Notes on Chapter 1: Data Collection (Sullivan text)

1.1 Introduction to practice of statistics.

• statistics — science of collecting, organizing, summarizing and analyzing data to draw conclusions or answer questions along with a measure of confidence in the conclusions and answers

• population — the complete set of individuals or objects under study or investigation
• sample — a subset of the population

• parameter — a numerical characteristic of a population such as a population mean, a population proportion, etc.
• statistic — a numerical summary of a sample

• descriptive statistics — organizing and summarizing data using tables, graphs, and numerical summaries
• inferential statistics — using sample data to draw conclusions or make inferences about a population along with a measure of reliability of the inference; includes confidence intervals and hypothesis testing (for population parameters)

• Process of statistics: 1. Identify research objective
  2. Collect data
  3. Describe data
  4. Perform inference

• qualitative variable — a variable which classifies data into categories or attributes; e.g., hair color, religious preference, gender, etc.
• quantitative variable — a numerical variable whose values allow meaningful arithmetic operations

• discrete variable — quantitative variable with a countable number of possible values
• continuous variable — quantitative variable with an uncountable number of possible values

• Levels of measurement: nominal level — categorical with no natural ranking
  ordinal level — can be ordered or naturally ranked at best
  interval level — numerical; allows meaningful subtraction
  ratio level — numerical with meaningful ratios of the data
1.2 Observational studies versus designed experiments

- explanatory (independent) variable — variable which "explains" another variable
- response (dependent) variable — variable whose value is related to value of the explanatory variable

- observational study — study in which researcher does not control the explanatory variable
- designed experiment — study in which researcher sets values for explanatory variable(s)

- confounding — occurs when the effect of an explanatory on the response variable cannot be discerned due to presence of another explanatory variable

- lurking variable — an explanatory variable not included in a study

- some types of observational studies: cross-sectional, case-control (retrospective), cohort (prospective)

- census — data collected on the whole population

1.3 Simple random variable

- random sampling — obtaining a sample by using chance to select individuals from a population
- simple random sample of size $n$ — every possible sample of size $n$ has an equal chance of occurring

- sample without replacement — once an individual is selected for sample, individual cannot be chosen again
- sample with replacement — once an individual is selected for sample, individual remains in population to be possibly selected again

1.4 Other sampling methods

- stratified sample — population classified into strata, then simple random sample taken from each stratum

- systematic sample — selecting every $k$th individual from the population; frame not needed.

- cluster sample — selecting all the individuals in one or more randomly selected clusters (group)

- convenience sample — easily obtained sample not based on randomness

- multistage sampling — a combination of different sampling methods arranged in stages
1.5 Bias in sampling

- bias — favoritism; sample is not representative of the population
- sampling bias — sample tends to favor one part of the population over the others, possibly from undercoverage due to incomplete frame
- nonresponse bias
- response bias — due to interviewer error, misrepresented answers, wording of questions, order of questions or words, type of question, or data-entry error

- above biases lead to nonsampling errors
- sampling error is the natural error that occurs when using estimates from the sample to make inferences about the population

1.6 The design of experiments

- An experiment is a controlled study conducted to determine the effect varying one or more explanatory variables (factors) has on a response variable. Any combination of the values of the factors is called a treatment.

- experimental unit — person (subject), object, or well-defined item upon which a treatment is applied
- control group — a baseline treatment group
- placebo — innocuous medication (treatment)

- single-blind experiment — subject unaware of treatment he or she is receiving
- double-blind experiment — both the subject and the person recording responses is unaware of treatment given

- Steps in designed experiment
  1. Identify problem
  2. Determine factors that may affect response variable
  3. Determine number of experimental units
  4. Determine the level of each factor
  5. Conduct the experiment
  6. Conduct the experiment

- completely randomized design — each experimental unit is randomly assigned to a treatment
- match-pairs design — experimental units are naturally paired (before/after, twins, etc.); two treatments involved