STAT 130 Midterm Part B

Wesley Chang

Fall 2020, Professor Subir Ghosh

Question 1

Please view the four videos, two each in Groops 1 and 2. Summarize the content of each video in bullet forms.

Group I:

How can a survey of 1,000 people tell you what the whole U.S. thinks?

This video begins by proposing a question: How can a survey of a small group of people measure the opinion of the whole United States? The speaker then introduces the idea of random sampling, in which conclusions about an entire population can be approximated from sampling from a smaller portion of the population, known as a sample, whose participants are chosen randomly. Although the most accurate way to gather information about a population is by asking the entire population, there many barriers and challenges that make this method impractical and too time-consuming to be useful to whoever wants to use this information. The speaker adds that “A nationally representative survey requires a random sample, in which each person in the United States have a chance of selection”. Conducting a truly representative survey also necessitates inclusion of different age groups, income levels, ethnic and racial backgrounds, and all political leanings. The speaker further adds that random sampling is found in many aspects of everyday life, even if not formally or consciously performed. For example, when making a pot of soup, once all ingredients are mixed together, we “sample” the soup by tasting a small sip of it, rather than drinking the whole soup. This allows us to have a good idea of what the rest of the soup will taste like so that we can know whether to make any adjustments, without having to consume the whole pot first. When conducting a sample by surveying participants, there is a great chance that the profiles of respondents will differ from the profiles of non-respondents. Therefore, pollsters use the technique of “weighting”, in which known characteristics about the population are used to adjust the data collected in a survey to control for variables, such as age, gender, and location. Through sampling and weighting, statisticians are able to use data from a small sample to infer data about the population at large.

What are nonprobability surveys?

For many years, the traditional way that public opinion has been measured in the United States has been through telephone polling, which is based on probability and gives every member of the population (in this case the United States) an equal chance to be polled. However, as telephone polling becomes outdated and less accurate, researchers have begun developing other methods of sampling, namely nonprobability surveys for polling public opinion. Traditional telephone polling is conducted by drawing random samples of a list of telephone numbers or addresses for all members of the population and surveying those individuals in the sample. In recent years, drastically declining response rates (36% in 1997 to 9% in 2016) have led researches to begin questioning the accuracy of data generated using this method. This has led reserachers to pursue polling through other avenues. The current ubiquity of the Internet has opened a channel through which polling can be conducted cheaply, convieniently, and faster. Since there is no “master list” of emails for every individual, sampling methods through the internet are nonprobability based (as opposed to telephone polling). These surveys are known as “opt-in”, and

participants are found through ways such as website advertisements, customer loyalty programs, or voluntary survey panels. However, this “opt-in” method has a perceived increased risk of non-response bias, due to the voluntary nature of the way these surveys are conducted. Although non-response bias is an issue with telephone polling, decades of research data have given researchers insight to how results from this type of sampling may be different from the entire population, which have allowed them to adjust sampled data to the population. With “opt-in” surveys, researchers do not have the same historical data to allow them to make corrections to data. Additionally, companies that engage in “opt-in” surveys also tend to keep their data private, which also reduces the amount of information researchers can use to figure out how to adjust sample data. Nevertheless, nonprobability and “opt-in” surveys are still able to generate useful information about public opinion, such as presidential approval. There is also a third newly developed way of taking public opinion surveys, which is online-based probability sampling. However, this method is expensive and difficult to conduct, but is becoming increasingly adopted by major polling organizations such as the Pew Research Center.

Group II:

