Stats Homework 5

Hanna Butsko

2023-10-07

Problem 1.1

a) Explanatory variable: marijuana sales in Colorado Response variable: traffic fatalities

This an observational study as it analyzes existing data rather than conducting a new experiment.

A lurking variable is any factor that influences both marijuana sales and traffic fatalities without being directly included in the analysis. Considering that Colorado was among the first states to legalize marijuana, it experienced a potential surge in tourism as people flocked to the state to partake in legal marijuana consumption and other related experiences. This tourism boost would inherently increase marijuana sales. Concurrently, the influx of tourists, who might be unfamiliar with Colorado's traffic regulations, local roads, or might be partying and potentially using other substances, could contribute to an uptick in traffic fatalities. Thus, the increase in tourism, driven by marijuana legalization, might act as a lurking variable when examining the relationship between marijuana sales and traffic fatalities.

b) Hypothetical experiment would require:

1. Randomly selecting the participants.
2. Randomly dividing the participants into two groups: experimental group(this group will consume marijuana) and control croup(this group will be administered a placebo).
3. Recording incidents of traffic fatalities while under the influence of marijuana.
4. Comparing the results between the two groups to determine if there is a difference in traffic fatality rates between those under the influence of marijuana and those not.

Conducting an experimental study to explore this link raises significant ethical concerns, as it involves potentially endangering participants' lives. Such study requires assigning groups of individuals to conditions where there's a potential for serious harm or death just to determine this association, making it an unethical and unrealistic approach.

Problem 1.2

The first video provides a clear instance of sampling bias, specifically convenience sampling. The platform developer selected colleagues and individuals with tech degrees as testers, expecting constructive feedback from those who could grasp the intricacies of his work. However, he overlooked a crucial aspect: the software is intended for the general public, who may not have the same level of understanding as software engineers. As a result, participants fount the interface challenging, leading to their frustration. Simultaneously, the developer became frustrated with them, as his expectations for their understanding and feedback were misaligned with their actual experience.

Another evident bias in this study is expectation bias. The developer anticipated that the selected sample group would resonate with his product in a manner similar to his colleagues or others with a tech background. This expectation might have shaped how he introduced the project and his receptiveness to feedback. Having previously received positive evaluations, he was confident in his product's perfection. Thus, when confronted with the participants' frustration, rather than recognizing potential flaws in his product, he became disgruntled with the sample group, mistakenly attributing their difficulties to their perceived incompetence.

The moment the CEO entered the room, the blinding requirement of the second video was compromised. This compromise highlights the issue of expectation bias: the CEO's own expectations regarding the study's outcomes could have inadvertently influenced the participants' responses. Now, any reactions from the participants cannot be solely attributed to their genuine experiences with the platform; they are potentially influenced by the CEO's comments and merely his presence in the room. The video doesn't show the full conversation between the CEO and the sample group, so it is hard to make definitive conclusions, however, there's a likelihood that the CEO's personal opinion influenced both his presentation of the platform and the way he framed his questions to the participants.

