Sampling / Descriptive Statistics 1 – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Populations and Bias

Each of the following scenarios involves collecting data to learn about a population. State (a) what population is involved, and (b) why the sample is biased. To demonstrate that a sample is biased, you must argue that certain members of the population are more or less likely to be sampled than others. Note: there will usually be many valid answers for parts (a) and (b), but your answer to part (b) will depend on how you define the population in part (a).

1. You need a survey on household spending patterns. You take a random sample from the customer list of the local brokerage firm.

   Solution: *Population: the spending patterns of all households.* People who have brokerage accounts are certainly more wealthy than average, so richer people are more likely to be sampled.

2. You want to learn about New York City residents' sentiments (positive or negative) towards their new mayor, Bill de Blasio. You search for "de Blasio" on Twitter and read the first 100 relevant search results.

   Solution: *Population: the sentiments towards de Blasio for all New York City residents.* People who use twitter are more likely to be sampled than people who don't. People who have tweeted about de Blasio are more likely to be sampled. Furthermore, even among all people who use twitter, and have tweeted about de Blasio, those with recent tweets are more likely to be sampled.

3. You need to know the opinions of Langone students with regard to some curriculum matters. You ask some of the people in your class.

   Solution: *Population: the opinions of all Langone students.* People more advanced in the program are less likely to be included.

4. You want to learn about the quality of the food at a local restaurant. You read the reviews for the restaurant on Yelp.com.

   Solution: *Population: opinions of all people who have eaten at the restaurant.* The people who write reviews on Yelp.com are more likely to be sampled than people who do not use Yelp, or do not have Yelp accounts. (Yelp reviewers tend to have extreme opinions towards food, or tend to be overly critical.)

5. You want to estimate the rate of growth of stocks over the last fifty years. You take a random sample of the stocks listed today on either the New York Stock Exchange or the Nasdaq. Some of these stocks did not exist fifty years ago; you set these aside. For the other stocks, you identify their prices fifty years ago, and you use this to compute the growth rate.

Solution: *Population: the rates of growth of all stocks over the last fifty years.* Companies that were listed fifty years ago but did not survive are not available to appear in your sample. This is an example of survival bias. You will seriously overestimate the growth rate!

Types of Data

6. The class survey asked each respondent to report the following information: gender; birth date; GMAT score; undergraduate major; time spent studying per week; interest level in the course; industry; job type; work hours per week; commute time; number of dinners out per month; number of pairs of shoes; cups of coffee consumed per week; and number of websites visited per day.

   (a) Which of the variables measured by the survey are categorical/qualitative?

   Solution: gender; major; industry; job type

   (b) Which of the variables measured by the survey are numerical/quantitative?

   Solution: birth date; GMAT score; study time; interest level; work hours; commute time; dinners out; shoes; cups of coffee; number of websites

7. What type of variable is the answer to the phone prompt "Enter '1' for English, '2' for Spanish."? Why?

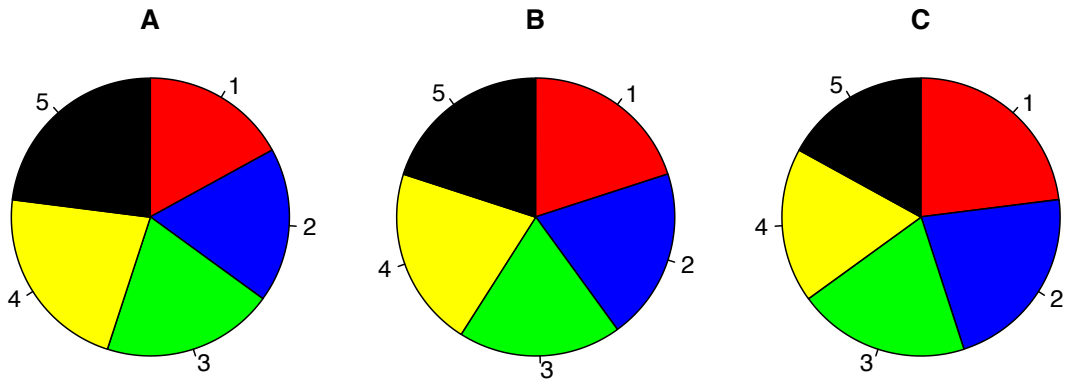
   Solution: Categorical. Even the variable is measured on a numerical scale (1 or 2), the scale is not meaningful.

8. Each Yelp restaurant includes a star rating (1–5). What type of variable is the star rating?

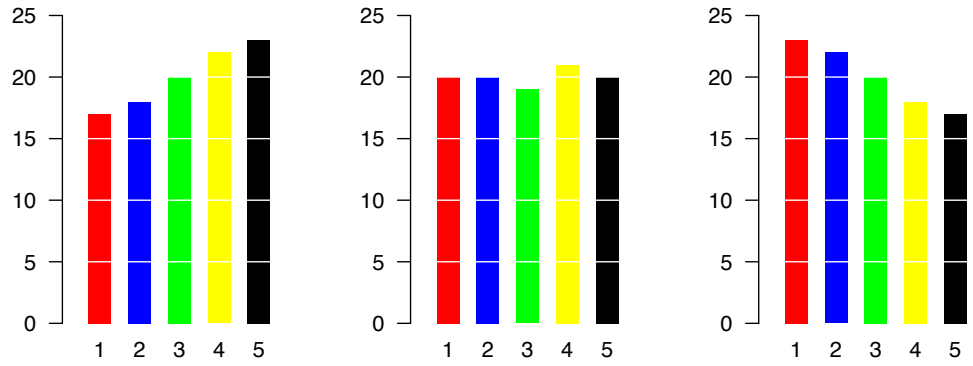
   Solution: An argument can be made for both Categorical and Quantitative. In some ways the numerical scale is meaningful (order matters), but in other ways it is not (the difference between 4 and 5 stars is different than the difference between 1 and 2 stars). Sometimes this type of data is called "Ordinal."

Describing Categorical (Qualitative) Data

9. Use the following pie charts to rank the categories (1–5) by size.



Solution: This is much easier if we have bar charts instead:



The relative ordering of the categories is obvious. The takeaway here is that you should never use a pie chart; a bar chart conveys the same information, and it is much easier to read.

10. List two methods to describe the reported undergraduate majors of the class survey respondents.

Solution: Frequency table; bar chart.

11. Draw what you think the bar chart for the birth months of the survey respondents will look like.

Solution:

Describing Numerical (Quantitative) Data

12. Draw what you think the histogram for "Websites Visited per Day" will look like.

Solution:

13. Draw what you think the histogram for "Dinners per Month" will look like.

Solution:

14. Draw what you think the histogram for "Interest in this Class" will look like.

Solution: