KEY_Practice17_Basic_Stats_III_Correlations

May 25, 2020

1 Practice 23: Correlations

Remember: * A **correlation** is a measure of the statistical relationship between two variables * Correlations can be **positive** or **negative**, and **strong** or **weak** * The output of the `corrcoef` function is a **correlation matrix**

First, import numpy and pandas:

[1]:
```
# load numpy and pandas and scipy.stats

import numpy as np
import pandas as pd
```

[2]:
```
#read in tips data
path = 'https://raw.githubusercontent.com/GWC-DCMB/ClubCurriculum/master/'
tips = pd.read_csv(path + 'SampleData/tips.csv')

tips.head()
```

[2]:

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

We want to calculate the correlations of `total_bill`, `tip` and `size`. Since we are getting the correlations for three variables, what size do we expect the resulting **correlation matrix** to be? How will the correlations be organized?

ANSWER: 3 x 3 matrix

_	total_bill	tip	size
total_bill	_	_	_
tip	_	_	_
size	_	_	_

[18]:
```
# create correlation matrix for total_bill, tip and size
# HINT: what parameter do we need to use when our observations are along the␣
 ↪rows?
corrs = np.corrcoef(tips[['total_bill', 'tip','size']], rowvar=False)
print(corrs)
```

```
[[1.         0.67573411 0.59831513]
 [0.67573411 1.         0.48929878]
 [0.59831513 0.48929878 1.        ]]
```

Which two variables have the strongest correlation?

ANSWER: `total_bill` and `tip`

Which two variables have the weakest correlation?

ANSWER: `tip` and `size`

CHALLENGE: What if these relationships are different between lunch and dinner? Create two subsets of tips for the lunch and dinner times and repeat the correlation analysis.

[19]:
```
# create two subsets of tips, one for lunch and one for dinner
lunch = tips.query('time == "Lunch"')
dinner = tips.query('time == "Dinner"')

# compute the correlation matrix for each of the data subsets you created
lunch_corrs = np.corrcoef(lunch[['total_bill', 'tip','size']], rowvar=False)
din_corrs = np.corrcoef(dinner[['total_bill', 'tip','size']], rowvar=False)

#print the correlations
print(lunch_corrs)
print(din_corrs)
```

```
[[1.         0.80542384 0.708662  ]
 [0.80542384 1.         0.64785392]
 [0.708662   0.64785392 1.        ]]
[[1.         0.63287125 0.55701503]
 [0.63287125 1.         0.42850163]
 [0.55701503 0.42850163 1.        ]]
```

What do you notice when you compare the results between the different times of day?

Answer: correlations are much stronger during the lunch shift than the dinner shift

Do you have a *hypothesis* for why this might be? *HINT*: Does it have to do with a difference in the amount people generally spend for those meals? Or maybe the number of *samples* we have for each condition (i.e. meal time)?

[20]:
```
# find the average meal price for each meal time
print(np.mean(lunch['total_bill']))
print(np.mean(dinner['total_bill']))
```

```
# find the number of samples we have for each meal time
# HINT: use len()
print(len(lunch))
print(len(dinner))
```

```
17.16867647058823
20.7971590909091
68
176
```

What do you notice here? When we look at the *sample size*, we find that we have a lot fewer *samples* from the lunch shift compared to the dinner shift. When our sample size is smaller, this means that each individual sample contributes more to our final statistic, here correlation. Also, we see that the average meal price is lower for lunch as well. Both of these differences are things we must consider when comparing the correlations of these two sets.

Nice job! You just practiced:

- Using statistical tests to determine if two groups are significantly different
- Using correlations to determine the relationships between variables