Collecting Data Unit Assignment

Jiho Min

MDM4U

Brandy Dobbin

Oct 20, 2019

1. Submit your spreadsheet containing your samples of m&m's.

- Attached in the dropbox

2. A research company wishes to perform an opinion poll ahead of a local election. They look at the breakdown of the town's population:

Age	Male	Female
0 to 9	193	185
10 to 19	197	206
20 to 29	245	243
30 to 39	228	224
40 to 49	254	245
50 to 59	243	258
60 to 69	165	171
70 to 79	96	112
Above 80	65	69

a. They wish to take a sample of 200 people

Write a proposal for the company explaining the following sample techniques, giving a detailed description of how each might be used in this situation

I. Stratified

- Stratified sampling is to divide the population into subgroups then draw out a certain number of people from the subgroups accordingly to the proportion. In this example, the population is already grouped into gender and age. It would be logical to exclude age 0-19 because their opinions do not influence the result of the local election. To execute stratified sampling method in this example, we would have to find the percentage of 20 to 29 male out of total population (245/2618*100=9.36%). Since we need to take a sample of 200 people, we need to multiply the percentage to figure out how many people should be chosen out of 245 males (200*9.36=19). Out of age range of 20 to 29 males, 19 people should

be chosen randomly. One can sample the rest of 13 subgroups using the same method of calculation.

II. Systematic
- Systematic sampling is to choose a random starting point than to draw out variables in a fixed interval. In this example, we can first organize the 2618 people based on their cell phone numbers or their house addresses. Laterwards one can divide the total number of people (2618) by 200 (sampling size) to calculate the interval between each sample, which would be 13.09. So, every 13th person can be chosen out of the list 2618 people, then survey can be conducted for each person.

III. Voluntary response
- Voluntary response sampling is to invite people to take part in the survey. In this example, surveyors can make an online poll or mail out questionnaires and wait for them to send back replies. The company can then collect the first 200 responses and use them to analyze data.

b. For each method, give advantages and disadvantages, keeping in mind the potential for bias

I. Stratified sampling has high precision compared to other sampling methods because it represents the entire population being studied. Each subgroup receives proportional representation which reduces sampling error. However, it is difficult and time-consuming to organize the population into subgroups. To identify individuals, it requires knowledge of the strata group prior. As a result of the complicated procedure, it would cost and take longer to execute the sampling method.

II. Systematic sampling is simple and has a low-risk factor compared to its simplicity. Also the researchers can control the coverage of the sample and the sample size as well. However, the sampling method has few flaws. Systematic sampling can not be executed when the size of the population is not reasonably approximated. It also has a greater risk of data manipulation. Stratified sampling can be affected by any hidden periodicity in the population. In this example, if the respondents figured that the 13th person's response is collected, they can make their way into the 13th position and manipulate the result.

III. Voluntary response sampling can reach a wide range of potential participants, but it has a high tendency of strongly opinionated people answering which can produce a biased result. It also can have non-response bias as well because some people can choose to answer. In this example, 20-30s who has low interest in politics are less likely to answer (non-response bias) and 40-50s are more likely to answer the questionnaire.

c. The company decides to use a stratified sample. However, after the election, they discover significant discrepancies between the predictions based on their findings, and the eventual outcome of the election.

- The stratified sample probably failed to identify correct subgroups that actually represents the opinion of the population. One's opinion about most elections is usually dependent on favored party, income, level of education, gender, and age. The stratified sample that the company conducted have taken age and gender into consideration, but they did not use other factors when identifying subgroups. This is probably why the result had so much discrepancies with the eventual outcome.

3. To study congestion in the core of a city, a survey measured traffic volume at a busy intersection between the hours of 1300 and 1600.

a. Describe the bias that could influence the results of this survey.

- The survey could have biased results due to several reasons. First, the survey only measured one location of the city, which is a source of undercoverage bias. Cores can't be generalized to one feature; some cores can be residential while, some cores are dense with office buildings. Also, the hours when the study was conducted created a timing bias because 1300 and 1600 are working hours in the core of a city when there isn't much traffic. Instead of measuring just 1300 to 1600, researchers should have measured all 24 hours.

b. Suggest ways in which the survey could be redesigned to reduce the possibility of bias

- In order to reduce the possibility of biased results, the survey should be taken at multiple locations in a city, and they should be of different types, such as busy intersection, residential area, and the business sector. Also, instead of collecting data in a certain period of time, one can collect data 24 hours to avoid timing bias.

