Introduction to Statistics

Justin Pounders

Objectives

By the end of the day you should be able to

- Describe where data comes from
- Define and calcualte the measures of central tendency (mean, median, mode)
- Describe how mean, median and mode are affected by skewness
- Define measures of variability (variance and standard deviation)

Warmup: Think, pair, share

Why learn statistics?

1. 60 sec: Think independently
2. 120 sec: Share with a partner
3. Volunteers will share with the class.

The goal of descriptive statistics is to summarize a collection of observations (aka data). Our observations are probabilistic or random.

Basic Probability Concepts

Let's define:
- categorical/discrete data vs. continuous data
- probability mass function
- probability density function
- random observation

Probability Concepts

— **probability mass function**: probabilities of **discrete/ categorical** data

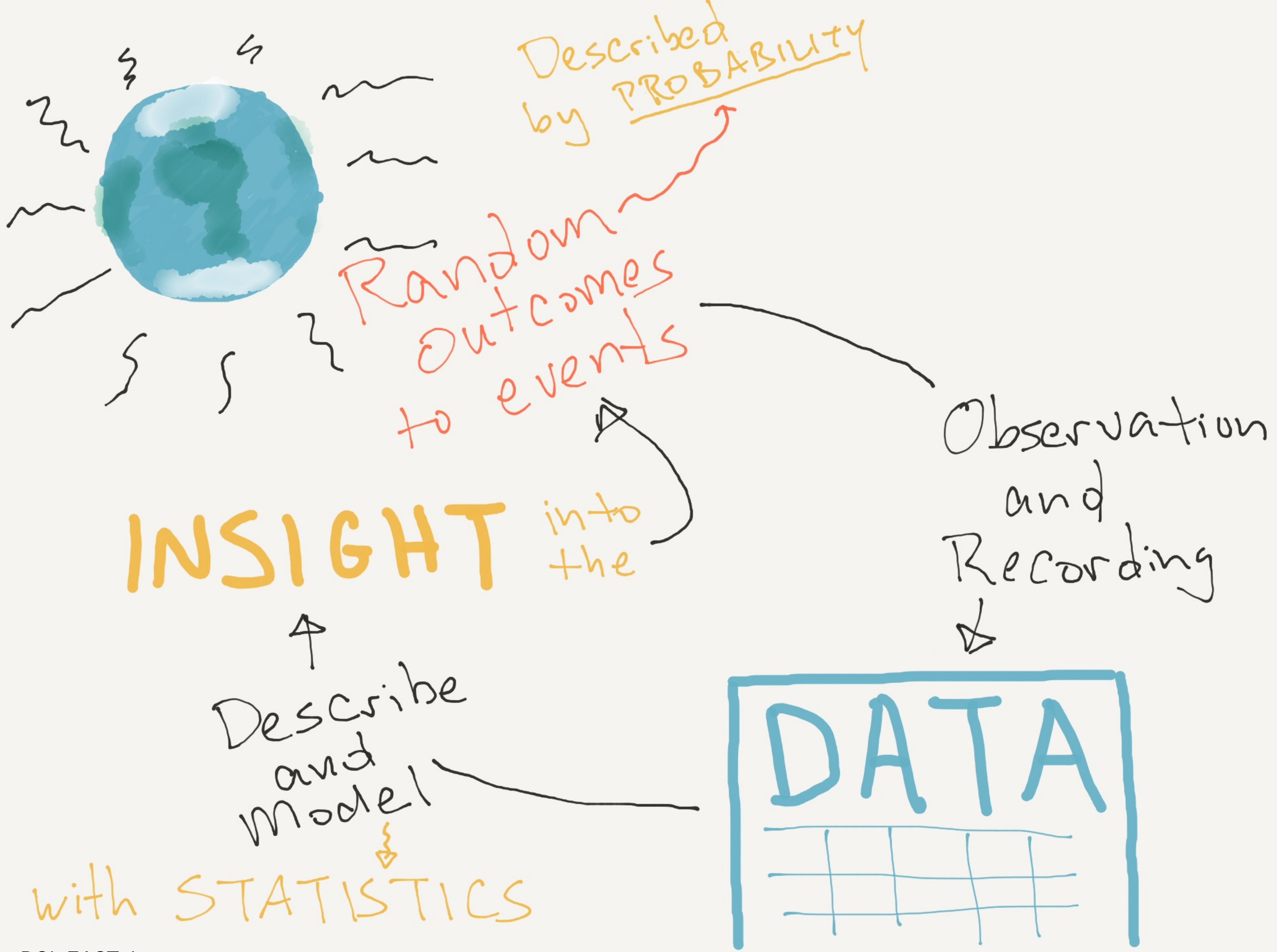
— **probability density function**: probabilities of **continuous** data

Why should I care?

