IP[y]: Notebook
(/tree/Udacity%20Data%20Analyst%20Nano%20Degree/1_Intro%20Data%20Analysis/3%20Data%20Analysis)

3_ProbSet 3.3 Mann-Whitney U

File Edit View Insert Cell Kernel Help

Code ▾ Cell Toolbar: None ▾

Below code used to calculate Mann-Whitney U Statistic for Udacity Data Analyst course for NanoDegree

I ran the code below locally with the following versions

- Windows 7 Professional on 64bit OS
- ipython-notebooks Version 2.2.0
- Anaconda Python Version: 2.7.8 Python Build: 0
- scipy version: 0.14.0
- pandas version: 0.14.1
- numpy version: 1.9.0

My local code outputs: (U, p) = (1924409167.0, 0.019309634413792565)

While Udacity's console outputs: (U, p) = (1924409167.0, 0.024999912793489721)

The p-values differ slightly.

The course TA's believe that this is due to logged here https://github.com/scipy/scipy/issues/4386 (https://github.com/scipy/scipy/issues/4386) and asked me to share my code.

In [15]:

```
## Practice for Prob 3.3

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

#def mann_whitney_plus_means(turnstile_weather):
'''
    This function will consume the turnstile_weather dataframe containing
    our final turnstile weather data.

    You will want to take the means and run the Mann Whitney U-test on the
    ENTRIESn_hourly column in the turnstile_weather dataframe.

    This function should return:
        1) the mean of entries with rain
        2) the mean of entries without rain
        3) the Mann-Whitney U-statistic and p-value comparing the number of entries
           with rain and the number of entries without rain

    You should feel free to use scipy's Mann-Whitney implementation, and you
    might also find it useful to use numpy's mean function.

    Here are the functions' documentation:
    http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
    http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

    You can look at the final turnstile weather data at the link below:
    https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv
'''

    ### YOUR CODE HERE ###

#open .csv file store in df
with open('turnstile_data_master_with_weather.csv', 'rb') as f:
    df = pd.read_csv(f)

with_rain_mean= np.mean(df[df.rain==1]["ENTRIESn_hourly"])
print ("with_rain_mean= " + str(with_rain_mean))

without_rain_mean= np.mean(df[df.rain==0]["ENTRIESn_hourly"])
```

```
without_rain_mean= np.mean(df[df.rain==0]["ENTRIESn_hourly"])
print ("without_rain_mean= " + str(without_rain_mean))

U,p = scipy.stats.mannwhitneyu(df[df.rain==1]["ENTRIESn_hourly"], df[df.rain==0]["ENTRIESn_hourly"])
# from lesson>>  U,p = scipy.stats.mannwhitneyu(sample1, sample2)
print "(U, p) = "
U,p

# return with_rain_mean, without_rain_mean, U, p # leave this line for the grader
```

```
with_rain_mean= 1105.44637675
without_rain_mean= 1090.27878015
(U, p) = 
```

```
Out[15]: (1924409167.0, 0.019309634413792565)
```