Python Note-Book: Day of Week Based Exercise Trends
A/B Testing of Significant Biased Variation of Time-Based Walking Activity

Tristen Bristow
04/08/2017

Abstract

This study is an application of A/B testing, where significant variation of walking activity is tested between weekdays and weekends in the Qualified Self Movement data from the FitBit device. Statistical testing involves the use of Student's t-Test. The direction of this difference in trends is tested by upper one-sided Student's t-Test. This study seeks to address the general question: do people walk a different number of steps on the weekend verses the weekday, on average, and if so, on which do they walk more.

```
import pandas as pd
import numpy as np
import datetime
import matplotlib.pyplot as plt
import math
import scipy.stats
import matplotlib.mlab as mlab
```

```
# Loading and preprocessing the data

# Activity monitoring data [52K] Data downloaded from:
# https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip

df=pd.read_csv('activity.csv')
fmt = '%Y-%m-%d'
df['date']=pd.to_datetime(df['date'],format=fmt)
np.random.seed(12345678)
```

Exploratory Investigation

The mean total number of steps taken for a daily exercise period of 27 minutes is calculated. This proceeds by aggregating all days by the date-labels, forming a table of unique dates. The mean value of steps taken per interval for that particular day is thus calculated.

```
# imputing missing values
print
print("total of na for each column:")
print(df.isnull().sum())
print("fract of na for each column:")
print(df.isnull().mean())
print


df=df.dropna()


mean_steps_day = pd.pivot_table(df, index=['date'], values=['steps'], aggfunc=np.mean)
print(mean_steps_day.describe())
mean_steps_day=mean_steps_day.dropna()
print(mean_steps_day)
x = mean_steps_day['steps'].mean()
plt.axvline(x,linewidth=3, color='k')
plt.text(x+0.2, .0355, r'$\mu$', size=20, rotation=90)


plt.hist(mean_steps_day['steps'], 10 , normed=True)

x = np.linspace(min(mean_steps_day['steps']), max(mean_steps_day['steps']), 100)
mean = np.mean(mean_steps_day['steps'])
variance = np.var(mean_steps_day['steps'])
sigma = np.sqrt(variance)
plt.plot(x, mlab.normpdf(x, mean, sigma), linewidth=3)

plt.xlabel('# Steps/day')
plt.ylabel('Frequency')
plt.title('Average Steps day Distribution')

plt.show()
```