Data collection is a probabilistic process!

— There is some process happening in the world
— I observe and record the outcome of this process
— My collection of observations is my data
— I can learn about the process from my **data**

The world is a random (probabilistic) place. **Statistics** is the tool allowing you to discover and quantify that randomness!


Described by PROBABILITY
Random outcomes to events
Observation and Recording
DATA
Describe and Model
with STATISTICS
INSIGHT into the

Let's do some Stats

Measures of Central Tendency

— mean
— median
— mode

A man walks into a bar...

Goal: estimate the average salary of a region.

Method: poll the people at a local bar

Data: salary responses


$80K
$97K
$73K
$67K

Average Salary

```
salaries = [80, 73, 97, 67]
mean = sum(salaries)/len(salaries)
```

mean = 79.25

Calculating the Mean

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

— $\bar{x}$ is the mean (or average)
— the x_i values are the data ($i = 1, 2, 3, \ldots$)
— N is the number of observations/data points

Mean as a Measure

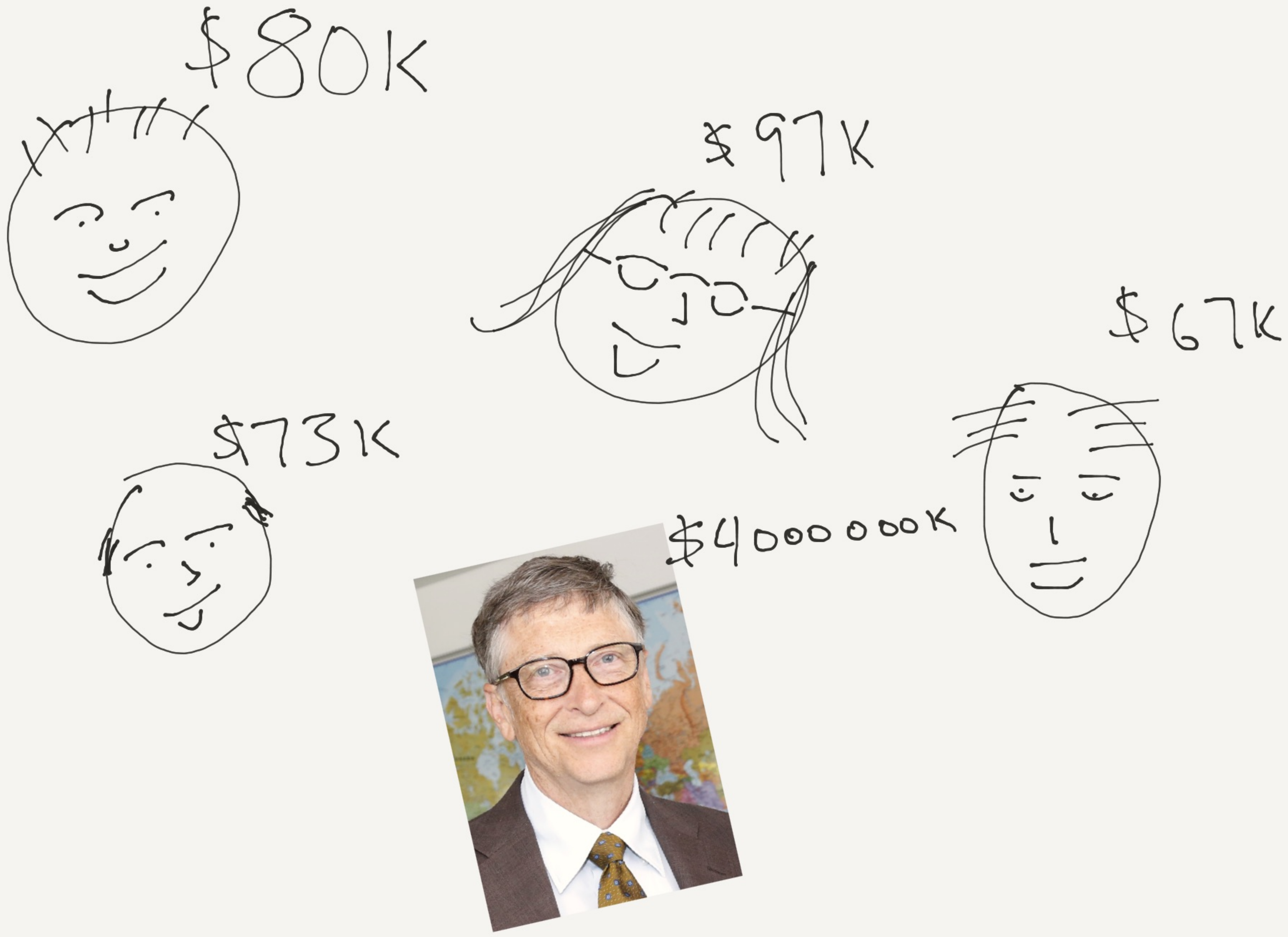
The mean measures the "center" of the data.