Problem 2.1

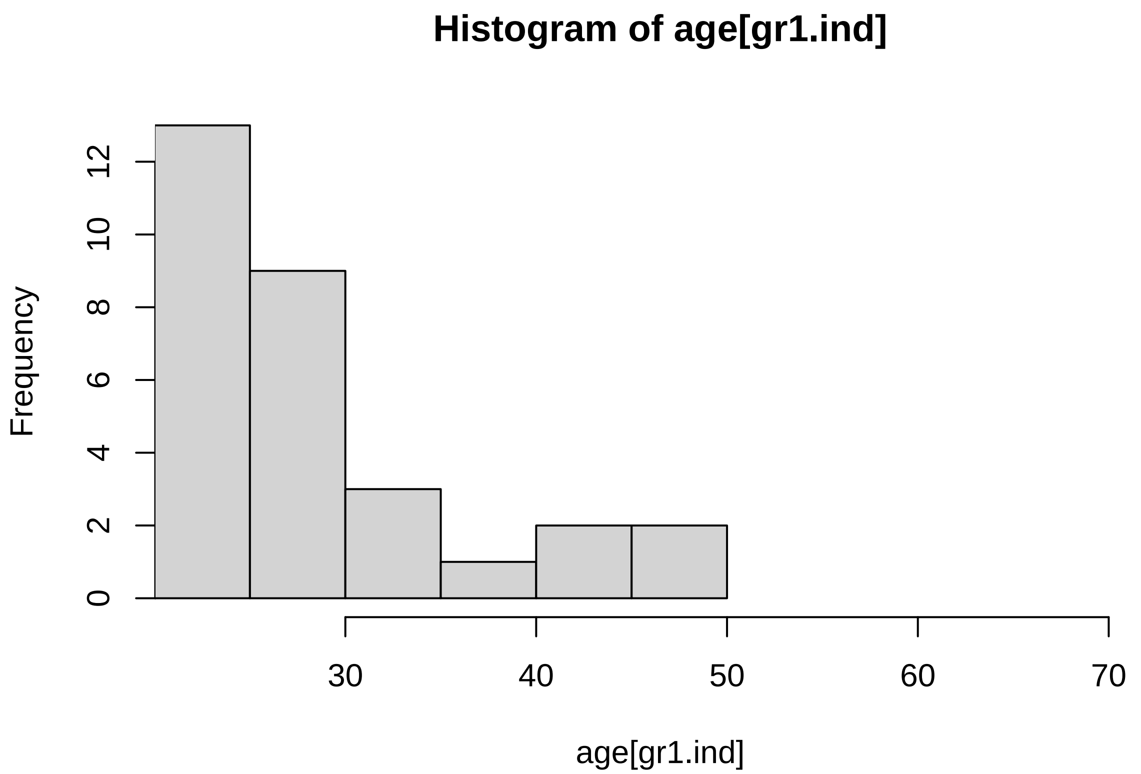
To eliminate bias, the teacher should employ random assignment, ensuring that each student has an equal chance of being placed in either group. Using the random assignment method is vital as it helps distribute both known and unknown lurking variables equally among the groups. Therefore, when comparing results, any observed differences can be attributed to the teaching methods, rather than to specific characteristics of the students or other confounding factors.

Loading the student survey data

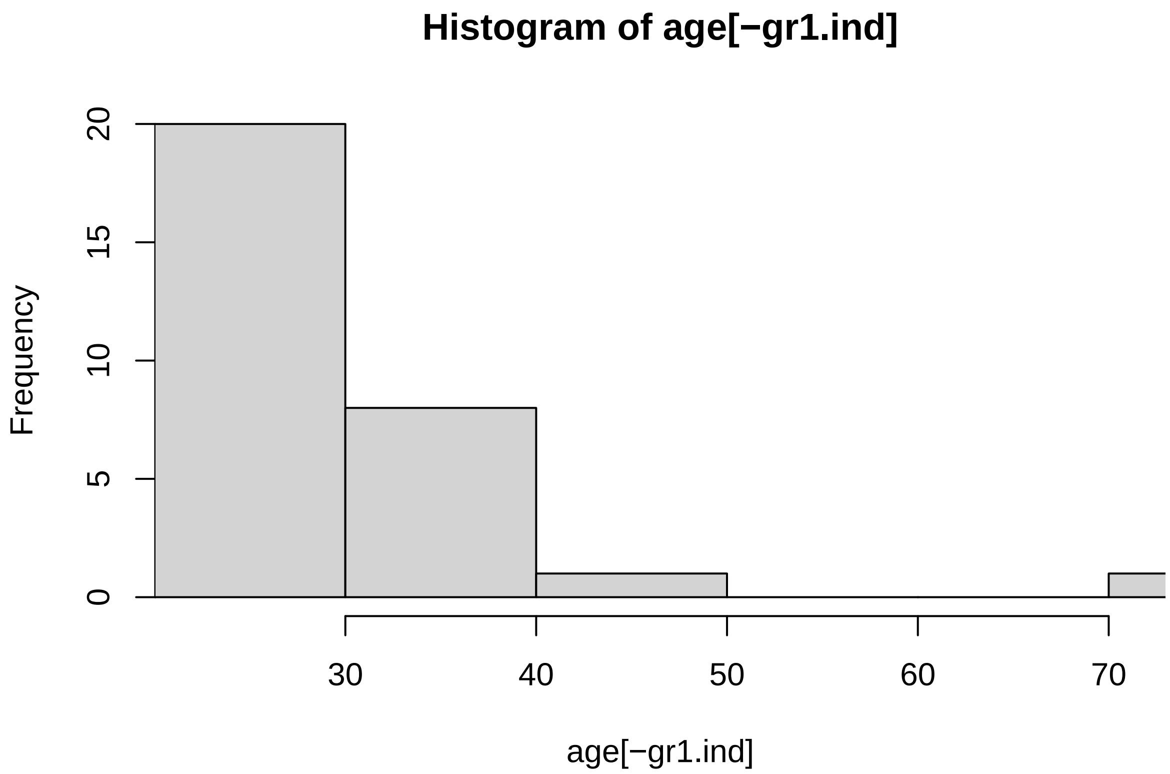
```r
fl_data <- read.csv("http://sites.williams.edu/bklingen/files/2015/07/fl_student_survey.csv")

# Attaching the dataset to make column names accessible without referencing the dataframe directly
attach(fl_data)

# Finding the total number of students (rows) in the dataset
n <- nrow(fl_data)

# Randomly selecting half of the students for the first group
gr1.ind <- sample(1:n, n/2)
```

