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, Darth 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 standout culprit in this example is observer bias. Having Palpatine deliver his speech and Vader deliver another clouds the effect. Is Palpatine just a better speaker than Vader? They should choose one person to use both slogans.

1. 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.

This sounds a bit like Simpson’s paradox. Averaging different sized populations can lead to averages that don’t reflect the sub-trends. It is possible that Windu had a higher conversion rate but spoke to more people on unfriendly planets (lower conversion) than Jar Jar. For example:

|  |  |  |
| --- | --- | --- |
|  | Jar Jar [Talked to] (Converted) | Mace Windu [Talked to] (Converted) |
| Planet 1 | [100] (20%) | [500] (35%) |
| Planet 2 | [100] (20%) | [500] (35%) |
| Planet 3 | [100] (20%) | [500] (35%) |
| Planet 4 | [100] (20%) | [500] (35%) |
| Planet 5 | [500] (80%) | [100] (90%) |
| Planet 6 | [500] (80%) | [100] (90%) |
| Planet 7 | [500] (80%) | [100] (90%) |
| Planet 8 | [500] (80%) | [100] (90%) |
| Total | 1680/2400 = 70% | 1060/2400 = 44.1% |

Clearly my numbers are a bit off but this is just to show what could have happened. Windu talking with more harder to convert people skewed his overall average down, whereas Jar Jar benefited from talking with more easier to convert people.

1. 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.

33% of HR workers work in each of the 3 countries they’re in while only 20% of the IT workers are in those same countries. That means that any positive or negative effects due to that country will be amplified in the HR worker satisfaction result.

I would suggest breaking down the results by country so as to avoid any Simpson’s paradox effects manifesting in the overall average.

1. 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.

Perhaps the only people opting in are people who are already motivated to work out or are of a certain demographic. To know if the app is effective for the whole population who downloads it we would need data on that whole population, not just the subset that opts in.

1. 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.

Are all of the questions the same? If not then that would be something to fix. Version B may have had on average higher achieving students taking the test than the other versions, or perhaps a row of friends all sitting together cheated and got higher scores because they were all taking the same test.