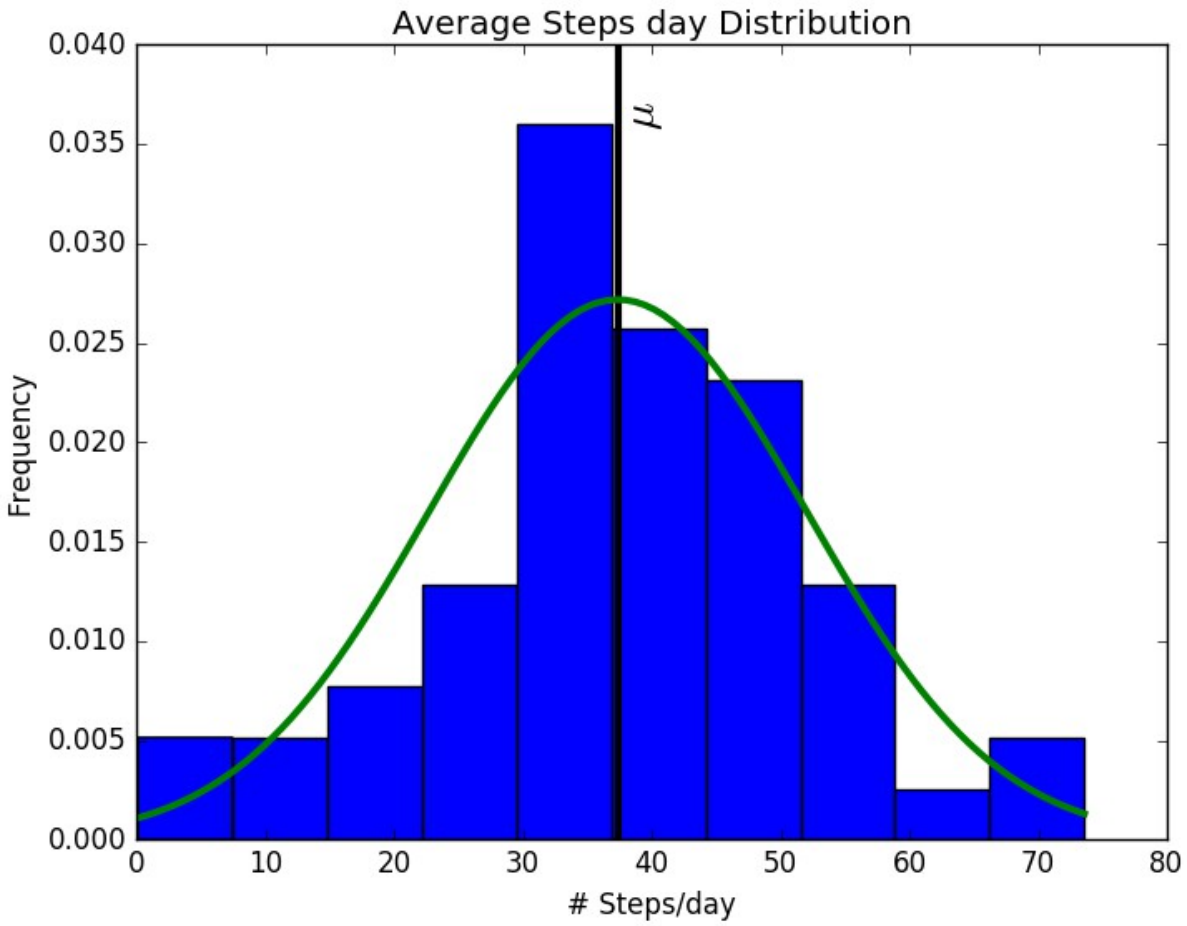

fig 1: Histogram of total number of steps over a day.

Summary Statistics

	steps
count	53.000000
mean	10766.188679
std	4269.180493
min	41.000000
25%	NaN
50%	NaN
75%	NaN
max	21194.000000

The mean value of the number of steps is centered in this plotted distribution, which indicates (by the Gaussian trend), that the data is sufficiently normal in its distribution for statistical methods chosen here (see *fig 1*). There are 320 total 5-second periods over which to aggregate the data for per-interval mean

values. The original recorded values are established in seconds. Though the originating data description claims per-minute values for the intervals in the raw data, the total-daily recorded periods implied are not possible with such a definition (in context of the data, this implies a calendar-labeled day contains 196.25 hours). It is judged that a mistake was made in the original documenting. The total period is sensible if only interpreted as being recorded in per-second sample intervals and not the per-minute intervals indicated by the data-description. Additionally, the step amounts do not match what common-sense claims is possible for human walking or running activity, indicating that the FitBit device output-units have been mis-reported in the documentation or the device is malfunctioning (see readme.md). The following processing steps aggregates the data into per-minute format.

```
# Average daily activity pattern. Bin over 5 sec intervals
# 288 total 5 sec periods over which to aggregate.

mean_steps_act = pd.pivot_table(df, index=['interval'], values=['steps'], aggfunc=np.mean)

print
plt.xlabel('Time (seconds)')
plt.ylabel('Number of Steps')
plt.plot(range(1, 5*288, 5), mean_steps_act)
plt.title('Average Number of Steps per Interval Over a Day-Period')
avy = mean_steps_act['steps'].mean()
print(avy )
plt.axhline(y=avy , hold=None)
plt.text(1, avy , r'$\mu$', size=20, rotation=90)
plt.show()
```

```
Total of na for each column:
steps       2304
date           0
interval       0
dtype: int64
fract. of na for each column:
steps       0.131148
date        0.000000
interval    0.000000
```

