Math 218, Mathematical Statistics
D Joyce, Spring 2016

Assignment. Chap. 3, exercises 1, 3, 4, 7, 8, 9, 12, 13, 15.

Selected answers.

1. In each of the following instances (i) tell whether the study is experimental or observational, (ii) tell whether the study is comparative or descriptive, and (iii) if the study is comparative, identify the response and explanatory variables.

a. Teenage students from suburban schools are interviewed to estimate the proportion with a gang affiliation.

It's observational. If the data from all the schools is merged, the study is descriptive; if the data is separated by school, then for each school, there will be a proportion with a gang affiliation, in which case it's comparative, and the response variable is the proportion while the explanatory variable is the school.

b. A study from hospital records found that women who had low weight gain during their pregnancy were more likely to have low birth weight babies than women with high weight gains.

Observational. Comparative. Mother's weight gain explanatory; baby's birth weight response.

c. A study monitors the the occurrence of heart disease over a 5 year period in men randomized to eat high fiber or low fiber diets.

It's experimental because the study assigned the men randomly into one of the two groups. Had it just classified the men according to the amount of fiber they had in their diet, it would have been observational. It's a comparative study. The explanatory variable is the diet treatment while the response variable is the level of heart disease.

d. Annual rates of return on investment are compared among different groups of mutual funds.

It's observational and comparative. The response variable is the rate of return; the explanatory variable the mutual fund.

4. Confounding is present in each of the following situations. Explain the nature of the confounding and why the conclusions drawn may not be valid.

a. A cross-country coach thinks that a particular technique will improve the times of his runners. As an experiment he offers an extra daily practice to work on this technique for runners who want to participate. At the end of the season the coach concludes that the technique is effective, because runners who participated in the extra practices have faster average times than those who did not.

The sample of participants was self-selected, that is to say, the runners who chose to participate may be the ones who run faster.

The coach didn't measure what he wanted to test. He thought the times would improve, but he didn't measure the improvement in times, but the times themselves. He should have recorded the average times at the beginning of the season and the average times at the end of the season so he could take the difference to see the improvement.

Also, perhaps any kind of practice—extra practice especially—would have been enough to improve performance.

b. In a geriatric study samples of people over 65 from a general community and from a retirement community are followed. At the end of two years it is found that a larger proportion of the people continue to reside in the retirement community than in the general community. It is concluded that the retirement community living allows elderly to maintain their independence longer than does the general community living.

Sounds like a non sequitur to me. What does independence have to do with how long someone stays in a community? Convicts in prison for life stay in their community for a long time, but they have no independence whatsoever.

7. In each of the following instances (i) tell whether the study is a survey, a prospective study, or a retrospective study, (ii) tell whether the study is comparative or descriptive, and (iii) if the study is comparative, identify the response and explanatory variables.

a. A sample of members listed in the directory from a professional organization is used to estimate the proportion of female members.

It's a descriptive survey.

b. To assess the effect of smoking during pregnancy on premature delivery, mothers of preterm infants are matched by age and number of previous pregnancies to mothers of full-term infants then both are asked about their smoking habits during pregnancy.

It's a retrospective study. It's comparative. I would say the response variable is preterm vs. full-term while the predictor variable is smoking habit.

c. A sociologist interviews juvenile offenders to find out what proportion live in foster care.

It's a descriptive survey.

d. A marketing study uses the registration card mailed back to the company following the purchase of a VCR to gauge to the percentage of purchasers who learned about that brand from advertising on radio, television, periodicals, or by word of mouth.

It's a descriptive survey.

8. The Nationwide Food Consumption Survey is the only survey on U.S. household food consumption and individual diets. Done once a decade, it is the basis for decision regarding food programs such as the multibillion dollar Food Stamps program. The survey conducted for the U.S. Department of Agriculture in 1987–1988 did extensive interviews that required from 80 minutes to over 5 hours to complete in exchange for a $2 reimbursement. As a result only about one-third of the people contacted participates. In addition, interviewers were inadequately trained, there was a high interviewer turnover, and the survey design resulted in data with low credibility, prompting the General Accounting Office to question whether or not the data were usable.

a. What are some potential sources of bias in this survey?

