Homework 1

Andre Ye

3 April 2021

Problem 1

Context: For each of the data sets below, (i) compute the median, first quartile, and third quartile; (ii) compute the arithmetic mean and sample standard deviation; and (iii) explain what conclusions one could reasonably draw based on that information.

Part A Question: `[1, 2, 5, 3, 2, 6, 1, 4, 3, 1, 2]`

Part A Solution: Sorting the array yields `[1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 6]`. This yields a first quartile of 1, a median of 2, and a third quartile of 4. Summing the array yields 30; dividing by 11 (the number of elements) yields a mean of about 2.73. To calculate the standard deviation, we begin by taking the difference of each element and the mean; this yields [-1.73, -1.73, -1.73, -0.73, -0.73, -0.73, 0.27, 0.27, 1.27, 2.27, 3.27]. Next, we must square each of these, yielding [2.98, 2.98, 2.98, 0.53, 0.53, 0.53, 0.07, 0.07, 1.62, 5.17, 10.71]. Summing these square differences yields 28.17; we need to divide this by one less than the length of the array, which comes to 10. Dividing yields 2.817; square-rooting this results in a standard deviation of ≈ 1.68. From this data, we can conclude that the data is right-skewed; there appears to be a greater "density" of smaller numbers. The general "Mode < Median < Mean" requirement for right-skewed distribution is also mostly met.

Part B Question: `[22.6, 88.3, 67.2, 93.6, 95.7, 89.7, 72.6, 99.9, 36.7]`

Part B Solution: Sorting the array yields `[22.6, 36.7, 67.2, 72.6, 88.3, 89.7, 93.6, 95.7, 99.9]`. This yields a first quartile of 51.95, a median of 88.3, and a third quartile of 94.65. Summing the array yields 666.3; dividing this by 9 (the length of the array) yields a mean of 74.03. Using the same process as above to calculate the standard deviation yields 27.52. From this, we can tell that the data is "denser" in higher numbers and is left-skewed.

Part C Question: `[18, 77, 19, 65, 22, 93, 18, 88]`

Part C Solution: Sorting the array yields `[18, 18, 19, 22, 65, 77, 88, 93]`. This yields a first quartile of 18.5, median of 43.5, and a third quartile of 82.5. Summing the array yields 400; dividing by 8 (the number of elements in the array) yields a mean of 50. Using the same process as outlined in Part A, we can derive a standard deviation of 33.89. From this data, we can tell that the distribution is right-skewed.

Problem 2

Context: In each of (a) through (c), make up an example of a pair of data sets matching the given condition, with at least five data points in each set. To explain your reasoning for these problems, describe the thought process you used for constructing your data sets.

Part A Question: The two data sets have the same median, but different means.

Part A Solution: There are a few ways we could approach this. For instance, we could have two identical arrays of n elements, then change the value of the nth index (or the 1st index). Of course, this assumes that $n \geq 3$. Changes to this value keep the median but have different means. Alternatively, we could construct two completely random arrays, calculate the index (or indices) of the median value, and set those two to the same value. This model has the slight disadvantage in that it is possible for these two randomly generated arrays to have the same mean.

Given this, we will opt for the first strategy. Let's begin with two arrays, `arr1` and `arr2`:

1. `arr1 = [1, 2, 3, 4, 5]`

2. `arr2 = [1, 2, 3, 4, 5]`

Changing the last element of `arr2` yields `[1, 2, 3, 4, 6]`. The median of both has remained 3, but the average has changed. Thus, the two data sets we will use are [1, 2, 3, 4, 5] and [1, 2, 3, 4, 6].

Part B Question: The two data sets have the same mean, but different medians.

Part B Solution: There are quite a few approaches we could take for this. Two include:

- We can have an array of the same value; for instance, `arr1 = [3, 3, 3, 3, 3]`. The mean of both arrays should be 3, although the median needs to be different. We can thus have `arr2 = [x, y, y, y, y]` where `y`≠median. We can then solve for `x` such that $\texttt{x} + 4\texttt{y} = 5 \cdot 3$.
- We can construct a dataset beginning from 0, like `arr1 = [0, 1, 2, 3, 4]`. We can decrease the median by 1; it remains the median. Because averages are concerned with the sum of elements, to account for this we can add 1 to the first element of the array, yielding `arr2 = [1, 1, 1, 3, 4]`. The two arrays have different medians but the same mean. The usage of 0 as an "empty element" that is present but "does not hold a value" (roughly speaking) helped make this "trick" work.

