FODS Assignment - 3

Abirami S
CB.EN.U4CSE20202

RACIAL AND ETHNIC DISPARITIES IN ALAMEDA COUNTY JURY POOLS

Hypothesis Testing:
Multiple Categories Problem

Attributes: Ethnicity , Eligible , Panels:

Null Hypothesis ['H0']: Panel was chosen at random from eligible population
Alternate Hypothesis ['H1']: Panel was not chosen at random from eligible population

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

```
jury = {"Ethnicity":["Asian", "Black", "Latino", "White", "Other"], "Eligible":[0.15, 0.18, 0.12, 0.54, 0.01], "Panels":[0.26, 0.08, 0.08, 0.54, 0.04]}
jury
```

```
{'Ethnicity': ['Asian', 'Black', 'Latino', 'White', 'Other'],
 'Eligible': [0.15, 0.18, 0.12, 0.54, 0.01],
 'Panels': [0.26, 0.08, 0.08, 0.54, 0.04]}
```

```
Alameda_df = pd.DataFrame(jury)
Alameda_df
```

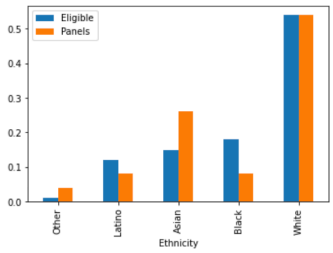
	Ethnicity	Eligible	Panels
0	Asian	0.15	0.26
1	Black	0.18	0.08
2	Latino	0.12	0.08
3	White	0.54	0.54
4	Other	0.01	0.04

```
Alameda_df = Alameda_df.set_index("Ethnicity")
Alameda_df
```

Ethnicity	Eligible	Panels
Asian	0.15	0.26
Black	0.18	0.08
Latino	0.12	0.08
White	0.54	0.54
Other	0.01	0.04

Plotting the bar for each group (Ethnicity and Panel)

```
Alameda_df.sort_values("Eligible").plot.bar()
plt.show()
```




```
Alameda_df['jury_diff'] = Alameda_df['Eligible'] - Alameda_df['Panels']
Alameda_df
```

Ethnicity	Eligible	Panels	jury_diff
Asian	0.15	0.26	-0.11
Black	0.18	0.08	0.10
Latino	0.12	0.08	0.04
White	0.54	0.54	0.00
Other	0.01	0.04	-0.03

- It can be observed that there is a significant difference for the Asians and Blacks as compared to the other ethnicities.
- It seems partial that the whites do not have any difference i.e, all the eligible whites are recruited for the panel unlike the other ethnicity.

Absolute Values Comparison

```
Alameda_df['Absolute_difference'] = abs(Alameda_df['Eligible'] - Alameda_df['Panels'])
Alameda_df
```

Ethnicity	Eligible	Panels	Absolute_difference
Asian	0.15	0.26	0.11
Black	0.18	0.08	0.10
Latino	0.12	0.08	0.04
White	0.54	0.54	0.00
Other	0.01	0.04	0.03

```
test_statistic = Alameda_df['Absolute_difference'].sum()
test_statistic
```

```
0.28
```

```
def total_variation_distance(distribution_1, distribution_2):
    return np.abs(distribution_1 - distribution_2).sum()/2
```

```
original= np.abs(Alameda_df['Eligible'] - Alameda_df['Panels']).sum()
original
```

```
0.28
```

- Both are the same value which implies excess and the difference values are the same

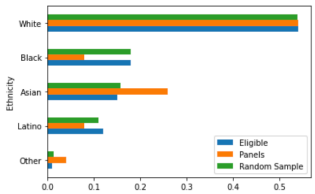
RANDOM SAMPLING:

```
Alameda_df_1 = pd.DataFrame(Alameda_df, columns = ['Eligible','Panels'])
```

```
Alameda_df_1['Random Sample'] = np.random.multinomial(1453, [0.15, 0.18, 0.12, 0.54, 0.01])/1453
Alameda_df_1
```

Ethnicity	Eligible	Panels	Random Sample
Asian	0.15	0.26	0.157605
Black	0.18	0.08	0.179628
Latino	0.12	0.08	0.110805
White	0.54	0.54	0.538885
Other	0.01	0.04	0.013076

```
Alameda_df_1.sort_values("Eligible").plot.barh()
plt.show()
```




```
TVD = (abs(Alameda_df_1['Eligible'] - Alameda_df_1['Random Sample'])).sum()/2
TVD
```

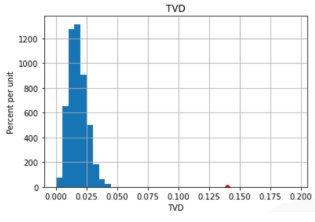
```
0.010681348933241585
```

```
simulations = 5000
panel_size = 1453
tvd=[]
for i in np.arange(simulations):
    Alameda_df_1["Random Sample"] = (np.random.multinomial(1453, [0.15, 0.18, 0.12, 0.54, 0.01]))/panel_size
    tvd.append(total_variation_distance(Alameda_df_1['Eligible'], Alameda_df_1['Random Sample']))
```

```
tvd_df = pd.DataFrame(tvd)
tvd_df.rename(columns={0:"TVD"}, inplace=True)
tvd_df.head()
```

	TVD
0	0.013187
1	0.026139
2	0.012705
3	0.015513
4	0.013469

```
tvd_df.hist(bins = np.arange(0, 0.2, 0.005))
plt.ylabel('Percent per unit')
plt.xlabel('TVD')
plt.scatter(0.14, 0, color='red', s=30)
plt.show()
```



- Random samples typically have lower TVD values than the 0.18 we got for the panel and eligible jurors.
- In this analysis, the data are not clouded by questions as in our previous analysis – the total numbers of people involved were relatively small, and the counting was done carefully for the Supreme Court case.
- Therefore, we can conclude that the panel was not representative of the population based on our analysis. It is hard to accept the Supreme Court's judgement that "the overall percentage disparity has been small.