.

1.2 Step1 : Import Packages

[1]:
```
import pandas as pd
from scipy import stats
```

Load Dataset

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

1.3 Step2 : Data Exploration

[3]:
```
df
```

[3]:
```
     Unnamed: 0  date_local            state_name           county_name  \
0             0  2018-01-01               Arizona              Maricopa
1             1  2018-01-01                  Ohio               Belmont
2             2  2018-01-01               Wyoming                 Teton
3             3  2018-01-01          Pennsylvania          Philadelphia
4             4  2018-01-01                  Iowa                  Polk
..          …         …                    …                    …
255         255  2018-01-01  District Of Columbia  District of Columbia
256         256  2018-01-01             Wisconsin                 Dodge
257         257  2018-01-01              Kentucky              Jefferson
258         258  2018-01-01              Nebraska                Douglas
259         259  2018-01-01        North Carolina                  Wake

         city_name                                      local_site_name  \
0          Buckeye                                             BUCKEYE
1        Shadyside                                           Shadyside
2    Not in a city  Yellowstone National Park - Old Faithful Snow …
3     Philadelphia                                North East Waste (NEW)
4       Des Moines                                           CARPENTER
..             …                                                   …
255     Washington                                           Near Road
256       Kekoskee                                HORICON WILDLIFE AREA
257     Louisville                                         CANNONS LANE
258          Omaha                                                 NaN
259  Not in a city                                          Triple Oak

      parameter_name   units_of_measure  arithmetic_mean  aqi
0    Carbon monoxide  Parts per million         0.473684    7
1    Carbon monoxide  Parts per million         0.263158    5
2    Carbon monoxide  Parts per million         0.111111    2
3    Carbon monoxide  Parts per million         0.300000    3
4    Carbon monoxide  Parts per million         0.215789    3
..               …                 …              …  …
255  Carbon monoxide  Parts per million         0.244444    3
256  Carbon monoxide  Parts per million         0.200000    2
257  Carbon monoxide  Parts per million         0.163158    2
258  Carbon monoxide  Parts per million         0.421053    9
259  Carbon monoxide  Parts per million         0.188889    2

[260 rows x 10 columns]
```

[4]:
```
df.describe(include = 'all')
```

[4]:

	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

[5]:
```
df.shape
```

[5]: (260, 10)

Points from the preceding data exploration

1. California state has highest count among states
2. Los Angles city has highest count among counties

3. There are 52 states and 159 cities in the dataset
4. All the readings are on the same day.
5. Majority of reading are not in the cities.
6. Mean aqi is approx 6.5
7. 75 % of the reading are equal or less than 9 aqi.

1.4 Step 3. Statistical Tests

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 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.

[10]:
```
# Create dataframes for each sample being compared
df_losangeles = df[df['county_name']=='Los Angeles']
df_california = df[(df['state_name']=='California') & (df['county_name']!='Los␣
↪Angeles')]
```

Formulate hypothesis: Formulate 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:

[7]:
```
# For this analysis, the significance level is 5%
significance_level = 0.05
```

Determine the appropriate test procedure: For comparing the sample means between two independent samples, utilize a **two-sample -test**.

Compute the P-value

[12]:
```
t_stat,p_val = stats.
↪ttest_ind(df_losangeles['aqi'],df_california['aqi'],equal_var=False)
```

[13]:
```
print('P-value for hypothesis 1: ',p_val)
print('T-Statistic for hypothesis 1: ',t_stat)

if p_val <= significance_level:
    print('Reject Null Hypothesis. There is a statistical evidence that there␣
 ↪is difference in the mean AQI between Los Angeles County and the rest of␣
 ↪California.')
else:
    print('Fail to reject Null Hypothesis. There not enough statistical␣
 ↪evidence that there is difference in the mean AQI between Los Angeles County␣
 ↪and the rest of California.')
```

```
P-value for hypothesis 1:  0.049839056842410995
T-Statistic for hypothesis 1:  2.1107010796372014
Reject Null Hypothesis. There is a statistical evidence that there is difference
in the mean AQI between Los Angeles County and the rest of California.
```

[ ]:
```

```

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?

[14]:
```
# Create dataframes for each sample being compared
df_newyork = df[df['state_name'] == 'New York']
df_ohio = df[df['state_name'] == 'Ohio']
```

Formulate hypothesis: Formulate 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: For comparing the sample means between two independent samples, utilize a **two-sample -test**.

Compute the P-value

[16]:
```
t_stat,p_val = stats.
 ↪ttest_ind(df_newyork['aqi'],df_ohio['aqi'],alternative='less',equal_var=False)
```

[17]:
```
print('P-value for hypothesis 2: ',p_val)
print('T-Statistic for hypothesis 2: ',t_stat)
```

```
if p_val <= significance_level:
    print('Reject Null Hypothesis. There is a statistical evidence that the␣
 ↪mean AQI of New York is below that of Ohio.')
else:
    print('Fail to reject Null Hypothesis. There not enough statistical␣
 ↪evidence that the mean AQI of New York is below that of Ohio')
```

```
P-value for hypothesis 2:  0.030446502691934697
T-Statistic for hypothesis 2:  -2.025951038880333
Reject Null Hypothesis. There is a statistical evidence that the mean AQI of New
York is below that of 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?

[23]:
```
# Create dataframes for each sample being compared
df_michigan = df[df['state_name'] == '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: comparing one sample mean relative to a particular value in one direction, utilize a **one-sample -test**.

Compute the P-value

[28]:
```
t_stat, p_value = stats.ttest_1samp(df_michigan['aqi'], 10,␣
 ↪alternative='greater')
```

[29]:
```
print('P-value for hypothesis 3: ',p_val)
print('T-Statistic for hypothesis 3: ',t_stat)

if p_val <= significance_level:
    print('Reject Null Hypothesis. There is a statistical evidence that the␣
 ↪mean AQI of Michigan is greater than 10')
else:
    print('Fail to reject Null Hypothesis. There not enough statistical␣
 ↪evidence that The mean AQI of Michigan is greater than 10')
```

```
P-value for hypothesis 3:  0.060893005383869395
T-Statistic for hypothesis 3:  -1.7395913343286131
Fail to reject Null Hypothesis. There not enough statistical evidence that The
mean AQI of Michigan is greater than 10
```

1.5 Step 4. Results and Evaluation

is 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.

Did New York or Ohio have a lower AQI?** Using a 5% significance level, New York has a lower AQI than Ohio based on the results.

Will Michigan be affected by the new policy impacting states with a mean AQI of 10 or greater?** it is unlikely that Michigan would be affected by the new policy.