Thus, the two data sets we will use are [0, 1, 2, 3, 4] and [1, 1, 1, 3, 4].

Part C Question: The two data sets have the same mean and median, but different sample standard deviations.

Part C Solution: Standard deviation describes how "spread apart" the data is. Thus in constructing our data, we should change how "spread out" the data is while keeping the mean and median the same. We can begin with the dataset `[0, 0, 0, 0, 0]` and make changes to create a second dataset. We can keep the median the same by not changing the third value. By decreasing the first element by a and increasing the fifth element by a, we have `[-a, 0, 0, 0, a]`; this keeps the mean and median the same, but the spread is vastly different as long as $a \neq 0$. Thus, the two data sets we will use are [0, 0, 0, 0, 0] and [1, 0, 0, 0, 1].

Problem 3

Context: In a certain neighborhood, the properties are all perfect squares. The lengths of each property, ordered by address, are given in the table below.

Part A Question: Find the mean, median, and standard deviation of the property length in this neighborhood.

Part A Solution: The property lengths in the neighborhood are [100, 150, 95, 170, 120, 200, 190, 180, 100, 150, 300, 30]. Sorting the array yields [30, 95, 100, 100, 120, 150, 150, 170, 180, 190, 200, 300]. It follows that the first quartile is 100 feet, the median is 150 feet, and the third quartile is 185 feet. The mean is 148.75 feet. The standard deviation is, using the method outlined in Problem 1A but using the population rather than the sample variant, is 65.32 feet.

Part B Question: Convert the property lengths into yards, and find the mean, median, and standard deviation of the property length in yards. How does this relate to your answer to (a)?

Part B Solution: Converting the property lengths into yards entails dividing each of the elements by 3, which yields [10.0, 31.67, 33.33, 33.33, 40.0, 50.0, 50.0, 56.67, 60.0, 63.33, 66.67, 100.0]. Because the indices of the locations of the quartiles remain the same, the new quartiles can be derived by dividing the quartiles of the property lengths in feet by 3. This yields a a first quartile of 33.33 yards, a median of 50 yards, and a third quartile of 61.67 yards. We can similarly approach the mean algebraically; the mean is $\frac{1}{N}\sum\frac{x_i}{3}$, where x_i is the dataset in feet as used in Part A. From sigma rules (or some strategic testing), the $\frac{1}{3}$ can be extracted; the mean becomes $\frac{1}{3N}\sum x_i$, which is one-third of the mean derived in Part A, yielding 49.58 yards.

For standard deviation, we have $\sigma = \sqrt{\frac{\sum(\bar{x}-x_i)^2}{N-1}}$. We need to prove that the standard deviation of data divided by 3 is the same as the standard deviation on data divided by 3. As an assumption, we will utilize the fact that the mean is linearly scaled when the data is linearly scaled.

$$\frac{\sqrt{\frac{\sum(\bar{x}-x_i)^2}{N-1}}}{3} = \sqrt{\frac{\sum\left(\frac{\bar{x}}{3}-\frac{x_i}{3}\right)^2}{N-1}}$$
$$\sqrt{\frac{\sum(\bar{x}-x_i)^2}{N-1}}}{9} = \sqrt{\frac{\sum\left(\frac{\bar{x}-x_i}{3}\right)^2}{N-1}}$$
$$\frac{\sum(\bar{x}-x_i)^2}{9(N-1)} = \frac{\sum\frac{(\bar{x}-x_i)^2}{9}}{N-1}$$
$$\frac{\sum(\bar{x}-x_i)^2}{9(N-1)} = \frac{\sum(\bar{x}-x_i)^2}{9(N-1)}$$

The two are equivalent. We can thus divide the standard deviation by 3, yielding 21.77 yards.

Part C Question: Find the mean, median, and standard deviation of the property area in this neighborhood. Is there a relationship between your answer and the answer to (a) similar to what you described in (b)? Why or why not?

