CS 234 Data, Analytics, and Visualization (Fall 2017)

09/05/2017 Class Worksheet Nr. 2

Topic: Python Review, Descriptive Statistics

In Python, we'll use the library `pandas` to calculate descriptive statistics. Below is an example:

```
import pandas as pd
numbers = [10, 14, 11, 11, 9, 8, 7, 10, 11]
numS = pd.Series(numbers)
print numS.describe()
```

This code will print the following result:

```
count     9.000000      # the size of the sample, N
mean     10.111111      # the average value
std       2.027588      # the standard deviation of the sample
min       7.000000      # minimum value
25%       9.000000      # 25% percentile
50%      10.000000      # median
75%      11.000000      # 75% percentile
max      14.000000      # maximum
```

Your Task

Using only the Python functions: **`len, sorted, sum`**, and **`math.sqrt`** write succinct code to generate the descriptive statistics shown above. **BONUS:** write efficient code to calculate the mode (no restriction in functions that you can use). The mode is the value that occurs the most in a series. In our example, it's 11.

Statistics Reminders:

The mean, median, the mode are known as measures of the **central tendency**. The standard deviation (or the variance), the min, max, and the quartiles are measures of **variability** (or dispersion).

Formulas

Assuming we are calculating the statistics of a population (all items), as opposed to a sample (only some of the items from a population), here are some useful formulas:

The mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The standard deviation (std for a sample has a **N-1 in the enumerator**)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

The percentiles are found by sorting a list of numbers, calculating the indices with the formula Index = k/100 * N (where k is the desired percentile), and then accessing the item with the calculated index.