Histograms for the 'age' variable for both groups

```r
# Group 1 (randomly selected students)
hist(age[gr1.ind], xlim=range(age))
```

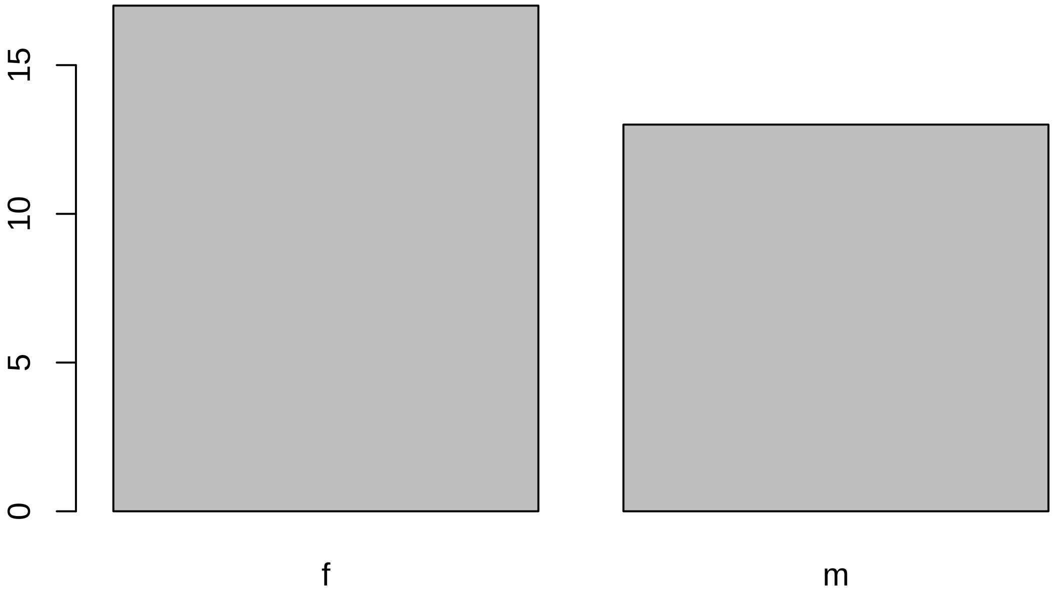


```r
# Group 2 (the remaining students)
hist(age[-gr1.ind], xlim=range(age))
```

