Homework 3

Andre Ye

19 April 2021

Problem 1

Context: This problem is intended to help conceptualize confidence intervals. Load up the links "Confidence Intervals I", "Confidence Intervals II", and "Confidence Intervals III", linked in the assignment description, and mess around with them enough to get a sense of what they're doing. Then do the following.

Part A Problem: Choosing any reasonable parameters you like, but keeping the Sample Size at default, run each simulation 50 times and note down how many times the true mean fell within the 80% confidence interval. Just tally; you don't need to record the intervals!

Part A Solution: The results are tallied below (parameters were kept on the defaults):

Simulation	Frequency	Proportion
Confidence Intervals I	39/50 sample means within interval	0.78
Confidence Intervals II	42/50 sample means within interval	0.84
Confidence Intervals III	41/50 sample means within interval	0.82

Part B Problem: Increase the Sample Size to 100, and run each simulation 50 times again, noting down how many times the true mean fell within the 80% confidence interval. Just tally; you don't need to record the intervals!

Part B Solution: The results are tallied below (parameters were kept on the defaults):

Simulation	Frequency	Proportion
Confidence Intervals I	42/50 sample means within interval	0.84
Confidence Intervals II	44/50 sample means within interval	0.88
Confidence Intervals III	44/50 sample means within interval	0.88

Part C Problem: What do you notice about your results? This is an open-ended question; give it some thought. You should be able to say something meaningful.

Part C Solution: I observe that – at least in my results – for all the simulations I ran in Part A and in Part B, somewhere near 80% of the sample means fall within the confidence interval. This makes sense, as it is the definition of a confidence interval: choosing a confidence interval of k% means that roughly k% of randomly sampled means will fall within the range.

This result is surprising because one may have expected confidence intervals to perform better on normal distributions. Confidence Intervals I is normally distributed, but Confidence Intervals II and III do not appear to be so. However, we can attribute that the proportion of sample means that lie within the confidence interval remains relatively similar across all simulations to the "versatility" or "universality" of the normal distribution. That Confidence Intervals II and III have a larger proportion may just be random chance.

I also notice an increase in the proportion of sample means that fall within the interval when each sample contains 100, rather than 10, sample sizes. This can likely be attributed to the fact that confidence intervals work better on larger distributions, and that the results in Part A used a small sample size that would have been better handled by a t-distribution.

Problem 2

Context: This problem is intended to help conceptualize F-tests. Load up the link "F-Tests", linked in the assignment description, and mess around with it enough to get a sense of what it's doing. Then do the following.

Part A Problem: Recall that the threshold for an F-test with 90% confidence with 10 samples is 3.29. Keeping the True Slope at 0, run the simulation 50 times and note down how many times the F value was above this threshold. Just tally; you don't need to record the values!

Part A Solution: Out of 50 times the simulation was run, the F-test was larger than 3.29 a total of 4 times. This is a frequency of 8%.

Part B Problem: Set the True Slope to 2 and the True Intercept to 1. Run the simulation 50 times and note down how many times the F value was above this threshold. Just tally; you don't need to record the values!

Part B Solution: Out of 50 times the simulation was run, the F-test was larger than 3.29 a total of 47 times. This is a frequency of 94%.

Part C Problem: Do (a) and (b) again with 100 samples instead of 10, using 2.71 as your threshold. Just tally; you don't need to record the values!

Part C Solution:

True Line	Number of Times F-Test > 2.71	Proportion
$y = 0$	8/50	0.16
$y = 2x + 1$	50/50	1

Part D Problem: What do you notice about your results? This is an open-ended question; give it some thought. You should be able to say something meaningful.

Part D Solution: We notice that as the sample size increases, the proportion of best-fit lines on the randomly generated data that satisfies the 90% F-test increases, regardless of the slope and y-intercept. This is likely because of the $(N-2)$ term in the $F = \frac{R^2}{1-R^2}(N-2)$ formula; as N increases, F increases, even if the R^2 remains the same (or even is slightly worse).

