GENERAL ASSEMBLY

INTRO TO NUMPY

Haley Boyan

Data Science Instructional Associate, General Assembly DC

LEARNING OBJECTIVES

- Be able to explain and utilize:
    - Measures of Central Tendency (mean, median, and mode)
    - How mean, median and mode are affected by skewness in data
    - Measures of variability (variance and standard deviation)
- Recognize and differentiate various visual descriptive data representations
- Explain pros and cons of SciPy and Numpy python libraries
- Use SciPy and Numpy to calculate descriptive statistics

TITLE

PRE-WORK

PRE-WORK REVIEW

- This should've been completed as pre-work before starting the course, but if you haven't, please watch Lesson 14 (Averages) and 15 (Variance)
    - https://www.udacity.com/course/intro-to-statistics--st101

- Watch: Khan Academy Intro to Central Limit Theorem
    - https://www.khanacademy.org/math/probability/statistics-inferential/sampling-distribution/v/central-limit-theorem

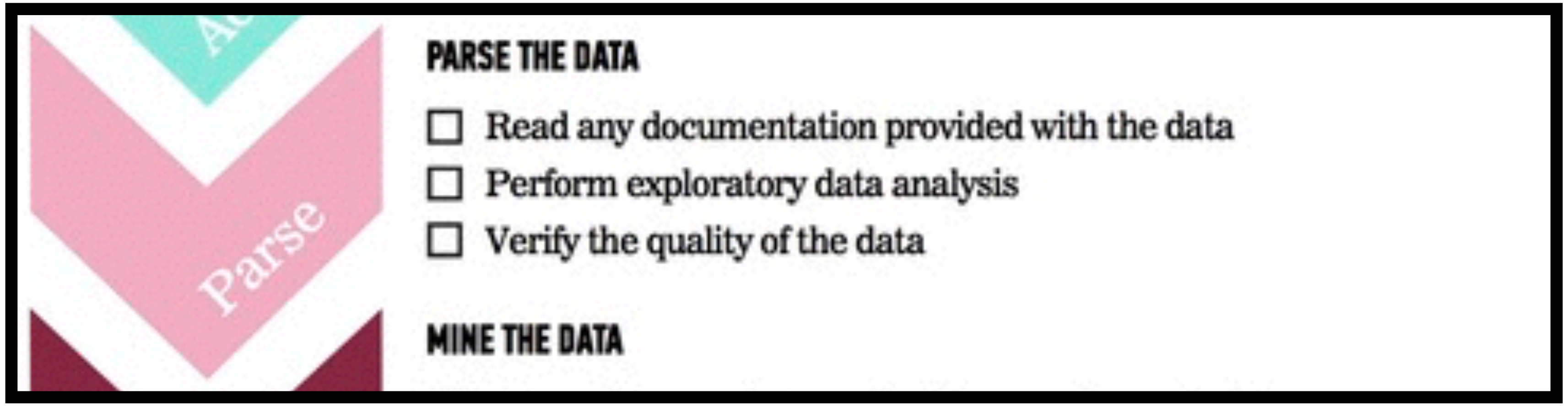
- Watch: Explaining Central Limit Theorem With Bunnies and Dragons
    - http://blog.minitab.com/blog/michelle-paret/explaining-the-central-limit-theorem-with-bunnies-and-dragons-v2

MEET YOUR DATA

FIELDS OF STATISTICS

- There are two fields of stats:
    - Inferential: estimation, hypothesis testing
    - Descriptive: describing, summarizing, and understanding data

- Today's focus is on Descriptive Stats



PARSE THE DATA

□ Read any documentation provided with the data

□ Perform exploratory data analysis

□ Verify the quality of the data

MINE THE DATA

INTRODUCTION: DESCRIPTIVE STATISTICS

DESCRIPTIVE STATS

- Measures of Central Tendency: Mean, Median, Mode
- Skewness
- Measures of Variability: Range, Variance, Standard Deviation

MEASURES OF CENTRAL TENDENCY

- Tell you about the center(s) of your data
- Mean:
  - The sum of the numbers divided by the length of the list
  - Often called “average”
- Median:
  - For odd-length lists: the middle number of the ordered list
  - For even-length lists: the average of the two middle numbers of the ordered list
- Mode:
  - The most frequently occurring number

