STAT 107 Final Project

STAT 107 Lecture AL1 | Discussion AYJ
Bryan Ge

In this project, I conduct a variety of hypothesis tests on the "flights.csv" dataset provided on Kaggle.

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
```

I read in the flights dataset from https://uofi.box.com/s/c9b4pdidfrzlhhm9himfm1p7nrr3ws6s by downloading the dataset and placing it in the same folder as my project program.

In [2]:
```
flights = pd.read_csv('flights.csv')
```

```
/Users/bryange/anaconda/anaconda3/lib/python3.7/site-packages/IPython/core/interactiveshell.py:3063: DtypeWarning: Co
lumns (7,8) have mixed types.Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [3]:
```
flights
```

Out[3]:

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE
0	2015	1	1	4	AS	98	N407AS	ANC	SEA	5
1	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	10
2	2015	1	1	4	US	840	N171US	SFO	CLT	20
3	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20
4	2015	1	1	4	AS	135	N527AS	SEA	ANC	25
...	...	...	...	...	...	...	...	...	...	..
5819074	2015	12	31	4	B6	688	N657JB	LAX	BOS	2359
5819075	2015	12	31	4	B6	745	N828JB	JFK	PSE	2359
5819076	2015	12	31	4	B6	1503	N913JB	JFK	SJU	2359
5819077	2015	12	31	4	B6	333	N527JB	MCO	SJU	2359
5819078	2015	12	31	4	B6	839	N534JB	JFK	BQN	2359

5819079 rows × 31 columns

Hypothesis Test 1

In this hypothesis test, I am comparing the average arrival delay between American Airlines and United Airlines.

First, I filter out the flights data for the average arrival delay on flights in December from O'Hare International Airport (Chicago) ORD to Los Angeles International Airport LAX for American Airlines and United Airlines. These are the 2 samples I will be working with.

I drop the missing values in the "ARRIVAL_DELAY" column in the flights.csv file.

In [4]:
```
american = flights.loc[(flights['MONTH'] == 12) & (flights['ORIGIN_AIRPORT'] == 'ORD') & (flights['DESTINATION_AIRPORT
'] == 'LAX') &
              (flights['AIRLINE'] == 'AA'), 'ARRIVAL_DELAY']
american = american.dropna()
american
```

Out[4]:
```
5341787   -33.0
5343348   -35.0
5344407     0.0
5346818   -20.0
5348394     2.0
           ...
5813669    -8.0
5815173   -16.0
5816347    -7.0
5817421    -8.0
5818331     5.0
Name: ARRIVAL_DELAY, Length: 280, dtype: float64
```

In [5]:
```
united = flights.loc[(flights['MONTH'] == 12) & (flights['ORIGIN_AIRPORT'] == 'ORD') & (flights['DESTINATION_AIRPORT']
== 'LAX') &
              (flights['AIRLINE'] == 'UA'), 'ARRIVAL_DELAY']
united = united.dropna()
united
```

Out[5]:
```
5342177   -34.0
5343938   -11.0
5345166   -41.0
5346638   -16.0
5347202    -4.0
           ...
5812349    -1.0
5813682    -2.0
5815399   -19.0
5816709   -14.0
5817847   -19.0
Name: ARRIVAL_DELAY, Length: 302, dtype: float64
```

Test Assumptions

I check that the test assumptions are satisfied by checking that 1 of the following 2 conditions are satisfied:

1. sample 1 and sample 2 are both large (in this course, greater than or equal to 30 means large)
2. population 1 and population 2 are approximately normal

