CS 133 - Introduction to Computational and Data Science

Instructor: Renzhi Cao
Computer Science Department
Pacific Lutheran University
Spring 2017

Updates of Project 1

- Due date: 3/25, 5 pm, Saturday.
- Careful about return statement in the function
- Understand the variable type (for i in Seq)
- Write main function
- Use single-letter amino acid for the last questions of Project 1:
  http://130.88.97.239/bioactivity/aacodefrm.html

Any other Questions?

Exercises on Numpy

-Solutions available on website

- Explore Numpy document with your partner.
- Read the data.txt and load the first column as list 1, and the second column as list2
- Use Numpy to calculate the mean, min, max of all data for each list. Write function to do that.
- Use Numpy to do a vector add, subtract, and multiply of this two lists.

Extra practices:
http://codingbat.com/python

Statistics

- In the previous class, we learned linear algebra.
- Today we are going to learn basic statistics

Statistics

- Reading (Data Science from Scratch):
  - Read chapter 6: Probability
  - A quiz will be given based on chapter 6

Statistics

This is a GIGANTIC topic.

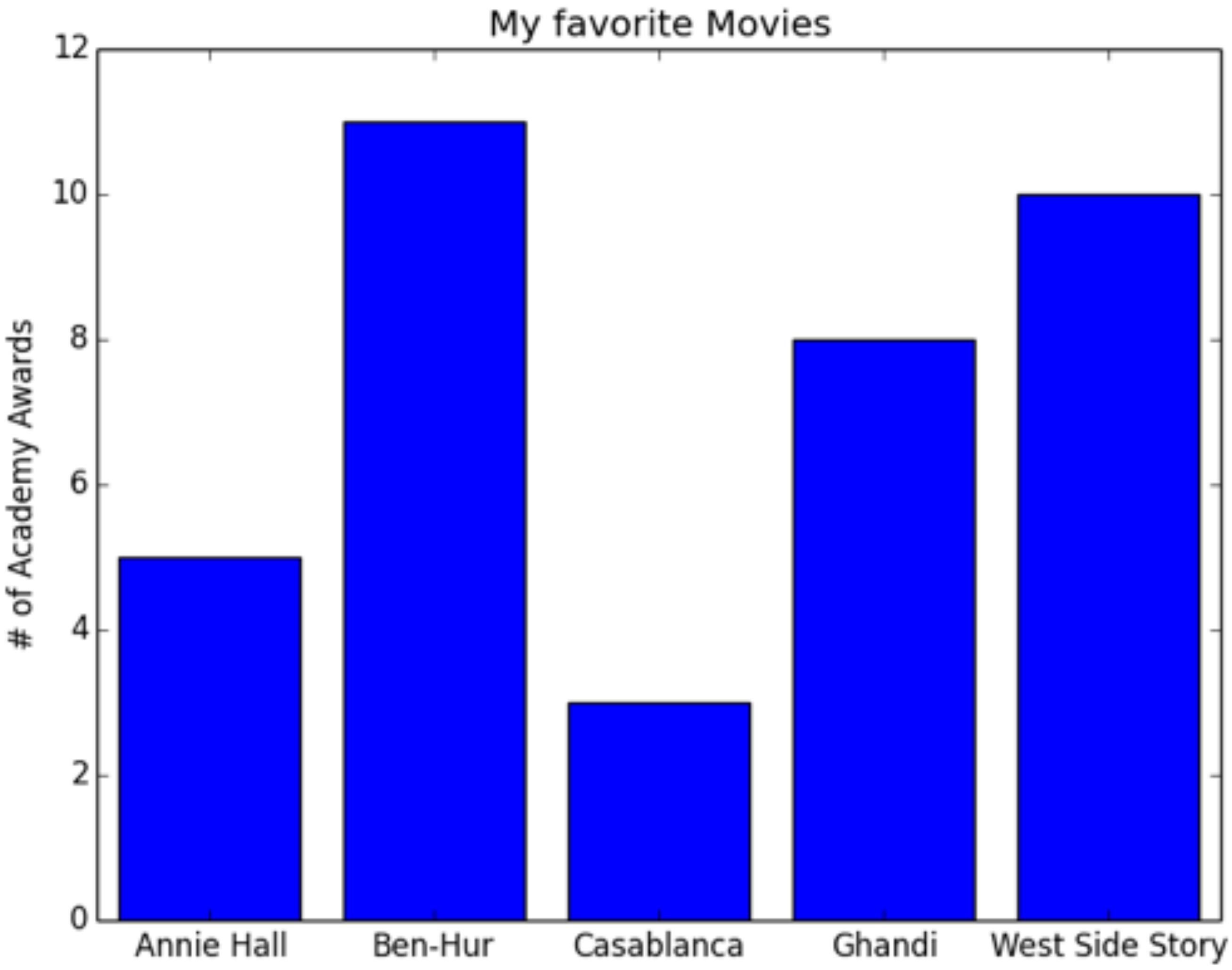
In this class, we will just cover the surface.

