KEY_Practice16_Intro_Stats_III

July 11, 2019

1 Practice with Statistics (Part 3)!

Remember: * Significance tests tell you the probability that a change happens purely by chance * A t-test is a significance test which compares *two groups*

First, import numpy and pandas and the scipy statistical module:

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

import numpy as np
import pandas as pd
import scipy.stats as stats
```

[ ]:
```
# mount Google Drive
from google.colab import drive
drive.mount('/content/gdrive')
path = '/content/gdrive/My Drive/SummerExperience-master/'
```

Load in the sample data from the Lesson:

[ ]:
```
# read the csv file: '../Lessons/SampleData/detroit_weather_2.csv'
data_table = pd.read_csv(path + 'Lessons/SampleData/detroit_weather_2.csv')
```

[ ]:
```
# Print the head of the table to remind you of the format:

data_table.head()
```

[ ]:

	YEAR	MONTH	DAY	Temperature
0	1937	1	1	0.50
1	1937	1	2	0.17
2	1937	1	3	-1.06
3	1937	1	4	-3.89
4	1937	1	5	-0.17

[ ]:
```
# Pick two decades that we didn't look at during the lesson
# extract the temperatures into numpy arrays

temps_1960 = np.array(data_table.query('YEAR >= 1960 and YEAR <␣
↪1970')["Temperature"] )
temps_today = np.array(data_table.query('YEAR >= 2010 and YEAR <␣
↪2020')["Temperature"] )
# note that the current decade isn't over...
```

```
# even though there isn't yet data for 2020, we include it in the query so our␣
 ↪code will still work in the year 2021!

print(temps_1960, temps_today)
```

```
[-2.33  0.11  1.28 ... -2.11 -0.67 -2.06] [ -3.78 -10.17 -12.11 ...   5.67
6.44  11.22]
```

[ ]:
```
# Calculate the means of your data to see if they differ

print(np.mean(temps_1960))
print(np.mean(temps_today))
```

```
8.915261428962499
9.666400235086689
0.7511388061241906
```

[ ]:
```
# Calculate the difference between the two means

print(np.mean(temps_today) - np.mean(temps_1960)) # .75 degree increase
```

[ ]:
```
# Perform a t-test of these data to calculate the p value

stats.ttest_ind(temps_1960, temps_today).pvalue
```

[ ]:
```
0.002958034438690926
```

Now, we will try using a different type of test, called an ANOVA (ANalysis Of VAriance) test. An ANOVA is similar to a t-test, but allows you to compare the variance of *multiple groups* in a single statistical test.

[ ]:
```
# Create an array where each element is the temperatures from a different␣
 ↪decade.

temps = [
    np.array( data_table.query('YEAR < 1940')["Temperature"] ), # 1930's
    np.array( data_table.query('YEAR < 1950 and YEAR >= 1940')["Temperature"]␣
 ↪), # 1940's
    np.array( data_table.query('YEAR < 1960 and YEAR >= 1950')["Temperature"]␣
 ↪), # 1950's
    np.array( data_table.query('YEAR < 1970 and YEAR >= 1960')["Temperature"]␣
 ↪), # 1960's
    np.array( data_table.query('YEAR < 1980 and YEAR >= 1970')["Temperature"]␣
 ↪), # 1970's
    np.array( data_table.query('YEAR < 1990 and YEAR >= 1980')["Temperature"]␣
 ↪), # 1980's
    np.array( data_table.query('YEAR < 2000 and YEAR >= 1990')["Temperature"]␣
 ↪), # 1990's
```

```
    np.array( data_table.query('YEAR < 2010 and YEAR >= 2000')["Temperature"]␣
 ↪), # 2000's
    np.array( data_table.query('YEAR >= 2010')["Temperature"] )  # 2010's
]
```

[ ]:
```
[array([  0.5 ,   0.17,  -1.06, ...,  -4.39,  -4.72, -10.33]),
 array([-8.33, -6.94, -6.67, ...,  1.61, -1.83,  0.61]),
 array([ 3.33,  5.89,  9.72, ..., -1.67, -3.39, -3.44]),
 array([-2.33,  0.11,  1.28, ..., -2.11, -0.67, -2.06]),
 array([-5.17, -4.56, -6.28, ...,  3.  ,  1.94,  0.17]),
 array([-0.56, -0.17, -1.83, ..., -1.83, -4.44,  1.39]),
 array([-1.17, -1.33,  2.22, ..., -3.5 ,  2.33, -0.56]),
 array([ 2.22,  8.44,  5.11, ..., -5.33, -5.44,  0.28]),
 array([ -3.78, -10.17, -12.11, ...,   5.67,   6.44,  11.22])]
```

[ ]:
```
# print the mean for each item in the list

list(map(np.mean, temps))
```

[ ]:
```
[9.681990867579907,
 9.483531344100738,
 9.631117196056953,
 8.915261428962499,
 10.01000822143053,
 10.385710375034218,
 10.658524096385543,
 10.51869112814896,
 9.666400235086689]
```

[ ]:
```
# Feed the items in your list into stats.f_oneway to perform an ANOVA test

stats.f_oneway(*temps).pvalue
```

[ ]:
```
2.93434096269204e-14
```

Since we are including a lot more data, this p value is much lower! We can conclude from this that the average temperature is changing over time more than what you would expect if it was truely random!

Nice job! You just practiced:

```
- Using statistical tests to determine if two groups are significantly different
- Using t-tests and ANOVA tests from the scientific python package
```