In [6]:
```
len(american) >= 30 and len(united) >= 30
```

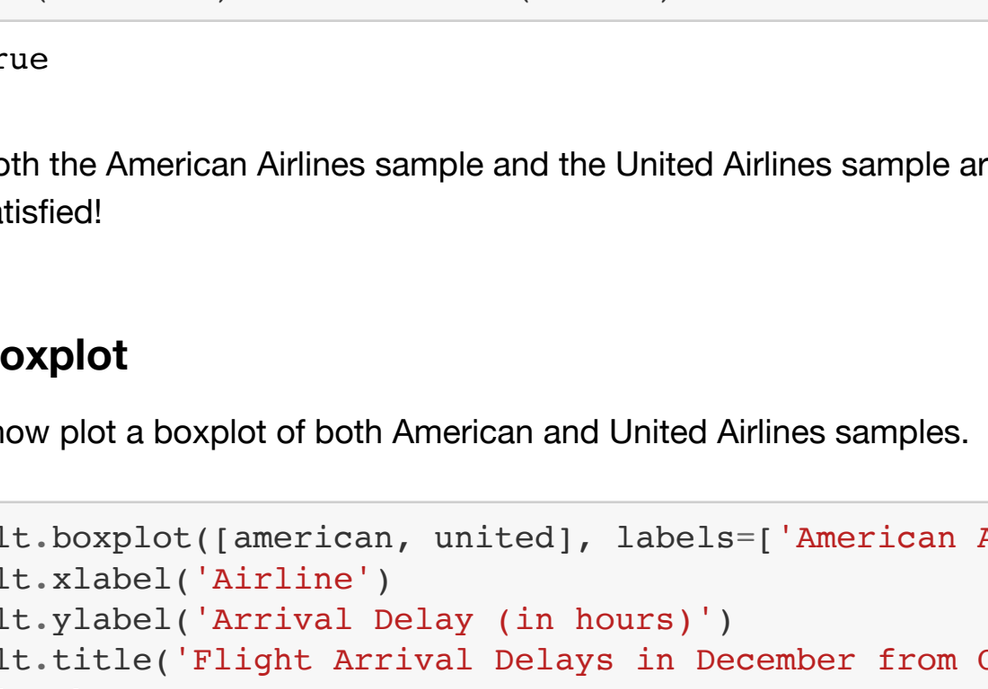
Out[6]:
```
True
```

Both the American Airlines sample and the United Airlines sample are large, or greater than or equal to 30. This means the test assumptions have been satisfied!

Boxplot

I now plot a boxplot of both American and United Airlines samples.

In [7]:
```
plt.boxplot([american, united], labels=['American Airlines', 'United Airlines'])
plt.xlabel('Airline')
plt.ylabel('Arrival Delay (in hours)')
plt.title('Flight Arrival Delays in December from ORD to LAX for Different Airline Companies')
plt.show()
```



Hypotheses Statements

H_0: mu_1 - mu_2 = 0
H_1: mu_1 - mu_2 != 0

Hypothesis Test

I conduct a hypothesis test at significance level alpha = 0.05. I am computing a 2-sample z-test.

First, I calculate the test statistic.

In [8]:
```
american_avg = american.mean()
american_std = american.std()
american_len = len(american)

united_avg = united.mean()
united_std = united.std()
united_len = len(united)
```

In [9]:
```
z_stat = ((american_avg - united_avg) - 0) / (((american_std)**2 / american_len) + ((united_std)**2 / united_len))**0.
5
z_stat
```

Out[9]:
```
-1.3926811323141544
```

Since the alternative hypothesis is the situation in which the flight arrival delay for American Airlines is not equal to the flight arrival delay for United Airlines, I solve for the p-value using the following formula:

p-value = P(Z > abs(z_stat)) + P(Z < -abs(z_stat)) = 2 * P(Z < -abs(z_stat))

In [10]:
```
test_1_p_value = 2 * norm.cdf(-abs(z_stat))
test_1_p_value
```

Out[10]:
```
0.16371623660393797
```

Conclusion

The p-value (0.163716...) is greater than the significance level (0.05) so we fail to reject the null hypothesis.

We conclude the data does not provide enough evidence that the average arrival delay on flights in December from ORD to LAX for American Airlines and United Airlines is not equal.

