Challenge: Evaluate an experiment analysis

Now it's time to flex your critical evaluation skills. Read the following descriptions of an experiment and its analysis, identify the flaws in each, and describe what you would do to correct them.

1. **The Sith Lords are concerned that their recruiting slogan, "Give In to Your Anger," isn't very effective. Darth Vader develops an alternative slogan, "Together We Can Rule the Galaxy." They compare the slogans on two groups of 50 captured droids each. In one group, Emperor Palpatine delivers the "Anger" slogan. In the other, Dark Vader presents the "Together" slogan. 20 droids convert to the Dark Side after hearing Palpatine's slogan, while only 5 droids convert after hearing Vader's. The Sith's data scientist concludes that "Anger" is a more effective slogan and should continue to be used.**

   The key flaw in this study is that other variables are introduced between the control and the treatment: the person delivering the slogan. The outcome metrics can't be relied upon because it's not clear if the presenter had an affect on whether or not the droids converted to the Dark Side.

   For future study's looking to assess recruiting messaging, the message should be delivered by the same person.

2. **In the past, the Jedi have had difficulty with public relations. They send two envoys, Jar Jar Binks and Mace Windu, to four friendly and four unfriendly planets respectively, with the goal of promoting favorable feelings toward the Jedi. Upon their return, the envoys learn that Jar Jar was much more effective than Windu: Over 75% of the people surveyed said their attitudes had become more favorable after speaking with Jar Jar, while only 65% said their attitudes had become more favorable after speaking with Windu. This makes Windu angry, because he is sure that he had a better success rate than Jar Jar on every planet. The Jedi choose Jar Jar to be their representative in the future.**

   The key flaw in this study is that the sample populations are completely different. It's not clear whether favorable attitudes were the result of the messenger or the result of the planets' relations with the Jedi's.

   For future work, it would make more sense to have Jar Jar Binks and Mace Windu go two friendly and non friendly planets.

3. **A company with work sites in five different countries has sent you data on employee satisfaction rates for workers in Human Resources and workers in Information Technology. Most HR workers are concentrated in three of the countries, while IT workers are equally distributed across worksites. The company requests a report on satisfaction for each job type. You calculate average job satisfaction for HR and for IT and present the report.**

   The key flaw in this study is that by aggregating job satisfaction for both HR and IT, you lose all the potential nuance in satisfaction that may or may not exist based on country. I would not be clear as to whether satisfaction differences were related to the profession being measured or the country that they're working in.

   For future work, it may make more sense to aggregate satisfaction for HR workers in the three countries they're most concentrated in and then aggregate satisfaction for IT workers in those same 3 three countries. In this way, you help significantly reduce the contextual bias that may have skewed your initial report.

4. **When people install the Happy Days Fitness Tracker app, they are asked to "opt in" to a data collection scheme where their level of physical activity data is automatically sent to the company for product research purposes. During your interview with the company, they tell you that the app is very effective because after installing the app, the data show that people's activity levels rise steadily.**

   The key flaw of this study is that the population is biased. Those that opt-in to data collection may be more inclined to stay physically active knowing that their activity levels will be measured. It's not clear whether the app is truly effective or if the study simply elicited the participation of those who tend to be more physically active.

   For future work, it may make more sense to require anonymous data collection to reduce bias in the sample. This would help give better indication as to whether the app is truly affective.

5. **To prevent cheating, a teacher writes three versions of a test. She stacks the three versions together, first all copies of Version A, then all copies of Version B, then all copies of Version C. As students arrive for the exam, each student takes a test. When grading the test, the teacher finds that students who took Version B scored higher than students who took either Version A or Version C. She concludes from this that Version B is easier, and discards it.**

   The key flaw in this study is that the population is not randomized. Because the tests were stacked in order from with A tests on top and C tests on bottom, it's not clear whether Version B was truly superior to tests A and C or if the population taking test B were simply more competent at taking the exam.

   To ensure that the population is truly randomized, the teacher should shuffle tests A, B and C.