Collecting Data Unit Assignment

Jin Hyung Park

Teacher: Armitage, Stephen
MDM4U

Virtual High School

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

Please refer to the file submitted in 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.**

- Stratified:

  By utilizing stratified sampling, one can divide the population into separate groups and extract a certain number of people from the proportion of the groups. When looking into the aforementioned graph, we can easily notice that the graph already grouped the population into age and gender. In order to introduce a stratified random technique, let us look at the 20-29 female group.

To begin with, getting the total population of the female group, which is 1322, by adding up from 20-29 to Above 80. I do not count the group of 0 to 9 and 10 to 19 since they do not have voting rights. (Although in many countries the minimum voting age is 19, we could not draw out the number of 19-year-old population from the data thus it would be better not to calculate their population)

Then, find the percentage of 20-29 female groups out of the total population, by calculating $(243/2618) * 100 \ = \ about\ 9.28\%$. In order to appreciate the exact number of population to be chosen out of total females, do calculate $200 * 0.0928 = 18.56$. This means that I need to get about 18 people from the total population randomly.

For the rest of the subgroups excluding 0-9 male, 0-9 female, 10-19 male, 10-19 female, do the same calculation for getting the number of samples that we need to get.

- Systematic:

  Systematic sampling is to draw out variables starting from the random point but in a fixed, periodic interval. To begin with, getting the contact information of 2618 people from the downtown administration data. After contacting them, we can calculate the interval of performing periodic surveys by dividing the total number of 2618 people by sampling size which is 200. The answer would be 13.09 by calculating $2618/200 = 13.09$ which means that we can choose every 13th person to be conducted with a survey from the total number of 2618.

- Voluntary Response:

  Voluntary Response consists of participants who self-registered into the survey which usually illicit unsolicited biased results compared to a random sampling method. In this example, the company might send an email or text message to promote 2618 people to join the survey. The company might collect the 200 responses from the replies and analyze the data.

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

- Stratified sampling:

  The advantage of this method is high accuracy when compared to other sampling techniques since it tries to represent the proportion of the total population that needs to be studied. The beauty of this method is to take consideration of the proportional representation of each subgroup to reduce the room for sampling error. The disadvantage of this method is clearly the other side of the aforementioned precision. In order to keep a strict proportion of each subgroup, the researcher is required to appreciate the exact number of prior strata groups. This means that a survey with this method requires a lot of time and sometimes too difficult to keep the trend. It is worth noting that the standard of dividing subgroups should have meaning when it comes to the result of the real world. In order to identify which types of stratification that we need to identify for subgroups, the researcher should obtain the population data in advance.

- Systematic sampling

  The advantage of systematic sampling is simplicity over other sampling techniques since the researcher can assure that the population will be evenly sampled without creating random numbers on a regular basis. In addition, the researcher has room for controlling the coverage of those who are surveyed and the size of sampling.

  The disadvantage of systematic sampling is data manipulation. It is worth noting that the method has interacted with a veiled periodic trait of the population. If a proportion of the population changes due to the periodicity of the trait, we cannot assure that the population has random representativeness. For example, if any plotter recognizes the pattern of doing a survey - every 13th person to be conducted with the survey - one can make any interruption to manipulate the result. In addition, the method cannot be utilized if the proportion of the population is not reasonably measured.

- Voluntary response

  The advantage of voluntary response is that the researcher can contact the wide range of potential participants of a survey. Those who voluntarily join the survey might have a strong willingness to respond to the questions meticulously. Another advantage for this is to spare less time and effort since the method is easy to apply.

  The disadvantage of voluntary response is the bias of the results since there are high possibilities of only strongly opinionated people to answer the survey. In addition, since only some people might answer based on their willingness, there are high tendencies for producing unsolicited non-response bias. For example, in this example, 20-29 female or female might not want to join the survey while 40-50 male has a high willingness to answer the questions.

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. Suggest reasons for these differences.

   The primary reason for the discrepancies is a misrepresentation of opinion based on subgroups that we have only gender and age. When considering the political issues, a number of factors play each role to voters for choosing their stance. Take consider a regional issue, socioeconomic status, level of education, income, and many more. Do more further stratification. These aforementioned elements were not taken into consideration for selecting subgroups.

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.

To begin with, when taking research for only one core of the city, it might cause undercoverage bias. What is the meaning of the core? Some locations are dense with residential apartments while others can be crowded with office buildings. In order to avoid bias, the researcher should provide a clear meaning of "core" when conducting the survey and study various locations.

The equally convincing argument is timing bias. Considering the city with lots of corporations. It might have not much traffic during the working time, between the hours of 1300 and 1600. In order for studying the congestion itself of the core location, the researcher should study all of the daily hours to get impartial results.

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

To begin with, in order to avoid undercoverage bias, the researcher should take multiple "core" locations of the city as mentioned above. Before doing that, the threshold of "core" should be also clearly stated. While doing that, the researcher might require considering lots of types among core sites. The example can be the following: residential area, business area, and sightseeing area. Secondly, the researcher should conduct research for 24 hours for avoiding timing bias as also stated before.