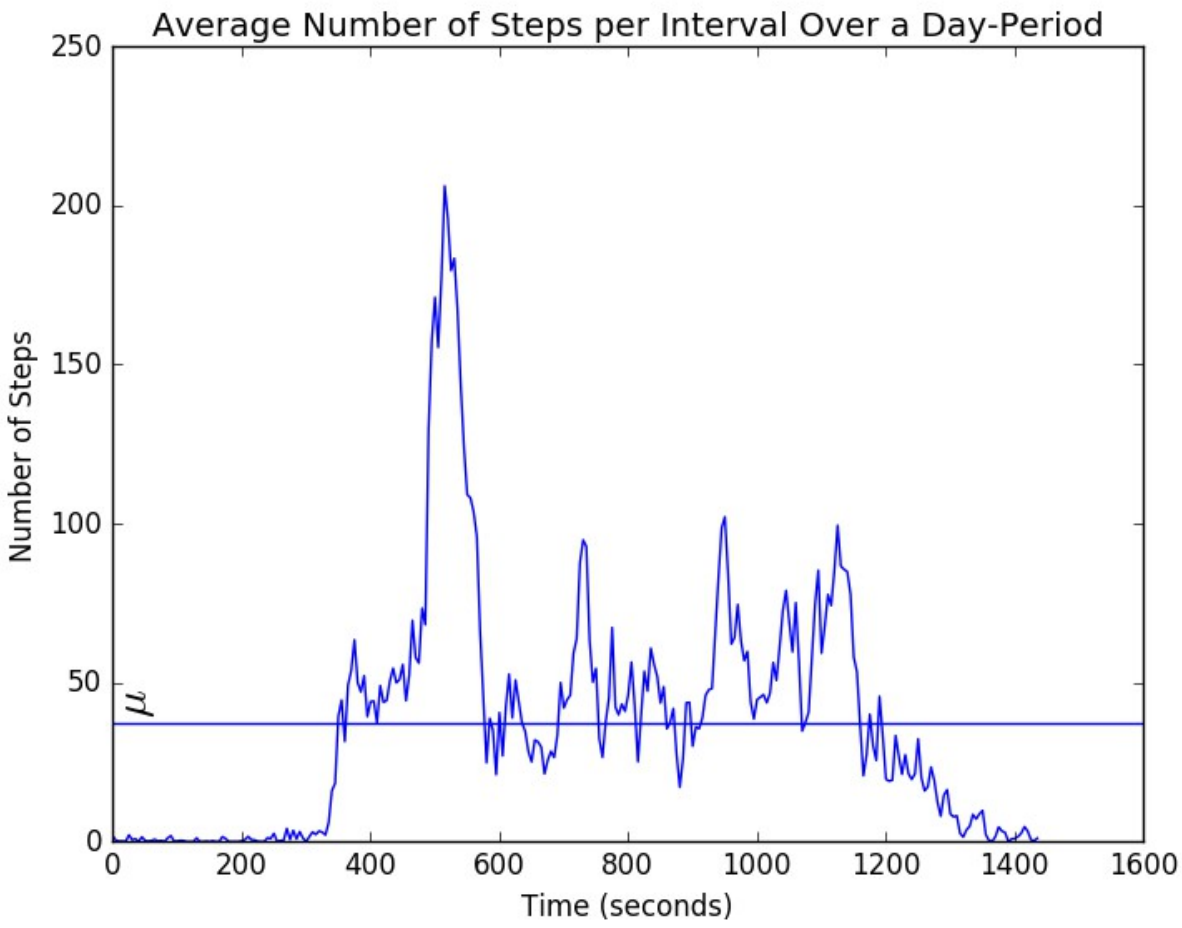

fig 2: Average timed-series, steps taken over a day-period.

13% of the values in the step column are missing (labeled NaN), thus it is apparent that imputing these values would be a ill-advised, as we do not have any additional information about the data that would justify such an approach. It is certain, based on this large percentage of missing values, that all statistical results will be affected by such a decision.