Hypothesis Test 2

In this hypothesis test, I am comparing the probability of flights delaying over an hour (in December from O'Hare International Airport (Chicago) ORD to Los Angeles International Airport LAX) between American Airlines and United Airlines.

Test Assumptions

I check that the test assumptions are satisfied by assuming that all observations are independent and that the following equalities prove true:

n p_hat >= 10 AND n (1-p_hat) >= 10

In other words, there needs to be at least 10 successes and 10 failures.

First, I calculate the sample proportions. I filter "american" and "united" for flights in which the arrival delay is great than 1 hour and determine how many of these flights exist.

In [11]:
```
american_delay_over_an_hour = american.loc[american > 60]
american_delay_over_an_hour
```

Out[11]:
```
5367174     89.0
5450821    141.0
5505345     76.0
5636478     68.0
5652683    303.0
5668984     70.0
5739095     64.0
5742894    118.0
5749086    187.0
5764393    330.0
5765763    317.0
5773051    225.0
5783647    153.0
5798199    149.0
5807879    178.0
Name: ARRIVAL_DELAY, dtype: float64
```

In [12]:
```
american_n = len(american)
american_n
```

Out[12]:
```
280
```

In [13]:
```
american_x = len(american_delay_over_an_hour)
american_x
```

Out[13]:
```
15
```

In [14]:
```
american_p_hat = american_x / american_n
american_p_hat
```

Out[14]:
```
0.05357142857142857
```

In [15]:
```
united_delay_over_an_hour = united.loc[united > 60]
united_delay_over_an_hour
```

Out[15]:
```
5459106    108.0
5476537     68.0
5519448     89.0
5568218     95.0
5593768    103.0
5627135    165.0
5634245    141.0
5698281    153.0
5743052     95.0
5744836    261.0
5755772     74.0
5756723    101.0
5757636    115.0
5760061     66.0
5765072    345.0
5769363    322.0
5771018     81.0
5772444    171.0
5773866     87.0
5775163    156.0
5776284    147.0
5780555    123.0
5785414    113.0
5787212    186.0
5800153    230.0
5805649    109.0
Name: ARRIVAL_DELAY, dtype: float64
```

In [16]:
```
united_n = len(united)
united_n
```

Out[16]:
```
302
```

In [17]:
```
united_x = len(united_delay_over_an_hour)
united_x
```

Out[17]:
```
26
```

In [18]:
```
united_p_hat = united_x / united_n
united_p_hat
```

Out[18]:
```
0.08609271523178808
```

I test that both American Airlines and United Airlines proportions satisfy the inequality:
n p_hat >= 10 AND n (1-p_hat) >= 10

In [19]:
```
american_n * american_p_hat >= 10 and american_n * (1-american_p_hat) >= 10
```

Out[19]:
```
True
```

In [20]:
```
united_n * united_p_hat >= 10 and united_n * (1-united_p_hat) >= 10
```

Out[20]:
```
True
```

Both proportions satisfy the inequality, meaning the test assumptions have been satisfied!

Hypotheses Statements

H_0: american_p_hat = united_p_hat
H_1: american_p_hat != united_p_hat

Hypothesis Test

I conduct a hypothesis test at significance level alpha = 0.05.

First, I compute the test statistic.

In [21]:
```
p_hat = (american_x + united_x) / (american_n + united_n)
p_hat
```

Out[21]:
```
0.070446735395189
```

In [22]:
```
z_stat = (american_p_hat - united_p_hat) / ((p_hat * (1 - p_hat) * ((1 / american_n) + (1 / united_n)))**0.5)
z_stat
```

Out[22]:
```
-1.5318672735156422
```

Now, I compute the p-value using the formula:

p-value = P(Z > abs(z_stat)) + P(Z < -abs(z_stat)) = 2 * P(Z < -abs(z_stat))

In [23]:
```
test_2_p_value = 2 * norm.cdf(-abs(z_stat))
test_2_p_value
```

Out[23]:
```
0.12555519084273978
```

Conclusion

The p-value (0.12555...) is greater than the significance level (0.05) so we fail to reject the null hypothesis.