PRACTICE

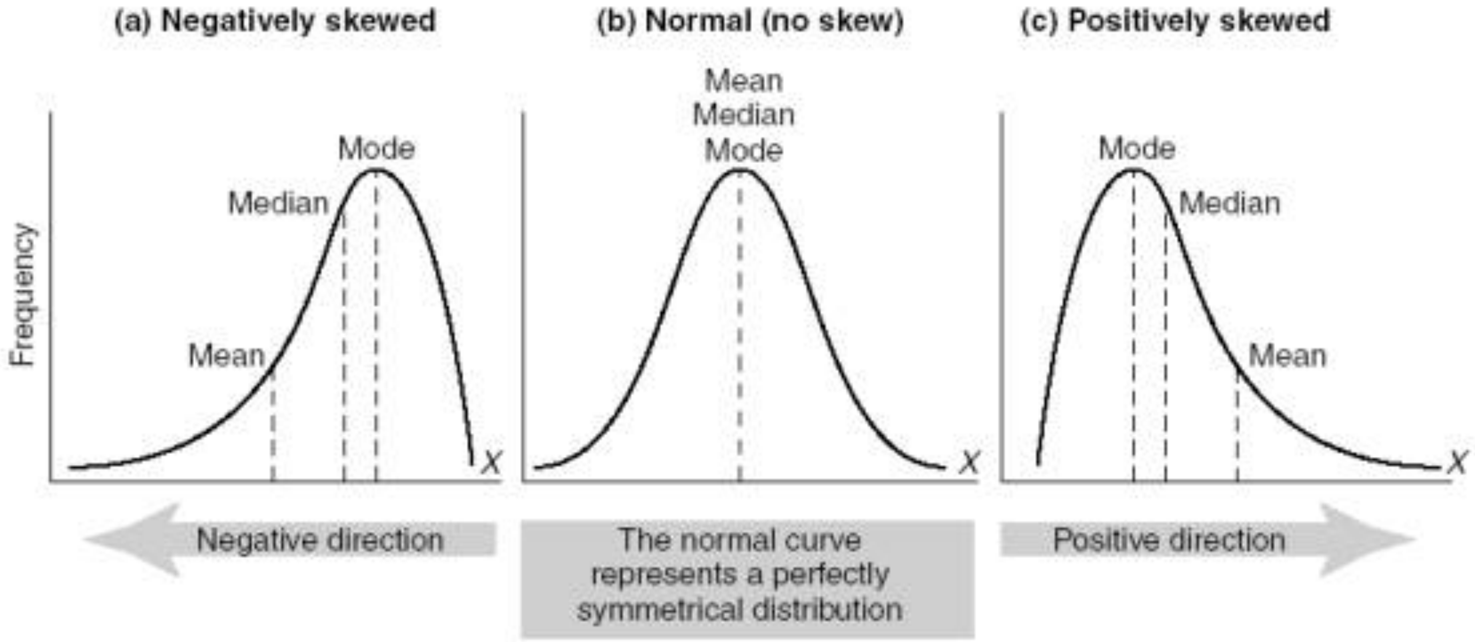
[3, 75, 98, 2, 10, 3, 14, 99, 44, 25, 31, 100, 356, 4, 23, 55, 327, 64, 6, 20]