We can also observe that generating data on the "true line" $y = 2x + 1$ rather than on $y = 0$ drastically increases the proportion, across both sample sizes. We can hypothesize that this is due to the nature of the R^2 metric used in the F-test. The R^2 metric "compares" how good the line of best fit is to a model that simply returns the average of the y-values. In the case of $y = 2x + 1$, any line of best fit that points in roughly the right direction would perform better than $y = \bar{x}$. However, when data is generated around $y = 0$, the best line of fit *is* a model that returns the average every time. Therefore, it is "more difficult" to "perform better than" the "baseline model" $(y = \bar{x})$, resulting in lower R^2 scores, and hence low F-tests.

Problem 3

Context: A survey among a school's track stars found the height (in meters) and average running speed (in meters per second) of a student runner. They produced the following results:

Height	Speed
1.67	3.7
1.7	3.8
1.5	3.5
1.65	3.8
1.6	3.6
1.75	3.9

Part A Problem: Find the correlation between height and running speed, and comment briefly on the strength of the correlation.

Part A Solution: The formula for Pearson's correlation coefficient, when simplified, is

$$\rho = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

The mean of the height and speed variables can be found as follows:

$$\bar{x} = \frac{1.67 + 1.7 + 1.5 + 1.65 + 1.6 + 1.75}{6} = 1.645$$

$$\bar{y} = \frac{3.7 + 3.8 + 3.5 + 3.8 + 3.6 + 3.9}{6} = 3.71\bar{6} \approx 3.717$$

We can thus find the numerator of ρ as follows:

$$(1.67 - 1.645)(3.7 - 3.717) + (1.7 - 1.645)(3.8 - 3.717) + (1.5 - 1.645)(3.5 - 3.717) + (1.65 - 1.645)(3.8 - 3.717)$$
$$+(1.6 - 1.645)(3.6 - 3.717) + (1.75 - 1.645)(3.9 - 3.717) = 0.0605$$

We can find the components of the denominator as follows:

$$\begin{aligned}\sqrt{\sum(x_i - \bar{x})} &= \sqrt{(1.67 - 1.645)^2 + (1.7 - 1.645)^2 + (1.5 - 1.645)^2 + (1.65 - 1.645)^2 + (1.6 - 1.645)^2 + (1.75 - 1.645)^2} \\ &= \sqrt{0.03775} \\ &\approx 0.1943\end{aligned}$$

$$\begin{aligned}\sqrt{\sum(y_i - \bar{y})} &= \sqrt{(3.7 - 3.717)^2 + (3.8 - 3.717)^2 + (3.5 - 3.717)^2 + (3.8 - 3.717)^2 + (3.6 - 3.717)^2 + (3.9 - 3.717)^2} \\ &= \sqrt{0.108334} \\ &\approx 0.3291\end{aligned}$$

$$\sqrt{\sum(x_i - \bar{x})}\sqrt{\sum(y_i - \bar{y})} \approx 0.1943 \times 0.3291 \approx 0.0639$$

Dividing the numerator by the denominator yields $\rho \approx 0.9468$.

Part B Problem: Based on this information, can we safely conclude greater height causes a person to be faster?

Part B Solution: Correlation does not make statements about causation. Therefore, we cannot safely conclude greater height causes a person to be faster.

Part C Problem: Based on this information, can we safely conclude running fast makes people taller?

Part C Solution: As stated in part B, correlation does not make statements about causation; therefore, we cannot safely conclude running fast makes people taller. From common sense, too, such an explanation would seem odd.

Part D Problem: Based on this information, what's the strongest conclusion we can reasonably make?

Part D Solution: Given a very high correlation, we can conclude that people who are tall tend to run faster. Such a statement does not make a causal claim, but simply asserts that there is a correlation between the height and speed of a person.

Problem 4

Problem: Is it possible to have three quantities, A, B, and C, so that A and B are correlated with $\rho > 0.5$, A and C are correlated with $\rho > 0.5$, but B and C are uncorrelated ($\rho \approx 0$)? If so, give an example. If not, explain why not.

Solution: We can devise a trick of sorts to construct the data sets A, B, and C. Since A should be correlated with B and C, but B should not be correlated with C, let us make A the simplest variable – it counts the natural numbers from 0 to $N - 1$ (inclusive): $A = [0, 1, ..., N - 1]$ such that it contains N elements. This makes it simpler to conceptually find correlated variables; a variable correlated with A should have the trend of increasing.

