Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

August 19, 2019

Notes

Topics

Notes

Objectives

Objectives:

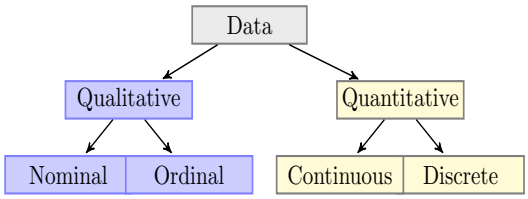
- Distinguish between quantitative and qualitative data. Distinguish between discrete and continuous (quantitative) variables. Distinguish between nominal and ordinal (qualitative) variables.
- Identify common types of selection bias in sampling.
- State how non-response can lead to biased survey results.

Notes

Introduction (1.1, 1.2)

- ***Statistics*** is the science of collecting, organizing, analyzing, and drawing conclusions from *data*.
- ***Data*** are observations (measurements) of a *variable*.
- A ***variable*** is a characteristic that varies from one individual to the next (e.g. height, age, income etc.).

Notes



Notes

- Variables can be either
  1. ***Quantitative*** (also called ***numerical***), i.e. possible values of the variable are numerical values, or
  2. ***Qualitative*** (also called ***categorical***), i.e. possible values of the variable are categories.

Notes

- Quantitative variables can be either
  1. ***Discrete***, i.e. possible values of the variable are isolated numbers with gaps between them, e.g. the integer values 0, 1, 2, 3, ..., or
  2. ***Continuous***, i.e. the set of possible values of the variable is an entire interval, or continuum, e.g. *all* values greater than 0, such as 47.46231.

Notes

- Qualitative variables can be either
  1. ***Ordinal***, i.e. possible values of the variable have a specific ordering to them (e.g. small, medium, and large), or
  2. ***Nominal***, i.e. the possible values of the variable have no specific ordering (e.g. the colors red, green, and blue, etc.).

Notes

Exercise

Determine whether each of these variables is **quantitative** or **qualitative**. If it's *quantitative*, determine whether it's **discrete** or **continuous**. If it's *qualitative*, determine whether it's **ordinal** or **nominal**.

a) The makes (e.g. Ford, Toyota, etc.) of cars in a certain downtown parking garage.

b) The amount of time it takes your bus to travel between two stops from day to day.

c) The numbers of flowers on plants after three months growth.

d) The heights of plants after three months growth.

Notes

e) The weights of babies born at a certain hospital.

f) The number of phone calls received by a store's customer service department from day to day.

g) The condition (good, fair, serious, critical) of patients admitted to a hospital.

h) The number of college credits students in this class have acquired.

i) The colors (blue, orange, red, etc.) of the shirts worn by students in this class.

Notes

- A ***population*** is an entire group of individuals about which we seek information.
- A ***sample*** is a subset of the population that's selected in some prescribed manner (usually randomly).
- Two uses of statistics are
  1. To describe (or summarize) a set of data (***descriptive statistics***).
  2. To draw conclusions about a population based on data in a sample (***inferential statistics***).

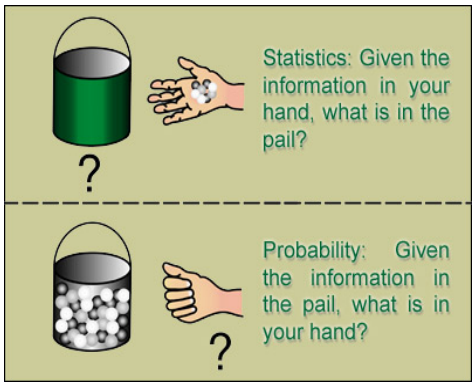
Notes

- Descriptive statistics can be used with any set of data, even if they aren't a random sample.
- Inferential statistics requires that the data be from either:
  1. A random sample (so that they'll be representative of the population), or
  2. A randomized experiment (so that we can draw conclusions about the effect of a treatment).

Notes

- This course has three main broad sections:
  1. Collecting data, summarizing data, and graphing data.
  2. Probability (which is used to quantify uncertainty).
  3. Inferential statistics (which uses probability to express uncertainty in the conclusions that we draw about a population).

Notes



Notes

Collecting Data: Sampling (1.2, 1.3)

Sampling versus Census Taking

- In a ***census***, data are obtained from *every* individual in the population (e.g. the U.S Census).
- Taking a census is time consuming and expensive. Instead we usually take a *sample* from the population, and then make generalizations (i.e. *statistical inferences*) about the population based on our sample.

Notes

Biased and Unbiased Sampling Schemes

- In order to generalize from sample data to the population, we'd like our sample to be *representative* of the population.

Notes

Notes

Some Biased Sampling Schemes

- A data collection method that systematically produces data that are *not* representative of the population is called a ***biased*** data collection method.

Notes

- Some examples of bias in sampling include:
  - ***Selection bias*** - Certain groups are systematically underrepresented in the sample (and others overrepresented). Examples include:
    - ***Convenience samples*** - The sample is selected in a way that's convenient, but may favor certain groups of the population (and exclude others). An example is standing on a street corner surveying passersby.
    - ***Voluntary response samples*** (also called ***self-selected samples***) - Sample selection is by "open invitation". Individuals decide for themselves whether or not to be included, and those who do may differ systematically from those who don't. Examples include online surveys conducted from websites, where individuals who have a strong interest in the survey's subject matter are more likely to respond than those who don't.

Notes

- Examples of bias in sampling (cont'd):
  - ***Nonresponse bias*** - One or more survey questions are skipped by some of the individuals surveyed (***item non-response***) or some surveys aren't returned at all (***unit non-response***). The individuals who respond may differ systematically from those who don't respond. Examples include:
    - Mailed surveys for which the return postage isn't prepaid. Low income people won't respond.
    - Phone surveys that take too long. People who are very busy don't participate or they hang up before the survey is completed.
    - Surveys that have questions about illegal activities (e.g. drug use) or other sensitive topics (e.g. sex) that some people don't answer.

Notes

Avoiding Selection Bias Via Simple Random Sampling

- **We can avoid selection bias** by taking a ***simple random sample*** (or ***SRS***) of size n, i.e. a sample chosen in such a way that every possible size-n subset of the population has the same chance of being selected.
- To take a SRS, either:
  1. Place the names of individuals from the population in a hat, and draw n of them without looking, or
  2. Create a list (called a ***sampling frame***) of all the individuals in the population, and then use a computer random sample generator to select n individuals from the list.

- Most of the statistical inference procedures we'll look at were developed for use with SRSs.

  The procedures that are used to *compare* two groups can also be used with randomized experiments.

Notes

Other Sampling Schemes

- Sometimes, though, other sampling schemes are used. These include:
  - ***Systematic random sampling*** - A sampling frame (i.e. list of all individuals from the population) is created, and then every kth individual from the list is selected, starting from a randomly selected individual among the first k on the list. Using a larger value of k will result in a smaller sample size.
  - ***Stratified random sampling*** - The population is first split into subpopulations called ***strata*** (e.g. based on age group, gender, or income bracket), and then a separate SRS is selected from each stratum. Care should be taken not to *oversample* from any particular stratum.

Notes

Notes

Notes