Find the:
mean
median
mode

Do this:
By hand
In python

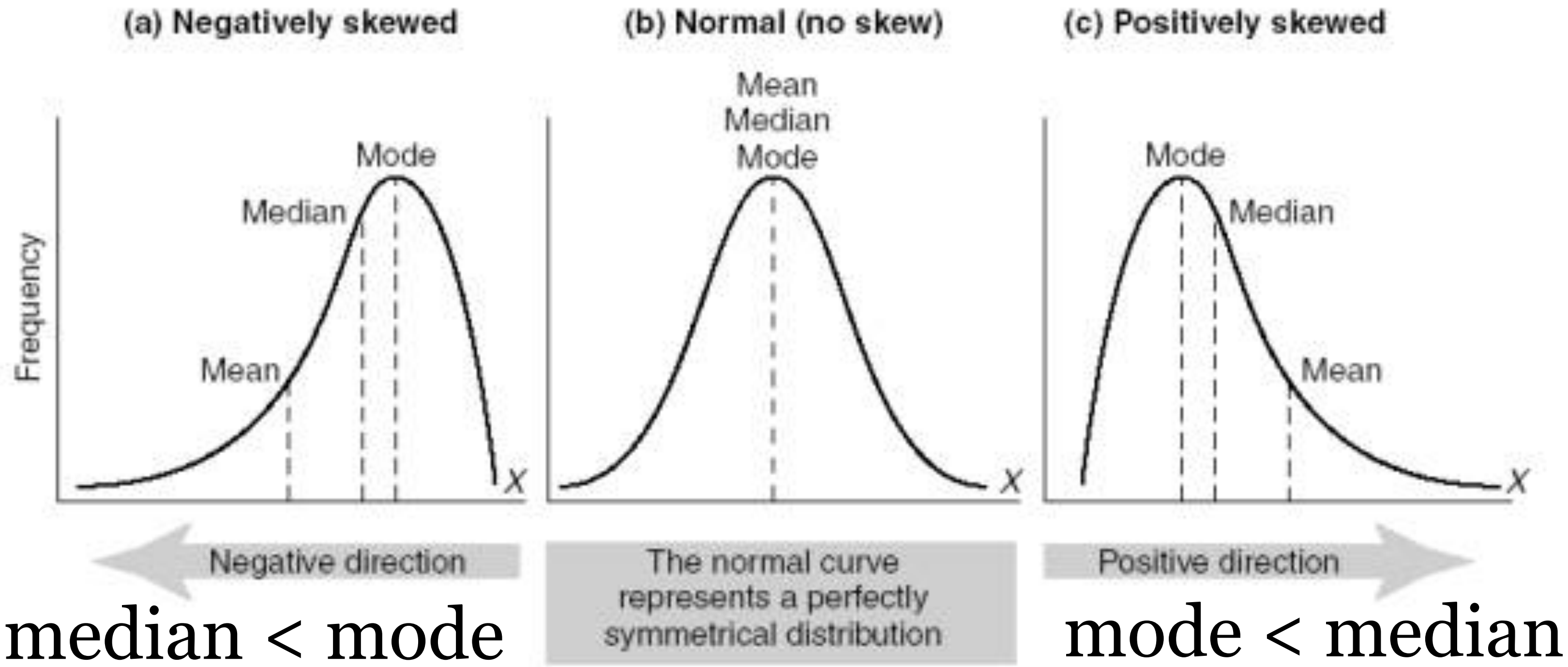
SKEWNESS

- Skewness is lack of symmetry in a distribution of data.
- Positive-skewed: the **right** side tail of the distribution is longer or fatter
- Negative-skewed: the **left** side tail is longer or fatter
- Symmetric distributions have no skewness



CENTRAL TENDENCY AND SKEW

- The mean, median, and mode are affected by skewness
- If the mean < median, the data are skewed left/negative
- If the mean > median, the data are skewed right/positive



mean < median < mode

mode < median < mean

mean = median = mode

MEASURES OF VARIANCE

- Tell you about the spread of your data
- Range: difference between the lowest and highest values of a distribution
- Variance:
    - the average of the sum of the squared distances of each number from the mean of the numbers
    - Shows how widely the numbers distribution varies
    - Squared to avoid negative numbers
- Standard Deviation:
    - Square root of the variance
    - Tells us approximately, on average, the distance of numbers in a distribution from the mean

PRACTICE

[3, 75, 98, 2, 10, 3, 14, 99, 44, 25, 31, 100, 356, 4, 23, 55, 327, 64, 6, 20]