Exploratory analysis concerns if any there are any potential differences in activity patterns between weekdays and weekends. To check for this potential relationship, a new class label for the data is created: weekend verses non-weekend.

```
wknd = df['date'].dt.dayofweek

wknd[wknd <= 4] = 0
wknd[wknd > 4] = 1
```

```
df['wknd'] = wknd

classA = df[df['wknd']==0]
classB = df[df['wknd']==1]
```

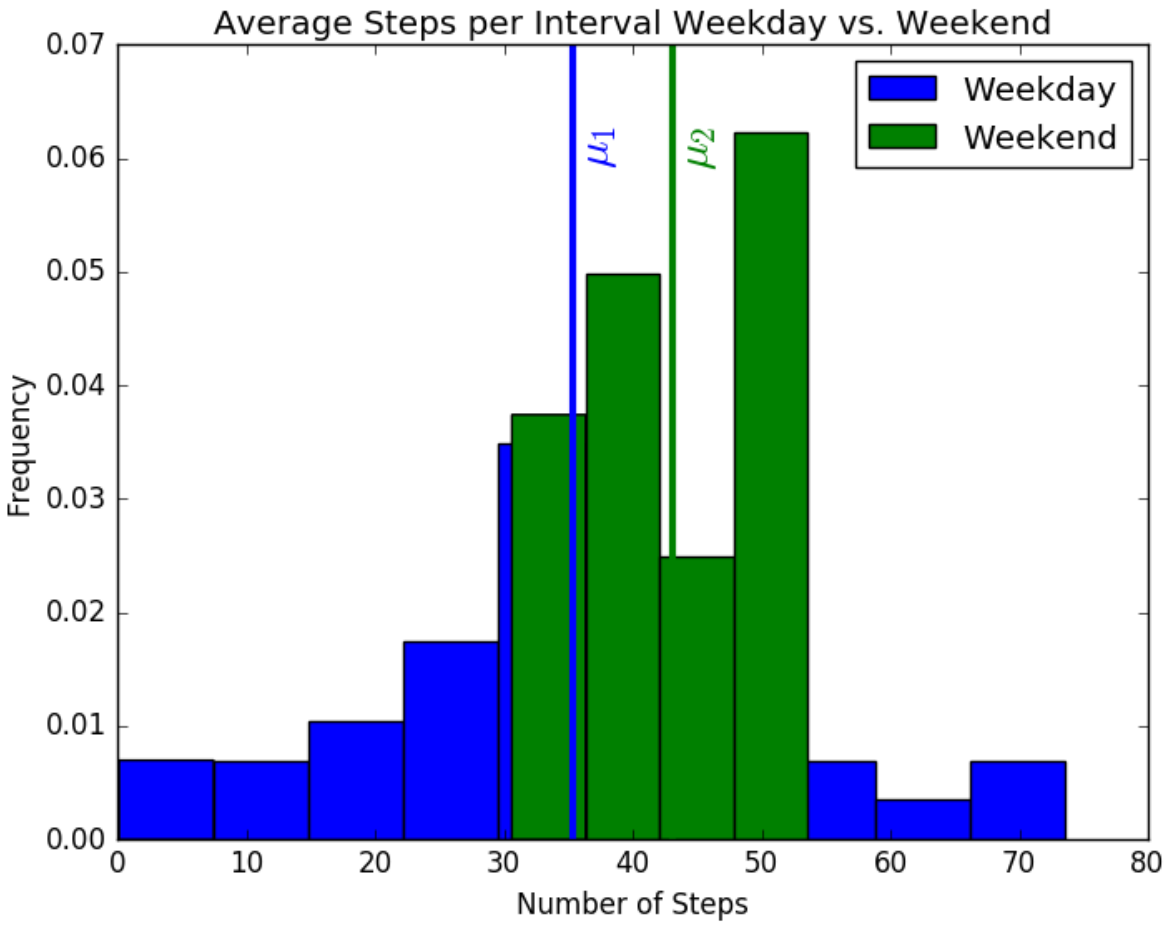

fig 3: Histogram of weekday and weekend step amount

Test I

This initial test concerns whether there is a significant difference between weekend and weekday activity. A Two-sample t-test is implemented. Should the results of the test indicate a rejection of the null-hypothesis (that no significant variation between the two exists), the study then proceeds to post-hoc testing.