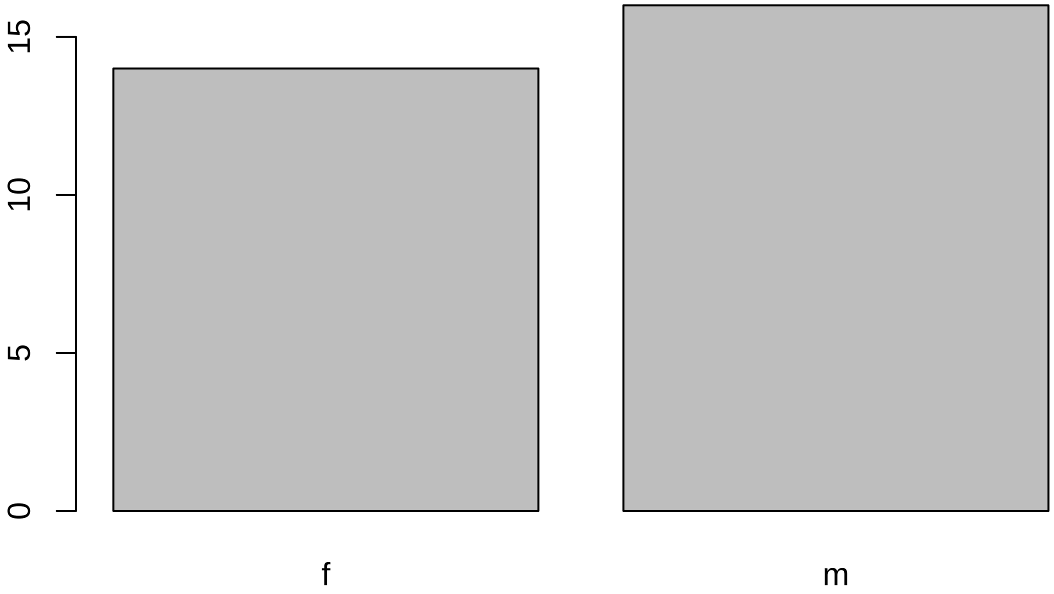


Barplots to visualize the distribution of the 'gender' variable for both groups

```
# Group 1
barplot(table(gender[gr1.ind]))
```

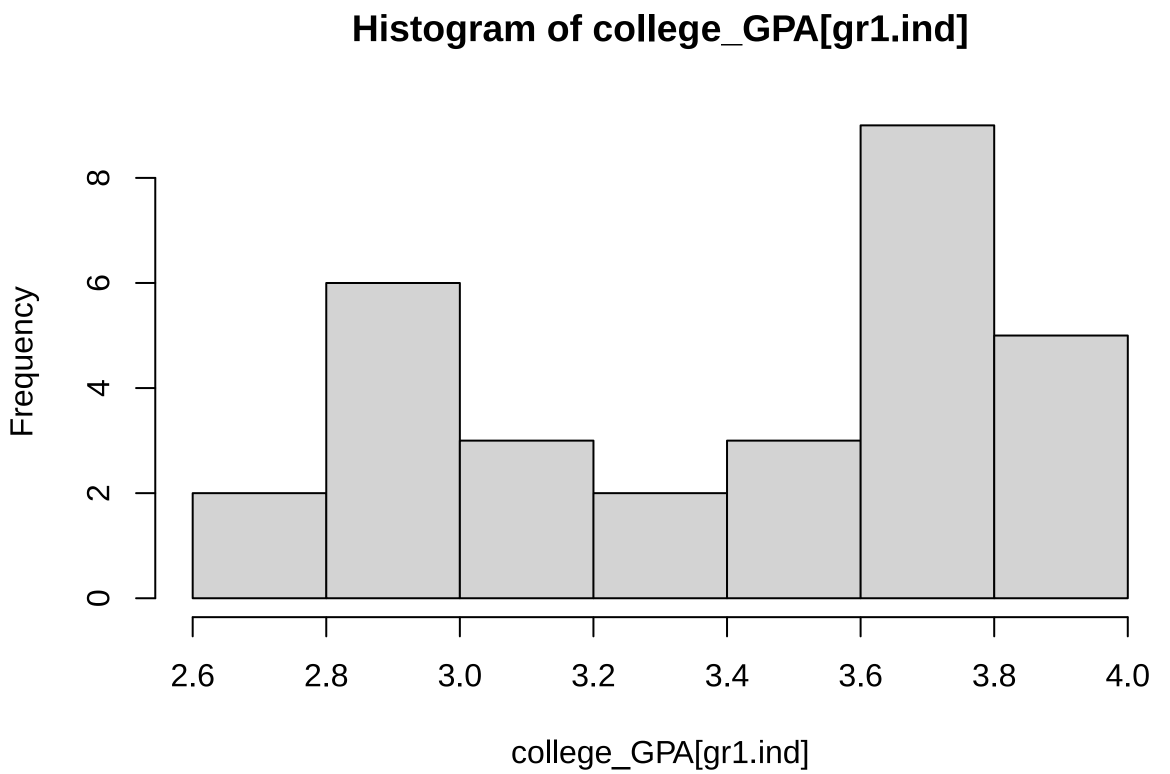


```
# Group 2
barplot(table(gender[-gr1.ind]))
```

