Homework 2

Sahana Sarangi

8 April 2024

Problem 1: In (a)-(c), a situation is described in which a researcher draws a conclusion based on statistical evidence.

(a) Jay is studying number theory, and they're fascinated by the prime numbers. A (positive whole) number is called prime if it is divisible only by 1 and itself; for technical reasons, 1 itself is not classified as prime. It is a well-known fact that the prime numbers are hard to "predict" - the sequence of prime numbers looks pretty random. Jay, however, heard that word "random" and realized they could try using their statistics knowledge on the problem! So Jay assembled a database of the first 10000 prime numbers, and the distance from each one to the next one. Based on that, they calculated that the average distance from one prime to the next in their sample was 10.48, with a standard deviation of 8.07. They therefore conclude that they can be 95% confident that the average distance between primes is between 10.32 and 10.64. Was this a reasonable conclusion? Why or why not?

(b) Dr. Lalonde is a confectionary scientist, studying chocolate consumption. She surveys a million people in the United States, collecting data on their chocolate-eating habits and their household income. She discovers that the Pearson correlation coefficient between the average reported volume of chocolate consumed per week by a household and the households total annual income is 0.862, which is a strong correlation. Dr. Lalonde then writes a book titled Chocolate: The Path to Wealth in which she claims that eating more chocolate will make you more energetic at work, and therefore you will get more raises and earn more in the long run. Does Dr. Lalonde's research support her conclusion? If so, explain your reasoning. If not, give an alternative hypothesis that explains the correlation Dr. Lalonde noticed.

(c) Dr. Strider is studying a worrying trend. In 2017, there were 52 Mondays. In 2018, there were 53. He constructs a line of best fit for this data, where t is the year and M is the number of Mondays:

$$M = t - 1965$$

This line has a root-mean-squared error (RMSE) of zero - it's a perfect match! Dr. Strider therefore concludes that the number of Mondays per year is on the rise. By the year 2330, the year will be entirely Mondays; the following year, there will be more Mondays than days, which suggests the existence of the fabled "double Monday." Are Dr. Strider's concerns supported by his statistical analysis? What flaws are present in his reasoning, if any?

Part (a) Solution: Jay's conclusion is unreasonable. When calculating confidence intervals, we make the assumption that our sample set is not "unusual," or does not fall in the extreme 5% of the total sorted dataset. In this case, the first 10000 prime numbers are part of the *most* "unusual" numbers of the total dataset. The amount of prime numbers is infinite, so taking the first 10000 of them would definitely be selecting the 5%, or the most unusual prime numbers. Hence it would be unreasonable to calculate a confidence interval using a sample set of the first 10000 primes, as the sample set does not satisfy the assumption that we make in order to calculate confidence intervals.

Part (b) Solution: No, Lalonde's research does not support her conclusion. When we calculate the correlation of two things, we are not determining any causation, as correlation does not equal causation.

Lalonde's calculations could be correct, but it does not mean that eating more chocolate necessarily *causes* people to earn more. Based on just the survey, she cannot conclude that eating chocolate causes people to have more energy or that more energy at work causes higher household incomes. An alternative hypothesis that explain's Lalonde's correlation is that people who have a higher household income would have more money to spend on less essential items such as chocolate, and hence people with higher household incomes are more likely to eat more chocolate.

Part (c) Solution: Strider's conclusion is unreasonable. The issue is that he is trying to construct a line of best fit using only two data points. When constructing a line of best fit using only two data points, the line will always go through those two points (as you only need two points to construct a line) and the resulting line will also be a perfect match. His line of best fit isn't really the line of best fit for all the data, it's only the line of best fit for the two data points he's taken. That's why he has come up with an extreme conclusion like the year being full of Mondays in 2330.

Problem 2: The table below shows the number of people fully vaccinated in Washington State and the number of reported new cases of COVID-19 in Washington State on the first day of each of several months in 2021.

Date	Fully Vaccinated	New Cases Reported
2/1/2021	126,672	1934
3/1/2021	621,171	1079
4/1/2021	1,403,264	1142
5/1/2021	2,493,486	1574
6/1/2021	3,553,192	1418
7/1/2021	4,163,605	457
8/1/2021	4,394,955	0

Based on this data, compute the correlation between the number of fully vaccinated people in Washington State and the number of reported new cases of COVID-19 in Washington State. What conclusion can you reasonably draw based on this? Why should we be cautious about this conclusion?

Solution: The formula to calculate the correlation between two datasets is

$$\frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum(x_i - \mu_x)^2}\sqrt{\sum(y_i - \mu_y)^2}}$$

where x_i is the ith element of the first dataset, μ_x is the mean of the first dataset, y_i is the ith element of the second dataset, and μ_y is the mean of the second dataset. We can take the numbers of fully vaccinated people to be our first dataset and the new cases reported to be our second dataset. Calculating the means of both the numbers of fully vaccinated people and the new cases reported, then plugging those means as well as the numbers in each dataset into the formula for correlation, we get -0.7031.

Having a correlation coefficient of -0.7031 means that there is a fairly strong negative correlation between the number of people that are fully vaccinated and the new cases reported. In other words, as the number of people that are vaccinated increases, the number of new cases reported is decreasing. It would seem as though the vaccine is effective in preventing COVID-19, as the number of new cases is going down. However, we have to be cautious about this conclusion because correlation does not imply causation. There could be other factors that are causing new case numbers to decrease in the summer months like new stricter quarantine regulations even though the correlation suggests that the vaccine is effective.

Problem 3: Is it possible to have three data sets, A, B, and C, so that all of the following conditions hold?

(i) A and B are correlated with correlation greater than 0.5.

(ii) B and C are correlated with correlation greater than 0.5.

(iii) A and C are uncorrelated, with correlation exactly 0.

Solution: Yes, it is possible to have these datasets. Dataset A would be $[2, 6, 4]$, dataset B would be $[2, 4, 4.5]$ and dataset C would be $[1, 1, 4]$. Using the formula for correlation in problem 2, the correlation between A and B is 0.7559, the correlation between B and C is 0.6547, and the correlation between A and C is 0.

Problem 4: The table below shows the number of new cases of COVID-19 reported in the United States each day of the first week of March 2022.

Date	New Cases
March 1, 2022	43,583
March 2, 2022	58,146
March 3, 2022	51,737
March 4, 2022	49,587
March 5, 2022	16,615
March 6, 2022	6,775
March 7, 2022	63,052

Here are three models for this data, where N is the number of cases and t is the number of days into March:

$$N = -2837.75t + 52707.42857$$

$$N = 4100(t - 5.5)^2 + 16000$$

$$N = \frac{130000}{t^{1.2}}$$

Which model is the best? Defend your answer.

Solution: To find which model is the best, we can calculate the RMSE for each one. The formula to find the RMSE of a model is

$$\sqrt{\frac{\sum(y_i - f(x_i))^2}{N}}$$

where y_i is the ith element of the new cases reported, x_i is the ith day into March, $f(x_i)$ is the predicted number of cases on the ith day of March, and N is the total number of values in the dataset containing the observed new cases. Finding all of these values for the linear model and plugging them into the formula, we get an RMSE of 18957.0119. Doing the same for the quadratic model, we get an RMSE of 27699.8136. Doing the same for the last model, we get an RMSE of 39645.7189.

Because RMSE calculates how "wrong" the outputs of a predicted model are compared to the actual data (or how "off" they are), the "best" model is the one with the lowest RMSE, as that is the model that is the most accurate. The linear model is the one with the smallest RMSE, so the $\boxed{\text{linear model is the best.}}$