Covering basic concepts that we will use in the future.

“Facts are stubborn, but statistics are more pliable”

Mark Twain

What do you want to know from this picture?



Measurements of central tendency and Dispersion

Mean

Median

Mode

Min and Max value

Percentiles

Range

Variance

Standard Deviation

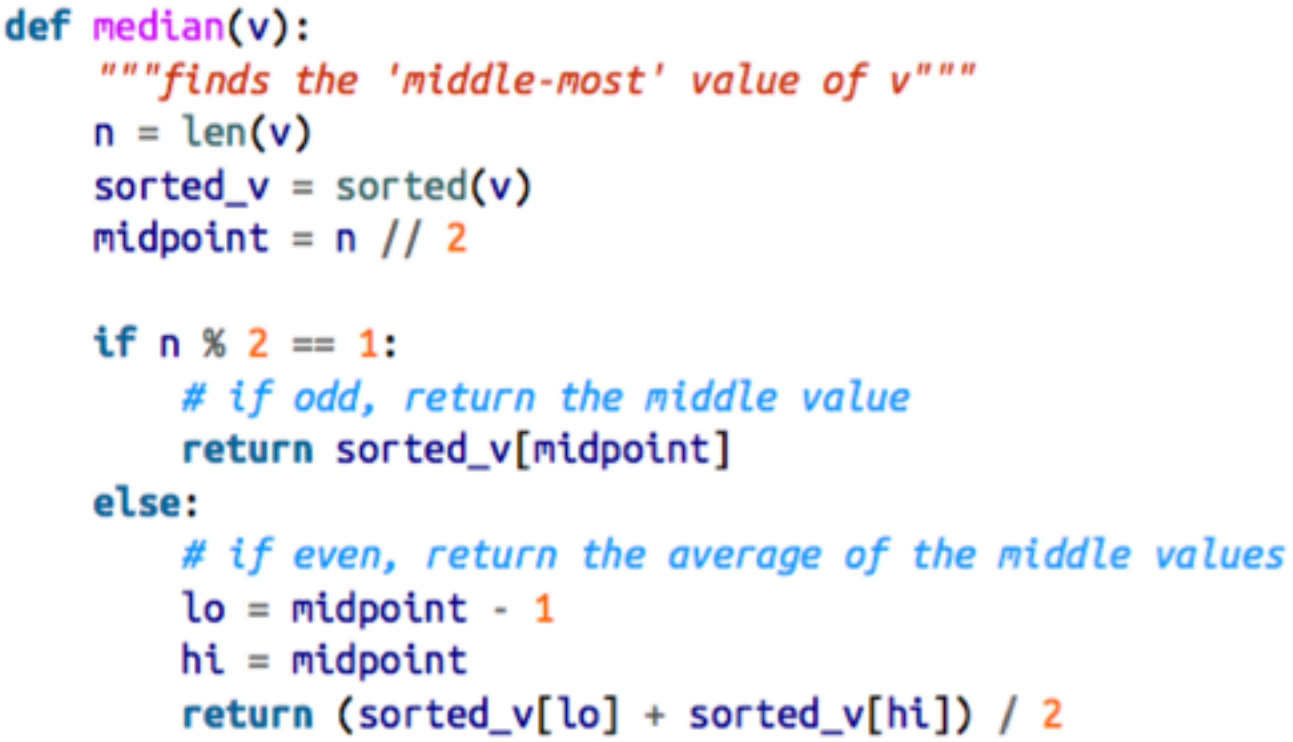
- movies = [3,5,2,4,7]
- How to write a function to calculate the mean?

```
def mymean(v):
    total = 0
    for i in range(len(v)):
        total = total + v[i]
    result = total/len(v)
    return result

mymean(movies)
```

Simple way, use numpy, numpy.mean(movies)