To construct B and C, we can begin with the objective of making their correlation as close to 0 as possible. It is well-known that points that form a square have a correlation of 0 – for instance, the points $(0, 0), (1, 0), (0, 1), (1, 1)$. As such, we could construct B and C as x and y axes like this:

- $B_1 = [0, 0, 0, 1, 1, 1, 2, 2, 2]$
- $C_1 = [0, 1, 2, 0, 1, 2, 0, 1, 2]$

This forms a square of side length $\sqrt{N}$ and has a Pearson correlation coefficient of 0. While B_1's correlation with A is very high, C_1's is not, since it exhibits a repeating behavior that does not align with the linearly increasing nature of A. What we would like to do is to somehow shuffle the order of B and C while keeping the order of A constant such that B and C still form a square when plotted out, but B and C both have the trend of increasing when mapped out against A. This entails first "zipping" B and C together into coordinate pairs, rearranging the coordinate pairs, and "unzipping" the coordinate pairs back into B and C.

To address this, let us begin by recognizing that there is no meaningful reason why B_1 should be correlated with A_1 and C_1 shouldn't; the listing of the points could be switched between B_1 and C_1 as follows:

- $B_2 = [0, 1, 2, 0, 1, 2, 0, 1, 2]$
- $C_2 = [0, 0, 0, 1, 1, 1, 2, 2, 2]$

Here, C_2 is strongly associated with A and B_2 is not. If we can merge the relationship between B_1 and C_1 with the relationship between B_2 and C_2, ideally the strong association between B_1 and A but not between C_1 and A in the first ordering and between C_2 and A but not between B_2 and A in the second ordering will get "distributed" across both variables B_3 and C_3.

To "merge" the two, we will take alternating indexes from the B_1-$C-1$ set and the B_2-C_2 set:

- $B_3 = [0, 1, 0, 0, 1, 2, 2, 1, 2]$
- $C_3 = [0, 0, 2, 1, 1, 1, 0, 2, 2]$

Now, in each set, smaller numbers appear more frequently earlier and larger numbers appear more frequently later. The correlation between B_3 and A is 0.7379, the correlation between C_3 and A is 0.5270, and the correlation between B_3 and C_3 is 0. Note that in order for correlation between B and C to be 0, N must be the square of an odd integer, like 3 or 5. When N is the square of an even integer, the merging algorithm excludes some points. Even in these cases, though, the correlation between B and C remains very small.

Interestingly, as N increases in size, the correlations between B_3 and A and between C_3 and A decrease towards 0.5, but never seem to go below 0.5. For context, see the following results:

Dimension ($\sqrt{N}$)	Correlation between A and B	Correlation between A and C
5	0.635416	0.541281
11	0.554438	0.531991
51	0.510271	0.509140
101	0.505071	0.504780
501	0.501002	0.500991
1001	0.500501	0.500498

This interesting phenomenon is perfect for this problem, since the requirement listed is that $\rho > 0.5$ for correlation between A and B and between A and C.

Sample Python code:

```
import numpy as np

dim = 3 # ideally, some odd number; dim**2 = N
A = range(dim**2)
B1, C1 = [], []
for i in range(dim):
    for j in range(dim):
        B1.append(i)
        C2.append(j)
zip1 = np.array([i for i in zip(B1,C1)])
zip2 = np.array([i for i in zip(C1,B2)])
new_zip = []
for i in range(dim**2):
    if i%2==0: new_zip.append(zip1[i])
    else: new_zip.append(zip2[i])
```

```
B = np.array(new_zip)[:,0]
C = np.array(new_zip)[:,1]

print(np.corrcoef(A,B)[0][1])
print(np.corrcoef(A,C)[0][1])
print(np.corrcoef(B,C)[0][1])
```

Thus, the scenario as outlined in the problem is possible.

Problem 5

Question: In the Homework Data spreadsheet, the table labeled "Problem 5" shows the deaths in early 1918 (before the beginning of the 1918 Spanish Influenza pandemic) and in late 1918 in each state that was affected by the disease. Compute the correlation of the two quantities, and give a reasonable explanation of any correlation you see.

Solution: The simplified formula for correlation is $\rho = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$.