H_0: X_wkdy == X_wknd
H_1: X_wkdy < > X_wknd

```
datestepA = pd.pivot_table(classA, index=['date'], values=['steps'], aggfunc=np.mean)
datestepB = pd.pivot_table(classB, index=['date'], values=['steps'], aggfunc=np.mean)
```

The test result from the 2-sample-t-test is p = 0.0002, where the statistical significance threshold for rejecting the null is set to $\alpha_1 = 0.05$. The test indicates $p < \alpha_1$, thus it is supported that a significant difference in means exists. H_1 is indicated over the null hypothesis, H_0.

Test II

Exploratory investigation suggests that walking activity is greater on a weekend than weekday, on average. This poses a new hypothesis that states an inequality of the two observed means. In the terms of A/B testing, we assign the first subject to weekday average number of steps, and the second to weekend average number of steps. We are interested in testing a hypothesis that the average associated with the weekend is greater than that associated with the weekday, verses the null hypothesis, that no discernible positive difference exists between mean values.

$$H_1 : X\_wkdy == X\_wknd$$
$$H_2 : X\_wknd > X\_wkdy$$

The new test is a Two-Sample upper one-sided t-test, so the test statistic is the same, except means are reversed to reflect this as an upper-one-sided test. The threshold for acceptance is $\alpha_2 = 0.025$. If $p < \alpha_2$, we reject the null in favor of alternate hypothesis.

```
xx = datestepA['steps'].mean()
yy = datestepB['steps'].mean()
plt.axvline(xx,linewidth=3, color='blue')
plt.axvline(yy,linewidth=3, color='green')
glab1 = r'$\mu_1$'
glab2 = r'$\mu_2$'
plt.text(xx+0.2, .06, glab1, size=20, rotation=90, color='blue')
plt.text(yy+0.2, .06, glab2, size=20, rotation=90, color='green')

plt.hist(datestepA['steps'], 10, normed=True, label='Weekday')
plt.hist(datestepB['steps'], 4, normed=True, label='Weekend')
plt.xlabel("Number of Steps")
plt.ylabel("Frequency")
plt.title('Average Steps per Interval Weekday vs. Weekend')
plt.legend(loc = 'upper right')
plt.show()
```

```
twosample_results = scipy.stats.ttest_ind(classA['steps'], classB['steps'],equal_var=False,axis=0)

# Test for difference in mean steps/day between weekday and weekend
print("****Is p < 0.05 ?")
print(twosample_results)
print("****")

twosample_results2 = scipy.stats.ttest_ind(classB['steps'],classA['steps'],equal_var=False,axis=0)

print("****Is p_2 < 0.025 ?")
print(twosample_results2)
print("****")
```

Output:
Ttest_indResult(statistic=3.7061655054800084, pvalue=0.00021206612638864717)

Results

$p_2 < \alpha_2$, ($p_2 = 0.0002$), and the test statistic output is positive for this upper one-sided t-test. Thus, the test result indicates that we reject the null hypothesis for the alternate hypothesis: that the number of steps taken on a weekend will be greater than the number of steps taken on a weekday, on average.

Conclusion

From the available data, it would appear that walking activity is significantly pronounced on the weekend over the weekday. It is expected any particular patterns beyond the general features discussed should be discarded. It is guessed that the device is not functioning properly enough to provide a more than rough estimate of what constitutes a human step. The general activity-pattern discussed here is, non-the-less, expected to be valid in the basic weekday verses weekend relationship. The sample size is large enough to at least compensate for variability owing to such device malfunction. The nature of the exercise trends is expected to vary in ways not encompassed in the source data. It is not believed that the specific magnitude of difference, or particularities of the time series are to be statistically meaningful in any context. Any future work that would attempt to discover such trends should specify a variety of factors (based on general physiological or biophysical profiles relevant to walking), to predict the magnitude of this difference or any particular behavioral trends.