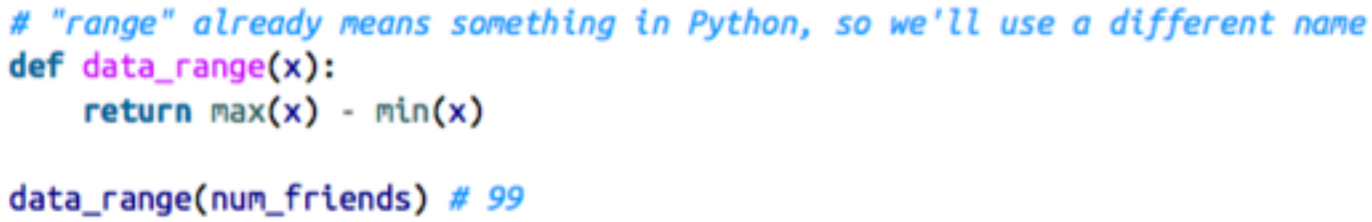
- movies = [3,5,2,4,7]
- How about median?



```python
def median(v):
    """finds the 'middle-most' value of v"""
    n = len(v)
    sorted_v = sorted(v)
    midpoint = n // 2

    if n % 2 == 1:
        # if odd, return the middle value
        return sorted_v[midpoint]
    else:
        # if even, return the average of the middle values
        lo = midpoint - 1
        hi = midpoint
        return (sorted_v[lo] + sorted_v[hi]) / 2
```

- movies = [3,5,2,4,7]
- How about data range?



```
# "range" already means something in Python, so we'll use a different name
def data_range(x):
    return max(x) - min(x)

data_range(num_friends) # 99
```

Example

Notes:

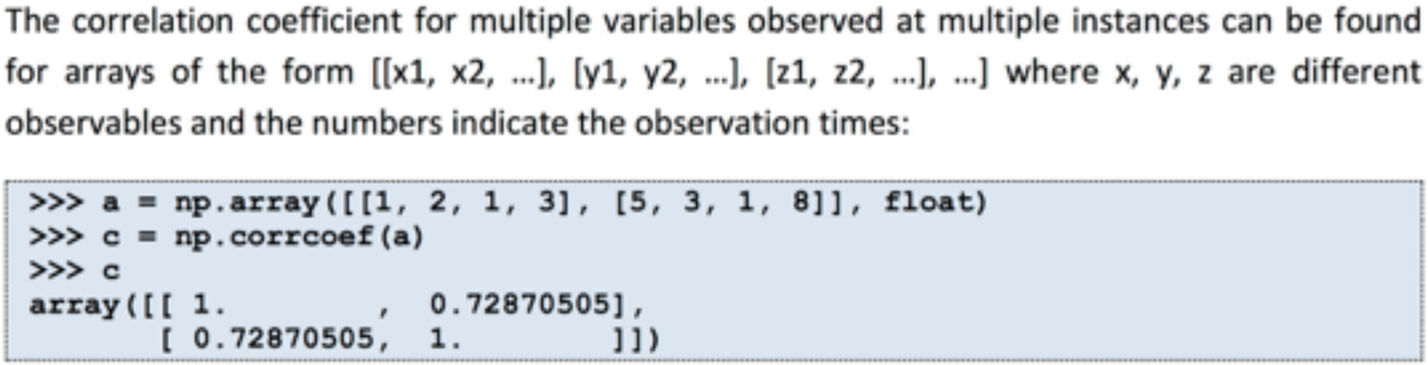
Statistics is a module only supported in 3.0

In 2.7 we can use numpy to calculate those values or create our own functions.

```
a = np.array([[0,2], [3, -1], [3, 5]], float)
a.mean(axis=0)      # will get [2,2]
a.mean(axis=1)      # will get [1,1,4]
l = [1,3,6,7,8]
np.median(l)        # get the median of l
```

Correlation

- In numpy, use numpy.corrcoef.



The correlation coefficient for multiple variables observed at multiple instances can be found for arrays of the form [[x1, x2, …], [y1, y2, …], [z1, z2, …], …] where x, y, z are different observables and the numbers indicate the observation times:

```
>>> a = np.array([[1, 2, 1, 3], [5, 3, 1, 8]], float)
>>> c = np.corrcoef(a)
>>> c
array([[ 1.        ,  0.72870505],
       [ 0.72870505,  1.        ]])
```

- A correlation of -1 means perfect anti-correlation
- A correlation of 1 means perfect positive correlation

- Indicates a relationship between the two parameters, lists, etc.
- Correlation does not imply causation!!!! Will be saying this many many times

Simpson's paradox

Confounding variables?

Number of friends for a group of scientists

Coast	# of Members	Avg.# of friends
West	101	8.2
East	103	6.5

Simpson's paradox

Coast	Degree	# of members	Avg. # of friends
West	Ph.D.	35	3.1
East	Ph.D.	70	3.2
West	No Ph.D.	66	10.9
East	No Ph.D.	33	13.4

Assumes all other values are equal
Always look at all confounding values!

Exercises on statistics

- Learning target: Numpy, and statistics based on Numpy
- Check the Sakai, and turn in your code on Sakai before 3/23/2017, 9:50am.
- Quiz about Probability this Thursday (Please read the book)