Problem Set 8

Survey Sampling Design. You decide to conduct a survey to measure physical activity in Boston. You plan to collect information on the amount of physical activity per week, history of diabetes, weight, height, age, and gender.

There are seventeen distinct neighborhoods in Boston, with substantial differences in race/ethnicity, socioeconomic status, and population density. You expect to observe variability in physical activity indicators by neighborhood. Unfortunately there is no specific information about within-neighborhood heterogeneity (variance); therefore, you design the survey based on the assumption that variances of indicators are equal across neighborhoods.

The table below displays population data for Boston in 2010. Assume that these population numbers are still accurate today.

Neighborhood	**Population - 2010**
South End	34,669
Central	30,901
Fenway - Kenmore	40,898
South Boston	33,688
Charlestown	16,439
Allston - Brighton	74,997
West Roxbury	30,445
Roxbury	59,790
East Boston	40,508
Jamaica Plain	39,897
Back Bay - Beacon Hill	27,476
Hyde Park	31,813
North Dorchester	28,384
South Dorchester	59,949
Roslindale	32,589
Mattapan	34,616
Harbor Islands	535
Boston	617,594

Source: http://www.bostonredevelopmentauthority.org/PDF/ResearchPublications//PDPercentChange.pdf

Consider the following questions:

1. Suppose you decided to randomly sample 1,700 people from the city of Boston (call this Design 1). For any given individual in South Dorchester, what is the probability of being selected in the survey? What is this probability for an individual in Harbor Islands?

2. What is likely the main challenge of implementing this survey?

(a) SRS is difficult to implement in practice, (b) SRS does not necessarily sample people from each neighborhood

3. Now, you randomly sample 100 people within each neighborhood (Design 2). What is the probability of a random individual in South Dorchester being sampled? What is the probability of a randomly selected individual in Harbor Islands being sampled?

4. What kind of survey design is Design 2?

   (a) stratified sample, (b) cluster sample, (c) simple random sample

5. Consider an alternate design (Design 3). In each neighborhood the number of individuals sampled is proportional to the population size of the neighborhood. Assuming that the sample size is fixed at 1,700, would you expect Design 2 or Design 3 to provide more precise estimates of the physical activity indicators among Boston residents?

   (a) Design 2, (b) Design 3

6. Once again, consider the probability of a random individual in South Dorchester being sampled; and the probability of a random individual in Harbor Islands being sampled. These probabilities are approximately the same as those in:

   (a) Design 1, (b) Design 2

7. Why might you want to use Design 2 compared to Design 3?

   (a) to increase precision, (b) if you wanted neighborhood-specific estimates, (c) both a and b

8. Next, you decide you do not want to visit all 17 neighborhoods, so you randomly sample 10 neighborhoods. Within each neighborhood selected, you randomly sample 170 people. Call this Design 4. What is the probability of a random individual in South Dorchester being included in the survey? What is the probability of a random individual in Harbor Islands being included in the survey?

9. A “self-weighting” design is a survey design for which every individual in the population has an equal probability of inclusion. Which survey designs are self-weighting (or approximately self-weighting)?

   (a) Design 1, (b) Design 2, (c) Design 3, (d) Design 4

10. Consider yet another design (Design 5), in which you again select 10 neighborhoods. Now, the probability of a neighborhood being included in the survey is proportional to its population size. Within each sampled neighborhood, you randomly sample 170 people. Would you expect Design 4 or Design 5 to provide more precise estimates of the physical activity indicators among Boston residents?

    (a) Design 4, (b) Design 5

Other aspects of survey design. Thus far, we have examined sampling design, which is only one element of designing a survey. Building and testing the survey instrument/questionnaire, anticipating and preparing for non-response, and training field teams to conduct the survey are several other critical aspects that we have not even touched on! Consider the following:

1. When designing your survey, you are trying to decide whether to use a 2-page questionnaire with 15 simple questions about whether an individual exercises, how many times per week, and whether they have a history of diabetes, along with some other very basic demographics (call this Survey 1). Your colleague says you could get much better information if you used a 10 page questionnaire with a more complete medical history and history of exercise and physical activity (call this Survey 2).

   Krosnick (1991) discusses “satisficing” (satisfy + suffice) in surveys. To paraphrase this discussion, satisficing occurs when, rather than optimizing their responses to best reflect reality, survey respondents try to reduce the cognitive burden associated with the survey and consequently may select a survey response haphazardly or even arbitrarily.

   Which survey would be more susceptible to satisficing?

   (a) Survey 1, (b) Survey 2

   Source: Krosnick, J. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5: 236.

2. You decide to implement a door-to-door survey, where you randomly sample addresses within a neighborhood and train a field team to ask the survey questions at the selected households. (You have a complete listing of addresses in Boston).

   Your colleague says you should obtain a listing of all land-line telephone numbers in Boston (a listing of cell phone numbers is not available) and randomly sample numbers from this list and ask the questions over the phone. For sake of argument, assume that you get a 100% response rate for both modes (door to door and phone calls), and that you construct survey weights using the table above.

   Would you expect the door-to-door or land-line method to produce unbiased results? (Think about which sampling frame is likely to be more complete.)

   (a) Door-to-door, (b) landline

   Would you expect the door-to-door or land-line method to be cheaper to implement?

   (a) Door-to-door, (b) landline

   Note: web-based surveys are also common. In order to minimize non-response, some surveys use multiple modes of response and follow-up (e.g. web + phone; or web + household follow-up).

3. Individuals can opt out of any part of your survey. High BMI and low physical activity are risk factors for diabetes, and you find that individuals with these characteristics are less likely to answer questions about history of diabetes (note that high BMI and low physical activity are risk factors for diabetes). In a complete case analysis, missing data is dropped, survey weights are recalculated, and data is analyzed assuming missing observations were never collected. In a complete case analysis, would you expect to obtain unbiased estimates of diabetes prevalence in Boston?

   (a) yes, (b) no

Case Control versus Cohort Study

1. One of the main advantages of case-control studies over cohort studies is that

   A. The investigator can identify exposed and unexposed people without concerns about selection bias.

   B. The investigator does not need to wait a long time for cases to occur.

   C. The investigator does not need to be concerned about issues of confounding.

   D. The investigator can examine whether a third factor modifies the relationship between exposure and outcome.

Alcohol Consumption and Cancer Study

Dr. Marks is interested in examining the association between alcohol consumption and a rare form of cancer. Since it is hard to identify cases, Dr. Marks designs a case-control study. Cases are identified from cancer treatment centers across the United States. Controls are selected on the day that the case is diagnosed with cancer from among the relatives of the cancer patient.

1. True or False: Dr. Marks should make sure to select the healthiest relative as the control to ensure that he will observe differences in risk between cases and controls.

2. The main concern of using relatives of cases as the controls is that
    A. Some cases may not have relatives that drink alcohol.
    B. Some cases may have relatives that live far away so not all relatives will be available to participate.
    C. Relatives of cases may be more likely to have levels of alcohol consumption that are more similar to the cases than the population that gave rise to the cancer cases.
    D. The controls may develop other diseases and will not be available to participate in this study.

3. Based on the study description above, Dr. Marks conducted a
    A. Density case control study
    B. Nested case-control study
    C. Case-cohort study
    D. Retrospective cohort study

4. True or False: Using this design, a relative who served as a control cannot be included as a case if he later develops the cancer of interest.

5. True or False: Using the data collected in this study, Dr. Marks will be able to estimate the rates for developing this cancer.