Find the:
range
variance
standard deviation

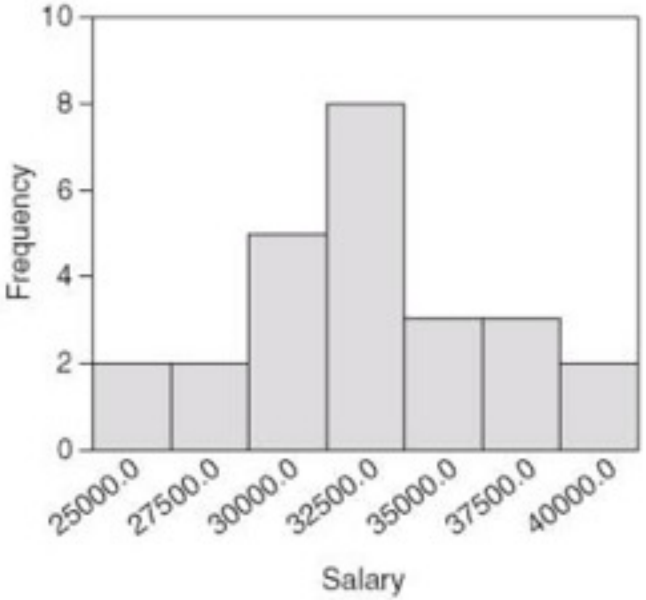
Do this:
By hand
In python

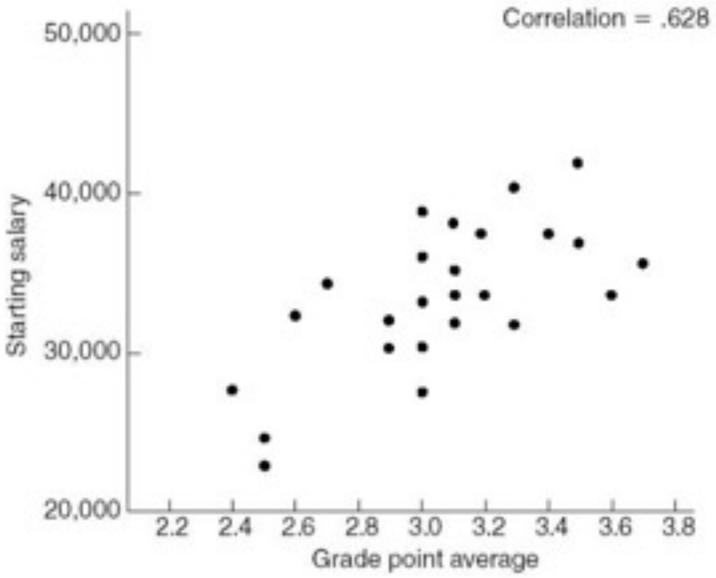
CHECKING IN

- What could a distribution with a large variance look like?
- A small variance?
- A variance of zero?

VISUAL DESCRIPTIVE STATS

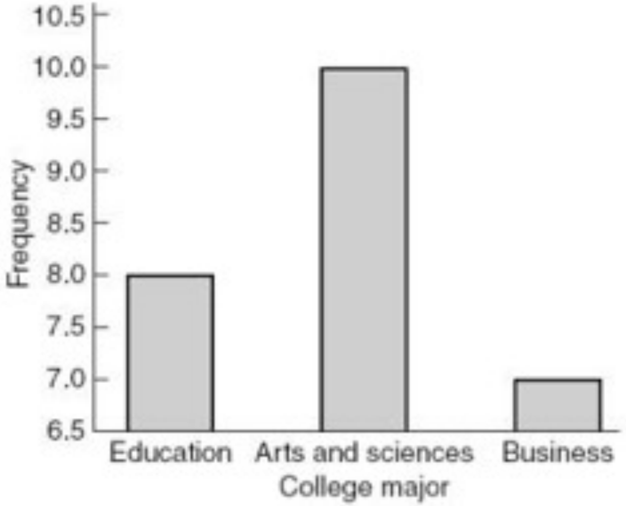
Histogram

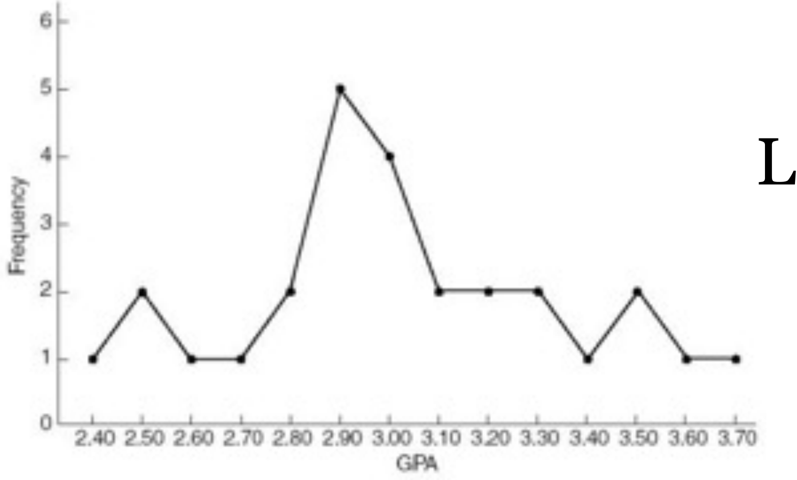




Scatter Plot

Bar Chart





Line Graph

SCIPY AND NUMPY

- Python can do this work for you

- NumPy: Python extension module that provides efficient operation on arrays of homogeneous data. It allows python to serve as a high-level language for manipulating numerical data.
- SciPy: a set of open source scientific and numerical tools for Python. It currently supports special functions, integration, ordinary differential equation (ODE) solvers, gradient optimization, parallel programming tools, an expression-to-C++ compiler for fast execution, and others. A good rule of thumb is that if it's covered in a general textbook on numerical computing (for example, the well-known Numerical Recipes series), it's probably implemented in scipy.
- In an ideal world, NumPy would contain nothing but the array data type and the most basic operations: indexing, sorting, reshaping, basic elementwise functions, et cetera. All numerical code would reside in SciPy. However, one of NumPy's important goals is compatibility, so NumPy tries to retain all features supported by either of its predecessors. Thus NumPy contains some linear algebra functions, even though these more properly belong in SciPy. In any case, SciPy contains more fully-featured versions of the linear algebra modules, as well as many other numerical algorithms. If you are doing scientific computing with python, you should probably install both NumPy and SciPy. Most new features belong in SciPy rather than NumPy.

- **In short: NumPy is simpler, easier to use, and less computationally expensive. If NumPy can't do it, SciPy probably can.**

DEMO: SCIPY AND NUMPY

INTRO TO STATS WITH NUMPY

- See demo_code notebook

GUIDED PRACTICE: STATS WITH NUMPY

PRACTICE STATS WITH NUMPY/SCIPY

EXERCISE

DIRECTIONS

1. Import numpy
2. Generate a series of random numbers with a normal distribution:
   ```
   numpy.random.randn(50)
   ```
3. Using numpy, find the:
   - mean
   - median
   - range
   - variance
   - standard deviation
4. Use scipy to find the mode
5. Using the mean, median, and standard variation, describe how you think the data will look if graphed
6. Plot the data to check your the actual distribution:
   ```
   import matplotlib.pyplot as plt
   %matplotlib inline
   plt.hist(numlist)
   ```

INDEPENDENT PRACTICE

WRITING A NUMPY FUNCTION

ACTIVITY: STATS IN A FUNCTION

EXERCISE

DIRECTIONS

Define a function that:

1. Allows the user to input a maximum number
2. Generates a set of 50 random integers between 0 and the max
3. Prints out the mean, median, mode, range, standard deviation, and variance
4. Describes the skewness of the data

DELIVERABLE

Python code delivered via google form (sent on Slack)

CONCLUSION

INTRO TO STATS

INTRO TO STATS

- What do we use to describe the center of a dataset?
- What do we use to describe the distribution of a dataset?
- How can we use statistics to interpret data without looking at it visually?

BEFORE NEXT CLASS

BEFORE NEXT CLASS

DUE DATE: 9AM TOMORROW (THURS)

- Homework: 4 of 9 practice problems in 1.09-Numpy-Lab
- Submitted as a pull request (create a folder titled YourNameNumpyLab) within the 1.09-Numpy-Lab folder

TITLE

CREDITS

CITATIONS

- http://www.southalabama.edu/coe/bset/johnson/lectures/lec15.htm
- https://www.scipy.org/scipylib/faq.html#what-is-the-difference-between-numpy-and-scipy

TITLE

Q & A