We conclude the data does not provide enough evidence that the probability of flights delaying over an hour in December from ORD to LAX for American Airlines and United Airlines is not equal.

Hypothesis Test 3

In this hypothesis test, I am interested in the **probability of flights occurring on Tuesdays (Day 2 out of 7) in December from O'Hare International Airport (Chicago) ORD to Los Angeles International Airport LAX**. I will compare this probability between American Airlines and United Airlines. I will conduct the test at a significance level of 0.05.

First, I filter the flights data for "DAY_OF_WEEK" for AA and UA flights.

In [24]:
```
american = flights.loc[(flights['MONTH'] == 12) & (flights['ORIGIN_AIRPORT'] == 'ORD') & (flights['DESTINATION_AIRPORT
'] == 'LAX') &
              (flights['AIRLINE'] == 'AA'), 'DAY_OF_WEEK']
american
```

Out[24]:
```
5341787    2
5343348    2
5344407    2
5346818    2
5348394    2
          ..
5813669    4
5815173    4
5816347    4
5817421    4
5818331    4
Name: DAY_OF_WEEK, Length: 291, dtype: int64
```

In [25]:
```
united = flights.loc[(flights['MONTH'] == 12) & (flights['ORIGIN_AIRPORT'] == 'ORD') & (flights['DESTINATION_AIRPORT']
== 'LAX') &
              (flights['AIRLINE'] == 'UA'), 'DAY_OF_WEEK']
united
```

Out[25]:
```
5342177    2
5343938    2
5345166    2
5346638    2
5347202    2
          ..
5812349    4
5813682    4
5815399    4
5816709    4
5817847    4
Name: DAY_OF_WEEK, Length: 310, dtype: int64
```

Test Assumptions

I check that the test assumptions are satisfied by assuming that all observations are independent and that the following equalities prove true:

n p_hat >= 10 AND n (1-p_hat) >= 10

In other words, there needs to be at least 10 successes and 10 failures.

First, I calculate the sample proportions. I filter "american" and "united" for flights in which the day of week they occur is Tuesday (DAY_OF_WEEK = 2) and determine how many of these flights exist.

In [26]:
```
american_tuesday = american.loc[american == 2]
american_tuesday
```

Out[26]:
```
5341787    2
5343348    2
5344407    2
5346818    2
5348394    2
5349931    2
5351408    2
5352843    2
5355375    2
5449267    2
5450821    2
5451860    2
5454277    2
5455837    2
5457351    2
5458803    2
5460227    2
5462656    2
5556653    2
5558215    2
5559260    2
5561672    2
5563229    2
5564751    2
5566201    2
5567633    2
5570098    2
5668984    2
5670315    2
5671357    2
5673832    2
5675201    2
5676885    2
5678331    2
5679809    2
5681384    2
5682502    2
5775778    2
5777100    2
5778135    2
5780606    2
5781967    2
5783647    2
5785088    2
5786556    2
5788125    2
5789248    2
Name: DAY_OF_WEEK, dtype: int64
```

In [27]:
```
american_n = len(american)
american_p_hat = len(american_tuesday) / american_n
american_p_hat
```

Out[27]:
```
0.16151202749140894
```

In [28]:
```
united_tuesday = united.loc[united == 2]
united_tuesday
```

Out[28]:
```
5342177    2
5343938    2
5345166    2
5346638    2
5347202    2
5350125    2
5351787    2
5353440    2
5354683    2
5355504    2
5449240    2
5451360    2
5452632    2
5454067    2
5454599    2
5457072    2
5457862    2
5459106    2
5460814    2
5462126    2
5556627    2
5558750    2
5560032    2
5561463    2
5561998    2
5564408    2
5565262    2
5566510    2
5568218    2
5569554    2
5667061    2
5668311    2
5669486    2
5670875    2
5673777    2
5674504    2
5676511    2
5677156    2
5678804    2
5680552    2
5681895    2
5682652    2
5773866    2
5775163    2
5776284    2
5777656    2
5780555    2
5781274    2
5783384    2
5783920    2
5785414    2
5787212    2
5788641    2
5789421    2
Name: DAY_OF_WEEK, dtype: int64
```