4. You are designing a questionnaire on the subject of attitudes to reality TV. Create one example of each of the following types of question;

a. Open question
- What are your suggestions to improve reality shows?

b. Closed question
- Do you agree that the show: Keeping Up with the Kardashian influences teenagers?
- Yes
- No

c. Multiple choice question
- What is your current favorite reality show?
- Option A. Keeping Up with the Kardashians
- Option B. The Bachelor
- Option C. Survivor
- Option D. Big Brother
- Option E. Other

d. A question collecting continuous quantitative data
- How many hours do you spend watching reality shows in a week?

e. A question collecting discrete quantitative data
- How many reality shows do you currently watch?

f. A question collecting ordinal qualitative data
- How much do you enjoy watching reality shows?
- Not at all
- Slightly
- Neutral
- Little
- A lot

5. Identify which type of bias might be present in each of the following situations. Explain your reasoning, and suggest improvements to the method:

a. In order to determine how people feel about healthy food choices in the cafeteria, copies of the survey are left beside the cash register that can be picked up as students pay for their food. On the survey it asks for completed surveys to be dropped off in the office.

- Since the copies are left beside the register, students who do not buy food from the cafeteria will not be participating in the survey which is a non-response bias. The school would mostly collect positive opinions from the kids who already pay for the food. Also, since the submission of survey is bothersome, it's more likely that students with strong opinions will submit the survey. In order to reduce biased results, the survey copies should be placed at a remote location such as the exit or entrance of the cafeteria so that everyone can take the survey easily. To simplify the completion process, the school could place a submission box next to the survey sheets.

b. In order to determine the popularity of a breakfast cereal, 100 people are surveyed: 65 six-year-olds, 20 university students, 15 pensioners

- First of all, the survey is most likely to result undercoverage bias because the respondent's age group is shifted to six years olds. Also, the surveyor didn't consider the fact that the main consumers of breakfast cereals are young male and females in the 20-30s. In order to reduce biased results, the researchers should study the cereal consuming population and design a stratified sampling that represents the correct proportion of each age group.

c. A questionnaire included the question "Do you agree with increased funding for drug-addicts?"

- The question is as a leading question due to its manipulative wording. Wording such as 'drug-addicts', or 'increased funding' can generate biased images which can result in negative results. Also, yes or no questions forces respondents to choose an extreme answer rather than giving them a wide range of options. In order to reduce biased results, surveyors should use more neutral words such as rehabs for drug overdose users. The questionnaire may be rephrased "Do you have any opinion regarding the government's policy about drug-overdose?"

d. Another questionnaire includes the question "On a scale of 1 to 5, how much do you like mathematics?"

- First of all, the question itself is vague, it is likely to result measurement bias. It's also not clear whether 1 is the least or the most. Even if the responses are recorded, the measurement are too broad since scale of 1 to 5 is subjective. In order to reduce biased results, they should specify the question such as "What is your favourite subject? List: mathematics, english, science, PE, art, music"

6. Look again at the health questionnaire, and its problems, as well as the rules for a good questionnaire. Recreate the health questionnaire, improving it so that it does not suffer from the same issues.

- I excluded some questions that are inappropriate or unnecessary.

Health Questionnaire

Age:
Gender: F, M, Prefer not to answer

How much water do you drink per day?
- Less than 300ml
- 300-500ml
- 501-800ml
- 801-1000ml
- More than 1L

How often do you work out in a week?
- None
- Once a week
- 2 to 3 times a week
- 4 to 6 times a week
- Everyday

How many fruits do you eat per week?
- None
- 1 to 3
- 4 to 6
- 6 to 8
- More than 8

Are you a vegetarian?
- Yes
- If yes, what kind of vegetarian are you?
- No

How often do you eat fast food in a week?
- None

- Once a week
- 2 to 3 times a week
- More than 4 times a week
- Everyday

Do you have any special physical conditions such as diseases or allergies?

- No
- Yes (please specify)

What is your favorite fast food chain?

- McDonald
- Chick-Fil-A
- Burger King
- Annies
- Other

Do you have any family members who drink alcohol?

- Yes
- No

How much do you weigh? Kg

How long can you hold your breath?

