# Tips Dataset Summary based on Hypothesis Testing

We are loading a built-in data set available in the Seaborn Python package. The given dataset doesn’t have any known outcome, so we are concluding that it is an unsupervised dataset. We have 7 columns.

Total\_bill & tip - continuous/regression data.

Sex, Day, Smoker, Time, Size - category/classification data.

From correlation test outcomes are:

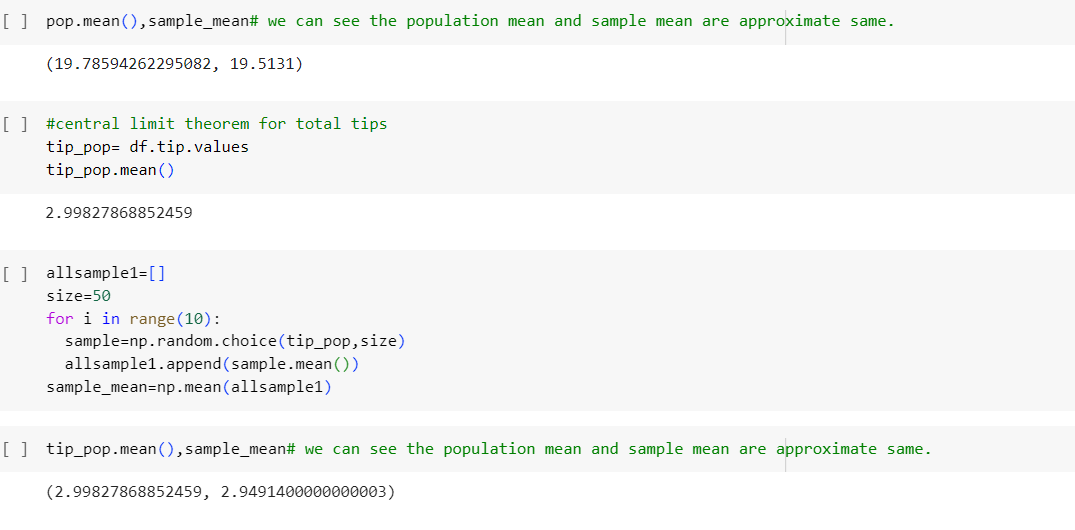
1. Total\_bill & tip are directly proportional. (There is a chance if bill value is high, Tip can be high also).
2. Total\_bill & size are directly proportional. (There is a chance if size value high, bill value can be high).
3. Tip & Size are directly proportional. (There is a chance if size value high, bill value can be high).

A screenshot of a computer

Description automatically generated

Outcomes of a central limit theorem.

The population mean, and Sample mean are approximately the same for the respective column.



One sample t-test hypothesis

One sample t-test for **total bills** says the data is distributed properly in the column.

One sample t-test for **Tips** says the data is distributed properly in the column.

Two sample T-tests are done for the Total\_bill and Tip column. (continuous columns), We can conclude that Total\_bill and Tip columns are independent from each other.

A screenshot of a computer program

Description automatically generated

From Two sample T-tests using central limit theorem for Total bill and Tip. We can say that both columns are independent of each other.

A screenshot of a computer

Description automatically generated

Chi-square tests are done for the Sex, Smoker, Size, Time, and Day columns. (Category columns). Outcomes are:

* Sex, smoker column doesn’t have any relationship.
* Sex, day columns have a relationship between them.
* Sex and time columns have a relationship between them.
* Sex and size columns don’t have a relationship between them.
* Smoker and day columns have a relationship between them.
* Smoker and time columns don’t have a relationship between them.
* Smoker and Size columns don’t have a relationship between them.
* Day and time columns have a relationship between them.
* Day and size columns have a relationship between them.
* Time and size Columns have a relationship between them.

Annova Tests are done for the Total\_bill, Tip, Sex, Smoker, Size, Time, and Day columns. (Category Columns vs continuous column).

* Total bill and smoker columns have a relationship between them.
* Total bill and sex columns don’t have a relationship between them.
* Total bill and Day columns don’t have a relationship between them.
* Total bill and time columns don’t have a relationship between them.
* Total bill and size columns don’t have a relationship between them.
* Tip and smoker columns have a relationship between them.
* Tip and sex columns have a relationship between them.
* Tip and day columns have a relationship between them.
* Tip and time columns have a relationship between them.
* Tip and size columns don’t have a relationship between them.

From graphs you can get better view.