Mean as a Measure

The mean measures the "center" of the data.

Can you think of a potential problem with the mean as a measure of centrality?

(Slack your response)


$80K
$97K
$67K
$73K
$4000000K

Average Salary

```
salaries = [80, 73, 97, 67, 4000000]
mean = sum(salaries)/len(salaries)
```

mean = 800063.4 = $800M

Median as a Measure

- The median is less sensitive to **outliers**
- Also measures the "center" of an observed data set

Median Calculation

```
1. sort data (e.g. smallest to largest)
2. if N is odd:
       median = middle number
   else if N is even:
       median = average of middle two numbers
```

Median Calculation

```
1. sort data (e.g. smallest to largest)
2. if N is odd:
       median = middle number
   else if N is even:
       median = average of middle two numbers
```

Example:

```
salaries_woGates = [67, 73, 80, 97]
salaries_wiGates = [67, 73, 80, 97, 4000000]
```

Mode

The mode is the **most common** value in the data.

- May be more than one: `[1, 2, 1, 2, 3]`
- May not be one: `[1, 2, 1, 2]`
- Most useful with categorical data:

```
[elephant, cat, dog, dog, cat, cat, whale]
```

Codealong

goto notebook
descriptive-statistics-and-numpy.ipynb
Section 1

Measures of Dispersion

— range
— variance
— standard deviation

Range

Range is simply

$$(\text{max value}) - (\text{min value})$$

```
salaries = [80, 73, 97, 67]
range_sal = max(salaries) - min(salaries)
```

range = 30

Variance

The variance of a (finite) population is

$$\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Variance

The variance of a (finite) **population** is

$$\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

The variance of a **sample**

$$\mathrm{Var}(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Variance

Population variance

```
N = len(salaries)
pop_var = np.sum((salaries - np.mean(salaries))**2)/N
# = 126.1875
```

Sample variance

```
N = len(salaries)
samp_var = np.sum((salaries - np.mean(salaries))**2)/(N - 1)
# = 126.1875
```

Standard Deviation

The standard deviation is

$$\sigma_x = \sqrt{\mathrm{Var}(x)}$$

Standard Deviation

The standard deviation is

$$\sigma_x = \sqrt{\mathrm{Var}(x)}$$

Q: What is the advantage of quoting standard deviation instead of variance?

Raise your hand to answer.

Codealong

goto notebook
descriptive-statistics-and-numpy.ipynb
Section 2

Covariance and Correlation

Covariance

Covariance answers the question, "How do two variables vary with respect to one another?"

Sample Covariance

Take N observations of two variables: x_i and y_i

$$\mathrm{Cov}(x) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

Covariance and Correlation

Take N observations of two variables: x_i and y_i

$$\mathrm{Cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{Corr}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$$

Correlation

- Correlation coefficients are **always** between -1 and 1.
- Sometimes labeled $\rho_{x,y}$

$$\rho_{x,y} = \mathrm{Corr}(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

Interpreting Correlation

— Values close to -1 or +1 indicate a strong, linear relationship between the two variables.

