EN 207 / MM 217: Data analysis and interpretation

Week 0: Assignment

For discussion during Week 1

1. For a general election to be held for choosing the elected representatives to the Maharashtra state assembly, if a sample survey is conducted in the Mumbai airport, will the survey results be reliable? Explain your reasoning.

2. You are asked to conduct a sample survey for a general election to be held for choosing the elected representatives to the Maharashtra state assembly. Which are the best public places to conduct it. Why? Suppose the choices are (a) banks, (b) post-offices, (c) places of worship, (d) public transportation, (e) traffic junction, and, (f) market. Order these choices in terms of your preference with explanation as to why.

3. Read the attached document on how the Lokniti-CSDS National Election Study was held post-poll in 2019. Then, answer the following questions:

   (a) The methodology document indicates ”Our investigators sat down in the homes of people whose names were selected from the electoral roll, and asked them a detailed set of questions”. It would have been easier to ask questions near the polling booth by randomly choosing voters. Why was this not done?

   (b) The methodology document indicates that at every stage (the constituency, the assembly segment, the polling booths, the voters) were all chosen randomly. Is this the correct way? Isn’t it known that some constituencies reflect the mood of the nation better and those should have been chosen for the study? Also, if more youngsters have voted in the election, does it not make sense to choose only voters below 30 for the poll?

(c) In the 2019 elections, about 60 crore Indians voted. The survey was conducted by talking to nearly 20000 voters. Do you think that this is enough?

(d) The survey was conducted after polling was completed but before results were announced. Suppose two more surveys are conducted – one before polling and another after the results are announced. Do you expect the results to vary? Why?

4. You are the MoodIndigo coordinator. You want to know the favourite Bollywood actor among college students in Mumbai. You approach Red Mirchi FM station to conduct the survey. The host of Red Mirchi FM asks the listeners to call and name their favourite Bollywood actor on a given day. (a) Do you think the sample is representative? Comment. (b) Can you come up with one more methodology to obtain a representative sample? Clearly indicate your methodology (like the CSDS-Lokniti document) and your assumptions.

5. In the CSDS-Lokniti survey, dummy votes were used for collecting data on voting. This was done in order to ensure that the voters remain anonymous. Now, consider the following problem and solution methodology from C R Rao's *Statistics and Truth*.

> Another interesting application of randomness is in eliciting responses to sensitive questions. If we ask a question like, "Do you smoke marijuana?", we are not likely to get a correct response. On the other hand, we can list two questions (one of which is innocuous) T: Does your telephone number end with an even digit? S: Do you smoke marijuana? and ask the respondent to toss a coin and answer S correctly if head turns up and T correctly if tail turns up. The investigator does not know which question the respondent is answering and the secrecy of information is maintained. From such responses, the true proportion of individuals smoking marijuana can be estimated.

Suppose you have such data collected from a set of respondents. How will you calculate the true proportion of individuals smoking marijuana?

6. What is **Hawthorne effect** in data collection? Look up the information online. Is this effect true? Did you ever notice Hawthorne effect?

7. In this course, we have announced that the self-evaluation quiz marks are not counted towards evaluation for grading. However, let us say that by mid-September Institute informs me that there is no possibility of in-person, proctored final examination. Will it be ethical, at that point, to inform that we will use the self-evaluation quiz marks towards evaluation for grading?

8. Suppose you visit a restaurant. While paying the bill, the restaurant manager asks you if you are satisfied with the service. You are. Then, he offers that if you give a good review on Google Maps and show the same to him, he will give you a 10% discount on your bill. Is is ethical for you to accept the offer? Is it ethical if you give the review and refuse the discount? Is it ethical if you give the review and disclose that you got the discount for the review?

9. Suppose you visit a restaurant. While paying the bill, the restaurant manager offers that if you review the restaurant on Google Maps and show the same to him, he will give you a 10% discount on your bill irrespective of whether the review is positive or negative. Is is ethical for you to accept the offer? Is it ethical if you give the review and refuse the discount? Is it ethical if you give the review and disclose that you got the discount for the review? How does this offer compare with the previous one in terms of ethics?

10. In the quiz 0 that we asked you to take, suppose a student gives height as 153.05 (in cms) or weight as 58.390 (in kgs). Is this correct? Why?

11. In a medical trial that you are administering, to half of the patients, you are giving no medicine (placebo). Is is ethical? Does it make a difference if you know who is getting placebo or if you are not aware of it?

12. In a sheet metal forming company, some sheets develop defects and so, the company wants you to analyse their process data and identify the

Age	Frequency
10	5
11	12
12	11
13	12
14	10
15	5
16	2
17	1
18	1

Table 1: The frequency table of ages of the members of a chess club in a school.

origin of the defect. You ask the company to supply you all the data (of both defective and non-defective plates). The shop floor manager believes that you need only data pertaining to defective sheets. Is she correct? If not, how will you convince her that you actually need all data?

13. Consider the frequency table (Table 1) which gives the age distribution of members of a chess club in a school:

    (a) How many students are there in the club? (b) What is the mean age of the members of the club? (c) What is the median age of the members of the club? (d) What is the mode of the given data set? (e) What is the standard deviation of the given data? (f) How does the mean age change if children above the age of 16 are not included in the calculation? (g) How does the median age change if children above the age of 16 are not included in the calculation? (i) Compare the changes in mean age and median age when children above the age of 16 are included and excluded. Which one changes more? Why? (j) Based on this data, somebody concludes that the chess playing abilities of students peak in the early teens and declines in the late teens. Do you think that is a correct conclusion?