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

- Open Question
  - What is your favorite reality show if there is any to prefer?
- Closed Question
  - Do you agree or disagree with the following statement: "Iron Chef Canada influences a lot of teenagers?"
    - Agree
    - DIsagree
- Multiple Choice Question

    - As of May 2020, which reality show programs do you think are the most famous in Canada?
        - Option A. The Amazing Race Canada
        - Option B. Backyard Builds
        - Option C. Canada’s Drag Race
        - Option D. Drag Heals
        - Option E. Fridge Wars
        - Option F. Others
- A question collecting continuous quantitative data
    - How many hours do you spend watching a reality show?
- A question collecting discrete quantitative data
    - How many different TV channels do you watch for reality shows?
- A question collecting ordinal qualitative data
    - How much do you like watching reality TV shows?
        - Very positive (5)
        - Positive (4)
        - Neutral (3)
        - Negative (2)
        - Very negative (1)

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 a 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.

   To begin with, the survey conducted research based on voluntary responses. This might cause non-response bias since those who do not buy and consume fast-food in cateria but packing their own food will not notice that the research happens. In other words, a non-response bias might happen. In addition, the submission is quite burdensome. The reason behind is that students willing to submit their opinion should visit the office and drop their survey off in the office. Thus, those who have an opinionated idea might be overrepresented in the result.

   In order for avoiding unsought biased results, the office should place the survey paper at numerous locations around the cafeteria. By doing this, anyone who uses the cafeteria every day might participate much easily so that the researcher can circumvent partial results. In addition to this, the researcher can take the survey online to simplify the submission process. By transforming into an internet-based form, students can state their opinion easily which means that the researcher can avoid non-response bias.

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

   To begin with, it is worth noting that the supermajority of respondents is 6-year-olds which might illicit undercoverage bias. When it comes to considering the main consumer of breakfast cereal, teenagers male and female would be the possible candidates since they are required to eat breakfast by their parents. Thus, in order to solve this bias issue, the researcher should adopt stratified sampling after studying the proportion of cereal consuming population.

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

   This question is the “leading question” that subtly prompts respondents to answer in a particular way where the researcher might want to get. In this example, using manipulative wording such as “increased funding” or “drug-addicts” might generate negative feelings toward a specific policy that the research targets. Thus, the result would incline to be negative as the research uses manipulative wording.

   In addition to the point, giving only options to choose between “Yes” or “No” would not represent a wide range of opinions. In order to avoid latent bias results, the question should be amended in the following statement: “Do you have any opinion regarding the government’s funding on the drug overdose treatment program?”

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

   To begin with, the question might throw information bias since the range given is too vague. What does 1 mean? dislike or like? What does 5 mean? The least? The most? Yes. Let us say the researcher amends the options by giving the following: “Highly dislike, dislike, neutral, like, Highly like”. Does any possibility of biased results purge? Not really. The participant might not have any clear comparison for answering the question. What does “dislike” mean? If the respondent dislikes PE and math together, which options to choose for providing “objective” results? In order to reduce biased results, the researcher should specify other subjects to compare. For example, the following question would be valid: “What are your favorite subjects of the following: PE, Math, Art, History, Computer Science, Other.”

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.

Health Questionnaire

Please answer the question following.

1. How much water do you drink everyday?

a. Less than 200ml
b. 200-400ml
c. 401-600ml
d. 601ml-800ml
e. More than 800ml

Age:

Sex: Male
Female
Other
Not to say

2. How often do you go for physical exercise for a week?

a. Not at all
b. Once per week
c. 2-3 times per week
d. 4-5 times per week
e. Almost everyday (more than 6 times per week)

3. How many fruits do you eat for a week?

a. Not at all
b. 1-3 fruit per week
c. 4-6 fruits per week
d. 6-8 fruits per week
e. More than 8 fruits per week

4-1. Are you a vegetarian?

a. Yes
b. No

4-2. If you answer “Yes” in 4-1, what kind of vegetarian are you?

a. Vegan
b. Lacto Vegetarian
c. Ovo Vegetarian
d. Flexitarian
e. Others

Go to Next Page.

Health Questionnaire

5. How often do you eat fast food per week?

a. Not at all

b. Once per week

c. 2-3 times per week

d. 4-5 times per week

e. Almost everyday (more than 6 times per week)

6. What is your favorite fast food chain?

a. McDonald

b. Burger King

c. Mom's touch

d. Lotte-Ria

e. Other

7. Do you have any special physical suffering such as allergies?

a. No

b. Yes

7-2. If you answer "Yes" in 7-1, please specify your physical suffering.