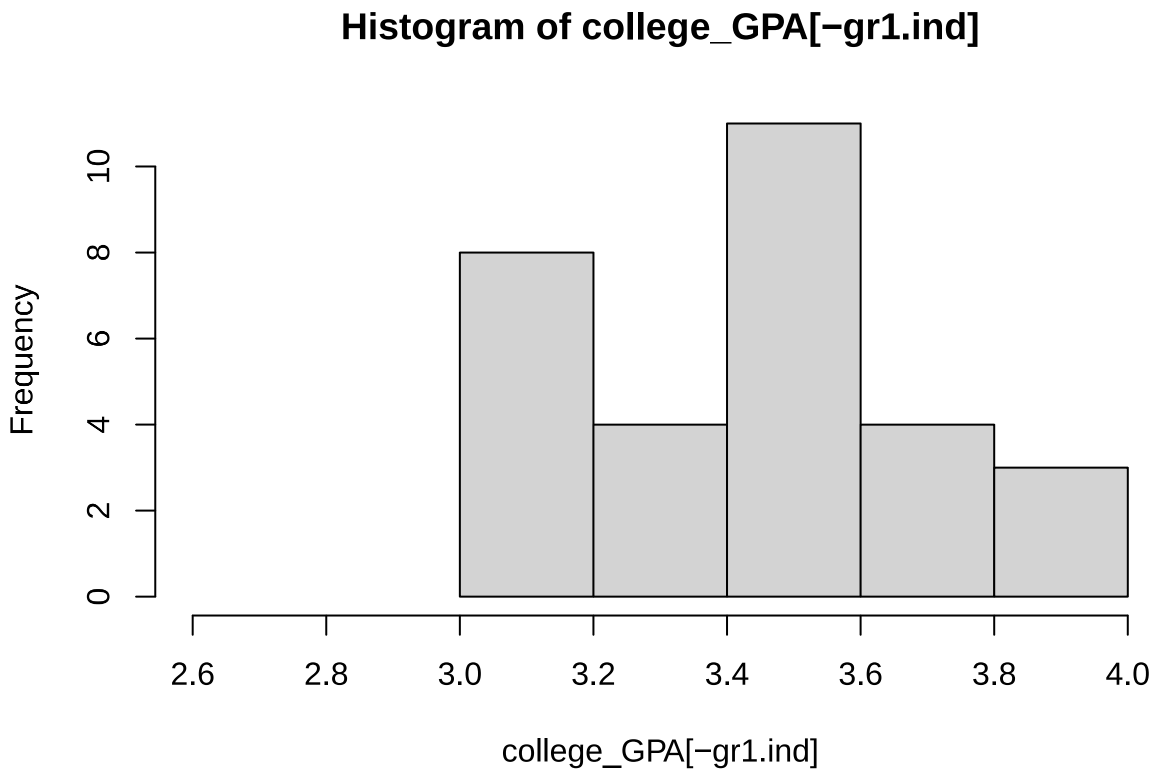


Histograms for the 'college_GPA' variable for both groups

```
# Group 1
hist(college_GPA[gr1.ind], xlim=range(college_GPA))
```



```
# Group 2
hist(college_GPA[-gr1.ind], xlim=range(college_GPA))
```



Problem 2.2

To maintain consistency while exposing students to different teaching methods, it's essential that both groups receive equal time and resources. For example, if one group is taught through video lectures and the other via in-person lectures, the sessions should be of identical length and should cover the same topics. As the experiment progresses, the teacher should periodically gather feedback from students about their learning experiences. The level of student engagement can be assessed by monitoring attendance, tracking participation in class discussions, and checking the completion rates of assignments. At the end of the experiment period, all students should be evaluated using a consistent assessment tool, such as a standardized test or a project assignment, ensuring the assessment includes all of the content taught during the experiment.

Problem 2.3

If the teacher selects a randomized blind sampling design and uses consistent evaluation tools at the conclusion of the experiment, then this study could potentially be used to generalize the measured effects of teaching methods onto a wider student population. Such a design minimizes biases and ensures that the differences noted are primarily attributable to the teaching methods, making the results more applicable to a larger population. However, 60 students might not be considered a large enough sample size in the context of educational experiments. The teacher should be cautious and may consider mentioning the sample size as a limitation when discussing the potential generalization of the study's results.

Problem 3

4.2 a) Since subjects were identified and followed for two decades without any experimental manipulation or intervention, I would classify this as an observational study rather than an experimental study. b) Explanatory Variable: High blood pressure in combination with binge drinking. Response Variable: Risk of death from stroke or a heart attack. c) No, it is not possible to definitively establish cause and effect in an observational study. There is always a possibility that some lurking variable could be responsible for the observed association. Therefore, while observational studies can suggest relationships and associations, they cannot prove causal relationships with absolute certainty.

4.3 a) Explanatory variable: Low-carb or low-fat diet. Response variable: Effect of the diet on weight loss. b) This study was experimental as the subjects were divided into groups and randomly assigned to a specific experimental condition, in this case, a low-carb or low-fat diet. Subsequently, outcomes were observed for the response variable, which is weight loss. c) It would not be appropriate to recommend that everyone should follow a low-carb diet over a low-fat diet solely based on the results of this one study. While the study suggests a stronger association between a low-carb diet and weight loss compared to a low-fat diet, it's important to acknowledge that various other variables could have influenced the study's results. These variables might include overall health, levels of physical activity, socioeconomic status, stress, age, and more. To make a recommendation that everyone who wishes to lose weight should prefer a low-carb diet, we would need to account for and carefully analyze all these potential influential variables.