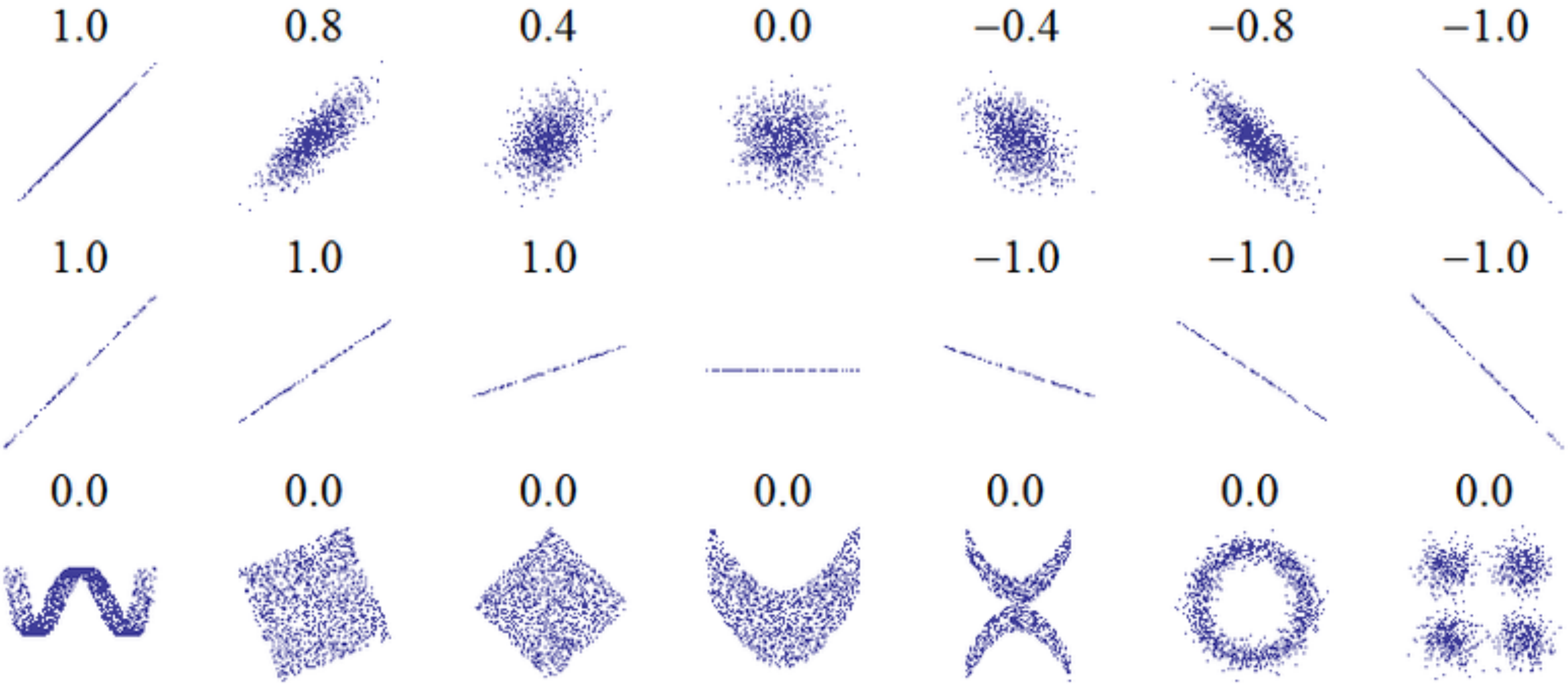
— Values close to 0 indicate a weak and/or nonlinear relationship between the two variables.

— Values above 0 indicate a positive relationship between the two variables.

— Values below 0 indicate a negative relationship between the two variables.

Codealong

```
goto notebook
descriptive-statistics-and-numpy.ipynb
Section 3
```


1.0
0.8
0.4
0.0
–0.4
–0.8
–1.0
1.0
1.0
1.0
–1.0
–1.0
–1.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0