Part C Solution: Converting the property lengths into areas entails squaring them, as the problem gives us that the properties are exact squares. This yields a data set of [900, 9025, 10000, 10000, 14400, 22500, 22500, 28900, 32400, 36100, 40000, 90000]. Because this transformation is not linear and this dataset happens to have quartiles that are averages of two adjacent elements, we cannot immediately square the quartiles. We can prove that assertion through contradiction:

$$\left(\frac{x_i+x_{i-1}}{2}\right)^2 = \frac{x_i^2+x_{i-1}^2}{2}$$
$$x_i^2+2x_ix_{i-1}+x_{i-1}^2 = x_i^2+x_{i-1}^2$$
$$2x_ix_{i-1} = 0$$

In our dataset, no element is 0, and thus squaring the quartiles can never be equivalent to the quartile of squared elements.

However, we can use the handy fact that in the dataset, $x_i = x_{i-1}$ in the first and second quartiles for this particular dataset. Under this assumption, we have:

$$\left(\frac{x_i+x_{i-1}}{2}\right)^2 = \frac{x_i^2+x_{i-1}^2}{2}$$
$$\left(\frac{2x_i}{2}\right)^2 = \frac{2x_i^2}{2}$$
$$x_i^2 = x_i^2$$

Thus, we can derive the first and second quartiles by squaring given this coincidence, this yields a first quartile of 1000 square feet and a median of 22500 square feet. For the third quartile, we more manually take the average of 32400 and 36100, yielding a third quartile of 34250 square feet.

Let us prove that we cannot find the average of squared elements by squaring the average.

$$\frac{x_1^2+x_2^2+\cdots+x_N^2}{N} = \left(\frac{x_1+x_2+\cdots+x_N}{N}\right)^2$$
$$x_1^2+x_2^2+\cdots+x_N^2 = \frac{(x_1+x_2+\cdots+x_N)^2}{N}$$

The $(x_1+x_2+\cdots+x_N)^2$ of the RHS will form "cross terms" like x_1x_2 or x_2x_N that are not "accounted for" in the LHS, which only contains "square terms". Furthermore, $N \neq 1$, complicating the possibility of the LHS equalling the RHS. In this case, we must find the average "manually" by finding the average of square elements, which yields 26393.75 square feet.

Lastly, let us attempt to prove that we cannot find the standard deviation by squaring the standard deviation. Let $\bar{x}_s$

represent "the average of the squared x_is".

$$\left(\sqrt{\frac{\sum(\bar{x}-x_i)^2}{N-1}}\right)^2 = \sqrt{\frac{\sum(\bar{x_s}-x_i^2)^2}{N-1}}$$

$$\frac{\sum(\bar{x}-x_i)^2}{N-1} = \sqrt{\frac{\sum(\bar{x_s}-x_i^2)^2}{N-1}}$$

$$\frac{\sum(\bar{x}-x_i)^2 \cdot \sum(\bar{x}-x_i)^2}{(N-1)^2} = \frac{\sum\left(\bar{x_s}-x_i^2\right)^2}{N-1}$$

$$\frac{\sum(\bar{x}-x_i)^2 \cdot \sum(\bar{x}-x_i)^2}{N-1} = \sum\left(\bar{x_s}-x_i^2\right)^2$$

$$\sum(\bar{x}-x_i)^2 \cdot \sum(\bar{x}-x_i)^2 = \sum(N-1)\left(\bar{x_s}-x_i^2\right)^2$$

At this point, the expression has gotten too complex and it is not clear where one should move on from here. Intuitively, the two expressions do not seem to be equivalent, although the dependency of the standard deviation on the mean, which cannot be calculated via simple squaring transformations, and the complicated nature of standard deviation itself, makes the task of proving this "formally" a complex one. One could disprove the validity of squaring the standard deviation to find the standard deviation of squared elements by posing a counterexample, which one imagines would not be difficult to come up with. For the purposes of this Collingwood, we will not take the "shortcuts" that could be taken advantage of in the linear transformations of Part B and simply take the standard deviation of square elements, yielding 22417.89 square feet.

Problem 4

Context: Over the last twenty years, the annual rainfall in Seattle and in Portland, Oregon are given below, measured in inches.

Part A Problem: Find the mean, median, and standard deviation of the rainfall in each location.

Part A Solution: Using methods established earlier in this homework, we can find the following:

City	Mean	Median	Standard Deviation
Seattle	39.05	38.44	6.65
Portland	36.87	34.96	8.33

Standard deviation was found using sample standard deviation, since the data samples is being used as a sample to make statements about the rainfall in general in Seattle and Portland.