8. Do you have any relatives who drink alcohol?

a. No

b. Yes

9. How much do you weight?

a. Less than 50kg

b. 50-70kg

c. 70-90kg

d. 90-110kg

e. More than 110kg

Health Questionnaire

10. How long can you hold your own breath?

a. Less than 10 seconds

b. 10-20 seconds

c. 21-30 seconds

d. 31-40 seconds

e. Over 40 seconds

11. Do you have any opinion on the following statement: “Government should or should not increase their funding on pediatric cancer over funding on drug overuse treatment program.

____

12-1. Have you had major operation in a recent year?

a. No

b. Yes

12-2. If you answer “Yes” in 12-1, please specify your operation.

____

13. What is your opinion regarding new teen obesity program?

a. Strongly Disagree

b. Disagree

c. Agree

d. Strongly Agree

e. No preference

14. How long do you sleep per day?

a. Less than 5 hours

b. 5 to 6 hours

c. 6 to 7 hours

d. 7 to 8 hours

e. More than 8 hours

Thanks for your answer. Please submit to vhswelove@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.

The well-known incident that the result was compromised due to the result of bias is the survey conducted by Literary Digest, the ex-American weekly magazine, in the 1936 presidential election. Although it was once an influential magazine in the states, the name of the magazine is remembered today as its demise. It conducted a straw poll which predicted Republican candidate Landon would win the election against Democratian Roosevelt by 57% to 43%. However, the result was totally inverse. President Roosevelt won the election by grasping over 43 states while the Republican only secured 2 states. The result was 62% to 38%(Upenn Math, 2020).

It is worth noting that the research made a mistake in a polling technique which recorded a 19% sampling error: selection bias and non-response bias. The problem happens when the magazine chooses respondents. The way choosing the magazine was to choose respondents based on magazine subscribers, which means that it surveyed their subscribers first. Although the magazine had polled millions of subscribers, they were the group of earning monthly income over the average of the states. This sampling error was accelerated by surveying based on the readily contactable lists like automobile and telephone users. Those who had cars and phones had high socioeconomic positions at that time. This excluded the opinion of low-income voters with low-response rates, which led to non-response bias. If the magazine attempts to survey with the magazine which gets high recognition in a lower-income group, the result would be different since the survey might acquire relatively more responses from them.

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?

   The reason for giving more credibility to printed materials than Internet material is that most of the printed materials are written and double-checked by credible authors. They have reputations within the domain of knowledge that the material is trying to describe. In addition, when publishing printed materials, the publisher conducts a reference check after proof-reading the contents of materials. On the other hand, when it comes to internet material, one can write anything that he/she thinks in their own blogs. No proof-reading process undergone, relatively less information to check the background of authors. Thus, we can confidently trust printed materials over the Internet.

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

- He states that the plethora of information exists on the internet. We can easily find opinions, stories, ideas, facts, statistics, news with extremely large quantities involved. The information can be found on the internet varies a lot when it comes to considering the level of quality, quantity, and reliability. One interesting fact is that the information on the Internet is modified and created by numerous writers together simultaneously or asynchronously.

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

Robert Harrison suggested some tips to check the reliability of the information on the Internet by looking at the information of authors. Users might check the name, occupation, career histories of the author. Checking the organization affiliated with authors or articles might help to discern the reliability of the source. In addition, checking the date of created and revised, the document version would help to indicate the accuracy of sources. CARS checklist would be helpful to assess the quality of sources.

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 stands for credibility, accuracy, reasonableness, and support. To begin with, credibility is to check the reliability of authors or sources itself. By doing that, we can confidently discern the trustworthiness of the given information. We can throw the following questions that must be asked to ourselves while reading.
    - Does the article have sufficient evidence to prove its statement?
    - Does the article provide logically compelling arguments or reasons?
- Secondly, accuracy is to check whether the information is confidently up-to-date, and is factual truth. We can ask the timeliness-related questions to assert that the information truly reflects the status-quo. If there is no date clearly elucidated on the document or an old date on information which fluctuates suddenly, we can assume that the document lacks accuracy. In addition to this, when the writing is too vague or obscure enough not to disclose the detailed truth, we might consider a lack of accuracy in documents.
- Thirdly, reasonableness is to examine the fairness and objectivity of the given information without conflict of interest. The reader might consider the lack of reasonableness when the author writes the blog post with financial support from the related corporates for the purpose of promotion. If an article seems to represent the interest of an author, the reader should selectively accept the information from the article.
- Last but not least, support is to get at least two references to confirm the validity and corroboration of sources. The key point of the question is whether the reader can double-check the information given by the article. A reader might ask the following question: Does the article state enough references for their suggestion?

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

- We might be able to check the top-level domain(TLD) from the Internet address. TLD represents the information on websites. It refers to the last element of URLs including .com, .org, .edu, and many more. For example, if a website address ends with .com, we can acknowledge the website relates to commercial purposes while those ending with .edu might be the sites of educational purposes.

References

1. (n.d.). Retrieved from https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html
2. (n.d.). Retrieved from https://www.virtualsalt.com/evalu8it.htm