Hypothesis Testing (TLC Project)

December 21, 2025

[1]:
```
#The goal is to apply descriptive statistics and hypothesis testing in Python.␣
 ↪The goal for this A/B test is to sample data and analyze whether there is a␣
 ↪relationship between payment type and fare amount. For example: discover if␣
 ↪customers who use credit cards pay higher fare amounts than customers who␣
 ↪use cash.
```

[2]:
```
import pandas as pd
from scipy import stats
```

[3]:
```
taxi_data = pd.read_csv("2017_Yellow_Taxi_Trip_Data.csv", index_col = 0)
```

[5]:
```
taxi_data ['payment_type'].value_counts()
```

[5]:
```
1    15265
2     7267
3      121
4       46
Name: payment_type, dtype: int64
```

[6]:
```
#In the dataset, payment_type is encoded in integers:
#1: Credit card
#2: Cash
#3: No charge
#4: Dispute
```

[7]:
```
# descriptive stats code for EDA
taxi_data.describe(include='all')
```

[7]:
```
            VendorID   tpep_pickup_datetime  tpep_dropoff_datetime  \
count   22699.000000                  22699                  22699
unique           NaN                  22687                  22688
top              NaN  07/03/2017 3:45:19 PM  10/18/2017 8:07:45 PM
freq             NaN                      2                      2
mean        1.556236                    NaN                    NaN
std         0.496838                    NaN                    NaN
min         1.000000                    NaN                    NaN
25%         1.000000                    NaN                    NaN
```

```
50%         2.000000                    NaN                    NaN
75%         2.000000                    NaN                    NaN
max         2.000000                    NaN                    NaN

        passenger_count  trip_distance    RatecodeID store_and_fwd_flag  \
count      22699.000000   22699.000000  22699.000000              22699
unique              NaN            NaN           NaN                  2
top                 NaN            NaN           NaN                  N
freq                NaN            NaN           NaN              22600
mean           1.642319       2.913313      1.043394                NaN
std            1.285231       3.653171      0.708391                NaN
min            0.000000       0.000000      1.000000                NaN
25%            1.000000       0.990000      1.000000                NaN
50%            1.000000       1.610000      1.000000                NaN
75%            2.000000       3.060000      1.000000                NaN
max            6.000000      33.960000     99.000000                NaN

        PULocationID  DOLocationID  payment_type   fare_amount         extra  \
count   22699.000000  22699.000000  22699.000000  22699.000000  22699.000000
unique           NaN           NaN           NaN           NaN           NaN
top              NaN           NaN           NaN           NaN           NaN
freq             NaN           NaN           NaN           NaN           NaN
mean      162.412353    161.527997      1.336887     13.026629      0.333275
std        66.633373     70.139691      0.496211     13.243791      0.463097
min         1.000000      1.000000      1.000000   -120.000000     -1.000000
25%       114.000000    112.000000      1.000000      6.500000      0.000000
50%       162.000000    162.000000      1.000000      9.500000      0.000000
75%       233.000000    233.000000      2.000000     14.500000      0.500000
max       265.000000    265.000000      4.000000    999.990000      4.500000

             mta_tax    tip_amount  tolls_amount  improvement_surcharge  \
count   22699.000000  22699.000000  22699.000000           22699.000000
unique           NaN           NaN           NaN                    NaN
top              NaN           NaN           NaN                    NaN
freq             NaN           NaN           NaN                    NaN
mean        0.497445      1.835781      0.312542               0.299551
std         0.039465      2.800626      1.399212               0.015673
min        -0.500000      0.000000      0.000000              -0.300000
25%         0.500000      0.000000      0.000000               0.300000
50%         0.500000      1.350000      0.000000               0.300000
75%         0.500000      2.450000      0.000000               0.300000
max         0.500000    200.000000     19.100000               0.300000

        total_amount
count   22699.000000
unique           NaN
top              NaN
```

```
freq              NaN
mean        16.310502
std         16.097295
min       -120.300000
25%          8.750000
50%         11.800000
75%         17.800000
max       1200.290000
```

[8]:

```
#We are interested in the relationship between payment type and the fare amount␣
↪the customer pays. One approach is to look at the average fare amount for␣
↪each payment type.
```

[9]:

```
taxi_data.groupby('payment_type')['fare_amount'].mean()
```

[9]:

```
payment_type
1    13.429748
2    12.213546
3    12.186116
4     9.913043
Name: fare_amount, dtype: float64
```

[10]:

```
#Based on the averages shown, it appears that customers who pay in credit card␣
↪tend to pay a larger fare amount than customers who pay in cash. However,␣
↪this difference might arise from random sampling, rather than being a true␣
↪difference in fare amount. To assess whether the difference is statistically␣
↪significant, you conduct a hypothesis test.
```

[11]:

```
#Null hypothesis: There is no difference in average fare between customers who␣
↪use credit cards and customers who use cash.
#Alternative hypothesis: There is a difference in average fare between␣
↪customers who use credit cards and customers who use cash

#The goal in this step is to conduct a two-sample t-test.

#State the null hypothesis and the alternative hypothesis
#Choose a signficance level
#Find the p-value
#Reject or fail to reject the null hypothesis
```

[12]:

```
# 0: There is no difference in the average fare amount between customers who␣
↪use credit cards and customers who use cash.
# : There is a difference in the average fare amount between customers who use␣
↪credit cards and customers who use cash.
#5% as the significance level and proceed with a two-sample t-test.
```

[13]:

```
#hypothesis test, A/B test
#significance level

credit_card = taxi_data[taxi_data['payment_type'] == 1]['fare_amount']
cash = taxi_data[taxi_data['payment_type'] == 2]['fare_amount']
stats.ttest_ind(a=credit_card, b=cash, equal_var=False)
```

[13]:

```
Ttest_indResult(statistic=6.866800855655372, pvalue=6.797387473030518e-12)
```

[14]:

```
#Since the p-value is significantly smaller than the significance level of 5%,␣
 ↪you reject the null hypothesis.
#Notice the 'e-12' at the end of the pvalue result.
#You conclude that there is a statistically significant difference in the␣
 ↪average fare amount between customers who use credit cards and customers who␣
 ↪use cash.
```

[15]:

```
#The key business insight is that encouraging customers to pay with credit␣
 ↪cards can generate more revenue for taxi cab drivers.

#This analysis relies on the assumption that passengers were randomly assigned␣
 ↪to a payment method and consistently followed it. The data was not collected␣
 ↪through a controlled A/B test, so random assignment had to be stimulated.␣
 ↪For example, riders might not carry lots of cash, so it's easier to pay for␣
 ↪longer/farther trips with a credit card. In other words, it's far more␣
 ↪likely that fare amount determines payment type.
```

[ ]:

```
```