We find the mean of the Early 1918 column to be 3,147.4583 and the mean of the Late 1918 column to be 12,162.25. Using these means, we can calculate the numerator to be 1,113,208,286. To calculate the denominator, we find $\sum(x_i-\bar{x})^2$ to be 295,571,382 and $\sum(y_i-\bar{y})^2$ to be 4,464,526,167, where x represents Early 1918 and y represents Late 1918. Multiplying these two sums and square-rooting yields a denominator of 1,148,732,418. Dividing the numerator by the denominator yields a correlation of 0.9691.

The dataset consists of deaths in Early 1918 and Late 1918. The correlation between the deaths in Early 1918 and the deaths in Late 1918 are very high, indicating that there is a strong linear relationship between the two – when the number of deaths for some state A in Early 1918 is low, the number of deaths in that same state A in late 1918 is also relatively low. One possible explanation for this high correlation is that states with larger populations have a higher number of deaths both in Early 1918 and in Late 1918, since one would expect viruses to spread similarly as a scientific phenomenon in all states. In this explanation, each state's increase in deaths is a population-scaled version of an "expected" pattern of increase.

Problem 6

Question: Using the link in the assignment description, find the US COVID-19 data. Using this data, and all appropriate statistical techniques at our disposal, can you conclude (with 90% confidence or better, if that is meaningful for the techniques you use) whether the rate of spread of COVID-19 in Washington state is decreasing over time?

Solution: After exporting the data for Washington, we find that there are 450 rows of data from 1/22/2020 to 4/15/2021. New cases is a good measure of the rate of spread of COVID-19 in Washington (although, the interpretation of what exactly "rate of spread" constitutes seems to be varied). We can think of rate of spread by looking at the total number of cases in Washington and analyzing the change in how fast it increases. This is equivalent to the number of new cases We can visualize the number of new cases over time to get a feel for the trends in Figures 1 and 2.

It seems that the number of new cases in Washington peaks roughly around the beginning of 2021. Although the problem does not seem to specify what "over time" constitutes, we will take a somewhat liberal interpretation that it means "even though the rate of spread increased at some parts, does it eventually start decreasing?" Let us restrict the focus of our analysis, then, to all dates on or after 1/1/2021.

The simplified formula for correlation is $\rho = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$. We find the mean of the dates to be 52 (after converting to number of days since 1/1/2021) and the mean of the number of new deaths to be 1285.4571. Using these means, we can calculate the numerator to be $-1,453,679$. To calculate the denominator, we find $\sum(x_i-\bar{x})^2$ to be $96,460$ and $\sum(y_i-\bar{y})^2$ to be $1,218,780,101.1$, where x represents the number of days since 1/1/2021 and y represents the number of new cases on that day. Multiplying these two sums and square-rooting yields a denominator of $3,428,753.833$. Dividing the numerator by the denominator yields a correlation of -0.4239. This suggests that the rate of spread is decreasing over time. However, the data – as demonstrated visually when plotted out – is very varied and seems to have a lot of noise, likely from difficulties in attaining accurate testing data and piling over. Interestingly, when a smoothing transformation with a window of just 5 days is applied to the number of new cases, the Pearson correlation coefficient increases (in absolute value) to -0.7174.

