Michael Uftring
Indiana University
V506 - Statistical Analysis, Summer 2018
Homework Exercise 1

Questions

1. Identify the following variable characteristics

- Type: quantitative or categorical (qualitative)
- Range: discrete or continuous
- Scale: interval, nominal, or ordinal

Variable	Type	Range	Scale
Occupation (plumber, teacher, secretary)	categorical	discrete	nominal
Occupational status (blue collar, white collar)	categorical	discrete	ordinal
Social status (lower, middle, upper class)	categorical	discrete	ordinal
Race	categorical	discrete	nominal
Statewide murder rate (number of murders per 1000 population)	quantitative	discrete	interval
County population size (number of people)	quantitative	discrete	interval
Population growth rate (in percentages)	quantitative	continuous	interval
Community size (rural, small town, large town, small city, large city)	categorical	discrete	ordinal
Annual income (thousands of dollars per year)	quantitative	discrete	interval
Attitude toward affirmative action (favorable, neutral, unfavorable)	categorical	discrete	ordinal

Lifetime number of sexual partners	quantitative	discrete	interval

2. Samples and Population

A sociologist wants to estimate the average age at marriage for women in New England in the early eighteenth century. She finds within her state archives marriage records for a large Puritan village for the years 1700-1730. She then takes a sample of those records, noting the age of the bride for each. The average age in the sample is 24.1 years. Using a statistical method from Chapter 5, the sociologist estimates the average age of brides at marriage for the population to be between 23.5 and 24.7 years.

- *(a) What part of this example is descriptive?*

  The average age in the sample is 24.1 years.

- *(b) What part of this example is inferential?*

  the sociologist estimates the average age of brides at marriage for the population to be between 23.5 and 24.7 years

- *(c) To what population does the inference refer?*

  married women in New England in the early eighteenth century

3. Statistics and Parameters

In the 2014 gubernatorial election in California, a CBS News exit poll of 1824 voters stated that 60.5% voted for the Democratic candidate, Jerry Brown. Of all 7.3 million voters, 60.0% voted for Brown.

- *(a) What was the population and what was the sample?*

  The population was the 7.3 million voters in California.

  The sample was 1824 voters in the exit poll.

- *(b) Identify a statistic and a parameter.*

  Statistic: 60.5% of 1824 voters in an exit poll voted for Jerry Brown.

Parameter: 60.0% of all 7.3 million voters voted for Jerry Brown.

4. Description and Inference.

The Current Population Survey (CPS) is a monthly survey of households conducted by the U.S. Census Bureau. A CPS of 68,000 households in 2013 indicated that of those households, 9.6% of the whites, 27.2% of the blacks, 23.5% of the Hispanics, and 10.5% of the Asians had annual income below the poverty level.

- *(a) Are these numbers statistics, or parameters? Explain.*

  These numbers are statistics because they are derived from a sample of 68,000 of all households in the United States. To find the corresponding actual parameter values, all households in the United States would need to be surveyed.

- *(b) A method from this text predicts that the percentage of all black households in the United States having income below the poverty level is at least 25% but no greater than 29%. What type of statistical method does this illustrate – descriptive or inferential? Why?*

  This is an inferential statistic because it is making a prediction about a population.

5. Simple Random Sampling

A local telephone directory has 400 pages with 130 names per page, a total of 52,000 names. Explain how you could choose a simple random sample of five names. Show how to select five random numbers to identify subjects for the sample.

The general process for selecting a sample of size 5 from this directory of 52,000 names using simple random sampling:

1. assign a number to each directory entry (name), from 1 to 52,000
2. generate a set of 5 numbers in the range 1 to 52,000 randomly
3. sample the subjects with the 5 numbers generated randomly

More specifically, we could use five digit long numbers and assign them to the names in the directory. Starting with 00001 for the first name, 00002 for the second, and so on up to 52000 for the final name. Next, we could consult a table of random numbers (e.g., in *Handbook of Tables for Probability and Statistics*). Pick a row and column starting point, select numbers in the range 00001 to 52000 (reading across rows or down columns, and skipping numbers outside the range) until we have five. This set of five represent the random selections from the

population to be included in the sample.

6. Observational versus Experimental Studies

A study is planned about whether passive smoking (being exposed to secondhand cigarette smoke on a regular basis) leads to higher rates of lung cancer.

- *(a) One possible study would take a sample of children, randomly select half of them for placement in an environment where they are passive smokers, place the other half in an environment where they are not exposed to smoke, and then 60 years later observe whether each person has developed lung cancer. Would this study be experimental or observational? Why?*

  This would be an experimental study because the treatment (placement in a passive smoking environment or not) is determined by a random process for all subjects.

- *(b) For many reasons, including time and ethics, it is not possible to conduct the study in (a). Describe a way that is possible, and indicate whether it would be an experimental, or observational, study.*

  An observational study approach would be to interview two groups of subjects in the same age range (e.g., 65-75 years old): one group which does not have lung cancer, and another group which does have lung cancer. Subjects who are or were smokers would be excluded. A single question could be asked: did you live in a passive smoking environment (i.e., exposed to second hand smoke)? Perhaps some alternate or follow-up questions might be: (1) how long did you live in a passive smoking environment? (2) how much second hand smoke were you exposed to over your lifetime (estimated in number of days, weeks, or years)?

7. Sampling Bias

Explain how the following had sampling bias, and explain what it means to call their samples "volunteer samples."

- *(a) The BBC in Britain requested viewers to call the network and indicate their favorite poem. Of more than 7500 callers, more than twice as many voted for Rudyard Kipling's "If" than for any other poem. The BBC reported that this was the clear favorite.*

- *(b) A mail-in questionnaire published in TV Guide posed the question "Should the President have the Line Item Veto to eliminate waste?" Of those who responded, 97% said*

yes. For the same question posed to a random sample, 71% said yes.

A volunteer sample is a group of subjects made up of people who have volunteered to be in a study. Volunteer subjects are assembled using a non-probabilistic sampling method whereby the probability of any member of the population being included is unknown. Non-probabilistic sampling, where volunteer sampling is a primary cause, results in sampling bias.

In both cases, the BBC TV viewers and TV Guide readers, providing a response was strictly voluntary. Neither sample was randomly chosen from a population. Those who volunteered are unlikely to be a representative sample of their respective populations. Those who responded are also likely to feel strongly enough to respond. Thus, there is almost certainly under-coverage from certain segments of the population. These are perfect examples of sampling bias.

8. Cluster Sampling

Refer to Problem #5 about selecting 5 of 52,000 names on 400 pages of a directory.

- *(a) Select five numbers to identify subjects for a systematic random sample of five names from the directory.*

  First, let `n` designate the sample size (5), and let `N` designate the population size (52,000). Next, let `k = N/n` and call it the skip number.

  Select a subject at random from the first `k` names, then skip ahead `k` names and select that one, and repeat until `n` subjects are selected.

  Essentially with **systematic random sampling** the population is portioned out into `n` segments of equal size `k`. A random number is generated between 1 and `k`, and that serves as an index into each segment for the selection of the subjects to be included in the sample.

- *(b) Is cluster sampling applicable? How could it be carried out, and what would be the advantages and disadvantages?*

  Cluster sampling could be applied here, but whether it is effective or not would depend on the nature of the study. Given that the entire population is available, cluster sampling may not be necessary (or even the best approach).

  The 52,000 directory entries could be clustered by street or neighborhood. Next, randomly select five of the clusters. Finally, perform simple random sampling on the

selected clusters to select one entry each to form the sample.

It is not clear what advantages and disadvantages are presented here with cluster sampling. The desired sample size is relatively small, and the entire population is known, and other sampling mechanisms seem more suitable.

The small desired sample size poses a disadvantage to true cluster sampling (without the second stage of simple random sampling within randomly selected clusters, noted in the example solution above) in that `52000/5` clusters would need to be formed, and just one cluster selected and all subjects within form the sample. How would we form that many clusters? What would be an effective means to partition the population?

9. Stratified Sampling

With a total sample size of 100, we want to compare Native Americans to other Americans on the percentage favoring legalized gambling. Why might it be useful to take a disproportional stratified random sample?

A disproportional stratified random sample would be useful here because the population size of Native Americans compared to all other Americans is very small. In order to obtain a meaningful representation of Native Americans, they should be represented in a number greater than their actual population percentage.