Week 3 Exercises: Statistics for Data Science

Part 1: Binomial Distribution

According to Forbes (https://www.forbes.com/sites/zackfriedman/2019/01/11/live-paycheck-to-paycheck-government-shutdown/#4693b2194f10), 78% of American workers live paycheck-to-paycheck.

1. You take a random sample of 10 American workers.
   a. What is the probability that between 6 and 9 people (inclusive) in your sample are living paycheck-to-paycheck?

   b. What is the probability that at least 9 people in your sample are living paycheck-to-paycheck?

   c. What is the probability that fewer that 75% of people in your sample are living paycheck-to-paycheck?

   d. What is the probability that fewer than 50% of people in your sample are living paycheck-to-paycheck?
2. You take a random sample of 25 American workers.
   a. What is the probability that fewer than 75% of people in your sample are living paycheck-to-paycheck?

   b. What is the probability that fewer than 50% of people in your sample are living paycheck-to-paycheck?
3. You take a random sample of 100 American workers.
   a. What is the probability that fewer than 75% of people in your sample are living paycheck-to-paycheck?

   b. What is the probability that fewer than 50% of people in your sample are living paycheck-to-paycheck?
4. You take a random sample of 1000 American workers.
   a. What is the probability that fewer than 75% of people in your sample are living paycheck-to-paycheck?

   b. What is the probability that fewer than 50% of people in your sample are living paycheck-to-paycheck?
5. What do you notice happens as the size of your sample increases?

Part 2: The Newton Problem

Now, you will attempt to outdo Isaac Newton. See this video by Vsauce 2 for an entertaining description and analysis of the problem: https://www.youtube.com/watch?v=RFlTawWwLZc. As described in the video, this is a problem which Isaac Newton got wrong. More precisely, he got the correct calculations, but his explanation was off. To be fair to Newton, the machinery of probability theory had not been developed yet in his time.

Use the binomial distribution to answer this question. Which of the following three propositions has the greatest chance of success?
a. Six fair dice are tossed independently and at least one “6” appears.
b. Twelve fair dice are tossed independently and at least two “6”s appear.
c. Eighteen fair dice are tossed independently and at least three “6”s appear.

Part 3: Binomial Distribution - Conceptual

1. Using what you know about the variance of a binomial random variable, which situation do you think would require a larger sample size to accurately estimate? Why?
   a. The proportion of voters who will vote for candidate A in an upcoming election, where you know that it is a close race between candidate A and candidate B.

   b. The proportion of people who will click on an ad on your company’s website, when you know that historically, the proportion of people clicking on ads is is very low ($< 2\%$).
2. The binomial distribution gives the probability of a particular number of successes. If instead of counting the total number of successes, you were looking at the *proportion* of successes (total number of successes / number of trials), what happens to the variance as the number of trials increases (keeping the probability of success fixed)?

Part 4: The Normal Distribution

1. Using a standard normal distribution (mean 0, standard deviation 1), answer the following questions.
   a. What percentage of outcomes are within one standard deviation of the mean?

   b. What percentage of outcomes are within two standard deviations of the mean?

   c. What percentage of outcomes are within three standard deviations of the mean?
2. How do your answers change to the previous question if you are using a normal distribution that has a different mean and/or standard deviation?
3. For this question, assume that the heights of men in the US are normally distributed with a mean of 70 inches and standard deviation of 3 inches.
   a. If a single man is chosen at random, what is the probability that his height is less than 66 inches tall?

   b. If a single man in chosen at random, what is the probability that his height is greater than 72 inches tall?

c. If a single man is chosen at random, what is the probability that his height is between 68 and 72 inches?

d. If two men are chosen at random, what is the probability that both of them are between 68 and 72 inches tall? (Hint: You'll need to use your answer from part a. plus the binomial distribution to answer this.)

e. If twenty-five men are chosen at random, what is the probability that all of them are between 68 and 72 inches tall?

f. Challenge Question - If two men are chosen at random, what is the probability that their **mean** height is between 68 and 72 inches tall? (Hint: You'll probably have to simulate this to get an approximate answer. Similar to what we did to generate binomial random variables, we can generate normal random variables using `norm.rvs()`. For example, `norm.rvs(loc = 70, scale = 3, size = 2)` will generate two observations from a normal random variable with mean 70 and standard deviation 3.)

g. Challenge Question, Part 2 - If twenty-five men are chosen at random, what is the probability that their **mean** height is between 68 and 72 inches tall?