The one-third of the people who completed the survey probably were not a representative sample. Only those people with enough time could complete it, and only those who found $2 a large enough compensation for their time would even consider it. It is not stated how the interviewees were selected, so there may be further bias due to the selection method.

b. Explain how the problems you listed in part a could influence the survey results.

Working families and single parent families may not be well represented due to time limitations. Unemployed and low-income families may be over-represented due to the low-compensation for completing the survey. The unknown selection method may have missed whole classes of people, like the homeless.

12. A large school district plans to survey 1000 out of 50,000 parents or guardians with enrolled children regarding their preferences on options to deal with growth in student enrollment. A complete alphabetical list of parent/guardian names is available. In each of the following instances name the sampling method used.

a. One of the first 50 names on the complete list is randomly chosen; that person and every 50th person on the list after that person are surveyed.

This is 1-in-50 systematic sampling.

b. The complete list of names is divided into separate lists by their oldest child's year in school (K through 12). Random numbers are assigned to names, and each list is ordered by the assigned random numbers. The first 2% of the ordered names in each list are surveyed.

Stratified random sampling.

c. A random number is assigned to each name using a computer random number generator. Names are ordered by the random numbers, and the first 1000 are surveyed.

SRS, Simple random sampling (without replacement).

d. A 50% SRS is taken of the high schools (grades 9–12) and of the elementary schools (grades K-8) from the district. Within each selected school a 50% SRS of grades is chosen. A sampling frame is composed of parents or guardians of the children in the selected grades and an SRS of those people is surveyed.

Multistage cluster sampling.

13. Comment on the advantages associated with each of the sampling plans of exercise 12.

a. Straightforward random sample.

b. Makes sure each of the 13 grades is equally represented.

c. Straightforward random sample. For all practical purposes, the same as the method in part a.

d. It may be easier to contact the interviewees if they're only from a few schools.

15. A survey of 10 million residential telephone customers in a state with 15 geographic "local" calling areas is planned to estimate the market share of long distance phone companies. One of the local calling areas is a major city in the state with 20% of the total customers. A list of residential phone numbers for each calling area is available. Describe how you would select a sample of 1000 residential customers using SRS, systematics sampling, stratified random sampling, and multistage cluster sampling. Comment on the advantages and disadvantages associated with each sampling plan.

Simple random sampling (SRS). To do SRS, pool al the lists together and randomly select 1000 from the whole list, perhaps using a random number generator to help with the selection. Advantages: every customer has the same probability of being selected, the 15 local calling areas will all be represented with the number of customers being selected approximately proportional to the total number of customers in the local calling area. Disadvantages: none in particular if the customers are going to be contacted by phone. Note that even personal computers are powerful enough to deal with a list of 10 million phone numbers.

Systematic sampling. Take the joined list mentioned above. Select 1 of the first 10000 at random in that list, then every 10000th name after that. That will give a list of 1000 customers for the sample. (Note that it isn't really necessary to join the 15 lists together to create this sample.) Advantages: slightly more representative for the 15 calling areas as the number of selected customers will be almost exactly proportional to the total number of customers in the calling area. Disadvantages: none in particular.

Stratified random sampling. The only strata mentioned are the local calling area and city versus noncity. The effect of stratifying on the local calling area is nil. A random sample for each of the local calling areas is make, perhaps with the size of the sample proportional to the number of customers in the calling area, but that's essentially what was done above. But if all 15 calling areas have the same number of customers selected (that would be about 67) you might get better information. Remember, the information is dependent on the size of the sample, not so much on the size of the population, so making all 15 samples as large as possible gives better information.

The city versus noncity divides the calling areas into two kinds. One calling area is the city, the other calling areas are noncity. It might make sense to stratify on that distinction.

Given other information, like the economic status of the customers, race and ethnicity, or family size, other stratifications may make more sense.

Multistage cluster sampling. There are only going to be two stages here. Of the 15 calling areas, randomly select some number of them, say 4. Then do an SRS on each of the 4 calling areas. An advantage is that if the customers are going to be contacted in some way that is easier if they're in the same locality, then cluster sampling will help.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`