Part B Problem: Based on this data, what conclusions can we make about the relative rainfall in Seattle and Portland with 90% confidence? With 95% confidence? (I.e., are they equal? Does it rain more in Portland?)

Part B Solution: The 90% confidence interval for Seattle can be found as:

$$\left[39.05 - 1.65\frac{6.65}{\sqrt{29}}, 39.05 + 1.65\frac{6.65}{\sqrt{29}}\right] \approx [37.01, 41.09]$$

The 90% confidence interval for Portland can be found as:

$$\left[36.87 - 1.65\frac{8.33}{\sqrt{29}}, 36.87 + 1.65\frac{8.33}{\sqrt{29}}\right] \approx [34.32, 39.42]$$

These two intervals intersect each other; thus, we cannot make a statement that Seattle and Portland's mean rainfall is more, less, or equal to each other. Since increasing the confidence interval broadens the intervals, a 95% confidence would not allow us to make a statement on relative rainfall either.

Problem 5

Question: Was Jay's conclusion a reasonable one? Why or why not?

Solution: No, it is not a reasonable conclusion. One can recall from formalized theorems or from a little bit of strategic experimentation prime numbers are less likely to occur in higher numbers. In fact, the distribution of prime numbers in most ranges is right-skewed; this does not resemble the symmetric nature of the normal distribution that is roughly needed for the application of confidence intervals. Furthermore, Jay's context is problematic in the sample is not representative of the entire data set of primes, as it is the *first* 10,000 prime numbers. Given that Jay's database was not randomly sampled from the entire distribution of prime numbers, Jay cannot make a statement about all primes. However, he could randomly sample, say, 2,000 random prime numbers from the first 10,000 primes, use transformations to form a normal distribution, and use confidence intervals to make claims about the mean of the first 10,000 primes.

Problem 6

Question: Open the CDC database on COVID-19 cases by state (the link may be found on Canvas). Washington and California were among the first states to be affected; locate their total case numbers for the week from March 7 to March 14. Based on that information, design and execute a method for determining which state had a faster rate of spread. What does your conclusion tell you about which state, if either, had a better response?

Solution: The following data will be used, collected from the provided database:

Date	WA Total Case Count	CA Total Case Count
3/7	102	56
3/8	136	110
3/9	162	135
3/10	267	152
3/11	366	175
3/12	457	224
3/13	568	264
3/14	642	311

We can begin by finding models that can represent each state well. Two forms of models seem fitting: exponential and polynomial (quadratic, cubic). Afterwards, we can take the derivative of both models to analyze change over time.

Firstly, let's define an error metric. We want to consider the error relative to the absolute value of the case count.

```
import numpy as np
def error(truth, pred):
    diff = np.abs(truth - pred)
    relative_diff = diff/truth
    return np.mean(relative_diff)
```

Fitting an exponential model:

```
import scipy
import scipy.optimize
x = np.array(range(0,8))
y = np.array([102, 136, 162, 267, 366, 457, 568, 642])
scipy.optimize.curve_fit(lambda t,a,b: a*np.exp(b*t),  x,  y)
```

Fitting yields this model:

```
def exp_model(x):
    return 125.92692507*np.exp(0.24182737*x)
```

The error is `0.12398322092759508`.

Fitting a quadratic model:

```
scipy.optimize.curve_fit(lambda t,a,b,c, d: a*(t**2) + b*t + c,  x,  y)
```

Fitting yields this model:

```
def quadratic_model(x):
    return 5.2976191*(x**2) + 45.34523755*x + 86.08333449
```

The error is `0.06546637225318994`.

Lastly, fitting a cubic model:

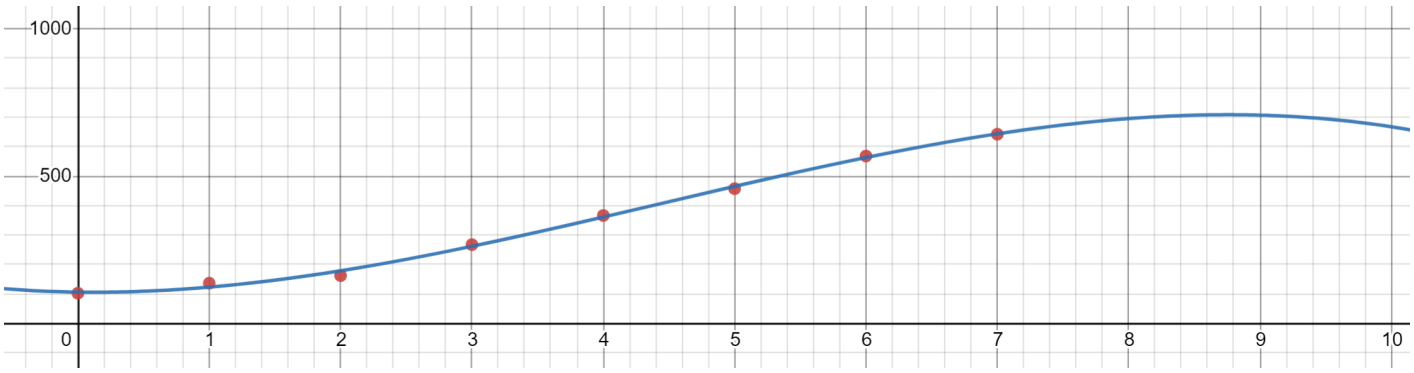
```
scipy.optimize.curve_fit(lambda t,a,b,c,d: a*(t**3) + b*(t**2) + c*t + d,  x,  y)
```

Fitting yields the model:

```
def cubic_model(x):
    return (-1.87373737)*(x**3) + 24.97186147*(x**2) + (-6.18253968)*x + 105.75757576
```

The error is `0.03744478665736465`.

This cubic model has the lowest error. Of course, we are testing on the same set we are training on – a bad practice, but this dataset is so small and the models are simple that we can ignore it. Visualizing as a sanity check shows that, at least within the range of [0,7], the model fits pretty well.



Using the same process on California data yields the following results:

Model	Error
Exponential	0.09686287315310975
Quadratic	0.0719820193552362
Cubic	0.04433825687344809

Thus, the two models are:

$$f_{\text{WA}}(x) = -1.87x^3 + 24.97x^2 - 6.18x + 105.76$$

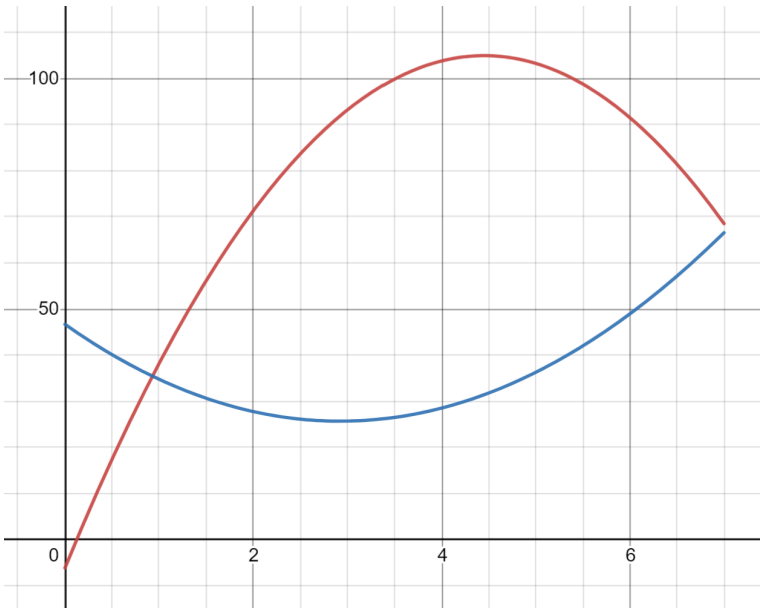
$$f_{\text{CA}}(x) = 0.82x^3 - 7.19x^2 + 46.67x + 60.89$$

Taking derivatives yields

$$f'_{\text{WA}}(x) = -5.61x^2 + 49.94x - 6.18$$

$$f'_{\text{CA}}(x) = 2.46x^2 - 14.38x + 46.67$$

Graphing these in the domain of [1,7] gives an interesting result (red = Washington, blue = California).



Of course, there are problems with the simplicity and assumptions of the models that we made. Our analysis really should focus only on the domain of [1,6] or even [2,5] (and even then with some hesitancy on the edges), because the ends of the domain are not "padded" on both sides by data points that guide a good model fit in that region. While we can make many claims from an analysis of the derivative, the only one we can say with a relatively high level of confidence is that in the first few days, Washington contained the spread worse than California did; cases spread much more quickly than in California.