5.6 Non-probability sampling: Quantitative methods

Non-probability sampling differs from probability sampling in that some elements of the sampling frame do not have the same probability of being selected as the rest of the sampling frame. This makes it impossible to determine the margin of error in non-probability samples, and how likely it is the the sample represents the rest of the population. The simplest form of nonprobability sampling is convenience sampling, in which participants are chosen from an already premade group, such as a group of university students from a specific university and a specific Bachelors program. This form of sampling suffers from a high risk of bias, as participants are not randomly selected, and have already been categorized in other ways, likely differentiating them from the wider population. However, this method can be used when it is impossible or impractical to determine the sampling frame for a particular research question, or when the research question is “universalistic causal”. A form of convenience sampling is “snowball sampling”, in which a small initial group of participants are selected and asked to recruit more participants into the sample. Those additional participants are further asked to recruit more, and this process is repeated such that the initial small sample can grow very large in a short amount of time. This can be useful when attempting to study characteristics of a specific population, such as patients with a rare form of cancer. It may be difficult for researchers to find a large sample of these patients, but it is possible that these patients are already connected to other patients. This allows researchers to find a smaller group of cancer patients and obtain a larger sample by utilizing the connections that the patients in the initial group may possess. However, snowball sampling also suffers from the same bias issues that convenience sampling has, if not more drastically. The next form of non-probability sampling is purposive sampling, which is primarily used for qualitative sampling. In this method, participants are selected based on the judgement of the researcher. This method also suffers from all the bias that convenience and snowball sampling possess, and may further suffer from bias introduced through the researcher’s judgement. Finally, there is quota sampling, in which known characteristics of a population are used to divide the population into separate categories with different weighted amounts of participants, similar to stratified sampling, but participants in each category are chosen through convenience sampling. This method suffers from the same possible biases of the above methods, such as the opinion of interviewers when determining which participants to include. Due to the biases in nonprobability sampling, all results generated through these methods should be approached carefully and with awareness of the effect of the biases.

5.3 Probability sampling: Quantitative methods

All forms of probability sampling share an essential feature: “For each element the probability is known and non-zero”. This means that any element of the population being studied should have an equal and possible chance of being selected. Probability sampling also requires the creation of a sampling frame, which is a list of all elements in a population that can be accessed, which is used to determine the probability of each element being selected. Probability sampling reduces the possible systemic bias in selecting participants by reducing the risk of over or underrepresentation of any subgroup that has extreme characteristics, which would affect the final results of the sample. The effectiveness of probability sampling can be explained through the example of random assignment. In the example shown in the video, multiple samples are randomly drawn with an experimental and control group. Even though the differences between a specific experimental and control group may be high, over the long run and over many samples, the characteristics of all the experimental and control groups should be nearly identical. Although the above example was applied towards random assignment with experimental and control groups, we can also use the same principle to conduct random selection for repeated samples of a single type, where over many samples, the characteristics of the many samples should match the characteristics of the population. The speaker also adds that, “In the long run, any specific particiapnt characteristics will be represented in the sample proportionally to their presence in the population”, which would constitute “representative sampling”. In addition to creating a representative sample in the long run, probability sampling also allows researchers to asses the accuracy of the sample. Over repeated samples, we can determine the margin of error for the sample, or what the value of the samples differ from the population in a certain proportion of the samples. Therefore, we can assess what the amount the sample may differ from the population by assessing the data from the whole representative sample. We will also be able to create a confidence interval that may tell use the range of values that a given sample may fall within in a specific percentage of times.

Question 2

From a population of 10,000 adults of a community, a simple random sample of 100 adults is selected without replacement. What is the chance of two population members, Bob and Fran, to be included in the sample? Present the formula and calculate its numerical value by the R program. Copy and paste the R output with your answer.

Answer

Given the population of 10,000 adults, the chance of any population member to be included in a sample of 100 is found by the formula:

$$\frac{\binom{9999}{99}}{\binom{10000}{100}}$$

which simplifies to:

$$\frac{\frac{9999!}{99!9900!}}{\frac{10000!}{100!9900!}}$$

and further into:

$$\frac{\frac{9999!}{99!}}{\frac{10000!}{100!}}$$

and equals:

$$\frac{1}{100}$$

This formula takes the ratio of the combinations of the sample with the desired and without the desired participants. Therefore, when applied to two population members, this formula changes to:

$$\frac{\binom{9998}{98}}{\binom{10000}{100}}$$

which then simplifies to:

$$\frac{\frac{9998!}{98!9900!}}{\frac{10000!}{100!9900!}}$$

and further into:

$$\frac{\frac{9998!}{98!}}{\frac{10000!}{100!}}$$

which equals:

$$\frac{9998!}{10000!} \times \frac{100!}{98!}$$

and then:

$$\frac{9998!}{9998! \times 9999! \times 1000} \times \frac{100 \times 99 \times 98!}{98!}$$

which equals:

$$\frac{100 \times 99}{10000 \times 9999}$$

which results in the value of:

```
answer <- (100 * 99) / (10000 * 9999)
answer
```

```
## [1] 9.90099e-05
```

or:

$$\frac{1}{10100}$$