4.9 a) An observational study is the more appropriate approach for investigating the effects of smoking on heart health. While experimental studies offer the advantage of controlling for confounding variables, leading to potentially more accurate results, conducting such an experiment in this context would be unethical. Asking participants to smoke for the sake of research could expose them to significant health risks, making the experimental approach not only impractical but also morally indefensible. b) Similar to the previous example, an observational study would be preferable approach when investigating whether higher SAT scores are positively correlated with higher college GPAs. While conducting an experimental study wouldn't directly harm participants, it's inappropriate and impractical to request that participants intentionally perform poorly on their exams to observe any potential association. c) To determine whether a special coupon attached to the catalog makes recipients order more products from a mail order company, an experimental study would be ideal. In this setup, two groups could be formed: one receiving catalogs with the special coupon and the other receiving catalogs without it. By tracking orders from both sets of recipients, the company can accurately measure the coupon's impact on purchasing behavior. This controlled environment eliminates many confounding variables and offers a clearer insight into the coupon's direct influence on customer orders.

4.22 a) The population for this survey might include all employers in the state of Michigan. It's crucial to differentiate between population and statistic in this context. The fact that the survey was answered by 6,500 employers represents a statistic. However, the population in question could include all employers in the state, or perhaps even the entire country. b) To calculate the nonresponse rate, we would need to know the total number of employers that were approached for this survey. We were not provided with this information, therefore, we cannot calculate the nonresponse rate. c) One source of bias in this survey is sampling bias. This is because the survey was directed towards employers who interacted with career services, yet it attempted to generalize hiring trends based on the degree to the entire population. Another source of bias we've already identified is nonresponse bias. While we don't have comprehensive information to determine the exact rate, it's reasonable to infer that some subjects either couldn't be reached or chose not to participate, given that the survey was based on a voluntary poll.

4.26 a) Sampling bias could arise from a fact that not all teenagers have an equal likelihood of responding to the survey. Since the survey was conducted online, it's unrealistic to determine whether teenagers who bought the alcohol online actually took part. b) Referring to the previous answer, we don't know how many teenagers skipped the survey and chose not to participate, leading to nonresponse bias. c) There's no way to verify if any respondents were untruthful in their answers, which would lead to response bias.

4.35 a) This study is an experiment due to its experimental setup. Two groups were formed: control and experimental. The subjects were observed over a period of time with two variables measured, and conclusions were drawn based on the results. b) The experimental units are the subjects - 689003 Facebook users. c) Explanatory variable: negatively manipulated facebook news feed. Response variables: the percentage of all words produced by a person that was positive and the percentage of all words produced by a person that was negative. d) The reason Facebook did not inform users that they were selected for the study was to maintain blinding. If users were aware of their participation in the study, they might intentionally or unintentionally alter their behavior. It is clear that not informing users that they are part of the study is not entirely ethical. Users did not give their consent to be a part of that study. Moreover, Facebook must have been tracking users' conversations to measure response variables, such as the amount of positive versus negative words produced by the person. This is clearly unethical to do without obtaining consent first.

4.40 a) An experiment would require: To randomly select the participants, I would ensure that everyone has an equal chance to be included in the study. This approach would reduce bias and help control for potential lurking variables. Participants will be randomly divided into two groups: the experimental group (which will receive vitamin C supplements) and the control group (which will be given a placebo). To achieve random assignment, I would use a random number generator in Python to evenly distribute the participants between the two groups. Making the study double-blind: I would first ensure that subjects are not informed about which treatment they are assigned to. Secondly, I would ensure that anyone who has contact with the subjects during the experiment is also unaware of the treatment information. b) The claim that people who take vitamin C get fewer colds might be misleading due to several factors. First, there could be lurking variables, such as the overall health and lifestyle of an individual, that could influence the results. Additionally, there's the potential for sampling bias and nonresponse bias. Sampling bias might arise if the participants aren't representative of the entire population. For example, if the study only sampled people from a particular age group or socio-economic status. In case of nonresponse bias, if a significant portion of people that were contacted chose not to participate, and if those who did not respond have different experiences with colds and vitamin C than the respondents, this could influence the results.