- Less than 10 seconds
- 10 to 20 seconds
- 21 to 30 seconds
- 31 to 40 seconds
- Over 40 seconds

Do you have any suggestions that the government should increase their funding on pediatric cancer more than rehab for avid drug users?

Have you ever had a major operation?

- Yes
- If yes, then what:
- No

What is your opinion about the new diabetes treatment?

- Strongly disagree

- Disagree
- No opinion
- Agree
- Strongly Agree

How long do you sleep per day?

- Less than 5 hours
- 5 to 6 hours
- 7 to 8 hours
- 9 hours or more

Please submit the questionnaire to response@gmail.com

7. Research an incident in which the findings of a study or survey were compromised as a result of bias in the sampling or in the collection of data.

- A magazine company, Literary Digest conducted a poll for the 1936 presidential election and resulted in a 19% sampling error. The magazine predicted that Landon would win the election against President Roosevelt by 57% to 43%. However, the results turned out that Roosevelt won the election by 62% to 38% (Ozan Özbey, 2018). The poll had two sampling errors: selection bias and nonresponse bias. The magazine selected their surveyors based on telephone directories, magazine subscribers, and club membership lists. As a result, the magazine only targeted middle and upper class voters considering the fact that owning a telephone was a luxury in the 1930s. This excluded lower-income voters as well as people who aren't subscribed to magazines and clubs. Thus, the magazine had a low response rate resulting non-response bias. They magazine intended to survey 10 million people but only 2.4 million people responded.

8. Use Robert Harris's article on Evaluating Internet Sources to answer the following questions:

    a. Why is print material considered more credible than Internet material?

    - Printed materials are considered more credible than internet materials. Most of the printed materials are written by credible authors, who usually has some reputation, education, training and experience in a field relevant to the information. On the other hand, the authors of internet materials are usually unknown. This means that misinformed or unknown people can spread just rumours or false information.

b. According to Robert Harris what kind of information exists on the internet?

- According to Rober Harrison, various information exists on the internet such as facts, opinion, story, interpretation and statistics. The information on the internet exist in extremely large quantities on many levels of quality and reliability. It is being created and revised by numerous authors even at this very moment.

c. What tip does Robert Harris offer to determine if a source is reliable/credible?

- Robert Harrsion gave some source selection tips such as checking the credibility of the author, and the quality of information. In order to determine whether the author is credible, users can check the author's name, title of position, organization affiliated with author, and contact information. Also checking date of page creation or version, and indication of accuracy can be helpful. Quality of information can be checked by CARS checklist.

d. Summarize the CARS checklist. Include important questions you must ask yourself and indicators of poor information when evaluating an Internet site for each of the topics.

- CARS checklist is designed ease of learning and use. CARS stands for credibility, accuray, reasonableness, and support. Credibility is to check whether it a trustworthy source and author and identify can the readers trust the information or not. Checking the competence and the style of the writing or the information can be an indicator of lack of credibility. It would be wise to question oneself how does the writer knows the written information, and what about the source makes it believable. Accuracy is to check is the information up to date and is the source giving out factual and detailed truth. If the writing is vague, sources and outdated information, and seems like a one sided review it can be an indicator of a lack of accuracy. Reasonableness it to check whether the information is balanced and fair with no conflict of interest. If the writing seems to be written the article on behalf of their own interests, readers should intemperate tones, and overclaims as indicators of a lack of reasonableness. the Support is to find the evidence for the stated claim and to collect support from at least two other sources. Indicators of a lack of support can be absence of source documentation and numbers/statistics on the writing does not identify sources.

e. How can you tell the motivation and source of a document from the Internet address?

- The motivation and source of a document can be checked by top-level domain (TLD). TLDs refers to the last segment of urls such as .com, .org, .net, .ca, .mil. It represents the objective of the website. For instance, .com is given to commercial businesses while .mil is given to military websites and .ca refers to canadian websites.

References

Evaluating Internet Research Sources. (2018). Retrieved September 30, 2019, from Virtualsalt.com website: https://www.virtualsalt.com/evalu8it.htm

Ozan Özbey. (2018, February 25). Two Lessons of Sampling Bias from 1936 US Elections. Retrieved from Medium website: https://medium.com/@ozanozbey/how-not-to-sample-11579793dac