In [29]:
```
united_n = len(united)
united_p_hat = len(united_tuesday) / united_n
united_p_hat
```

Out[29]:
```
0.17419354838709677
```

I test that both American Airlines and United Airlines proportions satisfy the inequality:
n p_hat >= 10 AND n (1-p_hat) >= 10

In [30]:
```
american_n * american_p_hat >= 10 and american_n * (1-american_p_hat) >= 10
```

Out[30]:
```
True
```

In [31]:
```
united_n * united_p_hat >= 10 and united_n * (1-united_p_hat) >= 10
```

Out[31]:
```
True
```

Both proportions satisfy the inequality, meaning the test assumptions have been satisfied!

Hypotheses Statements

H_0: american_p_hat = united_p_hat
H_1: american_p_hat != united_p_hat

Hypothesis Test

I conduct a hypothesis test at significance level alpha = 0.05.

First, I compute the test statistic.

In [32]:
```
p_hat = (len(american_tuesday) + len(united_tuesday)) / (american_n + united_n)
p_hat
```

Out[32]:
```
0.16805324459234608
```

In [33]:
```
z_stat = (american_p_hat - united_p_hat) / (p_hat * (1 - p_hat) * ((1 / american_n) + 1 / united_n))**0.5
z_stat
```

Out[33]:
```
-0.41551837129691793
```

Now, I compute the p-value using the formula:

p-value = P(Z > abs(z_stat)) + P(Z < -abs(z_stat)) = 2 * P(Z < -abs(z_stat))

In [34]:
```
test_3_p_value = 2 * norm.cdf(-abs(z_stat))
test_3_p_value
```

Out[34]:
```
0.6777624692569595
```

Conclusion

The p-value (0.67776...) is greater than the significance level (0.05) so we fail to reject the null hypothesis.

We conclude the data does not provide enough evidence that the probability of flights occurring on Tuesdays in December from ORD to LAX for American Airlines and United Airlines is not equal.

Multiple Comparisons

I now use Bonferroni correction to conduct the 3 hypothesis tests I have done above with a family-wise error rate of alpha = 0.05.

I calculate that the significance level for each hypothesis test is alpha / number of hypothesis tests = 0.05 / 3.

In [35]:
```
sig_lvl = 0.05 / 3
sig_lvl
```

Out[35]:
```
0.016666666666666666
```

In [36]:
```
test_1_p_value
```

Out[36]:
```
0.16371623660393797
```

In [37]:
```
test_1_p_value < sig_lvl
```

Out[37]:
```
False
```

Since the above inequality is false, we fail to reject the null hypothesis. We conclude the data does not provide enough evidence that the average arrival delay on flights in December from ORD to LAX for American Airlines and United Airlines is not equal.

In [38]:
```
test_2_p_value
```

Out[38]:
```
0.12555519084273978
```

In [39]:
```
test_2_p_value < sig_lvl
```

Out[39]:
```
False
```

Since the above inequality is false, we fail to reject the null hypothesis. We conclude the data does not provide enough evidence that the probability of flights delaying over an hour in December from ORD to LAX for American Airlines and United Airlines is not equal.

In [40]:
```
test_3_p_value
```

Out[40]:
```
0.6777624692569595
```

In [41]:
```
test_3_p_value < sig_lvl
```

Out[41]:
```
False
```

Since the above inequality is false, we fail to reject the null hypothesis. We conclude the data does not provide enough evidence that the probability of flights occurring on Tuesdays in December from ORD to LAX for American Airlines and United Airlines is not equal.