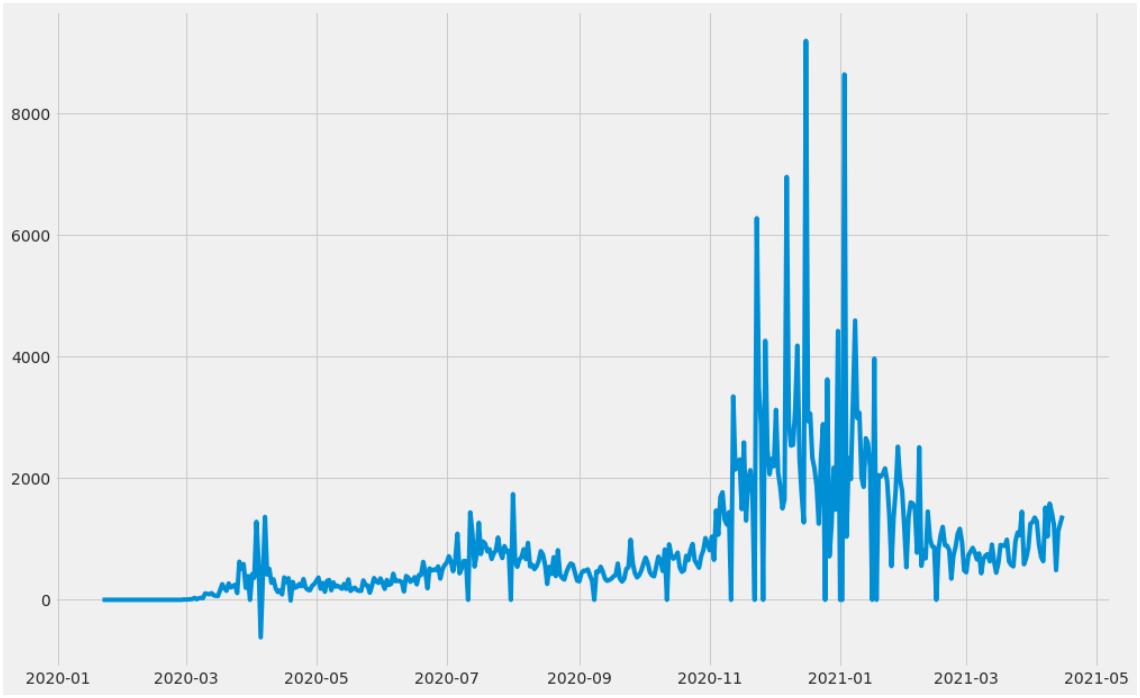

Figure 1: Number of new cases over time in Washington.

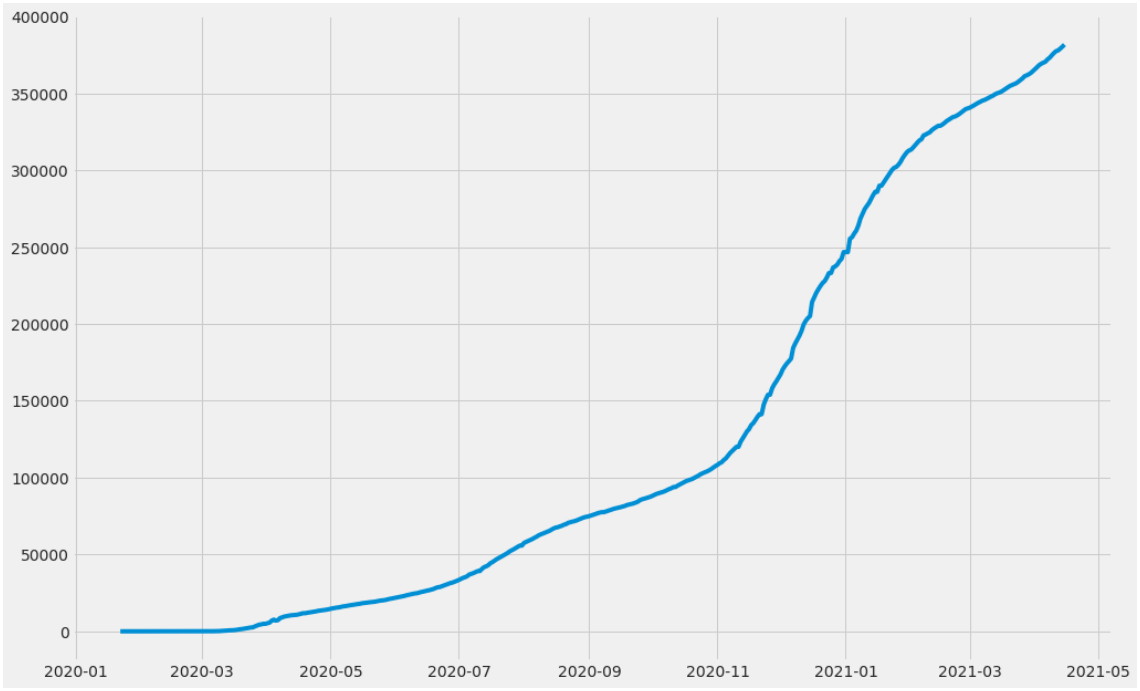

Figure 2: Total number of cases over time in Washington.